# Topics in Statistical Theory

Quentin Berthet[*]

## Foreword

These are informal lecture notes, for a course given in Part III at the University of Cambridge in Lent 2016. No originality is claimed for any of its contents. Some parts are based on lecture notes by Philippe Rigollet for the course 18.S997: *High Dimensional Statistics* at MIT (see Rigollet, 2016), and some on the author's Ph.D. thesis (Berthet, 2014). Please let me know of any errors.

## CONTENTS

[*] Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge. Email: q.berthet@statslab.cam.ac.uk

## 1. Deviation bounds.

### 1.1. *Markov and generalization.*

THEOREM 1.1 (Markov's inequality). *Let $X$ be a real-valued, nonnegative random variable such that $\mathbf{E}[X] < +\infty$. We have, for all $t > 0$*

$$\mathbf{P}(X > t) \leq \frac{\mathbf{E}[X]}{t}\,.$$

PROOF. We decompose $X$ in the following way

$$X = X\mathbf{1}\{X \leq t\} + X\mathbf{1}\{X > t\}\,.$$

As a consequence,

$$X \geq X\mathbf{1}\{X > t\} \geq t\mathbf{1}\{X > t\}\,.$$

Taking expectation yields $\mathbf{E}[X] \geq t\mathbf{P}(X > t)$, and the result. □

This inequality can seem almost trivial in itself. In layman's term it means that in a population, not more than half (or a third, or a fourth) of the individuals can have a wealth superior to twice (or three, or four times) the mean wealth, otherwise these individuals together would be worth more than the overall population. It is however a very powerful tool because of the following direct consequence.

THEOREM 1.2. *Let $X$ be a real-valued random variable, and $f : \mathbf{R} \to \mathbf{R}+$ increasing such that $\mathbf{E}[f(X)] < +\infty$. We have, for all $t \geq 0$*

$$\mathbf{P}(X > t) \leq \frac{\mathbf{E}[f(X)]}{f(t)}\,.$$

PROOF. As $f$ is increasing, we have $\mathbf{P}(X > t) = \mathbf{P}(f(X) > f(t))$. The result is a direct consequence of Markov's inequality, taking $X' = f(X)$ and $t' = f(t)$. □

By judicious choices of $f$, this inequality can be applied to obtain numerous inequalities.

EXAMPLE 1.1 (Chebyshev's inequality). Let $X$ be a real-valued random variable, such that $\mathbf{E}[X^2] < +\infty$. We have, for all $t \geq 0$

$$\mathbf{P}(|X - \mathbf{E}[X]| > t) \leq \frac{\mathbf{Var}[X]}{t^2}\,.$$

EXAMPLE 1.2 (Chernoff bounds). Let $X$ be a real-valued random variable, such that $\mathbf{E}[e^{\lambda X}] < +\infty$ for all $\lambda \geq 0$. We have, for all $t \geq 0$

$$\mathbf{P}(X > t) \leq e^{-\lambda t}\,\mathbf{E}[e^{\lambda X}]\,.$$

Clearly, the more quickly $f$ grows, the more powerful these inequalities are for large values of $t$. Ideally, we want a function that is as "explosive" as possible. For $\lambda > 0$, $f : x \mapsto e^{\lambda x}$ is a good example, for variables whose moment generating function $\mathbf{E}[e^{\lambda X}]$ is well-defined.

Consider the case of $X \sim \mathcal{N}(0, \sigma^2)$. Its moment-generating function is $e^{\frac{\lambda^2\sigma^2}{2}}$, and the Chernoff bound yields the following, for all $\lambda \geq 0$.

$$\mathbf{P}(X > t) \leq e^{-\lambda t}\,\mathbf{E}[e^{\lambda X}] \leq e^{-\lambda t + \frac{\lambda^2\sigma^2}{2}}\,.$$

Taking $\lambda = t/\sigma^2$ to minimize the term in the exponential yields

$$\mathbf{P}(X > t) \le e^{-\frac{t^2}{2\sigma^2}} .$$

This inequality can be directly recovered by using the explicit form of the Gaussian distribution, but by proving it this way, we see that it can be extended to any random variable that has a moment-generating function smaller than one of some Gaussian.

1.2. *Class of sub-Gaussian random variables.*

Definition 1.1 (sub-Gaussian). A real-valued random variable $X$ is said to be *sub-Gaussian* with parameter $\sigma^2 > 0$, denoted $X \in \mathsf{sG}(\sigma^2)$, if for all $\lambda \in \mathbf{R}$ it holds that

$$\mathbf{E}[e^{\lambda X}] \le e^{\frac{\lambda^2\sigma^2}{2}} .$$

In simpler term, the class $\mathsf{sG}(\sigma^2)$ is the set of random variables whose moment-generating function is less than the moment-generating function of a variable with distribution $\mathcal{N}(0, \sigma^2)$. This simple bound yields directly several properties for sub-Gaussian random variables.

1.2.1. *Basic properties.*

Proposition 1.1. *For all $X \in \mathsf{sG}(\sigma^2)$, we have $\mathbf{E}[X] = 0$ and $\mathbf{Var}(X) = \mathbf{E}[X^2] \le \sigma^2$.*

Proof. We decompose $e^{\frac{\lambda^2\sigma^2}{2}}$ and $\mathbf{E}[e^{\lambda X}]$ in power series in $\lambda$

$$\begin{aligned} e^{\frac{\lambda^2\sigma^2}{2}} &= \sum_{k\ge 0} \Big(\frac{\sigma^2\lambda^2}{2}\Big)^k \frac{1}{k!} \\ \mathbf{E}[e^{\lambda X}] &= \sum_{k\ge 0} \frac{\lambda^k \mathbf{E}[X^k]}{k!} , \end{aligned}$$

by Fubini. By putting all terms of order greater than 2 on the right hand side of the inequality $\mathbf{E}[e^{\lambda X}] \le e^{\frac{\lambda^2\sigma^2}{2}}$, we have for all $\lambda \in \mathbf{R}$

$$1 + \lambda\mathbf{E}[X] + \frac{\lambda^2}{2}\mathbf{E}[X^2] \le 1 + \frac{\lambda^2}{2}\sigma^2 + o(\lambda^2) .$$

By subtracting 1 on both sides, dividing by $\lambda > 0$ and letting $\lambda \to 0^+$ we obtain $\mathbf{E}[X] \le 0$. By doing the same thing for $\lambda < 0$ and $\lambda \to 0^-$, we obtain $\mathbf{E}[X] \ge 0$, so $\mathbf{E}[X] = 0$. As a consequence, by subtracting by 1, dividing by $\lambda^2/2$ and letting $\lambda$ go to 0, we have $\mathbf{E}[X^2] \le \sigma^2$. □

This class of random variables is of course not empty: for $X \sim \mathcal{N}(0, \sigma^2)$, $X \in \mathsf{sG}(\sigma^2)$ more or less by definition. Here are a few other useful examples.

Example 1.3.

- Let $X$ take value $-1$ and 1 with probability 1/2, often called a Rademacher random variable.

$$\mathbf{E}[e^{\lambda X}] \le \frac{e^{-\lambda} + e^{\lambda}}{2} = \cosh(\lambda) \underset{\text{(Taylor series)}}{\le} e^{\frac{\lambda^2}{2}} .$$

  As such, $X \in \mathsf{sG}(1)$.

- For $a > 0$, let $X$ be uniform over $[-a, a]$.

$$\mathbf{E}[e^{\lambda X}] \leq \int_{-a}^{a} e^{\lambda x} \frac{dx}{2a} = \frac{e^{-\lambda a} - e^{\lambda a}}{2\lambda a} \leq \frac{\sinh(\lambda a)}{\lambda a} \underset{\text{(Taylor series)}}{\leq} e^{\frac{\lambda^2 a^2}{2}} .$$

  By definition, $X \in \mathsf{sG}(a^2)$

- For any random variable $B$ be a taking values in $[a, b]$, $B - \mathbf{E}[B] \in \mathsf{sG}((b-a)^2/4)$ (see exercise).

1.2.2. *Chernoff and Hoeffding-type bounds.* These random variables have been introduced in order to obtain deviation bounds similar to those of Gaussian random variables. We will mimic the proof above to obtain, via a Chernoff bound the following inequality.

Theorem 1.3 (Hoeffding-type bound). *For any variable $X \in \mathsf{sG}(\sigma^2)$, it holds for all $t \geq 0$ that*

$$\mathbf{P}(X > t) \leq e^{-\frac{t^2}{2\sigma^2}} .$$

Proof.

$$\mathbf{P}(X > t) \leq e^{-\lambda t}\, \mathbf{E}[e^{\lambda X}] \leq e^{-\lambda t + \frac{\lambda^2 \sigma^2}{2}} .$$

Taking $\lambda = t/\sigma^2$ to minimize the term in the exponential yields

$$\mathbf{P}(X > t) \leq e^{-\frac{t^2}{2\sigma^2}} .$$

□

This bound is particularly useful when considering sums of independent sub-Gaussian random variables. Recall that for independent Gaussian variables $X_1, \ldots, X_n \sim \mathcal{N}(0, \sigma^2)$

$$\sum_{i=1}^{n} a_i X_i \sim \mathcal{N}(0, \sigma^2 |a|_2^2) .$$

More generally, Gaussian vectors are characterised by their projections $a^\top X$, which are also Gaussian. In a similar manner, we can define sub-Gaussian vectors

Definition 1.2 (sub-Gaussian vectors). A random variable $X$ taking values in $\mathbf{R}^d$ is said to be a *sub-Gaussian vector* with parameter $\sigma^2$, denoted $X \in \mathsf{sG}_d(\sigma^2)$, if $u^\top X \in \mathsf{sG}(\sigma^2)$ for all $u \in \mathcal{S}^{d-1}$.

Theorem 1.4. *For independent r.v. $X_1, \ldots, X_n$ such that $X_i \in \mathsf{sG}(\sigma^2)$, we have $X \in \mathsf{sG}_n(\sigma^2)$.*

Proof.

$$\begin{aligned} \mathbf{E}[e^{\lambda u^\top X}] &= \mathbf{E}\Big[\exp\Big(\lambda \sum_{i=1}^{n} u_i X_i\Big)\Big] = \prod_{i=1}^{n} \mathbf{E}[e^{\lambda u_i X_i}] \\ &\leq \prod_{i=1}^{n} e^{\lambda u_i^2 \sigma^2/2} \leq \exp\Big(\frac{\lambda^2 \sigma^2}{2} |u|_2^2\Big) . \end{aligned}$$

□

This result is very useful to deal with sums of independent sub-Gaussian random variables.

Corollary 1.1. *For independent r.v. $X_1, \dots, X_n$ such that $X_i \in \mathsf{sG}(\sigma^2)$, we have $a^\top X \in \mathsf{sG}(|a|_2^2\sigma^2)$, and as a consequence*

$$\mathbf{P}\Big(\sum_{i=1}^n a_i X_i > t\Big) \le e^{-\frac{t^2}{2\sigma^2|a|_2^2}}\,.$$

As an example, if the variables $X_i - \mathbf{E}[X_i]$ are independent and in $\mathsf{sG}(\sigma^2)$, noting $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, it holds that

$$\mathbf{P}(\bar{X} - \mathbf{E}[\bar{X}] > t) \le e^{-\frac{nt^2}{2\sigma^2}}\,.$$

Remark. All bounds have been written in the form $\mathbf{P}(X > t) \le \varepsilon(t)$. For $X \in \mathsf{sG}(\sigma^2)$, $-X$ is also in $\mathsf{sG}(\sigma^2)$, we can equivalently write $\mathbf{P}(X < -t) \le \varepsilon(t)$ and, from a union bound obtain that $\mathbf{P}(|X| > t) \le 2\varepsilon(t)$. Moreover, these inequalities can be equivalently rewritten in the inverse form $X \le r(\delta)$, with probability $1-\delta$, for any function $r$ such that $\varepsilon(r(\delta)) \le \delta$.

For the variables $X_i$ above, by the result of Corrolary 1.1, the deviation bound can be equivalently rewritten as

$$\bar{X} \le \mathbf{E}[\bar{X}] + \sigma\sqrt{\frac{2\log(1/\delta)}{n}}\,.$$

Note that this bound does not depend on the variances of these variables. For bounded *i.i.d.* variables in $[-1,1]$ with variance $v \ll 1$, it can seem suboptimal, with deviations around the mean of order $\sqrt{1/n}$. However, the central limit theorem tells us that in the limit $\sqrt{n}(\bar{X} - \mathbf{E}[\bar{X}]) \to \mathcal{N}(0, v)$, and that we should therefore expect deviations of order $\sqrt{v/n}$. We will see that this is not possible in general (see exercise), but that weaker versions of such an inequality exist, by studying the moments of sub-Gaussian random variables.

### 1.2.3. *Bernstein-type bound.*

Theorem 1.5. *Let $X$ be in $\mathsf{sG}(\sigma^2)$. For all integers $k \ge 1$ it holds that*

$$\mathbf{E}[|X|^k] \le 2(\sqrt{2}\sigma)^k\,\Gamma\Big(\frac{k}{2}+1\Big)\,,$$

*where the Gamma function is defined as $\Gamma(x) = \int_0^{+\infty} u^{x-1}e^{-u}du$.*

Proof. For $X \in \mathsf{sG}(\sigma^2)$, let $Y = \frac{X}{\sqrt{2}\sigma}$. By defintion, $Y \in \mathsf{sG}(1/2)$ and

$$\mathbf{E}[|X|^k] = (\sqrt{2}\sigma)^k\mathbf{E}[|Y|^k]\,.$$

We derive a bound for this last term, by using the fact that

$$|Y|^k = \int_0^{|Y|^k} dt = \int_0^{+\infty}\mathbf{1}\{|Y|^k > t\}dt = \int_0^{+\infty}\mathbf{1}\{|Y| > t^{1/k}\}dt\,.$$

Taking expectations on both sides and applying Fubini yields

$$\begin{aligned}\mathbf{E}[|Y|^k] &= \int_0^{+\infty}\mathbf{P}(|Y| > t^{1/k})dt \le 2\int_0^{+\infty} e^{-t^{\frac{2}{k}}}dt\\ &\le k\int_0^{+\infty} u^{\frac{k}{2}-1}e^{-u}du = k\,\Gamma\Big(\frac{k}{2}\Big) = 2\,\Gamma\Big(\frac{k}{2}+1\Big)\,.\end{aligned}$$

□

This can also be used to define equivalently (up to constants) the class of sub-Gaussian random variables (see exercise). Some other moment conditions will give different bounds, as in the following theorem.

Definition 1.3. Let $X$ be a real-valued random variable with $\mathbf{E}[X] = 0$. We say that it satisfies a Bernstein moment condition with parameters $(v, b)$ if for all integers $k \geq 2$

$$\mathbf{E}[|X|^k] \leq \frac{1}{2} k!\, v\, b^{k-2}\,.$$

Theorem 1.6 (Bernstein's inequality). *For any random variable $X$ that satisfies a Bernstein moment condition with parameters $(v, b)$, it holds for all $t \geq 0$ that*

$$\mathbf{P}(X > t) \leq \exp\Big( - \frac{t^2}{2v + 2bt}\Big)\,.$$

Proof. We derive a bound on the moment generating function of such a variable, based on the moment condition. By the series expansion and Fubini, we have

$$\begin{aligned}\mathbf{E}[e^{\lambda X}] &\leq 1 + \frac{\lambda^2 \mathbf{E}[X^2]}{2} + \sum_{k\geq 3} \frac{\lambda^k \mathbf{E}[X^k]}{k!} \\ &\leq 1 + \frac{\lambda^2 v}{2} + \frac{\lambda^2 v}{2} \sum_{k \geq 3} \lambda^{k-2} b^{k-2} \\ &\leq 1 + \frac{\lambda^2 v/2}{1 - b|\lambda|} \leq e^{\frac{\lambda^2 v/2}{1-b|\lambda|}} \quad \text{for all } \lambda < 1/b\,.\end{aligned}$$

We apply a Chernoff bound to obtain a deviation bound

$$\mathbf{P}(X > t) \leq e^{-\lambda t} \mathbf{E}[e^{\lambda X}] \leq e^{-\lambda t + \frac{\lambda^2 v/2}{1 - b|\lambda|}}\,.$$

Using $\lambda = \frac{t}{bt+v} < \frac{1}{b}$ yields the desired result. □

Note that this moment condition is particularly interesting for bounded random variables. Indeed, for all random variables such that $|X| \leq a$ for some $a > 0$ and $\mathbf{E}[X] = 0$, one obtains directly for all integers $k \geq 2$

$$\mathbf{E}[|X|^k] \leq \mathbf{E}[X^2 a^{k-2}] \leq \mathbf{E}[X^2] a^{k-2} \leq \frac{1}{2} k!\, \mathbf{E}[X^2] a^{k-2}\,.$$

From Bernstein's inequality, we obtain for such variables the inequality

$$\mathbf{P}(X > t) \leq \exp\Big( - \frac{t^2}{2\mathbf{E}[X^2] + 2at}\Big)\,.$$

For small values of $t$, this is strictly better than a Hoeffding type inequality in $e^{-\frac{t^2}{2\sigma^2}}$, as $\mathbf{E}[X^2] \leq \sigma^2$. For a fixed sub-Gaussian constant, the variance of a random variable can be arbitrarily small (see exercise). On the other hand, for large values of $t$, the tail is of order $e^{-\frac{t}{2a}}$, which is not as good. This is a necessary sacrifice: as mentioned above, for a sub-Gaussian variable with variance $v$, it is not possible to bound the tail by an expression of order $e^{-\frac{t^2}{2v}}$ (see exercise).

1.3. *Class of sub-exponential random variables.* We introduced the class of sub-Gaussian random variables as morally those who have a tail of order $e^{-t^2}$. Similarly, we can consider those who have, as in the previous inequality, a tail of order $e^{-t}$.

Definition 1.4 (sub-exponential random variables). A real-valued random variable $X$ is said to be *sub-exponential* with parameters $\nu^2 > 0$ and $a > 0$, denoted $X \in \mathsf{sE}(\nu^2, a)$, if it holds that

$$\mathbf{E}[e^{\lambda X}] \le e^{\frac{\lambda^2 \nu^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{a}.$$

These classes are obviously not empty, as for all $a > 0$, $\mathsf{sG}(\sigma^2) \subset \mathsf{sE}(\sigma^2, a)$. They are different though: consider $X$ with Laplace distribution with parameter 1, such that $\mathbf{P}(|X| > t) = e^{-t}$. It is clearly not sub-Gaussian for any parameter $\sigma^2 > 0$. However, its moment generating function can be easily derived

$$\mathbf{E}[e^{\lambda X}] = \frac{1}{1 - \lambda^2} \quad \text{for all } |\lambda| < 1.$$

Observe that this yields

$$\mathbf{E}[e^{\lambda X}] \le e^{2\lambda^2} \quad \text{for all } |\lambda| < \frac{1}{2},$$

so $X \in \mathsf{sE}(2, 2)$. It is straightforward that for all $t > 0$, $X \in \mathsf{sE}(\nu^2, a)$ implies $tX \in \mathsf{sE}(t^2\nu^2, ta)$.

1.3.1. *Basic properties.* Here are some properties of these variables.

