Chapter 1

Subgaussian Random Variables and Concentration

1.1 The law of large numbers and the central limit theorem

The law of large numbers is possibly one of the most celebrated results in probability, second probably only to the central limit theorem. Let us see what each of them says and where their limitations lie.

Theorem 1 (Strong Law of large numbers)**.** *Suppose that* $\{X_i : i \geq \mathbb{N}\}$ *is an i.i.d. collection of random variables such that* $\mathbb{E}[X_i] = \mu$*. Then we have that*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu, \quad \textit{almost surely},$$

as $n \to \infty$*.*

We also have the Weak Law of Large numbers.

Theorem 2 (Weak Law of large numbers)**.** *Suppose that* $\{X_i : i \geq \mathbb{N}\}$ *is an i.i.d. collection of random variables such that* $\mathbb{E}[X_i] = \mu$*. Then we have that*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu, \quad \textit{almost surely},$$

as $n \to \infty$*.*

Here $\xrightarrow{\mathbb{P}}$ denotes *convergence in probability.*

Before we move on let us briefly recall two of the three basic modes of convergence of random variables.

Definition 1 (Convergence in probability)**.** *a sequence of random variables* $\{X_i : i \in \mathbb{N}\}$ *defined on a common probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is said to converge to a random variable* X*, also on* $(\Omega, \mathcal{F}, \mathbb{P})$ *if for any* $\epsilon > 0$ *we have*

$$\mathbb{P}\left[|X_n - X| > \epsilon\right] \to 0,$$

as $n \to \infty$*.*

Contrast this with almost sure convergence.

Definition 2 (Convergence in probability)**.** *a sequence of random variables* $\{X_i : i \in \mathbb{N}\}$ *defined on a common probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is said to converge almost surely to a random variable* X*, also on* $(\Omega, \mathcal{F}, \mathbb{P})$*, denoted* $X_n \to X$*, if for any* $\epsilon > 0$ *we have*

$$\mathbb{P}\left[\lim_{n\to\infty} X_n \to X\right] = 1.$$

The difference in the two definitions is the order of the limit operation and the probability operator.

Example 1. *Let* $\Omega = [0, 1)$*,* $\mathcal{F}$ *be the Borel* σ*-algebra, the one generated by open sets, and* $\mathbb{P} = \mathsf{Leb}$*, the standard Lebesgue measure on* $[0, 1)$*. Define the sequence of random variables* $\{X_n : n \in \mathbb{N}\}$ *as follows:*

$$\begin{aligned}
X_0(\omega) &\equiv 1 \quad \textit{for all } \omega \in [0,1],\\
X_1(\omega) &= \mathbb{1}_{[0,1/2)}(\omega),\\
X_2(\omega) &= \mathbb{1}_{[1/2,1)}(\omega),\\
X_3(\omega) &= \mathbb{1}_{[0,1/3)}(\omega),\\
X_4(\omega) &= \mathbb{1}_{[1/3,2/3)}(\omega),\\
X_5(\omega) &= \mathbb{1}_{[2/3,1)}(\omega),
\end{aligned}$$

and so on. Take a moment to visualise what is happening and to convince yourselves that $X_n \xrightarrow{\mathbb{P}} 0$ *but* $X_n \nrightarrow X$ *almost surely.*

The strong law of large numbers can be proven using Ergodic theory or a clever application of truncation and analysis and is beyond our scope. Under additional moment assumptions it can also be proven from the weak law combined with the Borel-Cantelli lemma.

The weak law however, at least with additional assumptions, is easily within our grasp using only elementary probability.

Suppose for now that $X_1, X_2, \ldots,$ are i.i.d. with mean μ and finite variance $\mathrm{var}(X_i) < \infty$. We may assume w.l.o.g. that $\mathrm{var}(X_1) = 1$. Then a simple application of Markov's inequality, or Chebyshev in this case, shows that for any $\epsilon > 0$

$$\begin{aligned}
\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i - n\mu\right| > n\epsilon\right] &\leq \frac{\mathrm{var}\left(\sum X_i\right)}{\epsilon^2 n^2}\\
&\leq \frac{n\,\mathrm{var}(X_1)}{\epsilon^2 n^2} \leq \frac{1}{\epsilon^2 n},
\end{aligned}$$

where we used that for independent variables the variance of the sum is the sum of the variance.

Clearly this vanishes as $n \to \infty$ for any $\epsilon > 0$, and we have proven the w.l.l.n. under finite second moments. In other words we have proven that the difference between the average of i.i.d. r.v.s and their mean vanishes in probability.

The central limit theorem, probably the most distinctive result of probability theory, quantifies the fluctuations of the average from its mean. First we need the following definition.

Definition 3 (Convergence in distribution, weak convergence)**.** *A sequence of random variables* $\{X_i : i \in \mathbb{N}\}$ *is said to converge weakly, or in distribution, to the random variable* X*, denoted* $X_n \xrightarrow{\mathsf{D}} X$*, if*

$$\mathbb{P}[X_n \leq t] \to \mathbb{P}[X \leq t], \quad \textit{as } n \to \infty,$$

for all continuity points t *of the distribution function of* X*,* $t \mapsto \mathbb{P}[X \leq t]$*.*

Now we are ready to state the C.L.T.

Theorem 3 (Central Limit Theorem)**.** *Suppose that* $\{X_i : i \geq \mathbb{N}\}$ *is an i.i.d. collection of random variables such that* $\mathbb{E}[X_i] = \mu$ *and* $\mathrm{var}(X_i) = \sigma^2$*. Then*

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right] \xrightarrow{\mathsf{D}} \mathcal{N}(0, \sigma^2).$$

In other words

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx \mu + \frac{\xi}{\sqrt{n}}, \quad \xi \sim \mathcal{N}(0, \sigma^2).$$

Based on this one would then expect that

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq t\right] \lesssim \mathrm{e}^{-nt^2/2\sigma^2}, \tag{1.1}$$ {eq:gaussian_co

but in fact the CLT only provides us information at a coarser scale:

$$\mathbb{P}\left[\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq t\right] \approx \mathrm{e}^{-t^2/2\sigma^2}.$$

To get (1.2) one could attempt to replace t with $\sqrt{n}y$, but this is quite different to the conclusion of the C.L.T.

Moreover, we cannot expect (1.2) to hold in general under just finite second moments, as the above inequality directly implies existence of infinite moments, and even existence of all exponential moments.

In fact assuming just second moments the best we can do is use Chebyshev's inequality, that is if $\mathrm{var}(Y) < \infty$,

$$\mathbb{P}\left[|Y - \mathbb{E}\, Y| \geq t\right] \leq \frac{\mathrm{var}(Y)}{t^2},$$

which decays much more slowly that e^{-ct^2}. To see that in general Chebyshev's inequality is sharp, at least up to logarithmic terms, consider $Y \sim f_Y(\cdot)$, where for $c > 0$ and $\delta \in (0, 1)$,

$$f(\pm y) = \frac{c}{y^3 \log(y)^{1+\delta}}, \qquad |y| > 1.$$

Obviously Y is symmetric so that $\mathbb{E}\, Y = 0$ and it can be verified that Y has finite second moments, but $\mathbb{E}\,|Y|^3 = \infty$. An easy calculation shows that

$$\mathbb{P}(|Y| > t) \geq \frac{c'}{t^2 \log(t)^{1+\delta}},$$

so that the polynomial order of the bound achieved by Chebyshev's inequality is sharp.

On the other hand (1.2) is true for Gaussian random variables. We will now see that it also holds much more generally, under appropriate moment conditions.

1.2 Chernoff bounds

As we saw, assuming just two finite moments, we cannot hope in general to obtain tails decaying faster than $1/t^2$. We can however, if we assume existence of higher moments.

Suppose for example that $\mathbb{E}[|X|^k] < \infty$, for $k \geq 1$, then we automatically get using Markov's inequality that

$$\begin{aligned}\mathbb{P}[|X - \mathbb{E}\,X| > t] &= \mathbb{P}[|X - \mathbb{E}\,X|^k > t^k] \\ &\leq \frac{\mathbb{E}\left[|X - \mathbb{E}\,X|^k\right]}{t^k},\end{aligned}$$

where the equality follows from the fact that for $k > 0$ and $x > 0$, $x \mapsto x^k$ is strictly increasing, and the inequality from Markov's inequality.

Obviously the higher k is the faster the tails decay but then potentially the numerator will also grow. In fact for a fixed t the optimum bound that can be obtained using the above approach comes from optimising over k, that is

$$\mathbb{P}[|X - \mathbb{E}\,X| > t] \leq \inf_k \frac{\mathbb{E}\left[|X - \mathbb{E}\,X|^k\right]}{t^k}.$$

However, optimising the above, even with X Gaussian becomes cumbersome. What if we try instead the mapping $x \mapsto \mathrm{e}^x$ or more generally $x \mapsto \mathrm{e}^{\lambda x}$ for $\lambda \in \mathbb{R}$? This family of bounds is known collectively as Chernoff bounds. First we need a definition.

Definition 4 (Moment- and Cumulant-Generating Function)**.** *Given a random variable X such that $\mathbb{E}[\exp(\lambda X)] < \infty$ for all $\lambda \in (-b, b)$ for some $b > 0$, we define its Moment Generating Function (MGF) and its Cumulant Generating Function (CGF) respectively through*

$$M_X(\lambda) := \mathbb{E}[\exp(\lambda X)], \qquad \psi(\lambda) := \log \mathbb{E}[\exp(\lambda X - \mathbb{E}\,X)].$$

Lemma 1.2.1 (Chernoff Bound)**.** *Suppose that $M_X(\lambda) < \infty$ for all $\lambda \in \mathbb{R}$, and define the Legendre dual of ψ as*

$$\psi^*(t) := \sup_{\lambda \geq 0}[\lambda t - \psi(\lambda)].$$

Then for all $t \geq 0$

$$\mathbb{P}[X - \mathbb{E}\,X \geq t] \leq \mathrm{e}^{-\psi^*(t)}.$$

Proof. We start off as promised by considering $x \mapsto \exp(\lambda x)$, that is

$$\begin{aligned}\mathbb{P}[X - \mathbb{E}\,X > t] &\leq \mathbb{E}\left[\exp\left\{\lambda(X - \mathbb{E}\,X)\right\} > \exp(\lambda t\right] \\ &\leq \frac{\mathbb{E}\left[\exp\left\{\lambda(X - \mathbb{E}\,X)\right\}\right]}{\exp(\lambda t)} \\ &= \mathrm{e}^{\psi(\lambda) - \lambda t}.\end{aligned}$$

Since this holds for all λ it also follows that

$$\mathbb{P}[X - \mathbb{E}\,X > t] \leq \inf_\lambda \mathrm{e}^{\psi(\lambda) - \lambda t} = \mathrm{e}^{-\psi^*(t)}. \qquad \square$$

Notice that the above gives only the upper tail, but letting $Y = -X$, and using the *union bound*

$$\begin{aligned}\mathbb{P}[|X - \mathbb{E}\,X| > t] &\leq \mathbb{P}[X - \mathbb{E}\,X > t] + \mathbb{P}[X - \mathbb{E}\,X < -t] \\ &\leq \mathbb{P}[X - \mathbb{E}\,X > t] + \mathbb{P}[Y - \mathbb{E}\,Y > t],\end{aligned}$$

and we can bound each of the above terms as before.

Example 2. *Suppose now that $X \sim \mathcal{N}(0, \sigma^2)$. Then we know that $M_X(\lambda) = \mathrm{e}^{\lambda^2\sigma^2/2}$ and thus that $\psi(\lambda) = \lambda^2\sigma^2/2$. From this we can compute*

$$\psi^*(t) = \sup_\lambda \lambda t - \frac{\lambda^2\sigma^2}{2} = \frac{t^2}{2\sigma^2},$$

and thus the corresponding Chernoff bound is

$$\mathbb{P}[X \geq t] \leq \mathrm{e}^{-t^2/2\sigma^2},$$

which indeed attains the desired $C\mathrm{e}^{-Ct^2}$ tail decay.

Example 3. *Going back to our running example suppose that $\{X_i : i = 1, \ldots, n\}$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ and let $Y = \frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \sigma^2/n)$. Then using independence we have that $\psi_Y(\lambda) = \lambda^2\sigma^2/2n$, whence $\psi_Y^*(t) = nt^2/(2\sigma^2)$ and thus*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq t\right] \leq 2\mathrm{e}^{-nt^2/2\sigma^2}. \tag{1.2}$$ {eq:gaussian_co

1.3 Sub-gaussian random variables.

Inspecting the proof and the two examples in the last section we can see that what buys as the correct rate is the behaviour of the CGF. That is suppose that we know that for some random variable X we have $\psi_X(\lambda) \leq \sigma^2\lambda^2/2$, then we conclude that

$$\psi_Y^*(t) = \sum_{\lambda \geq 0} \left[\lambda t - \psi_Y(\lambda)\right] \geq \sup_\lambda \left[\lambda t - \sigma^2\lambda^2\right] = \frac{t^2}{2\sigma^2}.$$

This motivates the following definition.

Definition 5 (Sub-gaussian random variable)**.** *Let $\sigma > 0$. The random variable X, with mean $\mu = \mathbb{E}[X]$, is said to be σ-sub-Gaussian, or sub-Gaussian with variance proxy σ, if*

$$\psi_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2}, \quad \text{for all } \lambda \in \mathbb{R}.$$

Definition 6 (Sub-gaussian random vector)**.** *Let $\sigma > 0$. The random vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$ is a σ^2-sub-Gaussian vector in $\mathbb{R}^n$ if for any unit vector $u \in \mathbb{S}^{n-1}$ the random variable $u^\mathsf{T}\boldsymbol{\epsilon}$ is σ^2-sub-Gaussian.*

The discussion above then shows that if X is σ-sub-Gaussian, then it satisfies the upper-deviation inequality

$$\mathbb{P}[X \geq \mu + t] \leq \mathrm{e}^{-\frac{t^2}{2\sigma^2}}.$$

From the definition it easily follows that $-X$ must also be σ-sub-Gaussian and thus we also have that

$$\mathbb{P}[X \leq \mu - t] = \mathbb{P}[-X \geq -\mu + t] \leq \mathrm{e}^{-\frac{t^2}{2\sigma^2}},$$

and thus we see that X must also satisfy the concentration inequality

$$\mathbb{P}[|X - \mu| \geq t] \leq 2\mathrm{e}^{-\frac{t^2}{2\sigma^2}}.$$

Example 4 (Rademacher random variables are sub-Gaussian)**.** *Let X be a Rademacher random variable, that is $\mathbb{P}(X = \pm 1) = 1/2$. Then X is 1-sub-Gaussian.*

Proof 1. It is obvious that $M_X(\lambda) = (\mathrm{e}^{\lambda} + \mathrm{e}^{-\lambda})/2 = \cosh(x)$, and thus that $\psi_X(\lambda) = \log(\mathrm{e}^{\lambda} + \mathrm{e}^{-\lambda}) - \log(2)$. Also notice that $\psi_X(0) = 0$ and $\psi_X(1) = 0$ and since X is symmetric $\psi_X(\lambda) = \psi_X(-\lambda)$.

We next compute for $\lambda > 0$

$$\psi'_X(x) = \frac{\mathrm{e}^x - \mathrm{e}^{-x}}{\mathrm{e}^x + \mathrm{e}^{-x}} = 1 - \frac{2\mathrm{e}^{-x}}{\mathrm{e}^x + \mathrm{e}^{-\lambda}}$$
$$\psi''_X(x) = 4\frac{\mathrm{e}^{2x}}{(\mathrm{e}^{2x}+1)^2} \leq 1$$

since the function $y \mapsto y/(y+1)^2$ is decreasing for $y \geq 1$.

