Lecture notes: kernel and operator-theoretic methods in machine learning

G. Blanchard, Université Paris-Saclay

May 7, 2026

Warning: these lecture notes are an incomplete work in progress. They likely contain many typos, inconsistencies and more serious errors! If you happen to stumble upon this document, and notice some errors, feel free to contact me.

Thanks: I want to warmly thank the following people, who gave me precious feedback on these notes: Simone Giancola, Hanqi Sun, Noé Garric, Antoine Roche...

Contents

1 Introduction (and conventions used in these notes)

1.1 Motivation: regression in Hilbert space

TODO

1.2 Notation and convention index

$\mathcal{H}$	separable Hilbert space
$\mathcal{B}(\mathcal{H}, \mathcal{H}')$	space of bounded linear operators from $\mathcal{H}$ to $\mathcal{H}'$; $\mathcal{B}(\mathcal{H}) := \mathcal{B}(\mathcal{H}, \mathcal{H})$
$\|A\|_{\text{op}}$	operator norm of bounded operator A, $\|A\|_{\text{op}} := \max_{\substack{x \in \mathcal{H} \\ \|x\|=1}} \|Ax\|$
$\text{HS}(\mathcal{H})$	space of Hilbert-Schmidt operators on $\mathcal{H}$
$\mathcal{K}(\mathcal{H}, \mathcal{H}')$	space of compact linear operators from $\mathcal{H}$ to $\mathcal{H}'$; $\mathcal{K}(\mathcal{H}) := \mathcal{K}(\mathcal{H}, \mathcal{H})$.
$\langle , \rangle_2$	if $A, B \in \text{HS}(\mathcal{H})$: $\langle A, B \rangle_2 = \text{Tr}(AB^*)$
$\|A\|_2$	Hilbert-Schmidt norm of A if $A \in \text{HS}(\mathcal{H})$: $\|A\|_2^2 = \langle A, A \rangle_2$
$\mathcal{B}_p(\mathcal{H})$	Schatten p-class of $\mathcal{H}$
$\|A\|_p$	Schatten p-norm of A if $A \in \mathcal{B}(\mathcal{H})$

2 Tools of operator theory and functional calculus

2.1 Basics on Hilbert spaces

[Source: [7, Chapter 1]]

Definition 2.1. A real resp. complex Hilbert space $\mathcal{H}$ is a $\mathbb{R}$-, resp $\mathbb{C}$-vector space with an inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{H}^2$ which is complete for the metric induced by that inner product's norm. We recall here the defining properties of an inner product:

- $\langle u, v \rangle \in \mathbb{R}$, resp $\mathbb{C}$;
- $\langle v, u \rangle = \overline{\langle u, v \rangle}$;
- $\langle \alpha u + \beta u', v \rangle = \alpha \langle u, v \rangle + \beta \langle u', v \rangle$;
- (consequence of the two previous items): $\langle u, \alpha v + \beta v' \rangle = \overline{\alpha} \langle u, v \rangle + \overline{\beta} \langle u, v' \rangle$.
- $\langle u, u \rangle \geq 0$;
- $\langle u, u \rangle = 0 \Rightarrow u = 0$.

The norm induced by the inner product is $\|u\| = \langle u, u \rangle^{\frac{1}{2}}$. The inner product satisfies the Cauchy-Schwarz inequality:

$$|\langle u, v \rangle| \leq \|u\| \|v\|.$$

Proposition/Definition 2.2.

- If $\langle u, v \rangle = 0$ we denote $u \perp v$.
- For a subset $\mathcal{A} \subseteq \mathcal{H}$ we denote $\mathcal{A}^\perp = \{u \in \mathcal{H} : \langle u, a \rangle = 0 \text{ for all } a \in \mathcal{A}\}$.
- $(\mathcal{A}^\perp)^\perp = [\mathcal{A}]$, where $[\mathcal{A}]$ is the closure of the linear span of $\mathcal{A}$.
- If $\mathcal{A}$ is a closed linear subspace of $\mathcal{H}$, then $\mathcal{A} \oplus \mathcal{A}^\perp = \mathcal{H}$, i.e. for any $u \in \mathcal{H}$ there exists a unique decomposition $u = u^\parallel + u^\perp$, where $u^\parallel \in \mathcal{A}$ and $u^\perp \in \mathcal{A}^\perp$.
  The component $u^\parallel$ is called orthogonal projection of U onto $\mathcal{A}$.

We will need the following definition:

Proposition/Definition 2.3. If $(u_i, i \in I)$ is a family of elements of $\mathbb{R}$, $\mathbb{C}$, or a normed vector space, we will say that it is Hilbert-summable with sum h if:

1. $u_i = 0$ for all i except on a finite or countable subset of indices $\widetilde{I} \subseteq I$;

2. The sum $\sum_{i\in\widetilde{I}} u_i$ converges to h regardless of the order of summation (unconditional convergence), i.e. for any ordering $(i_k)_{k\in\mathbb{N}}$ of the elements of $\widetilde{I}$, we have

$$\left\| \sum_{k=1}^{n} u_{i_k} - h \right\| \to 0, \qquad \text{as } n \to \infty.$$

For real or complex-valued sums (and more generally finite-dimensional vector spaces), the above is equivalent to the corresponding sum being absolutely convergent.