Proposition 1.2. *For all $X \in \mathsf{sE}(\nu^2, a)$, we have $\mathbf{E}[X] = 0$ and $\mathbf{Var}(X) = \mathbf{E}[X^2] \le \nu^2$.*

Proof. As in sub-Gaussian random variables, as only values of $\lambda$ close to 0 are used. □

Theorem 1.7. *For independent random variables $X_1, \dots, X_n$ such that $X_i \in \mathsf{sE}(\nu_i^2, a_i)$, we have*

$$\sum_{i=1}^n X_i \in \mathsf{sE}\Big(\sum_{i=1}^n \nu_i^2, \max_i a_i\Big).$$

Proof. We have

$$\mathbf{E}[\exp\Big(\lambda \sum_{i=1}^n X_i\Big)] \le \prod_{i=1}^n \mathbf{E}[e^{\lambda X_i}] \le \prod_{i=1}^n e^{\frac{\lambda^2 \nu_i^2}{2}} \quad \text{for } |\lambda| < 1/\max_i a_i\,.$$

□

As a consequence, for i.i.d variables such that $X_i - \mathbf{E}[X] \in \mathsf{sE}(\nu^2, a)$, applying the theorem above yields that $\bar{X} - E[X] \in \mathsf{sE}(\nu^2/n, a/n)$.

1.3.2. *Bounds.* For sub-Gaussian variables, this type of result is used to derive a tail bound on the empirical average. We can obtain similar results for sub-exponential random variables.

Theorem 1.8 (Sub-exponential deviation bounds). *Let $X \sim \mathsf{sE}(\nu^2, a)$. It holds for all $t \ge 0$ that*

$$\mathbf{P}(X > t) \le \exp\Big(-\frac{t^2}{2\nu^2} \wedge \frac{t}{2a}\Big).$$

Proof. We use a Chernoff bound

$$\mathbf{P}(X > t) \leq e^{-\lambda t}\mathbf{E}[e^{\lambda X}] \leq e^{-\lambda t + \frac{\lambda^2 \nu^2}{2}} \quad \text{for } 0 \leq \lambda < 1/a$$

To optimize the term in the exponential, we can consider two cases: if $t < \nu^2/a$, we can pick the global minimum $\lambda^\star = t/\nu^2 < 1/a$, and obtain the bound $e^{-\frac{t^2}{2\nu^2}}$. If $t \geq \nu^2/a$, since the term in the exponential is decreasing over $[0, \lambda^\star)$, we can pick $\lambda = 1/a$ and obtain the bound $e^{-\frac{t}{2a}}$. □

As an example, for the i.i.d variables such that $X_i - \mathbf{E}[X] \in \mathsf{sE}(1,1)$ considered above, this yields that

$$\mathbf{P}(\bar{X} - \mathbf{E}[X] > t) \leq \exp\Big(-\frac{nt^2}{2\nu^2} \wedge \frac{nt}{2a}\Big)\,.$$

Inverting this inequality yields that with probability $1-\delta$

$$\bar{X} \leq \mathbf{E}[X] + \sqrt{\frac{2\nu^2 \log(1/\delta)}{n}} + \frac{2a\log(1/\delta)}{n}\,.$$

Sub-exponential variables show up in high-dimensional statistic mainly as squares of sub-Gaussian random variables, as shown in the following.

Theorem 1.9. *For any variable $X \in \mathsf{sG}(\sigma^2)$, $X^2 - E[X^2] \in \mathsf{sE}(128\sigma^4, 8\sigma^2)$*

Proof. We derive an upper bound on the moment generating function of $X^2 - \mathbf{E}[X^2]$, from bounds on the moments of $X$.

$$\begin{aligned}
\mathbf{E}[e^{\lambda(X^2 - \mathbf{E}[X^2])}] &= 1 + \sum_{k\geq 2} \frac{\lambda^k \mathbf{E}[(X^2 - E[X^2])^k]}{k!} \\
&\leq 1 + \sum_{k\geq 2} \frac{|\lambda|^k 2^k \mathbf{E}[|X|^{2k}]}{k!} \\
&\leq 1 + \sum_{k\geq 2} |\lambda|^k 2^k 2(\sqrt{2}\sigma)^{2k} \\
&\leq 1 + 32\lambda^2\sigma^4 \sum_{k\geq 0} (4|\lambda|\sigma^2)^k \\
&\leq 1 + \frac{32\lambda^2\sigma^4}{1 - 4|\lambda|\sigma^2} \qquad \text{for } |\lambda| < \frac{1}{4\sigma^2} \\
&\leq 1 + 64\lambda^2\sigma^4 \qquad \text{for } |\lambda| < \frac{1}{8\sigma^2} \\
&\leq e^{64\lambda^2\sigma^4}\,.
\end{aligned}$$

□

Note that the constants 128 and 8 have not been optimised, the scaling in $\sigma^4$ and $\sigma^2$ is the important part.

1.4. *Maximal inequalities.* Often in statistical problems, we will be interested not only in controlling some random variables, or their linear combinations, but in controlling their maximum over some set.

### 1.4.1. *Maximum over a finite set.*

Theorem 1.10. *Let $X_1, \dots, X_N$ be $N$ random variables such that $X_i \in \mathsf{sG}(\sigma^2)$.*

$$\mathbf{E}[\max_{1\le i\le N} X_i] \le \sigma\sqrt{2\log(N)} \quad \textit{and} \quad \mathbf{E}[\max_{1\le i\le N} |X_i|] \le \sigma\sqrt{2\log(2N)}\,.$$

*Moreover, for all $t > 0$*

$$\mathbf{P}(\max_{1\le i\le N} X_i > t) > Ne^{-\frac{t^2}{2\sigma^2}} \quad \textit{and} \quad \mathbf{P}(\max_{1\le i\le N} |X_i| > t) > 2Ne^{-\frac{t^2}{2\sigma^2}}$$

Note that the variables are not required to be independent.

Proof. To obtain the first inequality, we derive for any $\lambda > 0$,

$$\begin{aligned}
\mathbf{E}[\max_{1\le i\le N} X_i] &= \frac{1}{\lambda}\mathbf{E}\big[\log\big(e^{\lambda \max_{1\le i\le N} X_i}\big)\big]\\
&= \frac{1}{\lambda}\log \mathbf{E}\big[\max_{1\le i\le N} e^{\lambda X_i}\big]\\
&\le \frac{1}{\lambda}\log\Big[\sum_{i=1}^N \mathbf{E}[e^{\lambda X_i}]\Big]\\
&\le \frac{1}{\lambda}\log\Big[\sum_{i=1}^N e^{\frac{\lambda^2\sigma^2}{2}}\Big]\\
&= \frac{\log(N)}{\lambda} + \frac{\lambda\sigma^2}{2}\,.
\end{aligned}$$

Taking $\lambda = \sqrt{2\log(N)/\sigma^2}$ yields the first inequality. The third inequality is a consequence of a union bound on $N$ events

$$\mathbf{P}\big(\max_{1\le i\le N} X_i > t\big) = \mathbf{P}\big(\bigcup_{1\le i\le N}\{X_i > t\}\big) \le \sum_{i=1}^N \mathbf{P}(X_i > t) \le Ne^{-\frac{t^2}{2\sigma^2}}\,.$$

The two other inequalities are a direct consequence of the fact that

$$\max_{1\le i\le N} |X_i| = \max_{1\le i\le 2N} Y_i\,,$$

where $Y_i = X_i$ and $Y_{i+N} = -X_i$ for $1 \le i \le N$. □

1.4.2. *Maximum over a polytope.* The results above only apply to a finite family of sub-Gaussian random variables. There are several cases where this can be extended to an infinite family, by reduction to the finite case. We consider here the special case of variables indexed by a polytope.

A *polytope* $P$ of $\mathbf{R}^d$ is the convex hull of a finite number of points, denoted by $\mathcal{V}(P)$. We consider here the family of $\theta^\top X$ for some sub-Gaussian random vector of $\mathbf{R}^d$, for all $\theta \in P$. The link with the results above is made evident by the following lemma

Lemma 1.1. *For a polytope $P \subset \mathbf{R}^d$ and any $x \in \mathbf{R}^d$, it holds that*

$$\max_{\theta\in P} \theta^\top x = \max_{v\in\mathcal{V}(P)} v^\top x\,.$$

Proof. Let $v_1, \ldots, v_N$ denote the elements of $\mathcal{V}(P)$. For any $\theta \in P$, there exists nonnegative reals $\lambda_1, \ldots, \lambda_N$ that sum to one such that $\theta = \lambda_1 v_1 + \ldots + \lambda_N v_N$. As a consequence, for all $x \in \mathbf{R}^d$ it holds that

$$\theta^\top x = \sum_{i=1}^N \lambda_i v_i^\top x \le \sum_{i=1}^N \lambda_i \max_{v \in \mathcal{V}(P)} v^\top x \le \max_{v \in \mathcal{V}(P)} v^\top x\,.$$

Taking maximum over $\theta \in P$ yields

$$\max_{\theta \in P} \theta^\top x \le \max_{v \in \mathcal{V}(P)} v^\top x\,.$$

The reversed inequality is a direct consequence of $\mathcal{V}(P) \subset P$. □

Combining the last two results, we obtain the following theorem.

Theorem 1.11. *Let $P$ be a polytope such that $|\mathcal{V}(P)| = N$, and $X \in \mathbf{R}^d$ be a random vector such that $v^\top X \in \mathsf{sG}(\sigma^2)$ for all $v \in \mathcal{V}(P)$. It holds that*

$$\mathbf{E}[\max_{\theta \in P} \theta^\top X] \le \sigma\sqrt{2\log(N)} \quad \text{and} \quad \mathbf{E}[\max_{\theta \in P} |\theta^\top X|] \le \sigma\sqrt{2\log(2N)}\,.$$

*Moreover, for all $t > 0$*

$$\mathbf{P}(\max_{\theta \in P} \theta^\top X) \le N e^{-\frac{t^2}{2\sigma^2}} \quad \text{and} \quad \mathbf{P}(\max_{\theta \in P} |\theta^\top X| > t) \le 2N e^{-\frac{t^2}{2\sigma^2}}$$

In particular, if $X \in \mathsf{sG}_d(\sigma^2)$ and the polytope is a subset of the unit Euclidean ball ($P \subset \mathcal{B}_2^d$), then $\|v\|_2 \le 1$ for all $v \in \mathcal{V}(P)$ and the theorem applies.

1.4.3. *Maximum over the Euclidean ball.* In the previous section, we showed an example of an infinite family of sub-Gaussian random variables for which we can control the maximum, since it is the same as the maximum of a related finite family. However, for some set $K$ that has infinitely many extreme points, the same approach would not work. We can nevertheless create a finite family whose maximum approximates the original maximum.

Definition 1.5. Let $K \subset \mathbf{R}^d$ and $\varepsilon > 0$. A set $\mathcal{N}$ is an $\varepsilon$-net of $K$ with respect to a distance $d(\cdot,\cdot)$ on $\mathbf{R}^d$ if $\mathcal{N} \subset K$ and for all $z \in K$, there exists $x \in \mathcal{N}$ such that $d(x,z) \le \varepsilon$

Lemma 1.2. *Let $K$ be a compact subset of $\mathbf{R}^d$, and $\mathcal{N}_\varepsilon$ be an $\varepsilon$-net of $K$ for the Euclidean norm. For all $X \in \mathbf{R}^d$, it holds that*

$$\max_{v \in K} v^\top X \le \max_{z \in \mathcal{N}_\varepsilon} z^\top X + \varepsilon \max_{r \in \mathcal{B}_2} r^\top X\,.$$

Proof. For all $v \in K$, there exists $z \in \mathcal{N}_\varepsilon$ and $r \in \varepsilon\mathcal{B}_2$ such that $v = z + r$. As a consequence, it holds that

$$\begin{aligned}\max_{v \in K} v^\top X &\le \max_{z \in \mathcal{N}_\varepsilon} z^\top X + \max_{r \in \varepsilon\mathcal{B}_2} r^\top X \\ &\le \max_{z \in \mathcal{N}_\varepsilon} z^\top X + \varepsilon \max_{r \in \mathcal{B}_2} r^\top X\end{aligned}$$

□

We consider the case of the Euclidean unit ball in dimension $d$, denoted by $\mathcal{B}_2^d$. For this set, the lemma above yields that

$$\max_{v\in\mathcal{B}_2^d} v^\top X \le \max_{z\in\mathcal{N}_\varepsilon} z^\top X + \varepsilon \max_{r\in\mathcal{B}_2} r^\top X\,.$$

As a consequence, for any $\varepsilon$-net $\mathcal{N}_\varepsilon$ of $\mathcal{B}_2^d$, it holds that

$$\max_{v\in\mathcal{B}_2^d} v^\top X \le \frac{1}{1-\varepsilon}\max_{z\in\mathcal{N}_\varepsilon} z^\top X\,.$$

For a finite $\varepsilon$-net, we therefore have a finite family whose maximum approximates the maximum over $\mathcal{B}_2^d$. In order to obtain a quantitative bound, we can study the cardinality of such a set.

Lemma 1.3. *For all $\varepsilon \in (0,1)$, there exists an $\varepsilon$-net $\mathcal{N}_\varepsilon$ of $\mathcal{B}_2^d$ such that $|\mathcal{N}_\varepsilon| \le (3/\varepsilon)^d$.*

Proof. We build the set in an inductive manner: let $x_1 = 0$. For all $i \ge 2$, we take $x_i$ to be any $x \in \mathcal{B}_2^d$ such that $|x_i - x_j| > \varepsilon$ for all $j < i$. If the process stops, the $x_i$ form an $\varepsilon$-net, by definition.

At any step $N$, consider the balls centred at $x_i$, with radius $\frac{\varepsilon}{2}$, denoted by $x_i + \frac{\varepsilon}{2}\mathcal{B}_2^d$. By triangular inequality, we obtain that these balls are disjoint, and that they are all subsets of $\left(1+\frac{\varepsilon}{2}\right)\mathcal{B}_2^d$. Taking volumes, we therefore have that

$$\sum_{i=1}^N \mathrm{vol}\Big(x_i + \frac{\varepsilon}{2}\,\mathcal{B}_2^d\Big) = \mathrm{vol}\Big(\bigcup_{i=1}^N x_i + \frac{\varepsilon}{2}\,\mathcal{B}_2^d\Big) \le \mathrm{vol}\Big(\big(1+\frac{\varepsilon}{2}\big)\,\mathcal{B}_2^d\Big)\,.$$

As a consequence,

$$N\Big(\frac{\varepsilon}{2}\Big)^d \le \Big(1+\frac{\varepsilon}{2}\Big)^d\,,$$

so $N \le (1+2/\varepsilon)^d \le (3/\varepsilon)^d$. Therefore, the processes stops and provides an $\varepsilon$-net with cardinality less than $(3/\varepsilon)^d$. □

Combining these results, we obtain the following

Theorem 1.12. *Let $X$ be a random vector of $\mathbf{R}^d$ such that $X \in \mathsf{sG}_d(\sigma^2)$. It holds that*

$$\mathbf{E}[\max_{\theta\in\mathcal{B}_2} \theta^\top X] \le 4\sigma\sqrt{d}$$

*Moreover, for any $\delta > 0$, it holds with probability $1-\delta$ that*

$$\max_{\theta\in\mathcal{B}_2} \theta^\top X \le 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}$$

Proof. Let $\mathcal{N}_{1/2}$ be a 1/2-net of $\mathcal{B}_2^d$ such that $|\mathcal{N}_{1/2}| \le 6^d$. As noted above, it holds that

$$\max_{\theta\in\mathcal{B}_2} \theta^\top X \le 2\max_{z\in\mathcal{N}_{1/2}} z^\top X\,.$$

As a consequence, we have that

$$\mathbf{E}[\max_{\theta\in\mathcal{B}_2} \theta^\top X] \le 2\mathbf{E}[\max_{z\in\mathcal{N}_{1/2}} z^\top X] \le 2\sigma\sqrt{2\log(6^d)} \le 4\sigma\sqrt{d}\,.$$

Moreover, for all $t > 0$, it holds that

$$\mathbf{P}(\max_{\theta\in\mathcal{B}_2} \theta^\top X > t) \le 6^d e^{-\frac{t^2}{2\sigma^2}}\,.$$

Taking $t = 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}$ yields the desired result. □

For a more in-depth survey of this subject, we refer to the very exhaustive (Boucheron et al., 2013).

**2. Estimation and detection in high-dimension.** One of the important subjects in modern statistical theory is the study of problems where the dimension of the observation or parameter space $d$ is much larger than the number of samples $n$. If we make no further assumptions, the following example shows that this setting can be very problematic.

EXAMPLE 2.1. Let $\theta \in \mathbf{R}^d$, and $y_i = \theta^* + z_i$, for $1 \le i \le n$, where the $z_i \in \mathsf{sG}_d(\sigma^2)$ are independent. To estimate the parameter $\theta^*$, we consider the empirical mean of the $y_i$: $\hat\theta = \sum_{i=1}^n y_i/n$. Note that this estimator can be interpreted as the *empirical risk minimizer* for the $\ell_2$ loss, or equivalently the maximum likelihood estimator for Gaussian noise

$$\hat\theta \in \operatorname*{argmin}_{\theta \in \mathbf{R}^d} \sum_{i=1}^n \|\theta - y_i\|_2^2 \,.$$

In this case, it is easy to analyse the performance of this estimator, as we have a closed form expression for it

$$\hat\theta = \theta^* + \frac{1}{n}\sum_{i=1}^n z_i \,.$$

In the case where $z_i \sim \mathcal{N}(0, \sigma^2 I_d)$, we therefore have $\mathbf{E}[|\hat\theta - \theta^*|^2] = \sigma^2 \frac{d}{n}$. In the more general case, by Theorem 1.12, it holds with probability $1 - \delta$ that

$$\|\hat\theta - \theta^*\|_2 \le 4\sigma\sqrt{\frac{d}{n}} + 2\sigma\sqrt{\frac{2\log(1/\delta)}{n}} \,.$$

In a setting where $d \gg n$, the error bound is arbitrarily bad. Moreover, as will be shown later in the course, any estimator will suffer a loss of this order, it is not specific to that one. In order to overcome this *curse of dimensionality*, we can focus on problems where some information about the parameter $\theta^*$ is available.

2.1. *Constrained estimation.* We consider estimation problems of the type described above, with the additional constraint that $\theta^* \in \mathcal{S}$, for some know set $\mathcal{S} \subset \mathbf{R}^d$. For simplicity, we will directly consider averages of our observations and study cases with one observation, and $z \in \mathsf{sG}_d(\sigma^2/n)$. We take in all the following $\hat\theta \in \operatorname{argmin}_{\theta \in \mathcal{S}} \|y - \theta\|_2^2$. As above, this coincides with the empirical risk minimiser and the maximum likelihood estimator for Gaussian noise.