Combining everything we have for $\lambda > 0$

$$\begin{aligned}
\psi_X(\lambda) &= \psi_X(0) + \int_0^{\lambda} \psi'_X(r)\mathrm{d}r \\
&= \psi_X(0) + \lambda\psi'_X(0) + \int_0^{\lambda}\int_0^{r} \psi''_X(s)\mathrm{d}s\mathrm{d}r \\
&= 0 + \lambda \times 0 + \int_0^{\lambda}\int_0^{r} \psi''_X(s)\mathrm{d}s\mathrm{d}r \\
&\leq \int_0^{\lambda}\int_0^{r} 1\mathrm{d}s\mathrm{d}r \leq \frac{\lambda^2}{2}.
\end{aligned}$$

□

Proof 2. (from Wainwright 2019). We have

$$\begin{aligned}
\frac{1}{2}\left[\mathrm{e}^{\lambda} + \mathrm{e}^{-\lambda}\right] &= \frac{1}{2}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} + \sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!} = \frac{1}{2}\sum_{k=0}^{\infty}\frac{\lambda^k + (-\lambda)^k}{k!} \\
&= \frac{1}{2}\sum_{k=0,k\in 2\mathbb{Z}}^{\infty}\frac{2\lambda^k}{k!} = \sum_{m=0}^{\infty}\frac{\lambda^{2m}}{(2m)!} \\
&= 1 + \sum_{m=1}^{\infty}\frac{\lambda^{2m}}{(2m)!} \leq 1 + \sum_{m=1}^{\infty}\frac{\lambda^{2m}}{2^m m!}
\end{aligned}$$

since clearly for $m \geq 1$, $(2m)! \geq 2^m m!$

$$= \sum_{m=0}^{\infty}\frac{\lambda^{2m}}{2^m m!} = \mathrm{e}^{\lambda^2/2}.$$

□

Example 5 (Bounded random variables)**.** *Let X take values in the interval $[a, b]$, for $a < b$. Then X is $(b-a)/2$-sub-Gaussian.*

Proof. (Symmetrisation argument, from Wainwright 2019). Let X' be an independent copy of X and notice that

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\lambda(X - \mathbb{E}[X])\right\}\right] &= \mathbb{E}\left[\exp\left\{\lambda(X - \mathbb{E}[X'])\right\}\right] \\
&= \mathbb{E}\left[\mathrm{e}^{\lambda X}\mathrm{e}^{\lambda\,\mathbb{E}[-X']}\right]
\end{aligned}$$

and since by Jensen's inequality $\mathrm{e}^{\lambda\,\mathbb{E}[-X']} \leq \mathbb{E}\left\{\mathrm{e}^{\lambda(-X')}\right\}$

$$\leq \mathbb{E}\left[\mathrm{e}^{\lambda X}\,\mathbb{E}\left\{\mathrm{e}^{\lambda(-X')}\right\}\right]$$

$$\leq \mathbb{E}\left[\mathrm{e}^{\lambda(X-X')}\right],$$

where the last equality follows by independence.

The important thing is that we now have to deal with the random variable $Y := X - X'$ whose distribution is symmetric, that is $Y \overset{\mathsf{D}}{=} -Y$. Stated different, if ξ is a Rademacher random variable, then $Y \overset{\mathsf{D}}{=} \xi Y$ and therefore

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{\lambda(X-X')}\right] &= \mathbb{E}\left[\mathrm{e}^{\lambda Y}\right] = \mathbb{E}\left[\mathrm{e}^{\lambda\xi Y}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left\{\mathrm{e}^{\lambda\xi Y}\middle| Y\right\}\right] \leq \mathbb{E}\left[\mathbb{E}\left\{\mathrm{e}^{\lambda^2 Y^2/2}\middle| Y\right\}\right] = \mathbb{E}\left[\mathrm{e}^{\lambda^2 Y^2/2}\right],\end{aligned}$$

where we applied the result of the previous example, that is $\mathbb{E}[\mathrm{e}^{\lambda'\xi}] \leq \mathrm{e}^{(\lambda')^2/2}$, conditionally on Y, with $\lambda' = \lambda Y$. Finally since $X, X' \in [a, b]$, we have that $|X - X'| \leq b - a$ and thus

$$\mathbb{E}\left[\mathrm{e}^{\lambda(X-X')}\right] \leq \mathbb{E}\left[\mathrm{e}^{\lambda^2 Y^2/2}\right] \leq \mathrm{e}^{\lambda^2(b-a)^2/2},$$

and thus $\psi_X(\lambda) \leq (b-a)^2\lambda^2/2$. □

Just like a linear combination of independent Gaussian random variables is also Gaussian, the sub-Gaussian property is also preserved by linear operations. It is an easy exercise to verify that if X_1, X_2 are σ_1- and σ_2-sub-Gaussian respectively then $X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$-sub-Gaussian. This leads us to the famous Hoeffding bound.

Theorem 4 (Hoeffding bound)**.** *Suppose that $\{X_i\}_{i=1}^n$ are independent random variables, where X_i has mean μ_i and is σ_i-sub-Gaussian. Then for all $t \geq 0$ we have*

$$\mathbb{P}\left[\sum_{i=1}^n (X_i - \mu_i) \geq t\right] \leq \exp\left(-\frac{t}{2\sum_{i=1}^n \sigma_i^2}\right).$$

1.3.1 Characterisations of sub-Gaussian random variables

We based our search for ways to generalise Gaussian tail bounds to non-Gaussian variables on the condition $\psi_X(\lambda) \leq \lambda^2\sigma^2/2$, which could be used to directly lower bound $\psi_X^*(t)$ in Chernoff's bound. However as we shall see now there is a number of equivalent characterisations of sub-Gaussian random variables. The presentation is largely taken from Van Handel 2016.

{thm:subGtfae}

Theorem 5. *Let X be a centred random variable. TFAE*

(a) There is a constant $\sigma \geq 0$ such that

$$\mathbb{E}[\mathrm{e}^{\lambda X}] \leq \mathrm{e}^{\lambda^2\sigma^2/2} \quad \textit{for all } \lambda \in \mathbb{R};$$

(b) There exist a universal constant $c > 0$ such that

$$\mathbb{P}[|X| \geq s] \leq 2\exp\left(-\frac{s^2}{2c\sigma^2}\right), \quad \textit{for all } s > 0;$$

(c) There exist a universal constant $c > 0$ such that

$$\mathbb{E}\left[\exp\left(\frac{X^2}{c\sigma^2}\right)\right] \leq 2.$$

(d) There exist a universal constant $c > 0$ such that

$$\mathbb{E}[X^{2k}] \leq (c\sigma^2)^q q!;$$

Proof. $(a) \Rightarrow (b)$: We have

$$\mathbb{P}[|X| \geq s] \leq \mathbb{P}[X \geq s] + \mathbb{P}[X \leq -s],$$

let us treat the first term; let $\lambda \in \mathbb{R}$

$$\begin{aligned}\mathbb{P}[X \geq s] \leq \mathbb{P}[\exp(\lambda X) \geq \lambda s] &\leq \frac{\mathbb{E}\exp(\lambda X)}{\mathrm{e}^{\lambda s}} \\ &\leq \exp\left\{\frac{\lambda^2\sigma^2}{2} - \lambda s\right\}\end{aligned}$$

and optimising the bound over λ, we set $\lambda^* = s/\sigma^2$ to get

$$\leq \exp\left\{-\frac{s}{2\sigma^2}\right\}.$$

Similarly we get

$$\mathbb{P}[X \leq -s] \leq \exp\left\{-\frac{s}{2\sigma^2}\right\},$$

by applying the previous calculation to $-X$ and (b) follows.

$(b) \Rightarrow (c)$: Suppose that

$$\mathbb{P}[|X| \geq s] \leq 2\exp\left(-\frac{s^2}{2\sigma^2}\right).$$

We will use the following fact, if Y is a positive random variable with distribution function F_Y, and f increasing and differentiable then

$$\begin{aligned}\mathbb{E}[f(Y)] = \int_0^\infty f(y)F(\mathrm{d}y) &= \int_0^\infty f(0)F(\mathrm{d}y) + \int_0^\infty \int_{s=0}^y f'(s)F(\mathrm{d}y) \\ &= f(0) + \int_{s=0}^\infty \int_{y=s}^\infty F(\mathrm{d}y)f'(s)\mathrm{d}s = f(0) + \int_{s=0}^\infty \mathbb{P}[Y \geq s]f'(s)\mathrm{d}s.\end{aligned}$$

Then we have, letting $Y = X^2$

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{X^2}{c\sigma^2}\right)\right] &= \int_{-\infty}^\infty \exp\left(\frac{x^2}{c\sigma^2}\right) F_X(\mathrm{d}x) \\ &= 1 + \int_0^\infty \frac{1}{c\sigma^2}\exp\left(\frac{y}{c\sigma^2}\right)\mathbb{P}[|X|^2 > y]\mathrm{d}y \\ &\leq 1 + \frac{2}{c\sigma^2}\int_0^\infty \exp\left(\frac{y}{c\sigma^2}\right)\exp\left(-\frac{y}{2\sigma^2}\right)\mathrm{d}y \\ &\leq 1 + \frac{2}{c\sigma^2}\int_0^\infty \exp\left[-\left(\frac{c-2}{2c}\right)\frac{y}{\sigma^2}\right]\mathrm{d}y\end{aligned}$$

which is finite for $c > 2$

$$= 1 + \frac{2}{c\sigma^2}\frac{2c\sigma^2}{c-2} = 1 + \frac{4}{c-2}.$$

Choosing $c = 6$ we get (c).

$(c) \Rightarrow (a)$: Assume w.l.o.g. that $c = 1$. Then using the fact that for $x > 0$, $\mathrm{e}^x \geq 1 + x^k/k!$

$$2 \geq \mathbb{E}\left[\exp\left(\frac{X^2}{\sigma^2}\right)\right] \geq 1 + \mathbb{E}\left[\frac{X^{2q}}{\sigma^{2q} q!}\right]$$

whence we have $\mathbb{E}[X^{2q}] \leq \sigma^{2q} q!$.

$(d) \Rightarrow (a)$: Let X' be an independent copy of X. Then for $Y = X - X'$, notice by the c_r-inequality that $\mathbb{E}[Y^{2k}] \leq 2^{2k}\,\mathbb{E}[X^{2k}]$. Thus

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{\lambda Y}] &= \sum_{k=0}^{\infty} \frac{\lambda^k\,\mathbb{E}[Y^k]}{k!} \\ &= \sum_{k=0}^{\infty} \frac{\lambda^{2k}\,\mathbb{E}[Y^{2k}]}{(2k)!} + \sum_{k=0}^{\infty} \frac{\lambda^{2k+1}\,\mathbb{E}[Y^{2k+1}]}{(2k+1)!}\end{aligned}$$

and since Y is symmetric

$$\begin{aligned}&\leq \sum_{k=0}^{\infty} \frac{\lambda^{2k}\,\mathbb{E}[Y^{2k}]}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{\lambda^{2k} 2^{2k} \sigma^{2k} k!}{(2k)!} \\ &\leq \sum_{k=0}^{\infty} \frac{\lambda^{2k} (4\sigma^2)^k k!}{k! k!} \\ &\leq \sum_{k=0}^{\infty} \frac{\lambda^{2k} (4\sigma^2)^k}{k!} \leq \exp\left(4\lambda^2\sigma^2\right).\end{aligned}$$

Also since $\mathbb{E}\,X = 0$, letting X' be an independent copy of X we have

$$\mathbb{E}[\mathrm{e}^{\lambda X}] = \mathbb{E}[\mathrm{e}^{\lambda X - \lambda\,\mathbb{E}[X']}] \leq \mathbb{E}[\mathrm{e}^{\lambda X - \lambda X'}] = \mathbb{E}[\mathrm{e}^{\lambda Y}],$$

by Jensen's inequality.

Combining everything we thus get

$$M_X(\lambda) \leq \exp\left(4\lambda^2\sigma^2\right). \qquad \square$$

Useful lemmas

{lem:subGvecto

Lemma 1.3.1. *Suppose that $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$, where the variables $\epsilon_i, i = 1, \ldots, n$ are independent and σ^2-sub-Gaussian. Then ϵ is a σ^2-sub-Gaussian vector in $\mathbb{R}^n$. That is for any $v \in \mathbb{S}^{n-1}$, $\epsilon^{\mathsf{T}} v$ is σ^2-sub-Gaussian.*

Proof of Lemma 1.3.1. Obvious since

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda \epsilon^{\mathsf{T}} v\right)\right] &= \mathbb{E}\left[\exp\left(\lambda \sum \epsilon_i v_i\right)\right] \\ &\leq \prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda \epsilon_i v_i)\right] \\ &\leq \prod_{i=1}^{n} \exp\left(\frac{\lambda^2\sigma^2}{2} v_i^2\right) = \exp(\frac{\lambda^2\sigma^2}{2}).\end{aligned}$$

$\square$

{lem:maxbound}

Lemma 1.3.2 (Maximum of n sub-Gaussian random variables)**.** *Let $X_1, \dots, X_n$ be n, centred, σ^2-sub-Gaussian random variables. Then for any $t > 0$ we have*

$$\mathbb{P}\left(\max_{i=1,\dots,N} X_i > t\right) \leq N\mathrm{e}^{-\frac{t^2}{2\sigma^2}}, \quad \mathbb{E}[\max_{i=1,\dots,N} X_i] \leq \sigma\sqrt{2\log N}.$$

Proof. The probability bound holds simply by a union bound, that is

$$\mathbb{P}\left(\max i = 1, \dots, nX_i > t\right) \leq \sum_{i=1}^{n} \mathbb{P}\left[X_i > t\right].$$

For the expectation we have

$$\begin{aligned}
\mathbb{E}[\max X_i] &= \frac{1}{s}\,\mathbb{E}[\log\exp(s\max X_i)] \\
&\leq \frac{1}{s}\log\mathbb{E}[\exp(s\max X_i)] \\
&= \frac{1}{s}\log\mathbb{E}[\max\exp(sX_i)] \\
&= \frac{1}{s}\log N\exp\left(\frac{\sigma^2 s^2}{2}\right) \\
&= \frac{1}{s}\log N + \frac{\sigma^2 s}{2},
\end{aligned}$$

and setting $s = \sqrt{2\log n/\sigma^2}$ gives the result. □

1.3.2 Gaussian Concentration

Theorem 6 (Gaussian Concentration)**.** *Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(0,1)$ and let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be K-Lipschitz. Then $f(X_1, \dots, X_n)$ is K-sub-Gaussian.* {thm:gauss_conc

The proof is taken from Lalley's notes.

Proof. It suffices to prove the result for smooth Lipschitz functions and extend it with an approximation argument.

Notice that we want to prove that

$$\log\mathbb{E}[\exp(\lambda f(\boldsymbol{X}) - \lambda\,\mathbb{E}\,f(\boldsymbol{X}))] \leq \frac{\lambda^2 K^2}{2}.$$

By the standard symmetrisation trick and Jensen's inequality it suffices to show that

$$\log\mathbb{E}[\exp(\lambda f(\boldsymbol{X}) - \lambda f(\boldsymbol{X}'))] \leq \frac{\lambda^2 K^2}{2},$$

where $\boldsymbol{X}'$ is an independent copy of $\boldsymbol{X}$. Now we will form a smooth path $\{\boldsymbol{X}_t : t \in [0,1]\}$ connecting $\boldsymbol{X}$ and $\boldsymbol{X}'$, and in particular we will take one such that $\boldsymbol{X}_t \sim \mathcal{N}(\mathbb{O}, \mathbb{1}_n)$. So let

$$\boldsymbol{X}_t := \cos(\pi t/2)\boldsymbol{X} + \sin(\pi t/2)\boldsymbol{X}',$$

and we compute

$$\frac{\mathrm{d}\boldsymbol{X}_t}{\mathrm{d}t} = -\frac{\pi}{2}\sin(\pi t/2)\boldsymbol{X} + \frac{\pi t}{2}\cos(\pi t/2)\boldsymbol{X}' =: \frac{\pi}{2}\boldsymbol{Y}_t.$$

What's also important is that $\boldsymbol{Y}_t$ is uncorrelated and thus independent of $\boldsymbol{X}_t$.

The FTC then gives

$$\begin{aligned}\mathbb{E}\{\exp[\lambda f(\boldsymbol{X}) - \lambda f(\boldsymbol{X}']\} &= \mathbb{E}\left\{\exp\left[\lambda \int_0^1 \mathrm{d}t \nabla f(\boldsymbol{X}_t) \frac{\mathrm{d}\boldsymbol{X}_t}{\mathrm{d}t}\right]\right\} \\ &= \mathbb{E}\left\{\exp\left[\frac{\pi}{2}\lambda \int_0^1 \mathrm{d}t \langle \nabla f(\boldsymbol{X}_t), Y_t\rangle\right]\right\} \\ &\leq \int_0^1 \mathrm{d}t\, \mathbb{E}\left\{\exp\left[\frac{\pi}{2}\lambda \langle \nabla f(\boldsymbol{X}_t), Y_t\rangle\right]\right\}.\end{aligned}$$

Since as explained earlier, $\boldsymbol{Y}_t$ is independent of $\boldsymbol{X}_t$, conditionally on $\boldsymbol{X}_t$, we have that $\langle \nabla f(\boldsymbol{X}_t), Y_t\rangle$ is $\|\nabla f(\boldsymbol{X}_t)\|^2$-sub-Gaussian, so overall K^2-sub-Gaussian. Therefore

$$\begin{aligned}\mathbb{E}\{\exp[\lambda f(\boldsymbol{X}) - \lambda f(\boldsymbol{X}']\} &\leq \int_0^1 \mathrm{d}t\, \mathbb{E}\left\{\exp\left[\frac{\pi}{2}\lambda \langle \nabla f(\boldsymbol{X}_t), Y_t\rangle\right]\right\} \\ &\leq \int_0^1 \mathrm{d}t \exp\left(\frac{\pi^2 \lambda^2 K^2}{2}\right),\end{aligned}$$

and the conclusion follows. □

Chapter 2

High-dimensional regression and Lasso

This chapter is based on Rigollet and Hütter 2017 and Wainwright (2019).

