

Chapter 1

Mathematical preliminaries

Chapter summary

- Linear algebra: a bag of tricks to avoid lengthy and faulty computations.
- Concentration inequalities: for n independent random variables, the deviation between the empirical average and the expectation is of order $O(1/\sqrt{n})$. What is in the big O , and how does it depend explicitly on problem parameters?

The mathematical analysis and design of machine learning algorithms require specialized tools beyond classic linear algebra, differential calculus, and probability. In this chapter, I will review these non-elementary mathematical tools used throughout the book: first, linear algebra tricks, then concentration inequalities. The chapter can be safely skipped since relevant results will be referenced when needed.

1.1 Linear algebra and differentiable calculus

This section reviews basic linear algebra and differential calculus results that will be used throughout the book. Using these may usually greatly simplify computations. Matrix notations will be used as much as possible.

1.1.1 Minimization of quadratic forms

Given a positive definite (and hence invertible) symmetric matrix $A \in \mathbb{R}^{n \times n}$ and a vector $b \in \mathbb{R}^n$, the minimization of quadratic forms with linear terms can be done in closed form as:

$$\inf_{x \in \mathbb{R}^n} \frac{1}{2} x^\top A x - b^\top x = -\frac{1}{2} b^\top A^{-1} b,$$

with minimizer $x_* = A^{-1}b$ obtained by zeroing the gradient $f'(x) = Ax - b$ of the function $f(x) = \frac{1}{2}x^\top Ax - b^\top x$. Moreover, we have:

$$\frac{1}{2}x^\top Ax - b^\top x = \frac{1}{2}(x - x_*)^\top A(x - x_*) - \frac{1}{2}b^\top A^{-1}b.$$

If A was not invertible (simply positive semi-definite) and b was not in the column space of A , then the infimum would be $-\infty$.

Note that this result is often used in various forms, such as

$$b^\top x \leq \frac{1}{2}b^\top A^{-1}b + \frac{1}{2}x^\top Ax \text{ with equality if and only if } b = Ax.$$

This form is exactly the Fenchel-Young inequality¹ for quadratic forms (see Chapter 5), and is often used in one dimension in the form $ab \leq \frac{a^2}{2\eta} + \frac{\eta b^2}{2}$, for any $\eta \geq 0$ (and equality if and only if $\eta = a/b$).

1.1.2 Inverting a 2×2 matrix

Solving small systems happens frequently, as well as inverting small matrices. This can be easily done in two dimensions. Let $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be a 2×2 matrix. If $ad - bc \neq 0$, then we may invert it as follows

$$M^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

This can be checked by multiplying the two matrices or using Cramer's rule,² and can be generalized to matrices defined by blocks, as we present next.

1.1.3 Inverting matrices defined by blocks, matrix inversion lemma

The example above may be generalized to matrices of the form $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$, with blocks of consistent sizes (note that A and D have to be square matrices). The inverse of M may be obtained by applying directly Gaussian elimination³ done in block form. Given the two matrices $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ and $N = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$, we may linearly combine lines (with the same coefficients for the two matrices). Once M has been transformed into the identity matrix, N has been transformed to the inverse of M .

We make the simplifying assumption that A is invertible; we use the notation $(M/A) = D - CA^{-1}B$ for the Schur complement of the block A and also assume that (M/A) is

¹See https://en.wikipedia.org/wiki/Convex_conjugate.

²See https://en.wikipedia.org/wiki/Cramer's_rule.

³See https://en.wikipedia.org/wiki/Gaussian_elimination.

invertible. We thus get by Gaussian elimination, referring to L_i , $i = 1, 2$, as the two lines of blocks, so that for the first matrix $M = \begin{pmatrix} L_1 \\ L_2 \end{pmatrix}$:

$$\begin{aligned}
\text{Original matrices:} & \quad \begin{pmatrix} A & B \\ C & D \end{pmatrix} & \quad \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \\
L_2 \leftarrow L_2 - CA^{-1}L_1 : & \quad \begin{pmatrix} A & B \\ 0 & (M/A) \end{pmatrix} & \quad \begin{pmatrix} I & 0 \\ -CA^{-1} & I \end{pmatrix} \\
L_2 \leftarrow (M/A)^{-1}L_2 : & \quad \begin{pmatrix} A & B \\ 0 & I \end{pmatrix} & \quad \begin{pmatrix} I & 0 \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix} \\
L_1 \leftarrow L_1 - BL_2 : & \quad \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} & \quad \begin{pmatrix} I + B(M/A)^{-1}CA^{-1} & -B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix} \\
L_1 \leftarrow A^{-1}L_1 : & \quad \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} & \quad \begin{pmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix}.
\end{aligned}$$

This shows that

$$M^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix}. \quad (1.1)$$

Moreover, by doing the same operations but by putting to zero first the upper-right block, and assuming D and $(M/D) = A - BD^{-1}C$ are invertible, we obtain:

$$M^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (M/D)^{-1} & -(M/D)^{-1}BD^{-1} \\ -D^{-1}C(M/D)^{-1} & D^{-1} + D^{-1}C(M/D)^{-1}BD^{-1} \end{pmatrix}. \quad (1.2)$$

By identifying the upper-left and lower-right blocks in Eq. (1.1) and Eq. (1.2), we obtain the identities (sometimes referred to as Woodbury matrix identities, or the *matrix inversion lemma*):

$$\begin{aligned}
(A - BD^{-1}C)^{-1} &= A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} \\
(D - CA^{-1}B)^{-1} &= D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1}.
\end{aligned}$$

Another classical formulation is:

$$(A - BD^{-1}C)^{-1}B = A^{-1}B(D - CA^{-1}B)^{-1}D.$$

These are particularly interesting when the blocks A and D have very different sizes, as the inverse of a large matrix may be obtained from the inverse of a small matrix.

The lemma is often applied when $C = B^\top$, $A = I$ and $D = -I$, which leads to

$$(I + BB^\top)^{-1} = I - B(I + B^\top B)^{-1}B^\top,$$

and, once right-multiplied by B , this leads to the compact formula (which is easier to rederive and remember):

$$(I + BB^\top)^{-1}B = B(I + B^\top B)^{-1}.$$

These equalities are commonly used both for theoretical and algorithmic purposes.