2.1.1. *General bounds.* The following proposition is purely deterministic, valid for any $z \in \mathbf{R}^d$

PROPOSITION 2.1. *Let $\theta^* \in \mathcal{S}$ and $y = \theta^* + z$. For $\hat\theta$ as defined above, it holds that*

$$\|\hat\theta - \theta^*\|^2 \le 2 \sup_{u \in \mathcal{S} - \mathcal{S}} \langle u, z \rangle$$
$$\|\hat\theta - \theta^*\| \le 2 \sup_{u \in \frac{\mathcal{S} - \mathcal{S}}{|\mathcal{S} - \mathcal{S}|}} \langle u, z \rangle \,,$$

*where $\mathcal{S} - \mathcal{S} = \{s - s' \mid s, s' \in \mathcal{S}\}$, and $\mathcal{A}/|\mathcal{A}| = \{a/|a| \mid a \in \mathcal{A}\}$.*

PROOF. We use the definition of $\hat\theta$ as a risk minimiser, as $\theta^* \in \mathcal{S}$

$$\begin{aligned} \|\hat\theta - y\|^2 &\le \|\theta^* - y\|^2 \\ \|(\hat\theta - \theta^*) + (\theta^* - y)\|^2 &\le \|\theta^* - y\|^2 \\ \|\hat\theta - \theta^*\|^2 &\le 2\langle \hat\theta - \theta^*, y - \theta^* \rangle \,. \end{aligned}$$

Applying directly this inequality by replacing $y - \theta^*$ by $z$ and $\theta - \theta^* \in \mathcal{S} - \mathcal{S}$ by the worst case, we obtain the first result. By dividing each side by $\|\hat{\theta} - \theta^*\|$, we have

$$\|\hat{\theta} - \theta^*\| \le 2\Big\langle \frac{\hat{\theta} - \theta^*}{\|\hat{\theta} - \theta^*\|}, y - \theta^* \Big\rangle \,.$$

Again, replacing $\frac{\hat{\theta}-\theta^*}{\|\hat{\theta}-\theta^*\|} \in \frac{\mathcal{S}-\mathcal{S}}{|\mathcal{S}-\mathcal{S}|}$ by the worst case, we obtain the desired inequality. □

2.1.2. *Examples.* The maximal deviation inequalities derived in the previous section can be applied to obtain estimation bounds over a variety of constraints sets. We will see that the choice of inequality to apply depends on the set $\mathcal{S}$ considered (in particular whether it is unbounded or not). In the following examples, we recall that $z \in \mathsf{sG}_d(\sigma^2/n)$. Further note that the bounds below are only derived in expectation, but can easily be extended to statements that hold with high probability.

Example 2.2 (No constraints). When $\mathcal{S} = \mathbf{R}^d$, note that the first bound of Proposition 2.1 is unbounded, and does not give any information. On the other hand, noting that $\mathcal{S} - \mathcal{S} = \mathbf{R}^d$ and $\frac{\mathcal{S}-\mathcal{S}}{|\mathcal{S}-\mathcal{S}|} = \mathcal{S}^{d-1} \subset \mathcal{B}_2^d$, we have

$$\mathbf{E}[\|\hat{\theta} - \theta^*\|] \le 2\,\mathbf{E}[\sup_{u \in \mathcal{B}_2^d} \langle u, z\rangle] \le 8\sigma\sqrt{\frac{d}{n}}\,.$$

Example 2.3 (Euclidean ball). Fix $\mathcal{S} = R\mathcal{B}_2^d$, for some radius $R > 0$. The second bound of the Proposition 2.1 yields the same guarantees as above, i.e. it is indifferent to $R$. In order to leverage this information, we use the first inequality, and obtain

$$\mathbf{E}[\|\hat{\theta} - \theta^*\|^2] \le 2\,\mathbf{E}[\sup_{u \in 2R\mathcal{B}_2^d} \langle u, z\rangle] \le 16\sigma R\sqrt{\frac{d}{n}}\,.$$

Example 2.4 (Linear subspace). Fix $\mathcal{S} = V$, for some radius linear subspace $V$ of $\mathbf{R}^d$, with $\dim(V) = k \ge 1$. As in the example with all of $\mathbf{R}^d$, the first bound is not useful. The second one yields

$$\mathbf{E}[\|\hat{\theta} - \theta^*\|] \le 2\,\mathbf{E}[\sup_{u \in V \cap \mathcal{B}_2^d} \langle u, z\rangle] \le 2\,\mathbf{E}[\sup_{u \in \mathcal{B}_2^k} \langle u, \Pi_V z\rangle] \le 8\sigma\sqrt{\frac{k}{n}}\,,$$

where $\Pi_V$ is the Euclidean projection on $V$, since $\Pi_V z \in \mathsf{sG}_k(\sigma^2/n)$.

Example 2.5 (Sparse vectors). We take to be $\mathcal{S}$ to be the set of vectors $\theta^*$ with at most $k$ nonzero coefficients. This is denoted $\|\theta\|_0 \le k$. This is a generalisation of the linear subspace case: the parameter $\theta^*$ belongs to one of the $k$-dimensional subspaces generated by $k$ elements of the canonical basis, but it is not know which one. By taking this point of view, the first bound is clearly not useful, and we use the second one

$$\mathbf{E}[\|\hat{\theta} - \theta^*\|] \le 2\,\mathbf{E}[\sup_{\substack{\|u\|_0 \le 2k \\ \|u\|_2 \le 1}} \langle u, z\rangle] \le 2\,\mathbf{E}[\max_{|S|=2k} \max_{u \in \mathcal{B}_2^S} \langle u, z\rangle] \le 4\sigma\sqrt{\frac{2\log(6^k\binom{d}{k})}{n}} \le 4\sigma\sqrt{\frac{2k\log(6ed/k)}{n}}\,.$$

It is interesting to compare the rates for the last two problems: not knowing the location of the $k$ coefficients makes us use a union bound over $6^k\binom{d}{k}$ events instead of $6^k$. This comes at a small price in the final rate: the "effective dimension" is multiplied by a term of order $\log(d/k)$.

2.1.3. *Convex sets.* The bounds above are very general, but better results can be obtained by making further assumptions on the constraint set. In this subsection, we consider the case where $\theta^* \in \mathcal{C}$, where $\mathcal{C}$ is a convex set. Aside from the stronger bounds found in the following proposition, it is also algorithmically efficient to project on convex sets.

Proposition 2.2. *Let $\theta^* \in \mathcal{C}$ and $y = \theta^* + z$. For $\hat\theta$ as defined above, it holds that*

$$\|\hat\theta - \theta^*\|^2 \le \sup_{u \in \mathcal{C}-\mathcal{C}} \langle u, z\rangle$$
$$\|\hat\theta - \theta^*\| \le \sup_{u \in T_{\mathcal{C}}(\theta^*) \cap \mathcal{B}_2^d} \langle u, z\rangle\,.$$

We recall the definition, for convex sets, of the tangent and normal cones.

Definition 2.1. Let $\mathcal{C}$ be a convex set, and $c \in \mathcal{C}$. The *tangent cone* to $\mathcal{C}$ at $x$ is defined as

$$T_{\mathcal{C}}(x) = \mathrm{cone}\{y - x \,|\, y \in \mathcal{C}\}\,.$$

The *normal cone* to $\mathcal{C}$ at $x$ is defined as

$$\mathcal{N}_{\mathcal{C}}(x) = \{h \in \mathbf{R}^d \,|\, \langle h, y - x\rangle \le 0 \;\forall\, y \in \mathcal{C}\}\,.$$

Note that $\mathcal{N}_{\mathcal{C}}(x) = T^*_{\mathcal{C}}(x)$.

Proof of Proposition 2.2. Note that $\hat\theta$ being the projection of $y$ on $\mathcal{C}$ ensures the crucial inequality $\langle y - \hat\theta, \theta^* - \hat\theta\rangle \le 0$. Indeed, consider $\theta_t = \hat\theta + t(\theta^* - \hat\theta)$, for $t \in [0,1]$ it holds that

$$\|\theta_t - y\|^2 = \|\hat\theta - y\|^2 + 2t\langle \hat\theta - y, \theta^* - \hat\theta\rangle + t^2\|\theta^* - \hat\theta\|^2\,.$$

Using the fact that $\|\theta_t - y\|^2 \ge \|\hat\theta - y\|^2$, dividing by $t > 0$ and letting it go to 0 yields the inequality. Taking $y = \theta^* + z$, we obtain

$$\|\hat\theta - \theta^*\|^2 \le \langle \hat\theta - \theta^*, z\rangle\,.$$

This yields the two desired inequalities, as in the proof of Proposition 2.1. □

Note that the bounds in the examples of the previous subsection yields better bounds when $\mathcal{S}$ is a convex set. As an example, we can obtain the following.

Example 2.6 ($\ell_1$ ball). Let $\mathcal{C} = R\mathcal{B}_1^d$. The first bound from the proposition above yields

$$\|\hat\theta - \theta^*\|^2 \le \sup_{u \in 2R\mathcal{B}_1^d} \langle u, z\rangle \le 2R \sup_{u \in \mathcal{B}_1^d} \langle u, z\rangle \le 2\sigma R\sqrt{\frac{\log(2d)}{n}}\,.$$

The last inequality is a result of the duality between $\ell_1$ and $\ell_\infty$ norms.

In order to reveal a useful connection between sparsity and the $\ell_1$ ball, we use the following lemma, given without proof

Lemma 2.1. *For a convex cone $K$, it holds that*

$$\sup_{u \in K \cap \mathcal{B}_2^d} \langle u, z\rangle = \mathrm{dist}(z, K^*)\,.$$

We study an example that highlights once more the links between the $\ell_1$ norm and sparsity:

Example 2.7. Let $\theta^*$ be a $k$-sparse vector (i.e. $|\theta^*|_0 \leq k$), such that $|\theta^*|_1 = 1$. When the sparsity $k$ is unknown, the second constraint can be exploited, and we can consider the estimator $\hat{\theta} \in \mathrm{argmin}_{\theta \in \mathcal{B}_1^d} \|y - \theta\|_2^2$. Unexpectedly, even though the value of $k$ is not used in the construction of the estimator, the error bounds will depend on $k$, in an *adaptive* manner. This is due to the local geometry of this convex set: the bound depends on $T_{\mathcal{B}_1^d}(\theta^*)$, and the geometry of this cone is dependent on $k$. For simplicity, we consider here that $z \sim \mathcal{N}(0, \sigma^2 I_d/n)$.

Proof. We have, by the result of Proposition 2.2, that

$$\|\hat{\theta} - \theta^*\| \leq \sup_{u \in T_{\mathcal{B}_1^d}(\theta^*) \cap \mathcal{B}_2^d} \langle u, z \rangle \,.$$

Furthermore, by Lemma 2.1, this yields

$$\|\hat{\theta} - \theta^*\| \leq \mathrm{dist}\big(z, \mathcal{N}_{\mathcal{B}_1^d}(\theta^*)\big) \,.$$

For $\theta^*$ that is $k$-sparse, this cone can be explicitly described. First, note that

$$\mathcal{N}_{\mathcal{B}_1^d}(\theta^*) = \{p \in \mathcal{R}^d \ : \ \forall\, \theta \in \mathcal{B}_1^d \,, \langle p, \theta - \theta^* \rangle \leq 0\} \,.$$

Let $S$ be the support of size $k$ of $\theta^*$. For any $i, j \in S$, consider $\theta = \theta^* + te_i - te_j$. For $t > 0$ small enough, $|\theta|_1 = |\theta^*|_1$. As a consequence, for any $p \in \mathcal{N}_{\mathcal{B}_1^d}(\theta^*)$, we have

$$\langle p, \theta - \theta^* \rangle = \langle p, te_i - te_j \rangle = p_i - p_j \leq 0 \,.$$

By changing the roles of $i$ and $j$, we obtain that $p_i = p_j$. For any $i \in S$ and $j \in S^c$, let $\theta = \theta^* - te_i \pm te_j$. For $t > 0$ small enough, $|\theta|_1 = |\theta^*|_1$, and we obtain

$$\langle p, \theta - \theta^* \rangle = \langle p, -te_i \pm te_j \rangle = -p_i \pm p_j \leq 0 \,.$$

Therefore, for any $p \in \mathcal{N}_{\mathcal{B}_1^d}(\theta^*)$, there exists $\lambda \geq 0$ such that

$$\begin{aligned} p_i &= \lambda \quad \text{for} \quad i \in S \\ |p_j| &\leq \lambda \quad \text{for} \quad i \notin S \,. \end{aligned}$$

This can be shown to be sufficient as well. As a consequence, we have that

$$\begin{aligned} \|\hat{\theta} - \theta^*\|^2 &\leq \mathrm{dist}\big(z, \mathcal{N}_{\mathcal{B}_1^d}(\theta^*)\big)^2 \\ &= \inf_{u \in \mathcal{N}_{\mathcal{B}_1^d}(\theta^*)} \|z - u\|_2^2 \\ &= \inf_{\substack{\lambda \geq 0 \\ |u_j| \leq \lambda, j \in S^c}} \sum_{i \in S} \big(z_i - \lambda \mathrm{sign}(\theta_i^*)\big)^2 + \sum_{j \in S^c} (z_j - u_j)^2 \\ &= \inf_{\lambda \geq 0} \sum_{i \in S} \big(z_i - \lambda \mathrm{sign}(\theta_i^*)\big)^2 + \sum_{j \in S^c} \mathrm{st}_\lambda(z_j)^2 \,, \end{aligned}$$

where $\mathrm{st}_\lambda$ is the *soft-thresholding* function such that

$$\mathrm{st}_\lambda(z) = \begin{cases} z - \lambda & \text{if } z < -\lambda, \\ 0 & \text{if } |z| \leq \lambda, \\ z + \lambda & \text{if } z > \lambda. \end{cases}$$

Therefore, taking expectations yields, for any $\lambda \geq 0$

$$\begin{aligned}\mathbf{E}[\|\hat{\theta} - \theta^*\|^2] &\leq \sigma^2 \frac{k}{n}(1+\lambda^2) + \mathbf{E}\big[\sum_{j \in S^c} \mathrm{st}_\lambda(z_j)^2\big] \\ &\leq \sigma^2 \frac{k}{n}(1+\lambda^2) + \sigma^2 \frac{d-k}{n} \frac{2}{\sqrt{2\pi}} \frac{1}{\lambda} \exp(-\lambda^2/2)\,,\end{aligned}$$

Taking $\lambda = \sqrt{2\log(d/s)}$ yields, after simplification

$$\mathbf{E}[\|\hat{\theta} - \theta^*\|^2] \leq 2\sigma^2 \frac{k \log(d/k)}{n} + \frac{5}{4}\sigma^2 \frac{k}{n}\,.$$

□

2.2. *Hypothesis testing.* Testing an hypothesis is a central idea of the scientific method. Checking wether an assumption is supported by the results of an experiment allows researchers to try new ideas, to gain a better understanding of their area of study. Hypothesis testing problems are the natural setting in statistical analysis in order to answer a yes-or-no question, by choosing the answer that is more likely, given the data.

From a mathematical point of view, a simple hypothesis testing problem, the goal is to identify the underlying distribution of a dataset. Given a random variable $X \in \mathcal{X}$ and two distributions $\mathbf{P}_0$ and $\mathbf{P}_1$ on $\mathcal{X}$, we aim to distinguish the two hypotheses

$$\begin{aligned} H_0 &: X \sim \mathbf{P}_0 \\ H_1 &: X \sim \mathbf{P}_1\,. \end{aligned}$$

A test is a measurable function of the data $\psi : \mathcal{X} \mapsto \{0,1\}$ that indicates wether the instance was generated with distribution $\mathbf{P}_0$ or $\mathbf{P}_1$. We define the probability of error of a test as the maximum of the probabilities of type I (mistakenly rejecting the null hypothesis) and type II error (mistakenly accepting the null hypothesis), formally, we say that a test has a probability of error less than $\delta \in (0,1)$ if

$$\mathbf{P}_0(\psi(X) = 1) \vee \mathbf{P}_1(\psi(X) = 0) \leq \delta\,.$$

In the case of problems with composite hypotheses, of the type

$$\begin{aligned} H_0 &: X \sim \mathbf{P}_0\,,\ \mathbf{P}_0 \in \mathcal{P}_0 \\ H_1 &: X \sim \mathbf{P}_1\,,\ \mathbf{P}_1 \in \mathcal{P}_1\,, \end{aligned}$$

where $\mathcal{P}_0$ and $\mathcal{P}_1$ are disjoint sets of distributions, this would be replaced by

$$\sup_{\mathbf{P}_0 \in \mathcal{P}_0} \mathbf{P}_0(\psi(X) = 1) \vee \sup_{\mathbf{P}_1 \in \mathcal{P}_1} \mathbf{P}_1(\psi(X) = 0) \leq \delta\,.$$

In particular, we will consider the case of detecting means of a certain type in a high-dimensional sub-Gaussian vector. Let $Z \in \mathsf{sG}_d(\sigma^2)$, and consider for some $\mathcal{C} \subset \mathbf{R}^d$ such that $0 \notin \mathcal{C}$.

$$\begin{aligned} H_0 &: X = Z \\ H_1 &: X = \mu + Z,\ \mu \in \mathcal{C}\,. \end{aligned}$$

2.2.1. *Statistics for testing.* We recall that any measurable function of the data is called a *statistic*. The following proposition shows how judiciously chosen statistics can be helpful to design tests.

Proposition 2.3. *Let $\varphi : \mathcal{X} \to \mathbf{R}$ such that there exists $\tau \in \mathbf{R}$ for which*

$$\begin{aligned} \mathbf{P}_0(\varphi(X) > \tau) \le \delta \,, & \quad \textit{for all} \quad \mathbf{P}_0 \in \mathcal{P}_0 \\ \mathbf{P}_1(\varphi(X) \le \tau) \le \delta \,, & \quad \textit{for all} \quad \mathbf{P}_1 \in \mathcal{P}_1 \,. \end{aligned}$$

*The test $\psi$ defined as*

$$\psi(X) = \mathbf{1}\{\varphi(X) > \tau\}$$

*has a probability of error less than $\delta$.*

This is a direct consequence of the definitions above.

In the problem of distinguishing $\mu = 0$ from $\mu \in \mathcal{C}$, one can consider, for some set $E \subset \mathbf{R}^d$, the statistic $\varphi(X) = \max_{u \in E} \langle u, X\rangle$, particularly useful whenever $\varphi(\mu)$ is constant for all $\mu \in \mathcal{S}$ (e.g., for $\mathcal{C} = \mathcal{S}^{d-1}$, $E = \mathcal{B}_2^d$).

Proposition 2.4. *Whenever $\max_{u \in E}\langle u, \mu\rangle = t_{\mathcal{C}}$, for all $\mu \in \mathcal{C}$, if $t_{\mathcal{C}} \ge 2 \max_{u\in E}\langle u, Z\rangle$ with probability $1-\delta$, then the test $\psi$ defined by*

$$\psi(X) = \mathbf{1}\{\max_{u \in E}\langle u, X\rangle > \frac{t_{\mathcal{C}}}{2}\}$$

*has a probability of error less than $\delta$.*

We study in the following subsection some examples, for specific choices of $\mathcal{C}$.

2.2.2. *Examples.* We will consider cases where

$$\mathcal{C} = \{\mu \mathbf{1}_S \,,\; S \in \mathcal{S}\}\,,$$

for $\mu > 0$ and some class $\mathcal{S}$ of $k$-sets (subsets of $\{1, \dots, d\}$ of size $k$). For more information on this problem, see Addario-Berry et al. (2010).

Example 2.8. Let $\mathcal{S}$ be the class of all $k$-sets. We take $E = \{\mathbf{1}_S \,,\; S \in \mathcal{S}\}$, which is a finite set with cardinality $|E| = \binom{d}{k}$. Note that for all $\mu \in \mathcal{C}$, $\max_{u\in E}\langle u, \mu\rangle = t_{\mathcal{C}} = \mu k$, and that for all $u \in E$, $|u|_2^2 = k$. We derive a bound one of the quantities of importance in this problem

$$\max_{u \in E}\langle u, Z\rangle = \sqrt{k} \max_{u \in E/\sqrt{k}}\langle u, Z\rangle$$

for which it holds that

$$\mathbf{P}_0(\max_{u \in E/\sqrt{k}}\langle u, Z\rangle > t) \le \binom{d}{k} e^{-t^2/2}\,.$$

As a consequence, applying Proposition 2.4, it is possible to distinguish the two hypotheses with probability of error less than $\delta$ whenever

$$\mu \ge 2\sqrt{2\log(ed/k)} + 2\sqrt{\frac{2}{k}\log(1/\delta)}\,.$$

See exercises for more examples, related to graphs.