Let $\theta^* \in \mathbb{R}^d$ be an unknown vector of parameters. Our observation model is the following

$$y = \boldsymbol{X}\theta^* + \epsilon, \tag{2.1}$$ {eq:regression}

where we observe the vector $y \in \mathbb{R}^n$ and the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, whereas ϵ is an $\mathbb{R}^n$ noise vector. Our standing assumption will be that $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ where ϵ_i is centred, σ^2-sub-Gaussian for some $\sigma^2 > 0$, and this will allow us to use the machinery we developed in Chapter 2.

We will frequently distinguish between *fixed design*, where $\boldsymbol{X}$ is deterministic, and *random design*, where $\boldsymbol{X}$ is random. We first consider fixed design.

2.0.1 Warm-up

In classical linear regression one considers the setup where $d \ll n$. There the typical approach is to use Ordinary Least Squares that is attempt to recover θ^* by solving the convex minimisation problem

$$\hat{\theta}^{\mathsf{LS}} := \arg\min_{\theta} (y - \boldsymbol{X}\theta)^{\mathsf{T}}(y - \boldsymbol{X}\theta) = \arg\min_{\theta} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{d} \boldsymbol{X}_{ij}\theta_j\Big)^2. \tag{2.2}$$ {eq:OLS}

Since the quadratic loss function is convex we only need to check the first order condition of optimality:

$$\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\hat{\theta}^{\mathsf{LS}} = \boldsymbol{X}^{\mathsf{T}}y.$$

Under the usual assumption that $\boldsymbol{X}$ has *full rank*, that is its columns are linearly independent, it follows easily that $X^{\mathsf{T}}X$ is positive definite. To see why notice that for any vector $v \in \mathbb{R}^d$, $\boldsymbol{X}v$ is a linear combination of the column vectors of $\boldsymbol{X}$ and is thus non-zero unless $v = 0$. This implies that $v\boldsymbol{X}^{\mathsf{T}}Xv > 0$ for any non-zero v and that $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ is invertible and thus we can write the solution as $\hat{\theta}^{\mathsf{LS}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}y$.

All of the above, depends crucially on the assumption that $\boldsymbol{X}$ has full rank. In fact, this can only happen if $n > d$.

When $d \geq n$ the solution the first order condition is still the same, but we can no longer hope to find a unique solution since the linear system in (2.2) is now under-determined. Instead we will get a linear subspace of solutions. To select one we may to impose an

additional criterion, or some form of *regularization*; e.g. we may ask for the solution of (2.2) with the minimum norm, that is the solution to the regularized problem

$$\min_{\theta} \|\theta\|_2^2, \qquad \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\hat{\theta}^{\mathsf{LS}} = \boldsymbol{X}^{\mathsf{T}}y.$$

It turns out that 2.0.1 always exists and is unique and is known as the *Moore-Penrose* inverse denoted by $(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{\dagger}$.

So let us now see how the performance of the least squares estimators changes with the dimension. At this point we should point out that there are different ways of measuring performance. In particular one may be mostly interested in recovering the vector θ^* or in predicting the value of the response variable from the values of the independent variable, that is the *prediction error*.

As motivation for what's to come we first consider the prediction error. We will focus on the *Mean Squared Error(MSE)* of the prediction that is

$$\mathsf{MSE}(\boldsymbol{X}\hat{\theta}^{\mathsf{LS}}) = \frac{1}{n}\|\boldsymbol{X}\theta^* - \boldsymbol{X}\hat{\theta}^{\mathsf{LS}}\|^2 = (\theta^* - \hat{\theta}^{\mathsf{LS}})^{\mathsf{T}}\frac{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}}{n}(\theta^* - \hat{\theta}^{\mathsf{LS}}).$$

First of all notice that since $\hat{\theta}^{\mathsf{LS}}$ solves (2.2) we have that

$$\|y - \boldsymbol{X}\hat{\theta}^{\mathsf{LS}}\|_2^2 \leq \|y - \boldsymbol{X}\theta^*\| = \|\epsilon^2\|, \tag{2.3}$$ {eq:fundamental

and that

$$\|y - \boldsymbol{X}\hat{\theta}^{\mathsf{LS}}\|_2^2 = \|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|_2^2 + \|\epsilon\|_2^2 - 2\epsilon^{\mathsf{T}}\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*),$$

whence we get that

$$\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|_2^2 \leq 2\epsilon^{\mathsf{T}}\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*) = 2\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|\frac{\epsilon^{\mathsf{T}}\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)}{\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|},$$

and after simplifying

$$\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\| \leq 2\frac{\epsilon^{\mathsf{T}}\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)}{\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|}.$$

There are two issues here; first $\hat{\theta}^{\mathsf{LS}}$ clearly depends on ϵ, and second the right hand side also depends on the unknown θ^*.

One way around this is to consider the worst case scenario, that is to use the bound

$$\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\| \leq 2\frac{\epsilon^{\mathsf{T}}\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)}{\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|} \leq \sup_{v\in\mathbb{R}^n:\|v\|=1} \epsilon^{\mathsf{T}}v = 2\epsilon^{\mathsf{T}}\epsilon = \|\epsilon\|_2^2. \tag{2.4}$$ {eq:takingsups}

Since ϵ is a σ^2-sub-Gaussian $\mathbb{R}^n$-vector we know from Theorem 5 that for some universal constant $c > 0$ we have

$$\mathbb{E}[\|\epsilon\|_2^2] = \sum_{i=1}^{n} \mathbb{E}[\epsilon_i^2] \leq c\sigma^2 n.$$

From this we conclude that

$$\mathbb{E}\left[\mathsf{MSE}(\boldsymbol{X}\hat{\theta}^{\mathsf{LS}})\right] \leq 2c\sigma^2\frac{n}{n} = 2c\sigma^2.$$

The question is how wasteful we have been, in particular in (2.4) when we took the supremum over the l^2-ball in $\mathbb{R}^n$. Thinking about it, we have implicitly assumed that

$\mathsf{Im}(\boldsymbol{X}) = \{y \in \mathbb{R}^n : y = \boldsymbol{X}v\}$ is all of $\mathbb{R}^n$, which is equivalent to $\boldsymbol{X}$ having rank n. What happens when the rank of $\boldsymbol{X}$ is lower than n, say r? Can we use this to get a better bound? The answer turns out to be positive.

Suppose then that $\mathsf{Im}(\boldsymbol{X})$ is r-dimensional. Let $\{\mathbb{e}_i; i = 1, \ldots, n\}$ be the standard basis and let $\{\mathbb{h}_i : i = 1, \ldots, n\}$ be an orthonormal basis such that the first r vectors $\{\mathbb{h}_i : i = 1, \ldots, r\}$ form an orthonormal basis of $\mathsf{Im}(\boldsymbol{X})$. That is any element of $\mathsf{Im}(\boldsymbol{X})$ expressed as a column vector in the $\mathbb{H}$ basis has its last $n - r$ elements all zero. Let $\mathbb{h}_j = \sum_{k=1}^n h_j^k \mathbb{e}_k$, define the row vectors $h_j = (h_j^i)_{i=1}^n$ and define the matrix

$$O = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{pmatrix},$$

that is the matrix with rows equal to $h_1, \ldots, h_n$. Then for any $v \in \mathbb{R}^n$, of the form $v = \sum_j \alpha_j \mathbb{e}_j$ the vector $O\alpha$ allows us to express v in terms of the $\mathbb{H}$ basis. In particular, by definition of O we have $P_r O X v = O X v$ for any $v \in \mathbb{R}^n$, where

$$P_r = \left(\begin{array}{c|c} \mathbb{1}_r & \mathbb{O} \\ \hline \mathbb{O} & \mathbb{O} \end{array}\right).$$

Also, since $\mathbb{H}$ is orthonormal, an easy calculation shows that O must be orthogonal and in particular for any vectors $v, w \in \mathbb{R}^n$, we have $v^\mathsf{T} w = (Ov)^\mathsf{T}(Ow)$.

Therefore going back to our calculation we have, since for any v, $\|Ov\| = \|v\|$ we have

$$\begin{aligned}
\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\| &\leq 2\frac{\epsilon^\mathsf{T}\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)}{\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|} = 2\frac{(O\epsilon)^\mathsf{T} O\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)}{\|O\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|} \\
&= 2\frac{(O\epsilon)^\mathsf{T} P_r O\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)}{\|P_r O\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|} = 2\frac{\epsilon^\mathsf{T} O^\mathsf{T} P_r O\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)}{\|P_r O\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|} \\
&= 2\frac{\epsilon^\mathsf{T} O^\mathsf{T} P_r^\mathsf{T} P_r O\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)}{\|P_r O\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|} \leq 2\|P_r O\epsilon\|,
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality.

Assume $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$, where ϵ_i are independent mean-zero, σ^2-sub-Gaussian variables. Then we can write $(O\epsilon)_j = \left(\sum_{k=1}^n h_j^k \epsilon_j\right)$ and therefore

$$\begin{aligned}
\mathbb{E}\left[\|P_r O\epsilon\|^2\right] &= \mathbb{E}\left[\sum_{j=1}^r \left(\sum_{k=1}^n h_j^k \epsilon_k\right)^2\right] \\
&= \sum_{j=1}^r \sum_{k=1}^n \sum_{l=1}^n h_j^k h_j^l \,\mathbb{E}\left[\epsilon_k \epsilon_l\right] = \sum_{j=1}^r \sum_{k=1}^n (h_j^k)^2 = \sum_{k=1}^r (O^\mathsf{T} O)_{kk} = r,
\end{aligned}$$

using independence and the zero mean property.

In the scenario where $\epsilon \sim \mathcal{N}(0, \sigma \mathbb{1}_n)$ we can go even further and conclude that $P_r O\epsilon$ is a σ^2-sub-Gaussian vector in $\mathbb{R}^r$; since O is orthogonal, $O\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_n)$ and therefore $P_r O\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_r)$. Therefore in the case where $\epsilon \sim \mathcal{N}(0, \sigma \mathbb{1}_n)$ we can prove that for some universal constant $c > 0$, with probability at least $1 - \delta$ we have

$$\mathsf{MSE}(\boldsymbol{X}\hat{\theta}^{\mathsf{LS}}) \leq c\sigma^2 \frac{\mathsf{rank}(\boldsymbol{X}) + \log(1/\delta)}{n}.$$

We will now see that we can get similar control in the general case where ϵ is centred σ^2-sub-Gaussian vector in $\mathbb{R}^n$.

To proceed recall that we are trying to control $\epsilon^\mathsf{T}\boldsymbol{X}\phi/\|\boldsymbol{X}\phi\|$, where $\mathsf{Im}(\boldsymbol{X})$ is r-dimensional. As before write $\mathbb{H} = (\mathbb{h})_{i=1}^r$ for an orthonormal basis of $\mathsf{Im}(\boldsymbol{X})$ where we can express the vectors $\mathbb{h}_i$ in the standard basis as

$$\mathbb{h}_i = \sum_{k=1}^n h_i^k \mathbb{e}_k, \quad i = 1, \dots, r.$$

For $i = 1, \dots, r$ define the column vectors $\phi_i = (h_i^1, \dots, h_i^n)^\mathsf{T}$ and let Φ be the $n \times r$ matrix $\Phi = (\phi_1, \dots, \phi_r)$. Notice that $(\Phi^\mathsf{T}\Phi)_{ij} = \langle \mathbb{h}_i, \mathbb{h}_j \rangle = \delta_{ij}$ and that for any $v \in \mathsf{Im}(\boldsymbol{X})$ we have

$$v = \sum_{j=1}^r \alpha_j \mathbb{h}_j = \sum_{j=1}^r \alpha_j \sum_{k=1}^n h_j^k \mathbb{e}_k = \sum_{k=1}^n \left(\sum_{j=1}^r \alpha_j h_j^l \right) \mathbb{e}_k.$$

Therefore we have that for any $v \in \mathsf{Im}(\boldsymbol{X})$

$$\Phi[v]_\mathbb{H} = [v]_\mathbb{E},$$

where for any basis $\mathbb{K}$, $[v]_\mathbb{K}$ denotes the vector of coefficients of v when expressed in the basis $\mathbb{K}$. In particular, any $v \in \mathsf{Im}(\boldsymbol{X})$ can be written as $\Phi\nu$ for some $\nu \in \mathbb{R}^r$.

Therefore, for any $\phi \in \mathbb{R}^n$, there is some $\nu \in \mathbb{R}^r$ such that

$$\frac{\epsilon^\mathsf{T}\boldsymbol{X}\phi}{\|\boldsymbol{X}\phi\|} = \frac{\epsilon^\mathsf{T}\Phi\nu}{\|\Phi\nu\|} = \frac{\langle \Phi^\mathsf{T}\epsilon, \nu \rangle}{\langle \Phi\nu, \Phi\nu \rangle^{1/2}} = \frac{\langle \Phi^\mathsf{T}\epsilon, \nu \rangle}{\langle \Phi^\mathsf{T}\Phi\nu, \nu \rangle^{1/2}} = \frac{\langle \Phi^\mathsf{T}\epsilon, \nu \rangle}{\|\nu\|},$$

since $\Phi^\mathsf{T}\Phi = \mathbb{1}_r$ and $\nu \in \mathbb{R}^r$.

Next for $u \in \mathbb{S}^{r-1}$ we can always write

$$\mathbb{E} \exp\left[\lambda \langle u, \Phi^\mathsf{T}\epsilon \rangle\right] = \mathbb{E} \exp\left[\lambda \langle \Phi u, \epsilon \rangle\right]$$

and since $\Phi^\mathsf{T}\Phi = \mathbb{1}_r$ we have that for any $u \in \mathbb{S}^{r-1}$ we have

$$\|\Phi u\|^2 = \langle \Phi u, \Phi u \rangle = \langle \Phi^\mathsf{T}\Phi u, u \rangle = \|u\|^2 = 1,$$

and therefore $\Phi u \in \mathbb{S}^{n-1}$. Since, Φu is a unit vector, and by assumption ϵ is a σ^2-sub-Gaussian random vector in $\mathbb{R}^n$, by definition we have that $\langle \Phi u, \epsilon \rangle$ is σ^2-sub-Gaussian and therefore

$$\mathbb{E} \exp\left[\lambda \langle u, \Phi^\mathsf{T}\epsilon \rangle\right] = \mathbb{E} \exp\left[\lambda \langle \Phi u, \epsilon \rangle\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

Since this holds for any $u \in \mathbb{S}^{r-1}$ we conclude that $\Phi^\mathsf{T}\epsilon$ is a σ^2-sub-Gaussian vector in $\mathbb{R}^r$ (where as ϵ is in $\mathbb{R}^n$).

Thus, finally we have to consider

$$\frac{\epsilon^\mathsf{T}\boldsymbol{X}\phi}{\|\boldsymbol{X}\phi\|} \leq \sup_{\nu \in \mathbb{S}^{r-1}} \frac{\langle \tilde{\epsilon}, \nu \rangle}{\|\nu\|} \leq \sup_{v \in \mathbb{R}^r, \|v\| \leq 1} \langle \tilde{\epsilon}, v \rangle,$$

where $\tilde{\epsilon}$ is σ^2-sub-Gaussian in $\mathbb{R}^r$. The conclusion comes from the following theorem.

Theorem 7. *Let $\boldsymbol{X}$ be a σ^2-sub-Gaussian vector in $\mathbb{R}^d$. Then for any $\delta > 0$, with probability at least $1 - \delta$ we have*

$$\sup_{\theta \in B_1^d(0)} \langle \theta, \boldsymbol{X} \rangle \leq 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}.$$

Proof. The idea is quite common for controlling suprema and is the following; let $\mathcal{N} \subset B_1^d(0)$ be a $1/2$-net of $B_1^d(0)$, that is $\mathcal{N} = \{x_1, \dots, x_{|\mathcal{N}|}\} \subset B_1^d(0)$ and for any $x \in B_1^d(0)$, $\min_{y \in \mathcal{N}} |x - y| \leq \epsilon$. In particular $B_1 \subset \cup_{z \in \mathcal{N}} B_{1/2}(z)$.