The first point allows to formally define uncountable Hilbert sums; the second indicates that the notion of summability is quite strong. For sums of (infinite-dimensional) Hilbert space elements however, Hilbert-summability does not imply that $\sum_{i\in\widetilde{I}}\|u_i\|$ is convergent.

In the rest of the section all infinite sums will be Hilbert-sums.

Proposition/Definition 2.4.

- An orthonormal set $(e_k)_{k\in I}$ is a family of unit norm, pairwise orthogonal vectors.
- A (Hilbert) basis is a maximal orthonormal set.
- For an orthonormal set $(e_k)_{k\in I}$, it holds for any $h \in \mathcal{H}$:

  $$\sum_{k\in I} |\langle h, e_k\rangle|^2 \leq \|h\|^2.$$

  (Bessel's inequality)
- Any orthonormal set can be completed to a basis containing it.
- Any basis of $\mathcal{H}$ has the same cardinality; if $\mathcal{H}$ is separable, then any basis of $\mathcal{H}$ is countable (or finite)

Proposition 2.5. *If $\mathcal{E} = (e_k)_{k\in I}$ is a Hilbert basis of $\mathcal{H}$, we have the following properties:*

- $\mathcal{E}^\perp = \{0\}$.
- $[\mathcal{E}] = \mathcal{H}$.
- *For any $h \in \mathcal{H}$, it holds $\sum_{k\in I}\langle h, e_k\rangle e_k = h$.*
- *For any $h, g \in \mathcal{H}$, it holds $\langle h, g\rangle = \sum_{k\in I}\langle h, e_k\rangle\overline{\langle g, e_k\rangle}$.*
- *(Consequence of the previous point) For any $h \in \mathcal{H}$, it holds $\|h\|^2 = \sum_{k\in I}|\langle h, e_k\rangle|^2$ (Parseval's identity).*

(All sums above are Hilbert sums.)

To link back to Def. 2.3, in an infinite-dimensional Hilbert space with countable orthonormal set $(e_k)_{k\in\mathbb{N}}$, the series $u_k = e_k/k$ is Hilbert-summable but not absolutely summable.

2.2 Bounded operators on Hilbert space

[Source: [7, Chapter 2]]

Proposition/Definition 2.6. For a linear operator A between Hilbert spaces $\mathcal{H}$ and $\mathcal{H}'$, the following are equivalent:

- A is continuous.
- A is bounded on the unit ball: $\sup_{h\in\mathcal{H},\|h\|\leq 1}\|Ah\| < \infty$.

In this case we define the operator norm

$$\|A\|_{\text{op}} := \sup_{h\in\mathcal{H},\|h\|\leq 1} \|Ah\|;$$

it is a norm on the space of bounded/continuous operators, denoted $\mathcal{B}(\mathcal{H},\mathcal{H}')$.
The space of bounded operators is complete for the metric induced by the operator norm.
It holds for any $h \in \mathcal{H}$:

$$\|Ah\| \leq \|A\|_{\text{op}}\|h\|.$$

It holds for bounded operators A, B s.t. the output space of B is the input space of A:

$$\|AB\|_{\text{op}} \leq \|A\|_{\text{op}}\|B\|_{\text{op}}.$$

Proposition/Definition 2.7. For an operator $A \in \mathcal{B}(\mathcal{H},\mathcal{H}')$ there exists a unique $A^* \in \mathcal{B}(\mathcal{H}',\mathcal{H})$ called adjoint of A, such that

$$\forall u \in \mathcal{H}, v \in \mathcal{H}' \qquad \langle Au, v\rangle = \langle u, A^*v\rangle.$$

If $A \in \mathcal{B}(\mathcal{H})$, it is said self-adjoint if $A^* = A$.

Proposition 2.8. *We have the following properties for bounded operators A, B (with appropriate compatibility for input/output spaces of operators):*

- $(\alpha A + B)^* = \overline{\alpha}A^* + B^*$.
- $(AB)^* = B^*A^*$.
- *(Consequence of the previous point) AA^* and A^*A are self-adjoint.*
- $(A^*)^* = A$.
- *If A is invertible in $\mathcal{B}(\mathcal{H},\mathcal{H}')$, then so is A^* and $(A^{-1})^* = (A^*)^{-1}$.*
- $\|A\|_{\text{op}} = \|A^*\|_{\text{op}} = \|AA^*\|_{\text{op}}^{\frac{1}{2}} = \|A^*A\|_{\text{op}}^{\frac{1}{2}}$.

Proof. We prove only the last statement: for any $h \in \mathcal{H}$ it holds

$$\|Ah\|^2 = \langle Ah, Ah\rangle = \langle A^*Ah, h\rangle \leq \|A^*Ah\|\|h\| \leq \|A^*A\|_{\text{op}}\|h\|^2,$$

this implies $\|A\|_{\text{op}}^2 \leq \|A^*A\|_{\text{op}} \leq \|A^*\|_{\text{op}}\|A\|_{\text{op}}$, and further $\|A\|_{\text{op}} \leq \|A^*\|_{\text{op}}$. But since $(A^*)^* = A$, we also obtain $\|A^*\|_{\text{op}} \leq \|A\|_{\text{op}}$. Thus we have equalities everywhere. □

Remark 2.9. Concerning the point about invertibility in $\mathcal{B}(\mathcal{H}, \mathcal{H}')$: in fact, the fundamental open mapping theorem / bounded inverse theorem (Banach-Schauder theorem) states that if $A \in \mathcal{B}(\mathcal{H}, \mathcal{H}')$ is bijective, then its inverse is also bounded i.e. element of $\mathcal{B}(\mathcal{H}', \mathcal{H})$. (This is true more generally of bounded linear operators between Banach spaces; completeness is an essential assumption). So we can just assume that A is bijective there, it automatically ensures the "with bounded inverse" part.

