Saarland University
Faculty of Mathematics and Computer Science
Department of Mathematics

High-Dimensional Analysis: Random Matrices and Machine Learning

Lecture notes

Summer Term 2023

(version 1 from August, 2023)

Prof. Dr. Roland Speicher
with the support of
Dr. Johannes Hoffmann

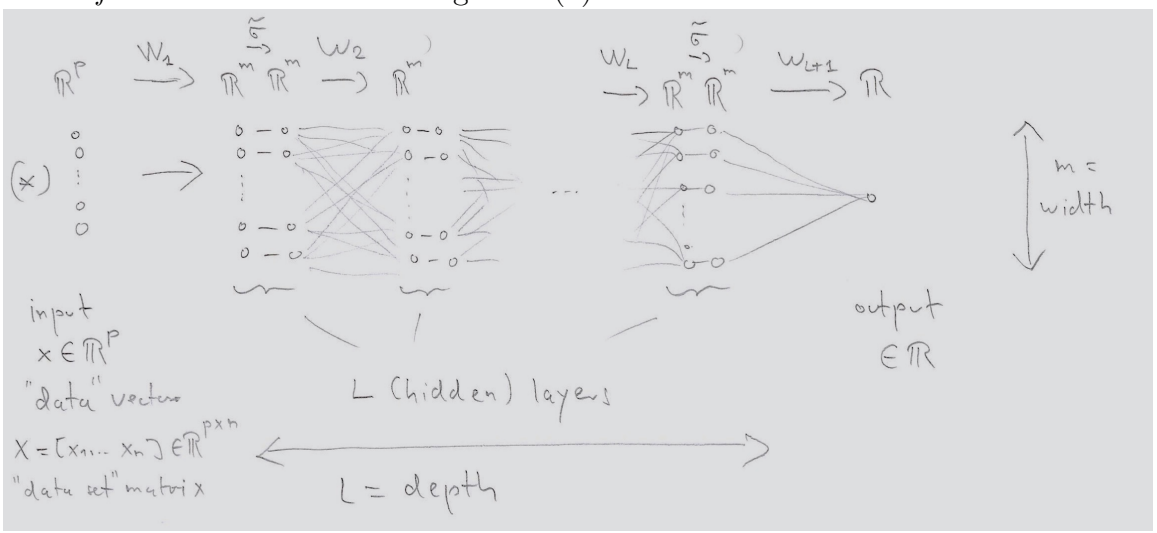
Contents

Introduction and Survey

The goal of this lecture series is to cover mathematical interesting aspects of neural networks, in particular, those related to random matrices. As the lecturer knows more about random matrices than about neural networks, the whole presentation is surely biased towards the former ones.

Neural networks

A neural network (more precisely, a fully connected feed-forward neural network) is a function $f : \mathbb{R}^p \to \mathbb{R}$ of the following form $(*)$



It is determined by

- a, in general non-linear, function $\sigma : \mathbb{R} \to \mathbb{R}$, which is usually fixed; this map σ is component-wise applied to vectors or matrices - in the above, we denoted it then by $\tilde{\sigma}$, but later we will just omit this distinction; a prominent example for such a σ is $\sigma = \mathrm{ReLu}$, which is just
$$\mathrm{ReLu}(t) = \max(0, t)$$
- all matrices $W_1, \dots, W_{L+1}$; those are the *network parameters* and they have to be chosen in the right way via training

The problem in this context is: our data set is high-dimensional – i.e., p is large – and we would like to characterize or distinguish data sets corresponding to different classes of objects. For example, $x \in \mathbb{R}^p$ can be a vector of a high-resolution picture and we have picture of cats and pictures of dogs,

$$\mathbb{R}^p \supset C \cup D = \{\text{pictures of cats}\} \cup \{\text{pictures of dogs}\},$$

and we would like to distinguish them. So we would like to have a function

$$f : \mathbb{R}^p \to \mathbb{R}, \qquad \text{such that} \qquad f(x) = \begin{cases} 1, & x \in C \\ 0, & x \in D \end{cases}$$

But the problem is that we have no idea how to define f directly, we can only hope to define it by what it is supposed to do, via an ansatz of the form $(*)$ and then find the parameters via training. That this really seems to work quite well, is a big mystery!

Consider our mappings

$$\mathbb{R}^p \overset{\sigma\circ W_1}{\to} \mathbb{R}^m \to \ldots \overset{\sigma\circ W_L}{\to} \mathbb{R}^m \overset{W_{L+1}}{\to} \mathbb{R}$$

The vector in the last hidden layer $\mathbb{R}^m$ is usually called the vector of *features*; and we can consider the mapping up to this point as an embedding of $\mathbb{R}^p$ into $\mathbb{R}^m$, which is often given by a kernel. In the last few years, the limit for $m \to \infty$ of this kernel, the so-called *neural tangent kernel*, has become quite prominent as a first (kind of linear) approximation for this embedding. If this kernel is fixed, and only W_{L+1} is learned, then this is a linear optimization problem, which is understood quite well. However, *feature learning* seems to be crucial for the success of neural networks. At the moment the mathematical description of this is still quite unclear.

Random vectors and random matrices

Our preferred way to model our high-dimensional data is by probability distributions on $\mathbb{R}^p$. The main distribution where we can really calculate something are Gaussian distributions

$$x = (x^{(1)}, \ldots, x^{(p)})^T \in \mathbb{R}^p$$

which depend on some mean vector

$$\mu = (E[x^{(1)}], \ldots, E[x^{(p)}])^T \in \mathbb{R}^p$$

and some covariance matrix

$$\Sigma = \left(E[x^{(i)}x^{(j)}] - E[x^{(i)}]E[x^{(j)}]\right)_{i,j=1}^{p} \in \mathbb{R}^{p\times p}.$$

Note that assuming a Gaussian distribution for real data is quite unrealistic. However: we can, for many statement, also compose Gaussian distributions with Lipschitz functions, which yields then the much more general class of concentrated random vectors. Those seem to be, in many respects, good models for real data; in particular, GANs (generative adversarial networks) are given in this way.

Our knowledge about the data is given by measurements or observations. Assume we have n observations of our p-dimensional data: $x_1, \ldots, x_n \in \mathbb{R}^p$. Then, assuming that the mean is zero, the canonical estimator for our "true" covariance Σ is given by the sample covariance matrix

$$\hat{\Sigma} := \frac{1}{n}\sum_{k=1}^{n} \underbrace{\left(x_k^{(i)}x_k^{(j)}\right)_{i,j=1}^{p}}_{x_k x_k^T} = \frac{1}{n}XX^T,$$

where $X = [x_1 \ldots x_n] \in \mathbb{R}^{p\times n}$ is the *data set matrix*.

In classical statistics, we fix p and let $n \to \infty$ and then we have that $\hat{\Sigma}$ converges to Σ. But now, in our modern setting, the size p of the data and the number n of observations are of the same order, $p \sim n$. This is a different regime than the classical one - more complicated, but still controllable.

In our neural network $(*)$ we can make precise asymptotic statements for the limit where the dimensions go to ∞, but we have to consider the regime that all sizes are of the same order, i.e., $p \sim n \sim m \to \infty$. The depth L is usually fixed in those investigations. Thus we will mainly talk about *wide networks*. The role of the depth L is not so clear, in particular, at least in this lecture notes we will not consider the deep limit $L \to \infty$.

A crucial ingredient for all our investigations and statements will be concentration phenomena in high dimensions. This is one side (the good one) of the two sides of working in high dimensions:

- We have the "curse of high dimensionality":
  high dimensional spaces are quite large and empty; sample sets in high dimensions are very thin; and usually one has bad convergence of estimators; in particular, it is impossible to sample the density of the distribution.

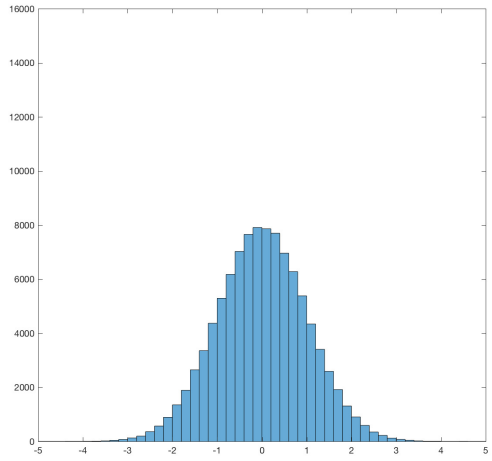
- But there is also a "blessing of high dimensionality", aka *concentration phenomena in high dimensions*:
  many random vectors and random matrices show in high dimensions a (close to) deterministic behaviour; and it is surprisingly simple to sample smooth 1-dimensional functions of the high-dimensional vectors.

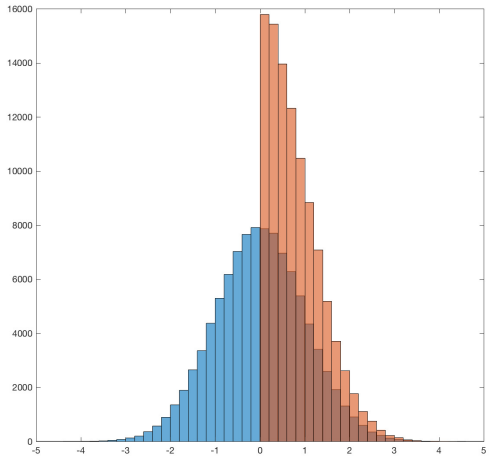
Let us consider a few numerical instances of such concentration behaviours.

Concentration of the norm of random vectors

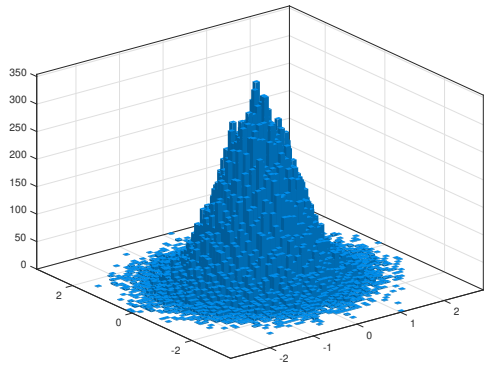
The following histograms show that Gaussian random vectors concentrate on the surface of the unit ball. Note that we have normalized the Gaussian random vectors such that in all dimensions the expectation of the square of their length is equal to 1.

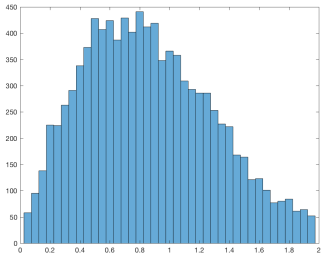
- distribution of a one-dimensional Gaussian vector $x \in \mathbb{R}^1$ and of its length $\|x\| \in \mathbb{R}$, with 100.000 samples



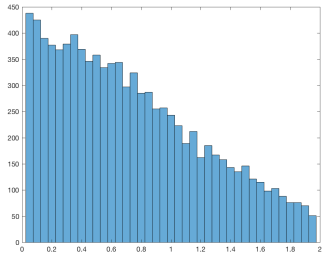


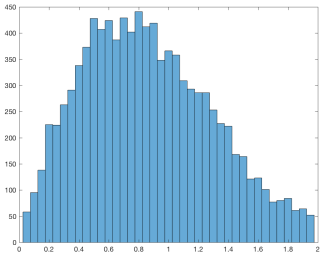
- distribution of a two-dimensional Gaussian vector $x \in \mathbb{R}^2$ and of its length $\|x\| \in \mathbb{R}$, with 100.000 samples

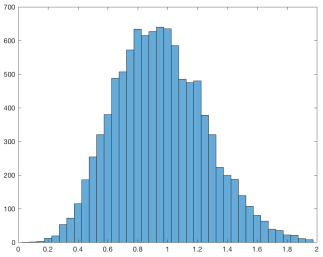


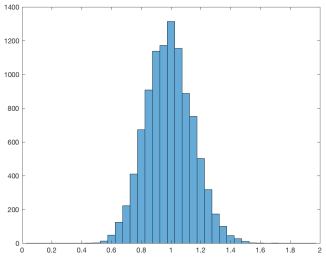


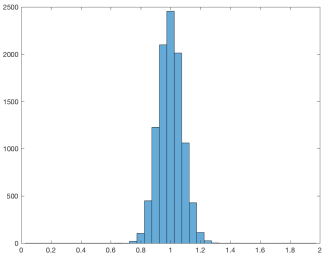
- the length $\|x\| \in \mathbb{R}$ of a Gaussian vector $x \in \mathbb{R}^p$, for $p = 1, 2, 5, 20, 80, 320$; with 10.000 samples in each case

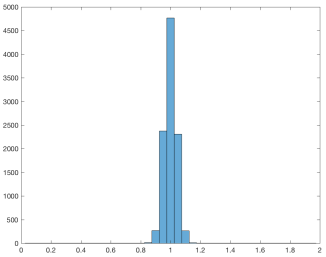






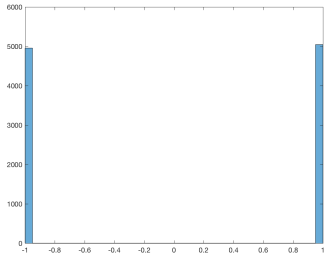
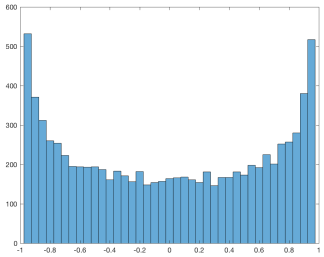
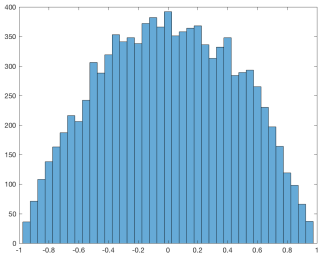
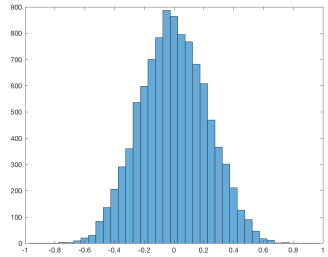
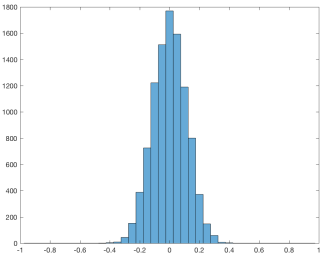
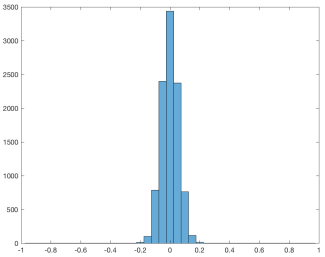






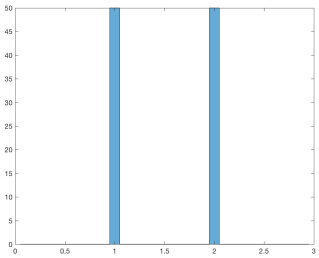
Two Gaussian random vectors are nearly orthogonal in high dimensions

For 10.000 samples of pairs (x_1, x_2) of independent Gaussian vectors $x_1, x_2 \in \mathbb{R}^p$, the histogram of the normalized inner product $\frac{\langle x_1, x_2 \rangle}{\|x_1\| \cdot \|x_2\|}$ is shown for $p = 1, 2, 5, 20, 80, 320$. This shows that two such vectors are close to orthogonal in high dimensions. Note that 0 is, also in small dimensions, the expectation of the normalized inner product.

Estimation of covariance matrix via sampling

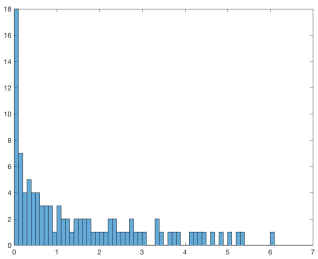
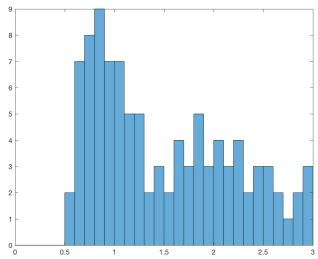
Consider Gaussian random vectors in $\mathbb{R}^p$ with independent components, and such that half of the components have variance 1, and the other half have variance 2. This means that the covariance matrix Σ has $p/2$ eigenvalues 1 und $p/2$ eigenvalues 2 and thus the eigenvalues of Σ look like this

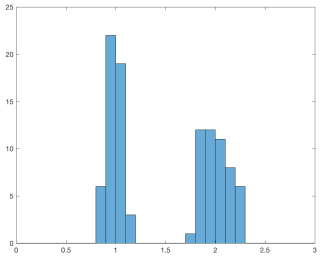
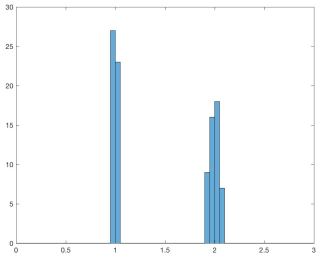


We observe now n samples $x_1, \ldots, x_n \in \mathbb{R}^p$ and plot the p eigenvalues of the sample covariance matrix

$$\hat{\Sigma} := \frac{1}{n} \sum_{k=1}^{n} x_k x_k^T.$$

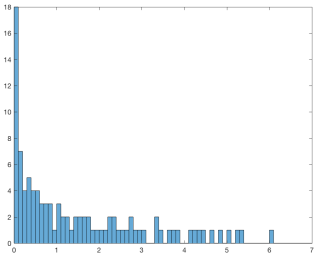
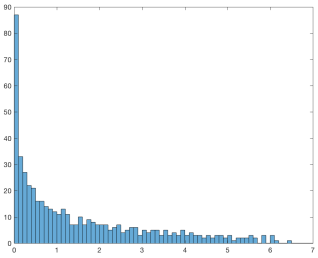
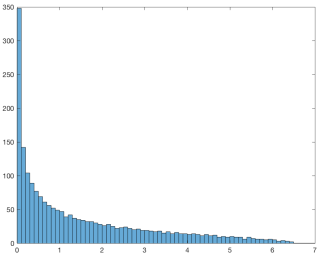
First, we consider the classical regime: we fix $p = 100$ and increase n as $n = 100$, $n = 1.000$, $n = 10.000$, $n = 100.000$

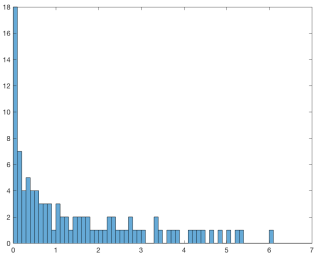



This shows the convergence of $\hat{\Sigma}$ to Σ in the classical regime.

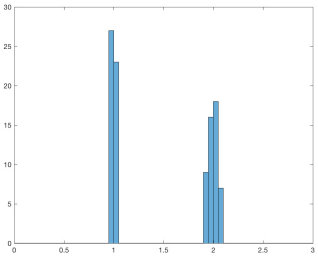
Next, we look at the modern regime, where n should scale proportional to p. We plot the eigenvalues of $\hat{\Sigma}$ for $p = n = 100$, $p = n = 500$, and $p = n = 2000$

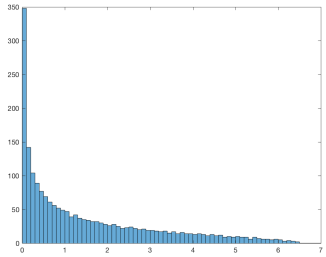
Assume now we are given a concrete situation $p = n = 100$, where we cannot change p or n, but just have to work with this specific instance. What can we get out of this specific histogram. According to the above we have two different asymptotics as possible approximations for our concrete situation $p = n = 100$? Which is more appropriate? Is



- like $p = 100$, $n \to \infty$



- or like $p = n$, $n \to \infty$



The first regime is clearly not appropriate! We need very large n to identify Σ; $n = 100$ is far from this limit; so the classical approach is useless here.

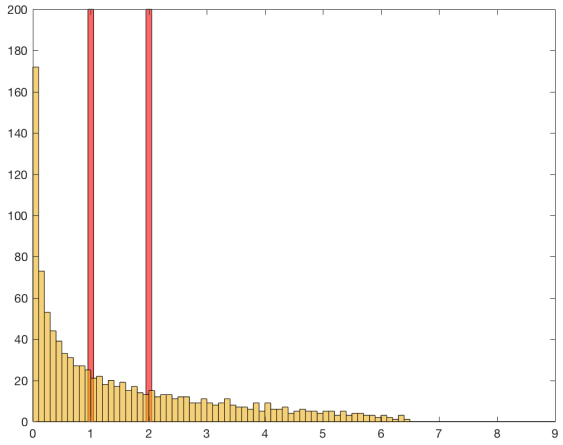
In the second regime, we already have for $n = 100$ a good approximation of the limit distribution; however, this is not given by Σ, but by a deterministic function of this. We will have to investigate the relation between Σ and $\hat{\Sigma}$ in this limit.

Signal-plus-noise models – can we observe the signal in the noise?

Consider Gaussian random vectors in $\mathbb{R}^{1000}$

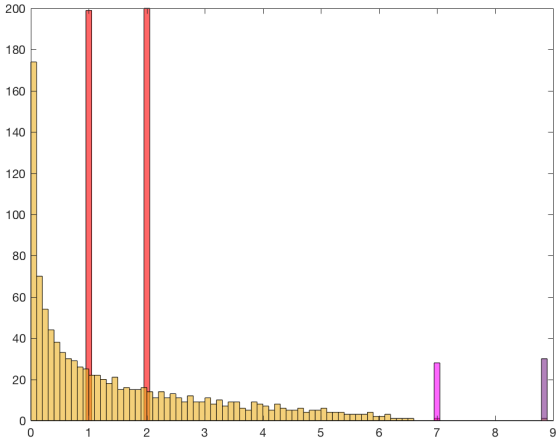
- with independent components,
- one component has "large" variance μ
- 499 of them have variance 1
- 500 of them have variance 2

If we put $\mu = 1$ then we are back to the situation before, with $p = n = 1.000$; we overlay in the following plot the histograms of the eigenvalues of Σ and of $\hat{\Sigma}$.

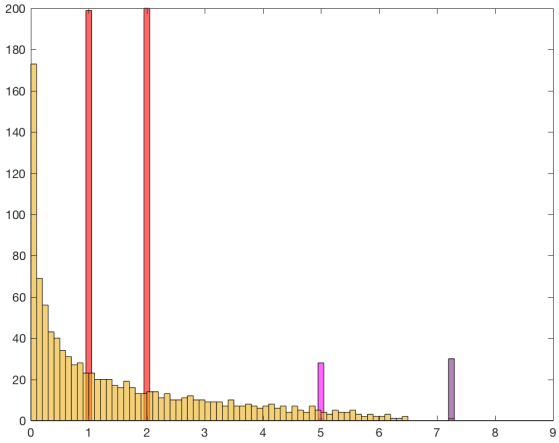


If we now choose $\mu = 7$, then Σ will have one eigenvalue at 7. We think of this eigenvalue as corresponding to some relevant information ("signal") in Σ, whereas the other eigenvalues are representing noise. Will we see a shadow of this signal eigenvalue in the spectrum of $\hat{\Sigma}$? Indeed, there is a clear eigenvalue λ corresponding to this - but this is not at 7, but somewhere around 9. Note that we have enlarged the contribution of the signal

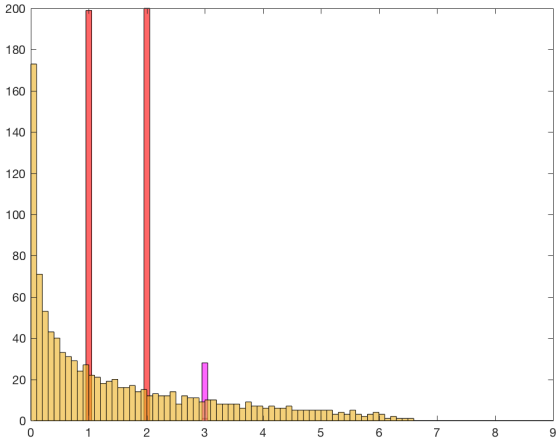
eigenvalue μ and its shadow λ in the histogram, to make them visible. There is always only one eigenvalue sitting at those positions.



If we decrease μ to 5, then we still see a corresponding eigenvalue λ in $\hat{\Sigma}$, sitting a bit above 7.



But when μ decreases to 3, its shadow in $\hat{\Sigma}$ will be swallowed by the bulk eigenvalues and is not visible any more



We will investigate the relation between the position μ of the signal eigenvalue in Σ and the position of the corresponding outlier λ in the eigenvalues of $\hat{\Sigma}$. In particular, we will see that one has tools for giving explicit formulas for the relation between them.

What else: double descent, non-linear random matrices, neural tangent kernel

In the following lectures we will make the above observations more precise. Understanding the by now quite well-established theories of Gaussian random vectors and random matrices in high dimensions is a main focus, but will also be complemented by more recent topics like "double descent", "non-linear random matrix models" or "neural tangent kernel". For the former topics we have benefited quite a bit from "standard" literature [CL22, RYH22, Ver18, Wai19] as well as from discussions with and lecture notes from Boris Hanin.

1. Volumes of Hyperballs and Hypercubes in High Dimensions

We want to understand functions on sets $A \subset \mathbb{R}^p$, for p large, in terms of their typical and averaged behavior. We describe the data sets by probability distributions in $\mathbb{R}^p$; the most basic ones are uniform distributions on sets A, given by the volume, or (later) by Gaussian distributions. For $A \subset \mathbb{R}^p$ we consider its volume (Lebesgue measure)

$$\mathrm{vol}(A) = \int_A \mathrm{d}x = \int_{\mathbb{R}^p} 1_A(x)\,\mathrm{d}x = \int\limits_{(t_1,\ldots,t_p)\in A} \ldots \int \mathrm{d}t_1 \ldots \mathrm{d}t_p.$$

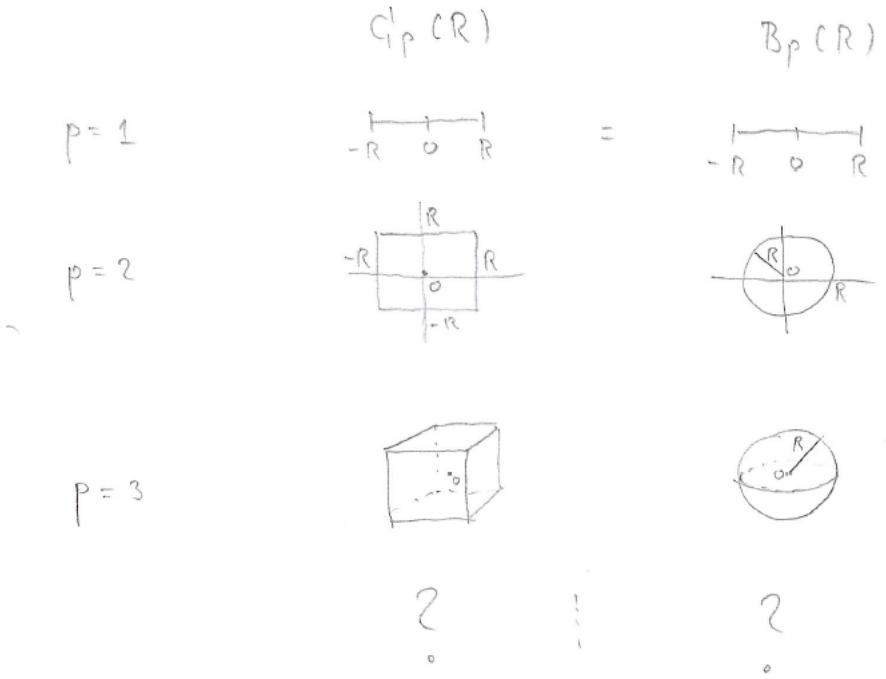
We consider in particular

$$\mathrm{B}_p(R) := \{x = (t_1, \ldots, t_p) \in \mathbb{R}^p : \underbrace{t_1^2 + \cdots + t_p^2}_{=:\|x\|^2} \leq R^2\},$$

the *p-dimensional (hyper)ball of radius R*, and

$$\mathrm{C}_p(R) := \underbrace{[-R, R] \times \ldots \times [-R, R]}_{p \text{ times}} = \{x = (t_1, \ldots, t_p) \in \mathbb{R}^p : |t_i| \leq R \ \forall i = 1, \ldots, p\},$$

the *p-dimensional hypercube with side length $2R$.*



1.1. Formula for volumes of the balls in any dimension

We clearly have for all p

$$\mathrm{vol}(\mathrm{C}_p(R)) = (2R)^p, \qquad \text{so in particular} \qquad \mathrm{vol}\left(\mathrm{C}_p\left(1/2\right)\right) = 1.$$

How about the volumes of hyperballs? We have

$$\begin{aligned}
\operatorname{vol}(\mathrm{B}_p(R)) &= \int\limits_{t_1^2+\ldots+t_p^2\le R} \ldots \int \mathrm{d}t_1 \ldots \mathrm{d}t_p \\
&= \int\limits_{(Rs_1)^2+\ldots+(Rs_p)^2\le R} \ldots \int \mathrm{d}(Rs_1) \ldots \mathrm{d}(Rs_p) \\
&= R^p \int\limits_{s_1^2+\ldots+s_p^2\le 1} \ldots \int \mathrm{d}s_1 \ldots \mathrm{d}s_p \\
&= R^p \operatorname{vol}(\mathrm{B}_p(1)).
\end{aligned}$$

Put $\mathrm{B}_p := \mathrm{B}_p(1)$, then

$$\operatorname{vol}(\mathrm{B}_1) = \int_{-1}^{1} \mathrm{d}t_1 = 2$$

and

$$\operatorname{vol}(\mathrm{B}_2) = \iint_{t_1^2+t_2^2\le 1} \mathrm{d}t_1\,\mathrm{d}t_2 = \int_{-1}^{+1} \int_{-\sqrt{1-t_1^2}}^{+\sqrt{1-t_1^2}} \mathrm{d}t_2\,\mathrm{d}t_1 = 2\int_{-1}^{+1} \sqrt{1-t^2}\,\mathrm{d}t = \ldots = \pi$$

by substitution. This works better in polar coordinates $t_1 = r\cos(\varphi)$ and $t_2 = r\sin(\varphi)$:

$$\iint_{t_1^2+t_2^2\le 1} \mathrm{d}t_1\,\mathrm{d}t_2 = \int_0^{2\pi} \int_0^1 r\,\mathrm{d}r\,\mathrm{d}\varphi = 2\pi \left[\frac{1}{2}r^2\right]_0^1 = 2\pi\frac{1}{2} = \pi.$$

For general p, we can derive a recursion by integrating out two variables in polar coordinates:

$$\begin{aligned}
\operatorname{vol}(\mathrm{B}_p) &= \int\limits_{t_1^2+\ldots+t_p^2\le 1} \ldots \int \mathrm{d}t_1 \ldots \mathrm{d}t_p \\
&= \int\limits_{t_1^2+\ldots+t_{p-2}^2+r^2\sin^2(\varphi)+r^2\cos^2(\varphi)\le 1} \ldots \int \mathrm{d}t_1 \ldots \mathrm{d}t_{p-2} r\,\mathrm{d}r\,\mathrm{d}\varphi \\
&= \int_0^{2\pi} \int_0^1 \left(\int\limits_{t_1^2+\ldots+t_{p-2}^2\le 1-r^2} \ldots \int \mathrm{d}t_1 \ldots \mathrm{d}t_{p-2} \right) r\,\mathrm{d}r\,\mathrm{d}\varphi \\
&= \int_0^{2\pi} \int_0^1 \operatorname{vol}(\mathrm{B}_{p-2}(\sqrt{1-r^2}))r\,\mathrm{d}r\,\mathrm{d}\varphi \\
&= \int_0^{2\pi} \int_0^1 \operatorname{vol}(\mathrm{B}_{p-2})(1-r^2)^{\frac{p-2}{2}} r\,\mathrm{d}r\,\mathrm{d}\varphi \\
&= \operatorname{vol}(\mathrm{B}_{p-2}) \cdot 2\pi \cdot \int_0^1 r(1-r^2)^{\frac{p-2}{2}}\,\mathrm{d}r \\
&= \operatorname{vol}(\mathrm{B}_{p-2}) \cdot 2\pi \cdot \underbrace{\left[(1-r^2)^{\frac{p}{2}} \left(-\frac{1}{2}\cdot\frac{2}{p}\right)\right]_0^1}_{=\frac{1}{p}} \\
&= \frac{2\pi}{p} \operatorname{vol}(\mathrm{B}_{p-2}).
\end{aligned}$$

This yields then for example the well-known volume of the three-dimensional ball

$$\mathrm{vol}(\mathrm{B}_3) = \frac{2\pi}{3}\underbrace{\mathrm{vol}(\mathrm{B}_1)}_{=2} = \frac{4\pi}{3}.$$

Iterating the recursion for even $p = 2k$ gives

$$\mathrm{vol}(\mathrm{B}_{2k}) = \frac{2\pi}{2k}\cdot\mathrm{vol}(\mathrm{B}_{2(k-1)}) = \frac{2\pi}{2k}\cdot\frac{2\pi}{2(k-1)}\cdot\mathrm{vol}(\mathrm{B}_{2(k-2)}) = \ldots = \frac{\pi^k}{k!}\cdot\mathrm{vol}(\mathrm{B}_0),$$

where $\pi = \mathrm{vol}(\mathrm{B}_2) = \frac{2\pi}{2}\,\mathrm{vol}(\mathrm{B}_0)$ yields $\mathrm{vol}(\mathrm{B}_0) = 1$. Thus for $p = 2k$

$$\mathrm{vol}(\mathrm{B}_{2k}) = \frac{\pi^k}{k!} = \frac{\pi^{\frac{p}{2}}}{\left(\frac{p}{2}\right)!}$$

For odd $p = 2k+1$, on the other hand, we can iterate down to one dimension:

$$\begin{aligned}\mathrm{vol}(\mathrm{B}_{2k+1}) &= \frac{2\pi}{2k+1}\cdot\frac{2\pi}{2k-1}\cdot\ldots\cdot\frac{2\pi}{3}\cdot\underbrace{\mathrm{vol}(\mathrm{B}_1)}_{=2}\\ &= \frac{\pi^k}{\left(k+\frac{1}{2}\right)\cdot\left(k-\frac{1}{2}\right)\cdot\ldots\cdot\frac{3}{2}\cdot\frac{1}{2}}\\ &= \frac{\pi^{\frac{p}{2}}}{\frac{p}{2}\cdot\left(\frac{p}{2}-1\right)\cdot\ldots\cdot\frac{3}{2}\cdot\frac{1}{2}\cdot\pi^{\frac{1}{2}}}.\end{aligned}$$

We can combine the two cases in a common formula for general p

$$\mathrm{vol}(\mathrm{B}_p) = \frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}+1\right)},$$

where Γ is Euler's gamma function, i.e.

$$\Gamma(k) = (k-1)! \quad\text{and}\quad \Gamma\left(k+\frac{1}{2}\right) = \left(k-\frac{1}{2}\right)\cdot\left(k-\frac{3}{2}\right)\cdot\ldots\cdot\frac{3}{2}\cdot\frac{1}{2}\cdot\sqrt{\pi}$$

and in general (for $s \in (0,\infty)$)

$$\Gamma(s) = \int_0^\infty t^{s-1}\exp(-t)\,\mathrm{d}t.$$

Stirling's approximation for the factorial

$$\Gamma(s+1) \sim \sqrt{2\pi s}\left(\frac{s}{\mathrm{e}}\right)^s \quad \text{for large } s$$

gives for large p the approximation

$$\mathrm{vol}(\mathrm{B}_p) = \frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}+1\right)} \sim \frac{\pi^{\frac{p}{2}}}{\sqrt{2\pi\frac{p}{2}}\left(\frac{p}{2\mathrm{e}}\right)^{\frac{p}{2}}} \sim \frac{1}{\sqrt{p\pi}}\cdot\left(\frac{2\pi\mathrm{e}}{p}\right)^{\frac{p}{2}}.$$

Theorem 1.1. The volume of the hyperball is given by

$$\mathrm{vol}(\mathrm{B}_p(R)) = \frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}+1\right)} \cdot R^p,$$

which behaves asymptotically as

$$\mathrm{vol}(\mathrm{B}_p(R)) \sim \frac{1}{\sqrt{p\pi}} \cdot \left(\frac{2\pi \mathrm{e} R^2}{p}\right)^{\frac{p}{2}} \quad \text{for} \quad p \to \infty.$$

Corollary 1.2. (1) For any fixed radius $R > 0$, the volume of $\mathrm{B}_p(R)$ goes to zero for high dimensions:

$$\lim_{p\to\infty} \mathrm{vol}(\mathrm{B}_p(R)) = 0.$$

(2) In order for $\mathrm{B}_p(R)$ to have volume 1, the radius has to scale with the dimension as $R \sim \sqrt{\frac{p}{2\pi \mathrm{e}}}$.

p	$\mathrm{vol}(\mathrm{B}_p)$	
1	2	
2	π	≈ 3.14
3	$\frac{4\pi}{3}$	≈ 4.19
4	$\frac{\pi^2}{2}$	≈ 4.93
5	$\frac{8\pi^2}{15}$	≈ 5.24
6	$\frac{\pi^3}{6}$	≈ 5.17
10	$\frac{\pi^5}{120}$	≈ 2.55
15	...	≈ 0.38

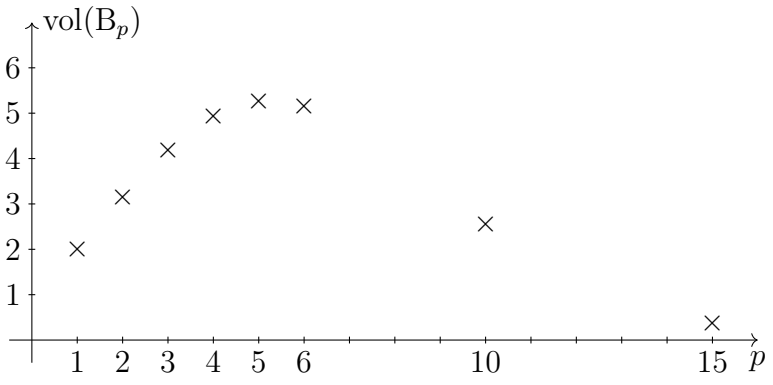


Figure 1: Comparison of the volumes of the hyperball in higher dimensions.

From the concrete formula for the volume one gets concentration phenomena for "random vectors" in the balls in high dimensions, according to the following slogans.

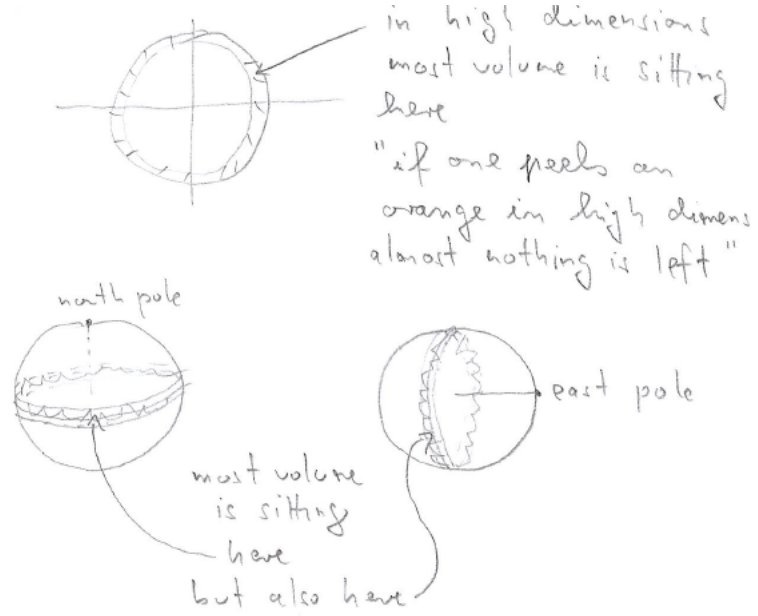
Slogans 1.3. In high dimensions:

(i) the volume of the ball is concentrated close to its surface;

(ii) almost all of the volume of the ball lies near its equator (note: there are many equators and this holds for all of them!);

(iii) any two vectors in the ball are almost orthogonal.

In small dimensions these statements are clearly not true.



In the following we will make those slogans mathematically precise and prove them. We will talk about probabilities P of events; here the probability is measured by the volume of the sets having a considered property, normalized by the volume of the unit ball.