Then we can write

$$\begin{aligned}\sup_{\theta \in B_1(0)} \langle \theta, \boldsymbol{X} \rangle &\leq \sup\{\langle z + y, \boldsymbol{X} \rangle : y \in \mathcal{N}, z \in B_{1/2}(0)\} \\ &\leq \sup_{z \in B_{1/2}(0)} \langle z, \boldsymbol{X} \rangle + \sup_{y \in \mathcal{N}} \langle y, \boldsymbol{X} \rangle \\ &= \frac{1}{2} \sup_{z \in B_1(0)} \langle z, \boldsymbol{X} \rangle + \sup_{y \in \mathcal{N}} \langle y, \boldsymbol{X} \rangle,\end{aligned}$$

and rearranging we obtain

$$\sup_{\theta \in B_1(0)} \langle \theta, \boldsymbol{X} \rangle \leq 2 \sup_{y \in \mathcal{N}} \langle y, \boldsymbol{X} \rangle.$$

Thus we have reduced the supremum over an uncountable set to a maximum over a finite set of size $\mathcal{N}$. We know that for each $y \in B_1(0)$, $\langle y, \boldsymbol{X} \rangle$ is σ^2-sub-Gaussian, and we can control the maximum of, possibly dependent, sub-Gaussian variables, very well using a union bound, as in Lemma 1.3.2. In particular we know that

$$\mathbb{P}[2 \sup_{y \in \mathcal{N}} \langle y, \boldsymbol{X} \rangle \geq t] \leq |\mathcal{N}| \mathrm{e}^{-t^2/8\sigma^2}.$$

We now have to control $\mathcal{N}$. Again the argument is classical. We construct an efficient ϵ-net with the following algorithm: initialise $\mathcal{N} := \{0\}$, $\mathcal{X} := B_1(0) \setminus \cup_{z \in \mathcal{N}} B_\epsilon(z)$. While $\mathcal{X} \neq \emptyset$ choose a point in $\mathcal{X}$ and add it to $\mathcal{N}$. By compactness, the algorithm will eventually terminate at which point for any $x, y \in \mathcal{N}$ we have $|x - y| > \epsilon$. Therefore if we replace the ϵ-balls with $\epsilon/2$-balls they will be disjoint and will satisfy

$$\cup_{z \in \mathcal{N}} B_{\epsilon/2}(z) \subset (1 + \tfrac{\epsilon}{2}) B_1(0).$$

Computing the volumes of the above sets we obtain

$$\begin{aligned}(1 + \tfrac{\epsilon}{2})^d \mathsf{Vol}(B_1(0)) = \mathsf{Vol}\left((1 + \tfrac{\epsilon}{2}) B_1(0)\right) &\geq \mathsf{Vol}\left(\cup_{y \in \mathcal{N}} B_{\epsilon/2}(y)\right) \\ &= \sum_{y \in \mathcal{N}} \mathsf{Vol}(B_{\epsilon/2}(y) = |\mathcal{N}| \left(\frac{\epsilon}{2}\right)^d \mathsf{Vol}(B_1(0)),\end{aligned}$$

using the fact that the $\epsilon/2$-balls are disjoint. Rearranging we obtain

$$|\mathcal{N}| \leq \left(\frac{1 + \epsilon/2}{\epsilon/2}\right) \leq (3/\epsilon)^d,$$

and choosing $\epsilon = 1/2$ we have $|\mathcal{N}| \leq 6^d$.

Going back to our union bound, we want to ensure that

$$\begin{aligned}\mathbb{P}[2 \sup_{y \in \mathcal{N}} \langle y, \boldsymbol{X} \rangle \geq t] &\leq |\mathcal{N}| \mathrm{e}^{-t^2/8\sigma^2} \\ &\leq 6^d \mathrm{e}^{-t^2/8\sigma^2} \leq \delta \\ \frac{t^2}{8\sigma^2} &\geq \log(1/\delta) + d \log 6,\end{aligned}$$

which is guaranteed if we choose $t = \sqrt{8\sigma^2 \log(6) d} + 2\sigma\sqrt{2 \log(1/\delta)}$. □

Going back to our estimation of the mean squared error, since

$$\mathsf{MSE}(\boldsymbol{X}\hat{\theta}^{\mathsf{LS}}) = \frac{1}{n}\|\boldsymbol{X}\theta^* - \boldsymbol{X}\hat{\theta}^{\mathsf{LS}}\|^2 \leq \frac{4}{n}\frac{\epsilon^{\mathsf{T}}\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)}{\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|}$$

we conclude that with probability at least $1-\delta$ we have

$$\mathsf{MSD}(\boldsymbol{X}\hat{\theta}^{\mathsf{LS}}) \lesssim \frac{\sigma^2}{n}\left(r + \log(1/\delta)\right).$$

Recalling that $r = \mathsf{rank}(\boldsymbol{X})$, we summarise what we have done so far we have the following result.

Theorem 8 (OLS for fixed design)**.** *Assume the model* (2.1) *holds, where* $\epsilon = (\epsilon_i)_{i=1}^n$ *is* σ^2*-sub-Gaussian random vector. Then, there exists a universal constant* $c > 0$*, such that the least squares estimator satisfies*

$$\mathbb{E}\left[\mathsf{MSE}(\boldsymbol{X}\hat{\theta}^{\mathsf{LS}})\right] \leq c\frac{\mathsf{rank}(\boldsymbol{X})}{n}\sigma^2,$$

and with probability at least $1-\delta$ *we have*

$$\mathsf{MSE}(\boldsymbol{X}\hat{\theta}^{\mathsf{LS}}) \leq c\sigma^2\frac{\mathsf{rank}(\boldsymbol{X}) + \log(1/\delta)}{n}.$$

Notice that in the case where $n \geq d$ and $\mathsf{rank}(\boldsymbol{X}) = d$ we can also control the recovery error in l_2, using

$$\lambda_{\mathsf{min}}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})\|\hat{\theta}^{\mathsf{LS}} - \theta^*\|_2^2 \leq \mathsf{MSE}(\boldsymbol{X}\hat{\theta}^{\mathsf{LS}}).$$

Although controlling the prediction error is easier than controlling the recovery error, we can already feel the effect of the dimension; in this case it is the dimension of the column space of the design matrix $\boldsymbol{X}$. In a sense this dimension controls the expressiveness of model (2.1) as it controls the dimension of the right hand side.

{ex:gaussian_se

Example 6 (Gaussian sequence)**.** *One example where this bound is actually tight is a finite dimensional version of the Gaussian sequence model, that is observations of the form*

$$y_i = \sqrt{n}\theta_i^* + \epsilon_i, \quad i = 1, \ldots, n.$$

Here clearly $n = d$ *and the design matrix takes the especially simple form* $\boldsymbol{X} = \sqrt{n}\mathbb{1}_n$*. Here clearly* $\hat{\theta}^{\mathsf{LS}} = n^{-1/2}y$*. One can then easily see that*

$$\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|_2^2 = \sum_{i=1}^n \frac{\epsilon_i^2}{n},$$

which is $\Omega(1)$ *for all* n*.*

So we can see that when $d \geq n$, even the prediction error can be quite large. We will see in the next section that the only way to make any progress in this case is if we impose some additional structure. Sparsity is the most common choice.

2.1 High-dimensional models and Sparsity

Despite its simplicity it is worth considering the deterministic linear model $y = \boldsymbol{X}\theta^*$. When $d > n$ then it defines an underdetermined linear system which then has a linear space of solutions. In particular there is no way of obtaining any meaningful information about θ^*, unless we impose some additional structure. This structure can often be introduced in the model in the form of a constraint. Before proceeding we introduce some standard notation.

Notation. *For a vector $v \in \mathbb{R}^d$ and $q > 0$ let $\|v\|_q$ denote the norm $\|v\|_q = \sum_{i=1}^n \|v_i\|^q$, l_q the normed space $(\mathbb{R}^d, \|\cdot\|_q)$ and $B_q(t)$ the l_q-ball of radius $t > 0$, that is $B_q(t) := \{v : \|v\|_q \leq t\}$. For $q = 0$, we define $\|v\|_0 := \sum_j \mathbb{1}\{v_j \neq 0\}$, which is not a norm but counts the number of non-zero entries of v.*

Suppose for example that we knew a priori that the solution θ^* belongs to some set K; K could be the set of k-sparse vectors, i.e. $B_0(k)$, or $B_q(1)$ for some $q > 0$. Then we could rephrase the problem as follows for example:

$$\hat{\theta} = \arg\min_{\theta \in K} \|y - \boldsymbol{X}\theta\|_2^2. \tag{2.5}$$ {eq:constrained

Let us first consider the case where $K = B_1(1)$ and $\theta^* \in K$.

Notice that since we assume $\theta^* \in B_1(1)$, the inequality (2.3) holds and thus we again have

$$\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|_2^2 \leq 2\frac{\epsilon^{\mathsf{T}}\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)}{\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|} \leq 2 \sup_{v,w \in B_1(1)} \epsilon^{\mathsf{T}}\boldsymbol{X}(v - w).$$

First notice that if $v, w \in B_1(1)$ then by the triangle inequality $v - w \in B_1(2)$ and thus by linearity we have

$$\|\boldsymbol{X}(\hat{\theta}^{\mathsf{LS}} - \theta^*)\|_2^2 \leq 2 \sup_{v' \in B_1(2)} \epsilon^{\mathsf{T}}\boldsymbol{X}v' \leq 4 \sup_{v' \in B_1(1)} \epsilon^{\mathsf{T}}\boldsymbol{X}v' = 4\sup\{\epsilon^{\mathsf{T}}\boldsymbol{X}v : v = \pm \mathtt{e}_j, j = 1, \ldots, d\},$$

where we used the standard fact that a linear form over a convex set is maximized at an extreme point, and that the extreme points of $B_1(1)$ are given by the standard basis vectors and their opposites (compare with $B_2(1)$. This has reduced the supremum from an uncountable to a finite set of cardinality d. Also notice that if the column vectors of $\boldsymbol{X}$ are given by $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n]$, then by Lemma 1.3.1, the variables $\epsilon^T \boldsymbol{X}_j$ are $\|\boldsymbol{X}_j\|_2^2\sigma^2$-sub-Gaussian. Thus using Lemma 1.3.2 we have that

$$\mathbb{E}\left[\mathsf{MSE}(\boldsymbol{X}\hat{\theta}^{\mathsf{LS}})\right] \leq \frac{4\sigma \max_j \|\boldsymbol{X}_j\|_2}{n}\sqrt{2\log(2d)},$$

and for any $t > 0$ we have

$$\mathbb{P}\left[\mathsf{MSE}(\boldsymbol{X}\hat{\theta}^{\mathsf{LS}}) > t\right] \leq \mathbb{P}\left[\max_{v = \pm \mathtt{e}_j} \epsilon^{\mathsf{T}}\boldsymbol{X}v > nt/4\right] \leq 2d\exp\left(\frac{-n^2t^2}{16\sigma^2 \max_j \|\boldsymbol{X}_j\|_2^2}\right).$$

In particular if the design matrix is normalized such that $\max_j \|\boldsymbol{X}_j\|_2 \leq \sqrt{n}$, then we have concentration at rate n.

Similar results hold if we let $K = B_0(k)$, for some integer k. However, there are computational issues with that choice, since to solve the problem one needs to compute a very large number of LS estimators, in particular ${}_dC_k$, one for every possible choice of *support*.

So it seems that indeed, additional structure can get us improved results. But we have cheated. Typically, one will not know a priori that θ^* belongs to $B_1(1)$, or that it is k-sparse, so we need methods that will automatically adapt to the structure of the problem.

Regularization

As we mentioned in the last section if we know that θ^* belongs to some set K then we can use this information to obtain sharper rates. Two particular cases of interest were $K = B_1(1)$ and $K = B_0(k)$. We saw that in the former case, under some additional assumptions on the normalization of $\boldsymbol{X}$ we can get much better rates by making use of the additional structure. The issue however is that typically we do not know that, for example, θ^* only has k non-zero entries. We may have reason to believe that it is sparse, but we need a methodology which does not need a priori knowledge of k, but rather adapts to it.

In this case instead of (2.5) we may modify (2.2) by adding a regularisation term that essentially penalises parameters with high l_0 norm, that is one could consider

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - \boldsymbol{X}\theta\|_2^2 + \lambda \|\theta\|_0, \right\} \tag{2.6}$$ {eq:hardsparse}

where λ is a user-set *regularisation* parameter.

Before we embark on our task however, it is quite interesting to first consider the noiseless setting, that is the model

$$y = \boldsymbol{X}\theta^*, \tag{2.7}$$ {eq:noiseless}

and try to understand when it is possible to *recover* θ^* exactly. To be more precise suppose that we know that θ^* is sparse, say k-sparse, where $k \ll d$ is unknown. Then we could try and solve the underdetermined problem (2.7) by adding a regularisation term. That is let us try to solve the minimisation problem

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_0, \qquad \boldsymbol{X}\theta = y. \tag{2.8}$$ {eq:initiall0pr

A solution to this would automatically give us the sparsest parameter θ solving (2.7) without assuming a priori anything about the sparsity of θ^*.

2.2 Recovery in the noiseless model

Although (2.8) directly controls the sparsity of the parameter, the loss function in the above problem is non-convex which makes our task quite hard. One would have to search all subsets of $\{1, \ldots, d\}$ and attempt to solve (2.7) restricted to that subspace. The computational cost grows exponentially in the sparsity parameter k.

So we may instead consider a convex relaxation of (2.8) by replacing the $\|\cdot\|_0$ regularisation term with a convex one

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1, \qquad \boldsymbol{X}\theta = y, \tag{2.9}$$ {eq:basispursui

Chen, Donoho, and Saunders (1998).

Notation. *For any integers $k < n$, let $[k : n] = \{k, k+1, \ldots, n\}$.*

The first question we will attempt to answer is whether (2.9) recovers the solution of (2.8). That is suppose that $y = \boldsymbol{X}\theta^*$, where

$$\theta_j \neq 0, j \in S \subset [1 : d], \qquad \theta_j = 0, j \in S^{\mathsf{c}} = [1 : d] \setminus S.$$

To understand when solving (2.9) will give us the solution of (2.8) we need to first think a little bit about the space of solutions of $\boldsymbol{X}\theta = y$. We know that θ^* is a solution, so the space of solutions will be given by

$$\mathcal{S}(\boldsymbol{X}, y) := \{\theta \in \mathbb{R}^d : \boldsymbol{X}\theta = y\} = \theta^* + \mathsf{ker}(\boldsymbol{X}),$$

where $\mathsf{ker}(\boldsymbol{X})$ is the kernel of $\boldsymbol{X}$.

Since can restate (2.9) as $\min_{\theta \in \mathcal{S}(\boldsymbol{X}, y)} \|\theta\|_1$ it is clear that (2.9) will recover the solution of (2.8) only when θ^* is the minimal element of $\mathcal{S}(\boldsymbol{X}, y)$ with respect to the $\|\cdot\|_1$ norm, or in other words, for any $v \in \mathsf{ker}(\boldsymbol{X})$ we have

$$\|\theta^* + v\|_1 \geq \|\theta^*\|.$$

In order to visualise the situation (see also Figure 2.2) consider $B(\|\theta^*\|_1)$, that is the l_1-ball with θ^* on its boundary. Notice that since θ^* is sparse, it will necessarily be an extreme point of $B(\|\theta^*\|_1)$, that is one of the "corners". Then (2.9) will correctly recover θ^* if and only if $\theta^* + \mathsf{ker}(\boldsymbol{X})$ only intersects $B(\|\theta^*\|_1)$ at θ^*.

Notation. *For $v \in \mathbb{R}^d$ and $S \subset [1 : d]$ we write v_S for the vector with entries $(v_S)_i = v_i$ if $i \in S$ and $(v_S)_i = 0$ otherwise.*

We define the following subset

$$\mathbb{C}(S) := \{v \in \mathbb{R}^d : \|v_{S^c}\|_1 \leq \|v_S\|_1\}.$$

We are now ready to have a closer look at Figure 2.2, where $d = 2$, $\theta^* = (0, 1)$, and thus its support is $S = \{2\}$. The shaded region in the figure represents $\mathbb{C}(S)$. We can see that there will be a unique solution if and only if $\theta^* + \mathsf{ker}(\boldsymbol{X})$ does not intersect the l_1 ball with θ^* on its boundary. Equivalently, by shifting everything by $-\theta^*$, we can see that we {fig:restricted

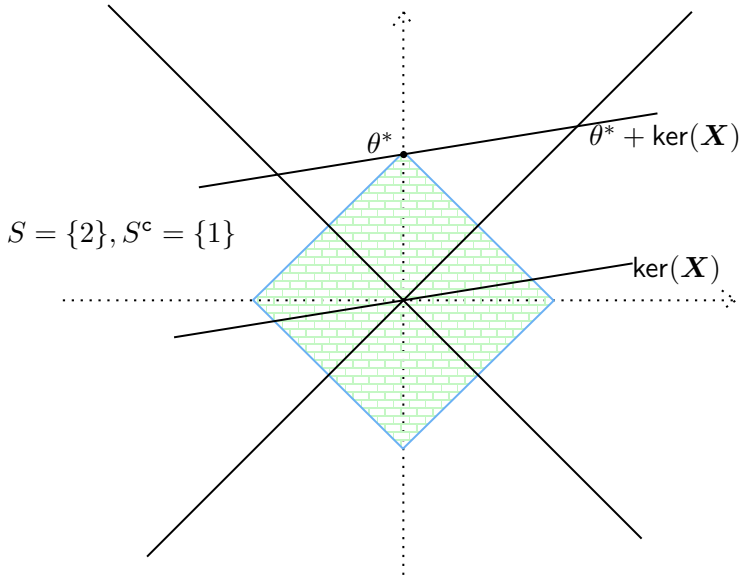

Figure 2.1: Here $S = \{2\}$. The gray shaded region represents $\mathbb{C}(S)$. The green shaded region is $B(\|\theta^*\|_1)$.

can recover θ^* if and only if $\mathsf{ker}(\boldsymbol{X})$ does not intersect the gray shaded region which is precisely $\mathbb{C}(S)$.