Proposition/Definition 2.10 (Rank-one and finite rank operators)**.** Given two elements $u \in \mathcal{H}$, $v \in \mathcal{H}'$, we denote $v \otimes u^*$ the linear mapping

$$v \otimes u^* : x \in \mathcal{H} \mapsto \langle x, u\rangle v.$$

It is a rank one operator. It holds $(v \otimes u^*)^* = u \otimes v^*$ and $\|v \otimes u^*\|_{\text{op}} = \|u\|\|v\|$.
The (finite) linear combinations of rank one operators form the space of finite rank operators.

2.3 Compact operators on a Hilbert space and spectral theorems

Proposition/Definition 2.11. An operator $A \in \mathcal{B}(\mathcal{H}, \mathcal{H}')$ is compact if the image of the unit ball of $\mathcal{H}$ by A is relatively compact in $\mathcal{H}'$.
The set of compact operators from $\mathcal{H}$ to $\mathcal{H}'$ is a closed linear subspace of $\mathcal{B}(\mathcal{H}, \mathcal{H}')$ denoted $\mathcal{K}(\mathcal{H}, \mathcal{H}')$ (in the literature is is sometimes denoted $\mathcal{B}_0$ instead of $\mathcal{K}$).

For a proof of the closedness (implying completeness) of $\mathcal{K}(\mathcal{H}, \mathcal{H}')$, see [7, Prop 4.2]

Proposition 2.12. *If $A \in \mathcal{K}(\mathcal{H}, \mathcal{H}')$ and $B \in \mathcal{B}(\mathcal{H}', \mathcal{H}'')$ then $BA \in \mathcal{K}(\mathcal{H}, \mathcal{H}'')$.*
If $A \in \mathcal{K}(\mathcal{H}', \mathcal{H}'')$ and $B \in \mathcal{B}(\mathcal{H}, \mathcal{H}')$ then $AB \in \mathcal{K}(\mathcal{H}, \mathcal{H}'')$.
Thus $\mathcal{K}(\mathcal{H})$ is an ideal of $\mathcal{B}(\mathcal{H})$.

Proof: straightforward from the definition.

Proposition 2.13. *If $A \in \mathcal{K}(\mathcal{H}, \mathcal{H}')$, then $A^* \in \mathcal{K}(\mathcal{H}', \mathcal{H})$.*
Furthermore, $A \in \mathcal{K}(\mathcal{H}, \mathcal{H}')$ iff A is the limit of a sequence of finite-rank operators, converging in operator norm.

Proof: see [5, Thm. VI.4] or [7, Thm. II.4.4].

Compact operators behave very much like finite matrices, and that we will mostly deal with such operators (or with operators of the form $I + K$ with K compact, which are called "compact pertubations of identity").

The following result is the cornerstone of the theory of compact operators on a Hilbert space. Let us recall here that for an operator A, $\lambda \in \mathbb{R}$ or $\mathbb{C}$ is an *eigenvalue* of A if $A - \lambda I$ is not injective, i.e. $\ker(A - \lambda I) \neq \{0\}$. The *spectrum* of A is the set of values λ such that $A - \lambda I$ is not bijective. Thus an eigenvalue belongs to the spectrum, but the converse does not hold.

Theorem 2.14 (Spectral theorem for compact self-adjoint operators)**.** *Let $A \in \mathcal{K}(\mathcal{H})$ be a compact, self-adjoint operator.*
Then there exists a finite or countably infinite orthonormal family $(e_k)_{k\in I}$ of eigenvectors of A and family of corresponding real nonzero eigenvalues $(\lambda_k)_{k\in I}$ (here $I = [\![n]\!]$ for some $n \in \mathbb{N}$ or $I = \mathbb{N}$; possibly $I = \emptyset$ when $A = 0$) with $\lambda_k \to 0$ if $I = \mathbb{N}$, such that

$$A = \sum_{k\in I} \lambda_k e_k \otimes e_k^*, \tag{2.1}$$

that is to say:

$$\forall u \in \mathcal{H} \qquad Au = \sum_{k\in I} \lambda_k \langle u, e_k\rangle e_k, \tag{2.2}$$

and the series in (2.1) converges in operator norm. It can be assumed, if needed, that the sequence $(\lambda_k)_{k\in I}$ is ordered by decreasing absolute value (which we will just call "ordered" for short).
If $I = \mathbb{N}$, the set $\sigma(A) := \{\lambda_k, k \in I\} \cup \{0\} \subset \mathbb{R}$ is the spectrum of A and has 0 as only accumulation value. If I is finite, we define $\sigma(A) := \{\lambda_k, k \in I\} \cup \{0\}$ if 0 is an eigenvalue of A and $\sigma(A) := \{\lambda_k, k \in I\}$ otherwise (the latter case can only happen if $\mathcal{H}$ is finite-dimensional).
If we group indices k corresponding to the same nonzero eigenvalue λ and denote $P_\lambda := \sum_{k:\lambda_k=\lambda} e_k \otimes e_k^*$ (necessarily a finite sum since $\lambda \neq 0$) then P_λ is the orthogonal projector on the eigenspace associated to λ; when rewritten in the form