Exercise 1.1 (♦) Show that we can “diagonalize” by blocks the matrices M and M^{-1} as:

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & (M/A) \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}$$

$$M^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} I & -A^{-1}B \\ 0 & I \end{pmatrix} \begin{pmatrix} A^{-1} & 0 \\ 0 & (M/A)^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -CA^{-1} & I \end{pmatrix}.$$

Conditional covariance matrices for Gaussian vectors (♦). The identities above can be used to compute conditional mean vectors and covariance matrices for Gaussian vectors (in this book, we will use the denominations “normal” and “Gaussian” interchangeably). If we have a Gaussian vector $\begin{pmatrix} x \\ y \end{pmatrix}$ with $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$, with mean vector defined by block as $\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$, and covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \succcurlyeq 0$ (defined with blocks of appropriate sizes), then the joint density $p(x, y)$ of (x, y) is proportional to

$$\exp \left(-\frac{1}{2} \begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix}^\top \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}^{-1} \begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix} \right).$$

By writing it as the product of a function of x and of a function of (x, y) , we can get that x is Gaussian with mean μ_x and covariance matrix Σ_x , and that given x , y is Gaussian with mean $\mu_{y|x} = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x)$ and covariance matrix $\Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$.

Exercise 1.2 (♦) Prove the identities $\mu_{y|x} = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x)$ and covariance matrix $\Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$.

1.1.4 Eigenvalue and singular value decomposition

In this book, we will often use eigenvalue decompositions of symmetric matrices. If $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix, there exists an orthogonal matrix $U \in \mathbb{R}^{n \times n}$ (that is, such that $U^\top U = UU^\top = I$), and a vector $\lambda \in \mathbb{R}^n$ of eigenvalues, such that $A = U \text{Diag}(\lambda) U^\top$. If $u_i \in \mathbb{R}^n$ denotes the i -th column of U , then we have $A = \sum_{i=1}^n \lambda_i u_i u_i^\top$, and $Au_i = \lambda_i u_i$. A symmetric matrix is said to be positive semi-definite if and only if all its eigenvalues are non-negative.

Given a rectangular matrix $X \in \mathbb{R}^{n \times d}$, such that $n \geq d$, there exists an orthogonal matrix $V \in \mathbb{R}^{d \times d}$ (that is, such that $V^\top V = VV^\top = I$), a matrix $U \in \mathbb{R}^{n \times d}$ with orthonormal columns (that is, such that $U^\top U = I$), a vector $s \in \mathbb{R}_+^d$ of singular values, such that $X = U \text{Diag}(s) V^\top$; this is often called the “economy-size” singular value decomposition (SVD) of the matrix X . If $u_i \in \mathbb{R}^n$ and $v_i \in \mathbb{R}^d$ denote the i -th columns of U and V , then we have $X = \sum_{i=1}^d s_i u_i v_i^\top$, and $Xv_i = s_i u_i$, $X^\top u_i = s_i v_i$.

There are several ways of relating eigenvalues and singular values. For example, if s_i is a singular value of X , then s_i^2 is an eigenvalue of XX^\top and $X^\top X$. Moreover, the eigenvalues of the matrix $\begin{pmatrix} 0 & X \\ X^\top & 0 \end{pmatrix}$ are zero, the singular values of X , and their opposites.

For further properties of eigenvalues and singular values, see [Golub and Loan \(1996\)](#), [Stewart and Sun \(1990\)](#) and [Bhatia \(2013\)](#).

Exercise 1.3 Express the eigenvectors of XX^\top and $X^\top X$ using the singular vectors of X .

Exercise 1.4 Express the eigenvectors of $\begin{pmatrix} 0 & X \\ X^\top & 0 \end{pmatrix}$ using the singular vectors of X .

1.1.5 Differential calculus

Throughout the book, we will compute gradients and Hessians of functions in almost all cases in matrix notations. Here are some classic examples:

- Quadratic forms: assuming $A = A^\top$, with $F(\theta) = \frac{1}{2}\theta^\top A\theta - b^\top \theta$, $F'(\theta) = A\theta - b$, $F''(\theta) = A$. If A is not symmetric, then we have $F'(\theta) = \frac{1}{2}(A + A^\top)\theta$ and $F''(\theta) = \frac{1}{2}(A + A^\top)$.
- Least-squares with $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$: $F(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$. Then $F'(\theta) = \frac{1}{n}X^\top(X\theta - y)$ and $F''(\theta) = \frac{1}{n}X^\top X$.

Exercise 1.5 Show that for the logistic regression objective function defined as $F(\theta) = \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i(X\theta)_i))$ with $X \in \mathbb{R}^{n \times d}$ and $y \in \{-1, 1\}^n$, then $F'(\theta) = \frac{1}{n}X^\top g$, where $g \in \mathbb{R}^n$ is defined as $g_i = -y_i \sigma(-y_i(X\theta)_i)$, with $\sigma(u) = (1 + e^{-u})^{-1}$ is the sigmoid function. Show that the Hessian is $\frac{1}{n}X^\top \text{Diag}(h)X$, with $h \in \mathbb{R}^n$ defined as $h_i = \sigma(y_i(X\theta)_i)\sigma(-y_i(X\theta)_i)$.

1.2 Concentration inequalities

All results presented in this textbook rely on the simple probabilistic assumption that data are independently and identically distributed (i.i.d.). The primary tool is then to relate empirical averages to expectations.

The key (very classical) insight behind probabilistic inequalities used in machine learning is that when you have n independent zero-mean random variables, the natural “magnitude” of their average is $1/\sqrt{n}$ times smaller than their average magnitude. The simplest instance of this phenomenon is that if $Z_1, \dots, Z_n \in \mathbb{R}$ are independent and identically distributed with variance $\sigma^2 = \mathbb{E}(Z - \mathbb{E}[Z])^2$, then the variance of the sum is the sum of variances, and

$$\text{var}\left(\frac{1}{n} \sum_{i=1}^n Z_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(Z_i) = \frac{\sigma^2}{n}.$$



Be careful with error measures or magnitudes: some are squared, some are not. Therefore, the $1/\sqrt{n}$ becomes $1/n$ after taking the square (this is trivial but typically leads to confusion).

The equality above can be interpreted as

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z]\right)^2\right] = \frac{\sigma^2}{n}, \quad (1.3)$$

which provides the simplest proof of the law of large numbers when variances exist and also highlights the convergence in squared mean of the random variable $\frac{1}{n}\sum_{i=1}^n Z_i$ to the constant $\mathbb{E}[Z]$.