1.2. The volume of the ball is concentrated close to its surface

Fix $\varepsilon > 0$, then we have for the random vectors $x \in \mathrm{B}_p$ that

$$\begin{aligned}\mathrm{P}(\|x\| > 1-\varepsilon) &= \frac{\mathrm{vol}(\{x \in \mathrm{B}_p : \|x\| > 1-\varepsilon\})}{\mathrm{vol}(\mathrm{B}_p)} \\ &= \frac{\mathrm{vol}(\mathrm{B}_p) - \mathrm{vol}(\{x \in \mathrm{B}_p : \|x\| \leq 1-\varepsilon\})}{\mathrm{vol}(\mathrm{B}_p)} \\ &= 1 - \frac{\mathrm{vol}(\mathrm{B}_p(1-\varepsilon))}{\mathrm{vol}(\mathrm{B}_p(1))}\end{aligned}$$

and thus

$$\mathrm{P}(\|x\| > 1-\varepsilon) = 1 - (1-\varepsilon)^p \xrightarrow{p\to\infty} 1 \quad \text{if } \varepsilon \text{ is fixed.}$$

If one wants the volume in the "skin" constant, one has to scale ε with p like $\varepsilon = \frac{1}{p}$:

$$\mathrm{P}\left(\|x\| > 1 - \frac{1}{p}\right) = 1 - \underbrace{\left(1 - \frac{1}{p}\right)^p}_{\approx \frac{1}{\mathrm{e}} \text{ for } p \text{ large}} \approx 1 - \frac{1}{\mathrm{e}} \approx 62\%.$$

Often, one likes to write the concentration via exponential estimates in the dimension. For this one has the following estimate.

Lemma 1.4. For $p \geq 1$ and $0 < \varepsilon \leq 1$ we have

$$(1-\varepsilon)^p \leq \exp(-\varepsilon p).$$

Proof. This follows from the obvious case $p = 1$

$$0 \leq 1 - \varepsilon \leq \exp(-\varepsilon)$$

by taking the p-th power. □

Then we can write our concentration of the norm estimate as follows.

Theorem 1.5. For $p \geq 1$ and $0 < \varepsilon \leq 1$ we have

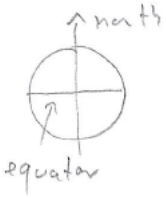
$$\mathrm{P}(x \in \mathrm{B}_p : \|x\| > 1 - \varepsilon) = 1 - (1-\varepsilon)^p \geq 1 - \exp(-\varepsilon p)$$

and

$$\mathrm{P}\left(x \in \mathrm{B}_p : \|x\| > 1 - \frac{1}{p}\right) \geq 1 - \frac{1}{\mathrm{e}}.$$

1.3. Almost all of the volume of the ball lies near its equator

Consider $x = (t_1, \ldots, t_p) \in \mathrm{B}_p$ and choose (arbitrarily) t_p as the north direction.



Then the equator $\{x \in \mathrm{B}_p \mid t_p = 0\}$ is a $(p-1)$-dimensional region and the probability of being close to the equator is given by

$$\begin{aligned}
\mathrm{P}(|t_p| \leq \varepsilon) &= \frac{\mathrm{vol}(\{x \in \mathrm{B}_p \mid t_p \in (-\varepsilon, \varepsilon)\})}{\mathrm{vol}(\mathrm{B}_p)} \\
&= \frac{1}{\mathrm{vol}(\mathrm{B}_p)} \int_{-\varepsilon}^{\varepsilon} \int \cdots \int_{t_1^2 + \ldots + t_{p-1}^2 \leq 1 - t^2} \mathrm{d}t_1 \ldots \mathrm{d}t_{p-1}\, \mathrm{d}t \\
&= \frac{1}{\mathrm{vol}(\mathrm{B}_p)} \int_{-\varepsilon}^{+\varepsilon} \mathrm{vol}(\mathrm{B}_{p-1}(\sqrt{1-t^2}))\, \mathrm{d}t \\
&= \frac{\mathrm{vol}(B_{p-1})}{\mathrm{vol}(\mathrm{B}_p)} \int_{-\varepsilon}^{+\varepsilon} (1-t^2)^{\frac{p-1}{2}}\, \mathrm{d}t \\
&= \frac{\frac{\pi^{\frac{p-1}{2}}}{\Gamma\left(\frac{p-1}{2}+1\right)}}{\frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}+1\right)}} \int_{-\varepsilon}^{+\varepsilon} (1-t^2)^{\frac{p-1}{2}}\, \mathrm{d}t \\
&= \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{p}{2}+1\right)}{\Gamma\left(\frac{p-1}{2}+1\right)} \int_{-\varepsilon}^{+\varepsilon} (1-t^2)^{\frac{p-1}{2}}\, \mathrm{d}t \\
&\sim \frac{1}{\sqrt{\pi}} \frac{\sqrt{2\pi\frac{p}{2}}\left(\frac{p}{2\mathrm{e}}\right)^{\frac{p}{2}}}{\sqrt{2\pi\frac{p-1}{2}}\left(\frac{p-1}{2\mathrm{e}}\right)^{\frac{p-1}{2}}} \int_{-\varepsilon}^{+\varepsilon} (1-t^2)^{\frac{p-1}{2}}\, \mathrm{d}t \\
&\sim \frac{1}{\sqrt{\pi}} \sqrt{\frac{p}{p-1}} \left(\frac{p}{p-1}\right)^{\frac{p}{2}} \left(\frac{p-1}{2\mathrm{e}}\right)^{\frac{1}{2}} \int_{-\varepsilon}^{+\varepsilon} (1-t^2)^{\frac{p-1}{2}}\, \mathrm{d}t
\end{aligned}$$

$$
\begin{aligned}
&\sim \frac{1}{\sqrt{\pi}}\sqrt{\frac{p}{2\mathrm{e}}}\ \underbrace{\frac{1}{\left(1-\frac{1}{p}\right)^{\frac{p}{2}}}}_{\sim\frac{1}{\exp\left(-\frac{1}{2}\right)}=\sqrt{\mathrm{e}}}\ \int_{-\varepsilon}^{+\varepsilon}(1-t^2)^{\frac{p-1}{2}}\,\mathrm{d}t\\
&\sim \frac{1}{\sqrt{\pi}}\sqrt{\frac{p}{2}}\int_{-\varepsilon}^{+\varepsilon}(1-t^2)^{\frac{p-1}{2}}\,\mathrm{d}t\\
&\sim \frac{1}{\sqrt{\pi}}\sqrt{\frac{p}{2}}\int_{-\varepsilon}^{+\varepsilon}(1-t^2)^{\frac{p}{2}}\,\mathrm{d}t\\
&= \frac{1}{\sqrt{\pi}}\int_{-\varepsilon\sqrt{\frac{p}{2}}}^{+\varepsilon\sqrt{\frac{p}{2}}}\underbrace{\left(1-\frac{2s^2}{p}\right)^{\frac{p}{2}}}_{\sim\exp(-s^2)}\mathrm{d}s \qquad \left(\text{substituting } t=\frac{s}{\sqrt{\frac{p}{2}}}\right)\\
&\sim \frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty}\exp(-s^2)\,\mathrm{d}s = 1
\end{aligned}
$$

and thus for any $\varepsilon > 0$:

$$\lim_{p\to\infty}\mathrm{P}(|t_p|\le\varepsilon) = 1.$$

By more precise estimates we can also control the speed of concentration as

$$\mathrm{P}(x=(t_1,\dots,t_p)\in\mathrm{B}_p : |t_p|\le\varepsilon) \ge 1-\exp\left(-\frac{p-1}{2}\cdot\varepsilon^2\right)\cdot\sqrt{2\pi}.$$

Scaling again ε with p, say $\varepsilon=\sqrt{\frac{2\cdot 2}{p-1}}$, we have

$$\mathrm{P}\left(x=(t_1,\dots,t_p)\in\mathrm{B}_p : |t_p|\le\frac{2}{\sqrt{p-1}}\right) \ge 1-\mathrm{e}^{-2}\cdot\sqrt{2\pi}\approx 66\%,$$

i.e. more than 66% of the volume of the unit ball are within a slice of width $2/\sqrt{p-1}$ around the equator. Note that this confirms the idea that because of $1\ge\|x\|^2=t_1^2+\cdots+t_p^2$ each coordinate t_i should typically be of size $t_i\lesssim 1/\sqrt{p}$ (since typically $\|x\|^2\approx 1$ because of the concentration of $\|\cdot\|$).

1.4. Any two vectors in the ball are almost orthogonal

Consider two vectors x, y in the unit ball. What is the probability that they are close to being orthogonal? We have

$$\mathrm{P}\left(x, y \in \mathrm{B}_p : \frac{|\langle x, y\rangle|}{\|x\| \cdot \|y\|} \le \varepsilon\right) = \frac{\mathrm{vol}\left(\left\{(x, y) \in \mathrm{B}_p \times \mathrm{B}_p : \frac{|\langle x,y\rangle|}{\|x\|\cdot\|y\|} \le \varepsilon\right\}\right)}{\underbrace{\mathrm{vol}(\mathrm{B}_p \times \mathrm{B}_p)}_{=\mathrm{vol}(\mathrm{B}_p)^2}}$$

$$= \frac{\mathrm{vol}\left(\left\{x \in \mathrm{B}_p : \frac{|\langle x,y\rangle|}{\|x\|\cdot\|y\|} \le \varepsilon\right\}\right)}{\mathrm{vol}(\mathrm{B}_p)}$$

for any fixed $y \in \mathrm{B}_p$, by the rotational invariance of the problem. Let's take $y = (0, \ldots, 0, 1) \in \mathrm{B}_p$, then $\|y\| = 1$ and

$$\mathrm{P}\left(x, y \in \mathrm{B}_p : \frac{|\langle x, y\rangle|}{\|x\| \cdot \|y\|} \le \varepsilon\right) = \frac{\mathrm{vol}\left(\left\{x \in \mathrm{B}_p : \frac{|\langle x,y\rangle|}{\|x\|\cdot\|y\|} \le \varepsilon\right\}\right)}{\mathrm{vol}(\mathrm{B}_p)}$$

$$= \frac{\mathrm{vol}\left(\left\{x = (t_1, \ldots, t_p) \in \mathrm{B}_p : \frac{|t_p|}{\|x\|} \le \varepsilon\right\}\right)}{\mathrm{vol}(\mathrm{B}_p)}.$$

Define η via $\frac{\eta}{1-\eta} = \varepsilon$ (note that $\varepsilon \approx \eta$ for small ε), then we know

$$\mathrm{P}(|t_p| \le \eta) \ge 1 - \exp\left(-\frac{p-1}{2} \cdot \eta^2\right) \cdot \sqrt{2\pi}$$

and

$$\mathrm{P}(\|x\| > 1 - \eta) \ge 1 - \exp\left(-p\eta\right),$$

thus

$$\mathrm{P}(|t_p| \le \eta \text{ and } \|x\| > 1 - \eta) \ge 1 - \exp\left(-\frac{p-1}{2} \cdot \eta^2\right) \cdot \sqrt{2\pi} - \exp\left(-p\eta\right).$$[1]

Since $|t_p| \le \eta$ and $\|x\| > 1 - \eta$ together imply

$$\frac{|t_p|}{\|x\|} \le \frac{\eta}{1-\eta} = \varepsilon,$$

we have

$$\mathrm{P}\left(x, y \in \mathrm{B}_p : \frac{|\langle x, y\rangle|}{\|x\| \cdot \|y\|} \le \frac{\eta}{1-\eta}\right) \ge 1 - \exp\left(-\frac{p-1}{2} \cdot \eta^2\right) \cdot \sqrt{2\pi} - \exp\left(-p\eta\right).$$

[1] If $\mathrm{P}(A^C) \le \alpha$ and $\mathrm{P}(B^C) \le \beta$, then $\mathrm{P}((A \cap B)^C) = \mathrm{P}(A^C \cup B^C) \le \mathrm{P}(A^C) + \mathrm{P}(B^C) \le \alpha + \beta$.

2. Gaussian random vectors and linear concentration of Chebyshev and Bernstein type

2.1. Gaussian random vectors

We now consider random vectors $x = (t_1, \ldots, t_p) \in \mathbb{R}^p$ given by a probability distribution $\psi(x) = \psi(t_1, \ldots, t_p)$ on $\mathbb{R}^p$, i.e.

- $\psi(x) \geq 0$ for all $x \in \mathbb{R}^p$ and
- $1 = \int_{\mathbb{R}^p} \psi(x)\,\mathrm{d}x = \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} \psi(t_1, \ldots, t_p)\,\mathrm{d}t_1 \ldots\, \mathrm{d}t_p$.

Thus, for $H \subset \mathbb{R}^p$, $\mathrm{P}\{x \in H\} = \int_H \psi(x)\,\mathrm{d}x$.

The coordinates of x are independent if ψ factorizes as $\psi = \psi_1 \times \ldots \times \psi_p$, i.e.

$$\psi(t_1, \ldots, t_p) = \psi_1(t_1) \cdot \psi_2(t_2) \cdots \psi_p(t_p),$$

i.e. for sets $H \subseteq \mathbb{R}^p$ of the form $H = H_1 \times \ldots \times H_p$, where $H_i \subset \mathbb{R}$ for each $i = 1, \ldots, p$, we have

$$\begin{aligned}
&\mathrm{P}\{(t_1, \ldots, t_p) \in \mathbb{R}^p \mid t_1 \in H_1, \ldots, t_p \in H_p\} \\
&\qquad = \mathrm{P}\{(t_1, \ldots, t_p) \in H_1 \times \ldots \times H_p\} \\
&\qquad = \int_{H_1 \times \ldots \times H_p} \psi(t_1, \ldots, t_p)\,\mathrm{d}t_1 \ldots\, \mathrm{d}t_p \\
&\qquad = \int_{H_1 \times \ldots \times H_p} \psi_1(t_1) \cdot \ldots \cdot \psi_p(t_p)\,\mathrm{d}t_1 \ldots\, \mathrm{d}t_p \\
&\qquad = \int_{H_p} \ldots \left(\int_{H_2} \left(\int_{H_1} \psi_1(t_1)\,\mathrm{d}t_1 \right) \psi_2(t_2)\,\mathrm{d}t_2 \right) \ldots \psi_p(t_p)\,\mathrm{d}t_p \\
&\qquad = \int_{H_1} \psi_1(t_1)\,\mathrm{d}t_1 \cdot \int_{H_2} \psi_1(t_2)\,\mathrm{d}t_2 \cdots \int_{H_p} \psi_1(t_p)\,\mathrm{d}t_p \\
&\qquad = \mathrm{P}\{t_1 \in H_1\} \cdot \mathrm{P}\{t_2 \in H_2\} \cdot \cdots \cdot \mathrm{P}\{t_p \in H_p\}.
\end{aligned}$$

Note that $\psi_1, \ldots, \psi_p$ are necessarily probability distributions on $\mathbb{R}$ for the components t_i of x.

We now consider Gaussian random vectors where the coordinates are i.i.d. (independent and identically distributed) standard Gaussians.

Definition 2.1. A *standard Gaussian random vector* $x = (t_1, \ldots, t_p) \in \mathbb{R}^p$ is given by a probability distribution such that

(i) all $t_1, \ldots, t_p$ are independent and

(ii) each t_i has a standard Gaussian distribution $t_i \sim N(0,1)$, i.e.,

$$\psi_i(t_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t_i^2}{2}\right)$$

and thus

$$\begin{aligned}
\psi(x) &= \psi(t_1, \ldots, t_p) \\
&= \psi_1(t_1) \cdot \ldots \cdot \psi_p(t_p) \\
&= \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}\left(t_1^2 + t_2^2 + \ldots + t_p^2\right)\right) \\
&= \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{\|x\|^2}{2}\right).
\end{aligned}$$

We denote this by $x \sim N(0, I_p)$, where 0, the vector of all zeros, is the *vector of means*, and I_p, the $p \times p$ identity matrix, is the *matrix of covariances*.

2.2. Concentration of the norm

In numerical simulations we also have seen concentration for such Gaussian random vectors, i.e., we expect

$$\mathrm{P}\left\{x \in \mathbb{R}^p : |f(x) - E[f(x]| \geq \varepsilon\right\}$$

to be small for large p and for some "nice" (still to be determined) classes of functions f. Consider $f(x) = \|x\|^2$; then

$$E[\|x\|^2] = E[t_1^2 + \ldots + t_p^2] = E[t_1^2] + \ldots + E[t_p^2] = 1 + \ldots + 1 = p,$$

so we expect that $\|x\|^2$ is close to p and thus $\|x\|$ is close to $\sqrt{p}$. Note that $E[\|x\|] \neq \sqrt{p}$: they only asymptotically approach each other. Often it is easier to compare $f(x)$ with easier quantities than $E[f(x)]$, which asymptotically of course have to approximate $E[f(x)]$ (e.g., the median might also be a good choice).

Let us now try to prove the following:

Theorem 2.2. Consider a p-dimensional standard Gaussian random vector $x \sim N(0, I_p)$. Then, for $0 \leq \varepsilon \leq \sqrt{p}$,

$$\mathrm{P}\left\{x \in \mathbb{R}^p : |\|x\| - \sqrt{p}| \geq \varepsilon\right\} \leq 2\exp\left(-\frac{\varepsilon^2}{16}\right),$$

or, in the rescaled version $\tilde{x} = \frac{1}{\sqrt{p}}x$, where $\tilde{x} = (\tilde{t_1}, \ldots, \tilde{t_p})$ with i.i.d. $t_i \sim N\left(0, \frac{1}{p}\right)$

$$\mathrm{P}\left\{\tilde{x} \in \mathbb{R}^p : |\|\tilde{x}\| - 1| \geq \varepsilon\right\} \leq 2\exp\left(-\frac{p\varepsilon^2}{16}\right).$$

Proof (getting started). We have

$$\begin{aligned}\mathrm{P}\left\{|\|x\| - \sqrt{p}| \geq \varepsilon\right\} &\leq \mathrm{P}\left\{|\|x\| - \sqrt{p}| \cdot (\|x\| + \sqrt{p}) \geq \varepsilon\sqrt{p}\right\} \\ &= \mathrm{P}\left\{\left|\|x\|^2 - p\right| \geq \varepsilon\sqrt{p}\right\}.\end{aligned}$$

So we have reduced a concentration question for $\|x\|$ to one for $\|x\|^2$. But $\|x\|^2 = t_1^2 + \ldots + t_p^2$ is the sum of independent variables, and for such there is hope. First, let us consider such situations in general ... proof to be continued later! □

2.3. Markov and Chebyshev inequality

Theorem 2.3 (Markov Inequality)**.** Let $y \in \mathbb{R}^p$ be a random vector with probability distribution ψ and $f : \mathbb{R}^p \to [0, \infty)$ a positive function. Then, for any $\alpha > 0$,

$$\mathrm{P}\left\{y \in \mathbb{R}^p : f(y) \geq \alpha\right\} \leq \frac{E[f(y)]}{\alpha}.$$

Proof. We have

$$\begin{aligned}E[f(y)] &= \int_{\mathbb{R}^p} f(y)\psi(y)\,\mathrm{d}y \\ &= \int_{y\in\mathbb{R}^p : f(y)\geq\alpha} f(y)\psi(y)\,\mathrm{d}y + \underbrace{\int_{y\in\mathbb{R}^p : f(y)<\alpha} f(y)\psi(y)\,\mathrm{d}y}_{\geq 0, \text{ since } f\geq 0 \text{ and } \psi\geq 0} \\ &\geq \int_{y\in\mathbb{R}^p : f(y)\geq\alpha} f(y)\psi(y)\,\mathrm{d}y \\ &\geq \alpha \int_{y\in\mathbb{R}^p : f(y)\geq\alpha} \psi(y)\,\mathrm{d}y \\ &= \alpha\,\mathrm{P}\left\{y \in \mathbb{R}^p : f(y) \geq \alpha\right\}.\end{aligned}$$

□

Unfortunately, this is not useful in our situation, since we cannot easily deal with $f(x) = E[\left|\|x\|^2 - p\right|]$.

Theorem 2.4 (Chebyshev Inequality)**.** Let $y \in \mathbb{R}^p$ be a random vector and $f : \mathbb{R}^p \to \mathbb{R}$ a function such that the average $E[f(y)]$ and the variance

$$V[f(y)] = E[(f(y) - E[f(y)])^2] = E[f(y)^2] - E[f(y)]^2$$

are finite. Then, for any $\varepsilon > 0$,

$$\mathrm{P}\left\{y \in \mathbb{R}^p : |f(y) - E[f(y)]| \geq \varepsilon\right\} \leq \frac{V[f(y)]}{\varepsilon^2}.$$

Proof. We will use the Markov Inequality 2.3 for $g(y) := (f(y) - E[f(y)])^2$. Note that $g \geq 0$ and

$$E[g(y)] = E[(f(y) - E[f(y)])^2] = V[f(y)].$$

Thus, for $\varepsilon > 0$, we have

$$\begin{aligned}
\mathrm{P}\left\{y \in \mathbb{R}^p : |f(y) - E[f(y)]| \geq \varepsilon\right\} &= \mathrm{P}\left\{y \in \mathbb{R}^p : (f(y) - E[f(y)])^2 \geq \varepsilon^2\right\} \\
&= \mathrm{P}\left\{y \in \mathbb{R}^p : g(y) \geq \varepsilon^2\right\} \\
&\leq \frac{E[g(y)]}{\varepsilon^2} \\
&= \frac{V[f(y)]}{\varepsilon^2}.
\end{aligned}$$

□

Now use this for our Gaussian random vectors $y = x \sim N(0, I_p)$ and $f(x) = \|x\|^2$: then $E[f(x)] = E[\|x\|^2] = p$ and

$$\begin{aligned}
V[f(x)] &= E\left[\left(\|x\|^2 - p\right)^2\right] \\
&= E\left[\left((t_1^2 - 1) + \ldots + (t_p^2 - 1)\right)^2\right] \\
&= pE\left[(t_1^2 - 1)^2\right] \\
&= pE\left[t_1^4 - 2t_1^2 + 1\right] \\
&= p\left(E[t_1^4] - 2E[t_1^2] + 1\right) \\
&= p(3 - 2 + 1) \\
&= 2p.
\end{aligned}$$

Alternatively, we can also calculate this directly in terms of the variance, by using the fact that the variance of a sum of independent variables is the sum of the variances:

$$\begin{aligned}
V[\|x\|^2] &= V(t_1^2 + \cdots + t_p^2) \\
&= V(t_1^2) + \cdots + V(t_p^2) \\
&= pV(t_1^2) \\
&= p(E[t_1^4] - E[t_1^2]^2) \\
&= p(3 - 1) \\
&= 2p
\end{aligned}$$

Applying now the Chebyshev inequality 2.4 to this setting, we get

$$\mathrm{P}\left\{x \in \mathbb{R}^p : \left|\|x\|^2 - p\right| \geq \varepsilon\sqrt{p}\right\} \leq \frac{V[\|x\|^2]}{\varepsilon^2 \cdot p} = \frac{2p}{\varepsilon^2 \cdot p} = \frac{2}{\varepsilon^2}.$$

Thus,

$$\mathrm{P}\left\{x \in \mathbb{R}^p : \left|\|x\| - \sqrt{p}\right| \geq \varepsilon\right\} \leq \frac{2}{\varepsilon^2},$$

which is still far away from our exponential estimate in Theorem 2.2.

2.4. Bernstein inequality

We have to strengthen Markov/Chebyshev by also taking higher moments into account. This can be done systematically by looking at all moments simultaneously via their generating power series.

Consider a random vector $y \in \mathbb{R}^p$ and a function $f : \mathbb{R}^p \to \mathbb{R}$. Then

$$E[\exp(\lambda f(y))] = E\left[\sum_{k=0}^{\infty} \frac{(\lambda f(y))^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} E[f(y)^k],$$

where $E[f(y)^k]$ is the k-th moment of $f(y)$, is the generating function in moments, where the choice of λ allows to weigh the contributions of the moments differently.

For each fixed $\lambda \in \mathbb{R}$, the function $y \mapsto \exp(\lambda f(y))$ is positive, so we can apply the Markov Inequality 2.3 to it, i.e. for each $\lambda > 0$ we have

$$\mathrm{P}\left\{f(y) \geq \alpha\right\} = \mathrm{P}\left\{\exp(\lambda f(y)) \geq \exp(\lambda\alpha)\right\} \leq \frac{E[\exp(\lambda f(y))]}{\exp(\lambda\alpha)}$$

(by using that $\exp(\lambda\cdot)$ is increasing for each $\lambda > 0$). We can optimize this by choosing the best $\lambda > 0$:

$$\mathrm{P}\left\{f(y) \geq \alpha\right\} \leq \inf_{\lambda>0} \exp(-\lambda\alpha) E[\exp(\lambda f(y))].$$

Without further information on f and y this is not very useful. Now invoke that

- $y = (s_1, \ldots, s_p)$ has independent coordinates, i.e. $\psi(s_1, \ldots, s_p) = \psi_1(s_1) \cdots \psi_p(s_p)$, and
- f is a sum of functions of the coordinates, i.e. $f(s_1, \ldots, s_p) = f_1(s_1) + \cdots + f_p(s_p)$.

Then

$$\exp(\lambda f(y)) = \exp\big(\lambda \cdot (f_1(s_1) + \cdots + f_p(s_p))\big) = \exp(\lambda f_1(s_1)) \cdots \exp(\lambda f_p(s_p)),$$

and, by independence,

$$E\big[\exp(\lambda f(y))\big] = E\big[\exp(\lambda f_1(s_1))\big] \cdots E\big[\exp(\lambda f_p(s_p))\big].$$

The right-hand side factors should not grow too fast; often they are of order $\sim e^{c\lambda^2}$, then we get good overall control. Note that for Gaussian distributions

$$\begin{aligned}\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp(\lambda t) \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{(t-\lambda)^2}{2}\right) \exp\left(\frac{\lambda^2}{2}\right) \mathrm{d}t \\ &= \exp\left(\frac{\lambda^2}{2}\right) \underbrace{\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{(t-\lambda)^2}{2}\right) \mathrm{d}t}_{=1} \\ &= \exp\left(\frac{\lambda^2}{2}\right).\end{aligned}$$

Now consider again our Gaussian random vectors $y = x \sim N(0, I_p)$, where

$$f(x) = \|x\|^2 - p = t_1^2 + \ldots + t_p^2 - p = \underbrace{(t_1^2 - 1)}_{=g(t_1)} + \ldots + \underbrace{(t_p^2 - 1)}_{=g(t_p)},$$

so all f_i are the same:

$$f_1 = \ldots = f_p = g \quad \text{with} \quad g(t) = t^2 - 1$$

and

$$\psi_1 = \ldots = \psi_p = \varphi \quad \text{where} \quad \varphi(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right).$$

If $\lambda < \frac{1}{2}$, then by substituting $s = t\sqrt{1-2\lambda}$ we get

$$\begin{aligned}E[\exp\big(\lambda g(t)\big)] &= E\left[\exp\big(\lambda(t^2-1)\big)\right] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\big(\lambda(t^2-1)\big) \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t \\ &= \frac{1}{\sqrt{2\pi}} \exp(-\lambda) \int_{-\infty}^{+\infty} \exp\left(-\frac{t^2}{2}(1-2\lambda)\right) \mathrm{d}t \\ &= \frac{1}{\sqrt{2\pi}} \exp(-\lambda) \underbrace{\int_{-\infty}^{+\infty} \exp\left(-\frac{s^2}{2}\right) \mathrm{d}s}_{=\sqrt{2\pi}} \cdot \frac{1}{\sqrt{1-2\lambda}} \\ &= \frac{\exp(-\lambda)}{\sqrt{1-2\lambda}}.\end{aligned}$$

Note that $E\left[\exp\big(\lambda(t^2-1)\big)\right]$ does not exist (or is equal to ∞) for $\lambda \geq \frac{1}{2}$. So now we get

$$\mathrm{P}\left\{f(x) \geq \alpha\right\} \leq \inf_{0<\lambda<\frac{1}{2}} \exp(-\lambda\alpha) E\left[\exp\big(\lambda g(t)\big)\right]^p = \inf_{0<\lambda<\frac{1}{2}} \exp(-\lambda\alpha) \frac{\exp(-\lambda p)}{(1-2\lambda)^{\frac{p}{2}}},$$

where the last factor behaves like (or at least can be estimated against) $\exp(\lambda^2 c)$.

Lemma 2.5. Let $t \sim N(0,1)$ be a one-dimensional standard Gaussian random variable. Then

$$E\left[\exp\bigl(\lambda(t^2-1)\bigr)\right] \leq \exp(4\lambda^2) \quad \text{for all} \quad |\lambda| \leq \frac{1}{4}.$$

Proof. Either use

(i) estimate for explicit calculation

$$E\left[\exp\bigl(\lambda(t^2-1)\bigr)\right] = \frac{\exp(-\lambda)}{\sqrt{1-2\lambda}} \leq \exp(4\lambda^2) \quad \text{for all} \quad |\lambda| \leq \frac{1}{4}$$

via curve discussion,

(ii) or estimate of moments of t^2-1: note that

$$\begin{aligned} \left|t^2-1\right|^k &= \begin{cases} (1-t^2)^k \leq 1, & \text{if } |t| \leq 1, \\ (t^2-1)^k \leq t^{2k}, & \text{if } |t| > 1, \end{cases} \\ &\leq t^{2k} + 1, \end{aligned}$$

so (for $k \geq 2$)

$$\begin{aligned} \left|E\left[(t^2-1)^k\right]\right| &\leq E\left[\left|t^2-1\right|^k\right] \\ &= \int_{\mathbb{R}} \left|t^2-1\right|^k \varphi(t)\,\mathrm{d}t \\ &\leq \int_{\mathbb{R}} (t^{2k}+1)\varphi(t)\,\mathrm{d}t \\ &= \underbrace{E[t^{2k}]}_{=(2k-1)!!} + \underbrace{E[1]}_{=1} \\ &= (2k-1)(2k-3)\cdot\ldots\cdot 5\cdot 3\cdot 1 + 1 \\ &\leq (2k)(2k-2)\cdot\ldots\cdot 6\cdot 4\cdot\frac{3}{4} + 1 \\ &\leq 2^{k-1}k!\cdot\frac{3}{4} + 2^{k-1}k!\cdot\frac{1}{4} \\ &= 2^{k-1}k! \end{aligned}$$

and thus

$$\begin{aligned}
E\left[\exp\big(\lambda(t^2-1)\big)\right] &= E\left[\sum_{k=0}^{\infty}\frac{\lambda^k(t^2-1)^k}{k!}\right] \\
&\le \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\big|E\left[(t^2-1)^k\right]\big| \\
&= 1+\lambda\underbrace{E[t^2-1]}_{=0}+\sum_{k=2}^{\infty}\frac{\lambda^k}{k!}\big|E\left[(t^2-1)^k\right]\big| \\
&\le 1+\frac{1}{2}\sum_{k=2}^{\infty}(2\lambda)^k \\
&= 1+\underbrace{\frac{(2\lambda)^2}{2(1-2\lambda)}}_{\le(2\lambda)^2 \text{ for } |\lambda|\le\frac{1}{4}} \\
&\le 1+4\lambda^2 \\
&\le \exp(4\lambda^2).
\end{aligned}$$

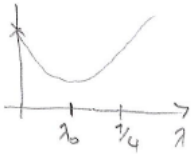
□

Proof of Theorem 2.2. We have to estimate $\mathrm{P}\left\{\big|\|x\|^2-p\big|\ge\varepsilon\sqrt{p}\right\}$, where we set

$$f(x)=\|x\|^2-p=g(t_1)+\ldots+g(t_p)$$

with independent $g(t_i)$ and $g(t)=t^2-1$ for $t\sim N(0,1)$. So, with $\alpha:=\varepsilon\sqrt{p}$,

$$\begin{aligned}
\mathrm{P}\left\{\big|\|x\|^2-p\big|\ge\varepsilon\sqrt{p}\right\} &= \mathrm{P}\left\{f(y)\ge\varepsilon\sqrt{p}\right\}+\mathrm{P}\left\{f(y)\le-\varepsilon\sqrt{p}\right\} \\
&= 2\,\mathrm{P}\left\{f(y)\ge\varepsilon\sqrt{p}\right\} \\
&\le 2\inf_{0<\lambda<\frac{1}{2}}\exp(-\lambda\alpha)\underbrace{E\left[\exp\big(\lambda(t^2-1)\big)\right]^p}_{\le\exp(4\lambda^2) \text{ for } \lambda\le\frac{1}{4}} \\
&\le 2\inf_{0<\lambda\le\frac{1}{4}}\exp(-\lambda\alpha+4\lambda^2 p).
\end{aligned}$$



Finding the minimum λ_0 of $h(\lambda)=-\lambda\alpha+4\lambda^2 p$ yields $-\alpha+8\lambda_0 p=h'(\lambda_0)=0$ and thus

$$\lambda_0=\frac{\alpha}{8p}=\frac{\varepsilon\sqrt{p}}{8p}=\frac{\varepsilon}{8\sqrt{p}}\le\frac{1}{4},$$

since we consider $0 \leq \varepsilon \leq \sqrt{p}$. Now

$$h(\lambda_0) = -\lambda_0\alpha + 4\lambda_0^2 p = -\frac{\alpha}{8p}\alpha + 4\left(\frac{\alpha}{8p}\right)^2 p = -\frac{\alpha^2}{8p} + \frac{\alpha^2}{16p} = -\frac{\alpha^2}{16p} = -\frac{\varepsilon^2 p}{16p} = -\frac{\varepsilon^2}{16}$$

and thus

$$\mathrm{P}\left\{x \in \mathbb{R}^p : |\|x\| - \sqrt{p}| \geq \varepsilon\right\} \leq \mathrm{P}\left\{x \in \mathbb{R}^p : \left|\|x\|^2 - p\right| \geq \varepsilon\sqrt{p}\right\} \leq 2\exp\left(-\frac{\varepsilon^2}{16}\right). \quad \square$$

Remark 2.6. Note that in

$$f(x) = g(t_1) + \ldots + g(t_p)$$

we only used that the $g(t_1), \ldots, g(t_p)$ are independent and that we could estimate their moments as

$$E\left[|g(t)|^k\right] \leq k! \cdot c^{k-1}$$

or, (more or less) equivalently,

$$E\left[\exp\big(\lambda g(t)\big)\right] \leq \exp(c\lambda^2)$$

on some interval. Distributions that satisfy these properties are called *sub-exponential* distributions; for them one has concentration estimates as above - those are usually called *Bernstein inequalities*. Other related inequalities go under the names of *Chernov inquality* and *Hoeffding inequality*.

3. Concentration of Gaussian random vectors for non-linear Lipschitz functions

3.1. Lipschitz functions

Up to now we considered concentration of linear sums of independent variables:

$$f(t_1, \ldots, t_p) = f_1(t_1) + \ldots + f_p(t_p).$$

Now we want to address more general, non-linear, functions

$$f(x) = f(t_1, \ldots, t_p).$$

Of course, they cannot be arbitrary; the guiding principle is:

> A random variable that depends (in a 'smooth' way) on the influence of many independent variables (but not too much on any of them) is essentially constant. (Michel Talagrand [Tal95])

Example: both

$$f(t_1, \ldots, t_p) = \frac{1}{p}(t_1^2 + \ldots + t_p^2) \quad \text{and} \quad g(t_1, \ldots, t_p) = t_1^2$$

satisfy $E[f(x)] = 1 = E[g(x)]$; f concentrates about 1 for large p; but g has, independent of p, always the same spread-out distribution.

Definition 3.1. (1) A function $f : \mathbb{R}^p \to \mathbb{R}$ is *L-Lipschitz*, with Lipschitz constant $L > 0$, if

$$|f(x) - f(y)| \leq L\|x - y\| \quad \text{for all} \quad x, y \in \mathbb{R}^p.$$

(2) More generally, a function $f : \mathbb{R}^p \to \mathbb{R}^m$ is *L-Lipschitz*, if

$$\|f(x) - f(y)\| \leq L\|x - y\| \quad \text{for all} \quad x, y \in \mathbb{R}^p.$$

Note:

(i) a smooth (i.e., differentiable) function f is Lipschitz if and only if its gradient vector

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial t_1} \\ \vdots \\ \frac{\partial f(x)}{\partial t_p} \end{pmatrix}$$

is bounded.

(ii) "interesting" non-differentiable Lipschitz functions for $p = 1$ are, e.g., the ReLU function with $L = 1$, see Figure 2.

(iii) componentwise application of Lipschitz functions is Lipschitz: if $f_1, \ldots, f_p : \mathbb{R} \to \mathbb{R}$ are L-Lipschitz, then

$$f : \mathbb{R}^p \to \mathbb{R}^p \quad \text{with} \quad f(t_1, \ldots, t_p) = \big(f_1(t_1), \ldots, f_p(t_p)\big)$$

is also Lipschitz with the same Lipschitz constant L: with $x = (t_1, \ldots, t_p)$ and $y = (s_1, \ldots, s_p)$ we have

$$\begin{aligned}
\|f(x) - f(y)\|^2 &= \left\| \begin{pmatrix} f_1(t_1) - f_1(s_1) \\ \vdots \\ f_p(t_p) - f_p(s_p) \end{pmatrix} \right\|^2 \\
&= \underbrace{|f_1(t_1) - f_1(s_1)|^2}_{\leq L^2 |t_1 - s_1|^2} + \ldots + \underbrace{|f_p(t_p) - f_p(s_p)|^2}_{\leq L^2 |t_p - s_p|^2} \\
&\leq L^2 \left(|t_1 - s_1|^2 + \ldots + |t_p - s_p|^2 \right) \\
&= L^2 \|x - y\|^2
\end{aligned}$$

and thus $\|f(x) - f(y)\| \leq L\|x - y\|$.

(iv) composition of Lipschitz functions is Lipschitz; e.g., if $f : \mathbb{R}^p \to \mathbb{R}^p$ is L_1-Lipschitz and $g : \mathbb{R}^p \to \mathbb{R}$ is L_2-Lipschitz, then

$$h := g \circ f : \mathbb{R}^p \to \mathbb{R}, \quad x \mapsto h(x) = g(f(x))$$

is also Lipschitz:

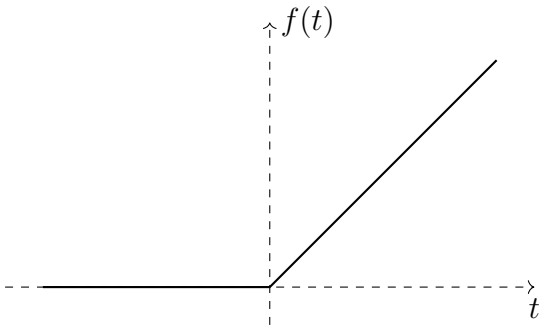
$$|h(x) - h(y)| = |g(f(x)) - g(f(y))| \leq L_2 \underbrace{\|f(x) - f(y)\|}_{\leq L_1 \|x - y\|} \leq L_1 \cdot L_2 \|x - y\|,$$

so h is $(L_1 \cdot L_2)$-Lipschitz. In particular, since if $L_1 = L_2 = 1$ then also $L_1 \cdot L_2 = 1$, compositions of 1-Lipschitz functions are 1-Lipschitz.

3.2. Concentration for Lipschitz functions of independent Gaussian variables

Theorem 3.2 (Gaussian concentration for Lipschitz functions)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$ be an L-Lipschitz function. Then we have for a standard Gaussian random vector $x \sim N(0, I_p)$ the following concentration estimate for any $\alpha \geq 0$:

$$\mathrm{P}\left\{x \in \mathbb{R}^p : |f(x) - E[f(x)]| \geq \alpha\right\} \leq 2 \exp\left(-\frac{\alpha^2}{2L^2}\right).$$



$$f(t) = \begin{cases} 0, & t \leq 0, \\ t, & t \geq 0 \end{cases}$$

Figure 2: the ReLU function (rectified linear unit).

Proof. (1) We start with some simplifications: by shifting $f \rightsquigarrow f - E[f(x)]$, the Lipschitz property is not affected, so we can assume that $E[f(x)] = 0$ and then we have to prove

$$\mathrm{P}\left\{x \in \mathbb{R}^p : |f(x)| \geq \alpha\right\} \leq 2\exp\left(-\frac{\alpha^2}{2L^2}\right).$$

Since

$$\{x \in \mathbb{R}^p : |f(x)| \geq \alpha\} = \{x \in \mathbb{R}^p : f(x) \geq \alpha\} \cup \{x \in \mathbb{R}^p : f(x) \leq -\alpha\}$$

and f and $-f$ have the same Lipschitz constant, it suffices to prove

$$\mathrm{P}\left\{x \in \mathbb{R}^p : f(x) \geq \alpha\right\} \leq \exp\left(-\frac{\alpha^2}{2L^2}\right).$$

Since one can approximate general Lipschitz functions by smooth functions, we can restrict to the case where f is smooth and $\|\nabla f(x)\| \leq L$ for all $x \in \mathbb{R}^p$ and we will prove a bit weaker estimate: instead of $\exp\left(-\frac{\alpha^2}{2L^2}\right)$ we will get $\exp\left(-\frac{2}{\pi^2}\frac{\alpha^2}{L^2}\right)$. The following elegant proof for this is due to Maurey and Pisier [Pis06].