**3. Information-theoretic lower bounds.** All of the results stated so far have been *upper bounds* on the statistical risk (expected distance from the true signal, probability of error, etc) of estimation or testing procedures. They have been positive results, stating what is possible to achieve in a statistical problem. Some of the results indicate that these bounds should be dependent on some parameters of the problem: the sample size $n$, the dimension $d$ of the problem, etc. It is not clear however if this is an artefact of our analysis, or if this is a consequence of some intrinsic difficulty of the problem. In order to address these questions, we will study *lower bounds* on the statistical risk, by showing that some problems are hard for all estimators. This is mostly based on an analysis of distances between distributions, using tools and methods from information theory.

3.1. *Bounds for hypothesis testing.* A simple starting observation is that there is a one to one correspondance between tests and events over the probability space: it is equivalent to define $\psi$ or $A = \psi^{-1}(\{1\})$ and $A^c = \psi^{-1}(\{0\})$. Furthermore, note that for an hypothesis testing problem over $\mathbf{P}_0$ and $\mathbf{P}_1$, we have

$$\begin{aligned}\mathbf{P}_0(\Psi(X)=1) \vee \mathbf{P}_1(\Psi(X)=0) &\geq \frac{1}{2}\big(\mathbf{P}_0(\Psi(X)=1) + \mathbf{P}_1(\Psi(X)=0)\big)\\ &\geq \frac{1}{2}\big(1 - (\mathbf{P}_1(\Psi(X)=1) - \mathbf{P}_0(\Psi(X)=1))\big)\\ &\geq \frac{1-(\mathbf{P}_1(A)-\mathbf{P}_0(A))}{2}\,.\end{aligned}$$

This term is related to an upper bound for the probability of error, as

$$\mathbf{P}_0(\Psi(X)=1) \vee \mathbf{P}_1(\Psi(X)=0) \leq \mathbf{P}_0(\Psi(X)=1) + \mathbf{P}_1(\Psi(X)=0) = 1 - (\mathbf{P}_1(A) - \mathbf{P}_0(A))\,.$$

Evidently, we wish for the event $A$ to have very different probabilities under $\mathbf{P}_0$ and $\mathbf{P}_1$. The natural question is: *How different can it be?* In order to answer these types of question, it is very natural to introduce a new tool, the *total variation distance*, which can be defined by maximising the difference $\mathbf{P}_1(A) - \mathbf{P}_0(A)$.

3.1.1. *Total variation distance.* The natural one-to-one link between tests and events on the sample space can be exploited to quantify the distance between two distributions.

Definition 3.1. For two distributions $\mathbf{P}$ and $\mathbf{Q}$ on a probability space $(\Omega, \mathcal{F})$, the total variation distance $d_{\mathsf{TV}}$ is defined as

$$d_{\mathsf{TV}}(\mathbf{P},\mathbf{Q}) = \sup_{A\in\mathcal{F}} |\mathbf{P}(A) - \mathbf{Q}(A)|\,.$$

As seen above, this definition relates distance between distributions and testing error. This is described formally by the Neyman-Pearson lemma. To clarify the notations, we consider two distributions $\mathbf{P}_0$ and $\mathbf{P}_1$ and a sigma finite measure $\nu$ such that $\mathbf{P}_0 \gg \nu$ and $\mathbf{P}_1 \gg \nu$ (e.g. take $\nu = \mathbf{P}_0 + \mathbf{P}_1$). We can then take $p_0$ and $p_1$ to be the Randon-Nykodym derivatives of these probability measures with respect to $\nu$. For any function $f$, we therefore write

$$\int f = \int f(x) d\nu$$

Lemma 3.1 (Neyman-Pearson). *Let $\mathbf{P}_0$ and $\mathbf{P}_1$ be two probability measures. Then for any test $\psi$, it holds*

$$\mathbf{P}_0(\psi = 1) + \mathbf{P}_1(\psi = 0) \geq \int \min(p_0, p_1)$$

*Moreover, equality holds for the Likelihood Ratio test* $\psi^\star = \mathbf{1}\{p_1 \geq p_0\}$.

Proof. Observe first that

$$\begin{aligned}
\mathbf{P}_0(\psi^\star = 1) + \mathbf{P}_1(\psi^\star = 0) &= \int_{\psi^*=1} p_0 + \int_{\psi^*=0} p_1 \\
&= \int_{p_1 \geq p_0} p_0 + \int_{p_1 < p_0} p_1 \\
&= \int_{p_1 \geq p_0} \min(p_0, p_1) + \int_{p_1 < p_0} \min(p_0, p_1) \\
&= \int \min(p_0, p_1)\,.
\end{aligned}$$

Next for any test $\psi$, define its rejection region $R = \{\psi = 1\}$. Let $R^\star = \{p_1 \geq p_0\}$ denote the rejection region of the likelihood ration test $\psi^\star$. It holds

$$\begin{aligned}
\mathbf{P}_0(\psi = 1) + \mathbf{P}_1(\psi = 0) &= 1 + \mathbf{P}_0(R) - \mathbf{P}_1(R) \\
&= 1 + \int_R p_0 - p_1 \\
&= 1 + \int_{R \cap R^\star} p_0 - p_1 + \int_{R \cap (R^\star)^c} p_0 - p_1 \\
&= 1 - \int_{R \cap R^\star} |p_0 - p_1| + \int_{R \cap (R^\star)^c} |p_0 - p_1| \\
&= 1 + \int |p_0 - p_1| \big[\mathbf{1}\{R \cap (R^\star)^c\} - \mathbf{1}\{R \cap R^\star\}\big]
\end{aligned}$$

The above quantity is clearly minimized for $R = R^\star$. □

We obtain the following property

Proposition 3.1. *The following definitions are equivalent*

$$\begin{aligned}
d_{\mathsf{TV}}(\mathbf{P}_0, \mathbf{P}_1) &= \sup_{R \in \mathcal{A}} |\mathbf{P}_0(R) - \mathbf{P}_1(R)| && (i) \\
&= \sup_{A \in \mathcal{F}} \Big| \int_R p_0 - p_1 \Big| && (ii) \\
&= \frac{1}{2} \int |p_0 - p_1| && (iii) \\
&= 1 - \int \min(p_0, p_1) && (iv) \\
&= 1 - \inf_{\psi} \big[\mathbf{P}_0(\psi = 1) + \mathbf{P}_1(\psi = 0)\big] && (v)
\end{aligned}$$

*where the infimum above is taken over all tests.*

Proof. Clearly $(i)$ and two are equivalent $(ii)$ and the Neyman-Pearson Lemma gives the same for $(iv)$ and $(v)$. Moreover, by identifying a test $\psi$ to its rejection region, we obtain equivalence between $(i)$ and $(v)$. Therefore it remains only to show that $(iii)$ is equal to any of the other expressions.

Hereafter, we show that $(iii) = (iv)$. To that end, observe that

$$\begin{aligned}\int |p_0 - p_1| &= \int_{p_1 \geq p_0} p_1 - p_0 + \int_{p_1 < p_0} p_0 - p_1 \\ &= \int_{p_1 \geq p_0} p_1 + \int_{p_1 < p_0} p_0 - \int \min(p_0, p_1) \\ &= 1 - \int_{p_1 < p_0} p_1 + 1 - \int_{p_1 \geq p_0} p_0 - \int \min(p_0, p_1) \\ &= 2 - 2 \int \min(p_0, p_1)\end{aligned}$$

□

These definitions are more intuitive if one keeps in mind the following picture, with densities $p_0$ and $p_1$ represented over $\mathbf{R}$.

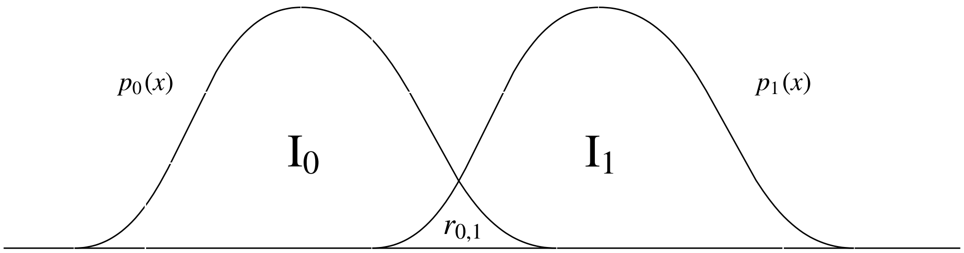

FIGURE 1. *The three zones are such that $I_0 + r_{0,1} = I_1 + r_{0,1} = 1$, and $r_{0,1} = \int \min(p_0, p_1)$. The total variation distance therefore satisfies $d_{\mathsf{TV}}(\mathbf{P}_0, \mathbf{P}_1) = I_0 = I_1 = 1 - r_{0,1}$.*

3.1.2. *Divergence between distributions.* The total variation distance between two distributions is often difficult to compute. Particularly, when the data is $n$ samples of i.i.d. variables, i.e. when $\mathbf{P} = \mathbf{Q}^{\otimes n}$ for some distribution $\mathbf{Q}$, there is no general way to relate $d_{\mathsf{TV}}(\mathbf{Q}_0^{\otimes n}, \mathbf{Q}_1^{\otimes n})$ to $d_{\mathsf{TV}}(\mathbf{Q}_0, \mathbf{Q}_1)$. This setting is very common in statistics, and other distances have been employed to address this issue. When $\mathbf{P}_1 \ll \mathbf{P}_0$, the total variation distance can be expressed as

$$d_{\mathsf{TV}}(\mathbf{P}_1, \mathbf{P}_0) = \frac{1}{2} \mathbf{E}_{\mathbf{P}_0} \Big[ \Big| \frac{\mathrm{d}\,\mathbf{P}_1}{\mathrm{d}\,\mathbf{P}_0} - 1 \Big| \Big] .$$

This a special case of an $f$-divergence (Csiszár, 1963), defined for any convex function on $(0, +\infty)$ such that $f(1) = 0$ as

$$D_f(\mathbf{P}_1, \mathbf{P}_0) = \mathbf{E}_{\mathbf{P}_0} \Big[ f\Big( \frac{\mathrm{d}\,\mathbf{P}_1}{\mathrm{d}\,\mathbf{P}_0} \Big) \Big] .$$

For the total variation distance, the choice of function is $f(t) = |t - 1|/2$.

DEFINITION 3.2. For different choices of divergence function $f$, the following distances

- The Hellinger distance $\mathsf{H}(\mathbf{P}_0, \mathbf{P}_1)$ coincides with the choice $f(t) = (\sqrt{t} - 1)^2$

$$\mathsf{H}(\mathbf{P}_1, \mathbf{P}_0) = \mathbf{E}_{\mathbf{P}_0}\Big[\Big(\sqrt{\frac{\mathrm{d}\,\mathbf{P}_1}{\mathrm{d}\,\mathbf{P}_0}} - 1\Big)^2\Big].$$

- The Kullback-Leibler divergence $\mathsf{KL}(\mathbf{P}_0, \mathbf{P}_1)$ coincides with the choice $f(t) = t\log(t)$

$$\mathsf{KL}(\mathbf{P}_1, \mathbf{P}_0) = \mathbf{E}_{\mathbf{P}_0}\Big[\frac{\mathrm{d}\,\mathbf{P}_1}{\mathrm{d}\,\mathbf{P}_0}\log\Big(\frac{\mathrm{d}\,\mathbf{P}_1}{\mathrm{d}\,\mathbf{P}_0}\Big)\Big].$$

- The $\chi^2$ divergence $\chi^2(\mathbf{P}_0, \mathbf{P}_1)$ coincides with the choice $f(t) = (t-1)^2$

$$\chi^2(\mathbf{P}_1, \mathbf{P}_0) = \mathbf{E}_{\mathbf{P}_0}\Big[\Big(\frac{\mathrm{d}\,\mathbf{P}_1}{\mathrm{d}\,\mathbf{P}_0} - 1\Big)^2\Big].$$

These divergences are always nonnegative, and equal to 0 if and only if $\mathbf{P}_0 = \mathbf{P}_1$ almost surely. This property that they share with the total variation distance is directly true for the Hellinger and $\chi^2$ divergence, as for any function $f$ that is positive except in 1. We study first the Kullback-Leibler divergence, its properties, and how it relates to the total variation distance

3.1.3. *Kullback-Leibler divergence.*

Definition 3.3. The Kullback-Leibler divergence between probability measures $\mathbf{P}_1$ and $\mathbf{P}_0$ is equivalently given by

$$\mathsf{KL}(\mathbf{P}_1, \mathbf{P}_0) = \begin{cases} \displaystyle\int \log\Big(\frac{\mathrm{d}\mathbf{P}_1}{\mathrm{d}\mathbf{P}_0}\Big)\mathrm{d}\mathbf{P}_1\,, & \text{if } \mathbf{P}_1 \ll \mathbf{P}_0 \\ \infty\,, & \text{otherwise}\,. \end{cases}$$

It can be shown Tsybakov (2009) that the integral is always well defined when $\mathbf{P}_1 \ll \mathbf{P}_0$ (though it can be equal to $+\infty$ even in this case). Unlike the total variation distance, the Kullback-Leibler divergence is not a distance. Actually, it is not even symmetric. Nevertheless, it enjoys properties that are very useful for our purposes.

Proposition 3.2. *Let $\mathbf{P}$ and $\mathbf{Q}$ be two probability measures. Then*

1. $\mathsf{KL}(\mathbf{P}, \mathbf{Q}) \geq 0$
2. *If $\mathbf{P}$ and $\mathbf{Q}$ are product measures, i.e.,*

$$\mathbf{P} = \bigotimes_{i=1}^n \mathbf{P}_i \quad \textit{and} \quad \mathbf{Q} = \bigotimes_{i=1}^n \mathbf{Q}_i$$

*then*

$$\mathsf{KL}(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^n \mathsf{KL}(\mathbf{P}_i, \mathbf{Q}_i)\,.$$

Proof. If $\mathbf{P}$ is not absolutely continuous then the result is trivial. Next, assume that $\mathbf{P} \ll \mathbf{Q}$ and let $X \sim \mathbf{P}$.

1. Observe that by Jensen's inequality,

$$\mathsf{KL}(\mathbf{P}, \mathbf{Q}) = -\mathbf{E}\log\Big(\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\mathbf{P}}(X)\Big) \geq -\log\mathbf{E}\Big(\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\mathbf{P}}(X)\Big) = -\log(1) = 0\,.$$

2. Note that if $X = (X_1, \dots, X_n)$,

$$\begin{aligned}
\mathsf{KL}(\mathbf{P}, \mathbf{Q}) &= \mathbf{E} \log \Big( \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}(X) \Big) \\
&= \sum_{i=1}^{n} \int \log \Big( \frac{\mathrm{d}\mathbf{P}_i}{\mathrm{d}\mathbf{Q}_i}(X_i) \Big) \mathrm{d}\mathbf{P}_1(X_1) \cdots \mathrm{d}\mathbf{P}_n(X_n) \\
&= \sum_{i=1}^{n} \int \log \Big( \frac{\mathrm{d}\mathbf{P}_i}{\mathrm{d}\mathbf{Q}_i}(X_i) \Big) \mathrm{d}\mathbf{P}_i(X_i) \\
&= \sum_{i=1}^{n} \mathsf{KL}(\mathbf{P}_i, \mathbf{Q}_i)
\end{aligned}$$

□

Point 2. in Proposition 3.2 is particularly useful in statistics where observations typically consist of $n$ independent random variables.

EXAMPLE 3.1. For any $\theta \in \mathbf{R}^d$, let $\mathbf{P}_\theta$ denote the distribution of $\mathbf{Y} \sim \mathcal{N}(\theta, \sigma^2 I_d)$. Then

$$\mathsf{KL}(\mathbf{P}_\theta, \mathbf{P}_{\theta'}) = \sum_{i=1}^{d} \frac{(\theta_i - \theta_i')^2}{2\sigma^2} = \frac{|\theta - \theta'|_2^2}{2\sigma^2} \,.$$

PROOF. We have that

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{P}_\theta}{\mathrm{d}\mathbf{P}_{\theta'}}(X) &= \exp\Big( - \frac{1}{2\sigma^2}|X - \theta|_2^2 + \frac{1}{2\sigma^2}|X - \theta'|_2^2 \Big) \\
&= \exp\Big( - \frac{1}{2\sigma^2}\big(|\theta|_2^2 - |\theta'|_2^2 - 2\langle X, \theta - \theta'\rangle\big) \Big)
\end{aligned}$$

By definition of the divergence, we obtain that

$$\mathsf{KL}(\mathbf{P}_\theta, \mathbf{P}_{\theta'}) = \mathbf{E}_\theta\Big[ - \frac{1}{2\sigma^2}\big(|\theta|_2^2 - |\theta'|_2^2 - 2\langle X, \theta - \theta'\rangle\big)\Big] = -\frac{1}{2\sigma^2}\big( - |\theta|_2^2 - |\theta'|_2^2 + 2\langle \theta, \theta'\rangle\big) \,,$$

hence the desired result. □

The Kullback-Leibler divergence is easier to manipulate than the total variation distance but only the latter is related to the minimax probability of error. Fortunately, these two quantities can be compared using Pinsker's inequality. We prove here a slightly weaker version of Pinsker's inequality that will be sufficient for our purpose. For a stronger statement, see Tsybakov (2009), Lemma 2.5.

LEMMA 3.2 (Pinsker's inequality.). *Let* $\mathbf{Q}$ *and* $\mathbf{Q}$ *be two probability measures such that* $\mathbf{P} \ll \mathbf{Q}$. *Then*

$$d_{\mathsf{TV}}(\mathbf{P}, \mathbf{Q}) \le \sqrt{\mathsf{KL}(\mathbf{P}, \mathbf{Q})} \,.$$

Proof. Note that

$$\begin{aligned}\mathsf{KL}(\mathbf{P},\mathbf{Q}) &= \int_{pq>0} p\log\Big(\frac{p}{q}\Big)\\ &= -2\int_{pq>0} p\log\Big(\sqrt{\frac{q}{p}}\Big)\\ &= -2\int_{pq>0} p\log\Big(\Big[\sqrt{\frac{q}{p}}-1\Big]+1\Big)\\ &\ge -2\int_{pq>0} p\Big[\sqrt{\frac{q}{p}}-1\Big] \qquad \text{(by Jensen)}\\ &= 2-2\int\sqrt{pq}\end{aligned}$$

Next, note that

$$\begin{aligned}\Big(\int\sqrt{pq}\Big)^2 &= \Big(\int\sqrt{\max(p,q)\min(p,q)}\Big)^2\\ &\le \int\max(p,q)\int\min(p,q) \qquad \text{(by Cauchy-Schwarz)}\\ &= \Big[2-\int\min(p,q)\Big]\int\min(p,q)\\ &= \big(1+d_{\mathsf{TV}}(\mathbf{P},\mathbf{Q})\big)\big(1-d_{\mathsf{TV}}(\mathbf{P},\mathbf{Q})\big)\\ &= 1-d_{\mathsf{TV}}(\mathbf{P},\mathbf{Q})^2\end{aligned}$$

The two displays yield

$$\mathsf{KL}(\mathbf{P},\mathbf{Q}) \ge 2-2\sqrt{1-d_{\mathsf{TV}}(\mathbf{P},\mathbf{Q})^2} \ge d_{\mathsf{TV}}(\mathbf{P},\mathbf{Q})^2\,,$$

where we used the fact that $0\le d_{\mathsf{TV}}(\mathbf{P},\mathbf{Q})\le 1$ and $\sqrt{1-x}\le 1-x/2$ for $x\in[0,1]$. □

More information on the subjects of information theory and distances, divergences between distributions can be found in Tsybakov (2009); Csiszár and Körner (2011); Cover and Thomas (1991).