From moments to deviation bounds. Given an (in)equality on the moments of a random variable, deviation bounds can be derived. Markov's inequality (see proof in Exercise 1.6 below) states that

$$\mathbb{P}(Y \geq \varepsilon) \leq \frac{1}{\varepsilon}\mathbb{E}[Y], \quad (1.4)$$

for all non-negative random variable Y with finite expectation and any scalar $\varepsilon > 0$. Chebyshev's inequality is obtained by applying Markov's inequality to the random variable $Y = (X - \mathbb{E}[X])^2$ for a random variable X with finite mean $\mathbb{E}[X]$ and variance $\text{var}[X]$, leading to

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) = \mathbb{P}(|X - \mathbb{E}[X]|^2 \geq \varepsilon^2) \leq \frac{1}{\varepsilon^2} \text{var}[X].$$

Thus, from the mean $\mathbb{E}[Z]$ and variance $\frac{\sigma^2}{n}$ of the random variable $\frac{1}{n}\sum_{i=1}^n Z_i$, computed in Eq. (1.3), we obtain the deviation bounds

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z]\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z]\right)^2\right] = \frac{\sigma^2}{n\varepsilon^2},$$

which implies convergence in probability.⁴

To characterize the deviations more finely, there are two classical tools: the *central limit theorem*, which states that $\frac{1}{n}\sum_{i=1}^n Z_i$ is approximately Gaussian with mean $\mathbb{E}[Z]$ and variance σ^2/n . This is an asymptotic statement: formally $\sqrt{n}(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z])$ converges in distribution to a Gaussian distribution with mean zero and variance σ^2 . Although it gives the correct scaling in n , in this textbook, we will look primarily at

⁴See https://en.wikipedia.org/wiki/Convergence_of_random_variables for convergences of random variables.

non-asymptotic results that quantify the deviation for any n .



In what follows, we will always provide versions of inequalities for *averages* of random variables (some authors equivalently consider sums).


Before describing various concentration inequalities, let us recall the classical *union bound*: given events indexed by $f \in \mathcal{F}$ (which can have a countably infinite number of elements), we have:

$$\mathbb{P}\left(\bigcup_{f \in \mathcal{F}} A_f\right) \leq \sum_{f \in \mathcal{F}} \mathbb{P}(A_f).$$

It has (among many other uses in machine learning) a direct application in upper-bounding the tail probability of the supremum of random variables:

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} Z_f > t\right) = \mathbb{P}\left(\bigcup_{f \in \mathcal{F}} \{Z_f > t\}\right) \leq \sum_{f \in \mathcal{F}} \mathbb{P}(Z_f > t).$$

We will only cover the most useful inequalities for machine learning. For more advanced inequalities, see, e.g., [Boucheron et al. \(2013\)](#); [Vershynin \(2018\)](#).

Homogeneity.  Random variables or vectors typically have a unit, and it is always helpful to perform some basic dimensional analysis⁵ to spot mistakes. For example, when performing linear predictions of the form $y = x^\top \theta$, the unit of y is the one of x times that of θ . Typically, these units are encapsulated in the constants describing the problem (such as the noise standard deviation for y or bounds for x and θ).

Exercise 1.6 Let Y be a non-negative random variable with finite expectation, and $\varepsilon > 0$. Show that $\varepsilon 1_{Y \geq \varepsilon} \leq Y$ almost surely and prove Markov's inequality in Eq. (1.4).

Exercise 1.7 Let Y be a non-negative random variable with finite expectation. Show that $\mathbb{E}[Y] = \int_0^\infty \mathbb{P}(Y \geq t) dt$.

1.2.1 Hoeffding's inequality

The simplest concentration inequality considers bounded real-valued random variables.

Proposition 1.1 (Hoeffding's inequality) If Z_1, \dots, Z_n are independent random variables such that $Z_i \in [0, 1]$ almost surely, then, for any $t \geq 0$,

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i] \geq t\right) \leq \exp(-2nt^2). \quad (1.5)$$

Proof The usual proof uses standard convexity arguments and is divided into two parts.

⁵https://en.wikipedia.org/wiki/Dimensional_analysis

- (1) Lemma: If $Z \in [0, 1]$ almost surely, then $\mathbb{E}[\exp(s(Z - \mathbb{E}[Z]))] \leq \exp(s^2/8)$ for any $s \geq 0$.

Proof: we can compute the first two derivatives of $\varphi : s \mapsto \log(\mathbb{E}[\exp(s(Z - \mathbb{E}[Z]))])$, which is a “log-sum-exp” function, often referred to as the “cumulant generating function”, so that the second derivative is related to a certain variance. We can compute the derivatives of φ as:

$$\begin{aligned}\varphi'(s) &= \frac{\mathbb{E}[(Z - \mathbb{E}[Z])e^{s(Z - \mathbb{E}[Z])}]}{\mathbb{E}[e^{s(Z - \mathbb{E}[Z])}]} \\ \varphi''(s) &= \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^2 e^{s(Z - \mathbb{E}[Z])}]}{\mathbb{E}[e^{s(Z - \mathbb{E}[Z])}]} - \left[\frac{\mathbb{E}[(Z - \mathbb{E}[Z])e^{s(Z - \mathbb{E}[Z])}]}{\mathbb{E}[e^{s(Z - \mathbb{E}[Z])}]} \right]^2.\end{aligned}$$

We thus get $\varphi(0) = \varphi'(0) = 0$, and $\varphi''(s)$ is the variance of some random variable $\tilde{Z} \in [0, 1]$, with distribution with density $z \mapsto e^{s(z - \mathbb{E}[Z])}$ with respect to μ , where μ is the distribution of Z . We recall that the variance of \tilde{Z} is the minimum squared deviation to a constant and can thus bound this variance as

$$\text{var}(\tilde{Z}) = \inf_{\nu \in [0, 1]} \mathbb{E}[(\tilde{Z} - \nu)^2] \leq \mathbb{E}[(\tilde{Z} - 1/2)^2] = \frac{1}{4} \mathbb{E}[(2\tilde{Z} - 1)^2] \leq \frac{1}{4},$$

since $2\tilde{Z} - 1 \in [-1, 1]$ almost surely. Thus, for all $s \geq 0$, $\varphi''(s) \leq 1/4$, and by Taylor’s formula, $\varphi(s) \leq \frac{s^2}{8}$.