As we can see this property refers only to the kernel of $\boldsymbol{X}$ and the support of the vector.

Definition 7. *We say that the matrix $\boldsymbol{X}$ satisfies the* ***restricted nullspace property*** *with respect to $S \subset [1:d]$, if $\mathbb{C}(S) \cap \ker(\boldsymbol{X}) = \{0\}$.*

Summarising the previous discussion we have the following result.

Theorem 9. *The following are equivalent:*

(a) For any vector θ^* supported on $S \subset [1:d]$, (2.9) has a unique solution θ^*.

(b) The matrix $\boldsymbol{X}$ satisfies the restricted nullspace property with respect to S.

Proof. $(b) \Rightarrow (a)$: By assumption $\boldsymbol{X}\theta^*$ so we need to prove that any other solution θ', has norm $\|\theta'\|_1 \geq \|\theta^*\|_1$. Let us write $\theta' = \theta^* + v$, where $v \in \ker(\boldsymbol{X})$; by (b) it must be the case that $\|v_{S^c}\|_1 \geq v_S\|_1$. Finally, recall that θ^* is supported on S, and thus by the triangle inequality

$$\begin{aligned}\|\theta'\|_1 = \|\theta^* + v\|_1 = \|\theta^*_{S^c} + v_{S^c}\|_1 + \|\theta^*_S + v_S\|_1 \\ = \|v_{S^c}\|_1 + \|\theta^*_S + v_S\|_1 \geq \|\theta^*_S\|_1 - \|v_S\|_1 + \|v_{S^c}\|_1 \geq \|\theta^*\|_1.\end{aligned}$$

$(a) \Rightarrow (b)$: Suppose that $\theta^* \in \ker(\boldsymbol{X})$, $\theta^* \neq 0$. Then $\boldsymbol{X}\theta^* = 0$, and thus

$$\boldsymbol{X}[\theta^*_S, 0]^\mathsf{T} = \boldsymbol{X}[0, -\theta^*_{S^c}]^\mathsf{T}.$$

But this means that the vector $[0, -\theta^*_{S^c}]$ solve the problem $\boldsymbol{X}\theta = \boldsymbol{X}[\theta^*_S, 0]$. Since by (a), (2.9) must recover θ^*_S uniquely, we must have that $\|\theta_S\|_1 < \|\theta_{S^c}\|_1$. □

Sufficient conditions for restricted nullspace property

For the restricted nullspace property to be useful we need checkable sufficient conditions. First we should develop a little intuition.

Let us first think about what it means for a vector $v = (v_1, \ldots, v_d)^\mathsf{T}$ to be in $\ker(X)$; by definition $\boldsymbol{X}v = 0$, and since $\boldsymbol{X}v = \sum_{j=1}^d v_j \boldsymbol{X}_j$, where $\boldsymbol{X}_j, j = 1, \ldots, d$ are the columns of $\boldsymbol{X}$, the condition $\boldsymbol{X}v = 0$ for $v \neq 0$, implies that the columns of $\boldsymbol{X}$ are not linearly independent. Therefore one obvious condition we could impose to ensure that $\ker(\boldsymbol{X}) = 0$, and therefore that the restricted nullspace property holds for all $S \subset [1:d]$ would be to require that the columns of $\boldsymbol{X}$ are linearly independent; one can see that this requires that $d \leq n$ so it restricts the dimension of the model. If the columns of $\boldsymbol{X}$ are linearly independent then by transforming the model (2.7), we can actually assume that the columns are orthonormal, that is $\langle \boldsymbol{X}_j, \boldsymbol{X}_k \rangle = \delta_{j,k}$, where for two vectors $v, w \in \mathbb{R}^d$ we write $\langle v, w \rangle$ for their inner product.

In fact when we consider the noisy model (2.1) it will be convenient to change the normalisation of the problem so that the corresponding assumption would be the following condition known

$$\frac{1}{n}\langle \boldsymbol{X}_j, \boldsymbol{X}_k \rangle = n\delta_{j,k}, \quad j, k, = 1, \ldots, d. \tag{2.10}$$ {eq:ORT}

However, as we mentioned before, this is quite restrictive as it can only hold for $d \leq n$.

One way around this problem would be to allow (2.10) to fail in a controlled way. That is we could require for the colums of $\boldsymbol{X}$ to be *almost* orthonormal. This leads us to the following definition.

Definition 8 (Pairwise incoherence parameter)**.** *For a $n \times d$-matrix $\boldsymbol{X}$, we define the pairwise incoherence parameter as*

$$\delta_{\mathsf{PW}}(\boldsymbol{X}) := \max_{j,k=1,\ldots,d} \left| \frac{1}{n}\langle \boldsymbol{X}_j, \boldsymbol{X}_k \rangle - \delta_{j,k} \right|.$$

The pairwise incoherence parameter allows us to quantify the degree to which (2.10) fails to hold.

Let us now see how we can use the pairwise incoherence parameter to check the restricted nullspace property. Let $S \subset [1:d]$, with $|S| = s$ and suppose that $\delta_{\mathsf{PW}}(\boldsymbol{X}) \leq \gamma/s$. Also let $\boldsymbol{X}_S = (\boldsymbol{X}_{i,j})_{i,j\in S}$. Suppose that λ is an eigenvalue of $\boldsymbol{X}_S^{\mathsf{T}}\boldsymbol{X}_S/n$. Then for some $w \in \mathbb{R}^s$

$$\begin{aligned}\left(\frac{1}{n}\boldsymbol{X}_S^{\mathsf{T}}\boldsymbol{X}_S - \mathbb{1}\right) w &= (\lambda - 1)w \\ |\lambda - 1|\|w\|_2 &\leq \left\|\frac{1}{n}\boldsymbol{X}_S^{\mathsf{T}}\boldsymbol{X}_S - \mathbb{1}\right\|_2 \|w\|_2 \\ &\leq \left\|\frac{1}{n}\boldsymbol{X}_S^{\mathsf{T}}\boldsymbol{X}_S - \mathbb{1}\right\|_{\mathsf{F}} \|w\|_2 \qquad (2.11)\end{aligned}$$ {eq:pwineq}

where for a matrix $A \in \mathbb{R}^{k\times k}$, the l_2-norm is defined as

$$\|A\|_2 := \sup w \in \mathbb{R}^k : \|w\|_2 = 1\|Aw\|_2,$$

and the Frobenius norm $\|A\|_{\mathsf{F}}$ as

$$\|A\|_{\mathsf{F}} := \sum_{i,j=1}^{k} A_{i,j}^2.$$

We also used the standard inequality $\|A\|_2 \leq \|A\|_{\mathsf{F}}$ which follows easily by checking the inequality for any $v = \sum_j v_j \mathbb{e}_j$, with $\sum c_j^2 = 1$. Thus continuing from (2.11) we have

$$|\lambda - 1|^2 \leq \sum_{j,k\in S}\left(\frac{1}{n}\langle \boldsymbol{X}_j, \boldsymbol{X}_k\rangle - \delta_{j,k}\right)^2 \leq |S|^2\delta_{\mathsf{PW}}(\boldsymbol{X})^2 \leq \gamma^2.$$

From this we easily deduce that if $\gamma \in (0,1)$ we must have $1 - \lambda \leq \gamma$ and therefore that $\lambda \geq 1 - \gamma > 0$.

This implies that for any $\theta \in \mathbb{R}^d$ we have

$$\theta_S^{\mathsf{T}}\frac{\boldsymbol{X}_S^{\mathsf{T}}\boldsymbol{X}_S}{n}\theta_S \geq (1-\gamma)\|\theta_S\|_2^2,$$

which after rearranging and noticing that $\|\boldsymbol{X}_S\theta_S\|_2 = \|\boldsymbol{X}\theta_S\|_2$, implies that

$$\|\theta_S\|_2^2 \leq \frac{1}{1-\gamma}\theta_S^{\mathsf{T}}\frac{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}}{n}\theta_S.$$

Now, suppose that $\theta \in \mathsf{ker}(\boldsymbol{X})$. Thus $\boldsymbol{X}\theta_S = -\boldsymbol{X}\theta_{S^{\mathsf{c}}}$. Therefore

$$\begin{aligned}\|\theta_S\|_2^2 &\leq \frac{1}{1-\gamma}\theta_S^{\mathsf{T}}\frac{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}}{n}\theta_S \\ &= \frac{1}{1-\gamma}\theta_S^{\mathsf{T}}\frac{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}}{n}\theta_{S^{\mathsf{c}}} \\ &= \frac{1}{1-\gamma}\theta_S^{\mathsf{T}}\left(\frac{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}}{n} - \mathbb{1}\right)\theta_{S^{\mathsf{c}}}\end{aligned}$$

since $\theta_S^{\mathsf{T}}\theta_{S^{\mathsf{c}}} = 0$

$$\leq \frac{1}{1-\gamma}\|\frac{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}}{n} - \mathbb{1}\|_\infty\|\theta_S\|_1\|\theta_{S^{\mathsf{c}}}\|_1$$

$$\leq \frac{1}{1-\gamma}\delta_{\mathsf{PW}}(\boldsymbol{X})\|\theta_S\|_1\|\theta_{S^c}\|_1,$$

where $\|A\|_\infty := \max_{i,k}|A_{i,j}|$. Finally, using the $l_1 - l_2$ inequality we have that $\|\theta_S\|_1^2 \leq |S|\|\theta_S\|_2^2$ and thus continuing from above

$$\|\theta\|_1^2 \leq s\|\theta_S\|_2^2 \leq s\frac{1}{1-\gamma}\frac{\gamma}{s}\|\theta_{S^c}\|_1$$
$$\|\theta\|_1 \leq \frac{\gamma}{1-\gamma}\|\theta_{S^c}\|_1.$$

We have essentially proven the following result.

Proposition 2.2.1. *If for some integer s we have*

$$\delta_{\mathsf{PW}}(\boldsymbol{X}) \leq \frac{1}{3s},$$

then the restricted nullspace property holds for all subsets S of cardinality at most s.

2.3 The Lasso estimator

We have developed some intuition about the way the design matrix $\boldsymbol{X}$ and the parameter θ^* must interact for (2.9) to recover the solution of (2.8) in the noiseless setting.

We now return to the noisy setting of Model (2.1). We aim to look for sparse solutions, but instead of considering (2.6) we aim for its convex relaxation

$$\hat{\theta}^{\mathsf{L}} \in \arg\min_{\theta\in\mathbb{R}^d}\left\{\frac{1}{2n}\|y-\boldsymbol{X}\theta\|_2^2 + \lambda\|\theta\|_1,\right\}, \tag{2.12}$$ {eq:lasso}

known as the Lasso estimator, Tibshirani (1996).

We will first analyse the Lasso estimator without imposing any structural assumptions on the design matrix. We will then see how controlling the pairwise incoherence parameter of $\boldsymbol{X}$ can allow us to get better rates.

2.3.1 Lasso-minimal assumptions

First we attempt to bound the means squared error of the Lasso estimator assuming only that the columns of $\boldsymbol{X}$ are normalized so that $\max_j \|\boldsymbol{X}_j\|_2^2 \leq n$. Recall that

$$\begin{aligned}\|Y - \boldsymbol{X}\hat{\theta}^{\mathsf{L}}\|_2^2 &= \|\boldsymbol{X}\theta^* + \epsilon - \boldsymbol{X}\hat{\theta}^{\mathsf{L}}\|_2^2 \\ &= \|\boldsymbol{X}(\theta^* - \hat{\theta}^{\mathsf{L}})\|_2^2 + 2\langle\epsilon, \boldsymbol{X}(\theta^* - \hat{\theta}^{\mathsf{L}})\rangle + \|\epsilon\|_2^2\end{aligned}$$

and thus

$$\|\boldsymbol{X}(\theta^* - \hat{\theta}^{\mathsf{L}})\|_2^2 = \|Y - \boldsymbol{X}\hat{\theta}^{\mathsf{L}}\|_2^2 - 2\langle\epsilon, \boldsymbol{X}(\theta^* - \hat{\theta}^{\mathsf{L}})\rangle - \|\epsilon\|_2^2. \tag{2.13}$$ {eq:lasso_slow1}

Directly from (2.12) we obtain

$$\frac{1}{n}\|Y - \boldsymbol{X}\hat{\theta}^{\mathsf{L}}\|_2^2 + 2\tau\|\hat{\theta}^{\mathsf{L}}\|_1^2 \leq \frac{1}{n}\|Y - \boldsymbol{X}\theta^*\|_2^2 + 2\tau\|\theta^*\|_1$$

which combined with (2.13) gives us

$$\|\boldsymbol{X}(\theta^* - \hat{\theta}^{\mathsf{L}})\|_2^2 \leq 2\langle\epsilon, \boldsymbol{X}(\hat{\theta}^{\mathsf{L}} - \theta^*)\rangle + 2n\tau\left(\|\theta^*\|_1 - \|\hat{\theta}^{\mathsf{L}}\|_1\right)$$

$$\leq 2\langle \boldsymbol{X}^\mathsf{T}\epsilon, \hat{\theta}^\mathsf{L}\rangle - 2n\tau\|\hat{\theta}^\mathsf{L}\|_1 + 2\left(\langle \boldsymbol{X}^\mathsf{T}\epsilon, \theta^*\rangle + n\tau\|\theta^*\|_1\right)$$
$$\leq 2\left(\|\boldsymbol{X}^\mathsf{T}\epsilon\|_\infty - n\tau\right)\|\hat{\theta}^\mathsf{L}\|_1 + \left(\|\boldsymbol{X}^\mathsf{T}\epsilon\|_\infty + n\tau\right)\|\theta^*\|_1.$$

A combination of a union bound with Lemma 1.3.1 gives

$$\mathbb{P}\left[\|\boldsymbol{X}^\mathsf{T}\epsilon\|_\infty \geq t\right] = \mathbb{P}\left[\max_{j=1,\ldots,d}\langle \epsilon, \boldsymbol{X}_j\rangle \geq t\right] \leq d\max_{j=1,\ldots,d}\mathbb{P}\left[\langle \epsilon, \boldsymbol{X}_j\rangle \geq t\right] \leq 2d\exp\left(-\frac{t}{2n\sigma^2}\right),$$

where we used the assumption that $\max_j \|\boldsymbol{X}_j\|_2^2 \leq n$.

We want the probability of this event to be smaller than $\delta > 0$ say, which can be guaranteed by letting

$$t = \sigma\sqrt{2n\log(2d)} + \sigma\sqrt{2n\log(1/\delta)}.$$

If we then choose the regularisation parameter τ to be such that $\tau = t/n$ then we have that with probability at least $1-\delta$ we have

$$\mathsf{MSE}(\boldsymbol{X}\hat{\theta}^\mathsf{L}) := \frac{\|\boldsymbol{X}(\theta^* - \hat{\theta}^\mathsf{L})\|_2^2}{n} \leq 4\tau\|\theta^*\|_1.$$

We have just proven the following result.

{thm:lasso_slow

Theorem 10. *Suppose that* (2.1) *holds and let* $\hat{\theta}^\mathsf{L}$ *the solution of* (2.12) *with*

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

Then with probability at least $1-\delta$ *we have*

$$\mathsf{MSE}(\boldsymbol{X}\hat{\theta}^\mathsf{L}) \leq \frac{4\|\theta^*\|_1}{\sqrt{n}}\left(\sqrt{2\log(2d)} + \sqrt{2\log(1/\delta)}\right).$$

We can see that if we only assume that $\max_j \|\boldsymbol{X}_j\|_2 \leq \sqrt{n}$, we get a $n^{-1/2}$ rate.