3.2. *Bounds for estimation.* The problem of estimating a parameter is intuitively harder than deciding between two values for this parameter. This can be argued formally in the following way, by *reducing* the problem of estimation to one of hypothesis testing: if there exists an estimator $\hat\theta$ of some parameter $\theta$ such that $|\hat\theta-\theta|\le r$ with probability $1-\alpha$, for some $r>0$ and $\alpha\in(0,1)$, then it is possible to distinguish the hypotheses $\theta=\theta_0$ and $\theta=\theta_1$ whenever $|\theta_1-\theta_0|>2r$. If this has already been proven to be impossible, we obtain a lower bound on the statistical performance of any estimator for $\theta$.

This simple argument can be made precise using the formalism of *statistical hypothesis testing*. To do so, we reduce our estimation problem to a testing problem. The reduction consists of two steps.

1. **Reduction to a finite number of *hypotheses*.** In this step the goal is to find the largest possible number of hypotheses $\theta_1,\dots,\theta_M\in\Theta$ under the constraint that

$$|\theta_j-\theta_k|_2^2 \ge 4\phi(\Theta)\,. \tag{1}$$

This problem boils down to a *packing* of the set $\Theta$. We can use the following trivial observations:

$$\inf_{\hat\theta}\sup_{\theta\in\Theta}\mathbf{P}_\theta\big[|\hat\theta-\theta|_2^2>\phi(\Theta)\big] \ge \inf_{\hat\theta}\max_{1\le j\le M}\mathbf{P}_{\theta_j}\big[|\hat\theta-\theta_j|_2^2>\phi(\Theta)\big]\,.$$

2. **Reduction to a testing problem.** In this second step, the necessity of the constraint (1) becomes apparent.
For any estimator $\hat{\theta}$, define the minimum distance test $\psi(\hat{\theta})$ that is associated to it by
$$\psi(\hat{\theta}) = \operatorname*{argmin}_{1 \le j \le M} |\hat{\theta} - \theta_j|_2 \,,$$
with ties broken arbitrarily.
Next observe that if, for some $j = 1, \dots, M$, $\psi(\hat{\theta}) \neq j$, then there exists $k \neq j$ such that $|\hat{\theta} - \theta_k|_2 \le |\hat{\theta} - \theta_j|_2$. Together with the reverse triangle inequality it yields
$$|\hat{\theta} - \theta_j|_2 \ge |\theta_j - \theta_k|_2 - |\hat{\theta} - \theta_k|_2 \ge |\theta_j - \theta_k|_2 - |\hat{\theta} - \theta_j|_2$$
so that
$$|\hat{\theta} - \theta_j|_2 \ge \frac{1}{2}|\theta_j - \theta_k|_2$$
Together with constraint (1), it yields
$$|\hat{\theta} - \theta_j|_2^2 \ge \phi(\Theta)$$
As a result,
$$\begin{aligned}\inf_{\hat{\theta}} \max_{1 \le j \le M} \mathbf{P}_{\theta_j}\big[|\hat{\theta} - \theta_j|_2^2 > \phi(\Theta)\big] &\ge \inf_{\hat{\theta}} \max_{1 \le j \le M} \mathbf{P}_{\theta_j}\big[\psi(\hat{\theta}) \neq j\big] \\ &\ge \inf_{\psi} \max_{1 \le j \le M} \mathbf{P}_{\theta_j}\big[\psi \neq j\big]\end{aligned}$$
where the infimum is taken over all tests based on $\mathbf{Y}$ and that take values in $\{1, \dots, M\}$.

Conclusion: it is sufficient for proving lower bounds to find $\theta_1, \dots, \theta_M \in \Theta$ such that $|\theta_j - \theta_k|_2^2 \ge 4\phi(\Theta)$ and
$$\inf_{\psi} \max_{1 \le j \le M} \mathbf{P}_{\theta_j}\big[\psi \neq j\big] \ge C' \,.$$
The above quantity is called *minimax probability of error*. In the next sections, we show how it can be bounded from below using arguments from information theory. For the purpose of illustration, we begin with the simple case where $M = 2$ in the next section.

3.2.1. *Fano's lemma.* The reduction to hypothesis testing from this section allows us to use more than two hypotheses. Specifically, we should find $\theta_1, \dots, \theta_M$ such that
$$\inf_{\psi} \max_{1 \le j \le M} \mathbf{P}_{\theta_j}\big[\psi \neq j\big] \ge C' \,,$$
for some positive constant $C'$. Unfortunately, the Neyman-Pearson Lemma no longer exists for more than two hypotheses. Nevertheless, it is possible to relate the minimax probability of error directly to the Kullback-Leibler divergence, without involving the total variation distance. This is possible using a well known result from information theory called *Fano's inequality*, that takes into account information about many hypotheses. We use it in a form that is tailored to our purposes and that is due to Lucien Birgé (Birgé, 1983) and builds upon an original result in (Ibragimov and Hasminskii, 1981).

Theorem 3.1 (Fano's inequality). *Let $P_1, \dots, P_M, M \ge 2$ be probability distributions such that $P_j \ll P_k$, $\forall\, j, k$. Then*
$$\inf_{\psi} \max_{1 \le j \le M} P_j\big[\psi(X) \neq j\big] \ge 1 - \frac{\frac{1}{M^2}\sum_{j,k=1}^{M} \mathsf{KL}(P_j, P_k) + \log 2}{\log(M-1)}$$
*where the infimum is taken over all tests with values in $\{1, \dots, M\}$.*

Proof. Let $Z \in \{1, \dots, M\}$ be a random variable such that $\mathbf{P}(Z = i) = 1/M$ and let $X \sim P_Z$. Note that $P_Z$ is a *mixture distribution* so that for any measure $\nu$ such that $P_Z \ll \nu$, we have

$$\frac{\mathrm{d}P_Z}{\mathrm{d}\nu} = \frac{1}{M} \sum_{j=1}^{M} \frac{\mathrm{d}P_j}{\mathrm{d}\nu}\,.$$

For all test $\psi$, we have

$$\begin{aligned}
&\sum_{j=1}^{M} \mathbf{P}(Z = j|X) \log[\mathbf{P}(Z = j|X)] = \\
&= \mathbf{P}(Z = \psi(X)|X) \log[\mathbf{P}(Z = \psi(X)|X)] + \sum_{j \neq \psi(X)} \mathbf{P}(Z = j|X) \log[\mathbf{P}(Z = j|X)] \\
&= (1 - \mathbf{P}(Z \neq \psi(X)|X)) \log[1 - \mathbf{P}(Z \neq \psi(X)|X)] \\
&\quad + \mathbf{P}(Z \neq \psi(X)|X) \sum_{j \neq \psi(X)} \frac{\mathbf{P}(Z = j|X)}{\mathbf{P}(Z \neq \psi(X)|X)} \log\Big[\frac{\mathbf{P}(Z = j|X)}{\mathbf{P}(Z \neq \psi(X)|X)}\Big] \\
&\quad + \mathbf{P}(Z \neq \psi(X)|X) \log[\mathbf{P}(Z \neq \psi(X)|X)] \\
&= h(\mathbf{P}(Z \neq \psi(X)|X)) + \mathbf{P}(Z \neq \psi(X)|X) \sum_{j \neq \psi(X)} q_j \log(q_j)\,,
\end{aligned}$$

where

$$h(x) = x \log(x) + (1 - x) \log(1 - x)$$

and

$$q_j = \frac{\mathbf{P}(Z = j|X)}{\mathbf{P}(Z \neq \psi(X)|X)}$$

is such that $q_j \geq 0$ and $\sum_{j \neq \psi(X)} q_j = 1$. It implies by Jensen's inequality that

$$\sum_{j \neq \psi(X)} q_j \log(q_j) = -\sum_{j \neq \psi(X)} q_j \log\Big(\frac{1}{q_j}\Big) \geq -\log\Big(\sum_{j \neq \psi(X)} \frac{q_j}{q_j}\Big) = -\log(M - 1)\,.$$

By the same convexity argument, we get $h(x) \geq -\log 2$. It yields

$$\sum_{j=1}^{M} \mathbf{P}(Z = j|X) \log[\mathbf{P}(Z = j|X)] \geq -\log 2 - \mathbf{P}(Z \neq \psi(X)|X) \log(M - 1)\,. \tag{2}$$

Next, observe that since $X \sim P_Z$, the random variable $\mathbf{P}(Z = j|X)$ satisfies

$$\mathbf{P}(Z = j|X) = \frac{1}{M} \frac{\mathrm{d}P_j}{\mathrm{d}P_Z}(X) = \frac{\mathrm{d}P_j(X)}{\sum_{k=1}^{M} \mathrm{d}P_k(X)}$$

It implies

$$
\begin{aligned}
&\int \Big\{ \sum_{j=1}^{M} \mathbf{P}(Z=j|X=x)\log[\mathbf{P}(Z=j|X=x)] \Big\} \mathrm{d}P_Z(x)\\
&= \sum_{j=1}^{M} \int \Big\{ \frac{1}{M}\frac{\mathrm{d}P_j}{\mathrm{d}P_Z}(x) \log\Big(\frac{1}{M}\frac{\mathrm{d}P_j}{\mathrm{d}P_Z}(x)\Big)\Big\}\mathrm{d}P_Z(x)\\
&= \frac{1}{M}\sum_{j=1}^{M}\int \log\Big(\frac{\mathrm{d}P_j(x)}{\sum_{k=1}^M \mathrm{d}P_k(x)}\Big)\mathrm{d}P_j(x)\\
&\le \frac{1}{M^2}\sum_{j,k=1}^{M}\int \log\Big(\frac{\mathrm{d}P_j(x)}{\mathrm{d}P_k(x)}\Big)\mathrm{d}P_j(x) - \log M \quad \text{(by Jensen)}\\
&= \frac{1}{M^2}\sum_{j,k=1}^{M} \mathsf{KL}(P_j,P_k) - \log M\,,
\end{aligned}
$$

Together with (2), it yields

$$\frac{1}{M^2}\sum_{j,k=1}^{M}\mathsf{KL}(P_j,P_k) - \log M \ge -\log 2 - \mathbf{P}(Z\ne\psi(X))\log(M-1)$$

Since

$$\mathbf{P}(Z\ne\psi(X)) = \frac{1}{M}\sum_{j=1}^{M}P_j(\psi(X)\ne j) \le \max_{1\le j\le M} P_j(\psi(X)\ne j)\,,$$

this implies the desired result. □

Fano's inequality leads to the following useful theorem.

Theorem 3.2. *Assume that $\Theta$ contains $M \ge 5$ hypotheses $\theta_1,\dots,\theta_M$ such that for some constant $0<\alpha<1/4$, it holds*

*(i)* $|\theta_j-\theta_k|_2^2 \ge 4\phi$

*(ii)* $|\theta_j-\theta_k|_2^2 \le \frac{2\alpha\sigma^2}{n}\log(M)$

*Then*

$$\inf_{\hat\theta}\sup_{\theta\in\Theta}\mathbf{P}_\theta\big(|\hat\theta-\theta|_2^2\ge\phi\big)\ge\frac{1}{2}-2\alpha\,.$$

Proof. in view of $(i)$, it follows from the reduction to hypothesis testing that it is sufficient to prove that

$$\inf_{\psi}\max_{1\le j\le M}\mathbf{P}_{\theta_j}\big[\psi\ne j\big]\ge\frac{1}{2}-2\alpha$$

If follows from $(ii)$ and Example 3.1 that

$$\mathsf{KL}(\mathbf{P}_j,\mathbf{P}_k)=\frac{n|\theta_j-\theta_k|_2^2}{2\sigma^2}\le\alpha\log(M)\,.$$

Moreover, since $M\ge5$,

$$\frac{\frac{1}{M^2}\sum_{j,k=1}^M\mathsf{KL}(\mathbf{P}_j,\mathbf{P}_k)+\log 2}{\log(M-1)}\le\frac{\alpha\log(M)+\log 2}{\log(M-1)}\le 2\alpha+\frac{1}{2}\,.$$

The proof then follows from Fano's inequality. □

3.2.2. *Packing techniques.* Theorem 3.2 indicates that we must take $\phi \le \frac{\alpha\sigma^2}{2n}\log(M)$. Therefore, the larger the $M$, the larger the lower bound can be. However, $M$ cannot be arbitrary larger because of the constraint $(i)$. We are therefore facing a *packing* problem where the goal is to "pack" as many Euclidean balls of radius proportional to $\sigma\sqrt{\log(M)/n}$ in $\Theta$ under the constraint that their centers remain close together (constraint $(ii)$). If $\Theta = \mathbf{R}^d$, this the goal is to pack the Euclidean ball of radius $R = \sigma\sqrt{2\alpha\log(M)/n}$ with Euclidean balls of radius $R\sqrt{2\alpha/\gamma}$. This can be done using the lemma below. It gives a a lower bound on the size of a packing of the discrete hypercube $\{0,1\}^d$ with respect to the *Hamming distance* defined by

$$\rho(\omega,\omega') = \sum_{i=1}^{d} \mathbf{1}(\omega_i \ne \omega'_j)\,, \qquad \forall\,\omega,\omega' \in \{0,1\}^d$$

LEMMA 3.3 (Varshamov-Gilbert). *For any $\gamma \in (0,1/2)$, there exist binary vectors $\omega_1,\dots\omega_M \in \{0,1\}^d$ such that*

$(i)$ $\rho(\omega_j,\omega_k) \ge \big(\frac{1}{2}-\gamma\big)d$ *for all* $j \ne k$ *,*

$(ii)$ $M = \lfloor e^{\gamma^2 d}\rfloor \ge e^{\frac{\gamma^2 d}{2}}$ *.*

PROOF. Let $\omega_{j,i}$, $1 \le i \le d, 1 \le j \le M$ to be i.i.d Bernoulli random variables with parameter $1/2$ and observe that

$$d - \rho(\omega_j,\omega_k) = X \sim \mathsf{Bin}(\mathsf{d},1/2)\,.$$

Therefore it follows from a union bound that

$$\mathbf{P}\big[\exists j \ne k\,,\rho(\omega_j,\omega_k) < \big(\frac{1}{2}-\gamma\big)d\big] \le \frac{M(M-1)}{2}\mathbf{P}\big(X - \frac{d}{2} > \gamma d\big)\,.$$

Hoeffding's inequality then yields

$$\frac{M(M-1)}{2}\mathbf{P}\big(X-\frac{d}{2} > \gamma d\big) \le \exp\Big(-2\gamma^2 d + \log\big(\frac{M(M-1)}{2}\big)\Big) < 1$$

as soon as

$$M(M-1) < 2\exp\Big(2\gamma^2 d\Big)$$

A sufficient condition for the above inequality to hold is to take $M = \lfloor e^{\gamma^2 d}\rfloor \ge e^{\frac{\gamma^2 d}{2}}$. For this value of $M$, we have

$$\mathbf{P}\big(\forall j \ne k\,,\rho(\omega_j,\omega_k) \ge \big(\frac{1}{2}-\gamma\big)d\big) > 0$$

and by virtue of the probabilistic method, there exist $\omega_1,\dots\omega_M \in \{0,1\}^d$ that satisfy $(i)$ and $(ii)$ □

We are now in a position to apply Theorem 3.2 by choosing $\theta_1,\dots,\theta_M$ based on $\omega_1,\dots,\omega_M$ from the Varshamov-Gilbert Lemma.

3.2.3. *Lower bounds for estimation.* Take $\gamma = 1/4$ and apply the Varshamov-Gilbert Lemma to obtain $\omega_1,\dots,\omega_M$ with $M = \lfloor e^{d/16}\rfloor \ge e^{d/32}$ and such that $\rho(\omega_j,\omega_k) \ge d/4$ for all $j \ne k$. Let $\theta_1,\dots,\theta_M$ be such that

$$\theta_j = \omega_j\frac{\beta\sigma}{\sqrt{n}}\,,$$

for some $\beta > 0$ to be chosen later. We can check the conditions of Theorem 3.2:

$$
\begin{aligned}
&(i)\ \ |\theta_j - \theta_k|_2^2 = \frac{\beta^2\sigma^2}{n}\rho(\omega_j, \omega_k) \geq 4\frac{\beta^2\sigma^2 d}{16n}\\
&(ii)\ \ |\theta_j - \theta_k|_2^2 = \frac{\beta^2\sigma^2}{n}\rho(\omega_j, \omega_k) \leq \frac{\beta^2\sigma^2 d}{n} \leq \frac{32\beta^2\sigma^2}{n}\log(M) = \frac{2\alpha\sigma^2}{n}\log(M)\,,
\end{aligned}
$$

for $\beta = \frac{\sqrt{\alpha}}{4}$. Applying now Theorem 3.2 yields

$$
\inf_{\hat\theta}\sup_{\theta\in\mathbf{R}^d} \mathbf{P}_\theta\big(|\hat\theta - \theta|_2^2 \geq \frac{\alpha}{256}\frac{\sigma^2 d}{n}\big) \geq \frac{1}{2} - 2\alpha\,.
$$

It implies the following corollary.

Corollary 3.1. *The minimax rate of estimation of estimating a vector of $\mathbf{R}^d$ with sub-Gaussian noise $z_i \in \mathsf{sG}_d(\sigma^2)$ is $\phi(\mathbf{R}^d) = \sigma^2 d/n$. Moreover, it is attained by the least squares estimator.*

As seen in the previous section, when the mean vector has sparsity $k$, we have to pay for an extra logarithmic term $\log(ed/k)$ for not knowing the sparsity pattern of the unknown $\theta^*$ (when compared to having $\theta^*$ belonging to a known $k$-dimensional space). In this section, we show that this term is unavoidable as it appears in the minimax optimal rate of estimation of sparse vectors.

Note that the vectors $\theta_1, \dots, \theta_M$ employed in the previous subsection are not guaranteed to be sparse because the vectors $\omega_1, \dots, \omega_M$ obtained from the Varshamov-Gilbert Lemma may themselves not be sparse. To overcome this limitation, we need a sparse version of the Varhsamov-Gilbert lemma (see examples sheet 2).