- (2) We recall Markov’s inequality for any non-negative random variable X and $a > 0$, which states $\mathbb{P}(X \geq a) \leq \frac{1}{a} \mathbb{E}[X]$. For any $t \geq 0$, and denoting $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$:

$$\begin{aligned}& \mathbb{P}(\bar{Z} - \mathbb{E}[\bar{Z}] \geq t) \\ &= \mathbb{P}(\exp(s(\bar{Z} - \mathbb{E}[\bar{Z}])) \geq \exp(st)) \text{ by monotonicity of the exponential,} \\ &\leq \exp(-st) \mathbb{E}[\exp(s(\bar{Z} - \mathbb{E}[\bar{Z}]))] \text{ using Markov’s inequality,} \\ &\leq \exp(-st) \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{s}{n}(Z_i - \mathbb{E}[Z_i])\right)\right] \text{ by independence,} \\ &\leq \exp(-st) \prod_{i=1}^n \exp\left(\frac{s^2}{8n^2}\right) = \exp\left(-st + \frac{s^2}{8n}\right), \text{ using the lemma above,}\end{aligned}$$

which is minimized for $s = 4nt$. We then get the result. ■

Note the difference with the central limit theorem, which states that when n goes to infinity, the probability in Eq. (1.5) is asymptotically equivalent to

$$\frac{1}{\sqrt{2\pi\sigma^2/n}} \int_t^\infty \exp\left(-\frac{nz^2}{2\sigma^2}\right) dz \text{ which can be shown to be less than } \exp\left(-\frac{nt^2}{2\sigma^2}\right),$$

where $\sigma^2 = \lim_{n \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^n \text{var}(Z_i)$. The central limit theorem is more precise (as it involves the variance of Z_i ’s and not an almost sure bound) but is asymptotic. Bernstein

inequalities (see Section 1.2.3) will be in between as they use both the variance and an almost sure bound.

Extensions. We get the following corollary by just applying the inequality to Z_i 's and $1 - Z_i$'s and using the union bound.

Corollary 1.1 (Two-sided Hoeffding's inequality) *If Z_1, \dots, Z_n are independent random variables such that $Z_i \in [0, 1]$ almost surely, then, for any $t \geq 0$,*

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i]\right| \geq t\right) \leq 2 \exp(-2nt^2). \quad (1.6)$$

We can make the following observations:

- Hoeffding's inequality can be extended to the assumption that $Z_i \in [a, b]$ almost surely, leading to

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i]\right| \geq t\right) \leq 2 \exp(-2nt^2/(a-b)^2).$$

- Such an inequality is often used “in the other direction”, starting from the probability and deriving t from it as follows. For any $\delta \in (0, 1)$, with probability greater than $1 - \delta$, we have:

$$\left|\frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i]\right| < \frac{|a-b|}{\sqrt{2n}} \sqrt{\log\left(\frac{2}{\delta}\right)}.$$

Note the dependence in n as $1/\sqrt{n}$ and the logarithmic dependence in δ (corresponding to the exponential tail bound in t).

Exercise 1.8 *Show the one-sided inequality: with probability greater than $1 - \delta$,*

$$\frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i] < \frac{|a-b|}{\sqrt{2n}} \sqrt{\log\left(\frac{1}{\delta}\right)}.$$

- When $Z_i \in [a_i, b_i]$ almost surely, with potentially different a_i 's and b_i 's, the probability upper-bound can be replaced by $2 \exp(-2nt^2/c^2)$, where $c^2 = \frac{1}{n} \sum_{i=1}^n (b_i - a_i)^2$.
- The result extends to martingales with essentially the same proof, leading to Azuma's inequality (see the exercise below).

Exercise 1.9 (Azuma's inequality (♦)) *Assume that the sequence of random variables $(Z_i)_{i \geq 0}$, satisfies $\mathbb{E}(Z_i | \mathcal{F}_{i-1}) = 0$ for an increasing sequence of increasing “ σ -fields” $(\mathcal{F}_i)_{i \geq 0}$,⁶ and $|Z_i| \leq c_i$ almost surely, for $i \geq 1$. Then*

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Z_i \geq t\right) \leq \exp\left(\frac{-n^2 t^2}{2(c_1^2 + \dots + c_n^2)}\right).$$

⁶See more details in https://en.wikipedia.org/wiki/Azuma's_inequality.

- Hoeffding's inequality is often applied to so-called “sub-Gaussian” random variables, that is, random variables X for which there exists $\tau \in \mathbb{R}_+$ such that the following bound on the Laplace transform of X holds:

$$\forall s \in \mathbb{R}, \mathbb{E}[\exp(s[X - \mathbb{E}[X]])] \leq \exp\left(\frac{\tau^2 s^2}{2}\right),$$

which is exactly what we used in the proof. In other words, a random variable with values in $[a, b]$ is sub-Gaussian with constant $\tau^2 = (b - a)^2/4$. For these sub-Gaussian variables, we have similar concentration inequalities (see next exercise). Moreover, for such sub-Gaussian random variables, we have the usual two versions of the tail bound:

$$\forall t \geq 0, \mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\tau^2}\right) \quad (1.7)$$

$$\Leftrightarrow \forall \delta \in (0, 1], |Z - \mathbb{E}[Z]| \leq \tau \sqrt{2 \log\left(\frac{2}{\delta}\right)} \text{ with probability } 1 - \delta. \quad (1.8)$$

Exercise 1.10 Show that a Gaussian random variable with variance σ^2 is sub-Gaussian with constant σ^2 .

Exercise 1.11 If Z_1, \dots, Z_n are independent random variables which are sub-Gaussian with constant τ^2 , then, $P(|\frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i]| \geq t) \leq 2 \exp(-\frac{nt^2}{2\tau^2})$ for any $t \geq 0$.

- Sub-Gaussian random variables can be defined in several other ways, equivalent to constants with the bound on the Laplace transform. See the exercises below.

Exercise 1.12 (♦) Let Z be a random variable which is sub-Gaussian with constant τ^2 . Then, by using the tail bound $\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq 2 \exp(-\frac{t^2}{2\tau^2})$ in Eq. (1.7), show that for any positive integer q , $\mathbb{E}[(Z - \mathbb{E}[Z])^{2q}] \leq (2q)q!(2\tau^2)^q$.

Exercise 1.13 (♦♦) Let Z be a random variable such that for any positive integer q , $\mathbb{E}[(Z - \mathbb{E}[Z])^{2q}] \leq (2q)q!(2\tau^2)^q$. Then show that Z is sub-Gaussian with parameter $24\tau^2$.

Exercise 1.14 Assume the random variable Z has almost surely non-negative values and finite second-order moment. Show that for any $s \geq 0$, $\log(\mathbb{E}[e^{-sZ}]) \leq -s\mathbb{E}[Z] + \frac{s^2}{2}\mathbb{E}[Z^2]$.

