Saarland University
Faculty of Mathematics and Computer Science
Department of Mathematics
Prof. Dr. Roland Speicher
Dr. Johannes Hoffmann

Exercises for the lecture
High Dimensional Analysis: Random Matrices and Machine Learning
Summer term 2023
Sheet 1
Hand-in: Friday, 28.04.2023, 22:00 Uhr via CMS

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)
- $\|x\|_\infty := \max\left\{|t_k| : 1 \leq k \leq p\right\}$ (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 $x_1, \dots, x_n \in \mathbb{R}^p$ are n independent copies of our p-dimensional random vectors x as above, and $X = [x_1 x_2 \dots x_n] \in \mathbb{R}^{p\times n}$ 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 $f(X) = f(x_1, \dots, x_n)$ 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 $X = [x_1 \dots x_n]$. 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.

Saarland University
Faculty of Mathematics and Computer Science
Department of Mathematics
Prof. Dr. Roland Speicher
Dr. Johannes Hoffmann

Exercises for the lecture
High Dimensional Analysis: Random Matrices and Machine Learning

Summer term 2023

Sheet 2

Hand-in: Friday, 12.05.2023, 22:00 Uhr via CMS

(Parts of) exercises that start with the indicator "Bonus" consider advanced or more philosophical questions; they come with extra points, but you cannot get more than 40 points per exercise sheet.

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 1 (6 points). Consider n independent copies $x_1, \dots, x_n \in \mathbb{R}^p$ 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 2 ($3+3+3^*+3$ 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 1, 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 $\gamma \in \left\{\frac{1}{4}, \frac{1}{2}, 1\right\}$ 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 3 ($3+3$ 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 4 ($3 + 3 + 3 + 2^*$ 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 5 ($5 + 5$ 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.

Saarland University
Faculty of Mathematics and Computer Science
Department of Mathematics
Prof. Dr. Roland Speicher
Dr. Johannes Hoffmann

Exercises for the lecture
High Dimensional Analysis: Random Matrices and Machine Learning

Summer term 2023

Sheet 3

Hand-in: Friday, 26.05.2023, 22:00 Uhr via CMS

Exercise 1 (2+5+3+5 points). Fix $p \in [0,1]$. Let $y_1, \dots, y_n$ 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).

please turn over

Exercise 2 (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

$$E[\exp(\lambda x)] \leq \exp(c^2\lambda^2) \quad \text{for all} \quad |\lambda| \leq \frac{1}{c}.$$

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 3 (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 4 (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.

Saarland University
Faculty of Mathematics and Computer Science
Department of Mathematics
Prof. Dr. Roland Speicher
Dr. Johannes Hoffmann

Exercises for the lecture
High Dimensional Analysis: Random Matrices and Machine Learning
Summer term 2023
Sheet 4
Hand-in: Friday, 09.06.2023, 22:00 Uhr via CMS

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 1 (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.

please turn over

Exercise 2 (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,

$$S_n(z) = E\left[\operatorname{tr}\big((X_n - zI_n)^{-1}\big)\right]$$

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 3 (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 $(Q + UV^T)^{-1} = Q^{-1} - Q^{-1}U(I_n + V^T Q^{-1} U)^{-1} V^T Q^{-1}$.

please turn over

Exercise 4 (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 $\varepsilon = 0.01$.[1] 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.

[1] 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).

Saarland University
Faculty of Mathematics and Computer Science
Department of Mathematics
Prof. Dr. Roland Speicher
Dr. Johannes Hoffmann

Exercises for the lecture
High Dimensional Analysis: Random Matrices and Machine Learning
Summer term 2023
Sheet 5
Hand-in: Friday, 23.06.2023, 22:00 Uhr via CMS

Recall that the ReLU function is defined as $\mathrm{ReLU}(t) = \max(0, t)$.

Exercise 1 (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:[1]

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!

please turn over

[1] Copy-friendly version of the y_k: `[-3, -3, -4, 1, -0.2, 0.1, 2, 1.8, 1.9, -0.2, 2]`

Exercise 2 (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 3 ((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 $\theta_2(\sigma)$.

please turn over

Exercise 4 ((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 $\gamma \in \left\{1, \frac{1}{2}, \frac{1}{4}\right\}$.

Saarland University
Faculty of Mathematics and Computer Science
Department of Mathematics
Prof. Dr. Roland Speicher
Dr. Johannes Hoffmann

Exercises for the lecture
High Dimensional Analysis: Random Matrices and Machine Learning
Summer term 2023
Sheet 6
Hand-in: Friday, 14.07.2023, 22:00 Uhr via CMS

Exercise 1 (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 2 (10 points)**.** Let $\{\alpha_n\}_{n\in\mathbb{N}}$ and $\{\kappa_n\}_{n\in\mathbb{N}}$ be two sequences that satisfy the relation

$$\alpha_n = \sum_{\pi\in\mathcal{P}(n)} k_\pi,$$

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,

$$\log\left(1 + \sum_{n=1}^{\infty} \alpha_n \frac{z^n}{n!}\right) = \sum_{n=1}^{\infty} \kappa_n \frac{z^n}{n!}. \tag{1}$$

(a) Show that by differentiating both sides of (1) it suffices to prove

$$\sum_{n=0}^{\infty} \alpha_{n+1} \frac{z^n}{n!} = \left(1 + \sum_{n=1}^{\infty} \alpha_n \frac{z^n}{n!}\right) \sum_{n=0}^{\infty} \kappa_{n+1} \frac{z^n}{n!}. \tag{2}$$

(b) By grouping the terms in $\sum_{\pi\in\mathcal{P}(n)} k_\pi$ according to the size of the block containing 1, show that

$$\alpha_n = \sum_{\pi\in\mathcal{P}(n)} k_\pi = \sum_{m=0}^{n-1} \binom{n-1}{m} \kappa_{m+1} \alpha_{n-m-1}.$$

(c) Use the result of (b) to prove (2).

please turn over

Exercise 3 (5 + 5 + 5 + 5 points). We consider, for $p = 1$, our 1 hidden layer neural network of width m,

$$f_m(x) = \frac{1}{\sqrt{m}} a^T \sigma(bx + c),$$

where a, b and c are independent standard Gaussian random vectors in $\mathbb{R}^m$. (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 $\sigma = \text{ReLU}$.

(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 $\sigma = \text{ReLU}$.