$$A = \sum_{\lambda\in\sigma(A)\setminus\{0\}} \lambda P_\lambda, \tag{2.3}$$

the decomposition is unique (we will call this the "canonical form" of the decomposition). Correspondingly every representation of the form (2.1) has the same ordered sequence $(\lambda_k)_{k\in I}$ (every nonzero eigenvalue of A is present in this sequence with its degree of multiplicity.)
If $\mathcal{H}$ is separable, we can complete $(e_k)_{k\in I}$ to a finite or countable Hilbert basis of $\mathcal{H}$. Defining $\lambda_\ell = 0$ for the added vectors of the completed basis, relations (2.1)-(2.2)-(2.3) still hold for this completed basis (for (2.3), the sum is then over $\sigma(A)$ and we include P_0, the orthogonal projector on the null space of A). We will call this the completed eigendecomposition of A.
This completion operation can also be made if $\mathcal{H}$ is nonseparable, but then we have to complete the orthonormal family $(e_k)_{k\in i}$ to an uncountable Hilbert basis.

For a proof, see e.g. [5, Section 6.4] or [7, Section II.5]. Note that there is a more general theory for the spectral decomposition of noncompact self-adjoint and even normal (commuting with their adjoint) operators, for which the sum is replaced by an integral in a suitable sense (see e.g. [7, Chapter IX]), be we will not consider it here.

Proposition 2.15. *If A is a compact, self-adjoint operator, it is positive if and only if its spectrum is nonnegative.*

Proof. Using the eigendecomposition (2.2) we see that $\lambda_k = \langle Ae_k, e_k\rangle \geq 0$ if A is positive. Conversely, if $u \in \mathcal{H}$ then $\langle Au, u\rangle = \sum_{k\in I} \lambda_k |\langle u, e_k\rangle|^2$ is nonnegative if $\lambda_k \geq 0$ for all $k \in I$. □

The following consequence is important:

Theorem 2.16 (Singular value decomposition of a compact operator)**.** *Let $A \in \mathcal{B}(\mathcal{H}, \mathcal{H}')$. If A is compact, then there exists a finite or countably infinite set I ($I = [\![n]\!]$ for some $n \in \mathbb{N}$ or $I = \mathbb{N}$) and:*

(1) a real positive sequence $(\sigma_k)_{k\in I}$, converging to 0 if $I = \mathbb{N}$;

(2) an orthonormal system $(e_k)_{k\in I}$ of $\mathcal{H}$;

(3) an orthonormal system $(f_k)_{k\in I}$ of $\mathcal{H}'$;

such that

$$A = \sum_{k\in I} \sigma_k f_k \otimes e_k^*, \tag{2.4}$$

that is to say:

$$\forall u \in \mathcal{H} \qquad Au = \sum_{k\in I} \sigma_k \langle u, e_k\rangle f_k, \tag{2.5}$$

and the series in (2.4) *converges in operator norm.*
The family $\{\sigma_k, k \in I\} \cup \{0\}$ is called set of singular values of A, denoted $\mathrm{sv}(A)$*, and it holds* $\mathrm{sv}(A) = \left\{\sqrt{\lambda}, \lambda \in \sigma(A^*A)\right\}$*, and* $\|A\|_{\mathrm{op}} = \max_k \sigma_k$*, and* $\mathrm{sv}(A^*) = \mathrm{sv}(A)$ *(with the same exception as before in the finite-dimensional case: 0 is not a singular value if A is bijective). Conversely, for any $(\sigma_k, e_k, f_k)_{k\in I}$ satisfying conditions (1)-(2)-(3), the series in* (2.4) *converges in operator norm and defines a compact operator.*

The fact that the theorem is "iff" shows that (2.4) is a complete characterization of compact operators between Hilbert spaces.

Proof. Assume that $A \in \mathcal{K}(\mathcal{H}, \mathcal{H}')$; hence $A^* \in \mathcal{K}(\mathcal{H}', \mathcal{H})$. The composition of compact operators is compact (Prop. 2.12), hence $A^*A \in \mathcal{K}(\mathcal{H})$ is compact and self-adjoint. We can apply Theorem 2.14 to find $(\lambda_k, e_k)_{k\in I}$ an eigendecomposition of A^*A. Obviously $A^*A \succeq 0$, so $\lambda_k > 0$ for all $k \in I$. Let us denote $\sigma_k := \sqrt{\lambda_k}$, and $f_k := \sigma_k^{-1} Ae_k$, $k \in I$.