1.2.2 McDiarmid's inequality

Given n independent random variables, it may be useful to concentrate other quantities than their average. What is needed is that the function of these random variables has “bounded variation”.

Proposition 1.2 (McDiarmid's inequality) Let Z_1, \dots, Z_n be independent random variables (in any measurable space \mathcal{Z}), and $f : \mathcal{Z}^n \rightarrow \mathbb{R}$ a function of “bounded variation”,

that is, such that for all $i \in \{1, \dots, n\}$, and all $z_1, \dots, z_n, z'_i \in \mathcal{Z}$, we have

$$|f(z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_n) - f(z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_n)| \leq c.$$

Then

$$\mathbb{P}(|f(Z_1, \dots, Z_n) - \mathbb{E}[f(Z_1, \dots, Z_n)]| \geq t) \leq 2 \exp(-2t^2/(nc^2)).$$

Proof (◆) The proof generalizes the formulation of Hoeffding's inequality in Eq. (1.6), which corresponds to $f(z) = \frac{1}{n} \sum_{i=1}^n z_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i]$ and $c = \frac{1}{n}$. We will only consider the one-sided inequality

$$\mathbb{P}(f(Z_1, \dots, Z_n) - \mathbb{E}[f(Z_1, \dots, Z_n)] \geq t) \leq \exp(-2t^2/(nc^2)),$$

which is sufficient to get the two-sided inequality using the union bound.

We introduce the random variables, for $i \in \{1, \dots, n\}$:

$$V_i = \mathbb{E}[f(Z_1, \dots, Z_n)|Z_1, \dots, Z_i] - \mathbb{E}[f(Z_1, \dots, Z_n)|Z_1, \dots, Z_{i-1}],$$

with $V_1 = \mathbb{E}[f(Z_1, \dots, Z_n)|Z_1] - \mathbb{E}[f(Z_1, \dots, Z_n)]$. We have $\mathbb{E}[V_i|Z_1, \dots, Z_{i-1}] = 0$. Moreover, given Z_1, \dots, Z_n , the maximal value of V_i minus the minimal value of V_i is almost surely less than c as a consequence of the bounded variation assumption since it is the difference of two terms that are conditional expectations of values of f taken at arguments that only differ in the i -th variable. Moreover, through a telescoping sum, we have $f(Z_1, \dots, Z_n) - \mathbb{E}[f(Z_1, \dots, Z_n)] = \sum_{i=1}^n V_i$. Using the same argument as in part (1) of the proof of Hoeffding's inequality, we get for any $s > 0$, $\mathbb{E}(e^{sV_i}|Z_1, \dots, Z_{i-1}) \leq e^{s^2 c^2/8}$, and we can obtain a proof with the same steps as part (2) of Hoeffding's inequality by being careful with conditioning, for any $s \geq 0$:

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^n V_i \geq t\right) &\leq \exp(-st) \cdot \mathbb{E}\left[\exp\left(s \sum_{i=1}^n V_i\right)\right] \text{ using Markov's inequality,} \\ &= \exp(-st) \cdot \mathbb{E}\left[\exp\left(s \sum_{i=1}^{n-1} V_i\right) \mathbb{E}[\exp(sV_n)|Z_1, \dots, Z_{n-1}]\right], \\ &\quad \text{since } V_1, \dots, V_{n-1} \text{ are in the } \sigma\text{-algebra generated by } Z_1, \dots, Z_{n-1}, \\ &\leq \exp(-st + s^2 c^2/8) \cdot \mathbb{E}\left[\exp\left(s \sum_{i=1}^{n-1} V_i\right)\right], \end{aligned}$$

using the bound above on $\mathbb{E}(e^{sV_n}|Z_1, \dots, Z_{n-1})$. Applying the same reasoning n times, we get a probability less than $\exp(-st + ns^2 c^2/8)$ and the desired result by minimizing with respect to s (leading to $s = 4t/(nc^2)$). ■

This inequality will be used to provide high-probability bounds on the estimation error in empirical risk minimization in Section 4.4.1.

Exercise 1.15 (♦) Use McDiarmid's inequality to prove a Hoeffding-type bound for vectors, that is, if $Z_1, \dots, Z_n \in \mathbb{R}^d$ are independent centered vectors such that $\|Z_i\|_2 \leq c$ almost surely, then with probability greater than $1 - \delta$, we have

$$\left\| \frac{1}{n} \sum_{i=1}^n Z_i \right\|_2 \leq \frac{c}{\sqrt{n}} \left(1 + \sqrt{2 \log \frac{1}{\delta}} \right).$$

1.2.3 Bernstein's inequality (♦)

As mentioned earlier, Hoeffding's inequality only uses an almost sure bound, but not explicitly the variance, as the central limit theorem is using (but only with an asymptotic result). Bernstein's inequality allows the use of variance to get a finer non-asymptotic result.

Proposition 1.3 (Bernstein's inequality) Let Z_1, \dots, Z_n be n independent random variables such that $|Z_i| \leq c$ almost surely and $\mathbb{E}[Z_i] = 0$. Then, for $t \geq 0$,

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i\right| \geq t\right) \leq 2 \exp\left(-\frac{nt^2}{2\sigma^2 + 2ct/3}\right), \quad (1.9)$$

where $\sigma^2 = \frac{1}{n} \sum_{i=1}^n \text{var}(Z_i)$. Moreover, for $\delta \in (0, 1)$, with probability greater than $1 - \delta$, we have:

$$\left|\frac{1}{n} \sum_{i=1}^n Z_i\right| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}} + \frac{2c \log(2/\delta)}{3n}. \quad (1.10)$$

Proof The proof is also divided into two parts, with first a lemma on the Laplace transform.

- (a) Lemma: if $|Z| \leq c$ almost surely, $\mathbb{E}[Z] = 0$, and $\mathbb{E}[Z^2] = \sigma^2$, then for any $s > 0$, we have $\mathbb{E}[e^{sZ}] \leq \exp\left(\frac{\sigma^2}{c^2}(e^{sc} - 1 - sc)\right)$.