(2) Now let's get started in the usual way: for each $\lambda > 0$ (to be determined later) we have

$$\begin{aligned} \mathrm{P}\left\{x \in \mathbb{R}^p : f(x) \geq \alpha\right\} &= \mathrm{P}\left\{x \in \mathbb{R}^p : \exp(\lambda f(x)) \geq \exp(\lambda\alpha)\right\} \\ &\leq \frac{E[\exp(\lambda f(x))]}{\exp(\lambda\alpha)}. \end{aligned}$$

So as before, we need to estimate $E[\exp(\lambda f(x))]$. To do so, we need two general ingredients, which we will recall first... our proof will then be continued later. □

Theorem 3.3 (Jensen Inequality)**.** Consider a random vector $x \in \mathbb{R}^p$ with probability density ψ, i.e.

$$E[f(x)] = \int_{\mathbb{R}} f(x)\psi(x)\,\mathrm{d}x \quad \text{for all} \quad f : \mathbb{R}^p \to \mathbb{R}.$$

Let $h : \mathbb{R} \to \mathbb{R}$ be a convex function. Then for any $f : \mathbb{R}^p \to \mathbb{R}$ we have

$$h(E[f(x)]) \leq E[h(f(x))].$$

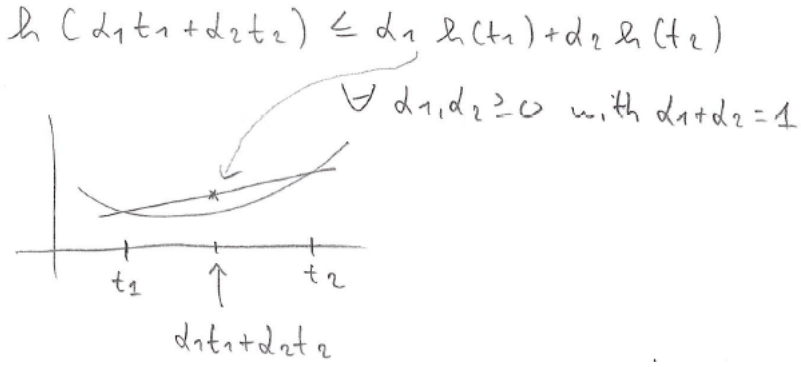
In particular, since for any λ the function $h(t) = \exp(\lambda t)$ is convex, we have

$$\exp(\lambda E[f(x)]) \leq E[\exp(\lambda f(x))] \quad \text{for any} \quad \lambda \in \mathbb{R}.$$

Recall:

(1) h convex means

$$h(\alpha_1 t_1 + \alpha_2 t_2) \leq \alpha_1 h(t_1) + \alpha_2 h(t_2) \quad \text{for all} \quad \alpha_1, \alpha_2 \geq 0 \quad \text{with} \quad \alpha_1 + \alpha_2 = 1.$$



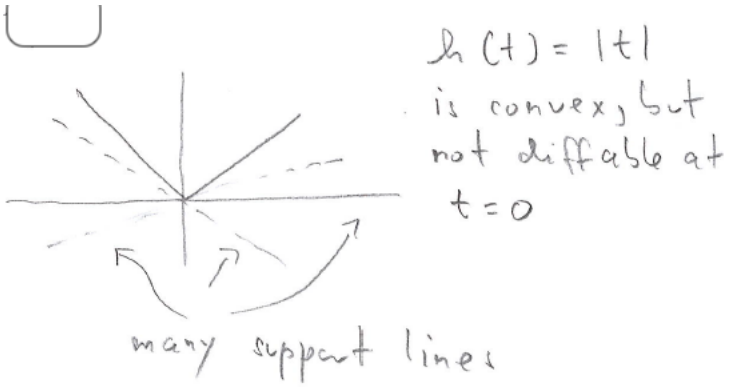
By induction, we get

$$h(\alpha_1 t_1 + \ldots + \alpha_m t_m) \leq \alpha_1 h(t_1) + \ldots + \alpha_m h(t_m)$$

for all t_i and $\alpha_i \geq 0$ such that $\alpha_1 + \ldots + \alpha_m = 1$. Jensen Inequality is the continuous version of this (where $\Sigma \rightsquigarrow \int$).

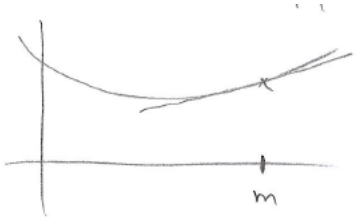
(2) At points where h is differentiable, h lies above its tangent.

but also at other points there are such "support" lines, but they don't have to be unique: $h(t) = |t|$ is convex, but not differentiable at $t = 0$.



Nevertheless, there are many "support" lines.

Proof of Theorem 3.3. Put $m := E[f(x)]$ and choose a support line at m, i.e. we have $a, b \in \mathbb{R}$ such that $h(t) \geq at + b$ for all t and $h(m) = am + b$.



Then we have

$$\begin{aligned} E[h(f(x))] &= \int_{\mathbb{R}} \underbrace{h(f(x))}_{\geq af(x)+b} \psi(x)\,\mathrm{d}x \\ &\geq a \underbrace{\int_{\mathbb{R}} f(x)\psi(x)\,\mathrm{d}x}_{=E[f(x)]=m} + b \underbrace{\int_{\mathbb{R}} \psi(x)\,\mathrm{d}x}_{=1} \\ &= am + b \\ &= h(m) \\ &= h(E[f(x)]). \end{aligned}$$

□

If $E[f(x)] = 0$, then we have in particular

$$E[\exp(-\lambda f(x))] \geq \exp(-\lambda E[f(x)]) = \exp(0) = 1$$

and thus we can estimate

$$\begin{aligned} E\left[\exp\big(\lambda f(x)\big)\right] &= E\left[\exp\big(\lambda f(x)\big)\right] \cdot 1 \\ &\leq E\left[\exp\big(\lambda f(x)\big)\right] \cdot E\left[\exp\big(-\lambda f(y)\big)\right] \\ &= E\left[\exp\big(\lambda f(x)\big) \cdot \exp\big(-\lambda f(y)\big)\right] \\ &= E\left[\exp\big(\lambda(f(x) - f(y))\big)\right], \end{aligned}$$

where y is a copy of x, which is independent from x, i.e. $\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^{2p}$ is a $2p$-dimensional Gaussian random vector.

Why do we introduce this y? Because we can write

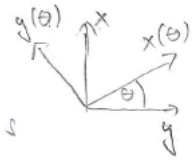
$$f(x) - f(y) = \int_0^{\frac{\pi}{2}} \frac{\partial}{\partial\theta} f\underbrace{\big(\cos(\theta)y + \sin(\theta)x\big)}_{=:x(\theta)} \,\mathrm{d}\theta = \int_0^{\frac{\pi}{2}} \nabla f(x(\theta)) \cdot x'(\theta)$$

and $x(\theta), x'(\theta)$ are, for each θ, a pair of independent standard Gaussian vectors.

Lemma 3.4. Let x and y be two independent standard Gaussian random vectors in $\mathbb{R}^p$. For $\theta \in \mathbb{R}$ we consider

$$x(\theta) := \cos(\theta)y + \sin(\theta)x \quad \text{and} \quad y(\theta) := -\sin(\theta)y + \cos(\theta)x = x'(\theta).$$

Then, for each $\theta \in \mathbb{R}$, $x(\theta)$ and $y(\theta)$ are also two independent standard Gaussian random vectors in $\mathbb{R}^p$.



Example. For $\theta = \frac{\pi}{2}$ this says that if $x, y \sim N(0, I_p)$ are independent, then also

$$\frac{1}{\sqrt{2}}(x+y) \sim N(0, I_p) \quad \text{and} \quad \frac{1}{\sqrt{2}}(x-y) \sim N(0, I_p)$$

are independent.

Proof of Lemma 3.4. By assumption, $(x, y) \in \mathbb{R}^{2p}$ has density

$$\begin{aligned}
\psi(x,y) &= \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}\|x\|^2\right) \cdot \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}\|y\|^2\right) \\
&= \frac{1}{(2\pi)^p} \exp\left(-\frac{1}{2}(\|x\|^2 + \|y\|^2)\right) \\
&= \frac{1}{(2\pi)^p} \exp\left(-\frac{1}{2}\left\|\begin{pmatrix} x \\ y \end{pmatrix}\right\|^2\right),
\end{aligned}$$

i.e. $\begin{pmatrix} x \\ y \end{pmatrix} \sim N(0, I_{2p})$. The replacement $\begin{pmatrix} x \\ y \end{pmatrix} \rightsquigarrow \begin{pmatrix} x(\theta) \\ y(\theta) \end{pmatrix}$ is a unitary transformation in $\mathbb{R}^{2p}$: with

$$U = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

as well as

$$x(\theta) = (t_1(\theta), \dots, t_p(\theta)) \quad \text{and} \quad y(\theta) = (s_1(\theta), \dots, s_p(\theta)),$$

we have

$$\begin{pmatrix} t_1(\theta) \\ s_1(\theta) \\ t_2(\theta) \\ s_2(\theta) \\ \vdots \\ t_p(\theta) \\ s_p(\theta) \end{pmatrix} = \underbrace{\begin{pmatrix} U & & & 0 \\ & U & & \\ & & \ddots & \\ 0 & & & U \end{pmatrix}}_{=:V} \cdot \begin{pmatrix} t_1 \\ s_1 \\ t_2 \\ s_2 \\ \vdots \\ t_p \\ s_p \end{pmatrix}$$

and thus by the behaviour of Gaussian random vectors under linear transformations (see Assignment 2, Exercise 4)

$$\begin{pmatrix} t_1(\theta) \\ s_1(\theta) \\ t_2(\theta) \\ s_2(\theta) \\ \vdots \\ t_p(\theta) \\ s_p(\theta) \end{pmatrix} \sim N(0, \underbrace{V I_{2p} V^T}_{=I_{2p}}),$$

i.e. for each θ we again have a standard Gaussian random vector in $\mathbb{R}^{2p}$. □

Now we continue:

Proof of Theorem 3.2. We have to estimate

$$\begin{aligned} E\left[\exp\big(\lambda f(x)\big)\right] &\le E\left[\exp\big(\lambda(f(x) - f(y)\big)\right] \\ &= E\left[\exp\left(\lambda \int_0^{\frac{\pi}{2}} \frac{\partial}{\partial\theta} f\big(x(\theta)\big)\, \mathrm{d}\theta\right)\right] \\ &= E\left[\exp\left(\lambda \int_0^{\frac{\pi}{2}} \nabla f\big(x(\theta)\big) \cdot x'(\theta)\, \mathrm{d}\theta\right)\right], \end{aligned}$$

where

- f is smooth and L-Lipschitz, i.e. $\|\nabla f(\cdot)\| \le L$,
- $x \sim N(0, I_p)$,
- y is an independent copy of x,
- $x(\theta) = \cos(\theta) y + \sin(\theta) x$, thus $x'(\theta) = -\sin(\theta) y + \cos(\theta) x$ and thus, by Lemma 3.4,

$$\big(x(\theta), x'(\theta)\big) \sim N(0, I_{2p}) \quad \text{for all} \quad \theta \in \mathbb{R}.$$

We use Jensen's inequality now for an average of the form $\frac{1}{\pi/2}\int_0^{\frac{\pi}{2}} g(\theta)\,\mathrm{d}\theta$ and the convex function $h(t)=\exp(ct)$, i.e.

$$\exp\left(c\cdot\frac{1}{\pi/2}\int_0^{\frac{\pi}{2}} g(\theta)\,\mathrm{d}\theta\right)\le\frac{1}{\pi/2}\int_0^{\frac{\pi}{2}}\exp\big(cg(\theta)\big)\,\mathrm{d}\theta$$

for $g(\theta)=\nabla f\big(x(\theta)\big)\cdot x'(\theta)$ and $c=\lambda\cdot\frac{\pi}{2}$ yielding

$$\begin{aligned}
&E\left[\exp\big(\lambda f(x)\big)\right]\\
&\le E\left[\frac{1}{\pi/2}\int_0^{\frac{\pi}{2}}\exp\Big(\lambda\cdot\frac{\pi}{2}\cdot\nabla f\big(x(\theta)\big)\cdot x'(\theta)\Big)\,\mathrm{d}\theta\right]\\
&=\frac{2}{\pi}\int_0^{\frac{\pi}{2}}\underbrace{E\left[\exp\Big(\lambda\frac{\pi}{2}\nabla f\big(x(\theta)\big)\cdot x'(\theta)\Big)\right]}_{=E\left[\exp\left(\lambda\frac{\pi}{2}\nabla f(y)\cdot x\right)\right]\text{ for all }\theta}\mathrm{d}\theta\\
&=E\left[\exp\Big(\lambda\frac{\pi}{2}\nabla f(y)\cdot x\Big)\right]\\
&=\frac{1}{(2\pi)^p}\int_{\mathbb{R}^p}\int_{\mathbb{R}^p}\exp\Big(\lambda\frac{\pi}{2}\nabla f(y)\cdot x\Big)\cdot\exp\Big(-\frac{1}{2}\|x\|^2\Big)\cdot\exp\Big(-\frac{1}{2}\|y\|^2\Big)\,\mathrm{d}y\,\mathrm{d}x.
\end{aligned}$$

For each fixed y, we have $\|\nabla f(y)\|\le L$ and $\nabla f(y)\cdot x$ is a Gaussian variable with variance at most L^2, thus

$$\frac{1}{(2\pi)^{\frac{p}{2}}}\int_{\mathbb{R}^p}\exp\Big(\lambda\frac{\pi}{2}\nabla f(y)\cdot x\Big)\cdot\exp\Big(-\frac{1}{2}\|x\|^2\Big)\le\exp\left(\frac{1}{2}\left(\lambda\frac{\pi}{2}L\right)^2\right)$$

for each y. Integrating over y yields

$$E\left[\exp\big(\lambda f(x)\big)\right]\le\exp\left(\frac{1}{2}\left(\lambda\frac{\pi}{2}L\right)^2\right)$$

and thus

$$\mathrm{P}\left\{f(x)\ge\alpha\right\}\le\frac{E\left[\exp\big(\lambda f(x)\big)\right]}{\exp(\lambda\alpha)}\le\exp\left(\frac{1}{2}\lambda^2\left(\frac{\pi}{2}L\right)^2-\lambda\alpha\right).$$

Minimizing the exponent function $h(\lambda)=\frac{1}{2}\lambda^2\left(\frac{\pi}{2}L\right)^2-\lambda\alpha$ yields $\lambda_0=\frac{4}{\pi^2}\frac{\alpha}{L^2}$ with

$$\begin{aligned}
h(\lambda_0)=\frac{1}{2}\lambda_0^2\left(\frac{\pi}{2}L\right)^2-\lambda_0\alpha&=\frac{1}{2}\left(\frac{4}{\pi^2}\frac{\alpha}{L^2}\right)^2\left(\frac{\pi}{2}L\right)^2-\frac{4}{\pi^2}\frac{\alpha}{L^2}\alpha\\
&=\frac{1}{2}\frac{16}{\pi^4}\frac{\alpha^2}{L^4}\frac{\pi^2}{4}L^2-\frac{4}{\pi^2}\frac{\alpha^2}{L^2}\\
&=\frac{2-4}{\pi^2}\frac{\alpha^2}{L^2}=-\frac{2}{\pi^2}\frac{\alpha^2}{L^2}
\end{aligned}$$

and thus

$$\mathrm{P}\left\{f(x)\ge\alpha\right\}\le\exp\left(-\frac{2}{\pi^2}\frac{\alpha^2}{L^2}\right).$$

□

3.3. Generalizations of concentration inequalities

The concentration inequalities are of the form

$$\text{input} \xrightarrow{\text{function}} \text{output},$$

where

- the input are independent variables, which up to now were Gaussian,
- the function is linear or Lipschitz, and
- the output concentrates like a Gaussian distribution $\sim \exp(-\alpha^2 \cdot c)$.

Note that in the linear situation the assumption on the Gaussianity of the input distributions can be generalized quite a bit: the main ingredient was the control of $E[\exp(\lambda g(t))]$.

Definition 3.5. Let x be a one-dimensional real random variable with $E[x] = 0$.

(1) x is called *sub-Gaussian*, if one of the following two equivalent properties is satisfied:

(i) There exists a $c > 0$ such that for all $\alpha \geq 0$ we have

$$\mathrm{P}\left\{|x| \geq \alpha\right\} \leq 2\exp\left(-\frac{\alpha^2}{c}\right).$$

(ii) There exists a $\tilde{c} > 0$ such that for all $\lambda \in \mathbb{R}$ we have

$$E\big[\exp(\lambda x)\big] \leq \exp(\tilde{c}\lambda^2).$$

(2) x is called *sub-exponential*, if one of the following two equivalent properties is satisfied:

(i) There exists a $c > 0$ such that for all $\alpha \geq 0$ we have

$$\mathrm{P}\left\{|x| \geq \alpha\right\} \leq 2\exp\left(-\frac{\alpha}{c}\right).$$

(ii) There exists a $\tilde{c} > 0$ such that for all $\lambda \in \mathbb{R}$ with $|\lambda| \leq \frac{1}{\tilde{c}}$ we have

$$E\big[\exp(\lambda x)\big] \leq \exp(\tilde{c}^2\lambda^2).$$

Example. (i) If x is sub-Gaussian, then x is also sub-exponential.

(ii) If $x \sim N(0,1)$, then x is sub-Gaussian. Furthermore, $x^2 - 1$ is sub-exponential, but not sub-Gaussian.

(iii) If x is bounded, then x is also sub-Gaussian.

By following the same ideas as in our proof of the linear concentration for Gaussian random variables one can then show concentration inequalities for sub-Gaussian or for sub-exponential distributions. Let us state a precise version for the former case and make a remark on what is different in the latter case.

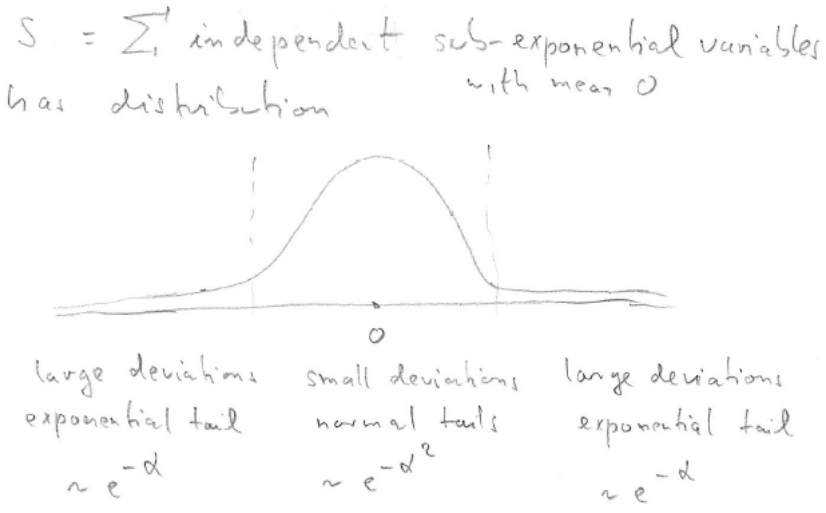
Theorem 3.6 (Hoeffding)**.** Let $t_1, \dots, t_p$ be real-valued independent sub-Gaussian random variables with parameters σ_i and μ_i for $i = 1, \dots, p$; i.e.

$$E\left[\exp\big(\lambda(t_i - \mu_i)\big)\right] \leq \exp\left(\sigma_i^2 \frac{\lambda^2}{2}\right).$$

Then, for all $\alpha \geq 0$,

$$\mathrm{P}\left\{\sum_{i=1}^{p}(t_i - \mu_i) \geq \alpha\right\} \leq \exp\left(-\frac{\alpha^2}{2\sum_{i=1}^{p}\sigma_i^2}\right).$$

Remark. For sub-exponential distributions one has a similar statement (the *Bernstein Inequality*), but then one has two different behaviours for small and for large deviations: if S is the sum of independent sub-exponential variables with mean zero, then small derivations (close to the mean) have normal tails ($\sim \exp(-\alpha^2)$), while large deviations (far from the mean) have exponential tails ($\sim \exp(-\alpha)$).



In the Lipschitz case, going away from normal distributions is more subtle, one needs stronger conditions on the function f. (One should in particular note that Lemma 3.4 is

only true for Gaussian distributions.) The following is a famous basic version of such an estimate, due to Talagrand. We will not address its proof; this would require new ideas.

Theorem 3.7 (Talagrand concentration inequality)**.** Let $t_1, \ldots, t_p$ be independent bounded random variables with $|t_i| \leq k$ for all $i = 1, \ldots, p$, for some $k > 0$. Let $f : \mathbb{R}^p \to \mathbb{R}$ be L-Lipschitz and convex. Then, for any $\alpha \geq 0$,

$$\mathrm{P}\left\{x = (t_1, \ldots, t_p) \in \mathbb{R}^p : |f(x) - E[f(\cdot)]| \geq \alpha k\right\} \leq c_1 \cdot \exp\left(-c_2 \cdot \frac{\alpha^2}{L^2}\right).$$

Note that there are counter-examples showing that the statement is not true without the convexity assumption.

4. Wishart Random Matrices

Consider data described by a Gaussian random vector $x \sim N(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p\times p}$ is the covariance matrix. Consider n independent draws from this distribution, i.e. $x_1, \ldots, x_n \in \mathbb{R}^p$ are independent with $x_k \sim N(0, \Sigma)$, and denote by $X = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \end{pmatrix} \in \mathbb{R}^{p\times n}$ the data matrix. Then

$$\hat{\Sigma} := \frac{1}{n} XX^T = \frac{1}{n} \sum_{k=1}^{n} x_k x_k^T,$$

called a *Wishart matrix*, is an estimator for Σ.

4.1. Concentration for the largest eigenvalue of Wishart matrices

We want to understand the spectral properties of $\hat{\Sigma}$, i.e. the eigenvalues of $\hat{\Sigma}$, or the singular values of X. Let us first restrict to the largest singular value, i.e. $\|X\|$. Recall that any matrix $X \in \mathbb{R}^{p\times n}$ can be identified with a map $X : \mathbb{R}^n \to \mathbb{R}^p$. As a map,

$$\|\cdot\| : \mathbb{R}^{p\times n} \cong \mathbb{R}^{pn} \to \mathbb{R}, \quad \|X\| := \sup_{\substack{v\in\mathbb{R}^n \\ \|v\|=1}} \|Xv\|$$

is Lipschitz: for $X_1, X_2 \in \mathbb{R}^{p\times n}$,

$$|\|X_1\| - \|X_2\|| \leq \|X_1 - X_2\| \leq \|X_1 - X_2\|_{\mathbb{R}^{pn}},$$

where we use the general estimate for the norm of a matrix compared to its norm as a vector: if $X = (x_{ij})_{\substack{i=1,\ldots,p \\ j=1,\ldots,n}}$, then

$$\|X\| \leq \|X\|_F := \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{n} |x_{ij}|^2},$$

where $\|\cdot\|_F$ is called the *Frobenius norm*. Thus $\|\cdot\|$ is 1-Lipschitz and we can apply our Gaussian concentration inequality for Lipschitz functions. Note that $X \sim N(0, I_{pn})$ corresponds to $\Sigma = I_p$.

Theorem 4.1. Suppose $X = (x_{ij})_{\substack{i=1,\ldots,p \\ j=1,\ldots,n}}$ is a standard Gaussian random matrix, i.e. all x_{ij} are independent and each $x_{ij} \sim N(0, 1)$ is a standard Gaussian variable. Then

$$\mathrm{P}\left\{|\|X\| - E[\|\cdot\|]| \geq \alpha\right\} \leq 2\exp\left(-\frac{\alpha^2}{2}\right).$$

This is a nice concentration about the expectation, but it does not tell us what the expectation is! Note that we can write

$$\|X\| = \max_{\substack{v\in\mathbb{R}^n \\ w\in\mathbb{R}^p \\ \|v\|=1=\|w\|}} \langle Xv, w\rangle$$

also as the maximum over (infinitely many!) terms which we can control better. We can go from infinitely many to finitely many conditions, by approximating all v and w by elements from ε-nets. It's a bit easier to do this for balls than for spheres, so let us write

$$\|X\| = \max_{\substack{\|v\|\le 1 \\ \|w\|\le 1}} \langle Xv, w\rangle,$$

and let $\mathcal{N}$ now be an ε-net for

$$\mathrm{B}_n = \{v \in \mathbb{R}^n : \|v\| \le 1\},$$

i.e. $\mathcal{N} \subset \mathrm{B}_n$ such that

$$\forall v \in \mathrm{B}_n \ \ \exists \tilde{v} \in \mathcal{N} : \|v - \tilde{v}\| < \varepsilon.$$

We want $\mathcal{N}$ to be as small as possible; it is easy to see that there exists an ε-net $\mathcal{N}$ with

$$|\mathcal{N}| \le \left(\frac{2}{\varepsilon} + 1\right)^n.$$

To see this, inductively construct an ε-net $v_1, v_2, v_3, \ldots$ by choosing a new point v_k such that

$$\mathrm{B}_n\left(v_k, \frac{\varepsilon}{2}\right) \cap \mathrm{B}_n\left(v_i, \frac{\varepsilon}{2}\right) = \emptyset \quad \text{for all} \quad i = 1, \ldots, k-1.$$

Note that all $\mathrm{B}_n\left(v_k, \frac{\varepsilon}{2}\right)$ are disjoint and

$$\bigcup_k \mathrm{B}_n\left(v_k, \frac{\varepsilon}{2}\right) \subset \mathrm{B}_n\left(0, 1 + \frac{\varepsilon}{2}\right),$$

since $v_k \in \mathrm{B}_n(0,1)$. Thus this inductive construction must stop after a finite number N of steps, for which we have

$$N \cdot \mathrm{vol}\left(\mathrm{B}_n\left(\frac{\varepsilon}{2}\right)\right) \le \mathrm{vol}\left(\mathrm{B}_n\left(1 + \frac{\varepsilon}{2}\right)\right),$$

therefore

$$N \le \frac{\mathrm{vol}\left(\mathrm{B}_n\left(1 + \frac{\varepsilon}{2}\right)\right)}{\mathrm{vol}\left(\mathrm{B}_n\left(\frac{\varepsilon}{2}\right)\right)} = \frac{\left(1 + \frac{\varepsilon}{2}\right)^n}{\left(\frac{\varepsilon}{2}\right)^n} = \left(\frac{2}{\varepsilon} + 1\right)^n.$$

Now $\mathcal{N} = \{v_1, \dots, v_N\}$ is an ε-net, since if there would be a point v with $\|v - v_k\| \geq \varepsilon$ for all $k = 1, \dots, N$, then this v could be chosen as v_{N+1} in our construction.

Let us fix for the following $\varepsilon = \frac{1}{4}$ and thus $\frac{2}{\varepsilon} + 1 = 9$. Then we can choose an ε-net $\mathcal{N}$ for B_n and an ε-net $\mathcal{M}$ for B_p with $|\mathcal{N}| \leq 9^n$ and $|\mathcal{M}| \leq 9^p$. Now we can estimate $\|X\|$ as a maximum where we only run over the finite nets $\mathcal{N}$ and $\mathcal{M}$. Let $v \in \mathrm{B}_n$ and $w \in \mathrm{B}_p$ be the maximizer for $\|X\| = \langle Xv, w\rangle$ (note that we are in finite dimensions, where the unit ball is compact, so that the supremum in the definition of the operator norm is indeed a maximum); then there exist $\tilde{v} \in \mathcal{N}$ and $\tilde{w} \in \mathcal{M}$ such that

$$\|v - \tilde{v}\| < \varepsilon \quad \text{and} \quad \|w - \tilde{w}\| < \varepsilon.$$

Then

$$\begin{aligned}
\langle Xv, w\rangle &= \langle X(\tilde{v} + (v - \tilde{v})), \tilde{w} + (w - \tilde{w})\rangle \\
&= \langle X\tilde{v}, \tilde{w}\rangle + \langle X\tilde{v}, w - \tilde{w}\rangle + \langle X(v - \tilde{v}), w\rangle \\
&\leq \langle X\tilde{v}, \tilde{w}\rangle + \|X\| \cdot \underbrace{\|\tilde{v}\|}_{=1} \cdot \underbrace{\|w - \tilde{w}\|}_{<\varepsilon} + \|X\| \cdot \underbrace{\|v - \tilde{v}\|}_{<\varepsilon} \cdot \underbrace{\|\tilde{w}\|}_{=1} \\
&\leq \langle X\tilde{v}, \tilde{w}\rangle + 2\varepsilon\|X\|,
\end{aligned}$$

thus

$$\|X\| \leq \max_{\substack{\tilde{v}\in\mathcal{N} \\ \tilde{w}\in\mathcal{M}}} \langle X\tilde{v}, \tilde{w}\rangle + 2\varepsilon\|X\|$$

and so

$$\|X\| \leq \frac{1}{1 - 2\varepsilon} \max_{\substack{v\in\mathcal{N} \\ w\in\mathcal{M}}} \langle Xv, w\rangle.$$

This means that if $\|X\| \geq \alpha$, then there exist $v \in \mathcal{N}$ and $w \in \mathcal{M}$ such that

$$\langle Xv, w\rangle \geq \alpha(1 - 2\varepsilon)$$

and thus

$$\begin{aligned}
\mathrm{P}\left\{\|X\| \geq \alpha\right\} &\leq \mathrm{P}\left\{ \bigcup_{\substack{v\in\mathcal{N} \\ w\in\mathcal{M}}} \left\{\langle Xv, w\rangle \geq \alpha(1 - 2\varepsilon)\right\} \right\} \\
&\leq \sum_{\substack{v\in\mathcal{N} \\ w\in\mathcal{M}}} \mathrm{P}\left\{\langle Xv, w\rangle \geq \alpha(1 - 2\varepsilon)\right\}.
\end{aligned}$$

Hence we need now the concentration inequality only for the finitely many summands on the right-hand side. Note that

$$\langle Xv, w\rangle = \sum_{i=1}^{p} \sum_{j=1}^{n} x_{ij} v_j w_i,$$

where all $x_{ij} \sim N(0,1)$ are independent, which yields for $\langle Xv, w\rangle$ a Gaussian distribution with variance

$$\sigma^2 = \sum_{i=1}^{p}\sum_{j=1}^{n} v_j^2 w_i^2 = \|v\|^2 \cdot \|w\|^2 \leq 1,$$

thus

$$\mathrm{P}\left\{\langle Xv, w\rangle \geq \alpha\right\} \leq \frac{1}{2}\exp\left(-\frac{\alpha^2}{2}\right),$$

or now with α replaced with $\alpha(1-2\varepsilon) = \frac{1}{2}\alpha$ for $\varepsilon = \frac{1}{4}$,

$$\mathrm{P}\left\{\langle Xv, w\rangle \geq \frac{\alpha}{2}\right\} \leq \frac{1}{2}\exp\left(-\frac{\alpha^2}{8}\right).$$

This then yields (since $9 \leq \exp(3)$)

$$\begin{aligned}
\mathrm{P}\left\{\|X\| \geq \alpha\right\} &\leq |\mathcal{N}| \cdot |\mathcal{M}| \cdot \frac{1}{2}\exp\left(-\frac{\alpha^2}{8}\right) \\
&= \frac{1}{2} \cdot 9^n \cdot 9^p \cdot \exp\left(-\frac{\alpha^2}{8}\right) \\
&\leq \frac{1}{2}\exp\left(3n + 3p - \frac{\alpha^2}{8}\right) \\
&= \frac{1}{2}\exp\left(-\frac{1}{8}(\alpha^2 - 24n - 24p)\right).
\end{aligned}$$

Put $\alpha = \sqrt{24}(\sqrt{n} + \sqrt{p}) + u$, then $\alpha^2 \geq u^2 + 24n + 24p$ and thus

$$\mathrm{P}\left\{\|X\| \geq \sqrt{24}(\sqrt{n} + \sqrt{p}) + u\right\} \leq \frac{1}{2}\exp\left(-\frac{u^2}{8}\right).$$

The term $\sqrt{24}(\sqrt{n}+\sqrt{p})$ gives the right order of $\|\cdot\|$, but the constants are off quite a bit.

In order to get a more precise understanding of the maximal eigenvalue (or its expected value), in particular in the asymptotic regime $n, p \to \infty$, one needs other tools. We will not pursue this any further, but instead we will now look at the collection of all singular values of X or of all eigenvalues of $\hat{\Sigma}$; i.e., we want now to understand the asymptotics of the histograms of the eigenvalues.

4.2. Eigenvalue distribution of Wishart matrices and Marchenko-Pastur law

Recall our setting: $x_1, \ldots, x_n \in \mathbb{R}^p$ are independent vectors with $x_k \sim N(0, \Sigma)$. We now restrict to the case $\Sigma = I_p$. With $X = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \end{pmatrix} \in \mathbb{R}^{p\times n}$ we have the Wishart

matrix

$$\hat{\Sigma} := \frac{1}{n} X X^T = \frac{1}{n} \sum_{k=1}^{n} x_k x_k^T \in \mathbb{R}^{p \times p}.$$

What can we say about the eigenvalues of $\hat{\Sigma}$?

Let $\lambda_1(A) \leq \lambda_2(A) \leq \ldots \leq \lambda_p(A)$ be the eigenvalues of a symmetric matrix $A = A^T \in \mathbb{R}^{p \times p}$. Then one has, as for the maximal eigenvalue,

$$|\lambda_i(A) - \lambda_i(B)| \leq \|A - B\| \leq \|A - B\|_F,$$

i.e. the maps $A \mapsto \lambda_i(A)$ for $i = 1, \ldots, p$ are Lipschitz and thus also the map

$$A \mapsto (\lambda_1(A), \ldots, \lambda_p(A))$$

is Lipschitz. However: since X is our matrix with independent Gaussian entries, we are interested in the mapping

$$X \mapsto \left(\lambda_1 \left(\frac{1}{n} X X^T \right), \ldots, \lambda_p \left(\frac{1}{n} X X^T \right) \right).$$

For this, the Lipschitz constant is modified as follows:

$$\begin{aligned}
\left| \lambda_i \left(\frac{1}{n} X X^T \right) - \lambda_i \left(\frac{1}{n} Y Y^T \right) \right| &\leq \frac{1}{n} \left\| X X^T - Y Y^T \right\| \\
&= \frac{1}{n} \left\| X X^T - X Y^T + X Y^T - Y Y^T \right\| \\
&\leq \frac{1}{n} \Big(\|X\| \cdot \underbrace{\left\| X^T - Y^T \right\|}_{\leq \|X - Y\|_F} + \underbrace{\|X - Y\|}_{\leq \|X - Y\|_F} \cdot \left\| Y^T \right\| \Big) \\
&\leq \frac{1}{n} \big(\|X\| + \|Y\| \big) \cdot \|X - Y\|_F \\
&\leq \frac{2}{n} \max\big(\|X\|, \|Y\| \big) \cdot \|X - Y\|_F.
\end{aligned}$$

Thus, the Lipschitz constant is bounded by

$$L = \frac{2}{n} \max\big(\|X\|, \|Y\| \big).$$

Note that the estimate $\|X\| \leq \|X\|_F = \|X\|_{\mathbb{R}^{pn}}$ is not helpful, since we know that $\|X\|_{\mathbb{R}^{pn}} \sim \sqrt{pn}$. But let us have a closer look on this, as it also reveals the difference between classical and modern regimes.

- In the classical regime, where p is fixed and $n \to \infty$, this would be okay, since then

$$L \sim 2\sqrt{p}\frac{1}{n}\sqrt{n} \sim c\frac{1}{\sqrt{n}},$$

  which would give good concentration.

- But we now are interested in the "modern regime", where $p \sim n \to \infty$, say $p = \gamma n$ for fixed γ. Then

$$L \sim 2\frac{1}{n}\sqrt{\gamma n \cdot n} = 2\sqrt{\gamma} \sim \text{constant},$$

  which does not give good concentration.

So let's keep the operator norm $\|X\|$ in L; for this we already know that we have good concentration around $\sim c(\sqrt{n} + \sqrt{p}) \sim c(1 + \sqrt{\gamma})\sqrt{n}$ and thus with high probability

$$L \sim 2\frac{1}{n}c(1 + \sqrt{\gamma})\sqrt{n} \sim \tilde{c}\frac{1}{\sqrt{n}}.$$

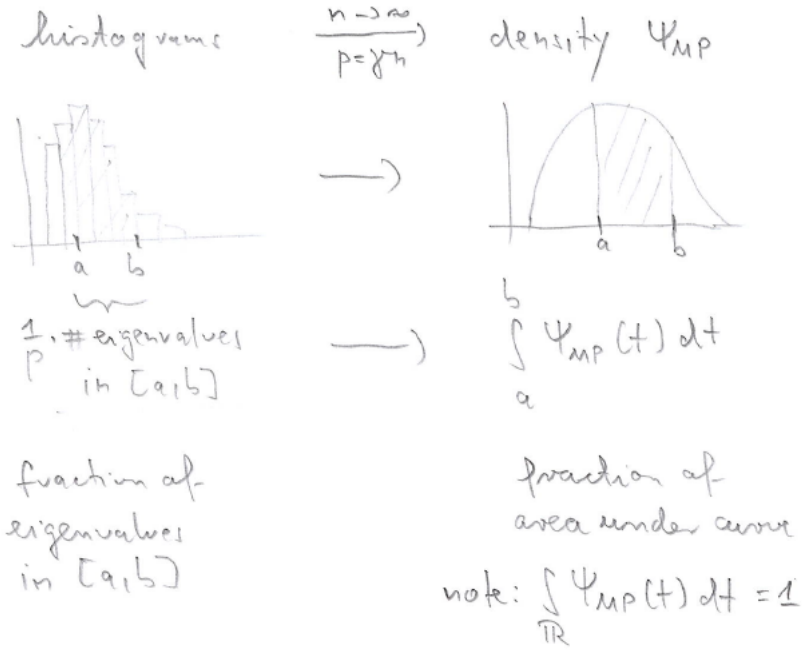
By Theorem 3.2, this then gives concentration of $\lambda_i(\hat{\Sigma})$ around its expected value with

$$2\exp\left(-\frac{\alpha^2}{2L^2}\right) \sim 2\exp\left(-\frac{n\alpha^2}{2\tilde{c}^2}\right)$$

as bound for the probability of deviation α from the mean.

This means: in the regime $p = \gamma n$ for fixed γ and $n \to \infty$, the eigenvalue distribution of $\hat{\Sigma}$ concentrates on its average. The scaling factor $\frac{1}{n}$ in $\hat{\Sigma}$ ensures that we have a limit for $n \to \infty$.

In numerical simulations we have seen that in this regime ($p = \gamma n \to \infty$) we have a nice asymptotic form for the (averaged and thus also typical) eigenvalue distribution.



Theorem 4.2 (Marchenko-Pastur Law, 1967, [MP67]). Let $X \in \mathbb{R}^{p\times n} \cong \mathbb{R}^{np}$ be our standard Gaussian random matrix (where all entries $x_{ij} \sim N(0,1)$). If $\frac{p}{n} \to \gamma \in (0,1]$ as $n \to \infty$, then the histogram of the eigenvalues of $\hat{\Sigma} = \frac{1}{n}XX^T$ converges to the Marchenko-Pastur density

$$\psi_{\mathrm{MP}}(t) = \frac{1}{2\pi\gamma t}\sqrt{(\gamma_+ - t)(t-\gamma_-)} \quad \text{on} \quad [\gamma_-, \gamma_+],$$

where

$$\gamma_- := (1-\sqrt{\gamma})^2 \quad \text{and} \quad \gamma_+ := (1+\sqrt{\gamma})^2.$$

Remark. Note that the statement is of the form

$$\frac{1}{p}\sum_{i=1}^{p} f(\lambda_i) \xrightarrow[p=\gamma n]{n\to\infty} \int f(t)\psi_{\mathrm{MP}}(t)\,\mathrm{d}t, \tag{1}$$

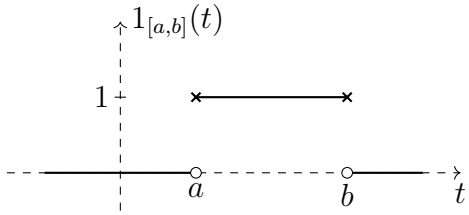
where $f = 1_{[a,b]}$ is the characteristic function of the interval $[a,b]$ (see Figure 3) and $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\hat{\Sigma}$. Proving (1) directly is not so clear, but can be achieved by proving analogous statements for other classes of functions. Instead of proving (1) for

(i) all $f = 1_{[a,b]}$ for all $a < b$,

one proves it for

(ii) all moments $f(t) = t^n$ for all $n \in \mathbb{N}$, or

(iii) all resolvents $f(t) = \frac{1}{t-z}$ for all $z \in \mathbb{C}_+$. ($\mathbb{C}_+$ denotes the complex upper half plane.)



$$1_{[a,b]}(t) = \begin{cases} 1, & t \in [a,b], \\ 0, & t \notin [a,b] \end{cases}$$

Figure 3: the characteristic function $1_{[a,b]}$ of an interval $[a,b]$.