Let us first prove that $(f_k)_{k\in I}$ is an orthonormal system of $\mathcal{H}'$:

$$\begin{aligned}\langle f_k, f_\ell\rangle &= \left\langle \sigma_k^{-1} A e_k, \sigma_\ell^{-1} A e_\ell \right\rangle \\ &= \sigma_k^{-1}\sigma_\ell^{-1}\langle e_k, A^* A e_\ell\rangle \\ &= \sigma_k^{-1}\sigma_\ell^{-1}\mathbf{1}\{k=\ell\}\sigma_k^2 \\ &= \mathbf{1}\{k=\ell\}.\end{aligned}$$

We now establish that $A = 0$ on the orthogonal of $\{e_k, k \in I\}$. Namely, if $u \in \mathcal{H}$ is orthogonal to all $e_k, k \in I$, then $A^*Au = 0$ by the eigendecomposition of A^*A, hence $\langle u, A^*Au\rangle = \|Au\|^2 = 0$. It follows that for any $v \in \mathcal{H}$:

$$Av = \sum_{k\in I}\langle v, e_k\rangle Ae_k = \sum_{k\in I}\sigma_k\langle v, e_k\rangle f_k.$$

This establishes (2.5) i.e. convergence in the weak sense.

We will now check that conversely, if $(\sigma_k, e_k, f_k)_{k\in I}$ satisfy (1)-(2)-(3), then the sum in (2.5) is well defined. Let $M^2 := \max_{k\in I}\sigma_k^2$. For any $u \in \mathcal{H}$, since $(e_k)_{k\in I}$ is an orthonormal system we have $\sum_{k\in I}|\langle u, e_k\rangle|^2 \le \|u\|^2$ (Bessel's inequality). Hence if $a_k := \sigma_k\langle u, e_k\rangle$ we have $|a_k|^2 \le M^2|\langle u, e_k\rangle|^2$, and further $\sum_{k\in I}|a_k|^2 \le M^2\|u\|^2$, hence the sum $\sum_{k\in I} a_k f_k$ is a well-defined element of $\mathcal{H}'$ since $(f_k)_{k\in I}$ is an orthonormal system. Linearity of this sum wrt. u is straightforward, and we also have as a byproduct that the resulting linear operator from $\mathcal{H}$ to $\mathcal{H}'$ has operator norm bounded by M. This also establishes strong convergence of (2.4) in operator norm, since by the same token, for any subset I' or I, it holds

$$\left\| \sum_{k\in I'} \sigma_k f_k \otimes e_k^* \right\|_{\mathrm{op}} \le \max_{k\in I'}\sigma_k,$$

and since $\sigma_k \to 0$ this implies the convergence in operator norm of (2.4).

Remark 2.17. It is worth noting that in contrast to the spectral theorem for self-adjoint operators, there is no such thing as a "completed" svd, which would tentatively include the singular value 0, and one would hope that the "input" and "output" orthonormal families $(e_k)_{k\in I}$ and $(f_k)_{k\in I}$ would both be bases. This is because the operator A could be surjective but not injective or vice-versa, so that one of the two families may be a basis but not the other, in which case it is impossible to extend the "deficient" family together with the one-to-one correspondence $e_k \leftrightarrow f_k$. Of course one can (for instance) complete the basis $(e_k)_{k\in I}$ by complementing it with a basis of the null space of A, but this cannot be written in the form (2.4) in general.

□

2.4 Functional calculus for compact self-adjoint operators

The following construction allows to apply a (bounded) real function to a compact self-adjoint operator:

Definition 2.18. Let $A \in \mathcal{K}(\mathcal{H})$ be a compact self-adjoint operator, and f be a bounded function $\sigma(A) \to \mathbb{R}$ or $\mathbb{C}$ (a real or complex function of real variable). Then if $(\lambda_k, e_k)_{k\in I}$ is a completed eigendecomposition of A, we define

$$f(A) := \sum_{k\in I} f(\lambda_k) e_k \otimes e_k^* = \sum_{\lambda\in\sigma(A)} f(\lambda) P_\lambda \in \mathcal{B}(\mathcal{H}), \tag{2.6}$$

where the above series converges for the weak topology, that is to say

$$\forall u \in \mathcal{H} \qquad f(A)u = \sum_{k\in I} f(\lambda_k)\langle u, e_k\rangle e_k = \sum_{\lambda\in\sigma(A)} f(\lambda) P_\lambda u. \tag{2.7}$$

Proof. We have to prove the claim that the series (2.7) converges, and that the operator defined this way is in $\mathcal{B}(\mathcal{H})$, so that this definition makes sense. The argument is the same as the one used in the proof of Thm. 2.16: for any $u \in \mathcal{H}$, we have $\sum_{k\in I} |\langle u, e_k\rangle|^2 \leq \|u\|^2$ (Bessel's inequality). Hence if $a_k := f(\sigma_k)\langle u, e_k\rangle$ we have $|a_k|^2 \leq M^2 |\langle u, e_k\rangle|^2$, where $M := \sup_{k\in I} |f(\sigma_k)|$ and further $\sum_{k\in I} |a_k|^2 \leq M^2 \|u\|^2$, hence the sum $\sum_{k\in I} a_k e_k$ is a well-defined element of $\mathcal{H}$. Furthermore it shows that the operator defined this way belongs to $\mathcal{B}(\mathcal{H})$, with operator norm bounded by M.
Finally, we check that the definition (2.6) is unique, because it can be rewritten as

$$f(A) = \sum_{\lambda\in\sigma(A)} f(\lambda) P_\lambda,$$

and the spectral decomposition of A in canonical form is unique. □

Note that we do **not** have convergence in operator norm in (2.6) general, since $(f(\sigma_k))_{k\in I}$ does not converge to 0 in general (in fact, Theorem 2.16 indicates that convergence in operator norm is equivalent to $(f(\sigma_k))_{k\in I}$ converges to 0 and $f(A)$ compact). For example, if f is the function identically equal to 1, $f(A)$ is the identity and the convergence is not in the strong sense.