Proof: using the power series expansion of the exponential, we get:

$$\begin{aligned} \mathbb{E}[e^{sZ}] &= 1 + \mathbb{E}[sZ] + \sum_{k=2}^{\infty} \frac{s^k}{k!} \mathbb{E}[Z^k] = 1 + \sum_{k=2}^{\infty} \frac{s^k}{k!} \mathbb{E}[Z^k] \text{ because } Z \text{ has zero mean,} \\ &\leq 1 + \sum_{k=2}^{\infty} \frac{s^k}{k!} \mathbb{E}[|Z|^{k-2} |Z|^2] \leq 1 + \sum_{k=2}^{\infty} \frac{s^k}{k!} c^{k-2} \sigma^2 = 1 + \frac{\sigma^2}{c^2} (e^{sc} - 1 - sc). \end{aligned}$$

Using the bound $1 + \alpha \leq e^\alpha$ valid for all $\alpha \in \mathbb{R}$ leads to the desired result.

(b) With $\sigma_i^2 = \text{var}(Z_i)$, we have the one-sided inequality:

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Z_i \geq t\right) &= \mathbb{P}\left(\exp\left(s \sum_{i=1}^n Z_i\right) \geq \exp(nst)\right) \\
&\quad \text{by monotonicity of the exponential,} \\
&\leq \mathbb{E}\left[\exp\left(s \sum_{i=1}^n Z_i\right)\right] e^{-nst} \text{ using Markov's inequality,} \\
&\leq e^{-nst} \prod_{i=1}^n \exp\left(\frac{\sigma_i^2}{c^2}(e^{sc} - 1 - sc)\right) = e^{-nst} \exp\left(\frac{n\sigma^2}{c^2}(e^{sc} - 1 - sc)\right),
\end{aligned}$$

using the lemma above. We now need to find an upper-bound on the minimal value (with respect to s) of $-nst + \frac{n\sigma^2}{c^2}(e^{sc} - 1 - sc) = \frac{n\sigma^2}{c^2}(e^{sc} - 1 - sc - \alpha sc)$, with $\alpha = ct/\sigma^2$. We first bound for $u = sc$, $e^u - 1 - u = \sum_{k=0}^{\infty} \frac{u^{k+2}}{(k+2)!} \leq \sum_{k=0}^{\infty} \frac{u^{k+2}}{2 \cdot 3^k}$, since $(k+2)! = 2 \cdot 3 \cdots (k+2) \geq 2 \cdot 3^k$. Thus, for $u \in (0, 3)$, we get

$$e^u - 1 - u \leq \frac{u^2}{2} \sum_{k=0}^{\infty} (u/3)^k = \frac{u^2}{2} \frac{1}{1 - u/3}.$$

Using the candidate $u = \frac{\alpha}{1+\alpha/3}$, we get $1 - u/3 = \frac{3}{\alpha+3}$, and thus:

$$e^u - 1 - u - \alpha u \leq \frac{u^2}{2} \frac{1}{1 - u/3} - \alpha u = \frac{\alpha^2}{2(1 + \alpha/3)^2} \frac{\alpha + 3}{3} - \frac{\alpha^2}{1 + \alpha/3} = -\frac{\alpha^2}{2(1 + \alpha/3)}.$$

This exactly leads to the one-sided version of Eq. (1.9).

To get Eq. (1.10) from the two sided-version of Eq. (1.9), we solve in t the equation $2 \exp\left(\frac{-nt^2}{2\sigma^2 + 2ct/3}\right) = \delta \Leftrightarrow \log \frac{2}{\delta} = \frac{nt^2}{2\sigma^2 + 2ct/3}$. Solving the quadratic equation in t leads to (using $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$):

$$t = \frac{1}{2} \left[\frac{2c}{3n} \log \frac{2}{\delta} + \left(\left(\frac{2c}{3n} \log \frac{2}{\delta} \right)^2 + \frac{8\sigma^2}{n} \log \frac{2}{\delta} \right)^{1/2} \right] \leq \frac{2c}{3n} \log \frac{2}{\delta} + \frac{1}{2} \left(\frac{8\sigma^2}{n} \log \frac{2}{\delta} \right)^{1/2},$$

which leads to Eq. (1.10). ■

Note here that we get the same dependence as for the central limit theorem for small deviations t (and a strict improvement on Hoeffding because the variance is essentially bounded by the squared diameter of the support). In contrast, for large t , the dependence in t is worse than Hoeffding's inequality.

Beyond zero mean random variables. Bernstein's inequality can also be applied when the random variables Z_i do not have zero means. Then Eq. (1.9) is replaced by

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i]\right| \geq t\right) \leq 2 \exp\left(-\frac{nt^2}{2\sigma^2 + 2ct/3}\right).$$

Exercise 1.16 (♦) *Prove the inequality above.*

1.2.4 Expectation of the maximum

Concentration inequalities bound the deviation from the expectation. Often, computing the expectation is tricky, particularly for maxima of random variables. In a nutshell, taking the maximum of n bounded random variables leads to an extra factor of $\sqrt{\log n}$. Note here that we do not impose independence. We will consider other tools such as Rademacher complexities in Section 4.5. See Figure 1.1 for an illustration.

⚠ This logarithmic factor appears many times in this textbook and can often be traced back to the expectation of a maximum and to the Gaussian decay of tail bounds.

⚠ The variables do not need to be independent.

Proposition 1.4 (Expectation of the maximum) *If Z_1, \dots, Z_n are (potentially dependent) zero-mean real random variables which are sub-Gaussian with constant τ^2 , then*

$$\mathbb{E}[\max\{Z_1, \dots, Z_n\}] \leq \sqrt{2\tau^2 \log n}.$$

Proof We have:

$$\begin{aligned} \mathbb{E}[\max\{Z_1, \dots, Z_n\}] &\leq \frac{1}{t} \log \mathbb{E}[e^{t \max\{Z_1, \dots, Z_n\}}] \text{ by Jensen's inequality,} \\ &= \frac{1}{t} \log \mathbb{E}[\max\{e^{tZ_1}, \dots, e^{tZ_n}\}] \\ &\leq \frac{1}{t} \log \mathbb{E}[e^{tZ_1} + \dots + e^{tZ_n}] \text{ bounding the max by the sum,} \\ &\leq \frac{1}{t} \log(ne^{\tau^2 t^2/2}) = \frac{\log n}{t} + \tau^2 \frac{t}{2} = \sqrt{2\tau^2 \log n} \text{ with } t = \tau^{-1} \sqrt{2 \log n}, \end{aligned}$$

using the definition of sub-Gaussianity in Section 1.2.1 (and the fact that the variables have zero means). ■

While we consider a direct proof using Laplace transforms above, we can prove a similar result using Gaussian tail bounds together with the union bound

$$\mathbb{P}(\max\{U_1, \dots, U_n\} \geq t) \leq \mathbb{P}(U_1 \geq t) + \dots + \mathbb{P}(U_n \geq t),$$

for well-chosen random variables U_1, \dots, U_n . In other words, the dependence in the probability δ as $\sqrt{\log(\frac{2}{\delta})}$ in Eq. (1.8) is directly related to the term $\sqrt{\log n}$ above (see exercise below). We will see a different dependence in n in Section 8.1.2 for the maximum of squared norms of Gaussians.