In our situation, either (ii) or (iii) are equivalent to (i). By concentration, it suffices to prove in each case the version for the average, i.e. one has to prove

$$\frac{1}{p}\sum_{i=1}^{p} E\big[f(\lambda_i)\big] \to \int f(t)\psi_{\mathrm{MP}}(t)\,\mathrm{d}t. \tag{2}$$

Note for this that $\frac{1}{t-z}$ and t^n (if we restrict them to a compact interval) are Lipschitz functions.

We will give in the following the main ideas for the proof of (2) in the case (iii).

Lemma 4.3. Let $A \in \mathbb{R}^{p\times p}$ be a symmetric matrix, i.e. $A = A^T$, with (necessarily real) eigenvalues $\lambda_1, \ldots, \lambda_p$. Then, for any $z \in \mathbb{C} \setminus \{\lambda_1, \ldots, \lambda_p\}$, we have

$$\frac{1}{p}\sum_{i=1}^{p} \frac{1}{\lambda_i - z} = \mathrm{tr}\left[(A - zI_p)^{-1}\right],$$

where $\mathrm{tr} = \frac{1}{p}\,\mathrm{Tr}$ is the normalized trace on $\mathbb{R}^{p\times p}$.

Proof. Since $A = A^T$, it can be diagonalized by an orthogonal matrix $\mathcal{O}$, i.e. $\mathcal{O}\mathcal{O}^T = I_p = \mathcal{O}^T\mathcal{O}$, in the form $A = \mathcal{O}D\mathcal{O}^T$, where

$$D = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{pmatrix}.$$

Then, $A - zI_p = \mathcal{O}(D - zI_p)\mathcal{O}^T$ and thus $(A - zI_p)^{-1} = \mathcal{O}(D - zI_p)^{-1}\mathcal{O}^T$, hence

$$\mathrm{tr}\left[(A - zI_p)^{-1}\right] = \frac{1}{p}\,\mathrm{Tr}\left[\mathcal{O}(D - zI_p)^{-1}\mathcal{O}^T\right] = \frac{1}{p}\,\mathrm{Tr}\left[(D - zI_p)^{-1}\right] = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i - z}$$

since

$$(D - zI_p)^{-1} = \begin{pmatrix} \lambda_1 - z & & & 0 \\ & \lambda_2 - z & & \\ & & \ddots & \\ 0 & & & \lambda_p - z \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} (\lambda_1 - z)^{-1} & & & 0 \\ & (\lambda_2 - z)^{-1} & & \\ & & \ddots & \\ 0 & & & (\lambda_p - z)^{-1} \end{pmatrix}. \qquad \square$$

Note that, independent of the specific form of $A = A^T$, this always makes sense if $z \notin \mathbb{R}$.

Definition 4.4. (1) For our Wishart matrices $\hat{\Sigma} = \frac{1}{n} X X^T \in \mathbb{R}^{p \times p}$ we define their *Stieltjes transform* as

$$S_n(z) := E\left[\operatorname{tr}\left[(\hat{\Sigma} - zI_p)^{-1}\right]\right] \quad \text{for} \quad z \in \mathbb{C} \setminus \mathbb{R}.$$

(2) For the Marchenko-Pastur distribution ψ_{MP} we define its *Stieltjes transform* as

$$S_{\mathrm{MP}}(z) := \int \frac{1}{t - z} \psi_{\mathrm{MP}} \, \mathrm{d}t \quad \text{for} \quad z \in \mathbb{C} \setminus \mathbb{R}.$$

So what we have to prove is the convergence of the Stieltjes transforms:

$$S_n(z) \xrightarrow[p = \gamma n]{n \to \infty} S_{\mathrm{MP}}(z) \quad \text{for all} \quad z \in \mathbb{C} \setminus \mathbb{R}.$$

Remark. We will derive an equation for

$$S(z) := \lim_{\substack{n \to \infty \\ p = \gamma n}} S_n(z);$$

of course, it has to be proven that this limit exists, but we will not bother about this. Then we have to check that $S(z) = S_{\mathrm{MP}}(z)$. Note that for $X = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}$,

$$S_n(z) = E\left[\operatorname{tr}\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_k x_k^T - zI_p\right)^{-1}\right]\right] = \frac{n}{p} E\left[\operatorname{Tr}\left[\left(\sum_{i=1}^{n} x_k x_k^T - nzI_p\right)^{-1}\right]\right],$$

where we can interpret

$$\sum_{i=1}^{n} x_k x_k^T = \left(\sum_{i=1}^{n-1} x_k x_k^T\right) + x_n x_n^T$$

as a deformation of $\sum_{i=1}^{n-1} x_k x_k^T$ with the rank-1-matrix $x_n x_n^T$.

Lemma 4.5 (Sherman-Morrison Formula)**.** Let $A \in \mathbb{R}^{p\times p}$ and $x, y \in \mathbb{R}^p$. Then

$$(A + xy^T)^{-1} = A^{-1} - \frac{A^{-1}xy^TA^{-1}}{1 + y^TA^{-1}x}$$

whenever A is invertible and $1 + y^TA^{-1}x \neq 0$.

(Note: $xy^T \in \mathbb{R}^{p\times p}$ is a $(p \times p)$-matrix of rank 1, $\langle y, x\rangle = y^Tx \in \mathbb{R}^{1\times 1}$ is a real number, and $y^TA^{-1}x = \langle y, A^{-1}x\rangle \in \mathbb{R}$.)

Proof. Either check that the RHS is the inverse of the LHS, or calculate formally

$$\begin{aligned}
(A + xy^T)^{-1} &= A^{-1}(I_p + xy^TA^{-1})^{-1} \\
&= A^{-1}\big(I_p - xy^TA^{-1} + x\underbrace{y^TA^{-1}x}_{\in\mathbb{R}}y^TA^{-1} \mp \ldots\big) \\
&= A^{-1}\left(I_p - xy^TA^{-1}\left(\sum_{k=0}^{\infty}(-y^TA^{-1}x)^k\right)\right) \\
&= A^{-1}\left(I_p - xy^TA^{-1}\cdot\frac{1}{1 + y^TA^{-1}x}\right) \\
&= A^{-1} - \frac{A^{-1}xy^TA^{-1}}{1 + y^TA^{-1}x}.
\end{aligned}$$

□

From Lemma 4.5 we also get

$$y^T(A + xy^T)^{-1}x = y^TA^{-1}x - \frac{(y^TA^{-1}x)(y^TA^{-1}x)}{1 + y^TA^{-1}x} = \frac{y^TA^{-1}x}{1 + y^TA^{-1}x}.$$

We will now choose $x = y = x_n$ and

$$A = \sum_{k=1}^{n-1} x_kx_k^T - nzI_p.$$

Note that A, and thus also A^{-1}, is independent from x_n.

Lemma 4.6. Let $x \sim N(0, I_p)$ be a standard Gaussian vector in $\mathbb{R}^p$ and $B \in \mathbb{R}^{p\times p}$ a deterministic or random matrix independent from x. Then

$$E_x[x^TBx] = \mathrm{Tr}(B).$$

Proof. Let $B = (b_{ij})_{i,j=1}^p$. We have

$$E_x[x^TBx] = E_x[\langle x, Bx\rangle] = \sum_{i,j=1}^p E[x_ib_{ij}x_j] = \sum_{i,j=1}^p b_{ij}E[x_ix_j] = \sum_{i,j=1}^p b_{ij}\delta_{ij} = \sum_{i=1}^p b_{ii}$$

$$= \mathrm{Tr}(B).$$ □

Proof of Theorem 4.2. Let us apply all this to

$$S_n(z) = \frac{n}{p} E\left[\operatorname{Tr}\left[\left(\sum_{i=1}^{n-1} x_k x_k^T - nzI_p + x_n x_n^T\right)^{-1}\right]\right],$$

and denote

$$A := \sum_{i=1}^{n-1} x_k x_k^T - nzI_p \quad \text{as well as} \quad B := A + x_n x_n^T.$$

Then

$$x_n^T B^{-1} x_n = \frac{x_n^T A^{-1} x_n}{1 + x_n^T A^{-1} x_n} \approx \frac{\operatorname{Tr}(A^{-1})}{1 + \operatorname{Tr}(A^{-1})}.$$

This is actually true if we replace x_n on the LHS by x_k for any $k = 1, \dots, n$. Note that A depends on the choice of k,[2] but by concentration, they are all close to $E[\dots]$, which is independent of k. Now

$$\frac{\operatorname{Tr}(A^{-1})}{1 + \operatorname{Tr}(A^{-1})} \approx x_k^T B^{-1} x_k = \operatorname{Tr}(\underbrace{x_k^T B^{-1} x_k}_{\in \mathbb{R}}) = \operatorname{Tr}(\underbrace{x_k x_k^T B^{-1}}_{\in \mathbb{R}^{p \times p}})$$

and thus

$$\begin{aligned}
n \cdot \frac{\operatorname{Tr}(A^{-1})}{1 + \operatorname{Tr}(A^{-1})} &\approx \sum_{k=1}^{n} x_k^T B^{-1} x_k \\
&= \sum_{k=1}^{n} \operatorname{Tr}(x_k x_k^T B^{-1}) \\
&= \operatorname{Tr}\left(\sum_{k=1}^{n} x_k x_k^T B^{-1}\right) \\
&= \operatorname{Tr}\left(\left(\sum_{k=1}^{n} x_k x_k^T\right) B^{-1}\right) \\
&= \operatorname{Tr}((B + nzI_p) B^{-1}) \\
&= \operatorname{Tr}(I_p + nzB^{-1}) \\
&= p + nz \operatorname{Tr}(B^{-1})
\end{aligned}$$

or

$$\frac{p}{n} + z \operatorname{Tr}(B^{-1}) \approx \frac{\operatorname{Tr}(A^{-1})}{1 + \operatorname{Tr}(A^{-1})}.$$

[2] For $k = 1$, e.g., A would be $\sum_{i=2}^{n} x_i x_i^T - nzI_p$ and thus $B = A + x_1 x_1^T$.

Now, for large n and by concentration

$$\mathrm{Tr}(B^{-1}) \approx E[\mathrm{Tr}(B^{-1})] = \frac{p}{n} S_n(z)$$

and

$$\begin{aligned}
\mathrm{Tr}(A^{-1}) &= \mathrm{Tr}\left(\sum_{k=1}^{n-1} x_k x_k^T - nzI_p\right) \\
&\approx E\left[\mathrm{Tr}\left(\sum_{k=1}^{n-1} x_k x_k^T - (n-1)zI_p\right)\right] \\
&= \frac{p}{n-1} S_{n-1}(z) \\
&\approx \frac{p}{n} S_n(z).
\end{aligned}$$

Thus, with $\gamma = \frac{p}{n}$:

$$\gamma + z\gamma S_n(z) \approx \frac{\gamma S_n(z)}{1 + \gamma S_n(z)}$$

or in the limit with $S(z) := \lim_{\substack{n\to\infty \\ p=\gamma n}} S_n(z)$[3]

$$1 + zS_n(z) = \frac{S(z)}{1 + \gamma S(z)},$$

i.e.

$$\gamma z S(z)^2 - (1 - z - \gamma) S(z) + 1 = 0.$$

This has the solution[4]

$$\begin{aligned}
S(z) &= \frac{1 - z - \gamma + \sqrt{(z + \gamma - 1)^2 - 4\gamma z}}{2\gamma z} \\
&= \frac{1 - z - \gamma + \sqrt{(z - (1 + \sqrt{\gamma})^2)(z - (1 - \sqrt{\gamma})^2)}}{2\gamma z} \\
&= \frac{1 - z - \gamma + \sqrt{(z - \gamma_+)(z - \gamma_-)}}{2\gamma z}.
\end{aligned}$$

Then check that this is the Stieltjes transform of ψ_{MP}; or, better, calculate ψ_{MP} from this $S(z)$ by the Stieltjes inversion formula! □

[3] One has to argue, by abstract compactness arguments, that the limit of $S_n(z)$ exists.

[4] Replacing the plus sign in front of the square root with a minus sign does not result in a solution, since for Stieltjes transforms, $S(z) \in \mathbb{C}_+$ for $z \in \mathbb{C}_+$.

Lemma 4.7 (Stieltjes Inversion Formula)**.** Let ψ be a continuous probability density on $\mathbb{R}$. Then its Stieltjes transform

$$S(z) := \int \frac{1}{t-z}\psi(t)\,\mathrm{d}t \quad \text{for} \quad z \in \mathbb{C}_+$$

has a continuous extension to $\mathbb{C}_+ \cup \mathbb{R}$ and

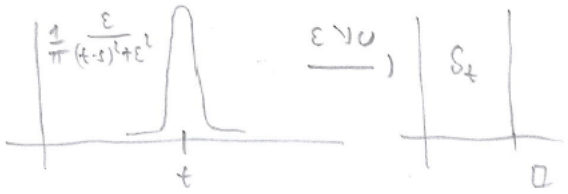
$$\psi(t) = \frac{1}{\pi}\operatorname{Im}(S(t)) = \frac{1}{\pi}\lim_{\varepsilon\to 0}\operatorname{Im}(S(t+\mathrm{i}\varepsilon)).$$

Proof. For all $z \in \mathbb{C}\setminus\{0\}$, we have

$$\operatorname{Im}\left(\frac{1}{z}\right) = \frac{1}{2\mathrm{i}}\left(\frac{1}{z}-\frac{1}{\overline{z}}\right) = \frac{1}{2\mathrm{i}}\cdot\frac{\overline{z}-z}{z\cdot\overline{z}} = -\frac{\operatorname{Im}(z)}{z\cdot\overline{z}}$$

and thus

$$\begin{aligned}\frac{1}{\pi}\operatorname{Im}(S(t+\mathrm{i}\varepsilon)) &= \frac{1}{\pi}\int \operatorname{Im}\left(\frac{1}{s-(t+\mathrm{i}\varepsilon)}\right)\psi(s)\,\mathrm{d}s \\ &= \frac{1}{\pi}\int \frac{\varepsilon}{(t-s)^2+\varepsilon^2}\psi(s)\,\mathrm{d}s \xrightarrow{\varepsilon\searrow 0} \int \delta_t(s)\psi(s)\,\mathrm{d}s = \psi(t). \qquad \square\end{aligned}$$

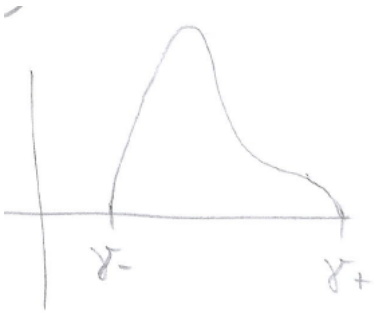


Apply this to

$$S(z) = \frac{1-z-\gamma+\sqrt{(z-\gamma_+)(z-\gamma_-)}}{2\gamma z},$$

then we get the form of the Marchenko-Pastur distribution as claimed in Theorem 4.2

$$\begin{aligned}\frac{1}{\pi}\operatorname{Im}(S(t)) &= \frac{1}{\pi}\frac{\operatorname{Im}(\sqrt{(t-\gamma_+)(t-\gamma_-)})}{2\gamma t} \\ &= \begin{cases}\dfrac{1}{2\pi\gamma t}\sqrt{(\gamma_+-t)(t-\gamma_-)}, & t\in[\gamma_-,\gamma_+], \\ 0, & \text{otherwise.}\end{cases}\end{aligned}$$



5. Spiked Signal+Noise Models

5.1. Statement of BBP transition

Consider a covariance of the form

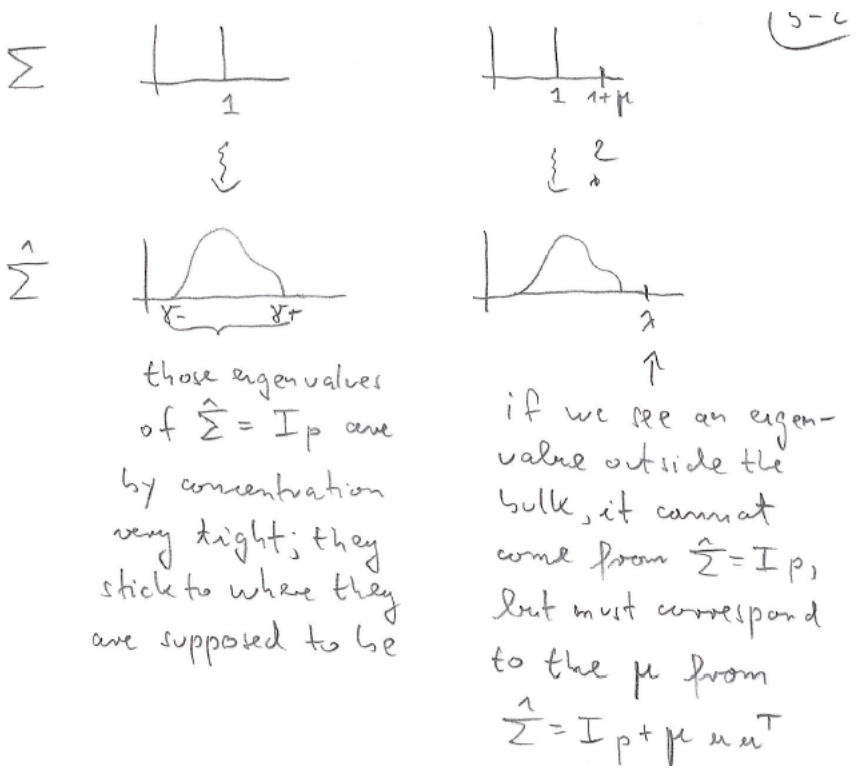
$$\Sigma = I_p + \mu u u^T,$$

where I_p is the "noise" and $\mu u u^T$ is the "signal" of strength $\mu > 0$ in direction u, where $\|u\| = 1$. In particular, $\mu u u^T$ is a rank 1 deformation of I_p.

We now want to address the following question: Can we see – and in particular, under which conditions – the signal in our corresponding Wishart matrix

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} x_k x_k^T$$

when we make n independent observations of $x \sim N(0, \Sigma)$? Does $\hat{\Sigma}$ have an eigenvalue λ that corresponds to μ and is it visible among the eigenvalues of $\hat{\Sigma}$ (in our usual regime $\frac{p}{n} \to \gamma$)?



In order to see the shadow λ of the eigenvalue μ, this λ must be at least γ_+, so that we see it as an outlier or a spike. Since we have concentration of the eigenvalues in the

Marchenko-Pastur bulk, none of those eigenvalues will appear as an outlier; thus, if we see something there, it must come from μ.

Our main question is now, whether we can control the relation between λ and μ? This is indeed the case; here is the main statement on this.

Theorem 5.1. Consider, for fixed $\mu \geq 0$, as above a covariance matrix of the form $\Sigma = I_p + \mu uu^T$.

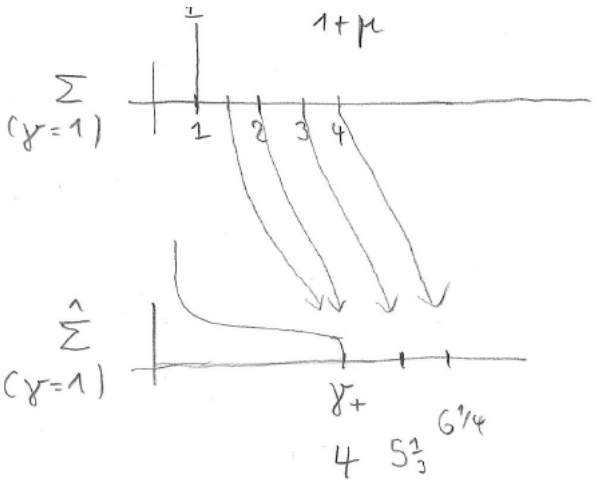
Then, in the asymptotic regime $\frac{p}{n} \to \gamma \in (0, 1]$, the largest eigenvalue λ of $\hat{\Sigma}$ is given by

$$\lambda = \begin{cases} 1 + \mu + \gamma\frac{1+\mu}{\mu}, & \text{if } \mu > \sqrt{\gamma}, \\ \gamma_+ = (1 + \sqrt{\gamma})^2, & \text{if } \mu \leq \sqrt{\gamma}. \end{cases}$$

Note that $1 + \mu + \gamma\frac{1+\mu}{\mu} > \gamma_+$ for $\mu > \sqrt{\gamma}$ and for $\mu = \sqrt{\gamma}$ we have

$$1 + \sqrt{\gamma} + \gamma\frac{1 + \sqrt{\gamma}}{\sqrt{\gamma}} = 1 + 2\sqrt{\gamma} + \gamma = (1 + \sqrt{\gamma})^2 = \gamma_+,$$

thus λ is visible as long as $\mu > \sqrt{\gamma}$ and for $\mu \leq \sqrt{\gamma}$ it is swallowed by the bulk. This theorem goes back to the works of Baik and Silverstein [BS06] as well as of Baik, Ben Arous and Péché [BBAP05] (where more spikes and also statements on the fluctuations of the outliers are treated). According to the latter work, the phenomenon is also known as *BBP transition*.



5.2. Proof of BBP transition

Proof of Theorem 5.1, part one. We write

$$\hat{\Sigma} = \frac{1}{n}XX^T = \frac{1}{n}\Sigma^{\frac{1}{2}}YY^T\Sigma^{\frac{1}{2}},$$

where $x_k = \Sigma^{\frac{1}{2}} y_k$ and $y_k \sim N(0, I_p)$, thus

$$X = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix} = \begin{pmatrix} \Sigma^{\frac{1}{2}} y_1 & \dots & \Sigma^{\frac{1}{2}} y_n \end{pmatrix} = \Sigma^{\frac{1}{2}} \begin{pmatrix} y_1 & \dots & y_n \end{pmatrix} = \Sigma^{\frac{1}{2}} Y.$$

Now $\frac{1}{n} YY^T$ is our Wishart matrix corresponding to covariance I_p and with distribution ψ_{MP} according to Theorem 4.2.

We are looking for an eigenvalue λ of $\hat{\Sigma}$, outside of the support $[\gamma_-, \gamma_+]$ of ψ_{MP}! Thus

$$\begin{aligned}
0 &= \det\left(\frac{1}{n} XX^T - \lambda I_p\right) \\
&= \det\left(\frac{1}{n} \Sigma^{\frac{1}{2}} YY^T \Sigma^{\frac{1}{2}} - \lambda I_p\right) \\
&= \det\left(\Sigma^{\frac{1}{2}} \left(\frac{1}{n} YY^T - \lambda \Sigma^{-1}\right) \Sigma^{\frac{1}{2}}\right) \\
&= \det(\Sigma) \cdot \det\left(\frac{1}{n} YY^T - \lambda \Sigma^{-1}\right) \\
&= \underbrace{\det(I_p + \mu uu^T)}_{\neq 0} \cdot \det\left(\frac{1}{n} YY^T - \lambda \Sigma^{-1}\right),
\end{aligned}$$

so $\det\left(\frac{1}{n} YY^T - \lambda \Sigma^{-1}\right) = 0$. By Lemma 4.5, we have

$$\Sigma^{-1} = (I_p + \mu uu^T)^{-1} = I_p - \mu \frac{uu^T}{1 + \mu uu^T u} = I_p - \mu \frac{uu^T}{1 + \mu},$$

since $u^T u = \|u\|^2 = 1$, hence

$$\begin{aligned}
0 &= \det\left(\frac{1}{n} YY^T - \lambda \Sigma^{-1}\right) \\
&= \det\left(\frac{1}{n} YY^T - \lambda I_p + \lambda \mu \frac{uu^T}{1 + \mu}\right) \\
&= \det\left(\left(\frac{1}{n} YY^T - \lambda I_p\right) \cdot \left(I_p + \left(\frac{1}{n} YY^T - \lambda I_p\right)^{-1} \cdot \lambda \mu \frac{uu^T}{1 + \mu}\right)\right) \\
&= \det\left(\frac{1}{n} YY^T - \lambda I_p\right) \cdot \det\left(I_p + \left(\frac{1}{n} YY^T - \lambda I_p\right)^{-1} \cdot \lambda \mu \frac{uu^T}{1 + \mu}\right).
\end{aligned}$$

Since $\frac{1}{n} YY^T - \lambda I_p$ is invertible for $\lambda \notin [\gamma_-, \gamma_+]$, its determinant is non-zero and we get

$$0 = \det\left(I_p + \frac{\lambda \mu}{1 + \mu} \left(\frac{1}{n} YY^T - \lambda I_p\right)^{-1} \cdot uu^T\right).$$

This is the determinant of a $(p \times p)$-matrix which we cannot control directly; but we can actually rewrite it as determinant of a (1×1)-matrix, which is accessible. Note that, for $A := \frac{1}{n} YY^T - \lambda I_p$, $A^{-1}u \in \mathbb{R}^{p \times 1}$ and $u^T \in \mathbb{R}^{1 \times p}$, thus $A^{-1}uu^T \in \mathbb{R}^{p \times p}$ but $u^T A^{-1} u \in \mathbb{R}^{1 \times 1}$. (To be continued...) □

Lemma 5.2 (Sylvester's Determinant Identity)**.** Consider $A \in \mathbb{R}^{p \times n}$ and $B \in \mathbb{R}^{n \times p}$. Then

$$\det(I_p + AB) = \det(I_n + BA).$$

Proof. Use "Schur complement":

$$\begin{aligned}
\begin{pmatrix} I_p & A \\ B & I_n \end{pmatrix} &= \begin{pmatrix} I_p & 0 \\ B & I_n \end{pmatrix} \begin{pmatrix} I_p & 0 \\ 0 & I_n - BA \end{pmatrix} \begin{pmatrix} I_p & A \\ 0 & I_n \end{pmatrix} \\
&= \begin{pmatrix} I_p & A \\ 0 & I_n \end{pmatrix} \begin{pmatrix} I_p - AB & 0 \\ 0 & I_n \end{pmatrix} \begin{pmatrix} I_p & 0 \\ B & I_n \end{pmatrix}.
\end{aligned}$$

Since the first and last factor are triangular matrices they have determinant 1, and we have

$$\begin{aligned}
\det(I_n - BA) &= \det(I_p) \cdot \det(I_n - BA) \\
&= \det \begin{pmatrix} I_p & 0 \\ 0 & I_n - BA \end{pmatrix} \\
&= \det \begin{pmatrix} I_p & A \\ B & I_n \end{pmatrix} \\
&= \det \begin{pmatrix} I_p - AB & 0 \\ 0 & I_n \end{pmatrix} \\
&= \det(I_p - AB) \cdot \det(I_n) \\
&= \det(I_p - AB).
\end{aligned}$$

□

Proof of Theorem 5.1, part two. Continuing the proof, we have

$$\begin{aligned}
0 &= \det\left(I_p + \frac{\lambda\mu}{1+\mu} \left(\frac{1}{n} YY^T - \lambda I_p \right)^{-1} uu^T \right) \\
&\overset{5.2}{=} \det\left(1 + \frac{\lambda\mu}{1+\mu} u^T \left(\frac{1}{n} YY^T - \lambda I_p \right)^{-1} u \right) \\
&= 1 + \frac{\lambda\mu}{1+\mu} u^T \left(\frac{1}{n} YY^T - \lambda I_p \right)^{-1} u \\
&= 1 + \frac{\lambda\mu}{1+\mu} \left\langle u, \left(\frac{1}{n} YY^T - \lambda I_p \right)^{-1} u \right\rangle \\
&\approx 1 + \frac{\lambda\mu}{1+\mu} S(\lambda),
\end{aligned}$$

where $S(\lambda)$ is the Stieltjes transform of the Marchenko-Pastur distribution of $\frac{1}{n}YY^T$. To justify the last step, recall that in the proof of Theorem 4.2 we have seen that

$$\operatorname{tr}\left(\left(\frac{1}{n}YY^T - \lambda I_p\right)^{-1}\right) \approx S(\lambda);$$

but actually for an orthonormal basis $u_1, \ldots, u_p$ we have

$$\begin{aligned}\operatorname{tr}\left(\left(\frac{1}{n}YY^T - \lambda I_p\right)^{-1}\right) &= \frac{1}{p}\sum_{k=1}^{p}\left\langle u_k, \left(\frac{1}{n}YY^T - \lambda I_p\right)^{-1} u_k\right\rangle \\ &= \left\langle u_1, \left(\frac{1}{n}YY^T - \lambda I_p\right)^{-1} u_1\right\rangle,\end{aligned}$$

since the summands are all the same due to rotational symmetry. But then we can choose any unit vector u as part $u_1 = u$ of an orthonormal basis.

The $\approx$ sign has of course to be understood as saying that with high probability one is close to an equality. In the following we will ignore these technicalities and just write "=".

Thus our condition on λ is

$$0 = 1 + \frac{\lambda\mu}{1+\mu}S(\lambda),$$

i.e.,

$$\lambda S(\lambda) = -\frac{1+\mu}{\mu}.$$

For a given μ, we have to check whether this has a solution λ with $\lambda > \gamma_+$. Recall that $S(\lambda)$ satisfies the equation

$$1 + \lambda S(\lambda) = \frac{S(\lambda)}{1+\gamma S(\lambda)} = \frac{1}{\frac{1}{S(\lambda)} + \gamma}$$

or

$$1 + \gamma S(\lambda) + \lambda S(\lambda) + \lambda\gamma S(\lambda)^2 = S(\lambda),$$

which is equivalent to

$$-S(\lambda)\big(1 - \lambda - \gamma\lambda S(\lambda)\big) = -\gamma S(\lambda) - 1$$

and

$$1 - \lambda - \gamma\lambda S(\lambda) = \frac{1}{S(\lambda)} + \gamma.$$

This leads to

$$1 + \lambda S(\lambda) = \frac{1}{1 - \lambda - \gamma\lambda S(\lambda)}.$$

With $\lambda S(\lambda) = -\frac{1+\mu}{\mu}$ we get

$$-\frac{1}{\mu} = 1 - \frac{1+\mu}{\mu} = \frac{1}{1-\lambda-\gamma\frac{1+\mu}{\mu}},$$

which implies

$$1 - \lambda - \gamma\frac{1+\mu}{\mu} = -\mu$$

and finally

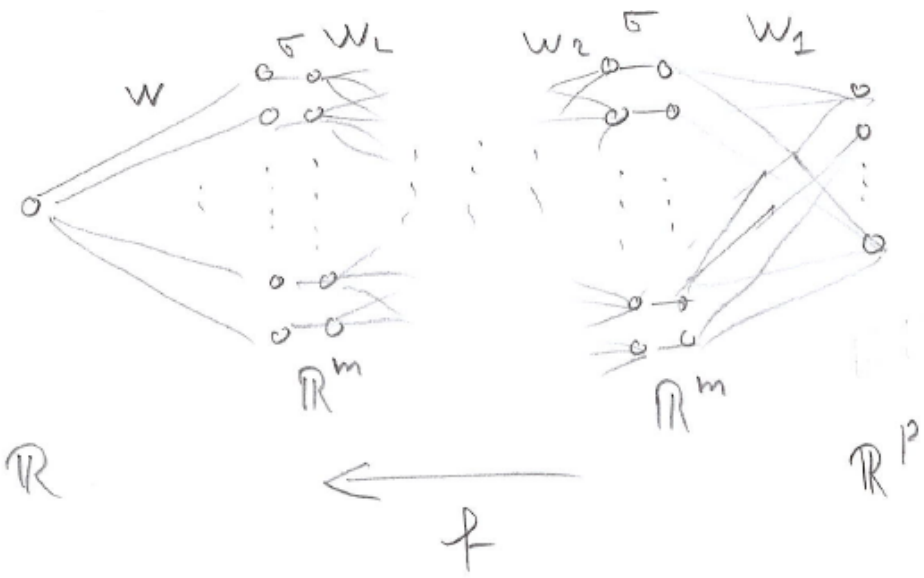
$$\lambda = 1 + \mu + \gamma\frac{1+\mu}{\mu}.$$

□

6. Neural Networks, Double Descent, and Linear Regression

6.1. Neural Networks

Neural networks are special functions, say $f : \mathbb{R}^p \to \mathbb{R}$, of the form



$$y = f(x) = w\sigma W_L \dots W_2 \sigma(W_1 x),$$

where

- $x \in \mathbb{R}^p$ is the input vector,
- $w \in \mathbb{R}^{1\times m}$ is a row vector,
- $W_2, \dots, W_L \in \mathbb{R}^{m\times m}$ are matrices,
- $W_1 \in \mathbb{R}^{m\times p}$ is a matrix, and
- $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear function that is applied entry-wise on vectors and matrices.

For multiple input vectors $x_1, \dots, x_n$, set

$$X := \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix} \in \mathbb{R}^{p\times n} \quad \text{and} \quad Y := \begin{pmatrix} y_1 & \dots & y_n \end{pmatrix} \in \mathbb{R}^{1\times n},$$

where $y_k := f(x_k)$ for all $k = 1, \dots, n$, then

$$Y = f(X) = w\sigma W_L \dots W_2 \sigma(W_1 X).$$

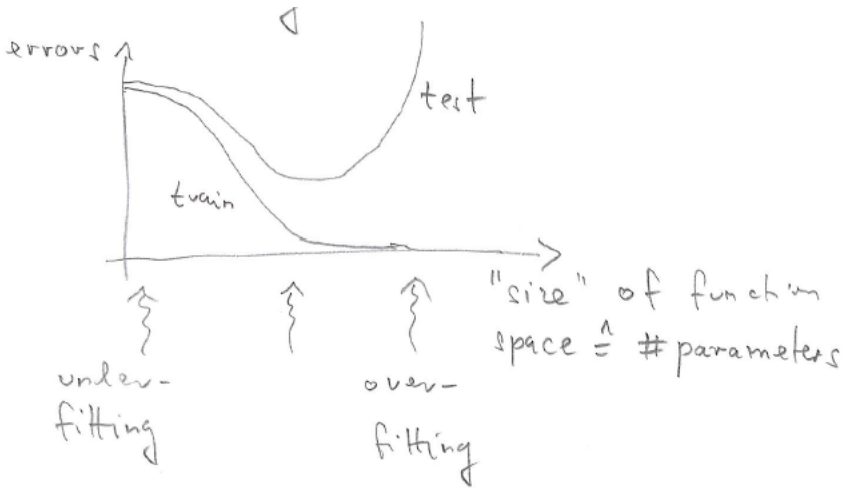
Two important questions on such neural networks are the following.

(1) Given such a neural network, what can we say about the relation between the input X and the output $Y = f(X)$; in particular, we are looking for probabilistic and asymptotic statements (if the width of the hidden layers goes to infinity)? The main rigorous statements of this form are saying that in this infinite width limit neural networks with random weights and biases converge to Gaussian processes. We will not say much on these aspects, but refer to the literature, like, e.g., [Han21, Yan19].

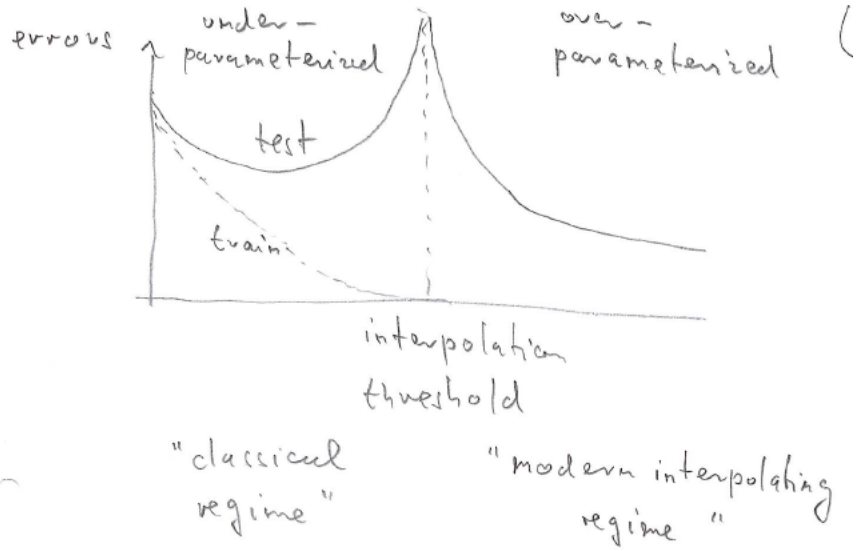
(2) How can we construct a neural network (i.e., choose its parameters) such that it approximates $f(X) = Y$ well for a given training set (X, Y), but also generalizes to unseen data. That's the question which we will address in the following.

6.2. The modern double descent picture

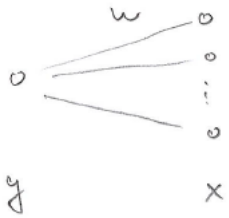
One has to note that neural networks challenge actually our classical ideas about learning; overparameterization (i.e., having much more parameters to adjust our functions than the number of data we want to learn) is considered bad classically.



But neural networks seem to indicate "the more the better" and in the last few years, the above picture has been replaced by the following "double descent" curve, see, e.g., [Bel21, MM22].



A kind of idea about this can be gotten already from the simplest neural network, without hidden layer and without a non-linearity.



Here, $y = f(x) = wx$ and $Y = wX$, where (X, Y) are given and we look for the "best" w to model this. This is then nothing else but the classical problem of "linear regression".

Assume we have a linear relation, but have noise in the measurements, thus we see

$$\hat{Y} = wX + N,$$

where N is some Gaussian noise. We want to find the best $\hat{w}$ such that we have

$$\hat{Y} = \hat{w}X.$$

Note that $X \in \mathbb{R}^{p\times n}$, $\hat{w} \in \mathbb{R}^{1\times p}$ and $\hat{Y} \in \mathbb{R}^{1\times n}$, i.e. we have a system of n linear equations for p variables. This is under-determined for $n < p$ and over-determined for $n > p$. These two cases will correspond to the two regimes in our double descent picture.