Lemma 3.4 (Sparse Varshamov-Gilbert). *There exist positive constants $C_1$ and $C_2$ such that the following holds for any two integers $k$ and $d$ such that $1 \leq k \leq d/8$. There exist binary vectors $\omega_1, \dots \omega_M \in \{0,1\}^d$ such that*

$$
\begin{aligned}
&(i)\ \ \rho(\omega_i, \omega_j) \geq \frac{k}{2}\ \textit{for all } i \neq j\,,\\
&(ii)\ \ \log(M) \geq \frac{k}{8}\log(1 + \frac{d}{2k})\,.\\
&(iii)\ \ |\omega_j|_0 = k\ \textit{for all } j\,.
\end{aligned}
$$

Apply the sparse Varshamov-Gilbert lemma to obtain $\omega_1, \dots, \omega_M$ with $\log(M) \geq \frac{k}{8}\log(1 + \frac{d}{2k})$ and such that $\rho(\omega_j, \omega_k) \geq k/2$ for all $j \neq k$. Let $\theta_1, \dots, \theta_M$ be such that

$$
\theta_j = \omega_j \frac{\beta\sigma}{\sqrt{n}}\sqrt{\log(1 + \frac{d}{2k})}\,,
$$

for some $\beta > 0$ to be chosen later. We can check the conditions of Theorem 3.2:

$$
\begin{aligned}
&(i)\ \ |\theta_j - \theta_k|_2^2 = \frac{\beta^2\sigma^2}{n}\rho(\omega_j, \omega_k)\log(1 + \frac{d}{2k}) \geq 4\frac{\beta^2\sigma^2}{8n}k\log(1 + \frac{d}{2k})\\
&(ii)\ \ |\theta_j - \theta_k|_2^2 = \frac{\beta^2\sigma^2}{n}\rho(\omega_j, \omega_k)\log(1 + \frac{d}{2k}) \leq \frac{2k\beta^2\sigma^2}{n}\log(1 + \frac{d}{2k}) \leq \frac{2\alpha\sigma^2}{n}\log(M)\,,
\end{aligned}
$$

for $\beta = \sqrt{\frac{\alpha}{8}}$. Applying now Theorem 3.2 yields

$$
\inf_{\hat\theta}\sup_{\substack{\theta\in\mathbf{R}^d\\ |\theta|_0 \leq k}} \mathbf{P}_\theta\big(|\hat\theta - \theta|_2^2 \geq \frac{\alpha^2\sigma^2}{64n}k\log(1 + \frac{d}{2k})\big) \geq \frac{1}{2} - 2\alpha\,.
$$

It implies the following corollary.

Corollary 3.2. *Recall that $\mathcal{B}_0(k) \subset \mathbf{R}^d$ denotes the set of all $k$-sparse vectors of $\mathbf{R}^d$. The minimax rate of estimation of estimating a $k$-sparse vector of $\mathbf{R}^d$ with sub-Gaussian noise $z_i \in \mathsf{sG}_d(\sigma^2)$ is $\phi(\mathcal{B}_0(k)) = \frac{\sigma^2 k}{n}\log(ed/k)$. Moreover, it is attained by the constrained least squares estimator.*

3.3. *Techniques for detection.* To derive lower bounds for hypothesis testing problems with multiple hypotheses, Fano's inequality (Theorem 3.1) is very useful. This is in particularly relevant when one wishes to reduce an estimation problem to a testing problem of this type. To apply this inequality, we need to derive the pairwise Kullback–Leibler divergences between the distributions in the hypotheses.

Another divergence mentioned in Definition 3.2 is the $\chi^2$ divergence, defined when $\mathbf{P}_1 \ll \mathbf{P}_0$ by

$$\chi^2(\mathbf{P}_1, \mathbf{P}_0) = \mathbf{E}\Big[\Big(\frac{\mathrm{d}\mathbf{P}_1}{\mathrm{d}\mathbf{P}_0} - 1\Big)^2\Big]\,.$$

This divergence has several very convenient properties. It is nonnegative, and it can be compared to the total variation distance, through a more straightforward inequality than in Pinsker's inequality (Lemma 3.2).

Lemma 3.5. *For any distributions $\mathbf{P}_0, \mathbf{P}_1$ such that $\mathbf{P}_1 \ll \mathbf{P}_0$, we have*

$$d_{\mathsf{TV}}(\mathbf{P}_1, \mathbf{P}_0) \le \frac{1}{2}\sqrt{\chi^2(\mathbf{P}_1, \mathbf{P}_0)}\,.$$

Proof. We apply the Cauchy–Schartwz inequality to a definition of the total variation distance

$$d_{\mathsf{TV}}(\mathbf{P}_1, \mathbf{P}_0) = \frac{1}{2}\mathbf{E}\Big[\Big|\frac{\mathrm{d}\mathbf{P}_1}{\mathrm{d}\mathbf{P}_0} - 1\Big|\Big] \le \frac{1}{2}\sqrt{\mathbf{E}\Big[\Big(\frac{\mathrm{d}\mathbf{P}_1}{\mathrm{d}\mathbf{P}_0} - 1\Big)^2\Big]} = \frac{1}{2}\sqrt{\chi^2(\mathbf{P}_1, \mathbf{P}_0)}\,.$$

□

As for other divergences, this allows us to obtain an explicit bound on the total variation distance, whenever it is possible to compute it. As for the Kullback–Leibler divergence, this is often possible in practice, largely due to the following properties.

Proposition 3.3. *Let $\mathbf{P}$ and $\mathbf{Q}$ be the distributions of $n$ independent samples, i.e. $\mathbf{P} = \otimes_{i=1}^n \mathbf{P}_i$ and $\mathbf{Q} = \otimes_{i=1}^n \mathbf{Q}_i$, with $\mathbf{P}_i \ll \mathbf{Q}_i$. We have that*

$$\chi^2(\mathbf{P}, \mathbf{Q}) = \prod_{i=1}^n \big(1 + \chi^2(\mathbf{P}_i, \mathbf{Q}_i)\big) - 1\,.$$

Proof. We compute the divergence, using independence of the samples.

$$\begin{aligned}
\chi^2(\mathbf{P}, \mathbf{Q}) &= \mathbf{E}\Big[\Big(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}} - 1\Big)^2\Big] \\
&= \mathbf{E}\Big[\Big(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}\Big)^2\Big] - 1 \\
&= \mathbf{E}\Big[\prod_{i=1}^n \Big(\frac{\mathrm{d}\mathbf{P}_i}{\mathrm{d}\mathbf{Q}_i}\Big)^2\Big] - 1 \\
&= \prod_{i=1}^n \mathbf{E}\Big[\Big(\frac{\mathrm{d}\mathbf{P}_i}{\mathrm{d}\mathbf{Q}_i}\Big)^2\Big] - 1 \\
&= \prod_{i=1}^n \big(1 + \chi^2(\mathbf{P}_i, \mathbf{Q}_i)\big) - 1
\end{aligned}$$

□

Proposition 3.4. *Let $\mathbf{P}_0$ and $\mathbf{P}_1$ be distributions, where $\mathbf{P}_1$ is a finite uniform mixture of $\mathbf{P}_\theta$ for $\theta \in \mathcal{C}$ such that $\mathbf{P}_1 = 1/|\mathcal{C}| \sum_{\theta\in\mathcal{C}} \mathbf{P}_\theta$, where $\mathbf{P}_\theta \ll \mathbf{P}_0$ for all $\theta \in \mathcal{C}$. We have*

$$\chi^2(\mathbf{P}_1, \mathbf{P}_0) = \frac{1}{|\mathcal{C}|^2} \sum_{\theta,\theta'\in\mathcal{C}} \mathbf{E}\Big[\frac{\mathrm{d}\mathbf{P}_\theta}{\mathrm{d}\mathbf{P}_0}\frac{\mathrm{d}\mathbf{P}_{\theta'}}{\mathrm{d}\mathbf{P}_0}\Big] - 1\,.$$

Proof. The result follows directly using definitions of the divergence and of $\mathbf{P}_1$

$$\chi^2(\mathbf{P}_1, \mathbf{P}_0) = \mathbf{E}\Big[\Big(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}\Big)^2\Big] - 1 = \mathbf{E}\Big[\Big(\frac{1}{|\mathcal{C}|}\sum_{\theta\in\mathcal{C}} \frac{\mathrm{d}\mathbf{P}_\theta}{\mathrm{d}\mathbf{P}_0}\Big)^2\Big] - 1 = \frac{1}{|\mathcal{C}|^2} \sum_{\theta,\theta'\in\mathcal{C}} \mathbf{E}\Big[\frac{\mathrm{d}\mathbf{P}_\theta}{\mathrm{d}\mathbf{P}_0}\frac{\mathrm{d}\mathbf{P}_{\theta'}}{\mathrm{d}\mathbf{P}_0}\Big] - 1\,.$$

□

This property is particularly useful to derive lower bounds for detection problems, when the alternative is composite. This folklore technique can be formally justified in the following lemma

Lemma 3.6. *Let $\mathbf{P}_0$ be a distribution and $\mathcal{D}$ be a set of distributions. We have, for any finite subset $\mathcal{D}' \subset \mathcal{D}$, and any test $\psi$*

$$\mathbf{P}_0(\psi = 1) \vee \max_{\mathbf{P}\in\mathcal{D}} \mathbf{P}(\psi = 0) \geq \mathbf{P}_0(\psi = 1) \vee \max_{\mathbf{P}\in\mathcal{D}'} \mathbf{P}(\psi = 0) \geq \mathbf{P}_0(\psi = 1) \vee \frac{1}{|\mathcal{D}|}\sum_{\mathbf{P}\in\mathcal{D}} \mathbf{P}(\psi = 0)$$

Proof. The first inequality holds by taking a maximum over a smaller set. The second one holds by upper bounding an average by the greatest term. □

If we denote by $\mathbf{P}_1$ this finite mixture (that may not be in $\mathcal{D}$), we obtain the following as a consequence of Lemma 3.5

$$\mathbf{P}_0(\psi = 1) \vee \frac{1}{|\mathcal{D}|}\sum_{\mathbf{P}\in\mathcal{D}} \mathbf{P}(\psi = 0) \geq \frac{1 - d_{\mathsf{TV}}(\mathbf{P}_1, \mathbf{P}_0)}{2} \geq \frac{1}{2} - \frac{\sqrt{\chi^2(\mathbf{P}_1, \mathbf{P}_0)}}{4}\,.$$

Proposition 3.4 provides a recipe to compute the last term. To illustrate this, we consider the following problem, mentioned in Section 2.2 and Example 2.2.2, of distinguishing two hypotheses for the distribution of a high-dimensional vector $X \in \mathbf{R}^d$

$$\begin{aligned} H_0 &: X = Z \\ H_1 &: X = \mu + Z,\ \mu \in \mathcal{C}\,. \end{aligned}$$

Here, $Z \in \mathsf{sG}_d(\sigma^2)$, and we take

$$\mathcal{C} = \{\mu \mathbf{1}_S\,,\ S \in \mathcal{S}\}\,,$$

for $\mu > 0$ and some class $\mathcal{S}$ of $k$-sets (subsets of $\{1, \ldots, d\}$ of size $k$). For more information on this problem, see Addario-Berry et al. (2010). We have derived some upper bounds for the probability of error of some tests - i.e. statements of the form "for $\mu$ taking these values, the probability of error is less than - in this problem in these sections and in Examples sheets. We focus now on lower bounds. We start by applying Lemma 3.6 with the subset $\mathcal{D}'$ of the set of distributions

$$\mathcal{D}' = \{\mathcal{N}(\mu\mathbf{1}_S, I_d)\,,\ S \in \mathcal{S}\}\,.$$

We can obtain a lower bound by computing the $\chi^2$ divergence between $\mathbf{P}_1$, uniform mixture of the $\mathcal{N}(\mu \mathbf{1}_S, I_d)$ over $\mathcal{S}$, and $\mathbf{P}_0$. By Proposition 3.4, we have that

$$\chi^2(\mathbf{P}_1, \mathbf{P}_0) = \frac{1}{|\mathcal{S}|^2} \sum_{S,S' \in \mathcal{S}} \mathbf{E}\Big[\frac{\mathrm{d}\mathbf{P}_S}{\mathrm{d}\mathbf{P}_0} \frac{\mathrm{d}\mathbf{P}_{S'}}{\mathrm{d}\mathbf{P}_0}\Big] - 1\,,$$

with the notation that $\mathbf{P}_S = \mathcal{N}(\mu \mathbf{1}_S, I_d)$ and $\mathbf{P}_0 = \mathcal{N}(0, I_d)$. Note that the likelihood ratio $\mathrm{d}\mathbf{P}_S/\mathrm{d}\mathbf{P}_0(X)$ in this expression can be directly computed, for $X \in \mathbf{R}^d$ and $S \in \mathcal{S}$

$$\frac{\mathrm{d}\mathbf{P}_S}{\mathrm{d}\mathbf{P}_0}(X) = \exp\big(-\frac{\mu^2 k}{2} + X_S\big)\,,$$

where $X_S = \langle \mathbf{1}_S, X\rangle$, i.e. the sum of the coefficients of $X$ in $S$. As a consequence, we have that

$$\begin{aligned}
\chi^2(\mathbf{P}_1, \mathbf{P}_0) &= \frac{1}{|\mathcal{S}|^2} \sum_{S,S' \in \mathcal{S}} \mathbf{E}\Big[\frac{\mathrm{d}\mathbf{P}_S}{\mathrm{d}\mathbf{P}_0} \frac{\mathrm{d}\mathbf{P}_{S'}}{\mathrm{d}\mathbf{P}_0}\Big] - 1 \\
&= \frac{1}{|\mathcal{S}|^2} \sum_{S,S' \in \mathcal{S}} e^{-k\mu^2} \mathbf{E}_0\big[e^{\mu(X_S + X_{S'})}\big] - 1 \\
&= \frac{1}{|\mathcal{S}|^2} \sum_{S,S' \in \mathcal{S}} e^{\mu^2 |S \cap S'|} - 1 \\
&= \mathbf{E}_Z\big[e^{\mu^2 Z}\big] - 1
\end{aligned}$$

where $Z = |S \cap S'|$ when $S, S'$ are chosen uniformly and independently in $\mathcal{S}$, and where the penultimate line is a direct computation based on the moment-generating function of a Gaussian variable. In order to obtain an explicit bound on the divergence, we make can some assumptions on the nature of $\mathcal{C}$, as in the following result.

Proposition 3.5. *Let $\mathcal{C}$ be such that*

- *$Z$ conditional on $S'$ has the same distribution for all $S$'.*
- *For a fixed $S_0 \in \mathcal{C}$, and $i \in S_0$, $\mathbf{P}(i \in S) = k/d$.*

*It holds that*

$$\mathbf{E}[e^{\mu^2 Z}] \le 1 + \frac{k}{d}\big(e^{k\mu^2} - 1\big)\,.$$

Proof. By the first assumption on $\mathcal{C}$, we can take without loss of generality $S' = \{1, \ldots, k\}$ in $Z = |S \cap S'|$. As a consequence, it holds by Holder's inequality that

$$\begin{aligned}
\mathbf{E}[e^{\mu^2 Z}] &= \mathbf{E}\Big[\prod_{i=1}^k e^{\mu^2 \mathbf{1}\{i \in S\}}\Big] \\
&\le \prod_{i=1}^k \Big(\mathbf{E}\big[e^{k\mu^2 \mathbf{1}\{i \in S\}}\big]\Big)^{\frac{1}{k}} \\
&= \mathbf{E}\big[e^{k\mu^2 \mathbf{1}\{1 \in S\}}\big]\,,
\end{aligned}$$

and the second assumption gives the result. □

As a consequence of this proposition, under minimal assumptions on $\mathcal{C}$, we have an upper bound on the divergence between the two distributions. For a concrete example, for any $\nu \in (0, 1/2)$, let

$$\mu \leq \sqrt{\frac{1}{k} \log\big(16\frac{\nu^2 d}{k} + 1\big)}$$

Simple computations yield that

$$\chi^{(}\mathbf{P}_1, \mathbf{P}_0) \leq \frac{k}{d}\big(e^{k\mu^2} - 1\big) \leq 16\nu^2$$

As a consequence of Lemma 3.6, we have that

$$\inf_{\psi} \mathbf{P}_0(\psi = 1) \vee \max_{\mathbf{P} \in \mathcal{D}} \mathbf{P}(\psi = 0) \geq \frac{1 - d_{\mathsf{TV}}(\mathbf{P}_1, \mathbf{P}_0)}{2} \geq \frac{1}{2} - \frac{\sqrt{\chi^2(\mathbf{P}_1, \mathbf{P}_0)}}{4} \geq \frac{1}{2} - \nu\,.$$

**4. Linear Regression.** In this section, we focus on the problem of *linear regression* where the data consists of $n$ observations $(X_i, y_i)$ where $y_i \in \mathbf{R}$ is a *response* and where $X_i \in \mathbf{R}^d$ are *covariates* or *regressors*, such that for some $\theta^* \in \mathbf{R}^d$, we have $y_i = X_i\theta^* + z_i$, or in matrix form

$$y = X\theta^* + z$$

with $y \in \mathbf{R}^n$, $X \in \mathbf{R}^{n \times d}$ and $z \in \mathbf{R}^n$ is a noise vector, often assumed to be centred. The goal of this problem is to estimate the explanatory vector $\theta^*$ or the "true response" $X\theta^*$.

4.1. *Parametric setting.* In this section, we assume that the rank of $X$ is $d$, which means that $d \leq n$. This is a low-dimensional, or parametric setting. The most popular method of estimation in linear regression is the method of least-squares. It can be motivated by its equivalence with maximum likelihood estimation when $z$ is normally distributed. Before this connection was made by Gauss, this method had already been used in astronomy as a way to find a linear model that best fits the data, for which there is an closed-form algebraic expression. Formally, we consider the function $\ell : \mathbf{R}^d \to \mathbf{R}$ defined by

$$\ell(\theta) = \|y - X\theta\|_2^2 = \|y\|_2^2 - 2\theta^\top X^\top y + \theta^\top X^\top X\theta\,.$$

This function is a positive definite quadratic form, as $X^\top X \in R^{d \times d}$ has rank $d$. It is therefore minimized at a unique point $\hat{\theta}$ such that $\nabla\ell(\hat{\theta}) = 0$, which yields

$$-2X^\top y + X^\top X\hat{\theta} = 0$$

In this parametric setting, this gives an algebraic expression for the least-squares estimator

$$\hat{\theta} = (X^\top X)^{-1}X^\top y = \theta^* + (X^\top X)^{-1}X^\top z\,.$$

If we are interested in estimating $y^* = X\theta^*$, we can propose $\hat{y} = X\hat{\theta}$, that takes the following form

$$\hat{y} = X\hat{\theta} = X(X^\top X)^{-1}X^\top y = X\theta^* + X(X^\top X)^{-1}X^\top z\,.$$

The matrix $X(X^\top X)^{-1}X^\top \in \mathbf{R}^{n \times n}$ is the orthogonal projection onto the span of $X$, and is often called the *hat matrix*, as it maps $y$ to $\hat{y}$.