Exercise 1.17 *Assume Z_1, \dots, Z_n are random variables that are sub-Gaussian with constant τ^2 and have zero means. Show that $\mathbb{E}[\max\{|Z_1|, \dots, |Z_n|\}] \leq \sqrt{2\tau^2 \log(2n)}$. Prove*

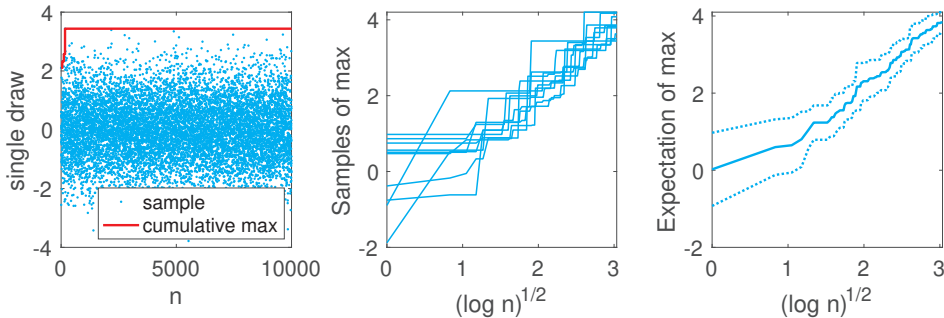


Figure 1.1: Expectation of the maximum of n independent standard Gaussian random variables. Left: illustration of the cumulative maximum $\max\{Z_1, \dots, Z_n\}$. Middle: 10 samples of the cumulative maximum as a function of $\sqrt{\log n}$. Right: mean and standard deviations from 1000 replications. Notice the linear growth in $\sqrt{\log n}$ compatible with our bounds.

the same result up to a universal constant using the tail bounds $\mathbb{P}(|Z_i| \geq t) \leq 2 \exp(-\frac{t^2}{2\sigma^2})$ together with the union bound, and the property $\mathbb{E}[|Y|] = \int_0^{+\infty} \mathbb{P}(|Y| \geq t) dt$ for any random variable Y such that $\mathbb{E}[|Y|]$ exists.

Exercise 1.18 (♦♦) Assume Z_1, \dots, Z_n are independent Gaussian random variables with mean zero and variance σ^2 . Provide a lower bound for $\mathbb{E}[\max\{Z_1, \dots, Z_n\}]$ of the form $c\sqrt{\log n}$ for $c > 0$.

Exercise 1.19 (♦♦) We consider a convex function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ such that $f(0) = 0$ and f is L -smooth with respect to the norm Ω , that is, f is continuously differentiable and for all $\theta, \eta \in \mathbb{R}^d$, $f(\theta) \leq f(\eta) + f'(\eta)^\top (\theta - \eta) + \frac{L}{2} \Omega(\theta - \eta)^2$. Let $Z_i \in \mathbb{R}^d$ be independent zero-mean random vectors with $\mathbb{E}[\Omega(Z_i)^2] \leq \sigma^2$, for $i = 1, \dots, n$. Show by induction in n that $\mathbb{E}[f(Z_1 + \dots + Z_n)] \leq nL\frac{\sigma^2}{2}$.

Exercise 1.20 Assume Z_1, \dots, Z_n are sub-Gaussian random variables with common sub-Gaussianity parameter τ , and potentially different means μ_1, \dots, μ_n . For a fixed set of non-negative weights π_1, \dots, π_n that sum to one, and $\delta \in (0, 1)$, show that with probability greater than $1 - \delta$, for all $i \in \{1, \dots, n\}$, $|z_i - \mu_i| \leq \tau\sqrt{2\log(1/\pi_i)} + \tau\sqrt{2\log(2/\delta)}$. If $\hat{i} \in \arg \min_{i \in \{1, \dots, n\}} \{z_i + \tau\sqrt{2\log(1/\pi_i)}\}$, show that with probability greater than $1 - \delta$, $\mu_{\hat{i}} \leq \min_{i \in \{1, \dots, n\}} \{\mu_i + \tau\sqrt{2\log(1/\pi_i)}\} + 2\tau\sqrt{2\log(2/\delta)}$.

1.2.5 Estimation of expectations through quadrature (♦)

In machine learning, the generalization error is an expectation of a function (the loss associated with a specific prediction function) of a random variable (the pair input/output). This generalization error is naturally approximated by an empirical average given some independent and identically distributed (i.i.d.) samples, with a convergence rate of $O(1/\sqrt{n})$ from n samples (as shown, for example, from Hoeffding's inequality).

In this section, we briefly present *quadrature* methods whose aim is to estimate the same expectation but with potentially non-random observations. For simplicity, we consider a random variable X uniformly distributed in $[0, 1]$, and the task of computing the expectation of a function $f : [0, 1] \rightarrow \mathbb{R}$, that is, $I = \mathbb{E}[f(X)] = \int_0^1 f(x)dx$, noting that there are many variants of such methods (see, e.g., [Davis and Rabinowitz, 1984](#); [Brass and Petras, 2011](#)), and that these techniques extend to higher dimensions ([Holtz, 2010](#)). Moreover, while we focus on equally spaced data in the interval, “quasi-random” methods lead to better convergence rates ([Niederreiter, 1992](#)).