6.3. Linear regression: over-determined case

In the case $n > p$, $\hat{Y} = \hat{w}X$ typically will have no solution and we approximate the non-linear[5] relation between $\hat{Y}$ and X by the method of least squares: instead of $\hat{Y} = \hat{w}X$ we

[5]The non-linearity here is created by the noise.

try to solve

$$\hat{Y}X^T = \hat{w}XX^T;$$

this is the "normal equation", characterizing a $\hat{w}$ such that $\|\hat{Y} - \hat{w}X\|$ is minimal. Typically, i.e., if $\operatorname{rank}(X) = p < n$, the matrix $XX^T \in \mathbb{R}^{p\times p}$ is invertible, thus

$$\hat{w} = \hat{Y}X^T(XX^T)^{-1}.$$

Thus we have made the error

$$\begin{aligned}
\|w - \hat{w}\|^2 &= \|w - \underbrace{\hat{Y}}_{=wX+N} X^T(XX^T)^{-1}\|^2 \\
&= \|w - w\underbrace{XX^T(XX^T)^{-1}}_{=1} - NX^T(XX^T)^{-1}\|^2 \\
&= \|NX^T(XX^T)^{-1}\|^2 \\
&= NX^T(XX^T)^{-1}(XX^T)^{-1}XN^T,
\end{aligned}$$

since $\|a\|^2 = aa^*$ for all $a \in \mathbb{R}^{1\times p}$. Assume now that $N \sim N(0, \sigma^2 I_n)$ is a Gaussian noise and average the error over N:

$$\begin{aligned}
E_N\left[\|w - \hat{w}\|^2\right] &= E_N\left[NX^T(XX^T)^{-2}XN^T\right] \\
&\overset{4.6}{=} \sigma^2 \cdot \operatorname{Tr}\left(X^T(XX^T)^{-2}X\right) \\
&= \sigma^2 \cdot \operatorname{Tr}\left(XX^T(XX^T)^{-2}\right) \\
&= \sigma^2 \cdot \operatorname{Tr}\left((XX^T)^{-1}\right) \\
&= \sigma^2 \frac{p}{n} \operatorname{tr}\left(\left(\frac{1}{n}XX^T\right)^{-1}\right).
\end{aligned}$$

Assume now that X is a standard Gaussian random matrix, then $\hat{\Sigma} = \frac{1}{n}XX^T$ is a Wishart matrix and the above error converges for $n, p = \gamma n \to \infty$ to $\sigma^2\gamma S(0)$, where

$$S(z) = \int \frac{1}{t - z}\psi_{\mathrm{MP}}\,\mathrm{d}t$$

is the Stieltjes transform of the Marchenko-Pastur distribution. We know from the proof of Theorem 4.2 that $S(z)$ satisfies the equation

$$1 + zS(z) = \frac{S(z)}{1 + \gamma S(z)},$$

i.e., for $z = 0$ (note that S has a continuous extension to $\mathbb{R}$ for $\gamma < 1$)

$$1 = \frac{S(0)}{1 + \gamma S(0)}, \quad \text{so} \quad S(0) = \frac{1}{1 - \gamma}$$

and thus

$$E\left[\|w-\hat{w}\|^2\right]=\sigma^2\frac{\gamma}{1-\gamma}.$$

6.4. Linear regression: under-determined case

In the case $n<p$, $\hat{Y}=\hat{w}X$ typically will have infinitely many solutions and we will choose the one with the smallest norm. This is actually given by the same formula as before if we replace the inverse by the pseudo-inverse

$$\hat{w}=\hat{Y}X^T(XX^T)^+=\hat{Y}(X^TX)^{-1}X^T,$$

where the last equation holds only if $\operatorname{rank}(X)=n<p$. This $\hat{w}$ then is a solution of $\hat{Y}=\hat{w}X$ and has smallest norm among all (infinitely many) solutions. Let us check this general linear algebra fact in the following lemma.

Lemma 6.1. Let $A\in\mathbb{R}^{n\times p}$ with $n<p$ and consider the system of linear equations

$$Ax=y$$

for given $y\in\mathbb{R}^{n\times 1}$. Assume that A has full rank, $\operatorname{rank}(A)=n$, i.e. $AA^T\in\mathbb{R}^{n\times n}$ is invertible. Then

$$x_0:=A^T(AA^T)^{-1}y$$

is a solution of $Ax=y$ and it has smallest norm among all solutions of $Ax=y$.

Proof. (i) We have

$$Ax_0=AA^T(AA^T)^{-1}y=y.$$

(ii) Assume that $Ax=y$, then (since $A(x-x_0)=Ax-Ax_0=y-y=0$)

$$\langle x-x_0,x_0\rangle=\left\langle x-x_0,A^T(AA^T)^{-1}y\right\rangle=\left\langle A(x-x_0),(AA^T)^{-1}y\right\rangle=0,$$

i.e. $x-x_0\perp x_0$. Thus

$$\begin{aligned}\|x\|^2=\langle x,x\rangle&=\langle x-x_0+x_0,x-x_0+x_0\rangle\\&=\langle x-x_0,x-x_0\rangle+\langle x_0,x_0\rangle+2\langle x-x_0,x_0\rangle\\&=\underbrace{\|x-x_0\|^2}_{\geq 0}+\|x_0\|^2\\&\geq\|x_0\|^2\end{aligned}$$

with equality if and only if $x=x_0$. □

Now back to $\hat{Y} = \hat{w}X$, i.e. $X^T\hat{w}^T = \hat{Y}^T$. Note that if we let $A = X^T$ and $x = \hat{w}^T$ in Lemma 6.1, then our minimal solution is

$$\hat{w} = \hat{Y}(X^TX)^{-1}X^T.$$

This $\hat{w}$ exactly matches the given noisy data $\hat{Y} = wX + N$. Let us again consider the error

$$\|w - \hat{w}\|^2 = \|w - \hat{Y}(X^TX)^{-1}X^T\|^2 = \|w - wX(X^TX)^{-1}X^T - N(X^TX)^{-1}X^T\|^2.$$

Note that

$$\begin{aligned}[w - wX(X^TX)^{-1}X^T](N(X^TX)^{-1}X^T)^T &= w[I_p - X(X^TX)^{-1}X^T]X(X^TX)^{-1}N^T \\ &= w[X - X(X^TX)^{-1}X^TX](X^TX)^{-1}N^T \\ &= w[X - X](X^TX)^{-1}N^T = 0,\end{aligned}$$

i.e. $w - wX(X^TX)^{-1}X^T) \perp N(X^TX)^{-1}X^T$. Thus

$$\begin{aligned}\|w - \hat{w}\|^2 &= \|w - wX(X^TX)^{-1}X^T - N(X^TX)^{-1}X^T\|^2 \\ &= \underbrace{\|w - wX(X^TX)^{-1}X^T\|^2}_{\text{“bias term”}} + \underbrace{\|N(X^TX)^{-1}X^T\|^2}_{\text{“variance term”}}.\end{aligned}$$

Consider first the variance term:

$$\begin{aligned}E_N\left[\|N(X^TX)^{-1}X^T\|^2\right] &= E_N\left[N(X^TX)^{-1}X^TX(X^TX)^{-1}N^T\right] \\ &= \sigma^2 \operatorname{Tr}\left((X^TX)^{-1}X^TX(X^TX)^{-1}\right) \\ &= \sigma^2 \operatorname{Tr}\left((X^TX)^{-1}\right) \\ &= \sigma^2\frac{n}{p}\operatorname{tr}\left(\left(\frac{1}{p}X^TX\right)^{-1}\right).\end{aligned}$$

This is the same as in the over-determined case, but the roles of p and n are exchanged, thus γ is replaced by $\frac{1}{\gamma}$ and the variance term converges for $p = \gamma n \to \infty$ to

$$\sigma^2\frac{\frac{1}{\gamma}}{1 - \frac{1}{\gamma}} = \sigma^2\frac{1}{\gamma - 1}.$$

Now consider the bias term:

$$\begin{aligned}\|w - wX(X^TX)^{-1}X^T\|^2 &= (w - wX(X^TX)^{-1}X^T)(w - wX(X^TX)^{-1}X^T)^T \\ &= (w - wX(X^TX)^{-1}X^T)(w^T - X(X^TX)^{-1}X^Tw^T) \\ &= ww^T - wX(X^TX)^{-1}X^Tw^T \\ &= \|w\|^2\left(1 - \frac{w}{\|w\|}X(X^TX)^{-1}X^T\frac{w^T}{\|w\|}\right).\end{aligned}$$

Note that

$$\frac{w}{\|w\|}X(X^TX)^{-1}X^T\frac{w^T}{\|w\|} = vX(X^TX)^{-1}X^Tv^T$$

for all $\|v\| = 1$, because of rotational symmetry. Thus

$$\begin{aligned}\frac{w}{\|w\|}X(X^TX)^{-1}X^T\frac{w^T}{\|w\|} &= \frac{1}{p}\operatorname{Tr}(X(X^TX)^{-1}X^T)\\ &= \frac{1}{p}\operatorname{Tr}(X^TX(X^TX)^{-1})\\ &= \frac{1}{p}\operatorname{Tr}(I_n)\\ &= \frac{n}{p}.\end{aligned}$$

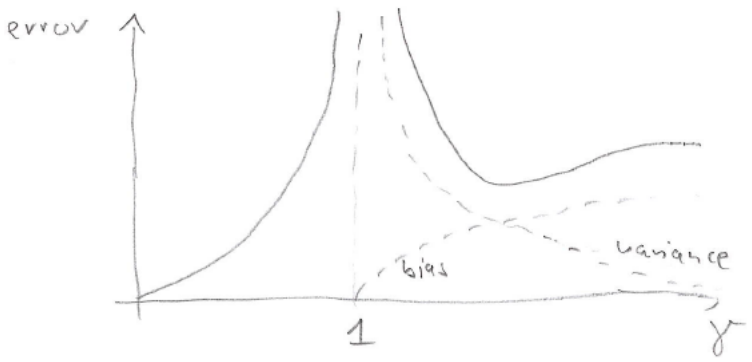
Continuing, we have

$$\begin{aligned}\|w - wX(X^TX)^{-1}X^T\|^2 &= \|w\|^2\left(1 - \frac{w}{\|w\|}X(X^TX)^{-1}X^T\frac{w^T}{\|w\|}\right)\\ &= \|w\|^2\left(1 - \frac{n}{p}\right)\\ &\xrightarrow[p=\gamma n]{n\to\infty} \|w\|^2\left(1 - \frac{1}{\gamma}\right).\end{aligned}$$

6.5. Double descent for linear regression

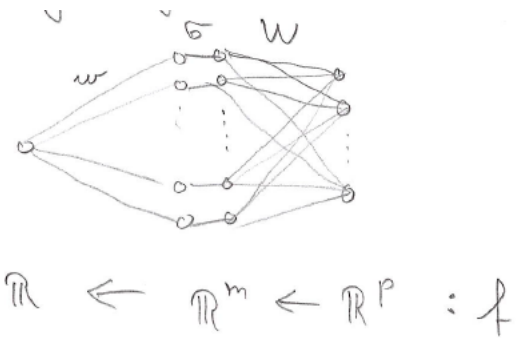
Combining all the results, we get

$$E\left[\|w - \hat{w}\|^2\right] = \begin{cases}\sigma^2\dfrac{\gamma}{1-\gamma}, & \gamma < 1,\\ \sigma^2\dfrac{1}{\gamma-1} + \|w\|^2\left(1 - \dfrac{1}{\gamma}\right), & \gamma > 1.\end{cases}$$



6.6. Adding layers and non-linearities

Consider now a more interesting neural network by adding one layer and non-linearities.



So we have now $f(x) = w\sigma Wx$ and $f(X) = w\sigma WX$. If we don't learn W, but only w, then this is still linear regression, but not on X, but on the "random features" $F := \sigma WX$ and the performance depends on the eigenvalue distribution of FF^T.

So we should now address the question whether we can understand the (combined) effect of

- multiplying two random matrices and
- applying non-linear functions entry-wise to random matrices

on the eigenvalue distribution? The first problem is part of classical random matrix theory, the second problem, in combination with the first one, is new and gives rise to "non-linear random matrix theory". Let us be a bit more precise on this.

(i) Consider $F = WX$, where both W and X are Gaussian random matrices (thus WW^T and XX^T are Wishart matrices), independent from each other. Then $FF^T = WXX^TW^T$ has the same distribution as a Wishart matrix YY^T, where Y is the data matrix of vectors $y = Wx$ where $x \sim N(0, I_p)$, so $y \sim N(0, WW^T)$. Put $\Sigma := WW^T \in \mathbb{R}^{m\times m}$. Then one can determine a fixed-point equation for the Stieltjes transform of FF^T; similar (but more general) as in Exercise 4 of Assignment 4.

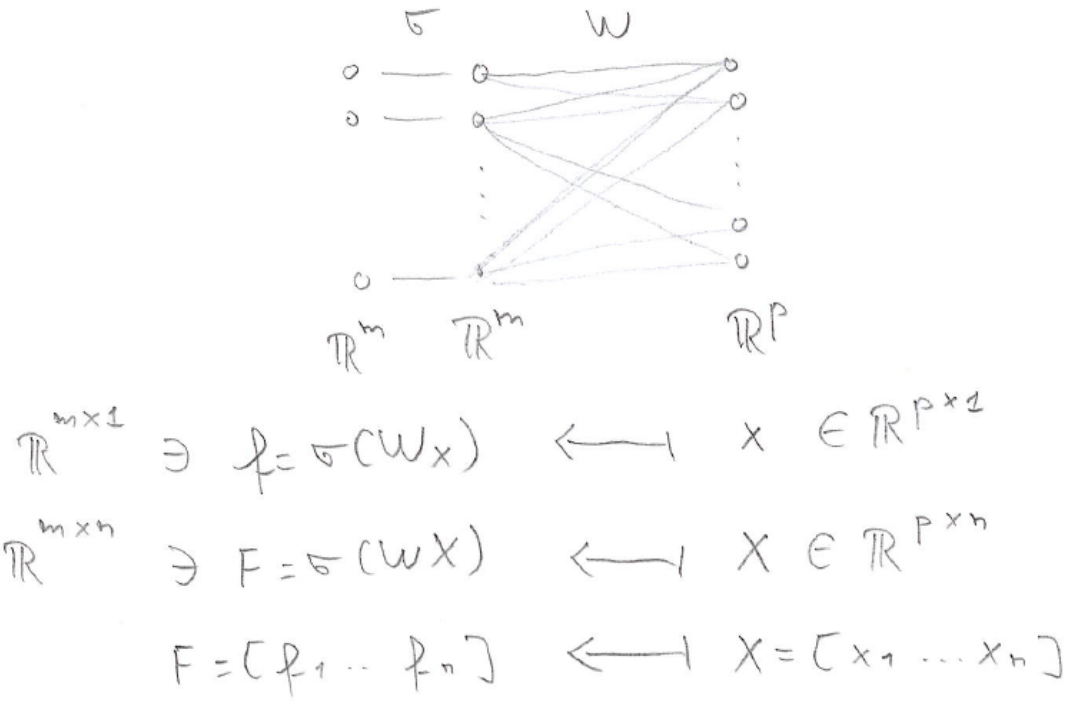
(ii) Consider $F = \sigma X$. Then the entries of F are still i.i.d., but their common distribution is not Gaussian any more, but the push-forward of Gauss under σ. However, for the validity of the Marchenko-Pastur law Theorem 4.2 one does not need the Gaussian distribution, the essential input is independence. Thus the distribution of F is still Marchenko-Pastur.

Both (i) and (ii) are thus within the realm of classical random matrix theory, but if we consider now the combination (iii) $F = \sigma WX$, then we have to move into new "non-linear" random matrix territory! We will address this in the next section.

7. Non-Linear Random Matrix Models: Resolvent Method and Cumulant Expansions

7.1. Distribution of the random features model

We consider our "random features" model



and we want to understand the distribution of the features f – in particular the eigenvalues of their covariance estimator FF^T – in the asymptotic regime where all sizes go, in a proportional way, to ∞: $p, n, m \to \infty$ such that $\frac{m}{n} \to \gamma$ and $\frac{p}{m} \to \tilde{\gamma}$.

We will try to understand the statement as well as the idea and the tools of the proof of the following theorem. For the proof we will follow the ideas from [PS21].

Theorem 7.1 (Pennington and Worah [PW17], Benigni and Peché [BP21])**.** Let $X \in \mathbb{R}^{p\times n}$ and $W \in \mathbb{R}^{m\times p}$ be standard Gaussian random matrices and $\sigma : \mathbb{R} \to \mathbb{R}$ be a sufficiently nice function (in particular, all derivatives have to exist), which is centered with respect to the Gaussian distribution, i.e.

$$\int_{\mathbb{R}} \sigma(t) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t = 0.$$

We put

$$F := \sigma\left(\frac{1}{\sqrt{p}} WX\right) \in \mathbb{R}^{m\times n} \quad \text{and} \quad M := \frac{1}{n} FF^T \in \mathbb{R}^{m\times m}.$$

Then, in the limit $p, n, m \to \infty$ such that $\frac{m}{n} \to \gamma$ and $\frac{p}{m} \to \tilde{\gamma}$, the Stieltjes transform of M converges to a limit

$$S(z) = \lim_{\substack{\frac{m}{n} \to \gamma \\ \frac{p}{m} \to \tilde{\gamma}}} E\left[\operatorname{tr}\left(\left(\frac{1}{n} F F^T - z I_m\right)^{-1}\right)\right]$$

and this $S(z)$ satisfies the following quartic equation:

$$1 + zS(z) = \theta_1 S(z)\Big(1 - \gamma\big(1 + zS(z)\big)\Big) - \frac{\theta_2}{\tilde{\gamma}}\big(1 + zS(z)\big)\Big(1 - \gamma\big(1 + zS(z)\big)\Big) + \frac{\theta_2(\theta_1 - \theta_2)}{\tilde{\gamma}} S(z)^2 \Big(1 - \gamma\big(1 + zS(z)\big)\Big)^2,$$

where

$$\theta_1 = \theta_1(\sigma) := \int_{\mathbb{R}} \sigma(t)^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$$

and

$$\theta_2 = \theta_2(\sigma) := \left(\int_{\mathbb{R}} \sigma'(t) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t\right)^2.$$

Remark. (1) If $\theta_2 = 0$ and $\theta_1 = 1$, this reduces to

$$1 + zS(z) = S(z)\Big(1 - \gamma\big(1 + zS(z)\big)\Big),$$

which is the equation for Marchenko-Pastur, compare our proof of Theorem 4.2. Thus in this case $F \mathrel{\hat{=}} Z$, where $Z \in \mathbb{R}^{m \times n}$ is a standard Gaussian random matrix.

(2) If $\theta_1 = \theta_2$ and $\theta_2 = 1$, this reduces to the cubic equation

$$1 + zS(z) = S(z)\Big(1 - \gamma\big(1 + zS(z)\big)\Big)\Big(1 - \frac{1}{\tilde{\gamma}}\big(1 + zS(z)\big)\Big).$$

Note that $\theta_1 = \theta_2$ is given for $\sigma(t) = \sqrt{\theta_2} \cdot t$,[6] thus this special non-linear case corresponds to the linear situation $F \mathrel{\hat{=}} \frac{\sqrt{\theta_2}}{\sqrt{p}} WX$.

(3) It is not obvious from the general form of the equation, but one can show (and we will come back to this in Section 7.6) that in general one actually has the "independent" combination of those two special cases, i.e.

$$F \mathrel{\hat{=}} \frac{\sqrt{\theta_2}}{\sqrt{p}} WX + \sqrt{\theta_1 - \theta_2}\, Z.$$

Thus the effect of the non-linearity is to produce some additional noise.

[6] Actually, $\theta_1 = \theta_2$ can only happen for linear σ, see Assignment 5.

7.2. Proof of Marchenko-Pastur law via Stein's identity

It is not clear how to generalize our proof of Marchenko-Pastur to the situation at hand. Thus, in the following we will first give another proof of the Marchenko-Pastur law, which has the potential for generalizations. This proof relies on Stein's identity, and we will see later that we can extend this to a cumulant expansion, which allows us then to also deal with products and non-linearities.

Second proof of the Marchenko-Pastur law, part one. Let $X = \big(x_{ij}\big) \in \mathbb{R}^{p\times n}$ be our standard Gaussian random matrix, then we want to calculate, in the limit $p, n \to \infty$ with $\frac{p}{n} \to \gamma$, the Stieltjes transform of $A = \frac{1}{n}XX^T$, given by

$$S(z) = E\left[\mathrm{tr}\left((A - zI_p)^{-1}\right)\right] = E\left[\mathrm{tr}\left(\left(\frac{1}{n}XX^T - zI_p\right)^{-1}\right)\right].$$

We put

$$R(z) := (A - zI_p)^{-1}, \qquad \text{thus} \qquad S(z) = E\Big[\mathrm{tr}\big(R(z)\big)\Big].$$

We have $(A - zI_p)R(z) = I_p$, i.e. $I_p + zR(z) = AR(z)$. Applying $E[\mathrm{tr}(\cdot)]$ to both sides, we get

$$1 + zS(z) = E\Big[\mathrm{tr}\big(AR(z)\big)\Big] = \frac{1}{np}E\Big[\mathrm{Tr}\big(XX^TR(z)\big)\Big] = \frac{1}{np}\sum_{\substack{i=1,\ldots,p\\ j=1,\ldots,n}} E\Big[x_{ij}\big[X^TR(z)\big]_{ji}\Big],$$

where $X^TR(z)$ is a function of all x_{kl}. Thus we need a formula to deal with such expectations. We will first present this "Stein's identity" and then later continue with our proof. □

Lemma 7.2 (Stein's Identity)**.** Let $t_1, \ldots, t_k$ be independent standard Gaussian random variables and $h : \mathbb{R}^k \to \mathbb{R}$ a nice function (like: continuously differentiable such that all partial derivatives are of polynomial growth). Then we have for $i = 1, \ldots, k$:

$$E\big[t_ih(t_1, \ldots, t_k)\big] = E\big[\partial_ih(t_1, \ldots, t_k)\big],$$

where $\partial_i = \frac{\partial}{\partial t_i}$ is the partial derivative with respect to the i-th variable t_i.

Proof. The main argument happens for $k = 1$; it is just partial integration:

$$\begin{aligned}
E\big[th(t)\big] &= \frac{1}{\sqrt{2\pi}} \int th(t) \exp\left(-\frac{t^2}{2}\right) \,\mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi}} \int h(t) \cdot \underbrace{t \exp\left(-\frac{t^2}{2}\right)}_{=\left(-\exp\left(-\frac{t^2}{2}\right)\right)'} \,\mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi}} \int h'(t) \exp\left(-\frac{t^2}{2}\right) \,\mathrm{d}t - \frac{1}{\sqrt{2\pi}} \left[h(t) \exp\left(-\frac{t^2}{2}\right)\right]_{-\infty}^{+\infty} \\
&= \frac{1}{\sqrt{2\pi}} \int h'(t) \exp\left(-\frac{t^2}{2}\right) \,\mathrm{d}t \\
&= E[h'(t)]
\end{aligned}$$

since the last summand in the partial integration is zero by the assumption on h. For general k, just do partial integration for the i-th coordinate. □

Second proof of the Marchenko-Pastur law, part two. In our setting, this now gives

$$\begin{aligned}
E\Big[x_{ij}\big[X^T R(z)\big]_{ji}\Big] &= E\Big[\partial_{ij}\big[X^T R(z)\big]_{ji}\Big] \\
&= E\left[\partial_{ij} \sum_{k=1}^{p} \big[X^T\big]_{jk} \big[R(z)\big]_{ki}\right] \\
&= E\left[\partial_{ij} \sum_{k=1}^{p} x_{kj} \big[R(z)\big]_{ki}\right] \\
&= E\left[\sum_{k=1}^{p} \frac{\partial x_{kj}}{\partial x_{ij}} \cdot \big[R(z)\big]_{ki} + x_{kj} \frac{\partial \big[R(z)\big]_{ki}}{\partial x_{ij}}\right] \\
&= E\left[\sum_{k=1}^{p} \delta_{ik} \big[R(z)\big]_{ki} + x_{kj} \frac{\partial \big[R(z)\big]_{ki}}{\partial x_{ij}}\right] \\
&= E\Big[\big[R(z)\big]_{ii}\Big] + \sum_{k=1}^{p} E\left[x_{kj} \frac{\partial \big[R(z)\big]_{ki}}{\partial x_{ij}}\right]
\end{aligned}$$

From Assigment 5 we know

$$\frac{\partial \big[R(z)\big]_{kl}}{\partial x_{ij}} = -\frac{1}{n}\Big(\big[R(z)\big]_{ki} \cdot \big[X^T R(z)\big]_{jl} + \big[R(z)X\big]_{kj} \big[R(z)\big]_{il}\Big),$$

thus, by setting $l = i$, we get

$$E\left[x_{kj} \frac{\partial \big[R(z)\big]_{ki}}{\partial x_{ij}}\right] = -\frac{1}{n} E\Big[x_{kj}\big[R(z)\big]_{ki}\big[X^T R(z)\big]_{ji} + x_{kj}\big[R(z)X\big]_{kj}\big[R(z)\big]_{ii}\Big]$$

and therefore

$$
\begin{aligned}
1 + zS(z) &= \frac{1}{np} \sum_{\substack{i=1,\dots,p \\ j=1,\dots,n}} E\Big[x_{ij}\big[X^T R(z)\big]_{ji}\Big] \\
&= \frac{1}{np} \sum_{\substack{i=1,\dots,p \\ j=1,\dots,n}} \left(E\Big[\big[R(z)\big]_{ii}\Big] + \sum_{k=1}^{p} E\left[x_{kj} \frac{\partial\big[R(z)\big]_{ki}}{\partial x_{ij}}\right]\right) \\
&= \frac{1}{np} \sum_{\substack{i=1,\dots,p \\ j=1,\dots,n}} E\Big[\big[R(z)\big]_{ii}\Big] - \frac{1}{n^2 p} \sum_{\substack{i=1,\dots,p \\ j=1,\dots,n \\ k=1,\dots,p}} E\left[x_{kj}\big[R(z)\big]_{ki}\big[X^T R(z)\big]_{ji}\right] \\
&\quad - \frac{1}{n^2 p} \sum_{\substack{i=1,\dots,p \\ j=1,\dots,n \\ k=1,\dots,p}} E\left[x_{kj}\big[R(z)X\big]_{kj}\big[R(z)\big]_{ii}\right].
\end{aligned}
$$

Consider the summands separately:

(1) For the first summand, we have

$$
\begin{aligned}
\frac{1}{np} \sum_{\substack{i=1,\dots,p \\ j=1,\dots,n}} E\Big[\big[R(z)\big]_{ii}\Big] = \frac{1}{np} \sum_{j=1}^{n} E\left[\sum_{i=1}^{p} \big[R(z)\big]_{ii}\right] &= \frac{1}{np} \cdot n \cdot E\Big[\operatorname{Tr}\big(R(z)\big)\Big] \\
&= \frac{1}{p} E\Big[p \operatorname{tr}\big(R(z)\big)\Big] \\
&= E\Big[\operatorname{tr}\big(R(z)\big)\Big] = S(z).
\end{aligned}
$$

(2) For the second summand, we have (by the cyclic property of the trace)

$$
\begin{aligned}
&\frac{1}{n^2 p} \sum_{\substack{i=1,\dots,p \\ j=1,\dots,n \\ k=1,\dots,p}} E\left[x_{kj}\big[R(z)\big]_{ki}\big[X^T R(z)\big]_{ji}\right] \\
&\qquad = \frac{1}{n^2 p} \sum_{\substack{i=1,\dots,p \\ j=1,\dots,n \\ k=1,\dots,p}} E\left[\big[X^T\big]_{jk}\big[R(z)\big]_{ki}\big[R(z)^T X\big]_{ij}\right] \\
&\qquad = \frac{1}{n^2 p} \sum_{j=1}^{n} E\left[\big[X^T R(z) R(z)^T X\big]_{jj}\right] \\
&\qquad = \frac{1}{n^2 p} E\Big[\operatorname{Tr}\big(X^T R(z) R(z)^T X\big)\Big] \\
&\qquad = \frac{1}{n^2 p} E\Big[\operatorname{Tr}\big(X X^T R(z) R(z)^T\big)\Big] \\
&\qquad = \frac{1}{n} E\left[\operatorname{tr}\left(\frac{X X^T}{n} R(z) R(z)^T\right)\right] \xrightarrow{n\to\infty} 0.
\end{aligned}
$$

(3) For the third summand, we have (again by the cyclic property of the trace)

$$\begin{aligned}
&\frac{1}{n^2 p} \sum_{\substack{i=1,\dots,p \\ j=1,\dots,n \\ k=1,\dots,p}} E\Big[x_{kj}\big[R(z)X\big]_{kj}\big[R(z)\big]_{ii}\Big] \\
&\qquad = \frac{1}{n^2 p} E\left[\sum_{k=1}^{p}\sum_{j=1}^{n} x_{kj}\big[R(z)X\big]_{kj}\sum_{i=1}^{p}\big[R(z)\big]_{ii}\right] \\
&\qquad = \frac{1}{n^2 p} E\left[\sum_{k=1}^{p}\sum_{j=1}^{n} x_{kj}\big[R(z)X\big]_{kj}\operatorname{Tr}\big(R(z)\big)\right] \\
&\qquad = E\left[\operatorname{tr}\big(R(z)\big)\frac{1}{n^2}\sum_{k=1}^{p}\sum_{j=1}^{n}\big[X^T\big]_{jk}\big[R(z)X\big]_{kj}\right] \\
&\qquad = E\left[\operatorname{tr}\big(R(z)\big)\frac{1}{n^2}\sum_{j=1}^{n}\big[X^T R(z)X\big]_{jj}\right] \\
&\qquad = E\left[\operatorname{tr}\big(R(z)\big)\frac{1}{n^2}\operatorname{Tr}\big(X^T R(z)X\big)\right] \\
&\qquad = E\left[\operatorname{tr}\big(R(z)\big)\frac{1}{n^2}\operatorname{Tr}\big(XX^T R(z)\big)\right] \\
&\qquad = E\left[\operatorname{tr}\big(R(z)\big)\frac{p}{n}\operatorname{tr}\left(\frac{XX^T}{n}R(z)\right)\right] \\
&\qquad \approx E\Big[\operatorname{tr}\big(R(z)\big)\Big]\cdot\frac{p}{n}E\left[\operatorname{tr}\left(\frac{XX^T}{n}R(z)\right)\right] \\
&\qquad = S(z)\cdot\gamma\cdot E\Big[\operatorname{tr}\big(AR(z)\big)\Big],
\end{aligned}$$

Note that we need concentration to asymptotically factorize the expectation of a product! Remember that we have $AR(z) = I_p + zR(z)$, so

$$E\Big[\operatorname{tr}\big(AR(z)\big)\Big] = E\Big[\operatorname{tr}\big(I_p + zR(z)\big)\Big] = 1 + zS(z),$$

so

$$\frac{1}{n^2 p} \sum_{\substack{i=1,\dots,p \\ j=1,\dots,n \\ k=1,\dots,p}} E\Big[x_{kj}\big[R(z)X\big]_{kj}\big[R(z)\big]_{ii}\Big] \approx S(z)\cdot\gamma\cdot\big(1 + zS(z)\big).$$

Putting everything together, we get in the limit

$$1 + zS(z) = S(z) - \gamma S(z)\cdot\big(1 + zS(z)\big),$$

i.e.

$$\gamma z S(z)^2 + (z + \gamma - 1)S(z) + 1 = 0.$$

This is the same equation as we derived in the proof of the Marchenko-Pastur law (cf. Theorem 4.2). □

7.3. Extension of Stein's identity to cumulant expansion

Now we want to extend this approach from X to $F = \sigma\left(\frac{1}{\sqrt{p}}WX\right) \in \mathbb{R}^{m\times n}$. So we start as before, with

$$S(z) = E\left[\operatorname{tr}\left(\left(\frac{1}{n}FF^T - zI_m\right)^{-1}\right)\right] = E\Big[\operatorname{tr}\big(R(z)\big)\Big]$$

and the equation

$$\begin{aligned} 1 + zS(z) = E\left[\operatorname{tr}\left(\frac{1}{n}FF^T R(z)\right)\right] &= \frac{1}{nm}E\Big[\operatorname{Tr}\big(FF^T R(z)\big)\Big] \\ &= \frac{1}{nm}\sum_{\substack{i=1,\dots,m\\ j=1,\dots,n}} E\Big[f_{ij}\big[F^T R(z)\big]_{ji}\Big], \end{aligned}$$

where $F = \big(f_{ij}\big)$ and $\big[F^T R(z)\big]_{ji}$ is a function of all f_{kl}.

The problem is now that the f_{ij} are neither Gaussian nor (and this is more serious) independent any more!

So we have to face the question: Do we still have a version of Stein's identity for such a general case?

Recall the one-dimensional case of Stein's identity: If t is a standard Gaussian random variable, then

$$E\big[th(t)\big] = E\big[h'(t)\big].$$

For general distributions, one can try to

- keep the RHS and change the LHS; this leads to the theory of score functions, which is an important subject, but not really relevant here;
- or keep the LHS and change the RHS; this leads to cumulant expansions and is what we need!

In order to get an idea what $E\big[th(t)\big]$ could be in general, we will consider it for special functions of the form $h_s(t) = \exp(\mathrm{i}ts)$ and then get the general case by Fourier decomposition. Defining constants κ_l via

$$\log\Big(E\big[\exp(\mathrm{i}ts)\big]\Big) =: \sum_{l=1}^{\infty} \frac{\kappa_\ell}{\ell!}(\mathrm{i}s)^\ell,$$

we have

$$\begin{aligned}
E\big[th_s(t)\big] = E\big[t\exp(\mathrm{i}ts)\big] &= E\left[-\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}s}\exp(\mathrm{i}ts)\right] \\
&= -\mathrm{i}\left(\frac{\mathrm{d}}{\mathrm{d}s}\log\Big(E\big[\exp(\mathrm{i}ts)\big]\Big)\right)\cdot E\big[\exp(\mathrm{i}ts)\big] \\
&= -\mathrm{i}\sum_{\ell=1}^{\infty}\frac{\kappa_\ell}{(\ell-1)!}\mathrm{i}(\mathrm{i}s)^{\ell-1}\cdot E\big[\exp(\mathrm{i}ts)\big] \\
&= E\left[\sum_{\ell=0}^{\infty}\frac{\kappa_{\ell+1}}{\ell!}(\mathrm{i}s)^\ell\cdot\exp(\mathrm{i}ts)\right] \\
&= E\left[\sum_{\ell=0}^{\infty}\frac{\kappa_{\ell+1}}{\ell!}\frac{\mathrm{d}^\ell\exp(\mathrm{i}ts)}{\mathrm{d}t^\ell}\right] \\
&= E\left[\sum_{\ell=0}^{\infty}\frac{\kappa_{\ell+1}}{\ell!}h_s^{(\ell)}(t)\right].
\end{aligned}$$

By Fourier decomposition, this then goes over to "arbitrary functions":

Lemma 7.3. Let t be a random variable such that all of its moments exist. Then we define its "cumulants" κ_ℓ by

$$\log\Big(E\big[\exp(\mathrm{i}ts)\big]\Big) = \sum_{\ell=1}^{\infty}\frac{\kappa_\ell}{\ell!}(\mathrm{i}s)^\ell,$$

and for a smooth function $h:\mathbb{R}\to\mathbb{R}$ we then have

$$E\big[th(t)\big] = \sum_{\ell=0}^{\infty}\frac{\kappa_{\ell+1}}{\ell!}E\big[h^{(l)}(t)\big].$$

Remark. Note that this is consistent with Stein's identity. If t is a standard Gaussian random variable, then we have

$$E\big[\exp(\mathrm{i}ts)\big] = \exp\left(-\frac{s^2}{2}\right),$$

thus

$$\log\Big(E\big[\exp(\mathrm{i}ts)\big]\Big) = -\frac{s^2}{2},$$

which means that all κ_ℓ are zero except $\kappa_2 = 1$; but then Lemma 7.3 reduces to

$$E\big[th(t)\big] = \frac{\kappa_2}{1!}E\big[h'(t)\big] = E\big[h'(t)\big].$$

We now need the multivariate version of all this. This works in the same way.

Definition 7.4. Let $t_1, \ldots, t_k$ be a collection of random variables. Their characteristic function (i.e. the Fourier transform of their density function) is

$$E\Big[\exp\big(\mathrm{i}(t_1s_1 + \ldots + t_ks_k)\big)\Big] = \int_{\mathbb{R}^k} \exp\big(\mathrm{i}(t_1s_1 + \ldots + t_ks_k)\big)\psi(t_1, \ldots, t_k)\,\mathrm{d}t_1 \ldots \mathrm{d}t_k$$

and the *cumulants* of $t_1, \ldots, t_k$ are defined as coefficients in the power series expansion of the logarithm of the characteristic function:

$$\log\Big(E\Big[\exp\big(\mathrm{i}(t_1s_1 + \ldots + t_ks_k)\big)\Big]\Big) = \sum_{\ell=0}^{\infty} \frac{\kappa_\ell}{\ell!}(\mathrm{i}s)^\ell,$$

where $\ell = (\ell_1, \ldots, \ell_k)$ is a multi-index, and we use the usual multi-index conventions, like $\ell! = \ell_1! \cdots \ell_k!$.

The κ are actually multi-linear mappings in the random variables, i.e. the coefficient of $s_{i_1} \cdots s_{i_m}$ is $\kappa(t_{i_1}, \ldots, t_{i_m})$. With this definition one also has the following multi-dimensional version of Lemma 7.3:

Proposition 7.5. Let $t_1, \ldots, t_k$ be a collection of random variables and κ their cumulants. For smooth functions $h : \mathbb{R}^k \to \mathbb{R}$ we then have

$$E\big[t_ih(t_1, \ldots, t_k)\big] = \sum_{\ell \geq 0} \sum_{i_1, \ldots, i_\ell = 1}^{k} \frac{\kappa(t_i, t_{i_1}, \ldots, t_{i_\ell})}{\ell!} E\big[\partial_{i_1} \ldots \partial_{i_\ell} h(t_1, \ldots, t_k)\big].$$

Remark. Note that the coefficients of the characteristic function power series expansion are essentially the moments of our random variables; the coefficients in the logarithm of the characteristic function are by definition the cumulants. This means that moments and cumulants are functions of each other. The combinatorial nature of this relation is revealed by Proposition 7.5, if we use monomials for h.

Example. (1) For $h = 1$, only $\ell = 0$ contributes, so

$$E[t_i] = E[t_i \cdot 1] = \frac{\kappa(t_i)}{0!} \underbrace{E[1]}_{=1} = \kappa(t_i).$$

(2) For $h(t_1, \ldots, t_k) = t_j$ we have

$$\begin{aligned} E[t_i t_j] &= \underbrace{\kappa(t_i) \cdot E[t_j]}_{\text{from } \ell=0} + \underbrace{\sum_{i_1=1}^{k} \kappa(t_i, t_{i_1}) E[\partial_{i_1} t_j]}_{\text{from } \ell=1} \\ &= \kappa(t_i) \cdot \kappa(t_j) + \sum_{i_1=1}^{k} \kappa(t_i, t_{i_1}) \delta_{i_1 j} \\ &= \kappa(t_i) \cdot \kappa(t_j) + \kappa(t_i, t_j), \end{aligned}$$

thus

$$E[t_i t_j] = \underbrace{\kappa(t_i, t_j)}_{\sqcup} + \underbrace{\kappa(t_i)\kappa(t_j)}_{\mid\ \mid},$$

where the two summands correspond to the two partitions of $\{t_i, t_j\}$ drawn below them.

(3) For $h(t_1, \ldots, t_k) = t_j t_r$ we have

$$\begin{aligned} E[t_i t_j t_r] &= \kappa(t_i) E[t_j t_r] + \sum_{i_1=1}^{k} \kappa(t_i, t_{i_1}) E[\partial_{i_1} t_j t_r] + \sum_{i_1, i_2=1}^{k} \frac{1}{2} \kappa(t_i, t_{i_1}, t_{i_2}) E[\partial_{i_1} \partial_{i_2} t_j t_r] \\ &= \kappa(t_i) E[t_j t_r] + \kappa(t_i, t_j) E[t_r] + \kappa(t_i, t_r) E[t_j] + \frac{1}{2}\big(\kappa(t_i, t_j, t_r) + \kappa(t_i, t_r, t_j)\big) \\ &= \kappa(t_i) E[t_j t_r] + \kappa(t_i, t_j) E[t_r] + \kappa(t_i, t_r) E[t_j] + \kappa(t_i, t_j, t_r) \\ &= \kappa(t_i, t_j, t_r) + \kappa(t_i)\kappa(t_j, t_r) + \kappa(t_i, t_j)\kappa(t_r) + \kappa(t_i, t_r)\kappa(t_j) + \kappa(t_i)\kappa(t_j)\kappa(t_r) \end{aligned}$$

(corresponding respectively to the partitions $\sqcup\!\sqcup$, $\mid \sqcup$, $\sqcup \mid$, $\sqcup\!\mid\!\sqcup$, and $\mid\mid\mid$ of $\{t_i, t_j, t_r\}$), since $E[\partial_{i_1} t_j t_r] = \delta_{i_1 j} t_r + \delta_{i_1 r} t_j$ and $E[\partial_{i_1} \partial_{i_2} t_j t_r] = \delta_{i_1 j}\delta_{i_2 r} + \delta_{i_1 r}\delta_{i_2 j}$.

This combinatorial relation is true in general.

Theorem 7.6. Moments and cumulants are related by the moment-cumulant formula

$$E[t_{i_1} \cdot t_{i_2} \cdot \ldots \cdot t_{i_n}] = \sum_{\pi \in \mathcal{P}(n)} \kappa_\pi(t_{i_1}, t_{i_2}, \ldots, t_{i_n}),$$

where $\pi = \{V_1, \ldots, V_r\}$ is a partition of the set $\{1, \ldots, n\}$ and

$$\kappa_\pi = \kappa_{V_1}(\ldots)\kappa_{V_2}(\ldots) \cdot \ldots \cdot \kappa_{V_r}(\ldots),$$

where the arguments are distributed according to π.

In order to use Proposition 7.5 we need the cumulants of our random feature matrices. For this we first have to get a more systematic understanding of cumulants.