4.1.1. *Gaussian and sub-Gaussian cases.* When $z \sim \mathcal{N}(0, \sigma^2 I_d)$, the distribution of these estimates can be directly obtained, and we have

$$\hat\theta \sim \mathcal{N}(\theta^*, \sigma^2 (X^\top X)^{-1})$$
$$X\hat\theta \sim \mathcal{N}(X\theta^*, \sigma^2 X (X^\top X)^{-1} X^\top)\,.$$

This is useful to derive bounds in expectation or high dimension on $\|\hat\theta - \theta^*\|_2^2$ or $\|X\hat\theta - X\theta^*\|_2^2$. More generally, the following is formulated for any vector $z \in \mathbf{R}^d$

Proposition 4.1. *Let $\hat\theta$ be the least-squares estimator in the linear regression problem, and note $V = span(X)$. For any noise vector $z \in \mathbf{R}^n$, it holds that*

$$\|X\hat\theta - X\theta^*\|_2 \le 2 \sup_{u \in V \cap \mathcal{B}_2^n} \langle u, z\rangle\,.$$

Proof. We start by using the definition of the least-squares estimator

$$\begin{aligned}
\|y - X\hat\theta\|_2^2 &\le \|y - X\theta^*\|_2^2 \\
\|(y - X\theta^*) + (X\hat\theta - X\theta^*)\|_2^2 &\le \|y - X\theta^*\|_2^2 \\
\|y - X\theta^*\|_2^2 + \|X\hat\theta - X\theta^*\|_2^2 - 2\langle X(\hat\theta - \theta^*), z\rangle &\le \|y - X\theta^*\|_2^2 \\
\|X\hat\theta - X\theta^*\|_2^2 &\le 2\langle X(\hat\theta - \theta^*), z\rangle \\
\|X\hat\theta - X\theta^*\|_2 &\le 2\Big\langle \frac{X(\hat\theta - \theta^*)}{\|X(\hat\theta - \theta^*)\|}, z\Big\rangle\,.
\end{aligned}$$

Taking $u = X(\hat\theta - \theta^*)/\|X(\hat\theta - \theta^*)\| \in V \cap \mathcal{B}_2^n$ gives the desired result. □

Most proofs in linear regression where the estimator is defined as the minimizer of some function $f$ start with the statement $f(\hat\theta) \le f(\theta^*)$, and try to isolate the term $\|X\hat\theta - X\theta^*\|_2^2$.

Corollary 4.1. *In the linear regression problem with $\textbf{\textit{rank}}(X) = r$ and noise vector $z \in \mathsf{sG}_n(\sigma^2)$, we have that for some constant $C > 0$,*

$$\frac{1}{\sqrt{n}} \mathbf{E}[\|X\hat\theta - X\theta^*\|_2] \le C\sigma\sqrt{\frac{r}{n}}\,.$$

Proof. We apply Proposition 4.1 to obtain

$$\mathbf{E}[\|X\hat\theta - X\theta^*\|_2] \le 2\mathbf{E}[\sup_{u \in V \cap \mathcal{B}_2^n} \langle u, z\rangle]\,.$$

The result follows directly, as in Example 2.4. □

Note that the scaling in $1/\sqrt{n}$ is related to the notion of mean-squared error $\|X\hat\theta - X\theta^*\|_2^2/n$, which is therefore of order $\sigma^2 r/n$. Bounds valid with high probability rather than in expectation can be derived similarly. We remark that if we are interested in estimating $\theta^*$ directly, we can use (in the full rank case) the inequality

$$\|\hat\theta - \theta^*\|_2^2 \le \frac{1}{\lambda_{\min}(X^\top X)} \|X\hat\theta - X\theta^*\|_2^2\,.$$

Note that in a high-dimensional setting this is not possible: the rank of $X$ is bounded by $n$, and $\lambda_{\min}(X^\top X) = 0$. It is also problematic for the problem of prediction (i.e. estimating $X\theta^*$). In general the spanof the design matrix $X$ is all of $\mathbf{R}^n$, and using the least-squares estimator leads to *overfitting*: If $y$ is in the span of $X$, we simply have $\hat y = y$, and the covariates do not provide any information.

4.2. *High-dimensional setting.* Note that if the parameter vector $\theta^*$ is known to belong to a set $\mathcal{C}$, we can adapt the techniques and proofs of Section 2 to these problems. We start by considering the constrained least-squares estimator

$$\hat\theta \in \operatorname*{argmin}_{\theta\in\mathcal{C}} \|y - X\theta\|_2^2\,.$$

We obtain, the following, by adapting the proof of Proposition 4.1

4.2.1. *Analysis of constrained estimator.*

Proposition 4.2. *Let $\hat\theta$ be the least-squares estimator constrained on $\mathcal{C}$ in the linear regression problem, and note $X(\mathcal{C}-\mathcal{C}) = \{Xv : v \in \mathcal{C}-\mathcal{C}\}$. For any noise vector $z \in \mathbf{R}^n$, it holds that*

$$\|X\hat\theta - X\theta^*\|_2^2 \le 2 \sup_{u\in X(\mathcal{C}-\mathcal{C})} \langle u, z\rangle$$
$$\|X\hat\theta - X\theta^*\|_2 \le 2 \sup_{u\in \frac{X(\mathcal{C}-\mathcal{C})}{X(\mathcal{C}-\mathcal{C})}} \langle u, z\rangle$$

Proof. We proceed as in the proof of Proposition 4.1 and obtain that

$$\|X\hat\theta - X\theta^*\|_2^2 \le 2\langle X(\hat\theta-\theta^*), z\rangle\,.$$

We can either take $u = X(\hat\theta-\theta^*) \in X(\mathcal{C}-\mathcal{C})$ or divide by its norm and obtain the second inequality. □

Example 4.1. Let $\mathcal{C} = \mathcal{B}_1^d$, i.e. we assume that $\|\theta^*\|_1 \le 1$. If we have $\|X^{(j)}\|_2 \le \sqrt{n}$ for all the columns of $X$ and $z \in \mathsf{sG}_n(\sigma^2)$, it holds that

$$\frac{1}{n}\mathbf{E}[\|X\hat\theta - X\theta^*\|] \le C\sigma\sqrt{\frac{\log(2d)}{n}}\,.$$

Indeed, in this case the supremum is taken over the image the $\ell_1$ ball of radius 2 by $X$, which is a polytope with $2d$ vertices or less, contained in the $\ell_2$ ball of radius $2\sqrt{n}$.

Example 4.2. If $\mathcal{C}$ is the set of vectors with sparsity $k$ or less, i.e. we assume that $\|\theta^*\|_0 \le k$, and $z \in \mathsf{sG}_n(\sigma^2)$, it holds with probab ility $1-\delta$ that

$$\frac{1}{n}\|X\hat\theta - X\theta^*\|_2^2 \le C\frac{\sigma^2 k}{n}\log(d/k) + C\frac{\sigma^2}{n}\log(1/\delta)\,.$$

Proof. We consider again the inequality $\|X\hat\theta - X\theta^*\|_2^2 \le 2\langle X(\hat\theta-\theta^*), z\rangle$, and note $\hat S = \mathrm{supp}(\hat\theta - \theta^*)$. We have $|\hat S| \le 2k$ and $X\hat\theta - X\theta^* = X_{\hat S}(\hat\theta-\theta^*)$, where $X_{\hat S} \in \mathbf{R}^{n\times|\hat S|}$. Let $\Phi_{\hat S}$ be the matrix of an orthonormal basis of the span of $X_{\hat S}$, with dimension $r_{\hat S}$ (i.e. the rank of $X_{\hat S}$). Take $u \in \mathcal{B}_2^{r_{\hat S}}$ such that

$$\Phi_{\hat S} u = \frac{X_{\hat S}(\hat\theta-\theta^*)}{\|X_{\hat S}(\hat\theta-\theta^*)\|}\,.$$

With these notations, we have that

$$\|X\hat\theta - X\theta^*\|_2 \le 2\langle \Phi_{\hat S} u, z\rangle = 2\langle u, \Phi_{\hat S}^\top z\rangle\,.$$

For the last term, we have $\Phi_{\hat S}^\top z \in \mathsf{sG}_{r_{\hat S}}(\sigma^2)$. It is therefore equivalent to formulate this inequality as

$$\|X\hat\theta - X\theta^*\|_2^2 \le 4 \max_{|S|=2k} \sup_{u \in \mathcal{B}_2^{r_S}} \big(\langle u, z\rangle\big)^2 .$$

Taking a 1/2 net, this yields, in two steps

$$\mathbf{P}\Big( \sup_{u \in \mathcal{B}_2^{r_S}} \big(\langle u, z\rangle\big)^2 > t\Big) \le 26^{2k} e^{-\frac{t}{8\sigma^2}}$$

and (by a union bound)

$$\mathbf{P}\big(\|X\hat\theta - X\theta^*\|_2^2 > t\big) \le 2\binom{d}{2k} 6^{2k} e^{-\frac{t}{8\sigma^2}} .$$

Setting the last term to $\delta$ yields the desired result. □

Notes: There are several issues with this estimator. The first one is that computing it is equivalent to solving an NP-hard problem, so it cannot be implemented in practice, in particular for large datasets. The other one is that it requires the sparsity to be known. One would wish to have an estimator for which the same guarantees hold, without having to specify or know it in advance. This property is called adaptivity.

4.2.2. *Analysis of the BIC estimator.* In order to address the latter issue, we consider the Bayesian information criterion (BIC) estimator, defined for any parameter $\tau > 0$ as

$$\hat\theta \in \operatorname*{argmin}_{\theta \in \mathbf{R}^d} \{ \frac{1}{n}\|y - X\theta\|_2^2 + \tau^2 \|\theta\|_0 \} .$$

Note that . While computationally hard to implement, the BIC estimator gives us a good benchmark for sparse estimation.

THEOREM 4.1. *In the linear regression problem, with $z \in \mathsf{sG}_n(\sigma^2)$, the BIC estimator $\hat\theta$ with regularization parameter*

$$\tau^2 = 16 \log(6) \frac{\sigma^2}{n} + 32 \frac{\sigma^2 \log(ed)}{n} . \tag{3}$$

*satisfies*

$$\frac{1}{n}\|X\hat\theta - X\theta^*\|_2^2 \le C\sigma^2 \, |\theta^*|_0 \frac{\log(ed/\delta)}{n}$$

*with probability at least $1-\delta$.*

PROOF. We begin as usual by noting that

$$\frac{1}{n}|Y - X\hat\theta|_2^2 + \tau^2 |\hat\theta|_0 \le \frac{1}{n}|Y - X\theta^*|_2^2 + \tau^2 |\theta^*|_0 .$$

It implies

$$|X\hat\theta - X\theta^*|_2^2 \le n\tau^2 |\theta^*|_0 + 2z^\top X(\hat\theta - \theta^*) - n\tau^2 |\hat\theta|_0 .$$

First, note that

$$\begin{aligned} 2z^\top X(\hat\theta - \theta^*) &= 2z^\top \Big( \frac{X\hat\theta - X\theta^*}{|X\hat\theta - X\theta^*|_2} \Big) |X\hat\theta - X\theta^*|_2 \\ &\le 2\Big[ z^\top \Big( \frac{X\hat\theta - X\theta^*}{|X\hat\theta - X\theta^*|_2} \Big) \Big]^2 + \frac{1}{2}|X\hat\theta - X\theta^*|_2^2 , \end{aligned}$$

where we use the inequality $2ab \le 2a^2 + \frac{1}{2}b^2$. Together with the previous display, it yields

$$|X\hat\theta - X\theta^*|_2^2 \le 2n\tau^2|\theta^*|_0 + 4\big[z^\top \mathcal{U}(\hat\theta - \theta^*)\big]^2 - 2n\tau^2|\hat\theta|_0 \tag{4}$$

where

$$\mathcal{U}(\hat\theta - \theta^*) = \frac{X\hat\theta - X\theta^*}{|X\hat\theta - X\theta^*|_2}$$

Next, we need to "sup out" $\hat\theta$. To that end, we decompose the sup into a max over cardinalities as follows:

$$\sup_{\theta \in \mathbf{R}^d} = \max_{1\le k\le d} \max_{|S|=k} \sup_{\mathrm{supp}(\theta)=S} .$$

Applied to the above inequality, it yields

$$\begin{aligned}
&4\big[z^\top \mathcal{U}(\hat\theta - \theta^*)\big]^2 - 2n\tau^2|\hat\theta|_0\\
&\quad\le \max_{1\le k\le d}\Big\{\max_{|S|=k}\sup_{\mathrm{supp}(\theta)=S} 4\big[z^\top \mathcal{U}(\theta - \theta^*)\big]^2 - 2n\tau^2 k\Big\}\\
&\quad\le \max_{1\le k\le d}\Big\{\max_{|S|=k}\sup_{u\in\mathcal{B}_2^{r_{S,*}}} 4\big[z^\top \Phi_{S,*}u\big]^2 - 2n\tau^2 k\Big\},
\end{aligned}$$

where $\Phi_{S,*} = [\phi_1, \ldots, \phi_{r_{S,*}}]$ is an orthonormal basis of the set $\{X_j, j \in S \cup \mathrm{supp}(\theta^*)\}$ of columns of $X$ and $r_{S,*} \le |S| + |\theta^*|_0$ is the dimension of this column span.

Using union bounds, we get for any $t > 0$,

$$\begin{aligned}
&\mathbf{P}\Big(\max_{1\le k\le d}\Big\{\max_{|S|=k}\sup_{u\in\mathcal{B}_2^{r_{S,*}}} 4\big[z^\top \Phi_{S,*}u\big]^2 - 2n\tau^2 k\Big\} \ge t)\\
&\quad\le \sum_{k=1}^d \sum_{|S|=k} \mathbf{P}\Big(\sup_{u\in\mathcal{B}_2^{r_{S,*}}} \big[z^\top \Phi_{S,*}u\big]^2 \ge \frac{t}{4} + \frac{1}{2}n\tau^2 k\Big)
\end{aligned}$$

Moreover, using the $\varepsilon$-net argument, we get for $|S| = k$,

$$\begin{aligned}
\mathbf{P}\Big(\sup_{u\in\mathcal{B}_2^{r_{S,*}}} \big[z^\top \Phi_{S,*}u\big]^2 \ge \frac{t}{4} + \frac{1}{2}n\tau^2 k\Big) &\le 2\cdot 6^{r_{S,*}} \exp\Big(-\frac{\frac{t}{4} + \frac{1}{2}n\tau^2 k}{8\sigma^2}\Big)\\
&\le 2\exp\Big(-\frac{t}{32\sigma^2} - \frac{n\tau^2 k}{16\sigma^2} + (k + |\theta^*|_0)\log(6)\Big)\\
&\le \exp\Big(-\frac{t}{32\sigma^2} - 2k\log(ed) + |\theta^*|_0\log(12)\Big)
\end{aligned}$$

where, in the last inequality, we used the definition (3) of $\tau$.

Putting everything together, we get

$$\begin{aligned}
&\mathbf{P}\Big(|X\hat\theta - X\theta^*|_2^2 \ge 2n\tau^2|\theta^*|_0 + t\Big) \le \\
&\sum_{k=1}^d \sum_{|S|=k} \exp\Big(-\frac{t}{32\sigma^2} - 2k\log(ed) + |\theta^*|_0\log(12)\Big)\\
&= \sum_{k=1}^d \binom{d}{k} \exp\Big(-\frac{t}{32\sigma^2} - 2k\log(ed) + |\theta^*|_0\log(12)\Big)\\
&\le \sum_{k=1}^d \exp\Big(-\frac{t}{32\sigma^2} - k\log(ed) + |\theta^*|_0\log(12)\Big)\\
&= \sum_{k=1}^d (ed)^{-k} \exp\Big(-\frac{t}{32\sigma^2} + |\theta^*|_0\log(12)\Big)\\
&\le \exp\Big(-\frac{t}{32\sigma^2} + |\theta^*|_0\log(12)\Big)\,.
\end{aligned}$$

To conclude the proof, choose $t = 32\sigma^2|\theta^*|_0\log(12) + 32\sigma^2\log(1/\delta)$ and observe that combined with (4), it yields with probability $1-\delta$,

$$\begin{aligned}
|X\hat\theta - X\theta^*|_2^2 &\le 2n\tau^2|\theta^*|_0 + t\\
&= 64\sigma^2\log(ed)|\theta^*|_0 + 64\log(12)\sigma^2|\theta^*|_0 + 32\sigma^2\log(1/\delta)\\
&\le 224|\theta^*|_0\sigma^2\log(ed) + 32\sigma^2\log(1/\delta)\,.
\end{aligned}$$

□

It follows from Theorem 4.1 that $\hat\theta$ *adapts* to the unknown sparsity of $\theta^*$. Moreover, this holds under no assumption on the design matrix $X$. However, it does not address the algorithmic problem.

4.2.3. *Slow rate for the Lasso estimator.* To obtain an estimator that is actually tractable, one approach is to replace the $\ell_0$ penalty by a convex surrogate, the $\ell_1$ norm. Formally, the *Lasso estimator* is defined, for $\tau > 0$ as

$$\hat\theta \in \operatorname*{argmin}_{\theta\in\mathbf{R}^d}\{\frac{1}{n}\|y - X\theta\|_2^2 + 2\tau\|\theta\|_1\}\,.$$

Lasso estimator is a bit more difficult because, by construction, it should more naturally adapt to the unknown $\ell_1$-norm of $\theta^*$. This can be easily shown as in the next theorem.

Theorem 4.2. *In the linear regression problem, with $z \in \mathsf{sG}_n(\sigma^2)$, assume that the columns of $X$ are normalized in such a way that $\max_j |X_j|_2 \le \sqrt{n}$. Then, the Lasso estimator $\hat\theta$ with regularization parameter*

$$2\tau = 2\sigma\sqrt{\frac{2\log(2d)}{n}} + 2\sigma\sqrt{\frac{2\log(1/\delta)}{n}}\,. \tag{5}$$

*satisfies*

$$\frac{1}{n}|X\hat\theta - X\theta^*|_2^2 \le 4|\theta^*|_1\sigma\sqrt{\frac{2\log(2d)}{n}} + 4|\theta^*|_1\sigma\sqrt{\frac{2\log(1/\delta)}{n}}$$

*with probability at least $1-\delta$. Moreover, there exists a numerical constant $C > 0$ such that*

$$\frac{1}{n}\mathbf{E}\big[|X\hat\theta - X\theta^*|_2^2\big] \le C|\theta^*|_1\sigma\sqrt{\frac{\log(2d)}{n}}\,.$$

Proof. From the definition of $\hat\theta$, it holds

$$\frac{1}{n}|Y - X\hat\theta|_2^2 + 2\tau|\hat\theta|_1 \le \frac{1}{n}|Y - X\theta^*|_2^2 + 2\tau|\theta^*|_1\,.$$

Using Hölder's inequality, it implies

$$\begin{aligned}|X\hat\theta - X\theta^*|_2^2 &\le 2\varepsilon^\top X(\hat\theta-\theta^*) + 2n\tau\big(|\theta^*|_1 - |\hat\theta|_1\big)\\ &\le 2|X^\top\varepsilon|_\infty|\hat\theta|_1 - 2n\tau|\hat\theta|_1 + 2|X^\top\varepsilon|_\infty|\theta^*|_1 + 2n\tau|\theta^*|_1\\ &= 2(|X^\top\varepsilon|_\infty - n\tau)|\hat\theta|_1 + 2(|X^\top\varepsilon|_\infty + n\tau)|\theta^*|_1\end{aligned}$$

Observe now that for any $t > 0$,

$$\mathbf{P}(|X^\top\varepsilon|_\infty \ge t) = \mathbf{P}(\max_{1\le j\le d}|X_j^\top\varepsilon| > t) \le 2de^{-\frac{t^2}{2n\sigma^2}}$$

Therefore, taking $t = \sigma\sqrt{2n\log(2d)} + \sigma\sqrt{2n\log(1/\delta)} = n\tau$, we get that with probability $1-\delta$,

$$|X\hat\theta - X\theta^*|_2^2 \le 4n\tau|\theta^*|_1\,.$$

The bound in expectation follows using the same argument. □

Notice that the regularization parameter (5) depends on the confidence level $\delta$. This not the case for the BIC estimator (see (3)).