2.3.2 ORT, thresholding and faster rates

Although we seem to have attained the $n^{-1/2}$ rate almost for free, it is quite interesting to revisit the Gaussian sequence Example 6. Recall that there $n = d$ and $\boldsymbol{X} = \sqrt{n}\mathbb{1}$. In this case (2.12) becomes

$$\arg\min_{\theta\in\mathbb{R}^n}\left[\frac{1}{2n}\|y - \boldsymbol{X}\theta\|_2^2 + \tau\|\theta\|_1.\right] \tag{2.14}$$

{eq:sft_gaussia

By multiplying (2.1) by $\boldsymbol{X}^\mathsf{T}/n$ the model now becomes

$$y' := \frac{1}{n}\boldsymbol{X}^\mathsf{T}y = \theta^* + \xi,$$

where $\xi = (\xi_1, \ldots, \xi_n)$, with ξ_i independent and (σ^2/n)-sub-Gaussian. Then notice that

$$\frac{1}{n}\|y - \boldsymbol{X}\theta\|_2^2 = \left\|\frac{1}{n}\boldsymbol{X}^\mathsf{T}(y - \boldsymbol{X}\theta)\right\|_2^2 = \|y' - \theta\|_2^2,$$

whence (2.14) becomes

$$\arg\min_{\theta\in\mathbb{R}^d}\left[\|y' - \theta\|_2^2 + 2\tau\|\theta\|_1\right] = \arg\min_{\theta_1,\ldots,\theta_d}\sum_{i=1}^d\left[(y_i' - \theta_i)^2 + 2\tau|\theta_i|\right]$$

$$= \sum_{i=1}^{d} \underset{\theta_1,\dots,\theta_d}{\arg\min} \left[(y_i' - \theta_i)^2 + 2\tau|\theta_i| \right].$$

For positive θ_i this becomes

$$\theta_i^2 - 2y_i'\theta_i + y_i'^2 + 2\tau\theta_i = \theta_i^2 + 2(\tau - y_i')\theta_i + y_i'^2,$$

which is optimized at $y_i' - \tau$, whereas for negative θ_i this is

$$\theta_i^2 - 2y_i'\theta_i + y_i'^2 - 2\tau\theta_i = \theta_i^2 - 2(\tau + y_i')\theta_i + y_i'^2,$$

which is minimised at $\tau + y_i'$. Therefore if $y_i' > \tau$, $\hat{\theta}_i = y_i - \tau$, and if $\tau + y_i < 0$, that is $y_i < -\tau$, $\hat{\theta}_i = y_i + \tau$. In the case where $|y_i'| \le \tau$, we have in either case no solution so the function is minimised at the boundary, that is $\theta_i = 0$.

We can summarise this as

$$\theta_i^{\mathsf{sft}} = T_\tau^{\mathsf{sft}}(y_i'), \quad i = 1, \dots, d,$$

where we used the *soft thresholding function*

$$T_\tau^{\mathsf{sft}}(x) := \begin{cases} \mathsf{sign}(x')(|x| - \tau) & \text{if } |x| \ge \tau, \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$\tau := \tau_n := \sigma\sqrt{\frac{8\log(2d/\delta)}{n}},$$

and define the event $\mathcal{A} := \{\max_i |\xi_i| \le \tau/2\}$. Since each ξ_i is (σ^2/n)-sub-Gaussian we get that

$$\mathbb{P}[\mathcal{A}^{\mathsf{c}}] \le 2d \exp\left(-n\frac{\tau^2}{8\sigma^2}\right) \le \delta.$$

On the event $\mathcal{A}$, we can thus estimate

$$\begin{aligned}
\|\theta^{\mathsf{sft}} - \theta^*\|_2^2 &= \sum_{j=1}^{d} \left| T_\tau^{\mathsf{sft}}(y_i') - \theta_i^* \right|^2 \\
&= \sum_{j=1}^{d} \Big[\mathbb{1}\{y_i' \ge \tau\}(\theta_i^* + \xi_i - \tau - \theta_i^*) \\
&\qquad + \mathbb{1}\{y_i' \le -\tau\}(\theta_i^* + \xi_i + \tau - \theta_i^*) - \mathbb{1}\{|y_i'| \le \tau\}|\theta_i^*| \Big]^2 \\
&= \sum_{j=1}^{d} \Big[\mathbb{1}\{y_i' \ge \tau\}(\xi_i - \tau) \\
&\qquad + \mathbb{1}\{y_i' \le -\tau\}(\xi_i + \tau) - \mathbb{1}\{|y_i'| \le \tau\}|\theta_i^*| \Big]^2 \\
&\le \sum_{j=1}^{d} \Big[2\tau\mathbb{1}\{|y_i'| \ge \tau\} + \mathbb{1}\{|y_i'| \le \tau\}|\theta_i^*| \Big]^2.
\end{aligned}$$

Also notice that on the event $\mathcal{A}$ it easily follows that $|y_j'| \ge \tau$ implies that

$$\frac{\tau}{2} \ge |\xi_j| \ge |\xi_j + \theta_j^*| - |\theta_j^*|,$$

whence we obtain $|\theta_j^*| \geq \tau/2$, and $|y'_j| \leq \tau$ implies that $|\theta_j^*| \leq 3\tau/2$. Overall we thus obtain

$$\begin{aligned}
\|\theta^{\mathsf{sft}} - \theta^*\|_2^2 &\leq \sum_{j=1}^d \left[2\tau \mathbb{1}\{|\theta_j^*| \geq \tau/2\} + \mathbb{1}\{|\theta_j^*| \leq 3\tau/2\}|\theta_j^*|\right]^2 \\
&\leq \sum_{j=1}^d \left[4\min\{|\theta_j^*|, \tau/2\}\right]^2 \\
&\leq \sum_{j=1}^d 16 \min\left\{|\theta_j^*|^2, \frac{\tau^2}{4}\right\} \leq 4\|\theta^*\|_0^2 \tau^2 \\
&\leq 4\|\theta^*\|_0^2 \sigma^2 \frac{8\log(2d/\delta)}{n} = \frac{32\|\theta^*\|_0^2 \sigma^2 \log(2d/\delta)}{n}.
\end{aligned}$$

Also notice that by construction the support of θ^{sft} is contained in the support of θ^*, and that with probability at least $1-\delta$, if $\theta_j^* > 3\tau_n/2$ for every j in the support of θ^*, then the support of θ^{sft} matches that of θ^*.

Notice that we can get a similar rate for the mean squared error of the original model (2.1) under the assumption (2.10). Indeed if $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}/n = \mathbb{1}$, then by multiplying the model $y = \boldsymbol{X}\theta^* + \epsilon$, where ϵ_i are independent σ^2-sub-Gaussian, by $\boldsymbol{X}^\mathsf{T}/n$ we obtain

$$y' = \frac{\boldsymbol{X}^\mathsf{T} y}{n} = \theta^* + \xi,$$

where $\xi = \boldsymbol{X}^\mathsf{T}\epsilon/n$. Notice that $\xi = (\xi_1, \ldots, \xi_n)$ where the ξ_i are (σ^2/n)-sub-Gaussian, although no longer independent. However, notice that all preceding calculations did not use independence at all, since we only relied on union bounds and the sub-Gaussian properties of the individual noise components ϵ_i. Since under assumption (2.10) the MSE of θ^{sft} for model (2.1) is given by

$$\mathsf{MSE}(\boldsymbol{X}\theta^{\mathsf{sft}}) = (\theta^* - \theta^{\mathsf{sft}})^\mathsf{T} \frac{\boldsymbol{X}^\mathsf{T}\boldsymbol{X}}{n} (\theta^* - \theta^{\mathsf{sft}}),$$

we thus conclude that similar bounds hold.

We thus have the following result.

{thm:lasso_ort_

Theorem 11. *For the observation model 2.1, under the assumption* (2.10)*, with probability at least* $1-\delta$*, we have*

$$\mathsf{MSE}(\boldsymbol{X}\theta^{\mathsf{sft}}) \leq \frac{32\|\theta^*\|_0^2 \sigma^2 \log(2d/\delta)}{n}.$$

where θ^{sft} *is the soft thresholding estimator with threshold*

$$\tau := \tau_n := \sigma\sqrt{\frac{8\log(2d/\delta)}{n}}.$$

When discussing exact recovery in the noiseless setting we saw that we could actually make progress by using the pairwise incoherence parameter to quantify the extend to which (2.10) failed. We will now see that this is also the case for the Lasso estimator.

2.3.3 The restricted eigenvalue condition

In the last Section we saw that under the assumption (2.10) we can obtain n^{-1} rates for the prediction error of the Lasso estimator, which in that case takes the special form of the soft thresholding estimator θ^{sft}. On the other hand, under only the normalisation assumption that $\max_j \|\boldsymbol{X}_j\|_2 \leq \sqrt{n}$, in Theorem 10 we were able to obtain the rate $n^{-1/2}$.

We will now attempt to obtain the n^{-1} rate without assumption (2.10). In the noiseless setting we saw that the restricted nullspace assumption was crucial in general for recovering the truth θ^*. In this section a similar, albeit stronger condition, will prove extremely useful. Before we introduce this condition, we need some notation. For $\alpha \geq 1$ and $S \subset [1:d]$, we define the set

$$\mathbb{C}_\alpha(S) := \{v \in \mathbb{R}^d : \|v_{S^\mathsf{c}}\|_1 \leq \alpha \|v_S\|_1\}.$$

Notice that for $\mathbb{C}_1(S)$ coincides with the set $\mathbb{C}(S)$ used in the restricted nullspace property.

Definition 9 (Restricted eigenvalue condition)**.** *The matrix $\boldsymbol{X}$ satisfies the (κ, α)-restricted eigenvalue condition (RE) over $S \subset [1:d]$ if for all $v \in \mathbb{C}_\alpha(S)$*

$$\kappa \|v\|_2^2 \leq \frac{1}{n}\|\boldsymbol{X}v\|_2^2.$$

{thm:lasso_fast

Theorem 12. *Suppose that (2.1) holds with θ^* supported on $S \subset [1:d]$, where $\epsilon = (\epsilon_i)_{i=1}^d$, with ϵ_i σ^2-sub-Gaussian. Suppose in addition that $\boldsymbol{X}$ is normalised so that $\max_j \|\boldsymbol{X}_j\|_2 \leq \sqrt{n}$ and that $\boldsymbol{X}$ satisfies the $(\kappa, 3)$-restricted eigenvalue condition with respect to S. Let $\hat{\theta}^{\mathsf{L}}$ solve (2.12) with*

$$\tau_n := \sqrt{\frac{8\sigma^2 \log(2d/\delta)}{n}}.$$

Then with probability at least $1 - \delta$

$$\mathsf{MSE}(\boldsymbol{X}\hat{\theta}^{\mathsf{L}}) \leq \frac{24\|\theta^*\|_0 \sigma^2 \log(2d/\delta)}{\kappa n}.$$

Proof. Define the event

$$\mathcal{A} := \left\{ \max_{j=1,\ldots,d} \frac{1}{n}\langle \boldsymbol{X}_j, \epsilon\rangle \leq \tau/2 \right\}.$$

Notice that

$$\mathbb{P}(\mathcal{A}^\mathsf{c}) \leq 2d \exp\left(-\frac{n^2\tau^2}{8\sigma^2 \max_j \|\boldsymbol{X}_j\|_2^2}\right) \leq 2d \exp\left(-\frac{n\tau^2}{8\sigma^2}\right) \leq \delta,$$

since by assumption $\max_j \|\boldsymbol{X}_j\|_2 \leq \sqrt{n}$, and the ϵ_i are σ^2-sub-Gaussian.

To ease notation let

$$L(\theta; \tau) := L_n(\theta; \tau) := \frac{1}{2n}\|y - \boldsymbol{X}\theta\|_2^2 + \tau\|\theta\|_1,$$

and let's write $\Delta := \hat{\theta}^{\mathsf{L}} - \theta^*$.

Since $y - \boldsymbol{X}\hat{\theta}^{\mathsf{L}} = -\boldsymbol{X}\Delta + \epsilon$, we get

$$L(\hat{\theta}^{\mathsf{L}}; \tau) = \frac{1}{2n}\|\boldsymbol{X}\Delta\|_2^2 - \frac{1}{n}\langle \boldsymbol{X}\Delta, \epsilon\rangle + \frac{\|\epsilon\|_2^2}{2n} + \tau\|\hat{\theta}^{\mathsf{L}}\|_1.$$

Note that by definition of $\hat{\theta}^{\mathsf{L}}$ we have that

$$L(\hat{\theta}^{\mathsf{L}};\tau) \leq L(\theta^*;\tau) = \frac{1}{2n}\|\epsilon\|_2^2 + \tau\|\theta^*\|_1,$$

which after re-arrangement gives

$$0 \leq \frac{1}{2n}\|\boldsymbol{X}\Delta\|_2^2 \leq \frac{1}{n}\langle \boldsymbol{X}\Delta, \epsilon\rangle + \tau\left[\|\theta^*\|_1 - \|\hat{\theta}^{\mathsf{L}}\|_1\right]. \tag{2.15}$$ {eq:fund_lagran

Since by definition θ^* is supported on S we have that

$$\begin{aligned}\|\theta^*\|_1 - \|\hat{\theta}^{\mathsf{L}}\|_1 = \|\theta^*_S\|_1 - \|\hat{\theta}^{\mathsf{L}}_S\|_1 - \|\hat{\theta}^{\mathsf{L}}_{S^{\mathsf{c}}}\|_1 = \|\theta^*_S\|_1 - \|\theta^*_S + \Delta_S\|_1 - \|\Delta_{S^{\mathsf{c}}}\|_1 \\ \leq \|\theta^*_S\|_1 - (\|\theta^*_S\|_1 - \|\Delta_S\|_1) - \|\Delta_{S^{\mathsf{c}}}\|_1 = \|\Delta_S\|_1 - \|\Delta_{S^{\mathsf{c}}}\|_1.\end{aligned}$$

Thus continuing from (2.15) we have

$$\begin{aligned}0 \leq \frac{1}{n}\|\boldsymbol{X}\Delta\|_2^2 &\leq \frac{2}{n}\langle \boldsymbol{X}^{\mathsf{T}}\epsilon, \Delta\rangle + 2\tau\left(\|\Delta_S\|_1 - \|\Delta_{S^{\mathsf{c}}}\|_1\right) \\ &\leq 2\|\Delta\|_1 \max_{j=1,\ldots,d}\frac{|\langle \boldsymbol{X}_j, \epsilon\rangle|}{n} + 2\tau\left(\|\Delta_S\|_1 - \|\Delta_{S^{\mathsf{c}}}\|_1\right)\end{aligned}$$

and on the event $\mathcal{A}$

$$\leq \tau\left(3\|\Delta_S\|_1 - \|\Delta_{S^{\mathsf{c}}}\|_1\right).$$

This implies that on the event $\mathcal{A}$, the error vector Δ belongs to $\mathbb{C}_3(S)$. Thus, continuing from above

$$\frac{1}{n}\|\boldsymbol{X}\Delta\|_2^2 \leq 3\tau\|\Delta_S\|_1 \leq 3\tau\sqrt{s}\|\Delta\|_2 \leq 3\tau\sqrt{s}\frac{\|\boldsymbol{X}\Delta\|_2}{\sqrt{\kappa n}} \tag{2.16}$$ {eq:fast_rate_i

by the $l_1 - l_2$-inequality and the assumption that $\boldsymbol{X}$ satisfies the $(\kappa, 3)$-RE condition with respect to S. Rearranging the above we finally get

$$\frac{1}{\sqrt{n}}\|\boldsymbol{X}\Delta\|_2 \leq 3\tau\sqrt{\frac{s}{\kappa}}, \tag{2.17}$$ {eq:fast_rate_i

whence we conclude that

$$\mathsf{MSE}(\boldsymbol{X}(\hat{\theta}^{\mathsf{L}}) \leq \frac{24\|\theta^*\|_0\sigma^2\log(2d/\delta)}{\kappa n}.$$

□

Remark 2.3.1. *comment about necessity of RE condition?*

2.4 Bounds on l_2-error

So far we have mainly discussed the prediction error, that is error incurred when using our estimator $\hat{\theta}$ to predict the response vector y for a fresh instance of the noise vector. However, one may be more interested in inference for the parameter vector θ^* itself. In this section we will obtain bounds on the l_2 error $\|\theta^* - \hat{\theta}^{\mathsf{L}}\|_2^2$.