7.4. Cumulants and their properties and uses

Definition 7.7. (1) We call $\pi = \{V_1, \ldots, V_r\}$ a *partition* of the set S if

- $V_i \neq \emptyset$ and $V_i \subset S$ for all i,
- $V_i \cap V_j = \emptyset$ for all $i \neq j$, and
- $V_1 \cup \ldots \cup V_r = S$.

We call $V_1, \ldots, V_r$ the *blocks* of π. Given two elements $p, q \in S$, we write $p \sim_\pi q$ if p and q belong to the same block of π.

(2) The set of all partitions of S is denoted by $\mathcal{P}(S)$. If $S = \{1, \ldots, n\}$, we write $\mathcal{P}(n) = \mathcal{P}(\{1, \ldots, n\})$. Note that $\mathcal{P}(n)$ has a "smallest" element

$$0_n := \{\{1\}, \ldots, \{n\}\} \in \mathcal{P}(n)$$

and a "largest" element

$$1_n := \{\{1, \ldots, n\}\} \in \mathcal{P}(n).$$

Often we use a graphical representation of a partition π like in Figure 4.

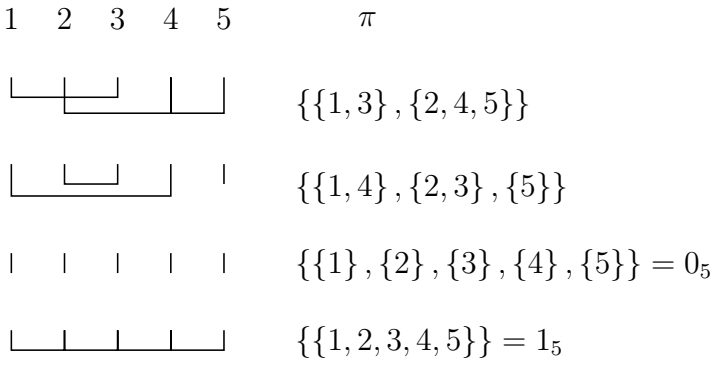


Figure 4: Examples for the graphical representation of partitions of five elements.

In the following we consider an algebra $\mathcal{A}$ of our random variables, for which the expectation $E : \mathcal{A} \to \mathbb{R}$ is defined. For example, if we have a collection of random variables $t_1, \ldots, t_r$, then

$$\mathcal{A} = \left\{ \sum_{n \geq 0} \sum_{i(1), \ldots, i(n) = 1}^{r} \alpha_{i(1), \ldots, i(n)} t_{i(1)} \cdot \ldots \cdot t_{i(n)} \mid \alpha_{i(1), \ldots, i(n)} \in \mathbb{C} \right\}$$

is the collection of all polynomials in the random variables. Given such $(\mathcal{A}, E)$ we know all moments $E(t)$ for all $t \in \mathcal{A}$. But for many questions it is better to rewrite the information about moments into other objects, so-called cumulants.

Notation 7.8. For $\pi = \{V_1, \ldots, V_r\} \in \mathcal{P}(n)$ and $t_1, \ldots, t_n \in \mathcal{A}$ we put

$$E_\pi(t_1, \ldots, t_n) = \prod_{V \in \pi} E(t_1, \ldots, t_n \mid V),$$

where for $V = \{i_1 < i_2 < \ldots < i_s\}$

$$E(t_1, \ldots, t_n \mid V) = E(t_{i_1} t_{i_2} \cdot \ldots \cdot t_{i_s}).$$

More general, if we have a collection $(\kappa_n)_{n \in \mathbb{N}}$ of n-linear functions

$$\kappa_n : \mathcal{A}^n \to \mathbb{R}, \quad (t_1, \ldots, t_n) \mapsto \kappa_n(t_1, \ldots, t_n),$$

then we define in the same way their multiplicative extension to all $\mathcal{P}(n)$ by

$$\kappa_\pi(t_1, \ldots, t_n) = \prod_{V \in \pi} \kappa_{\#V}(t_1, \ldots, t_n \mid V),$$

where for $V = \{i_1 < i_2 < \ldots < i_s\}$

$$\kappa(t_1, \ldots, t_n \mid V) = \kappa_s(t_{i_1}, t_{i_2}, \ldots, t_{i_s}).$$

Definition 7.9. Given $(\mathcal{A}, E)$ we define the *cumulants* of the random variables in $\mathcal{A}$ as $\kappa_n : \mathcal{A}^n \to \mathbb{R}$ by the moment-cumulant formulas: for all $n \in \mathbb{N}$ and all $t_1, \ldots, t_n \in \mathcal{A}$

$$E(t_1 \cdot \ldots \cdot t_n) = \sum_{\pi \in \mathcal{P}(n)} \kappa_\pi(t_1, \ldots, t_n). \tag{MCF}$$

These MCF define the κ_n's recursively as n-linear functionals.

Example 7.10. (1) $n = 1$: we have $E(t_1) = \kappa_1(t_1)$ and thus $\kappa_1(t_1) = E(t_1)$, corresponding to the partition $\sqcap$ of $\{t_1\}$.

(2) $n = 2$: we have

$$E(t_1 t_2) = \kappa_2(t_1, t_2) + \kappa_1(t_1)\kappa_1(t_2)$$

and thus

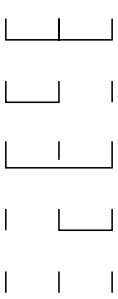
$$\kappa_2(t_1, t_2) = E(t_1 t_2) - \kappa_1(t_1)\kappa_1(t_2) = E(t_1 t_2) - E(t_1)E(t_2),$$

which is the covariance of (t_1, t_2).

(3) $n = 3$: we have

$$\begin{aligned} E(t_1t_2t_3) &= \kappa_3(t_1,t_2,t_3) \\ &\quad + \kappa_2(t_1,t_2)\kappa_1(t_3) \\ &\quad + \kappa_2(t_1,t_3)\kappa_1(t_2) \\ &\quad + \kappa_2(t_2,t_3)\kappa_1(t_1) \\ &\quad + \kappa_1(t_1)\kappa_1(t_2)\kappa_1(t_3), \end{aligned}$$

$t_1 \quad t_2 \quad t_3$



so

$$\begin{aligned} &\kappa_3(t_1,t_2,t_3) \\ &\quad = E(t_1t_2t_3) - \kappa_2(t_1,t_2)\kappa_1(t_3) - \dots \\ &\quad = E(t_1t_2t_3) - \big(E(t_1t_2) - E(t_1)E(t_2)\big)E(t_3) - \dots \\ &\quad = \underbrace{E(t_1t_2t_3)}_{\sqcup\!\sqcup\,\mid} - \underbrace{E(t_1t_2)E(t_3)}_{\sqcup\,\mid\,\mid} - \underbrace{E(t_1t_3)E(t_2)}_{\sqcup\!\sqcup} - \underbrace{E(t_2t_3)E(t_1)}_{\mid\,\sqcup} + 2\,\underbrace{E(t_1)E(t_2)E(t_3)}_{\mid\,\mid\,\mid} \\ &\quad = E_{\sqcup\!\sqcup}(t_1,t_2,t_3) - E_{\sqcup\,\mid}(\dots) - E_{\sqcup\!\sqcup}(\dots) - E_{\mid\,\sqcup}(\dots) + 2E_{\mid\,\mid\,\mid}(\dots). \end{aligned}$$

We see that we can write the cumulants, similar as in the MCF, via a summation over $\mathcal{P}(n)$, but now we get non-trivial coefficients. This rewriting of the MCF is a general version of the inclusion-exclusion principle, abstractly known as Möbius inversion.

Theorem 7.11. The recursive definition of the cumulants via the MCF is equivalent to the explicit formula

$$\kappa_n(t_1,\dots,t_n) = \sum_{\pi\in\mathcal{P}(n)} (-1)^{\#\pi-1}(\#\pi-1)!E_\pi(t_1,\dots,t_n). \tag{CMF}$$

The relevance of the cumulants is that they characterize independence.

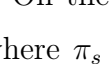
Theorem 7.12. Consider in $(\mathcal{A},E)$ subsets $T_i \subset \mathcal{A}$ $(i \in I)$ of random variables. Then the following are equivalent:

(i) the T_i are independent;

(ii) mixed cumulants in the T_i vanish: $\kappa_n(t_1,\dots,t_n) = 0$ whenever $t_j \in T_{i(j)}$ and there exist ℓ, k such that $i(\ell) = i(k)$.

"Proof". Let us only check (ii)⇒(i); namely that vanishing of mixed cumulants gives us factorization of moments. We do this via a telling example; consider $E[t_1s_1t_2s_2s_3t_3t_4]$, where mixed moments in $\{t_1, t_2, t_3, t_4\}$ and $\{s_1, s_2, s_3\}$ vanish. We have

$$E[t_1s_1t_2s_2s_3t_3t_4] = \sum_{\pi} \kappa_\pi,$$

but in the sum partitions like  are not included, since blocks are not allowed to connect a t_i with an s_j. On the other hand, partitions like  are included. So we have $\pi = \pi_s \cup \pi_t$, where π_s is a partition of $\{s_1, s_2, s_3\}$ and π_t is a partition of $\{t_1, t_2, t_3, t_4\}$. Continuing the computation, we get

$$\begin{aligned}
E[t_1s_1t_2s_2s_3t_3t_4] &= \sum_{\pi} \kappa_\pi(t_1, s_1, t_2, s_2, s_3, t_4, t_5) \\
&= \sum_{\pi_s \cup \pi_t} \kappa_{\pi_s \cup \pi_t}(t_1, s_1, t_2, s_2, s_3, t_4, t_5) \\
&= \sum_{\pi_s \cup \pi_t} \kappa_{\pi_s}(s_1, s_2, s_3)\kappa_{\pi_t}(t_1, t_2, t_3, t_4) \\
&= \left(\sum_{\pi_s} \kappa_{\pi_s}(s_1, s_2, s_3)\right)\left(\sum_{\pi_t} \kappa_{\pi_t}(t_1, t_2, t_3, t_4)\right) \\
&= E(s_1s_2s_3) \cdot E(t_1t_2t_3t_4). \qquad \square
\end{aligned}$$

Remark. (1) Note that, as for moments, cumulants do not change under permutation of arguments; e.g.

$$\kappa_3(t_1, t_2, t_3) = \kappa_3(t_1, t_3, t_2) = \kappa_3(t_2, t_1, t_3)$$

etc. since the terms in the CMF are mapped to each other under such permutations.

(2) Cumulants seem to be more complicated than moments. So, why do we want to use them? Here are some answers to this question.

- Expansions around special situations (like independent Gaussians) are easier to deal with.
- Almost factorization of moments is hard to work with, almost vanishing (i.e. smallness) of cumulants is much better for estimates.

(3) In order to be able to make really good use of cumulants, we also have to understand their multiplicative structure.

Note: the multiplicative structure for moments is easy, they are "associative", i.e.

$$E_2[t_1t_2, t_3] = E[(t_1t_2)t_3] = E[t_1(t_2t_3)] = E_2[t_1, t_2t_3].$$

This is not true for cumulants:

$$\kappa_2(t_1t_2, t_3) = E[t_1t_2t_3] - E[t_1t_2]E[t_3] \neq E[t_1t_2t_3] - E[t_1]E[t_2t_3] = \kappa_2(t_1, t_2t_3),$$

but there is a replacement here:

$$\kappa_2(t_1t_2, t_3) = E[t_1t_2t_3] - E[t_1t_2]E[t_3] = \kappa_3(t_1, t_2, t_3) + \kappa_1(t_1)\kappa_2(t_2, t_3) + \kappa_2(t_1, t_3)\kappa_1(t_2).$$

In particular, we see that from all possible partitions $\sqcup\!\sqcup\!\sqcup$, $\sqcup\,|$, $|\,\sqcup$, $\sqcup\!\sqcup$, and $|\,|\,|$, only the partitions $\sqcup\!\sqcup\!\sqcup$, $|\,\sqcup$, and $\sqcup\!\sqcup$ make a contribution. But these are exactly those partitions that connect the groups $\{t_1, t_2\}$ and $\{t_3\}$ of multiplied variables.

Theorem 7.13. Consider n random variables and multiply them together in m groups

$$\begin{aligned} T_1 &= t_1t_2 \cdot \ldots \cdot t_{i(1)}, \\ T_2 &= t_{i(1)+1} \cdot \ldots \cdot t_{i(2)}, \\ &\vdots \\ T_m &= t_{i(m-1)+1} \cdot \ldots \cdot t_{i(m)}, \end{aligned}$$

i.e.

$$\underbrace{t_1t_2 \cdot \ldots \cdot t_{i(1)}}_{T_1} \cdot \underbrace{t_{i(1)+1} \cdot \ldots \cdot t_{i(2)}}_{T_2} \cdot \ldots \cdot \underbrace{t_{i(m-1)+1} \cdot \ldots \cdot t_{i(m)}}_{T_m}.$$

Then we have

$$\kappa_m(T_1, \ldots, T_m) = \sum_{\substack{\pi \in \mathcal{P}(n) \\ \pi \text{ connects all the} \\ \text{groups together}}} \kappa_\pi(t_1, t_2, \ldots, t_n).$$

Example. (i) No products: if $T_i = t_i$ for all i, then

$$\kappa_n(t_1, \ldots, t_n) = \sum_{\substack{\pi \in \mathcal{P}(n) \\ \pi \text{ connects all the} \\ \text{groups together}}} \kappa_\pi(t_1, t_2, \ldots, t_n) = \kappa_n(t_1, \ldots, t_n),$$

since only 1_n connects everything.

(ii) One product: if $T = t_1 \cdot \ldots \cdot t_n$, then

$$\kappa_1(T) = \sum_{\substack{\pi \in \mathcal{P}(n) \\ \pi \text{ connects all the} \\ \text{groups together}}} \kappa_\pi(t_1, t_2, \ldots, t_n) = \sum_{\pi \in \mathcal{P}(n)} \kappa_\pi(t_1, \ldots, t_n) = E(t_1 \cdot \ldots \cdot t_n),$$

since all partitions "connect" the only block. This agrees with

$$\kappa_1(T) = E(T) = E(t_1 \cdot \ldots \cdot t_n).$$

(iii) We have

$$\kappa_3(t_1 t_2, t_3, t_4) = \kappa_4(t_1, t_2, t_3, t_4) + \kappa_3(t_1, t_3, t_4)\kappa_1(t_2) + \ldots$$

according to the connecting partitions  ,  ,  ,  , and  .

Note also that our definition of the cumulants via the MCF leads directly to our cumulant expansion from Proposition 7.5 if we choose h as a moment. Namely, take

$$h(t_1, \ldots, t_k) = t_{r(1)} \cdot \ldots \cdot t_{r(n)},$$

then

$$E\big[t_i h(t_1, \ldots, t_k)\big] = E[t_i t_{r(1)} \cdot \ldots \cdot t_{r(n)}] = \sum_{\pi \in \mathcal{P}(n+1)} \kappa_\pi(t_i, t_{r(1)}, \ldots, t_{r(n)}).$$

Now write $\pi = V \cup (\pi \setminus V)$, where $V = \{1, j_1, \ldots, j_\ell\}$ is the block of π containing 1. Then $E\big[t_i h(t_1, \ldots, t_k)\big]$

$$= \sum_{\pi \in \mathcal{P}(n+1)} \kappa_\pi(t_i, t_{r(1)}, \ldots, t_{r(n)})$$

$$= \sum_{\ell \geq 0} \sum_{j_1, \ldots, j_\ell} \sum_{\pi \setminus V} \kappa_{\ell+1}(t_i, t_{r(j_1)}, \ldots, t_{r(j\ell)}) \kappa_{\pi \setminus V}(t_{r(1)}, \ldots, t_{r(n)} \mid \{1, \ldots, n\} \setminus V)$$

$$= \sum_{\ell \geq 0} \sum_{j_1, \ldots, j_\ell} \kappa_{\ell+1}(t_i, t_{r(j_1)}, \ldots, t_{r(j\ell)}) E(t_{r(1)}, \ldots, t_{r(n)} \mid \{1, \ldots, n\} \setminus V)$$

$$= \sum_{\ell \geq 0} \sum_{i_1, \ldots, i_\ell} \sum_{j_1, \ldots, j_\ell} \kappa_{\ell+1}(t_i, t_{i_1}, \ldots, t_{i_\ell}) \cdot E\left[t_{r(1)} \cdot \ldots \cdot \frac{\partial t_{r(j_1)}}{\partial t_{i_1}} t_{r(j_1+1)} \cdot \ldots \cdot \frac{\partial t_{r(j_\ell)}}{\partial t_{i_\ell}} \cdot \ldots \cdot t_{r(n)}\right]$$

$$= \sum_{\ell \geq 0} \sum_{i_1, \ldots, i_\ell = 1}^{k} \kappa_{\ell+1}(t_i, t_{i_1}, \ldots, t_{i_\ell}) \cdot \frac{1}{\ell!} E[\partial_{i_1} \ldots \partial_{i_\ell} t_{r(1)} \ldots t_{r(n)}]$$

$$= \sum_{\ell \geq 0} \sum_{i_1, \ldots, i_\ell = 1}^{k} \frac{\kappa_{\ell+1}(t_i, t_{i_1}, \ldots, t_{i_\ell})}{\ell!} E[\partial_{i_1} \ldots \partial_{i_\ell} h(t_1, \ldots, t_k)].$$

7.5. Cumulants and Stieltjes transform for the random feature model

Now let us go back to our random feature model

$$(f_{ij}) = F := \sigma\left(\frac{W}{\sqrt{p}} X\right).$$

In order to use our cumulant expansion, we need (asymptotic) information about the cumulants of $\{f_{ij}\}$! For W and X this is easy, so let us start with those; then we will consider

$$(g_{ij}) = G := \frac{W}{\sqrt{p}} X$$

and finally $(f_{ij}) = F = \sigma(G)$.

Proposition 7.14. Let $X = (x_{ij}) \in \mathbb{R}^{p\times n}$ be a standard Gaussian random matrix. Then all cumulants in $\{x_{ij}\}$ are zero with the exception of the second order ones:

$$\kappa_2(x_{ij}, x_{k\ell}) = \delta_{ik}\delta_{j\ell}.$$

Proof. Since the x_{ij} are independent, mixed cumulants in them vanish by Theorem 7.12, thus the only possibly non-zero cumulants are $\kappa_n(x_{ij}, x_{ij}, \ldots, x_{ij})$. As we have seen in the Remark after Lemma 7.3, for a Gaussian random variable as x_{ij}, only $\kappa_2 \neq 0$. □

So for X and also for W the cumulants are easy. Now consider

$$(g_{ij}) = G := \frac{1}{\sqrt{p}} W \cdot X \in \mathbb{R}^{m\times n},$$

where

$$g_{ij} = \sum_{k=1}^{p} \frac{1}{\sqrt{p}} w_{ik} x_{kj}.$$

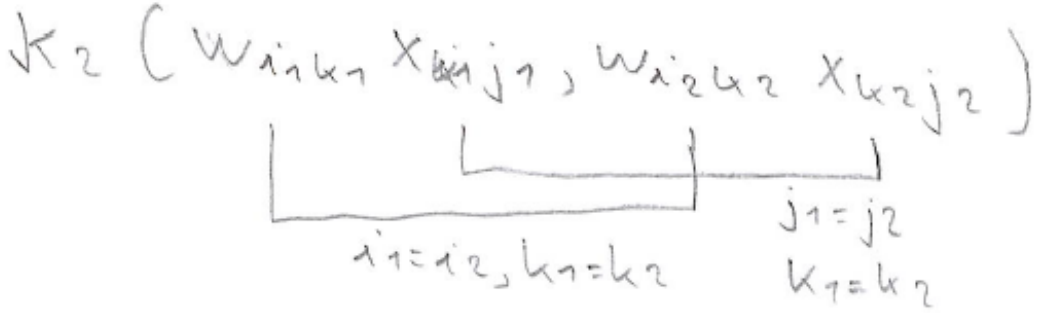
We have:

(i)

$$\kappa_1(g_{ij}) = \sum_{k=1}^{p} \frac{1}{\sqrt{p}} \kappa_1(w_{ik}x_{kj}) = \sum_{k=1}^{p} \frac{1}{\sqrt{p}} \big(\underbrace{\kappa_2(w_{ik}, x_{kj})}_{=0 \text{ (independ.)}} + \underbrace{\kappa_1(w_{ik})}_{=0} \underbrace{\kappa_1(x_{kj})}_{=0} \big) = 0.$$

(ii)

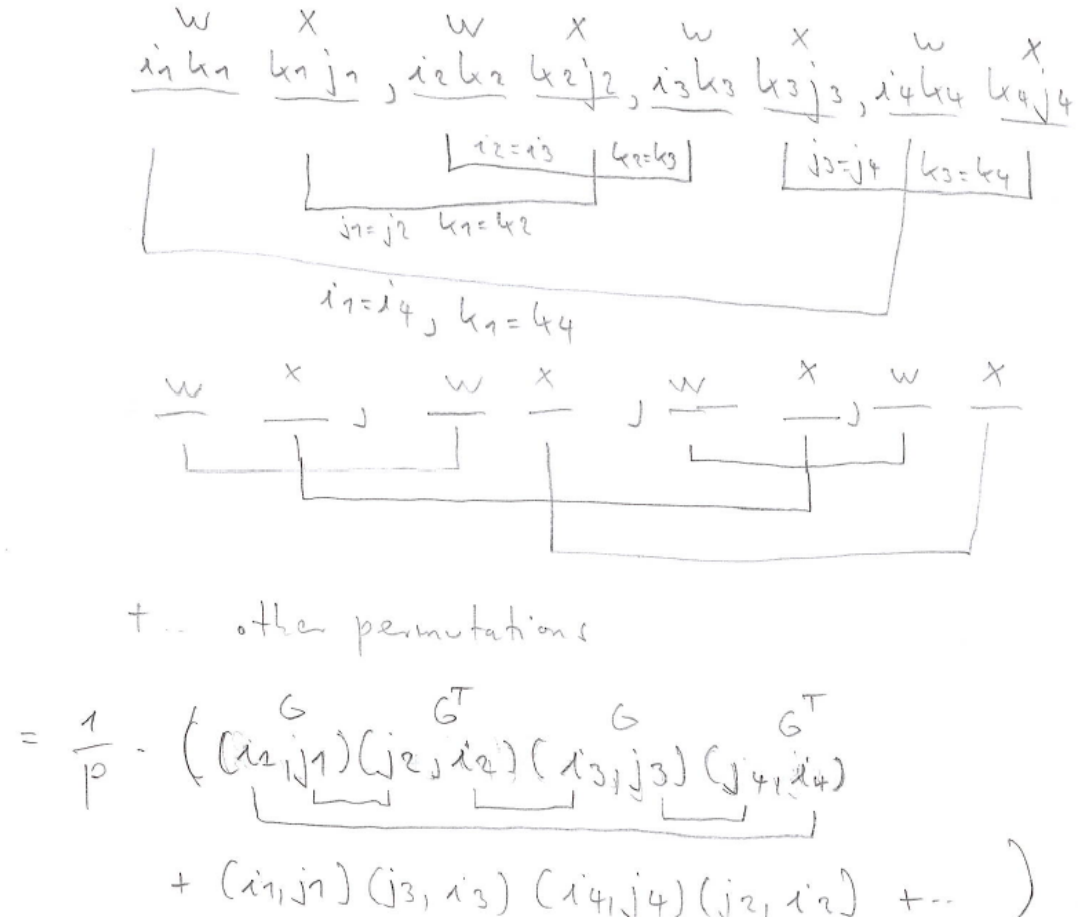
$$\kappa_2(g_{i_1j_1}, g_{i_2j_2}) = \frac{1}{p} \sum_{k_1,k_2=1}^{p} \kappa_2(w_{i_1k_1}x_{k_1j_1}, w_{i_2k_2}x_{k_2j_2}) = \delta_{i_1i_2}\delta_{j_1j_2} \cdot \frac{1}{p} \sum_{k_1=1}^{p} 1 = \delta_{i_1i_2}\delta_{j_1j_2}.$$



(iii) We have $\kappa_3(g_{i_1j_1}, g_{i_2j_2}, g_{i_3j_3}) = 0$ as well as all odd cumulants since we need an even number of x_{ij} (and of w_{ij}) to get a non-vanishing contribution.

(iv)

$$\kappa_4(g_{i_1j_1}, \ldots, g_{i_4j_4}) = \frac{1}{p^2} \sum_{k_1,\ldots,k_4=1}^{p} \kappa_4(w_{i_1k_1}x_{k_1j_1}, \ldots, w_{i_4k_4}x_{k_4j_4})$$



The same arguments lead to the corresponding result for higher cumulants.

Proposition 7.15. Let $W \in \mathbb{R}^{m\times p}$ and $X \in \mathbb{R}^{p\times n}$ be independent standard Gaussian matrices. Then the cumulants of

$$(g_{ij}) = G := \frac{1}{\sqrt{p}} W \cdot X \in \mathbb{R}^{m\times n}$$

are given by

$$\kappa_r(g_{i_1j_1}, g_{i_2j_2}, \ldots, g_{i_rj_r}) = \frac{1}{p^{\frac{r}{2}-1}} \cdot M,$$

where M is the number of permutations $\sigma \in S_r$ such that

$$(i_{\sigma(1)}\, j_{\sigma(1)})(j_{\sigma(2)}\, i_{\sigma(2)})(i_{\sigma(3)}\, j_{\sigma(3)}) \ldots\ldots\ldots (j_{\sigma(r)}\, i_{\sigma(r)})$$

has a cyclic structure, i.e.

$$j_{\sigma(1)} = j_{\sigma(2)}, \quad i_{\sigma(2)} = i_{\sigma(3)}, \quad j_{\sigma(3)} = j_{\sigma(4)}, \quad \dots \quad i_{\sigma(r)} = i_{\sigma(1)}.$$

Note that $M := 0$ for r odd.

Let us now use this structure of the cumulants in the cumulant expansion to calculate

$$S(z) = E\left[\operatorname{tr}\Big(\underbrace{\Big(\frac{GG^T}{n} - zI_m\Big)^{-1}}_{=:R(z)}\Big)\right]$$

via

$$\begin{aligned}
&1 + zS(z) \\
&\quad = \frac{1}{nm}\sum_{i,j} E\Big(g_{ij}\big[G^T R(z)\big]_{ji}\Big) \\
&\quad = \frac{1}{nm}\sum_{i,j}\sum_{\ell}\frac{1}{\ell!}\kappa_{\ell+1}(g_{ij}, g_{p_1q_1}, g_{p_2q_2}, \dots, g_{p_\ell q_\ell}) \cdot E\Big(\partial_{p_1q_1}\partial_{p_2q_2}\dots\partial_{p_\ell q_\ell}\big[G^T R(z)\big]_{ji}\Big) \\
&\quad = \frac{1}{nm}\sum_{r}\frac{1}{p^{r-1}}\sum_{\substack{i_1,\dots,i_r\\ j_1,\dots,j_r}} E\Big(\partial_{i_2j_1}\partial_{i_2j_2}\dots\partial_{i_1j_r}\big[G^T R(z)\big]_{j_1i_1}\Big).
\end{aligned}$$

This results in a term for $r = 1$, which was treated in the second proof of the Marchenko-Pastur law after Lemma 7.2, and terms for $r > 1$. For the latter ones, one does two of the partial integrations and reduces it to versions of $1 + zS(z)$. After quite a bit of approximations and technicalities (which we prefer not to do here), this finally gives the equation

$$1 + zS(z) = S(z) \cdot \Big(1 - \gamma\big(1 + zS(z)\big)\Big) \cdot \Big(1 - \frac{1}{\tilde{\gamma}}\big(1 + zS(z)\big)\Big),$$

which is the special case $\theta_1 = \theta_2 = 1$ of Theorem 7.1.

Now let us finally consider the effect of the non-linearity σ. The main observation is that the qualitative cumulant structure of G is preserved for $F = \sigma(G) = (f_{ij})$.

Proposition 7.16. In leading order we have for the cumulants of the $\{f_{ij}\}$:

(i) only cumulants with a cyclic structure are different from zero,

(ii) odd cumulants are zero,

(iii) we have

$$\kappa_2(f_{ij}, f_{ij}) = \theta_1(\sigma) = \int_{\mathbb{R}} \sigma(t)^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t,$$

(iv) and we have

$$\kappa_{2r}(f_{i_1j_1}, f_{i_2j_1}, f_{i_2j_2}, \ldots, f_{i_1j_r}) = \frac{1}{p^{r-1}} \theta_2(\sigma)^r$$

for disjoint $i_1, j_1, \ldots, i_r, j_r$ for $r > 1$, where

$$\theta_2(\sigma) = \left(\int_{\mathbb{R}} \sigma'(t) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t \right)^2.$$

Proof. (i+ii) This follows because of the cyclicity of the cumulants for the g_{ij}; only κ_1 might be problematic. For $f_{ij} = \sigma(g_{ij})$, note that

$$g_{ij} = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} w_{ik} x_{kj}.$$

Since $w_{ik}x_{kj}$ are independent for different k and also centered, g_{ij} is approximately Gaussian for large p by the central limit theorem. Thus

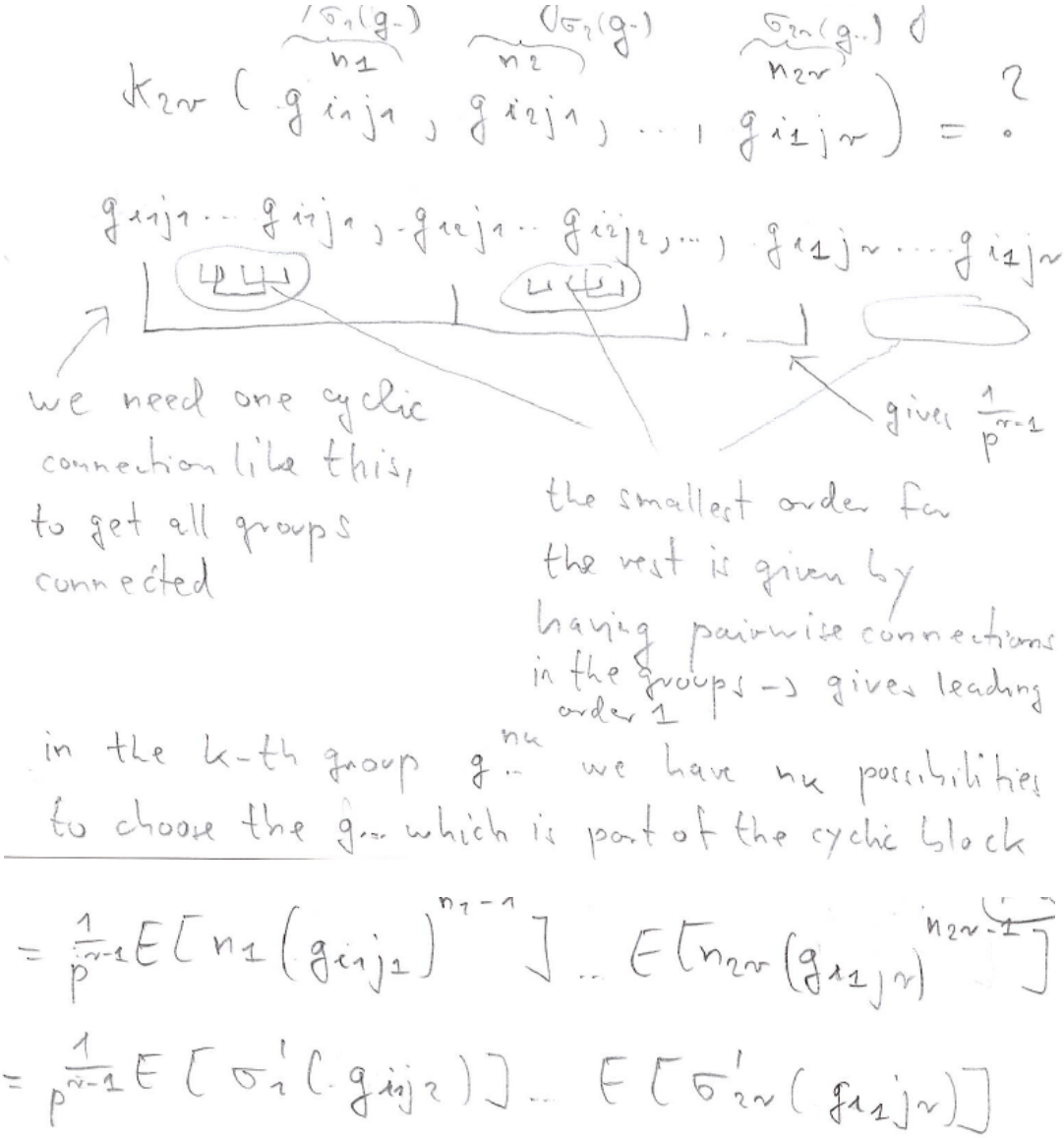
$$\kappa_1(f_{ij}) = E\big[\sigma(g_{ij})\big] \approx \int \sigma(t) \exp\left(-\frac{t^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \mathrm{d}t = 0$$

by our assumption in Theorem 7.1.

(iii) We have

$$\kappa_2(f_{ij}, f_{ij}) = E[f_{ij}^2] - \underbrace{E[f_{ij}]^2}_{=0} = E\big[\sigma(g_{ij})^2\big] = \int \sigma(t)^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t = \theta_1(\sigma).$$

(iv) Check for monomials and extend then by linearity and continuity.



and thus

$$\begin{aligned}\kappa_{2r}\big(\sigma(g_{i_1j_1}),\dots,\sigma(g_{i_1j_r})\big) &= \frac{1}{p^{r-1}}E\big[\sigma'(g_{i_1j_2})\big]^{2r} \\ &= \frac{1}{p^{r-1}}\left(\int_{\mathbb{R}}\sigma'(t)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)\,\mathrm{d}t\right)^{2r} \\ &= \frac{1}{p^{r-1}}\theta_2(\sigma)^r. \end{aligned}$$

□

Using this in the cumulant expansion yields then after quite some work the claimed equation for $S(z)$ in Theorem 7.1. For details of the calculations we refer to the paper by Piccolo and Schröder [PS21].

7.6. The Gaussian equivalence principle for the non-linear random feature model

Note that getting the final formula for $S(z)$ out of the cumulants might not be easy, but from Proposition 7.5 on the form of the cumulants it is very easy to see that

$$F = \sigma\left(\frac{W}{\sqrt{p}}X\right) \quad \text{and} \quad \tilde{F} = \sqrt{\theta_2}\frac{W}{\sqrt{p}}X + \sqrt{\theta_1 - \theta_2}Z = (\tilde{f}_{ij})$$

have in leading order the same cumulants and thus $\frac{1}{n}FF^T$ and $\frac{1}{n}\tilde{F}\tilde{F}^T$ have the same asymptotic eigenvalue distribution. Such statements are known as Gaussian equivalence principle. Let us check this. We have for $r > 1$

$$\kappa_{2r}(\tilde{f}_{i_1j_1}, \ldots, \tilde{f}_{i_1j_r}) = \theta_2^r \cdot \kappa_{2r}(g_{i_1j_1}, \ldots, g_{i_1j_r}) = \theta_2^r \cdot \frac{1}{p^{r-1}}$$

and

$$\kappa_2(\tilde{f}_{ij}, \tilde{f}_{ij}) = \theta_2 \underbrace{\kappa_2(g_{ij}, g_{ij})}_{=1} + (\theta_1 - \theta_2) \underbrace{\kappa_2(z_{ij}, z_{ij})}_{=1} = \theta_2 + \theta_1 - \theta_2 = \theta_1.$$

Note that mixed cumulants in WX and Z vanish and that Z only has second-order cumulants.

Note also that in the asymptotic calculation in the summations over $i_1, j_1, \ldots, i_r, j_r$ one can neglect terms where some of the indices are the same. That's good because we do not really have much control over terms like $\kappa_4(f_{11}, f_{11}, f_{12}, f_{12})$.

8. Gradient Descent and Neural Tangent Kernel

Consider our random feature neural network function

$$Y = f(X) = w\sigma(W_1 X),$$

where $w \in \mathbb{R}^{1\times m}$, $W_1 \in \mathbb{R}^{m\times p}$, and $X \in \mathbb{R}^{p\times n}$. We have a kind of an understanding of how the statistical properties of X influence the statistical properties of Y if W_1 and w are fixed (e.g., W_1 is randomly chosen). But the crux of a neural network is to find w and W_1 through "learning" such that a given set of data $\hat{Y} = f(X)$ is described best. In the above we can

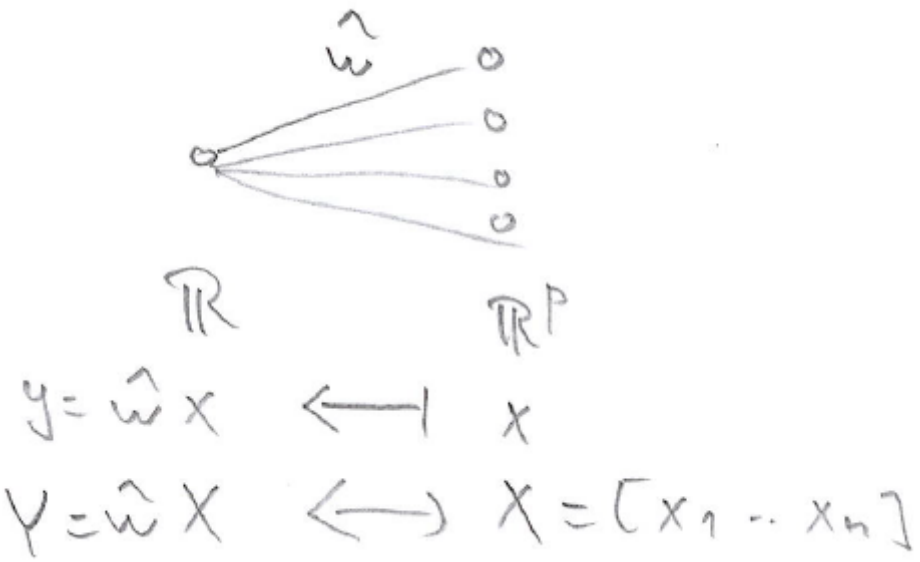
(i) fix W_1 (deterministically or randomly) and only learn w (*linear regression*), or

(ii) learn both w and W_1 (*feature learning*).

8.1. Gradient descent for linear regression

For linear regression we have an explicit solution for the best w, but in general this won't be the case and we need an algorithmic way for learning. The basic algorithm is *gradient descent*. Even for linear regression this has some value. Let us reconsider our linear regression problem in the over-parameterized ($\hat{=}$ under-determined) case (compare Section 6).

Given $X \in \mathbb{R}^{p\times n}$ and $\hat{Y} \in \mathbb{R}^{1\times n}$, we seek $\hat{w} \in \mathbb{R}^{1\times p}$ such that $\hat{Y} = \hat{w}X$ in the case $n < p$.



In the under-determined case $n < p$, there are typically (i.e. if $\operatorname{rank}(X) = n$) infinitely many solutions and the "best" (i.e., the one with the smallest norm) is given by

$$\hat{w} = \hat{Y}(X^T X)^{-1} X^T. \tag{3}$$

Since taking inverses of big matrices is usually not a good idea, even here it is better to have an algorithm to approximate the solution via iterations. Note that $\hat{Y} = \hat{w}X$ is the same as $\|\hat{Y} - \hat{w}X\|^2 = 0$, i.e. $\hat{w}$ minimizes

$$\|\hat{Y} - wX\|^2 = (\hat{Y} - wX)(\hat{Y} - wX)^T = \hat{Y}\hat{Y}^T - \hat{Y}X^T w^T - wX\hat{Y}^T + wXX^T w^T$$

and

$$\nabla_w \|\hat{Y} - wX\|^2 = -\hat{Y}X^T \cdot 2 + wXX^T \cdot 2.$$

Let us check the last equation on the gradient by explicit calculations with the components of the vectors and matrices. We also should decide whether we want to represent the gradient as a row or as a column vector. We prefer here the first possibility; i.e., we take ∇ as row vector and denote

$$w = \begin{pmatrix} w_1 & \dots & w_p \end{pmatrix}, \quad \hat{Y}X^T = a = \begin{pmatrix} a_1 & \dots & a_p \end{pmatrix}, \quad XX^T = A = (a_{ij})_{i,j=1}^p = A^T.$$

Then

$$\nabla_w(wa^T) = \nabla_w(aw^T) = \nabla_w(a_1 w_1 + \ldots + a_p w_p) = \begin{pmatrix} a_1 & \dots & a_p \end{pmatrix} = a$$

and

$$\begin{aligned}
\nabla_w(wAw^T) &= \nabla_w \left(\sum_{i,j=1}^{p} w_i a_{ij} w_j \right) \\
&= \left(\sum_{j=1}^{p} (a_{1j} w_j + w_j a_{j1}) \quad \dots \quad \sum_{j=1}^{p} (a_{pj} w_j + w_j a_{jp}) \right) \\
&= \left(\sum_{j=1}^{p} 2 w_j a_{j1} \quad \dots \quad \sum_{j=1}^{p} 2 w_j a_{jp} \right) \\
&= 2wA.
\end{aligned}$$

So $\nabla_w \|\hat{Y} - \hat{w}X\|^2 = 0$ gives the "normal" equation $\hat{w}XX^T = \hat{Y}X^T$, which gives the least square solution $\hat{w} = \hat{Y}X^T(XX^T)^{-1}$ in the over-determined case, and the best solution $\hat{w} = \hat{Y}X^T(XX^T)^{+} = \hat{Y}(X^T X)^{-1}X^T$ in the under-determined case, where A^{+} is the pseudo-inverse of A.