We have the following properties:

Proposition 2.19. *Let $A \in \mathcal{K}(\mathcal{H})$ be a compact self-adjoint operator, and f, g be bounded functions $\sigma(A) \to \mathbb{R}$. Then:*

(a) For any $\lambda, \mu \in \mathbb{C}$: $(\lambda f + \mu g)(A) = \lambda f(A) + \mu g(A)$;

(b) $(fg)(A) = f(A)g(A)$, implying in particular $f(A)g(A) = g(A)f(A)$;

(c) $\|f(A)\|_{\text{op}} = \sup_{t\in\sigma(A)} |f(t)|$;

(d) If f is the constant function equal to 1, then $f(A) = \boldsymbol{I}$;

(e) If f is the identity function, then $f(A) = A$.

(f) If $f \geq 0$, then $f(A)$ is a positive operator.

The following proposition gives a useful trick:

Proposition 2.20 (Shift formula)**.** *Let $A \in \mathcal{K}(\mathcal{H}, \mathcal{H}')$ and $g : \{\lambda^2 | \lambda \in \mathrm{sv}(A)\} \to \mathbb{R}$ be a bounded function. Then it holds*

$$g(AA^*)A = Ag(A^*A) \text{ and } A^*g(AA^*) = g(A^*A)A^*. \tag{2.8}$$

Proof. Let $(\sigma_k, e_k, f_k)_{k\in I}$ be an SVD of A, i.e. $A = \sum_{k\in I} \sigma_k f_k \otimes e_k^*$. Then $A^* = \sum_{k\in I} \sigma_k e_k \otimes f_k^*$ and $AA^* = \sum_{k\in I} \sigma_k^2 f_k \otimes f_k^*$, which is an eigendecomposition of AA^*. Let us also denote P_0 the orthogonal projector onto the null space of AA^*.

$$g(AA^*)A = \left(\sum_{k\in I} g(\sigma_k^2) f_k \otimes f_k^* + g(0)P_0\right)\left(\sum_{\ell\in I} \sigma_\ell f_\ell \otimes e_\ell\right) = \sum_{k\in I} \sigma_k g(\sigma_k^2) f_k \otimes e_k^*,$$

it can be checked that $Ag(AA^*)$ leads to the the same formula. The other part of the claim is proved similarly. □

2.5 Hilbert-Schmidt operators

Source: [8, Chap.3]

Proposition/Definition 2.21. If $(e_k)_{k\in I}$ and $(f_k)_{k\in J}$ are Hilbert bases of $\mathcal{H}, \mathcal{H}'$ and $A \in \mathcal{B}(\mathcal{H}, \mathcal{H}')$, it holds

$$\sum_{k\in I} \|Ae_k\|^2 = \sum_{\ell\in J} \|A^* f_\ell\|^2 = \sum_{(k,\ell)\in I\times J} |\langle Ae_k, f_\ell\rangle|^2, \tag{2.9}$$

meaning that if any of the sums is convergent the other also are, and their value is independent of the choice of basis. If these sums are convergent, the operator A is called Hilbert-Schmidt operator.

Proof. By Parseval's identity we have $\|Ae_k\|^2 = \sum_{\ell\in J} |\langle Ae_k, f_\ell\rangle|^2 = \sum_{\ell\in J} |\langle e_k, A^* f_\ell\rangle|^2$. Summing over $k \in I$ and using Fubini's relation yields the two first equalities. □

Proposition/Definition 2.22. The set of Hilbert-Schmidt operators from $\mathcal{H}$ to $\mathcal{H}'$, denoted $\mathrm{HS}(\mathcal{H}, \mathcal{H}')$ (or sometimes $\mathcal{B}_2(\mathcal{H}, \mathcal{H}')$ in the literature), is a closed linear subspace of $\mathcal{K}(\mathcal{H}, \mathcal{H}')$, and a Hilbert space, once endowed with the Hilbertian product

$$\langle A, B\rangle_2 := \sum_k \langle Ae_k, Be_k\rangle = \sum_\ell \langle B^* f_\ell, A^* f_\ell\rangle = \sum_{(k,\ell)\in I\times J} \langle Ae_k, f_\ell\rangle \overline{\langle Be_k, f_\ell\rangle}, \tag{2.10}$$

where $(e_k)_{k\in I}, (f_\ell)_{\ell\in J}$ are any orthonormal bases of $\mathcal{H}, \mathcal{H}'$.
This definition does not depend on the choice of bases.
The associated Hilbert norm satisfies

$$\|A\|_{\mathrm{op}} \le \|A\|_2.$$

Proof. Let us fix $(e_k)_{k\in I}, (f_\ell)_{\ell\in J}$ orthonormal bases of $\mathcal{H}, \mathcal{H}'$. It is easy to check that $\mathrm{HS}(\mathcal{H},\mathcal{H}')$ is a vector space, using the definition and the fact that

$$|\langle (A+B)e_k, f_\ell\rangle|^2 \leq \big(|\langle Ae_k, f_\ell\rangle| + |\langle Be_k, f_\ell\rangle|\big)^2 \leq 2\big(|\langle Ae_k, f_\ell\rangle|^2 + |\langle Be_k, f_\ell\rangle|^2\big).$$