We consider uniformly spaced grid points on $[0, 1]$, as it can serve as an idealization of random sampling when studying regression models, in particular in Chapter 6 and Chapter 7. That is, we consider $x_i = \frac{i}{n}$ for $i \in \{0, \dots, n\}$ (with $n + 1$ points). The classical trapezoidal rule considers the approximation

$$\hat{I} = \frac{1}{n} \left[\frac{1}{2} f(x_0) + \sum_{i=1}^{n-1} f(x_i) + \frac{1}{2} f(x_n) \right].$$

The error $|I - \hat{I}|$ then depends on the regularity of f . We have a decomposition of the error as the integral between f and its piecewise affine interpolant:

$$\begin{aligned} I - \hat{I} &= \sum_{i=1}^n \left(\int_{x_{i-1}}^{x_i} f(x)dx - \frac{x_i - x_{i-1}}{2} [f(x_i) + f(x_{i-1})] \right) \\ &= \sum_{i=1}^n \left(\int_{x_{i-1}}^{x_i} f(x)dx - \int_{x_{i-1}}^{x_i} \left\{ \frac{x_i - x}{x_i - x_{i-1}} f(x_{i-1}) + \frac{x - x_{i-1}}{x_i - x_{i-1}} f(x_i) \right\} dx \right). \end{aligned}$$

If f is twice differentiable and has a second-derivative bounded by L uniformly in absolute value, then we have the bound (which can be obtained by Taylor’s formula):

$$|I - \hat{I}| \leq \sum_{i=1}^n \frac{L}{2} \int_{x_{i-1}}^{x_i} (x_i - x)(x - x_{i-1})dx = \sum_{i=1}^n \frac{L}{12} (x_i - x_{i-1})^3 dx = \frac{L}{12n^2}.$$

We thus have an error bound in $O(1/n^2)$ if we assume two bounded derivatives. We typically get an error of $O(1/n^s)$ for such numerical integration methods if we assume s bounded derivatives (with the appropriate rule, such as Simpson’s rule, which makes a piecewise quadratic interpolation). See the exercises below.

Exercise 1.21 *Show that the trapezoidal rule leads to an error in $O(1/n)$ if we only assume one bounded derivative.*

Exercise 1.22 (♦) *Show that for 1-periodic functions, the trapezoidal rule leads to an error in $O(1/n^s)$ if we assume s bounded derivatives.*

1.2.6 Concentration inequalities for matrices (♦♦)

It turns out the concentration inequalities that have been presented in this chapter apply equally well to matrices with the positive semi-definite order. The following bounds are

adapted from [Tropp \(2012\)](#) and presented without proofs, with the following notations: $\lambda_{\max}(M)$ denote the largest eigenvalue of the symmetric matrix M . In contrast, $\|M\|_{\text{op}}$ denotes the largest singular value of a potentially rectangular matrix M , and $A \preceq B$ if and only if $B - A$ is positive semi-definite.

Proposition 1.5 (Matrix Hoeffding bound) ([Tropp, 2012, Theorem 1.3](#)) *Given n independent symmetric matrices $M_i \in \mathbb{R}^{d \times d}$, such that for all $i \in \{1, \dots, n\}$, $\mathbb{E}[M_i] = 0$, $M_i^2 \preceq C_i^2$ almost surely. Then for all $t \geq 0$,*

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n} \sum_{i=1}^n M_i\right) \geq t\right) \leq d \cdot \exp\left(-\frac{nt^2}{8\sigma^2}\right),$$

for $\sigma^2 = \lambda_{\max}\left(\frac{1}{n} \sum_{i=1}^n C_i^2\right)$.

Proposition 1.6 (Matrix Bernstein bound) ([Tropp, 2012, Theorem 1.4](#)) *Given n independent symmetric matrices $M_i \in \mathbb{R}^{d \times d}$, such that for all $i \in \{1, \dots, n\}$, $\mathbb{E}[M_i] = 0$, $\lambda_{\max}(M_i) \leq c$ almost surely. Then for all $t \geq 0$,*

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n} \sum_{i=1}^n M_i\right) \geq t\right) \leq d \cdot \exp\left(-\frac{nt^2/2}{\sigma^2 + ct/3}\right),$$

for $\sigma^2 = \lambda_{\max}\left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}[M_i^2]\right)$.

We can make the following observations:

- Note the similarity with the corresponding bounds for scalar random variables when $d = 1$. McDiarmid's inequality can also be extended ([Tropp, 2012, Corollary 7.5](#)).
- These bounds apply as well to rectangular matrices $M_i \in \mathbb{R}^{d_1 \times d_2}$ by considering the symmetric matrices $\widetilde{M}_i = \begin{pmatrix} 0 & M_i \\ M_i^\top & 0 \end{pmatrix} \in \mathbb{R}^{(d_1+d_2) \times (d_1+d_2)}$, whose eigenvalues are plus and minus the singular values of M_i ; see Section 1.1.4 and [Stewart and Sun \(1990, Theorem 4.2\)](#).

Exercise 1.23 *Assume the matrices $M_i \in \mathbb{R}^{d_1 \times d_2}$ are independent, have zero mean, and such that $\|M_i\|_{\text{op}} \leq c$ almost surely for all $i \in \{1, \dots, n\}$. Show that*

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_{i=1}^n M_i\right\|_{\text{op}} \geq t\right) \leq (d_1 + d_2) \cdot \exp\left(-\frac{nt^2}{8c^2}\right).$$

Moreover, with $\sigma^2 = \max\left\{\lambda_{\max}\left(\frac{1}{n} \sum_{i=1}^n M_i^\top M_i\right), \lambda_{\max}\left(\frac{1}{n} \sum_{i=1}^n M_i M_i^\top\right)\right\}$, show that

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_{i=1}^n M_i\right\|_{\text{op}} \geq t\right) \leq (d_1 + d_2) \cdot \exp\left(-\frac{nt^2/2}{\sigma^2 + ct/3}\right).$$

Infinite-dimensional covariance operators (♦♦). As used within Chapter 7, we will need to extend the results above, which depend on the underlying dimension, to the notion of “intrinsic dimension”, which can still be finite if the underlying dimension is infinite. That is, we have this bound from [Minsker \(2017, Eq. \(3.9\)\)](#):

Proposition 1.7 (Matrix Bernstein bound - intrinsic dimension) *Given n independent random bounded self-adjoint operators M_i on a Hilbert space, such that for all $i \in \{1, \dots, n\}$, $\mathbb{E}[M_i] = 0$, $\lambda_{\max}(M_i) \leq c$ almost surely, and $\frac{1}{n} \sum_{i=1}^n \mathbb{E}[M_i^2] \preceq V$. Then for all $t \geq 0$,*

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n} \sum_{i=1}^n M_i\right) \geq t\right) \leq d \cdot \left(1 + \frac{6}{n^2 t^4} (\sigma^2 + ct/3)^2\right) \exp\left(-\frac{nt^2/2}{\sigma^2 + ct/3}\right),$$

for $\sigma^2 \geq \lambda_{\max}(V)$ and $d = \frac{\text{tr}(V)}{\sigma^2}$. When $t \geq \frac{c}{3n} + \frac{\sigma}{\sqrt{n}}$, then we get the upper-bound $7d \exp\left(-\frac{nt^2/2}{\sigma^2 + ct/3}\right)$.