From an algorithmic point of view, the gradient gives the direction to improve an approximation, i.e. starting from some $w^{(0)}$, take

$$w^{(t+1)} := w^{(t)} - \eta \nabla_w \|\hat{Y} - w^{(t)} X\|^2 = w^{(t)} - \eta \cdot 2\big(w^{(t)} X X^T - \hat{Y} X^T\big),$$

where $t \in \mathbb{N}_0$ is the time and η is the "step size" or "learning rate". For η small enough this algorithm will converge. In the over-determined case, it converges to the least square solution, in the under-determined case it converges to a solution, but this might not be the best one. Thus it is advantageous to improve the algorithm by regularization.

Example 8.1. For $x = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \in \mathbb{R}^{2\times 1}$ and $w = \begin{pmatrix} \theta_1 & \theta_2 \end{pmatrix} \in \mathbb{R}^{1\times 2}$, consider a function

$$f(x) = wx = \begin{pmatrix} \theta_1 & \theta_2 \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = \theta_1 t_1 + \theta_2 t_2.$$

Assume we are given that

$$f\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = 1, \quad \text{i.e.} \quad X = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \hat{Y} = \begin{pmatrix} 1 \end{pmatrix}.$$

Thus we want to find the best $\hat{w} = \begin{pmatrix} \hat{\theta}_1 & \hat{\theta}_2 \end{pmatrix}$ such that $\hat{Y} = \hat{w} X$, i.e. $1 = \hat{\theta}_1 \cdot 0 + \hat{\theta}_2 \cdot 1$. Our solution (3) yields

$$\hat{w} = \hat{Y}(X^T X)^{-1} X^T = 1 \cdot \left(\begin{pmatrix} 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right)^{-1} \cdot \begin{pmatrix} 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \end{pmatrix},$$

i.e., $\hat{\theta}_1 = 0$ and $\hat{\theta}_2 = 1$, which is clearly the solution with the smallest norm. But now try to get a solution via gradient descent, where $w^{(t)} = \begin{pmatrix} \theta_1^{(t)} & \theta_2^{(t)} \end{pmatrix}$ and

$$\begin{aligned} \begin{pmatrix} \theta_1^{(t+1)} & \theta_2^{(t+1)} \end{pmatrix} = w^{(t+1)} &= w^{(t)} - \eta \nabla_w \|\hat{Y} - w^{(t)} X\|^2 \\ &= w^{(t)} - \eta \cdot 2 \left(w^{(t)} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & 1 \end{pmatrix} \right) \\ &= \begin{pmatrix} \theta_1^{(t)} & \theta_2^{(t)} - 2\eta(\theta_2^{(t)} - 1) \end{pmatrix}, \end{aligned}$$

thus $\theta_1^{(t+1)} = \theta_1^{(t)}$ and $\theta_2^{(t+1)} - 1 = (\theta_2^{(t)} - 1)(1 - 2\eta)$. Hence, if $|1 - 2\eta| < 1$, θ_2 converges to the right solution $\hat{\theta}_2 = 1$, but θ_1 does not change at all. So if we start with $\theta_1^{(0)} \neq 0$, we will not converge to the solution with the smallest norm.

Remark 8.2 (Ridge Regression)**.** Instead of minimizing

$$\mathcal{L}(w) := \|\hat{Y} - wX\|^2,$$

we add a penalty term for large norms, i.e. we want to minimize

$$\mathcal{L}_\lambda(w) := \|\hat{Y} - wX\|^2 + \lambda\|w\|^2 \quad \text{for some} \quad \lambda > 0.$$

The minimizer of this is determined by

$$0 \overset{!}{=} \nabla_w \mathcal{L}_\lambda(w) = 2(wXX^T - \hat{Y}X^T) + \lambda \underbrace{\nabla_w ww^T}_{=2w} = 2\Big(w(XX^T + \lambda I) - \hat{Y}X^T\Big),$$

thus $\hat{w}(XX^T+\lambda I) = \hat{Y}X^T$. Note that for $\lambda > 0$, the matrix $XX^T+\lambda I$ is always invertible even if XX^T is not invertible, since XX^T is positive semi-definite. So we have a unique solution

$$\hat{w} = \hat{Y}X^T(XX^T + \lambda I)^{-1},$$

which

- does not interpolate exactly anymore, but we allow small (depending on λ) errors, and
- is "nice" (has small norm, or more general some good smoothness properties).

Gradient descent with the "ridged" gradient converges to this solution. Note also that

$$\lim_{\lambda \searrow 0} X^T(XX^T + \lambda I)^{-1} = X^+,$$

the pseudo-inverse of X, and the ridge regression solution converges for $\lambda \searrow 0$ to the regression solution.

Example 8.3 (Ridge regression for Example 8.1)**.** The unique solution $\hat{w}_\lambda$ for λ is

$$\begin{aligned}\hat{w}_\lambda = \hat{Y}X^T(XX^T + \lambda I)^{-1} &= 1 \cdot \begin{pmatrix}0 & 1\end{pmatrix} \cdot \left(\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix} + \begin{pmatrix}\lambda & 0\\ 0 & \lambda\end{pmatrix}\right)^{-1} \\ &= \begin{pmatrix}0 & 1\end{pmatrix} \cdot \begin{pmatrix}\frac{1}{\lambda} & 0\\ 0 & \frac{1}{1+\lambda}\end{pmatrix} = \begin{pmatrix}0 & \frac{1}{1+\lambda}\end{pmatrix} \xrightarrow{\lambda \searrow 0} \begin{pmatrix}0 & 1\end{pmatrix} = \hat{w}.\end{aligned}$$

Gradient descent iteration gives now

$$w_\lambda^{(t+1)} = w_\lambda^{(t)} - 2\eta\left(\begin{pmatrix}0 & \theta_2^{(t)}\end{pmatrix} + \lambda w_\lambda^{(t)} - \begin{pmatrix}0 & 1\end{pmatrix}\right),$$

so

$$\theta_1^{(t+1)} = \theta_1^{(t)} - 2\eta\lambda\theta_1^{(t)]} = \theta_1^{(t)}(1 - 2\eta\lambda)$$

and

$$\theta_2^{(t+1)} = \theta_2^{(t)} - 2\eta\big(\theta_2^{(t)]}(1+\lambda) - 1\big).$$

In particular, we have

$$\theta_1^{(t)} \to 0 \quad \text{if} \quad |1 - 2\eta\lambda| < 1$$

and

$$\theta_2^{(t)} \to \frac{1}{1+\lambda} \quad \text{if} \quad |1 - 2\eta(1+\lambda)| < 1,$$

since

$$\theta_2^{(t+1)} - \frac{1}{1+\lambda} = \left(\theta_2^{(t)} - \frac{1}{1+\lambda}\right) \cdot \Big(1 - 2\eta(1+\lambda)\Big).$$

8.2. Gradient descent for feature learning

Now let's go beyond linear regression and allow all parameters to be trained. We consider

$$f(x) = w \cdot \sigma\left(\frac{1}{\sqrt{p}} W_1 x\right)$$

as before, but rename

$$w = a^T \quad \text{and} \quad \frac{1}{\sqrt{p}} W_1 = W$$

(our vectors, like a, should be column vectors). So we have

$$f_\theta(x) = a^T \cdot \sigma(Wx)$$

for $a \in \mathbb{R}^m$, $W \in \mathbb{R}^{m \times p}$, and $x \in \mathbb{R}^p$, depending on the parameters $\theta := \{a, W\}$ (a $(m + m \cdot p)$-dimensional vector). We are given n observations

$$(\hat{x}_1, \hat{y}_1), \ldots, (\hat{x}_n, \hat{y}_n) \in \mathbb{R}^p \times \mathbb{R},$$

i.e. we want

$$f_\theta(\hat{x}_k) = \hat{y}_k \quad \text{for all} \quad k = 1, \ldots, n.$$

We measure deviation from this by the loss function

$$\mathcal{L}(\theta) = \frac{1}{2} \sum_{k=1}^{n} \big(f_\theta(\hat{x}_k) - \hat{y}_k\big)^2.$$

We want to minimize this by changing θ via gradient descent:

$$\theta(t+1) = \theta(t) - \eta \nabla_\theta \mathcal{L}\big(\theta(t)\big),$$

or, by renaming,

$$\theta(t+\Delta t) = \theta(t) - \eta \Delta t \nabla_\theta \mathcal{L}\big(\theta(t)\big),$$

which, after rearranging and for $\Delta t \searrow 0$, becomes

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -\eta \nabla_\theta \mathcal{L}\big(\theta(t)\big).$$

We also write $\theta_t = \theta(t)$. Any change in θ induces a change in $f_t := f_{\theta(t)}$, which is our main concern. We have

$$\frac{\mathrm{d}f_t(x)}{\mathrm{d}t} = \frac{\mathrm{d}f_{\theta(t)}(x)}{\mathrm{d}t} = \nabla_\theta f_{\theta(t)}(x)^T \cdot \frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -\eta \nabla_\theta f_{\theta(t)}(x)^T \cdot \nabla_\theta \mathcal{L}(\theta_t).$$

Now

$$\nabla_\theta \mathcal{L}(\theta_t) = \sum_{k=1}^{n} \nabla_\theta \frac{1}{2}\big(f_{\theta_t}(\hat{x}_k) - \hat{y}_k\big)^2 = \sum_{k=1}^{n} \big(f_t(\hat{x}_k) - \hat{y}_k\big) \nabla_\theta f_t(\hat{x}_k)$$

and thus

$$\frac{\mathrm{d}f_t(x)}{\mathrm{d}t} = -\eta \sum_{k=1}^{n} \nabla_\theta f_t(x)^T \cdot \nabla_\theta f_t(\hat{x}_k) \cdot \big(f_t(\hat{x}_k) - \hat{y}_k\big).$$

8.3. Neural tangent kernel

We now define the *neural tangent kernel*

$$k_t(x, \tilde{x}) := \nabla_\theta f_t(x)^T \cdot \nabla_\theta f_t(\tilde{x}),$$

which was introduced by Jacot, Gabriel, and Hongler [JGH18] in 2018.

Note that a priori k_t is a probabilistic object which depends on time t. Unless we can say more about it, the above is just a compact and useless way of writing down the time-evolution in an abstract way.

However, it turns out that in the large width limit $m \to \infty$, k_t converges to a limit object k, which

- is deterministic (which we can believe by concentration),
- is independent of time (which is not so clear right now, maybe later there will be more on this), and

- stays away from zero, i.e. $k \geq \delta I$ for some $\delta > 0$, thus has only strictly positive eigenvalues.

Note that with

$$\hat{X} = \begin{pmatrix} \hat{x}_1 & \dots & \hat{x}_n \end{pmatrix} \quad \text{and} \quad \hat{Y} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix}$$

we have

$$\begin{aligned} \frac{\mathrm{d}f_t(x)}{\mathrm{d}t} &= -\eta \nabla_\theta f_t(x)^T \cdot \sum_{k=1}^{n} \nabla_\theta f_t(\hat{x}_k) \cdot \big(f_t(\hat{x}_k) - \hat{y}_k\big) \\ &= -\eta \nabla_\theta f_t(x)^T \cdot \nabla_\theta f_t(\hat{X}) \cdot \big(f_t(\hat{X}) - \hat{Y}\big) \end{aligned} \tag{4}$$

and thus

$$\frac{\mathrm{d}f_t(\hat{X})}{\mathrm{d}t} = -\eta \nabla_\theta f_t(\hat{X})^T \cdot \nabla_\theta f_t(\hat{X}) \cdot \big(f_t(\hat{X}) - \hat{Y}\big),$$

so

$$\frac{\mathrm{d}\big(f_t(\hat{X}) - \hat{Y}\big)}{\mathrm{d}t} = -\eta \nabla_\theta f_t(\hat{X})^T \cdot \nabla_\theta f_t(\hat{X}) \cdot \big(f_t(\hat{X} - \hat{Y})\big),$$

where

$$\nabla_\theta f_t(\hat{X})^T \cdot \nabla_\theta f_t(\hat{X}) \approx k(\hat{X}, \hat{X})$$

is constant and $\geq \delta I$, so $f_t(\hat{X})$ converges exponentially to $\hat{Y}$:

$$\big(f_t(\hat{X}) - \hat{Y}\big) = \exp\big(-\eta t k(\hat{X}, \hat{X})\big) \cdot \big(f_0(\hat{X}) - \hat{Y}\big),$$

8.4. Test error in the random feature model

Thus the training error goes to zero; but how about the test error? What is the prediction for $t \to \infty$ for arbitrary, "unseen" data x? By (4), we have

$$\frac{\mathrm{d}f_t(x)}{\mathrm{d}t} = -\eta k(x, \hat{X})\big(f_t(\hat{X}) - \hat{Y}\big) = -\eta k(x, \hat{X}) \exp\big(-\eta t k(\hat{X}, \hat{X})\big) \cdot \big(f_0(\hat{X}) - \hat{Y}\big),$$

so

$$f_t(x) = k(x, \hat{X}) \cdot k(\hat{X}, \hat{X})^{-1} \exp\big(-\eta t k(\hat{X}, \hat{X})\big) \cdot \big(f_0(\hat{X}) - \hat{Y}\big) + C.$$

For $t = 0$, we have

$$f_0(x) = C + k(x, \hat{X}) \cdot k(\hat{X}, \hat{X})^{-1}\big(f_0(\hat{X}) - \hat{Y}\big),$$

so

$$\begin{aligned} f_t(x) &= f_0(x) - k(x,\hat{X}) \cdot k(\hat{X},\hat{X})^{-1} \cdot \big(f_0(\hat{X}) - \hat{Y}\big) \\ &\quad + k(x,\hat{X}) \cdot k(\hat{X},\hat{X})^{-1} \exp\big(-\eta t k(\hat{X},\hat{X})\big) \cdot \big(f_0(\hat{X}) - \hat{Y}\big) \\ &= f_0(x) + k(x,\hat{X}) \cdot k(\hat{X},\hat{X})^{-1} \Big(\exp\big(-\eta t k(\hat{X},\hat{X})\big) - 1\Big) \cdot \big(f_0(\hat{X}) - \hat{Y}\big) \end{aligned}$$

and thus for $t \to \infty$

$$f_\infty(x) = f_0(x) + k(x,\hat{X}) \cdot k(\hat{X},\hat{X})^{-1} \cdot \big(\hat{Y} - f_0(\hat{X})\big).$$

Applying some centering we can restrict to the case where $f_0 = 0$, thus

$$f_\infty(x) = k(x,\hat{X}) k(\hat{X},\hat{X})^{-1} \hat{Y}.$$

We now have to prescribe some model for the unseen data, like

$$y = g(x) = a_T^T \sigma_T(W_T x) + N,$$

where the subindex T stands for "teacher" and N is some noise. Then the test error is

$$E_{\text{test}} = E_x\Big[\big(g(x) - f_\infty(x)\big)^2\Big] = E_x\Big[\big(g(x) - k(x,\hat{X}) k(\hat{X},\hat{X})^{-1} \hat{Y}\big)^2\Big].$$

In principle, this can be expressed as a complicated, but manageable (namely rational) function in the involved random matrices. In particular, note:

(i) The quantities $\nabla_\theta f_\theta(x)$ and thus $k(x,\tilde{x})$ can be given explicitly: from

$$f_\theta(x) = a^T \sigma(Wx)$$

we get as in linear regression

$$\nabla_a f_\theta(x) = \sigma(Wx).$$

But what is $\nabla_w f_\theta(x)$? Write

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} w_1^T \\ \vdots \\ w_m^T \end{pmatrix} \quad \text{for} \quad w_i \in \mathbb{R}^p,$$

then

$$f_\theta(x) = \sum_{i=1}^m a_i \sigma(w_i^T x)$$

and

$$\nabla_{w_i} f_\theta(x) = a_i \sigma'(w_i^T x) \cdot x,$$

thus

$$\begin{aligned} k(x, \tilde{x}) &= \nabla_a f_\theta(x)^T \cdot \nabla_a f_\theta(\tilde{x}) + \nabla_w f_\theta(x)^T \cdot \nabla_w f_\theta(\tilde{x}) \\ &= \sigma(Wx)^T \sigma(W\tilde{x}) + \sum_{i=1}^m a_i^2 \sigma'(w_i^T x)^T \sigma'(w_i^T \tilde{x}) x^T \tilde{x} \end{aligned}$$

and thus for the data matrices

$$\begin{aligned} k(X, \tilde{X}) &= \sigma(WX)^T \sigma(W\tilde{X}) + X^T \tilde{X} \odot \sum_{i=1}^m a_i^2 \sigma'(w_i^T X)^T \sigma'(w_i^T \tilde{X}) \\ &= \sigma(WX)^T \sigma(W\tilde{X}) + X^T \tilde{X} \odot \sigma'(WX)^T \operatorname{diag}(a) \sigma'(W\tilde{X}), \end{aligned}$$

where $\odot$ is the Hadamard product and

$$\operatorname{diag}(a) = \begin{pmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_m \end{pmatrix}.$$

(ii) By the Gaussian equivalence principle from Section 7.6, we can replace non-linear random matrices like $\sigma(WX)$ by linear+noise random matrices $\alpha WX + \beta Z$.

We will not go more into those calculations. For details one should see [AP20].

8.5. Concentration of the neural tangent kernel

We still should get a better understanding of the claimed asymptotic properties of the neural tangent kernel (NTK). Consider for simplicity the model

$$f_\theta(x) = \frac{1}{\sqrt{m}} a^T \sigma(Wx) = \frac{1}{\sqrt{m}} \sum_{i=1}^m a_i \sigma(w_i^T x),$$

where we only optimize over W and keep a fixed, thus $\theta = \{W\}$. For asymptotic statements we have to be precise about our normalizations; we choose:

- a_i uniformly on $\{-1, +1\}$, so that $\|a\| \sim 1$,
- W as a standard Gaussian random matrix, i.e. each $w_i \sim N(0, I_p)$, and
- $\|x\| = 1$.

The kernel k is then given by

$$k(x,\tilde{x}) = \nabla_w f_\theta(x)^T \cdot \nabla_w f_\theta(\tilde{x}) = x^T\tilde{x} \cdot \frac{1}{m}\sum_{i=1}^{m} \underbrace{a_i^2}_{=1} \sigma'(w_i^T x)\sigma'(w_i^T\tilde{x}).$$

Note that $\sigma'(w_i^T x)\sigma'(w_i^T\tilde{x})$ is independent for different i and has the same distribution for each i, thus they are i.i.d.. In particular, for $v \sim N(0, I_p)$, we have

$$\frac{1}{m}\sum_{i=1}^{m} \underbrace{a_i^2}_{=1} \sigma'(w_i^T x)\sigma'(w_i^T\tilde{x}) \xrightarrow{m\to\infty} E_v\big[\sigma'(v^T x)\sigma'(v^T\tilde{x})\big]$$

by the law of large numbers. Define the limiting NTK k^* by

$$k^*(x,\tilde{x}) = E_v\big[\sigma'(v^T x)\sigma'(v^T\tilde{x})\big]x^T\tilde{x}, \tag{5}$$

then by the above and by concentration, with high probability for sufficiently large m we have

$$|k(x,\tilde{x}) - k^*(x,\tilde{x})| < \varepsilon.$$

Example. In some cases one can also calculate the limiting NTK k^*. Let us consider $\sigma = \mathrm{ReLU}$, then

$$\sigma'(t) = \begin{cases} 1, & t > 0, \\ 0, & t < 0 \end{cases}.$$

Let $\|x\| = 1 = \|\tilde{x}\|$. What is $E_v[1_{\{v^T x>0\}} \cdot 1_{\{v^T\tilde{x}>0\}}]$? Note that $t_1 = v^T x$ and $t_2 = v^T\tilde{x}$ are two Gaussian vectors with covariance

$$\Sigma = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \quad \text{where} \quad \alpha = x^T\tilde{x},$$

thus

$$\Sigma^{-1} = \frac{1}{1-\alpha^2}\begin{pmatrix} 1 & -\alpha \\ -\alpha & 1 \end{pmatrix}$$

and we have the joint density

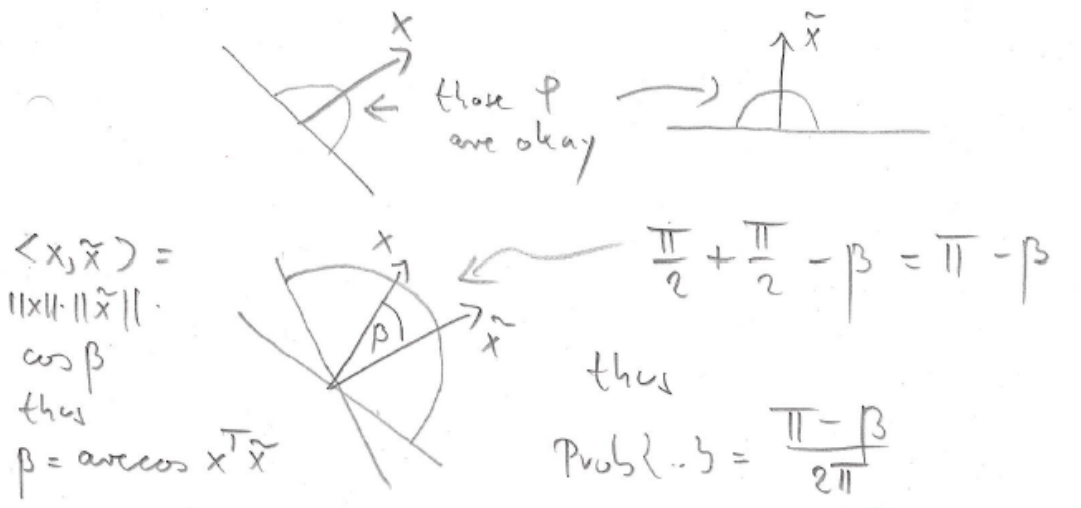
$$\begin{aligned}\psi(t_1,t_2) &= \frac{1}{2\pi\big(\det(\Sigma)\big)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left\langle \begin{pmatrix} t_1 \\ t_2 \end{pmatrix}, \Sigma^{-1}\begin{pmatrix} t_1 \\ t_2 \end{pmatrix}\right\rangle\right) \\ &= \frac{1}{2\pi\sqrt{1-\alpha^2}} \exp\left(-\frac{1}{2}\frac{t_1^2 + t_2^2 - 2\alpha t_1 t_2}{1-\alpha^2}\right).\end{aligned}$$

So we have to calculate

$$E_v[1_{\{v^T x>0\}} \cdot 1_{\{v^T\tilde{x}>0\}}] = \int_0^\infty \int_0^\infty \psi(t_1,t_2)\,\mathrm{d}t_1\,\mathrm{d}t_2.$$

This can be done by manipulating the integrals, but we prefer here another approach without explicit integration: This is the same problem if we restrict to the plane spanned by x and $\tilde{x}$, so we can assume that $x, \tilde{x}, v \in \mathbb{R}^2$ and $v \sim N(0, I_2)$. Since $v^T x > 0$ if and only if $\frac{v^T}{\|v\|} \cdot x > 0$, we can replace $v \sim N(0, I_2)$ by $\frac{v^T}{\|v\|}$ from the uniform distribution on $\mathcal{S}^1 = \{\exp(\mathrm{i}\varphi) \mid 0 \leq \varphi \leq 2\pi\}$. So what we want is, for given $x, \tilde{x} \in \mathbb{R}^2$:

$$\mathrm{P}\left\{\exp(\mathrm{i}\varphi) : \langle \exp(\mathrm{i}\varphi), x\rangle \geq 0 \text{ and } \langle \exp(\mathrm{i}\varphi), \tilde{x}\rangle \geq 0\right\}.$$



So

$$k^*(x, \tilde{x}) = x^T \tilde{x} \frac{\pi - \arccos(x^T \tilde{x})}{2\pi}.$$

8.6. Evolution of the neural tangent kernel under training

Now consider the evolution under training. First we show that the weight vectors do not change much. Recall that we have (put $\eta = 1$)

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -\nabla_\theta \mathcal{L}\big(\theta(t)\big) = -\sum_{k=1}^{n} \big(f_t(\hat{x}_k) - \hat{y}_k\big) \cdot \nabla_\theta f_t(\hat{x}_k).$$

For $\theta = w_i$ we get

$$\nabla_{w_i} f_t(\hat{x}_k) = a_i \sigma'(w_i^T \hat{x}_k) \hat{x}_k \cdot \frac{1}{\sqrt{m}},$$

thus

$$\frac{\mathrm{d}w_i}{\mathrm{d}t} = -\frac{1}{\sqrt{m}} \sum_{k=1}^{n} \big(f_t(\hat{x}_k) - \hat{y}_k\big) \cdot a_i \sigma'(w_i^T \hat{x}_k) \hat{x}_k.$$

Now consider the evolution of the weights: since $w_i^T \hat{x}_k$ is a Gaussian variable of variance $\|\hat{x}_k\| = 1$, we have

$$\begin{aligned}
\|w_i(t) - w_i(0)\|_2 &= \left\| \int_0^t \frac{\mathrm{d}w_i(\tau)}{\mathrm{d}\tau} \, \mathrm{d}\tau \right\|_2 \\
&= \left\| \int_0^t \frac{1}{\sqrt{m}} \sum_{k=1}^{n} \big(f_t(\hat{x}_k) - \hat{y}_k\big) \cdot a_i \sigma'(w_i^T \hat{x}_k) \hat{x}_k \, \mathrm{d}\tau \right\|_2 \\
&\leq \frac{1}{\sqrt{m}} \sum_{k=1}^{n} \int_0^t [\text{order } 1] \, \mathrm{d}\tau \\
&\sim \text{order } \frac{t \cdot n}{\sqrt{m}},
\end{aligned}$$

which is small if $m \to \infty$ for fixed t and n.

Now consider the change in the kernel: since $x^T \tilde{x} \leq \|x\| \cdot \|\tilde{x}\| = 1$, we have

$$\begin{aligned}
&|k_t(x, \tilde{x}) - k_0(x, \tilde{x})| \\
&= \left| x^T \tilde{x} \frac{1}{m} \sum_{i=1}^{m} \Big(\sigma'\big(w_i(t)^T x\big) \sigma'\big(w_i(t)^T \tilde{x}\big) - \sigma'\big(w_i(0)^T x\big) \sigma'\big(w_i(0)^T \tilde{x}\big) \Big) \right| \\
&\leq \frac{1}{m} \sum_{i=1}^{m} \Big| \sigma'\big(w_i(t)^T x\big) \sigma'\big(w_i(t)^T \tilde{x}\big) - \sigma'\big(w_i(t)^T x\big) \sigma'\big(w_i(0)^T \tilde{x}\big) \\
&\qquad + \sigma'\big(w_i(t)^T x\big) \sigma'\big(w_i(0)^T \tilde{x}\big) - \sigma'\big(w_i(0)^T x\big) \sigma'\big(w_i(0)^T \tilde{x}\big) \Big| \\
&\leq \frac{1}{m} \sum_{i=1}^{m} \Big(\max\big\{ \big|\sigma'\big(w_i(t)^T x\big)\big| \big\} \cdot \big| \sigma'\big(w_i(t)^T \tilde{x}\big) - \sigma'\big(w_i(0)^T \tilde{x}\big) \big| \\
&\qquad + \big| \sigma'\big(w_i(t)^T x\big) - \sigma'\big(w_i(0)^T x\big) \big| \cdot \max\big\{ \big|\sigma'\big(w_i(0)^T \tilde{x}\big)\big| \big\} \Big) \\
&\leq \frac{1}{m} \sum_{i=1}^{m} \Big(\max\big\{ \big|\sigma'\big(w_i(t)^T x\big)\big| \big\} \cdot \max\{\sigma''(\ldots)\} \cdot \big\| \big(w_i(t)^T - w_i(0)^T\big) \tilde{x} \big\| \\
&\qquad + \max\{\sigma''(\ldots)\} \cdot \big\| \big(w_i(t)^T - w_i(0)^T\big) x \big\| \cdot \max\big\{ \big|\sigma'\big(w_i(0)^T \tilde{x}\big)\big| \big\} \Big).
\end{aligned}$$

Since

$$\big\| \big(w_i(t)^T - w_i(0)^T\big) \tilde{x} \big\| \sim \frac{t \cdot n}{\sqrt{m}} \cdot \|\tilde{x}\| = \frac{t \cdot n}{\sqrt{m}},$$

we have that $|k_t(x, \tilde{x}) - k_0(x, \tilde{x})|$ is of order $\frac{t \cdot n}{\sqrt{m}}$, which goes to zero for $m \to \infty$ for fixed t and n.

If we put $k_{ij} = k(\hat{x}_i, \hat{x}_j)$ such that $k = (k_{ij})$ is an $n \times n$-matrix, then also in operator norm

$$\|k(t) - k(0)\| \leq \|k(t) - k(0)\|_F = \left(\sum_{i,j=1}^{n} |k_{ij}(t) - k_{ij}(0)|^2 \right)^{\frac{1}{2}}$$
$$\leq \left(\sum_{i,j=1}^{n} \left[\text{order } \frac{t \cdot n}{\sqrt{m}} \right] \right)^{\frac{1}{2}} \sim \frac{tn^2}{\sqrt{m}} \xrightarrow{m \to \infty} 0$$

for fixed t and n.

8.7. Boundedness away from zero of the neural tangent kernel

In order to see that the limiting NTK k^* according to Equation (5) (and thus also its approximations in high dimensions) has only positive eigenvalues which are bounded away from zero, i.e., $k^* \geq \delta \cdot I$, one should note:

- k^* is essentially diagonal, $k^*(x, \tilde{x}) \approx 0$ since two vectors in high dimension are with high probability almost orthogonal, and
- $k^*(x, x) = E_v\left[\sigma'(v^T x)\sigma'(v^T x)\right] \geq \delta$ in general.

As an example for the latter statement, let us check this concretely for the case $\sigma = \text{ReLU}$; then

$$k^*(x, x) = \frac{\pi - \arccos(1)}{2\pi} = \frac{\pi - \frac{\pi}{2}}{2\pi} = \frac{1}{4}.$$

9. (Operator-valued) Free Probability Theory

We have seen that the calculation of the eigenvalue distribution or the Stieltjes transform of polynomials or even rational functions in several random matrices is relevant. *Free probability theory*, which was introduced by Dan Voiculescu in the 1980's, provides powerful tools for dealing with this. In the following we will give an appetizer for this; for more details on those topics, see [MS17].

9.1. Free cumulants and freeness

In Section 7 we have seen that our matrices often have some special structure (at least asymptotically) for the cumulants of their entries. In order not to have to bother with the transpose for general rectangular matrices, we consider now symmetric square matrices:

$$X = (x_{ij})_{i,j=1}^{n}, \quad X = X^T \quad (\text{i.e. } x_{ij} = x_{ji}).$$

For those, typically in leading order only cumulants with cyclic index structure survive:

$$\kappa_\ell(x_{i(1)i(2)}, x_{i(2)i(3)}, \ldots, x_{i(\ell)i(1)}) \sim n^{-(l-1)}.$$

Their value is independent of $i(1), \ldots, i(\ell)$ for distinct $i(1), \ldots, i(\ell)$. So let us put

$$r_\ell := \lim_{n\to\infty} n^{\ell-1} \kappa_\ell(x_{i(1)i(2)}, x_{i(2)i(3)}, \ldots, x_{i(\ell)i(1)}).$$

Note that the order $n^{-(l-1)}$ is the right one for a ℓ-th cumulant to make a contribution in the calculation of

$$\begin{aligned} E\big[\mathrm{tr}(X^\ell)\big] = E\left[\frac{1}{n}\mathrm{Tr}(X^\ell)\right] &= \frac{1}{n}\sum_{i(1),\ldots,i(\ell)=1}^{n} E[x_{i(1)i(2)}x_{i(2)i(3)} \cdot \ldots \cdot x_{i(n)i(1)}] \\ &= \frac{1}{n}\sum_{i(1),\ldots,i(\ell)=1}^{n} \sum_{\pi\in\mathcal{P}(\ell)} \kappa_\pi(x_{i(1)i(2)}, x_{i(2)i(3)}, \ldots, x_{i(n)i(1)}). \end{aligned}$$

Since $\kappa_\pi(x_{i(1)i(2)}, x_{i(2)i(3)}, \ldots, x_{i(n)i(1)})$ is at most of order $n^{-(\ell-1)}$ and the sum over the $i(j)$ has about n^ℓ terms, we get $E\big[\mathrm{tr}(X^\ell)\big] \sim 1$.

If one collects the leading order contributions in this, one gets a "non-commutative" version of a moment-cumulant relation between the $m_\ell := \lim_{n\to\infty} \mathrm{tr}(X^\ell)$ and the r_ℓ.

Example. If $\ell = 1$, then

$$E\big[\mathrm{tr}(X)\big] = \sum_i \frac{1}{n} E[x_{ii}] = \sum_i \frac{1}{n}\kappa_1(x_{ii}) \to r_1$$

and

$$E\big[\mathrm{tr}(X^2)\big] = \sum_{i,j} \frac{1}{n} E[x_{ij}x_{ji}] = \sum_{i,j} \frac{1}{n} \big(\underbrace{\kappa_1(x_{ij})}_{\sim \delta_{ij} r_1} \kappa_1(x_{ij}) + \underbrace{\kappa_2(x_{ij}, x_{ji})}_{\sim \frac{1}{n} r_2} \big) \to r_1 r_1 + r_2,$$

this looks like the normal moment-cumulant relation for the usual cumulants. But consider now $\ell = 4$, and assume that odd cumulants don't contribute, i.e. $r_1 = r_3 = 0$. Then

$$E\big[\mathrm{tr}(X^4)\big] = \frac{1}{n} \sum_{i(1),\ldots,i(4)=1}^{n} \underbrace{x_{i(1)i(2)} x_{i(2)i(3)} x_{i(3)i(4)} x_{i(4)i(1)}}_{\kappa_{\sqcup\sqcup\sqcup\sqcup} + \kappa_{\sqcup\,\sqcup} + \kappa_{\sqcup\,\sqcup} + \kappa_{\sqcup\,\sqcup}} \cdot$$

Looking at the summands in detail, we have

$$\begin{aligned}
\kappa_{\sqcup\sqcup\sqcup\sqcup} &\sim r_4 n^{-3} \rightsquigarrow r_4, \\
\kappa_{\sqcup\,\sqcup} &= \kappa_2(x_{i(1)i(2)}, x_{i(2)i(3)}) \cdot \kappa_2(x_{i(3)i(4)}, x_{i(4)i(1)}) \\
&\sim \delta_{i(1)i(3)} r_2 n^{-1} \cdot \delta_{i(3)i(1)} r_2 n^{-1} \\
&\sim r_2 \cdot r_2 n^{-2} \delta_{i(1)i(3)} \\
&\rightsquigarrow r_2 \cdot r_2, \\
\kappa_{\sqcup\,\sqcup} &\rightsquigarrow r_2 \cdot r_2, \\
\kappa_{\sqcup\,\sqcup} &= \kappa_2(x_{i(1)i(2)}, x_{i(3)i(4)}) \cdot \kappa_2(x_{i(2)i(3)}, x_{i(4)i(1)}) \\
&\sim \delta_{i(1)i(4)} \delta_{i(2)i(3)} \delta_{i(1)i(3)} \delta_{i(2)i(4)} r_2 \cdot (\ldots) r_2 \\
&\sim r_2 \cdot r_2 n^{-2} \cdot \delta(\text{at least two indices equal}) \\
&\rightsquigarrow 0.
\end{aligned}$$

Thus $\sqcup\!\!\sqcup\!\!\sqcup$ does asymptotically not contribute; this is true in general, *crossing* partitions do not contribute and we have the "free" moment-cumulant relation:

$$m_\ell = \sum_{\pi \in \mathrm{NC}(\ell)} r_\pi,$$

where $\mathrm{NC}(\ell) \subset \mathcal{P}(\ell)$ is the set of all non-crossing partitions. The r_ℓ are called *free cumulants*. Note that one also has a multivariate version of this, then the vanishing of classical mixed cumulants of entries of two independent matrices implies vanishing of the corresponding mixed free cumulants, which gives a notion of "free independence" or "freeness", hence the name free probability theory.

9.2. Linearization of non-linear problems

Vanishing of mixed cumulants implies additivity of free cumulants for the sum, and allows sums of "asymptotically free" random matrices to be treated. This looks nice, but how

about polynomials or rational functions, as they showed up in Section 7? Those can also be addressed via the following linearization trick. The idea is to reformulate a polynomial (non-linear) problem into a linear one with matrix coefficients, which makes it an operator-valued linear problem.

We will give the idea of this only via a concrete example.

Example. Consider the polynomial $P = p(X,Y) = XY + YX + X^2$. We want its Stieltjes transform

$$S_p(z) = E\Big[\mathrm{tr}\big((P - z\cdot 1)^{-1}\big)\Big].$$

How do we deal with the inverse of $P - z\cdot 1$? For this, we embed the problem in matrices:

$$\begin{pmatrix} XY+YX+X^2-z\cdot 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & Y+\frac{X}{2} & X \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} -z & X & Y+\frac{X}{2} \\ X & 0 & -1 \\ Y+\frac{X}{2} & -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ Y+\frac{X}{2} & 1 & 0 \\ X & 0 & 1 \end{pmatrix}$$

and since the triangular matrices are always invertible, we have

$$\begin{pmatrix} (P-z\cdot 1)^{-1} & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ Y+\frac{X}{2} & 1 & 0 \\ X & 0 & 1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} -z & X & Y+\frac{X}{2} \\ X & 0 & -1 \\ Y+\frac{X}{2} & -1 & 0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 1 & Y+\frac{X}{2} & X \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ * & 1 & 0 \\ * & * & 1 \end{pmatrix} \cdot \begin{pmatrix} -z & X & Y+\frac{X}{2} \\ X & 0 & -1 \\ Y+\frac{X}{2} & -1 & 0 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 1 & * & * \\ 0 & 1 & * \\ 0 & 0 & 1 \end{pmatrix}$$

and thus

$$(P - z\cdot 1)^{-1} = \Big[\big(\hat{P} - \Lambda(z)\big)^{-1}\Big]_{1,1},$$

where

$$\hat{P} := \begin{pmatrix} 0 & X & Y+\frac{X}{2} \\ X & 0 & -1 \\ Y+\frac{X}{2} & -1 & 0 \end{pmatrix} \quad \text{and} \quad \Lambda(z) := \begin{pmatrix} z & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Now $\hat{P}$ is a linear polynomial in X and Y with matrix coefficients:

$$\hat{P} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} \cdot 1 + \begin{pmatrix} 0 & 1 & \frac{1}{2} \\ 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{pmatrix} \cdot X + \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \cdot Y.$$

This can then be dealt with by *operator-valued free probability theory*, which allows to calculate the distribution of such matrix-valued linear combinations.

Note also that the above factorization of P might have looked very special, but indeed for any polynomial (and actually also for any non-commutative rational function) P there exists (via a concrete algorithm) such a linearization $\hat{P}$.

10. Assignments

10.1. Assignment 1

Exercise 1 (5 points)**.** Show that

$$\int_{\mathbb{R}} \exp(-t^2)\,\mathrm{d}t = \sqrt{\pi}.$$

Hint: start by showing that

$$\left(\int_{\mathbb{R}} \exp(-t^2)\,\mathrm{d}t\right)^2 = \int_{\mathbb{R}}\int_{\mathbb{R}} \exp(-t^2 - s^2)\,\mathrm{d}t\,\mathrm{d}s$$

and compute the double integral using polar coordinates.