4.2.4. *Fast rate for the Lasso.* Note that the rate is sub-optimal, but that the estimator can now be efficiently computed, as it is a minimizer of a convex function. The result can nevertheless be improved, and we can obtain a fast rate, while maintaining adaptivity in $\|\theta^*\|_0$. This requires to modify the proof, and to actually show that $\|\hat\theta - \theta^*\|_2$ is small. It is not the case in any of the previous proofs, and actually sometimes not the case. In particular, if $X$ has two identical columns, it is impossible to estimate $\theta^*$ accurately: the problem is ill-posed. However, if an assumption is made on the design matrix $X$, it will be possible to prove the desired result.

Definition 4.1. Let $A \in \mathbf{R}^{n\times d}$. It is said to satisfy the *restricted isometry property* (RIP) for sparsity $k$, with parameter $\alpha \in (0,1)$ if for all $v \in \mathbf{R}^d$ such that $\|v\|_0 \le k$, it holds that

$$(1-\alpha)\|v\|_2^2 \le \|Xv\|_2^2 \le (1+\alpha)\|v\|_2^2$$

Some aspects of this property are studied in Examples sheet 3.

Theorem 4.3. *Fix $n \ge 2$, in the linear regression problem where $z \in \mathsf{sG}_n(\sigma^2)$. Moreover, assume that $|\theta^*|_0 \le k$ and that $X/\sqrt{n}$ satisfies the restricted isometry property, for sparsity $k$, and parameter $\alpha$. The Lasso estimator $\hat\theta$ with regularization parameter defined by*

$$2\tau = 8\sigma\sqrt{\frac{\log(2d)}{n}} + 8\sigma\sqrt{\frac{\log(1/\delta)}{n}}$$

*satisfies, for some $C_\alpha > 0$*

$$\frac{1}{n}|X\hat\theta - X\theta^*|_2^2 \le C_\alpha k\sigma^2\frac{\log(2d/\delta)}{n}$$

*and*

$$|\hat\theta - \theta^*|_1 \le C_\alpha k\sigma\sqrt{\frac{\log(2d/\delta)}{n}}\,.$$

*with probability at least* $1-\delta$*. Moreover,*

$$\frac{1}{n}\mathbf{E}\big[|X\hat\theta - X\theta^*|_2^2\big] \le C_\alpha k\sigma^2\frac{\log(2d)}{n}\,, \quad \textit{and} \quad \mathbf{E}\big[|\hat\theta - \theta^*|_1\big] \lesssim k\sigma\sqrt{\frac{\log(2d/\delta)}{n}}\,.$$

Proof. From the definition of $\hat\theta$, it holds

$$\frac{1}{n}|Y - X\hat\theta|_2^2 \le \frac{1}{n}|Y - X\theta^*|_2^2 + 2\tau|\theta^*|_1 - 2\tau|\hat\theta|_1\,.$$

Adding $\tau|\hat\theta - \theta^*|_1$ on each side and multiplying by $n$, we get

$$|X\hat\theta - X\theta^*|_2^2 + n\tau|\hat\theta - \theta^*|_1 \le 2\varepsilon^\top X(\hat\theta - \theta^*) + n\tau|\hat\theta - \theta^*|_1 + 2n\tau|\theta^*|_1 - 2n\tau|\hat\theta|_1\,.$$

Applying Hölder's inequality and using the same steps as in the proof of Theorem 4.2, we get that with probability $1-\delta$, we get

$$\begin{aligned}\varepsilon^\top X(\hat\theta - \theta^*) &\le |\varepsilon^\top X|_\infty|\hat\theta - \theta^*| \\ &\le \frac{n\tau}{2}|\hat\theta - \theta^*|_1\,,\end{aligned}$$

where we used the fact that $|X_j|_2^2 \le n + 1/(14k) \le 2n$. Therefore, taking $S = \mathrm{supp}(\theta^*)$ to be the support of $\theta^*$, we get

$$\begin{aligned}|X\hat\theta - X\theta^*|_2^2 + n\tau|\hat\theta - \theta^*|_1 &\le 2n\tau|\hat\theta - \theta^*|_1 + 2n\tau|\theta^*|_1 - 2n\tau|\hat\theta|_1 \\ &= 2n\tau|\hat\theta_S - \theta^*|_1 + 2n\tau|\theta^*|_1 - 2n\tau|\hat\theta_S|_1 \\ &\le 4n\tau|\hat\theta_S - \theta^*|_1 \qquad (6)\end{aligned}$$

Using now the Cauchy-Schwarz inequality the RIP on $X/\sqrt{n}$ respectively, we get since $|S| \le k$,

$$|\hat\theta_S - \theta^*|_1 \le \sqrt{|S|}|\hat\theta_S - \theta^*|_2 \le \frac{1}{\sqrt{1-\alpha}}\sqrt{\frac{k}{n}}|X\hat\theta - X\theta^*|_2\,.$$

Combining this result with (6) yields the desired results. The bound in expectation follows using the same argument. □

**5. Matrix problems.** In many high-dimensional problems, the data is available in the form of a matrix. They can be vectorized and the results of the previous sections ca be applied, but it is often better to exploit notions specific to matrices (rank, eigenvalues, etc) in these problems. Matrices can also be used in problems that do not directly appear to be related to them, e.g. involving graphs or dependencies between variables of a vector.

5.1. *Notations.* For a real valued matrix $A \in \mathbf{R}^{m\times n}$, the rank of $A$ is the dimension of the span of $A$. We have $\mathbf{rank}(A) = r \le \min(m,n)$. The *singular value decomposition* of $A$ is the factorization

$$A = UDV^\top = \sum_{j=1}^r \sigma_j u_j v_j^\top\,,$$

where $D$ is the diagonal of the $\sigma_j > 0$, and the columns of $U$ and $V$ are orthonormal, formed by the eigenvectors of $AA^\top$ and $A^\top A$

$$AA^\top u_j = \sigma_j^2 u_j \quad \text{and} \quad A^\top A u_j = \sigma_j^2 v_j \,.$$

The largest singular value $\sigma_{\max}(A)$ therefore satisfies

$$\sigma_{\max}(A) = \max_{x \in \mathbf{R}^n \setminus \{0\}} \frac{\|Ax\|}{\|x\|} = \max_{\substack{y \in \mathcal{S}^{m-1} \\ x \in \mathcal{S}^{n-1}}} y^\top A x \,.$$

For a matrix $\Sigma$ that is symmetric and positive semidefinite (e.g. and i.e., a covariance matrix), the singular values are eigenvalues and we have

$$\lambda_{\max}(\Sigma) = \max_{x \in \mathcal{S}^{n-1}} x^\top A x \,.$$

The operator norm of a matrix is the value of its largest singular value. The Frobenius norm is the $\ell_2$ norm of all the singular values, it is also equal to the $\ell_2$ norm of the matrix, treated as a vector.

5.2. *Additive models.* In this section, we focus on problems of the type

$$Y = \Theta^* + Z \,,$$

where the observation is the matrix $Y$, the true signal is $\Theta^*$, and the noise is $Z \in \mathsf{sG}_{m \times d}(\sigma^2/n)$, representing $n$ i.i.d observations of $Y$ with noise level $\sigma^2$. If no assumption can be made about $\Theta^*$, and our estimate is $Y$ - corresponding to the average of observations, or least-squares estimate with $n$ independent observations - the error is, measured in Frobenius norm

$$\|Y - \Theta^*\|_F^2 \le \|Z\|_F^2 \approx \sigma^2 \frac{md}{n}$$

5.2.1. *Low-rank signals.* In high-dimensional settings where $m, d$ can be much larger than $n$, this can be very problematic. If $\Theta^*$ is assumed to have a simple structure, e.g. has small rank, we can use a constrained estimator of the type

$$\hat{\Theta} \in \operatorname*{argmin}_{\mathbf{rank}(\Theta) \le r} \|Y - \Theta\|_F^2 \,.$$

One can check that that in this case the SVD of $\hat{\Theta}$ and $Y$ are directly related. If we have

$$Y = \sum_{j=1}^{\min(m,d)} \hat{\sigma}_j \hat{u}_j \hat{v}_j^\top \,,$$

the decomposition of $\hat{\Theta}$ can be obtained by truncating the decomposition of $Y$, as we have

$$\hat{\Theta} = \sum_{j=1}^{r} \hat{\sigma}_j \hat{u}_j \hat{v}_j^\top \,.$$

It is actually easier to analyze a *thresholded* (rather than truncated) version of the decomposition of $Y$, defined by some real $\tau > 0$ as

$$Y = \sum_{j=1}^{\min(m,d)} \hat{\sigma}_j \mathbf{1}\{\hat{\sigma}_j > 2\tau\} \, \hat{u}_j \hat{v}_j^\top \,,$$

Proposition 5.1. *Let $\hat\Theta$ be the thresholded SVD estimator in the linear model, and let*

$$2\tau = 8\sigma\sqrt{\frac{(d \vee m)\log(12)}{n}} + 4\sigma\sqrt{\frac{2\log(1/\delta)}{n}}\,.$$

*For some constant $C > 0$, it holds with probability at least $1-\delta$ that*

$$\|\hat\Theta - \Theta^*\|_F^2 \le C\,\frac{\sigma^2\mathbf{rank}(\Theta^*)}{n}\big((d\vee m) + \log(1/\delta)\big)\,.$$

We will use the following lemma, given without proof. It can be established by the same techniques as a problem in the Examples sheet.

Lemma 5.1. *Let $Z \in \mathsf{sG}_{m\times d}(\sigma^2/n)$, and $\tau$ as in Proposition 5.1. There is an event $\mathcal{A}$, with probability $1-\delta$, on which it holds that $\|Z\|_{op} \le \tau$.*

We also use the following result

Theorem 5.1 (Weyl's inequality). *Denoting by $\sigma_j$ the singular values of $\Theta^*$ and by $\hat\sigma_j$ the singular values of $Y$ (in nondecreasing order), it holds for all $j$ that*

$$|\sigma_j - \hat\sigma_j| \le \|Y - \Theta^*\|_{op}\,.$$

It is useful in this model, as $Y - \Theta^* = Z$, and the operator norm is bounded by $\tau$ on the event $\mathcal{A}$. Based on these observations, we can prove Proposition 5.1

Proof. Assume without loss of generality that the singular values of $\Theta^*$ and $y$ are arranged in a non increasing order: $\sigma_1 \ge \sigma_2 \ge \dots$ and $\hat\sigma_1 \ge \hat\sigma_2 \ge \dots$. Define the set $\hat S = \{j\,:\,|\hat\sigma_j| > 2\tau\}$.

Observe first that it follows from Lemma 5.1 that $\|Z\|_{\mathrm{op}} \le \tau$ for $\tau$ chosen as in (**??**) on an event $\mathcal{A}$ such that $\mathbf{P}(\mathcal{A}) \ge 1-\delta$. The rest of the proof is on $\mathcal{A}$.

It follows from Weyl's inequality that $|\hat\lambda_j - \lambda_j| \le \|Z\|_{\mathrm{op}} \le \tau$. It implies that $\hat S \subset \{j\,:\,|\lambda_j| > \tau\}$ and $\hat S^c \subset \{j\,:\,|\lambda_j| \le 3\tau\}$.

Next define the oracle $\bar\Theta = \sum_{j\in\hat S}\lambda_j u_j v_j^\top$ and note that

$$\|\hat\Theta - \Theta^*\|_F^2 \le 2\|\hat\Theta - \bar\Theta\|_F^2 + 2\|\bar\Theta - \Theta^*\|_F^2 \tag{7}$$

Using Cauchy-Schwarz, we control the first term in (7) as follows

$$\|\hat\Theta - \bar\Theta\|_F^2 \le \mathbf{rank}(\hat\Theta - \bar\Theta)\|\hat\Theta - \bar\Theta\|_{\mathrm{op}}^2 \le 2|S|\|\hat\Theta - \bar\Theta\|_{\mathrm{op}}^2$$

Moreover,

$$\begin{aligned}\|\hat\Theta - \bar\Theta\|_{\mathrm{op}} &\le \|\hat\Theta - y\|_{\mathrm{op}} + \|y - \Theta^*\|_{\mathrm{op}} + \|\Theta^* - \bar\Theta\|_{\mathrm{op}}\\ &\le \max_{j\in S^c}|\hat\lambda_j| + \tau + \max_{j\in S^c}|\lambda_j| \le 6\tau\,.\end{aligned}$$

Therefore,

$$\|\hat\Theta - \bar\Theta\|_F^2 \le 72|S|\tau^2 = 72\sum_{j\in S}\tau^2\,.$$

The second term in (7) can be written as

$$\|\bar\Theta - \Theta^*\|_F^2 = \sum_{j\in S^c}|\lambda_j|^2\,.$$

Plugging the above two displays in (7), we get

$$\|\hat{\Theta} - \Theta^*\|_F^2 \leq 144 \sum_{j \in S} \tau^2 + \sum_{j \in S^c} |\lambda_j|^2$$

Since on $S$, $\tau^2 = \min(\tau^2, |\lambda_j|^2)$ and on $S^c$, $|\lambda_j|^2 \leq 3 \min(\tau^2, |\lambda_j|^2)$, it yields,

$$\begin{aligned} \|\hat{\Theta} - \Theta^*\|_F^2 &\leq 432 \sum_j \min(\tau^2, |\lambda_j|^2) \\ &\leq 432 \sum_{j=1}^{\mathbf{rank}(\Theta^*)} \tau^2 \\ &= 432\, \mathbf{rank}(\Theta^*) \tau^2 \,. \end{aligned}$$

□

5.3. *Spectral methods in estimation.* In this section, we focus on symmetric matrices that are semidefinite positive, i.e. $\Sigma \in \mathbf{R}^{n \times n}$ such that for all $u \in \mathbf{R}^n$, we have $u^\top \Sigma u \geq 0$. Their eigendecomposition can be written in the form

$$\Sigma = \sum_{i=1}^n \lambda_i \, v_i v_i^\top \,,$$

where the $v_i$ are orthonormal eigenvectors of $\Sigma$, with associated nonnegative eigenvalues $\lambda_i$. In many statistical problems, we can construct from the observations a matrix $\hat{\Sigma}$ that has for mean a SDP matrix $\Sigma$, where the leading eigenvector $v = v_1 \in \mathcal{S}$ (for some parameter set $\mathcal{S}$) is important. It can either be the signal that we wish to estimate, or it can give information about it (dependence structure among coefficients, clustering in a network, etc). The vector $v$ is a maximizer of $u^\top \Sigma u$ over the unit sphere, and therefore on $\mathcal{S}$. It is natural to try to estimate $v$ by its analogue for $\hat{\Sigma}$

$$\hat{v} \in \operatorname*{argmax}_{u \in \mathcal{S}} u^\top \hat{\Sigma} u \,.$$

The following theorem allows us to control how good of an estimate it is.

Theorem 5.2 (Davis–Kahan). *Let $\Sigma$ be a semidefinite positive matrix with $\lambda_2 \leq 1$, $\lambda_1 = 1 + \theta$. It holds that*

$$\min_{\varepsilon \in \{\pm 1\}} |\varepsilon \hat{v} - v|^2 \leq \|\hat{v}\hat{v}^\top - vv^\top\|_F^2 \leq \frac{2\sqrt{2}}{\theta} \|\hat{\Sigma} - \Sigma\|_{op, \mathcal{S}} \,.$$

There are a few interesting points about this result. The norm $\|\cdot\|_{\mathrm{op},\mathcal{S}}$ is a proxy of the operator norm adapted for $\mathcal{S}$; we discuss it later. The error is measured between $v$ and $\pm\hat{v}$ because of the ambiguity up to a sign of eigenvectors. This ambiguity disappears when considering $vv^\top$, which is the orthonormal projector on the space generated by $v$. This error is governed by two quantities: the *spectral gap* $\theta$, which measures the curvature of the quadratic norm, and the operator norm of the difference between the two matrices. It makes sense that the spectral gap is involved: if it is too small, the vectors $v_2, v_3$, etc., who are orthonormal to $v_1$, are "almost maximizers" of the quadratic form generated by $\Sigma$, and could therefore be close to maximizers of the one generated by $\hat{\Sigma}$.

Proof. We establish the first inequality

$$\min_{\varepsilon \in \{\pm 1\}} |\varepsilon \hat{v} - v|^2 = 2 - 2|\hat{v}^\top v| \leq 2 - 2(\hat{v}^\top v)^2 = \|\hat{v}\hat{v}^\top - vv^\top\|_F^2 \,.$$

We establish the second one by using the definition of $v$ as leading eigenvector of $\Sigma$ and the fact that $u^\top \Sigma u \leq 1 + \theta(v^\top u)^2$ for all unit vectors $u$. We therefore have

$$v^\top \Sigma v - \hat{v}^\top \Sigma \hat{v} \geq \theta\big(1 - (v^\top \hat{v})^2\big) = \frac{\theta}{2}\|\hat{v}\hat{v}^\top - vv^\top\|_F^2 .$$

We should therefore only need to bound this quantity

$$\begin{aligned} v^\top \Sigma v - \hat{v}^\top \Sigma \hat{v} &= v^\top \hat{\Sigma} v - \hat{v}^\top \Sigma \hat{v} - v^\top(\hat{\Sigma} - \Sigma)v \\ &\leq \hat{v}^\top \hat{\Sigma} \hat{v} - \hat{v}^\top \Sigma \hat{v} - v^\top(\hat{\Sigma} - \Sigma)v \\ &\leq \langle \hat{\Sigma} - \Sigma, \hat{v}\hat{v}^\top - vv^\top \rangle \\ &\leq \|\hat{\Sigma} - \Sigma\|_{\mathrm{op},\mathcal{S}} \|\hat{v}\hat{v}^\top - vv^\top\|_1 \\ &\leq \sqrt{2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op},\mathcal{S}} \|\hat{v}\hat{v}^\top - vv^\top\|_F \end{aligned}$$

Together, this yields the desired inequality. □

The norm $\|\cdot\|_{\mathrm{op},\mathcal{S}}$ can be taken as the norm such that the following inequality is satisfied

$$\langle A, \hat{v}\hat{v}^\top - vv^\top \rangle \leq \|A\|_{\mathrm{op},\mathcal{S}} \|\hat{v}\hat{v}^\top - vv^\top\|_1 ,$$

when $v, \hat{v}$ are two unit vectors of $\mathcal{S}$. If $\mathcal{S}$ is the whole unit sphere, it is the usual operator norm. Otherwise, it can be more restrictive: e.g. if $\mathcal{S}$ is the set of unit vectors with a sparsity less than $k$, it is the maximum of the operator norms of all the submatrices of $A$ of size $2k$. It can naturally be modified for other cases.

## References.

Addario-Berry, L., Broutin, N., Devroye, L. and Lugosi, G. (2010). On combinatorial testing problems. *Ann. Statist.* **38** 3063–3092.

Berthet, Q. (2014). *Computational and statistical tradeoffs in high-dimensional problems.* Ph.D. thesis, University of Princeton.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press.

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory.* Wiley-Interscience, New York, NY, USA.

Csiszár, I. (1963). Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutato Int. Kozl* **8** 85–108.

Csiszár, I. and Körner, J. (2011). *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Cambridge Univ. Press.

Ibragimov, I. A. and Hasminskii, R. Z. (1981). *Statistical estimation, volume 16 of Applications of Mathematics.* Springer-Verlag, New York.

Rigollet, P. (2016). *High Dimensional Statistics.*

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation.* Springer Series in Statistics, Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.