Similarly, the sums in (2.10) are absolutely convergent if A and B are Hilbert-Schmidt due to

$$\Big|\langle Ae_k, f_\ell\rangle\overline{\langle Be_k, f_\ell\rangle}\Big| \leq \frac{1}{2}\big(|\langle Ae_k, f_\ell\rangle|^2 + |\langle Be_k, f_\ell\rangle|^2\big)$$

for the last sum, and to the Cauchy-Schwarz inequality (to apply twice) for each of the two first sums.
It is straightforward to check that it is sesquilinear. Furthermore $\langle A, A\rangle_2 = \sum_{k,\ell}|\langle Ae_k, f_\ell\rangle|^2 = \sum_k \|Ae_k\|^2$ is 0 iff $A = 0$. It is thus a Hibertian product and induces a norm on $\mathrm{HS}(\mathcal{H})$. The sums involved in Formula (2.10) do not depend on the chosen basis (since one can change one the two bases while keeping the other one fixed).

If u is a unit vector, we can complete it to an orthonormal basis $(u_k)_{k\in I}$, and we have

$$\|Au\|^2 \leq \sum_{k\in I}\|Au_k\|^2 = \|A\|_2^2,$$

which implies $\|A\|_{\mathrm{op}} \leq \|A\|_2$ by taking the supremum in u.
From this, we deduce that any operator $A \in \mathrm{HS}(\mathcal{H},\mathcal{H}')$ can be arbitrarily approximated by a finite rank operator in HS-norm: for fixed $\varepsilon > 0$ given, let I_ε be a finite subset of I such that $\sum_{k\in I\setminus I_\varepsilon}\|Ae_k\|^2 \leq \varepsilon$. Define the operator $B_\varepsilon = AP_\varepsilon$, where P_ε is the orthogonal projector on $[e_k, k \in I_\varepsilon]$. Then B_ε is finite-rank and $\|B_\varepsilon - A\|_2^2 = \sum_{k\in I\setminus I_\varepsilon}\|Ae_k\|^2 \leq \varepsilon$.
Since the operator norm is dominated by the HS-norm, this proves in particular that any Hilbert-Schmidt operator can be arbitrarily approximated in operator norm by finite rank operators, and hence is compact.
It remains to justify that $\mathrm{HS}(\mathcal{H},\mathcal{H}')$ is complete for its norm. Because the Hilbert-Schmidt norm dominates the operator norm, a Cauchy sequence A_n in HS norm is Cauchy in operator norm and converges in operator norm towards some limit A_∞, since $\mathcal{B}(\mathcal{H})$ is complete. Thus $\|A_n e_k\|^2$ converges pointwise to $\|A_\infty e_k\|^2$ for every k. By Fatou's Lemma, $\sum_k\|A_\infty e_k\|^2 \leq \liminf_n \|A_n\|_2^2 < \infty$ since the sequence A_n is Cauchy in HS norm, hence bounded. This implies that A_∞ is Hilbert-Schmidt. Reiterating the same kind of argument, for any fixed n_0, $\|(A_{n_0} - A_n)e_k\|^2$ converges pointwise to $\|(A_{n_0} - A_\infty)e_k\|^2$ for every k. By Fatou's Lemma, $\sum_k\|(A_{n_0} - A_\infty)e_k\|^2 \leq \liminf_n \|A_{n_0} - A_n\|_2^2 \leq \sup_{n\geq n_0}\|A_{n_0} - A_n\|_2^2 = \varepsilon(n_0)$, and $\varepsilon(n_0) \to 0$ as $n_0 \to \infty$ by the Cauchy property; hence $\lim_{n\to\infty}\|A_n - A_\infty\|_2^2 = 0$. □

Proposition 2.23. *We have the following properties:*

1. *For any $u, v \in \mathcal{H}$ and $w, x \in \mathcal{H}'$, and $A \in \mathrm{HS}(\mathcal{H},\mathcal{H}')$, it holds*

$$\langle A, w\otimes u^*\rangle_2 = \langle Au, w\rangle;$$
$$\langle w\otimes u^*, x\otimes v^*\rangle_2 = \langle w, x\rangle\overline{\langle u, v\rangle}.$$

2. *If* $(e_k)_{k\in I}$ *and* $(f_k)_{k\in J}$ *are bases of* $\mathcal{H}, \mathcal{H}'$*, the family of rank-one operators* $(f_\ell \otimes e_k^*)_{(k,\ell)\in I\times J}$ *forms an orthonormal basis of* $\mathrm{HS}(\mathcal{H}, \mathcal{H}')$*.*