Definition. A real random variable x is a *Gaussian random variable* with *mean* $\mu \in \mathbb{R}$ and *variance* $\sigma^2 \in (0, \infty)$, denoted by $x \sim N(\mu, \sigma^2)$, if its probability density function ψ is given by

$$\psi : \mathbb{R} \to \mathbb{R}, \quad t \mapsto \psi(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right).$$

If $\mu = 0$ and $\sigma = 1$, then x is also called a *standard* Gaussian.

For a function $f : \mathbb{R} \to \mathbb{R}$, the *expectation* of $f(x)$ is

$$E[f(x)] = \int_{\mathbb{R}} f(t)\psi(t)\,\mathrm{d}t;$$

the *n-th moment* of x is given by

$$E[x^n] = \int_{\mathbb{R}} t^n \psi(t)\,\mathrm{d}t.$$

Exercise 2 $(3+4+3+3+2$ points$)$**.** Let $x \sim N(\mu, \sigma^2)$.

(a) Use Exercise 1 to compute $E[x^0]$ and $E[x^1]$. Explain the results.

(b) Show that x satisfies the moment recursion

$$E[x^n] = \mu E[x^{n-1}] + (n-1)\sigma^2 E[x^{n-2}] \quad \text{for all integers} \quad n \geq 2.$$

(c) Find the higher moments $E[x^2]$, $E[x^3]$, and $E[x^4]$.

(d) Give an explicit formula for the moments of x in the case $\mu = 0$.

(e) Calculate for the standard Gaussian $x \sim N(0,1)$ the first central moment

$$E\left[|x|\right] = \int_{\mathbb{R}} |t| \psi(t)\, \mathrm{d}t.$$

Exercise 3 $(3+3+4$ points$)$**.** We know from class that

$$\begin{aligned} P\left\{(t_1, \ldots, t_p) \in B_p : |t_p| \geq \varepsilon\right\} &= \frac{2\int_\varepsilon^1 \mathrm{vol}[B_{p-1}(\sqrt{1-t^2})]\, \mathrm{d}t}{\mathrm{vol}[B_p]} \\ &= 2\frac{\mathrm{vol}[B_{p-1}]}{\mathrm{vol}[B_p]} \int_\varepsilon^1 (1-t^2)^{\frac{p-1}{2}}\, \mathrm{d}t. \end{aligned}$$

Note that this includes also in particular for $\varepsilon = 0$ a formula for the ratio of the unit balls of consecutive dimensions:

$$1 = 2\frac{\mathrm{vol}[B_{p-1}]}{\mathrm{vol}[B_p]} \int_0^1 (1-t^2)^{\frac{p-1}{2}}\, \mathrm{d}t.$$

By estimating the integrals we want to show from this an estimate for

$$P\left\{(t_1, \ldots, t_p) \in B_p : |t_p| \geq \varepsilon\right\}.$$

(a) Prove for $y \geq 0$ the estimate

$$\int_y^\infty \exp(-t^2)\, \mathrm{d}t \leq \frac{\sqrt{\pi}}{2} \exp(-y^2).$$

Hint: treat the cases $y \leq 1$ and $y > 1$ separately.

(b) Let $p \geq 3$. Prove that

$$\int_0^1 (1-t^2)^{\frac{p-1}{2}}\,\mathrm{d}t \geq \int_0^{\frac{1}{\sqrt{p-1}}} (1-t^2)^{\frac{p-1}{2}}\,\mathrm{d}t \geq \frac{1}{2\sqrt{p-1}}.$$

Hint: Bernoulli's inequality states that $(1+a)^b \geq 1+ab$ *for all real numbers* $b \geq 1$ *and* $a \geq -1$.

(c) Let $p \geq 3$. Show that

$$P\left\{(t_1,\ldots,t_p) \in B_p : |t_p| \geq \varepsilon\right\} \leq \sqrt{2\pi}\exp\left(-\varepsilon^2\frac{p-1}{2}\right),$$

and thus

$$P\left\{(t_1,\ldots,t_p) \in B_p : |t_p| \leq \varepsilon\right\} \geq 1-\sqrt{2\pi}\exp\left(-\varepsilon^2\frac{p-1}{2}\right).$$

Hint: use Lemma 1.4: for $p \geq 1$ *und* $0 < \varepsilon \leq 1$ *we have* $(1-\varepsilon)^p \leq \exp(-\varepsilon p)$.

Definition. Let $x = (t_1,\ldots,t_p) \in \mathbb{R}^p$. We define the following norms:

- $\|x\|_2 := \sqrt{\sum_{k=1}^{p} t_k^2}$ (Euclidean norm, length, 2-norm)
- $\|x\|_1 := \sum_{k=1}^{p} |t_k|$ (ℓ_1 norm, Manhattan norm, 1-norm)
- $\|x\|_\infty := \max\{|t_k| : 1 \leq k \leq p\}$ (maximum norm, infinity norm)

Exercise 4 $(4+3+3$ points$)$**.** In this exercise, you are tasked with performing some numerical experiments and presenting the results as a histogram similar to the ones shown in the slides of the first lecture. You are free to choose your tools to do this, for example, you can use computer algebra systems with integrated plotting like MATLAB, Maple, or Mathematica, or use a programming language of your choice to compute the values and combine it with some visualization tool to plot the histogram.

As the slides in class, this exercise should give you a feeling for the concentration phenomena. We consider in the following Gaussian random vectors $x \in \mathbb{R}^p$ with independent

standard Gaussians as components; i.e., every component of the vector is a Gaussian random variable with mean zero and variance 1 and the components are independent from each other. Such vectors show concentration.

The concentration property says roughly that for our high-dimensional vector $x = (t_1, \ldots, t_p) \in \mathbb{R}^p$ any function $f(x) = f(t_1, \ldots, t_p)$ that depends (in a 'smooth' way) on the components (but not too much on any of them) is essentially constant, and thus close to the average value $E[f(x)]$ of the function. (Later in the course the parentheticals will be made more precise via the notion of Lipschitz functions.) In part (a) we consider the relatively simple situation where the function f is essentially a sum of independent components. In that case the expectation is also quite easy to determine. In part (b), the function f is much more non-linear, and its expectation is not directly clear. In part (c), we arrange our vectors in a matrix form and take as function f the largest eigenvalue of those matrices – these are very non-linear (and not very concrete) functions of the matrix entries, but still 'smooth enough', so that we also have concentration of the eigenvalues.

(a) For f we take here the 1-norm $f(x) = \|x\|_1$ and the 2-norm $f(x) = \|x\|_2$. For each of the two cases plot a histogram of $f(x)$ for $1,000$ realizations of the vector $x \in \mathbb{R}^p$. Do this for $p = 1$, $p = 100$, and $p = 10,000$. You should recognize in those plots the dependence of $E[f(x)]$ on p. Can you explain those values? (For the case of the 1-norm, Exercise 2(e) should be relevant.)

(b) For f we take now the maximum norm $f(x) = \|x\|_\infty$. Plot a histogram of $f(x)$ for 1,000 realizations of the vector $x \in \mathbb{R}^p$. Do this for $p = 1$, $p = 10$, $p = 10,000$, and $p = 100,000$.

The value of $E[f(x)]$ will probably not become clear from the plots. Instead, we can look at some estimates for the concentration: let M be the median of the $f(x_j)$, then for all $\varepsilon > 0$ we have

$$P\left(f(x) > (1+\varepsilon)M\right) \leq \sqrt{\frac{2}{\pi}} \frac{p}{(1+\varepsilon)M} \exp\left(-\frac{1}{2}(1+\varepsilon)^2 M^2\right).$$

Check for some reasonable values for ε whether this is compatible with your data.

(c) We consider now a sample covariance matrix

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} x_k x_k^T = \frac{1}{n} X X^T,$$

where $x_1, \ldots, x_n \in \mathbb{R}^p$ are n independent copies of our p-dimensional random vectors x as above, and $X = [x_1 x_2 \ldots x_n] \in \mathbb{R}^{p \times n}$ is the corresponding data matrix. (Such random matrices $\hat{\Sigma}$ are called *Wishart matrices*.) We take as our function f now the largest eigenvalue of $\hat{\Sigma}$ (which is the same as the square of the largest singular value of the matrix $X/\sqrt{n}$.) This $f(X) = f(x_1, \ldots, x_n)$ is a very non-linear (and not explicit) function of the $p \times n$ independent standard Gaussian entries of the data matrix X. Plot a histogram of $f(X)$ for $1,000$ realizations of the data matrix $X = [x_1 \ldots x_n]$. Do this for $p = n = 1$, $p = n = 10$, $p = n = 50$, $p = n = 100$.

In this case concentration estimates are quite complicated and not very explicit, so let us just quote the following simple rules of thumb (according to the paper "On the distribution of the largest eigenvalue in principal component analysis" by Iain Johnstone): define

$$\mu := \frac{1}{n}\left(\sqrt{n-1} + \sqrt{p}\right)^2 \quad \text{and} \quad \sigma := \frac{1}{n}\left(\sqrt{n-1} + \sqrt{p}\right)\left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}}\right)^{\frac{1}{3}}.$$

Then, about 83% of the distribution is less than μ, about 95% lies below $\mu + \sigma$, and about 99% lies below $\mu + 2\sigma$.

Check whether this is compatible with your data.

Further experimentation is encouraged.

10.2. Assignment 2

Definition. A random vector $x \in \mathbb{R}^p$ is a *Gaussian random vector* with mean vector $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$, denoted $x \sim N(\mu, \Sigma)$, if its probability density function ψ is given by

$$\psi(x) = \frac{1}{(2\pi)^{p/2}\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}\langle x-\mu, \Sigma^{-1}(x-\mu)\rangle\right).$$

The mean μ can be an arbitrary vector in $\mathbb{R}^p$, but the covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ has to be positive definite.

If $\mu = 0$ and $\Sigma = I_p$, then x is also called a *standard Gaussian random vector*.

Exercise 5 (6 points)**.** Consider n independent copies $x_1, \ldots, x_n \in \mathbb{R}^p$ of Gaussian random vectors with mean zero, where the components of each x_k are independent and half of them has variance 1 and the other half has variance 2. Plot a histogram of the p eigenvalues of the sample covariance matrix

$$\hat{\Sigma} := \frac{1}{n}\sum_{k=1}^{n} x_k x_k^T \in \mathbb{R}^{p\times p}$$

for the following parameters:

(i) $p = 100$, $n = 400$

(ii) $p = 100$, $n = 4000$

(iii) $p = 100$, $n = 40000$

(iv) $p = 500$, $n = 2000$

(v) $p = 1000$, $n = 4000$

in the domain $[0, 4]$. Choose $\frac{1}{10}$ as the width of the bars (or *bins*) in the histogram.

Further experimentation is encouraged.

Exercise 6 $(3+3+3^*+3$ points$)$**.** In this exercise, let $p = 1{,}000$.

(a) Consider n independent copies $x_1, \dots, x_n \in \mathbb{R}^p$ of standard Gaussian random vectors, i.e., $x_i \sim N(0, I_p)$. As in Exercise 5, plot the histogram for the p eigenvalues of the sample covariance matrix and compare this with the Marchenko-Pastur distribution, which is given by the density

$$\psi(t) = \frac{1}{2\pi} \frac{\sqrt{(\gamma_+ - t)(t - \gamma_-)}}{\gamma t} \qquad \text{on the interval } [\gamma_-, \gamma_+],$$

where

$$\gamma = \frac{p}{n}, \qquad \gamma_- = (1 - \sqrt{\gamma})^2, \qquad \gamma_+ = (1 + \sqrt{\gamma})^2.$$

Do this for $\gamma = \frac{1}{4}$, $\gamma = \frac{1}{2}$ and $\gamma = 1$.

Hint: functions that draw histograms often can also automatically rescale the data to mimic a probability density function, which allows to draw actual densities like Marchenko-Pastur on top for easier comparison.

(b) The above is for $\gamma \leq 1$. How does the formula change for $\gamma > 1$? Plot the cases $\gamma = 2$ and $\gamma = 4$ like above.

(c) Bonus: what is the relation between the case γ and the case $\frac{1}{\gamma}$?
Hint: how are the eigenvalues of XX^T and X^TX for a rectangular matrix X related?

(d) Now change in $x_i \sim N(0, I_p)$ the covariance matrix from I_p to Σ by replacing the $(1,1)$-entry 1 with $1 + \beta$ and plot again the histograms from above for all combinations of $\gamma \in \{\frac{1}{4}, \frac{1}{2}, 1\}$ and $\beta \in \{1, 2\}$.

The BBP (Baik, Ben Arous, Péché) transition predicts that (in the limit $n \to \infty$) the eigenvalue $1 + \beta$ of Σ survives as a visible outlier in the eigenvalues of $\hat{\Sigma}$, as long as $\beta \geq \sqrt{\gamma}$, and then sits at the position $(1 + \beta)(1 + \frac{\gamma}{\beta})$. Check whether this is confirmed by your data!

Exercise 7 $(3 + 3$ points$)$**.** Let $x \in \mathbb{R}^p$ be a random vector with probability density function $\psi : \mathbb{R}^p \to \mathbb{R}$, then the expectation of x is

$$E[x] = \int_{\mathbb{R}^p} x\psi(x)\, \mathrm{d}x \in \mathbb{R}^p.$$

and the covariance of x is

$$\Sigma(x) = E[xx^T] - E[x]E[x]^T \in \mathbb{R}^{p\times p}.$$

Let $A \in \mathbb{R}^{p\times p}$ and $b \in \mathbb{R}^p$.

(a) Show that E is linear in the sense that $E[Ax + b] = AE[x] + b$.

(b) Write $\Sigma(Ax + b)$ in terms of $\Sigma(x)$.

Exercise 8 $(3+3+3+2^*$ points$)$**.**

(a) Show that for a standard Gaussian random variable $x \sim N(0, I_p)$ we have $E[x] = 0$ and $\Sigma(x) = I_p$.

(b) Let $y = Ax + b$ be an affine transformation of $x \sim N(\mu, \Sigma)$ by an invertible matrix $A \in \mathbb{R}^{p\times p}$ and an arbitrary vector $b \in \mathbb{R}^p$. Find $\tilde{\mu}$ and $\tilde{\Sigma}$ such that $y \sim N(\tilde{\mu}, \tilde{\Sigma})$.

(c) Conclude that for $x \sim N(\mu, \Sigma)$ we have $E[x] = \mu$ and $\Sigma(x) = \Sigma$.

(d) Bonus: the affine transformation $y = Ax + b$ for $x \sim N(0, I_p)$ also makes sense for arbitrary matrices A that are not necessarily invertible. It seems appropriate to also call this a Gaussian random vector. Are there uniform descriptions which support this point of view?

Exercise 9 $(5+5$ points$)$**.** We will address here concentration estimates for the law of large numbers, and see that control of higher moments allows stronger estimates. Let x_i be a sequence of independent and identically distributed random variables with common mean $\mu = E[x_i]$ and write $X := (x_1, x_2, \dots)$ We put

$$S_n(X) = S_n(x_1, \dots, x_n) := \frac{1}{n}\sum_{i=1}^{n} x_i.$$

(a) Assume that the variance $V[x_i]$ is finite. Prove that we have then the weak law of large numbers, i.e., convergence in probability of S_n to the mean: for any $\varepsilon > 0$

$$P\left\{(x_1, \dots, x_n) : |S_n(X) - \mu| \geq \varepsilon\right\} \xrightarrow{n\to\infty} 0.$$

(b) Assume that the fourth moment of the x_i is finite, i.e. $E[x_i^4] < \infty$ (note that this implies that also all moments of smaller order are finite). Show that we then have

$$\sum_{n=1}^{\infty} P\left\{(x_1, \dots, x_n) : |S_n(X) - \mu| \geq \varepsilon\right\} < \infty.$$

(Note: by the Borel-Cantelli Lemma, this then implies the strong law of large numbers, i.e., $S_n \to \mu$ almost surely.)

One should also note that our assumptions for the weak and strong law of large numbers are far from optimal. Even the existence of the variance is not needed for them, but for proofs of such general versions one needs other tools than our simple consequences of the Chebyshev/Markov inequalities.

10.3. Assignment 3

Exercise 10 (2+5+3+5 points)**.** Fix $p \in [0,1]$. Let $y_1, \ldots, y_n$ be independent Bernoulli random variables with

$$\mathrm{P}\{y_i = 1\} = p, \qquad \mathrm{P}\{y_i = 0\} = 1 - p$$

and consider $y := y_1 + \ldots + y_n$. Let $\delta > 0$.

(a) Show that $E[\exp(\lambda y_i)] \leq \exp(p(\exp(\lambda) - 1))$ holds for every $\lambda > 0$.

(b) Conclude the following classic Chernoff bound:

$$\mathrm{P}\{y \geq (1+\delta)np\} \leq \left(\frac{\exp(\delta)}{(1+\delta)^{1+\delta}}\right)^{np}.$$

Hint: we know from class that

$$\mathrm{P}\{y \geq \alpha\} \leq \exp(-\lambda\alpha) \prod_{i=1}^{n} E[\exp(\lambda y_i)] \quad \textit{for any} \quad \lambda > 0.$$

(c) Assume you are rolling a fair six-sided dice n times. Apply (b) to estimate the probability to roll a six at least 70% of the experiments.

(d) Compare the estimate of (b) with the estimates from the Markov and the Chebyshev Inequalities. Run a simulation of the experiment in (c) to test how tight the predictions of the three bounds are for $n \in \{1, 5, 25, 100\}$ (use 1,000 repetitions of each experiment to get sensible data).

Exercise 11 (6 + 6 points)**.**

(a) Let x be a sub-exponential centred random variable, i.e. a one-dimensional real random variable with mean zero and such that there exists a constant $c > 0$ satisfying

$$E[\exp(\lambda x)] \leq \exp(c^2\lambda^2) \quad \text{for all} \quad |\lambda| \leq \frac{1}{c}.$$

Let $\alpha > 0$. Prove that we then have

$$\mathrm{P}\{x \geq \alpha\} \leq \begin{cases} \exp\left(-\dfrac{\alpha^2}{4c^2}\right), & \text{if } \alpha \leq 2c, \\ \exp\left(-\dfrac{\alpha}{2c}\right), & \text{if } \alpha > 2c. \end{cases}$$

(b) In the proof of Theorem 2.2. we have shown that for a standard Gaussian random vector $x \sim N(0, I_p)$ we have the concentration

$$\mathrm{P}\left\{\left|\|x\|^2 - p\right| \geq \varepsilon\sqrt{p}\right\} \leq 2\exp\left(-\frac{\varepsilon^2}{16}\right).$$

However, this was only for the case where $\varepsilon\sqrt{p} \leq p$, but the proof actually works for all $\varepsilon\sqrt{p} \leq 2p$. Complement this now by a corresponding estimate also for the case of large deviations $\varepsilon\sqrt{p} > 2p$.

Exercise 12 (7 points)**.** Show that every bounded random variable is sub-Gaussian: let x be a real random variable that is bounded, i.e., for some $a, b \in \mathbb{R}$ we have

$$\mathrm{P}\{a \leq x \leq b\} = 1.$$

Assume also that x is centred, i.e., $E[x] = 0$. Then there exists a $c \in \mathbb{R}$ such that we have for all λ

$$E[\exp(\lambda x)] \leq \exp(c\lambda^2).$$

The best constant is actually given by $c = \frac{(b-a)^2}{8}$, but here we are satisfied with any bound.

Hint: for symmetric distributions the situation is easy; in the non-symmetric case one might try to symmetrize the situation by going over, as in our proof of Theorem 3.2., from $E[\exp(\lambda x)]$ *to* $E[\exp(\lambda(x-y))]$*, where* y *is an independent copy of* x*.*

Exercise 13 (2 + 4 points)**.** Consider the following statement: if $h := f \circ g$ is the composition of two convex functions $f, g : \mathbb{R} \to \mathbb{R}$, then h is also convex.

(a) Give a counterexample to show that the statement is not true in general.

(b) Repair the statement by introducing an additional assumption on f and g and prove the statement under this assumption.

10.4. Assignment 4

Besides Wishart matrices the other important random matrix ensemble is given by Wigner matrices. A symmetric matrix $X = X^T \in \mathbb{R}^{n\times n}$ is a *Wigner matrix* if, apart from the symmetry condition, all its entries are independent and identically distributed according to a centred Gaussian distribution (this can be more general, but let us restrict here to Gaussians). In order to have an asymptotic distribution for $n \to \infty$ we have to normalize the entries to have variance $1/n$, i.e., our Wigner matrix has the form

$$X_n = \frac{1}{\sqrt{n}}(x_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n},$$

where

- $x_{ij} \sim N(0,1)$ for all i, j,
- $\{x_{ij} : 1 \leq i \leq j \leq n\}$ is independent, and
- $x_{ji} = x_{ij}$ for all i, j.

Their asymptotic eigenvalue distribution was determined by Wigner in 1955; this was the first and still most fundamental (asymptotic) result about random matrices. In the following two exercises we will address Wigner's semicircle law from a numerical and a theoretical perspective.

Exercise 14 (6 points)**.** Generate histograms of the eigenvalues of an $n \times n$ Wigner matrix, where $n \in \{10, 100, 1000, 2000\}$. Do this in each case for at least two realizations, in order to convince yourself that also in this case we have concentration of the eigenvalues around a deterministic asymptotic distribution. This asymptotic distribution is Wigner's semicircle, which has density

$$\psi(t) = \frac{1}{2\pi}\sqrt{4 - t^2} \quad \text{on} \quad [-2, 2].$$

Compare your histograms with this semicircle distribution.

Exercise 15 $(3+3+3+3$ points$)$**.** We will now determine the form of the semicircle in an analytic way relying on the Stieltjes transform, similar as we did it in class for the Marchenko-Pastur distribution. Denote by S_n the Stieltjes transform of our Wigner matrices,

$$S_n(z) = E\left[\mathrm{tr}\big((X_n - zI_n)^{-1}\big)\right]$$

We will try to derive an equation for the limiting Stieltjes transform (assuming that it exists) $S(z) := \lim_{n\to\infty} S_n(z)$, by writing X_n in the form

$$X_n = \frac{1}{\sqrt{n}} \begin{pmatrix} x_{11} & x^T \\ x & Y \end{pmatrix},$$

where $Y \in \mathbb{R}^{(n-1)\times(n-1)}$ contains the last $n-1$ rows and columns of X_n and $x \in \mathbb{R}^{n-1}$ is the vector $x = (x_{21}, \ldots, x_{n1})^T$. The replacement of the Sherman-Morrison formula in this case is given by Schur's complement formula, which says that for a decomposition of $M \in \mathbb{R}^{n\times n}$ in the form

$$M = \begin{pmatrix} a & v^T \\ v & D \end{pmatrix} \qquad D \in \mathbb{R}^{(n-1)\times(n-1)}, \quad v \in \mathbb{R}^{n-1}, \quad a \in \mathbb{R},$$

the inverse of M exists if D is invertible and $a - v^T D^{-1} v \neq 0$, and in this case the $(1,1)$-entry of M^{-1} is given by

$$[M^{-1}]_{11} = \frac{1}{a - v^T D^{-1} v}.$$

(a) Prove the formula above for the $(1,1)$-entry of M^{-1}.

Hint: it might be good to also find formulas for the other entries of M^{-1}*.*

(b) By applying the formula above to $M = X_n - zI_n$ show that

$$[M^{-1}]_{11} \approx \frac{1}{-z - S_n(z)}.$$

(c) By doing the same with splitting off the k-th row and column in M, show that the Stieltjes transform of our Wigner matrix satisfies in the limit $n \to \infty$ the equation

$$S(z) = \frac{1}{-z - S(z)}.$$

(d) Solve the equation for $S(z)$ and derive from this, by Stieltjes inversion formula, the formula for the density of the semicircle.

Exercise 16 $(4 + 4$ points$)$**.** Let $Q \in \mathbb{R}^{p\times p}$ and $U, V \in \mathbb{R}^{p\times n}$ be deterministic matrices such that both Q and $Q + UV^T$ are invertible.

(a) Show that $I_n + V^T Q^{-1} U$ is also invertible.

(b) Show that $(Q + UV^T)^{-1} = Q^{-1} - Q^{-1}U(I_n + V^T Q^{-1} U)^{-1} V^T Q^{-1}$.

Exercise 17 $(3+5+6$ points$)$**.** Let $p, n \in \mathbb{N}$ with p even and $\gamma := \frac{p}{n}$. In Assignment 2, Exercise 1 we looked on Wishart matrices where Σ is not the identity matrix, but has one half of its eigenvalues equal to $t_1 = 1$ and the other half equal to $t_2 = 2$. Let us now consider such a situation with arbitrary $t_1, t_2 \in \mathbb{R}$, i.e., our data matrix is of the form

$$\begin{pmatrix} X \\ Y \end{pmatrix} \in \mathbb{R}^{p \times n},$$

where

- the columns of $X \in \mathbb{R}^{\frac{p}{2} \times n}$ are $N(0, t_1 I_{\frac{p}{2}})$-distributed,
- the columns of $Y \in \mathbb{R}^{\frac{p}{2} \times n}$ are $N(0, t_2 I_{\frac{p}{2}})$ distributed, and
- all these column vectors are independent.

Thus the Wishart matrix is of the form

$$\hat{\Sigma} = \frac{1}{n} \begin{pmatrix} X \\ Y \end{pmatrix} \begin{pmatrix} X^T & Y^T \end{pmatrix} = \frac{1}{n} \begin{pmatrix} XX^T & XY^T \\ YX^T & YY^T \end{pmatrix} \in \mathbb{R}^{p \times p}.$$

(a) Recall that, apart from some zeros, $\hat{\Sigma}$ has the same eigenvalues as

$$\check{\Sigma} = \frac{1}{n} \begin{pmatrix} X^T & Y^T \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \frac{1}{n}(X^T X + Y^T Y) \in \mathbb{R}^{n \times n}.$$

Give, for $p \leq n$, the relation between the Stieltjes transforms of $\hat{\Sigma}$ and of $\check{\Sigma}$.

(b) By following the same ideas as in class for the determination of the Marchenko-Pastur law, show that the limit $\check{S}(z)$ of the Stieltjes transform for this $\check{\Sigma}$ satisfies

$$1 + z\check{S}(z) = \frac{\gamma}{2} \frac{t_1 \check{S}(z)}{1 + t_1 \check{S}(z)} + \frac{\gamma}{2} \frac{t_2 \check{S}(z)}{1 + t_2 \check{S}(z)}.$$

(c) If we put $S(z) := \check{S}(z)/\gamma$, then this satisfies the equation

$$S(z) = -\frac{1}{\gamma z} + \frac{1}{2z} \frac{t_1 \gamma S(z)}{1 + t_1 \gamma S(z)} + \frac{1}{2z} \frac{t_2 \gamma S(z)}{1 + t_2 \gamma S(z)}.$$

This $S(z)$ gives us then the density ψ of the asymptotic eigenvalue distribution of $\hat{\Sigma}$ via the Stieljes inversion formula

$$\psi(t) = \lim_{\varepsilon \to 0} \frac{1}{\pi} \operatorname{Im}\big(S(t + i\varepsilon)\big).$$

Let $t_1 = 3$, $t_2 = 15$ and $\gamma = \frac{1}{5}$. In the same diagram, plot the following:

(i) The graph of ψ, obtained by numerically applying a fixed-point iteration to calculate $\psi(t) \approx \frac{1}{\pi}\operatorname{Im}\big(S(t+i\varepsilon)\big)$ for $\varepsilon = 0.01$.[7] As a starting point, any point in the complex upper half-plane will work and result in a solution in the complex upper half-plane. Use enough values for t to get a smooth curve!

(Note that there will be an additional pole at 0, coming from the difference between $\hat{\Sigma}$ and $\check{\Sigma}$.)

(ii) A histogram of the eigenvalues of a numerical simulation of the corresponding Wishart matrix with $p = 500$, normalized to fit the density.

[7] Although the equation for $S(z)$ is a cubic one and might thus be solved explicitly, it is easier to solve the equation numerically as a fixed-point equation (especially in more general situations).

10.5. Assignment 5

Recall that the ReLU function is defined as $\mathrm{ReLU}(t) = \max(0, t)$.

Exercise 18 (10 points)**.** We now investigate a one-layer perceptron with random features and n parameters: given an input $x \in \mathbb{R}$, the neural network computes $y = w\sigma(ax + b)$, where

- $a \sim N(0, I_n)$ is the *weight* and $b \sim N(0, I_n)$ is the *bias*,
- $\sigma : \mathbb{R} \to \mathbb{R}$ is the (non-linear) activation, applied component-wise,
- $w \in \mathbb{R}^{1\times n}$ is the linear regression of the training data.

Consider the following eleven (training) data points:[8]

x_k	-5	-4	-3	-2	-1	0	1	2	3	4	5
y_k	-3	-3	-4	1	-0.2	0.1	2	1.8	1.9	-0.2	2

For each $n \in \{5, 10, 11, 30, 300, 1000\}$ and each $\sigma \in \{\mathrm{ReLU}, \sin\}$ do the following:

(a) For two d-dimensional standard Gaussian vectors $a, b \sim N(0, I_n)$, compute the feature matrix

$$F = \begin{pmatrix} f_1 & \dots & f_{11} \end{pmatrix} \in \mathbb{R}^{n\times 11}, \quad \text{where} \quad f_k = \sigma(a \cdot x_k + b).$$

(b) Perform linear regression on the so-obtained features in order to fit the data given above: $w = YF^T(FF^T)^+$, where $Y = \begin{pmatrix} y_1 & \dots & y_{11} \end{pmatrix} \in \mathbb{R}^{1\times 11}$ and A^+ is the pseudo-inverse of A.

(c) Plot the output of your neural network on the grid from -5 to 5 with step size 0.1. For comparison, also plot the original data points. It suffices to hand in the plots, no need to print out all the intermediate data.

Compare the plots and describe what you see. This is an instance of the so-called double-descent!

[8] Copy-friendly version of the y_k: `[-3, -3, -4, 1, -0.2, 0.1, 2, 1.8, 1.9, -0.2, 2]`

Exercise 19 (7 points)**.** Consider the entries x_{ij} of our matrix $X = (x_{ij}) \in \mathbb{R}^{p\times n}$ as formal variables. For fixed $z \in \mathbb{C}$, we put

$$R = R(z) = \left(\frac{1}{n}XX^T - zI_p\right)^{-1} \in \mathbb{R}^{p\times p}.$$

Show that we have

$$\left[\frac{\partial R}{\partial x_{ij}}\right]_{kl} = -\frac{1}{n}\Big(R_{ki}[X^T R]_{jl} + [RX]_{kj}R_{il}\Big).$$

Exercise 20 ($(3+4)+(2+3)$ points)**.** For a function $\sigma : \mathbb{R} \to \mathbb{R}$ we denote

$$\theta_1(\sigma) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \sigma(t)^2 \exp\left(-\frac{t^2}{2}\right)\,\mathrm{d}t$$

and

$$\theta_2(\sigma) := \left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \sigma'(t) \exp\left(-\frac{t^2}{2}\right)\,\mathrm{d}t\right)^2 = \left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} t\sigma(t) \exp\left(-\frac{t^2}{2}\right)\,\mathrm{d}t\right)^2.$$

(a) Let $\sigma : \mathbb{R} \to \mathbb{R}$ be such that $\theta_1(\sigma)$ and $\theta_2(\sigma)$ are finite.

(i) Show that $\theta_2(\sigma) \le \theta_1(\sigma)$.

(ii) Show that $\theta_2(\sigma) = \theta_1(\sigma)$ if and only if σ is a linear function, i.e., $\sigma(t) = \beta t$ for some $\beta \in \mathbb{R}$.

(b) Let $\alpha \in \mathbb{R}$ be a constant and consider the shifted ReLU function

$$\sigma(t) = \mathrm{ReLU}(t) - \alpha.$$

(i) Determine α such that

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \sigma(t) \exp\left(-\frac{t^2}{2}\right)\,\mathrm{d}t = 0.$$

(ii) Determine for this σ the quantities $\theta_1(\sigma)$ and $\theta_2(\sigma)$.

Exercise 21 ((4 + 4) + 3 points)**.** Like in class, consider standard Gaussian random matrices $X \in \mathbb{R}^{p\times n}$ and $W \in \mathbb{R}^{p\times p}$ together with a non-linearity $\sigma : \mathbb{R} \to \mathbb{R}$. Let

$$F := \sigma\left(\frac{1}{\sqrt{p}}WX\right) \in \mathbb{R}^{p\times p} \quad \text{and} \quad M := \frac{1}{n}FF^T \in \mathbb{R}^{p\times p}.$$

(a) Consider $\sigma_1(t) = t^2 - 1$ and $\sigma_2(t) = t^3 - 3t$. For each $\sigma \in \{\sigma_1, \sigma_2\}$ do the following:

(i) Compute $\theta_1(\sigma)$ and show that $\theta_2(\sigma) = 0$.

(ii) For $p = 2000$ and each $\gamma \in \left\{1, \frac{1}{2}, \frac{1}{4}\right\}$, draw a diagram including a histogram of the eigenvalues of M and the corresponding Marchenko-Pastur distribution. Re-scale σ such that the distribution matches the histogram.

(b) From class we know that in general, F behaves like

$$\tilde{F} = \frac{\sqrt{\theta_2}}{\sqrt{p}}WX + \sqrt{\theta_1 - \theta_2}Z$$

for (independent) standard Gaussian matrices $W \in \mathbb{R}^{p\times p}$ and $X, Z \in \mathbb{R}^{p\times n}$. For $\sigma(t) = \mathrm{ReLU}(t) - \alpha$ from the previous exercise, compare a histogram of the eigenvalues of M with a histogram of the eigenvalues of $\tilde{M} := \frac{1}{n}\tilde{F}\tilde{F}^T$. Again, use $p = 2000$ and consider each $\gamma \in \left\{1, \frac{1}{2}, \frac{1}{4}\right\}$.

10.6. Assignment 6

Exercise 22 (5 + 5 points)**.**

(a) Let t be Poisson-distributed with rate $\lambda > 0$, i.e. t is a discrete random variable supported on $\mathbb{N}_0$ with distribution

$$\mathrm{P}(t = k) = \frac{\lambda^k \exp(-\lambda)}{k!}.$$

Compute the cumulants of t using their definition as coefficients in the logarithm of the characteristic function.

(b) Let t be χ^2-distributed with $k \in \mathbb{N}$ degrees of freedom, i.e. $t = \sum_{j=1}^{k} x_j^2$, where the $x_j \sim N(0, 1)$ are independent. Compute the cumulants of t using Theorem 7.13.

Exercise 23 (2 + 4 + 4 points)**.** Let $\{\alpha_n\}_{n\in\mathbb{N}}$ and $\{\kappa_n\}_{n\in\mathbb{N}}$ be two sequences that satisfy the relation

$$\alpha_n = \sum_{\pi\in\mathcal{P}(n)} \kappa_\pi,$$

where $\kappa_\pi = \kappa_1^{r_1} \cdot \ldots \cdot \kappa_n^{r_n}$ and r_j is the number of blocks of π of size j. We want to show that, as formal power series,

$$\log\left(1 + \sum_{n=1}^{\infty} \alpha_n \frac{z^n}{n!}\right) = \sum_{n=1}^{\infty} \kappa_n \frac{z^n}{n!}. \tag{6}$$

(a) Show that by differentiating both sides of (6) it suffices to prove

$$\sum_{n=0}^{\infty} \alpha_{n+1} \frac{z^n}{n!} = \left(1 + \sum_{n=1}^{\infty} \alpha_n \frac{z^n}{n!}\right) \sum_{n=0}^{\infty} \kappa_{n+1} \frac{z^n}{n!}. \tag{7}$$

(b) By grouping the terms in $\sum_{\pi\in\mathcal{P}(n)} \kappa_\pi$ according to the size of the block containing 1, show that

$$\alpha_n = \sum_{\pi\in\mathcal{P}(n)} \kappa_\pi = \sum_{m=0}^{n-1} \binom{n-1}{m} \kappa_{m+1} \alpha_{n-m-1}.$$

(c) Use the result of (b) to prove (7).

Exercise 24 $(5+5+5+5$ points$)$**.** We consider, for $p=1$, our 1 hidden layer neural network of width m,

$$f_m(x) = a^T \sigma(bx+c).$$

Initially, $\tilde{a}, b, c \in \mathbb{R}^m$ are independent standard Gaussian random vectors and $a := \frac{\tilde{a}}{\sqrt{m}}$. (Note that we include here also a bias c in the argument of σ). We want to use this to learn the function $g : \mathbb{R} \to \mathbb{R}$ given by

$$g(x) = \sqrt{|x|} + \sin(10x),$$

restricted to the interval $[-1,1]$.

Choose randomly 15 data points x_i, drawn from the uniform distribution on the interval $[-1,1]$, and let $y_i := g(x_i)$. From this data we try to recover g: Use gradient descent to train the parameters $\{a, b\}$ (we don't train the bias c, but keep this fixed) with respect to the loss function

$$\mathcal{L}(a,b) = \frac{1}{2}\sum_{i=1}^{15}\big(y_i - f_m(x_i)\big)^2,$$

for varying widths m. It is actually advisable to use *stochastic gradient descent*; that is, in each step one uses only the gradient of $\big(y_i - f_m(x_i)\big)^2$, with respect to a and to b, for a randomly chosen i. Train until the loss function is less than 0.01 (in the case $m > 15$) or until it does not decrease any more (in the case $m \leq 15$). Plot then the trained function $f_m(x)$ against the target function $g(x)$ for 2000 points x sampled evenly from the interval $[-1,1]$, for $m \in \{1,2,5,10,15,30,100,500\}$. Show also the 15 data points $(x_i, g(x_i))$ in this plot. As learning rate you might choose any $\eta \in (0.001, 0.01)$.

(a) Do this for $\sigma(x) = \sin(8x)$.

(b) Do this for $\sigma = \mathrm{ReLU}$.

(c) Check in those cases also what happens if you switch off the bias (i.e., put $c = 0$).

(d) Explain why it is a bad idea to switch off the bias in the case of $\sigma(x) = \sin(8x)$. Explain why it is an even worse idea to do this in the case of $\sigma = \mathrm{ReLU}$.

References

[AP20] Ben Adlam and Jeffrey Pennington, *The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization*, International Conference on Machine Learning, PMLR, 2020, pp. 74–84.

[BBAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché, *Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices*, Annals of Probability (2005), 1643–1697.

[Bel21] Mikhail Belkin, *Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation*, Acta Numerica **30** (2021), 203–248.

[BP21] Lucas Benigni and Sandrine Péché, *Eigenvalue distribution of some nonlinear models of random matrices*, Electronic Journal of Probability **26** (2021), 1–37.

[BS06] Jinho Baik and Jack W Silverstein, *Eigenvalues of large sample covariance matrices of spiked population models*, Journal of multivariate analysis **97** (2006), no. 6, 1382–1408.

[CL22] Romain Couillet and Zhenyu Liao, *Random matrix methods for machine learning*, Cambridge University Press, 2022.

[Han21] Boris Hanin, *Random neural networks in the infinite width limit as gaussian processes*, arXiv preprint arXiv:2107.01562 (2021).

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, Advances in neural information processing systems **31** (2018).

[MM22] Song Mei and Andrea Montanari, *The generalization error of random features regression: Precise asymptotics and the double descent curve*, Communications on Pure and Applied Mathematics **75** (2022), no. 4, 667–766.

[MP67] VA Marchenko and LA Pastur, *The distribution of eigenvalues in certain sets of random matrices math*, Math. USSR-Sbornik **1** (1967), 457–483.

[MS17] James A Mingo and Roland Speicher, *Free probability and random matrices*, Fields Institute Monographs (2017).

[Pis06] Gilles Pisier, *Probabilistic methods in the geometry of banach spaces*, Probability and Analysis: Lectures given at the 1st 1985 Session of the Centro Internazionale Matematico Estivo (CIME) held at Varenna (Como), Italy May 31–June 8, 1985, Springer, 2006, pp. 167–241.

[PS21] Vanessa Piccolo and Dominik Schröder, *Analysis of one-hidden-layer neural networks via the resolvent method*, Advances in Neural Information Processing Systems **34** (2021), 5225–5235.

[PW17] Jeffrey Pennington and Pratik Worah, *Nonlinear random matrix theory for deep learning*, Advances in neural information processing systems **30** (2017).

[RYH22] Daniel A Roberts, Sho Yaida, and Boris Hanin, *The principles of deep learning theory*, Cambridge University Press Cambridge, MA, USA, 2022.

[Tal95] Michel Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques **81** (1995), 73–205.

[Ver18] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018.

[Wai19] Martin J Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48, Cambridge university press, 2019.

[Yan19] Greg Yang, *Wide feedforward or recurrent neural networks of any architecture are gaussian processes*, Advances in Neural Information Processing Systems **32** (2019).