We will again be making use of the Restricted Eigenvalue condition, and perhaps here it will be easier to visualise its importance. To simplify things suppose that instead of

(2.12) we consider the constrained problem (2.5) with $K = B_1(R)$, where $R = \|\theta^*\|_1$, so that the truth is feasible. That is we consider the problem

$$\hat{\theta} := \arg\min_{\theta \in B_1(R)} L_n(\theta),$$

where $L_n(\cdot)$ is the empirical loss

$$L_n(\theta) := \frac{1}{2n}\|y - \boldsymbol{X}\theta\|_2^2.$$

Notice that $\hat{\theta}$ above is a minimiser of the empirical loss, whereas θ^* by definition minimises the population loss, that is

$$L(\theta) := \mathbb{E}\left[\|y - \boldsymbol{X}\theta\|_2^2\right] = \mathbb{E}\left[\|\boldsymbol{X}(\theta^* - \theta) + \epsilon\|_2^2\right].$$

When the sample size n is large, one would expect that the empirical and population errors should be close, and therefore that θ^* should almost minimise the empirical loss; that is one would expect that $L_n(\theta^*) \approx L_n(\hat{\theta})$, and indeed this is what we observed in the last section. When does $|L_n(\theta^*) - L_n(\hat{\theta})|$ being small also imply that $|\theta^* - \hat{\theta}|$ is also small? The reverse situation is probably easier to parse: can we have $|\theta^* - \hat{\theta}|$ large and $|L_n(\theta^*) - L_n(\hat{\theta})|$ small? The answer is of course yes, when L_n is flat. In the opposite direction, if f is strongly convex, that is we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \kappa\|y - x\|_2^2,$$

for all x, y, x^* is a minimiser, whence $\nabla f(x^*) = 0$, and $f(x') - f(x^*) \leq \epsilon$, then

$$\kappa\|x' - x^*\|_2^2 \leq f(x') - f(x^*) < \epsilon,$$

concluding that $\|x' - x^*\|_2^2 \leq \epsilon/\kappa$.

Although the quadratic form of the loss may raise some hope that L_n is indeed strongly convex, a more careful examination shows that in the high-dimensional case that is of interest to us this is impossible. In fact, the loss will be completely flat along any direction in the nullspace of $\boldsymbol{X}$ and the nullspace is certainly non-empty when $d \gg n$. Therefore in the high-dimensional case, there is only hope of $\|\theta^* - \hat{\theta}\|_2^2$ being small if θ^* does is not too aligned with the nullspace of $\boldsymbol{X}$. The restricted eigenvalue condition quantifies the above intuition.

Example 7. *Consider the following simple scenario. Let $d = 2$ and $n = 1$ and suppose for simplicity that $\boldsymbol{X} = [1, 0]$. Thus we have a single observation given by*

$$y = \boldsymbol{X}[\theta_1^*, \theta_2^*] + \epsilon = \theta_1^* + \epsilon.$$

We are trying to infer θ^*. Clearly there isn't much hope in recovering any information about θ_2^*. Suppose first that $\theta^* = [1, 0]$, so that $S = \{1\}$ and $\boldsymbol{X}$ satisfies the $(\kappa, 3)$-RE condition: let $v \in \mathbb{C}_3(S)$, that is $3|v_1| \geq |v_2|$. Then

$$\begin{aligned}\|\boldsymbol{X}v\|_2^2 = v_1^2 &= (1-\epsilon)v_1^2 + \epsilon v_1^2 \\ &\geq (1-\epsilon)v_1^2 + \frac{\epsilon}{9}v_2^2 \geq \kappa\|v\|_2^2,\end{aligned}$$

where $\kappa := \min\{(1-\epsilon), \epsilon/9\}$.

Let $\hat{\theta}$ solve (2.12). Then notice that

$$(y - \boldsymbol{X}\hat{\theta})^2 + \lambda\|\hat{\theta}\|_1 = (1 + \epsilon - \hat{\theta}_1)^2 + \lambda|\hat{\theta}_1| + \lambda|\hat{\theta}_2|,$$

which is clearly minimised at $\hat{\theta} = [1 + \epsilon - \lambda/2, 0]$, so that the prediction and l_2(squared) error are both given by $(\epsilon - \lambda)^2$. In particular for a small regularisation parameter λ and small $\operatorname{var}\epsilon$, both errors will be small.

Now, suppose that $\theta^* = [0, 1]$, so that $\boldsymbol{X}$ no longer satisfies the $(\kappa, 3)$-RE for any $\kappa > 0$. Then $y = \epsilon$ and thus

$$(y - \boldsymbol{X}\hat{\theta})^2 + \lambda\|\hat{\theta}\|_1 = (\epsilon - \hat{\theta}_2)^2 + \lambda|\hat{\theta}_1| + \lambda|\hat{\theta}_2|,$$

which, at least for small enough λ is minimised at, $\hat{\theta} = [0, \epsilon - \lambda/2]$. Now notice that the prediction error is then

$$\|\boldsymbol{X}(\theta^* - \hat{\theta})\|_2^2 = 0,$$

where as the l_2-error is

$$\|\theta^* - \hat{\theta}\|_2^2 = \left(1 - \epsilon + \frac{\lambda}{2}\right)^2,$$

which is much larger.

{thm:lasso_l2}

Theorem 13. *Under the assumptions of Theorem 12, with probability at least $1 - \delta$ we have*

$$\|\hat{\theta}^{\mathsf{L}} - \theta^*\|_2 \leq \frac{3}{\kappa}\sqrt{\frac{8|S|\sigma^2 \log(2d/\delta)}{n}}.$$

Proof. The proof is essentially contained in the proof of Theorem 12. Recall just before (2.16) that on an event $\mathcal{A}$ of probability at least $1 - \delta$, the error vector $\Delta = \hat{\theta}^{\mathsf{L}} - \theta^*$ belongs to $\mathbb{C}_3(S)$. Therefore from (2.17) and the $(\kappa, 3)$-RE condition, we conclude from

$$\sqrt{\kappa}\|\Delta\|_2 \leq \frac{1}{\sqrt{n}}\|\boldsymbol{X}\Delta\|_2 \leq 3\tau\sqrt{\frac{|S|}{\kappa}}.$$

□

2.5 Random design

So far we have studied the fixed design case under assumptions like the restricted nullspace or restricted eigenvalue condition. Now we will show that for certain random designs, mainly Gaussian, these conditions hold with high probability. The results will be based on a few auxiliary results, most importantly some comparison inequalities for Gaussian processes.

{lem:interpolat

Lemma 2.5.1 (Interpolation)**.** *Suppose $\boldsymbol{X} \sim \mathcal{N}(\mathbb{O}, \Sigma^{\boldsymbol{X}})$ and $\boldsymbol{Y} \sim \mathcal{N}(\mathbb{O}, \Sigma^{\boldsymbol{Y}})$ are independent Gaussian vectors and define their interpolation*

$$\boldsymbol{Z} = \sqrt{t}\boldsymbol{X} + \sqrt{1 - t}\boldsymbol{Y}, \quad t \in [0, 1].$$

Then for every smooth function f we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}[f(\boldsymbol{Z}_t)] = \frac{1}{2}\sum_{i,j=1}^{n}\left(\Sigma_{ij}^{\boldsymbol{X}} - \Sigma_{ij}^{\boldsymbol{Y}}\right)\mathbb{E}\left[\frac{\partial^2}{\partial x_j \partial x_i} f(\boldsymbol{Z}_t)\right].$$

Before proving the lemma we will an auxiliary result which is a multivariate form of Stein's lemma.

Lemma 2.5.2 (Multivariate Stein's Lemma)**.** *Let* $\boldsymbol{X} \sim \mathcal{N}(\mathbb{0}, \Sigma)$*. Then*

$$\mathbb{E}[X_i f(\boldsymbol{X})] = \sum_j \Sigma_{ij} \, \mathbb{E}\left[\frac{\partial f}{\partial x_j} f(\boldsymbol{X})\right],$$

for any function f *for which both sides make sense.*

Proof. We can write $\boldsymbol{X}$ as $\boldsymbol{X} = \Sigma^{1/2}\boldsymbol{\xi}$, where ξ is a standard normal vector. Then we have

$$\begin{aligned}
\mathbb{E}[X_i f(x)] &= \sum_k \Sigma_{ik}^{1/2} \, \mathbb{E}[\xi_k f(\Sigma^{1/2}\boldsymbol{\xi})] \\
&= \sum_k \Sigma_{ik}^{1/2} \, \mathbb{E}\left\{\mathbb{E}\left[\xi_k f(\Sigma^{1/2}\boldsymbol{\xi}) \mid \boldsymbol{\xi}_{-i}\right]\right\}
\end{aligned}$$

and applying Stein's lemma on the function $\xi_k \mapsto f(\Sigma^{1/2}\xi)$

$$\begin{aligned}
&= \sum_k \Sigma_{ik}^{1/2} \sum_j \mathbb{E}\left\{\mathbb{E}\left[\frac{\partial f}{\partial x_j}(\Sigma^{1/2}\boldsymbol{\xi})\Sigma_{jk}^{1/2} \mid \boldsymbol{\xi}_{-i}\right]\right\} \\
&= \sum_k \sum_j \Sigma_{ik}^{1/2}\Sigma_{jk}^{1/2} \, \mathbb{E}\left\{\frac{\partial f}{\partial x_j}(\boldsymbol{X})\right\} \\
&= \sum_j \mathbb{E}\left\{\frac{\partial f}{\partial x_j}(\boldsymbol{X})\right\} \sum_k \Sigma_{ik}^{1/2}\Sigma_{jk}^{1/2} \\
&= \sum_j \mathbb{E}\left\{\frac{\partial f}{\partial x_j}(\boldsymbol{X})\right\} \Sigma_{ij},
\end{aligned}$$

where we used the fact that by definition of the square root of the symmetric matrix Σ, we have $\sum_k \Sigma_{jk}^{1/2}\Sigma_{ik}^{1/2} = \Sigma_{ij}$. □

Proof of Lemma 2.5.1. We can easily see that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}[f(\boldsymbol{Z}_t)] &= \sum_{i=1}^n \mathbb{E}\left[\frac{\partial}{\partial x_i} f(\boldsymbol{Z}_t) \frac{\mathrm{d}Z_i(t)}{\mathrm{d}t}\right] \\
&= \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[\frac{\partial}{\partial x_i} f(\sqrt{t}\boldsymbol{X} + \sqrt{1-t}\boldsymbol{Y})\left(\frac{X_i}{\sqrt{t}} - \frac{Y_i}{\sqrt{1-t}}\right)\right] \\
&= \frac{1}{2}\sum_{j=1}^n \sum_{i=1}^n \Sigma_{ij}^{\boldsymbol{X}} \, \mathbb{E}\left[\frac{\partial^2}{\partial x_j \partial x_i} f(\sqrt{t}\boldsymbol{X} + \sqrt{1-t}\boldsymbol{Y})\right] \\
&\quad - \frac{1}{2}\sum_{j=1}^n \sum_{i=1}^n \Sigma_{ij}^{\boldsymbol{Y}} \, \mathbb{E}\left[\frac{\partial^2}{\partial x_j \partial x_i} f(\sqrt{t}\boldsymbol{X} + \sqrt{1-t}\boldsymbol{Y})\right] \\
&= \frac{1}{2}\sum_{j=1}^n \sum_{i=1}^n \left(\Sigma_{ij}^{\boldsymbol{X}} - \Sigma_{ij}^{\boldsymbol{Y}}\right) \mathbb{E}\left[\frac{\partial^2}{\partial x_j \partial x_i} f(\boldsymbol{Z}_t)\right].
\end{aligned}$$

□

We next start present the first Gaussian comparison inequality, due to Sudakov and Fernique.

Theorem 14 (Sudakov-Fernique)**.** *Let $\boldsymbol{X}$, $\boldsymbol{Y}$ be Gaussian vectors such that $\mathbb{E}[X_i] = \mathbb{E}[Y_i] = \mu_i$ for all i, and $\mathbb{E}[(X_i - X_j)^2] \leq \mathbb{E}[(Y_i - Y_j)^2]$ whenever $i \neq j$. Then we have*

$$\mathbb{E}[\max X_i] \leq \mathbb{E}[\max Y_i].$$

Proof. We want to apply the Interpolation Lemma 2.5.1, but the max is not really a smooth function. We instead first consider a soft-max function given by

$$f_\beta(x) = \frac{1}{\beta} \log \sum_i \mathrm{e}^{\beta x_i}.$$

An easy calculation shows that

$$\begin{aligned}
\frac{\partial}{\partial x_i} f_\beta(x) &= \frac{\mathrm{e}^{\beta x_i}}{\sum \mathrm{e}^{\beta x_i}} =: p_i(x), \\
\frac{\partial^2}{\partial_j \partial x_i} f_\beta(x) &= \delta_{i,j} \frac{\beta \mathrm{e}^{\beta x_i}}{\sum \mathrm{e}^{\beta x_i}} - \frac{\beta \mathrm{e}^{\beta x_i} \mathrm{e}^{\beta x_j}}{\left(\sum \mathrm{e}^{\beta x_i}\right)^2} = \beta \left[\delta_{ij} p_i(x) - p_i(x) p_j(x)\right],
\end{aligned}$$

where the choice of notation obviously hints at the fact that the $p_i(x)$ form a probability vector for each x.

Applying Lemma 2.5.1 to f_β we thus get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}[f_\beta(\boldsymbol{Z}_t)] &= \frac{\beta}{2} \sum_{i,j=1}^n \left(\Sigma_{ij}^{\boldsymbol{X}} - \Sigma_{ij}^{\boldsymbol{Y}}\right) \mathbb{E}\left[\delta_{ij} p_i\left(\boldsymbol{Z}(t)\right)\right] \\
&\quad - \frac{\beta}{2} \sum_{i,j=1}^n \left(\Sigma_{ij}^{\boldsymbol{X}} - \Sigma_{ij}^{\boldsymbol{Y}}\right) \mathbb{E}\left[p_i\left(\boldsymbol{Z}(t)\right) p_j\left(\boldsymbol{Z}(t)\right)\right] \\
&= \frac{\beta}{2} \sum_{i}^n \left(\Sigma_{ii}^{\boldsymbol{X}} - \Sigma_{ii}^{\boldsymbol{Y}}\right) \mathbb{E}\left[p_i\left(\boldsymbol{Z}(t)\right)\right] \\
&\quad - \frac{\beta}{2} \sum_{i,j=1}^n \left(\Sigma_{ij}^{\boldsymbol{X}} - \Sigma_{ij}^{\boldsymbol{Y}}\right) \mathbb{E}\left[p_i\left(\boldsymbol{Z}(t)\right) p_j\left(\boldsymbol{Z}(t)\right)\right]
\end{aligned}$$

next using the fact that $\sum_j p_j(x) = 1$

$$\begin{aligned}
&= \frac{\beta}{2} \sum_{i,j=1}^n \left(\Sigma_{ii}^{\boldsymbol{X}} - \Sigma_{ii}^{\boldsymbol{Y}}\right) \mathbb{E}\left[p_i\left(\boldsymbol{Z}(t)\right) p_j\left(\boldsymbol{Z}(t)\right)\right] \\
&\quad - \frac{\beta}{2} \sum_{i,j=1}^n \left(\Sigma_{ij}^{\boldsymbol{X}} - \Sigma_{ij}^{\boldsymbol{Y}}\right) \mathbb{E}\left[p_i\left(\boldsymbol{Z}(t)\right) p_j\left(\boldsymbol{Z}(t)\right)\right].
\end{aligned}$$

After noticing that terms with $i = j$ cancel we can write

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}[f_\beta(\boldsymbol{Z}_t)] &= \frac{\beta}{2} \sum_{i<j}^n \left(\Sigma_{ii}^{\boldsymbol{Y}} - \Sigma_{ii}^{\boldsymbol{X}} + \Sigma_{jj}^{\boldsymbol{Y}} - \Sigma_{jj}^{\boldsymbol{X}} + 2\Sigma_{ij}^{\boldsymbol{X}} - 2\Sigma_{ij}^{\boldsymbol{Y}}\right) \mathbb{E}\left[p_i\left(\boldsymbol{Z}(t)\right) p_j\left(\boldsymbol{Z}(t)\right)\right] \\
&= \frac{\beta}{2} \sum_{i<j}^n \left(\gamma_{ij}^{\boldsymbol{Y}} - \gamma_{ij}^{\boldsymbol{X}}\right) \mathbb{E}\left[p_i\left(\boldsymbol{Z}(t)\right) p_j\left(\boldsymbol{Z}(t)\right)\right],
\end{aligned}$$

where

$$\gamma_{ij}^{\boldsymbol{X}} = \Sigma_{ii}^{\boldsymbol{X}} + \Sigma_{jj}^{\boldsymbol{X}} - 2\Sigma_{ij}^{\boldsymbol{X}} = \mathbb{E}\left[(X_i - \mu_i - X_j + \mu_j)^2\right] = \mathbb{E}[(X_i - X_j)^2] - (\mu_i - \mu_j)^2$$

$$\gamma_{ij}^{\boldsymbol{Y}} = \Sigma_{ii}^{\boldsymbol{Y}} + \Sigma_{jj}^{\boldsymbol{Y}} - 2\Sigma_{ij}^{\boldsymbol{Y}} = \mathbb{E}\left[(Y_i - \mu_i^Y - Y_j + \mu_j^Y)^2\right] = \mathbb{E}[(Y_i - Y_j)^2] - (\mu_i - \mu_j)^2.$$

Therefore, the assumption implies that $\gamma_{ij}^{\boldsymbol{X}} \leq \gamma_{ij}^{\boldsymbol{Y}}$ and thus since $p_i \geq 0$ we conclude that $\mathbb{E}[f_\beta(\boldsymbol{Z}_t)]$ is increasing in t and therefore that

$$\mathbb{E}[f_\beta(\boldsymbol{X})] \leq \mathbb{E}[f_\beta(\boldsymbol{Y})].$$

Letting $\beta \to \infty$, the dominated convergence theorem gives us the desired result. □

Theorem 15. *Consider a random matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with i.i.d. $\mathcal{N}(0,1)$ entries. Then there are universal positive constants $c_1 < 1 < c_2$ such that*

$$\frac{\|\boldsymbol{X}\theta\|_2^2}{n} \geq c_1\|\theta\|_2^2 - c_2\frac{\log d}{n}\|\theta\|_1^2, \quad \theta \in \mathbb{R}^d,$$

with probability at least $1 - \mathrm{e}^{-n/32}/(1 - \mathrm{e}^{-n/32})$.

Proof. First notice it suffices to only consider $\theta \in \mathbb{S}^{d-1}$. Let $g(t) := 2\sqrt{\frac{\log d}{n}}t$ and define the event

$$\mathcal{E} := \left\{\boldsymbol{X} \in \mathbb{R}^{n\times d} : \inf_{\theta\in\mathbb{S}^{d-1}} \frac{\|\boldsymbol{X}\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2g(\|\theta\|_1)\right\}.$$

Exercise: Show that on the complement of the event $\mathcal{E}$ the desired bound holds.

To complete the proof we have to obtain an upper bound for $\mathbb{P}[\mathcal{E}]$.

Although we have total control of the l_2 norm of $\|\theta\|_2^2$, since we are in high-dimensions, its l_1 can actually vary up to $\sqrt{n}$. Thus we would like to split the event $\mathcal{E}$ into smaller events that allow us finer control on the size of $\|\theta\|_1$. For $0 < r < s$ let

$$K(r,s) = \{\theta \in \mathbb{S}^{d-1} : g(\|\theta\|_1) \in [r,s]\},$$

and consider the family of events

$$\mathcal{A}(r,s) := \left\{\inf_{K(r,s)} \frac{\|\boldsymbol{X}\theta\|_2}{\sqrt{n}} \leq \frac{1}{2} - 2s\right\}.$$

Notice that we have the following inclusion

$$\mathcal{E} \subset \mathcal{A}\left(0, \frac{1}{4}\right) \cup \left(\bigcup_{l=1}^{\infty} \mathcal{A}\left(\frac{2^{l-1}}{4}, \frac{2^l}{4}\right)\right).$$

To see why let θ be the vector where the infimum in the event $\mathcal{E}$ is attained. Then θ must belong to either $K(0,1/4)$ or one of the $K(2^{l-1}/4, 2^l/4)$. In the first case we have $g(\|\theta\|_1) \leq 1/4$ and for $\boldsymbol{X} \in \mathcal{E}$, we have

$$\frac{\|\boldsymbol{X}\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2g(\|\theta_1\|) \leq \frac{1}{4} = \frac{1}{2} - \frac{1}{4},$$

and thus $\boldsymbol{X} \in \mathcal{A}(0,1/4)$. Otherwise $\theta \in K(2^{l-1}/4, 2^l/4)$ for some $l \geq 1$ and then

$$\frac{\|\boldsymbol{X}\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2g(\|\theta_1\|) \leq \frac{1}{2} - 2 \times \frac{2^{l-1}}{4} \leq \frac{1}{2} - \frac{2^l}{4},$$

and thus $\boldsymbol{X} \in \mathcal{A}(2^{l-1}/4, 2^l/4)$.

Thus a simple union bound gives

$$\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{A}(0,1/4)] + \sum_{l=1}^{\infty} \mathbb{P}\left[\mathcal{A}\left(\frac{2^{l-1}}{4}, \frac{2^l}{4}\right)\right].$$

We now bound the probability of the event $\mathcal{A}(r,s)$. In fact we will now show that for any $0 < r < s$ we have

$$\mathbb{P}[\mathcal{A}(r,s)] \leq \mathrm{e}^{-n/32}\mathrm{e}^{-ns^2/2}.$$

We can equivalently aim for a high-probability lower bound on the quantity

$$T(r,s) := -\inf_{\theta \in K(r,s)} \frac{\|\boldsymbol{X}\theta\|_2}{\sqrt{n}}.$$

Using the variational representation of the l_2 norm we have

$$T(r,s) = -\inf_{\theta \in K(r,s)} \sup_{u \in \mathbb{S}^{n-1}} \frac{\langle u, \boldsymbol{X}\theta\rangle}{\sqrt{n}} = \sup_{\theta \in K(r,s)} \inf_{u \in \mathbb{S}^{n-1}} \frac{\langle u, \boldsymbol{X}\theta\rangle}{\sqrt{n}}.$$

Let's write $X_{u,\theta} := \langle u, \boldsymbol{X}\theta\rangle$ for the Gaussian process indexed by $(u,\theta) \in \mathbb{S}^{n-1} \times \mathbb{S}^{d-1}$. Notice that $X_{u,\theta} \sim \mathcal{N}(0, n^{-1})$.

To obtain this bound we will use the following result from Gordon (1985) which we will state without proof.

Theorem 16 (Theorem 1.4 from Gordon 1985)**.** *Let $\{X_{i,j} : i \in [n], j \in [m]\}$, $\{Y_{i,j} : i \in [n], j \in [m]\}$ be two arrays of centred Gaussian random variables such that*

$$\mathbb{E}\left[(Y_{i,j} - Y_{i,k})^2\right] \leq \mathbb{E}\left[(X_{i,j} - X_{i,k})^2\right], \quad \textit{for all } i,j,k, \textit{ and}$$
$$\mathbb{E}\left[(Y_{i,j} - Y_{l,k})^2\right] \geq \mathbb{E}\left[(X_{i,j} - X_{l,k})^2\right], \quad \textit{for any } i \neq l.$$

Then

$$\mathbb{E}\left[\min_{i\in[n]} \max_{j\in[m]} Y_{i,j}\right] \leq \mathbb{E}\left[\min_{i\in[n]} \max_{j\in[m]} X_{i,j}\right].$$

First notice that if both conditions above hold, then they will also both hold for $(-X_{i,j})$ and $(-Y_{i,j})$. The conclusion of the theorem then holds for these processes too, that is

$$\begin{aligned}
&\mathbb{E}\left[\min_{i\in[n]} \max_{j\in[m]} (-Y_{i,j})\right] \leq \mathbb{E}\left[\min_{i\in[n]} \max_{j\in[m]} (-X_{i,j})\right] \\
&\Leftrightarrow \mathbb{E}\left[\min_{i\in[n]} (-\min_{j\in[m]} Y_{i,j})\right] \leq \mathbb{E}\left[\min_{i\in[n]} (-\min_{j\in[m]} X_{i,j})\right] \\
&\Leftrightarrow \mathbb{E}\left[-\max_{i\in[n]} \min_{j\in[m]} Y_{i,j}\right] \leq \mathbb{E}\left[-\max_{i\in[n]} \min_{j\in[m]} X_{i,j}\right] \\
&\Leftrightarrow \mathbb{E}\left[\max_{i\in[n]} \min_{j\in[m]} Y_{i,j}\right] \geq \mathbb{E}\left[\max_{i\in[n]} \min_{j\in[m]} X_{i,j}\right]
\end{aligned}$$

We will apply the max – min version of the above result to upper bound $\mathbb{E}[\sup_\theta \inf_u X_{u,\theta}]$ by comparing $\{X_{u,\theta}\}$ with the process

$$Y_{u,\theta} := \frac{\langle u, \boldsymbol{\xi}\rangle}{\sqrt{n}} + \frac{\langle \theta, \boldsymbol{\zeta}\rangle}{\sqrt{n}},$$

where $\boldsymbol{\xi}, \boldsymbol{\zeta}$ are Gaussian i.i.d. random vectors in $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively. Let us first consider the case where the first index is the same, that is

$$\begin{aligned}\mathbb{E}[(X_{u,\theta} - X_{u,\phi})^2] &= \mathbb{E}[\langle u, \boldsymbol{X}(\theta - \phi)\rangle^2] \\ &= \sum_{i,j} u_i^2(\theta_i - \phi_i)^2 = \|u\|_2^2\|\theta - \phi\|_2^2 = \|\theta - \phi\|_2^2.\end{aligned}$$

On the other hand

$$\begin{aligned}\mathbb{E}[(Y_{u,\theta} - Y_{u,\phi})^2] &= \mathbb{E}\left[\langle \boldsymbol{\zeta}, \theta - \phi\rangle^2\right] \\ &= \|\theta - \phi\|_2^2.\end{aligned}$$

Now let's consider the case $u \neq w$.

$$\begin{aligned}\mathbb{E}[(Y_{u,\theta} - Y_{w,\phi})^2] &= \mathbb{E}\left[(\langle u - w, \boldsymbol{\xi}\rangle + \langle \theta - \phi, \boldsymbol{\zeta}\rangle)^2\right] \\ &= \|u - w\|_2^2 + \|\theta - \phi\|_2^2.\end{aligned}$$

Next we compute for the process $X_{u,\theta}$ and find

$$\begin{aligned}\mathbb{E}[(X_{u,\theta} - X_{w,\phi})^2] &= \mathbb{E}\left[(\langle u, \boldsymbol{X}\theta\rangle - \langle w, \boldsymbol{X}\phi\rangle)^2\right] \\ &= \mathbb{E}\left[(\langle u, \boldsymbol{X}\theta\rangle - \langle w, \boldsymbol{X}\phi\rangle)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{i,j} \boldsymbol{X}_{ij}(u_i\theta_j - w_i\phi_j)\right)^2\right] \\ &= \sum_{i,j}(u_i\theta_j - w_i\phi_j)^2 = \|u\theta^\mathsf{T} - w\phi^\mathsf{T}\|_\mathsf{F}^2\end{aligned}$$

where $\|\cdot\|_\mathsf{F}$ denotes the Frobenius norm of a matrix, that is for $A = (a_{i,j})$, $\|A\|_\mathsf{F}^2 = \sum_{ij} a_{ij}^2$. For two matrices A, B of the same dimensions, we write $\langle A, B\rangle_\mathsf{F}$ for the Frobenius inner product, that is

$$\langle A, B\rangle_\mathsf{F} = \sum_{ij} A_{ij}B_{ij}.$$

Continuing from above we have

$$\begin{aligned}\mathbb{E}[(X_{u,\theta} - X_{w,\phi})^2] &\leq \|u\theta^\mathsf{T} - w\phi^\mathsf{T}\|_\mathsf{F}^2 \\ &= \|u(\theta - \phi)^\mathsf{T} + (u - w)\phi^\mathsf{T}\|_\mathsf{F}^2 \\ &= \|u(\theta - \phi)^\mathsf{T}\|_\mathsf{F} + \|(u - w)\phi^\mathsf{T}\|_\mathsf{F}^2 + 2\langle u(\theta - \phi)^\mathsf{T}, (u - w)\phi^\mathsf{T}\rangle_\mathsf{F}.\end{aligned}$$

Expanding the correlation term, after straightforward calculations we obtain

$$\begin{aligned}\langle u(\theta - \phi)^\mathsf{T}, (u - w)\phi^\mathsf{T}\rangle_\mathsf{F} &= \|u\|_2^2\langle\theta, \phi\rangle - \langle u, w\rangle\langle\theta, \phi\rangle - \|u\|_2^2\|\phi\|_2^2 + \langle u, w\rangle\|\phi\|_2^2 \\ &= \|u\|_2^2\left(\langle\theta, \phi\rangle - \|\phi\|_2^2\right) - \langle u, w\rangle\left(\langle\theta, \phi\rangle - \|\phi\|_2^2\right) \\ &= \left(\|u\|_2^2 - \langle u, w\rangle\right)\left(\langle\theta, \phi\rangle - \|\phi\|_2^2\right) \\ &\leq 0,\end{aligned}$$

since, $\|u\|_2^2 = \|w\|_2^2 = \|\theta\|_2^2 = \|\phi\|_2^2 = 1$, and therefore the first factor is positive and the second negative. Therefore we have

$$\mathbb{E}[(X_{u,\theta} - X_{w,\phi})^2] \leq \|u(\theta - \phi)^\mathsf{T}\|_\mathsf{F}^2 + \|(u - w)\phi^\mathsf{T}\|_\mathsf{F}^2$$

$$= \|u\|_2^2\|\theta-\phi\|_2^2 + \|\phi\|_2^2\|u-w\|_2^2 = \|u-w\|_2^2 + \|\theta-\phi\|_2^2 = \mathbb{E}[(Y_{u,\theta} - Y_{w,\phi})^2]$$

We can now apply Gordon's inequality to obtain

$$\begin{aligned}
\sqrt{n}\,\mathbb{E}[T(r,s)] &= \mathbb{E}\left[\sup_{\theta\in K(r,s)}\inf_{u\in\mathbb{S}^{d-1}} X_{u,v}\right] \\
&\leq \mathbb{E}\left[\sup_{\theta\in K(r,s)}\inf_{u\in\mathbb{S}^{d-1}} Y_{u,v}\right] \\
&= \mathbb{E}\left[\sup_{\theta\in K(r,s)}\langle\boldsymbol{\zeta},\theta\rangle\right] + \mathbb{E}\left[\inf_{u\in\mathbb{S}^{d-1}}\langle\boldsymbol{\xi},u\rangle\right] \\
&= \mathbb{E}\left[\sup_{\theta\in K(r,s)}\langle\boldsymbol{\zeta},\theta\rangle\right] - \mathbb{E}\left[\|\boldsymbol{\xi}\|_2\right] \\
&\leq \sup_{\theta\in K(r,s)}\|\theta\|_1\,\mathbb{E}\left[\|\boldsymbol{\zeta}\|_\infty\right] - \sqrt{\frac{2n}{\pi}}.
\end{aligned}$$

Now for $\theta \in K(r,s)$ we have that $g(\|\theta\|_1) \leq s$, or in other words that

$$\|\theta\|_1 \leq \frac{s}{2\sqrt{\log d/n}}.$$

In addition, from Lemma 1.3.2, we have that

$$\mathbb{E}[\|\boldsymbol{\zeta}\|_\infty] \leq \sqrt{2\log d},$$

and thus overall we have

$$\begin{aligned}
\mathbb{E}[T(r,s)] &\leq \sup_{\theta\in K(r,s)}\|\theta\|_1\,\mathbb{E}\left[\frac{\|\boldsymbol{\zeta}\|_\infty}{\sqrt{n}}\right] - \sqrt{\frac{2}{\pi}} \\
&\leq \frac{s}{2\sqrt{\log d/n}} \times \sqrt{2\log d/n} - \sqrt{\frac{2}{\pi}} \\
&\leq s - \sqrt{\frac{2}{\pi}}. \qquad (2.18)
\end{aligned}$$

{eq:mean_bound}

Finally, from Theorem 6 it follows that

$$\mathbb{P}\left(T(r,s) \geq \mathbb{E}[T(r,s)] + \delta\right) \leq \mathrm{e}^{-n\delta^2/2},$$

for all $\delta > 0$. Letting $\delta = \sqrt{2/\pi} - 1/2 + s$ and using the upper bound in (2.18) we have that

$$\begin{aligned}
\mathrm{e}^{-n\delta^2/2} &\geq \mathbb{P}\left(T(r,s) \geq \mathbb{E}[T(r,s)] + \delta\right) \\
&= \mathbb{P}\left(T(r,s) \geq \mathbb{E}[T(r,s)] + \sqrt{\frac{2}{\pi}} - \frac{1}{2} + s\right) \\
&\geq \mathbb{P}\left(T(r,s) \geq s - \sqrt{\frac{2}{\pi}} + \sqrt{\frac{2}{\pi}} - \frac{1}{2} + s\right) \\
&\geq \mathbb{P}\left(T(r,s) \geq 2s - \frac{1}{2}\right).
\end{aligned}$$

The result follows since $\delta^2 > C^2 + s^2$. □

Bibliography

[1] Peter J Bickel, Ya?acov Ritov, Alexandre B Tsybakov, et al. "Simultaneous analysis of Lasso and Dantzig selector". In: *The Annals of statistics* 37.4 (2009), pp. 1705–1732.

[2] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[3] SS Chen, DL Donoho, and MA Saunders. *Atomic decomposition by basis pursuit: SIA M Journal on Scientific Computing, 20, 33–61*. 1998.

[4] Yehoram Gordon. "Some inequalities for Gaussian processes and applications". In: *Israel Journal of Mathematics* 50.4 (1985), pp. 265–289.

[5] P. Rigollet and J.C. Hütter. *High Dimensional Statistics. Massachusetts Institute of Technology*. 2017.

[6] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[7] R. Van Handel. *Probability in High Dimension*. 2016. URL: `https://web.math.princeton.edu/~rvan/APC550.pdf`.

[8] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.