3. *If* A *is Hilbert-Schmidt, and* $(\sigma_k, e_k, f_k)_{k\in I}$ *is a svd of* A*, then*

$$\|A\|_2^2 = \sum_{k\in I} \sigma_k^2; \tag{2.11}$$

conversely if the sum in (2.11) *converges, then* A *is Hilbert-Schmidt.*

4. *If* $A \in \mathcal{B}_2(\mathcal{H}, \mathcal{H}')$ *and* $B \in \mathcal{B}(\mathcal{H}', \mathcal{H}'')$*, then* $BA \in \mathcal{B}(\mathcal{H}, \mathcal{H}'')$*. Similarly if* $A \in \mathcal{B}_2(\mathcal{H}', \mathcal{H}'')$ *and* $B \in \mathcal{B}(\mathcal{H}, \mathcal{H}')$*, then* $AB \in \mathcal{B}(\mathcal{H}, \mathcal{H}'')$*. (In particular* $\mathcal{B}_2(\mathcal{H})$ *is a left and right ideal of* $\mathcal{B}(\mathcal{H})$*.)*

Proof. For the first point, without loss of generality we can assume $\|w\| = \|u\| = 1$, and we can choose a Hilbert basis of $\mathcal{H}, \mathcal{H}'$ with u, resp. v as their first element. Using this basis to develop the product $\langle A, w \otimes u^* \rangle_2$ as per (2.10) yields the first point, which can be easily specialized to the case $A = x \otimes v^*$.
It results that if $(e_k)_{k\in I}$ and $(f_k)_{k\in J}$ are bases of $\mathcal{H}, \mathcal{H}'$, then
$\langle e_i \otimes f_j^*, e_k \otimes f_\ell^* \rangle_2 = \mathbf{1}\{i = j\}\mathbf{1}\{k = \ell\}$, so we have an orthonormal family. If it was not a basis, we could find a nonzero HS operator A in their orthogonal, but then by using the formula (2.10) and the previous point we would have $\|A\|_2 = 0$, a contradiction.
For the third point we can use the characterization (2.9) of the squared HS norm and apply it to the completion to a basis $(\widetilde{e}_k)_{k\in I}$ of the orthonormal family $(e_k)_{k\in I}$ entering into the svd $(\sigma_k, e_k, f_k)_{k\in I}$ of A.
Finally, for the last point we can use the defining property (2.9) and $\|BAe_k\| \le \|B\|_{\mathrm{op}} \|Ae_k\|$ (in the first case) and $\|(AB)^* e_k\| \le \|B^*\|_{\mathrm{op}} \|A^* e_k\|$ (in the second case). □

2.6 Schatten p-classes

Using functional calculus, if A is a compact, self-adjoint operator such that $A \succeq 0$, then the square root $A^{\frac{1}{2}}$ of A is well-defined (and satisfies the expected relation $(A^{\frac{1}{2}})^2 = A$).

Definition 2.24. If $A \in \mathcal{K}(\mathcal{H}, \mathcal{H}')$ is a compact operator, we define $|A| := (A^* A)^{\frac{1}{2}}$.

Proposition 2.25. *If* $A \in \mathcal{K}(\mathcal{H}, \mathcal{H}')$*, we have the following properties:*

- *It holds* $\mathrm{sv}(A) = \sigma(|A|)$*.*
- *If* $A = \sum_{k\in I} \sigma_k f_k \otimes e_k^*$ *is an svd of* A*, then*

$$|A| = \sum_{k\in I} \sigma_k e_k \otimes e_k^*.$$

- *For any* $u \in \mathcal{H}$*,*

$$\|Au\| = \||A|u\|.$$

- *With the notation of the previous item, if $W : \mathcal{H} \to \mathcal{H}'$ is the operator defined as $We_k = f_k, k \in I$ and $Wu = 0$ for $u \in [e_k, i \in I]^\perp$, then $A = W|A|$ and W is a partial isometry with the same null space as A (i.e. an isometry from the orthogonal of its null space onto its image). This is called the polar decomposition of A.*

Proof: left to the reader.

Remark: In the polar decomposition one cannot in general assume that W is an isometry. This is directly related to the fact that in a svd, one cannot in general assume that the "input" and "output" orthonormal families are bases, see Remark 2.17.

From now on, we will restrict our attention to endomorphisms of $\mathcal{H}$ i.e. elements of $\mathcal{B}(\mathcal{H})$ (instead of operators from $\mathcal{H}$ to $\mathcal{H}'$). Some of the definitions below can be extended to $\mathcal{B}(\mathcal{H}, \mathcal{H}')$ but some are specific to endomorphisms (for instance the notion of trace).

Proposition/Definition 2.26 (Schatten p-class)**.** Let $p \in [1, \infty)$. An operator $A \in \mathcal{K}(\mathcal{H})$ is said to in a Schatten p-class if either of the following equivalent properties is satisfied:

1. $|A|^{\frac{p}{2}} \in \mathrm{HS}(\mathcal{H})$.
2. $\sum_{k\in I} \sigma_k^p < \infty$, where $(\sigma_k)_{k\in I}$ are the singular values of A (with multiplicity).
3. There exists a basis $(e_k)_{k\in I}$ of $\mathcal{H}$ such that $\sum_{k\in I} \langle |A|^p e_k, e_k\rangle < \infty$.
4. For any basis $(e_k)_{k\in I}$ of $\mathcal{H}$, it holds $\sum_{k\in I} \langle |A|^p e_k, e_k\rangle < \infty$ and the value of this quantity is independent of the choice of basis.
5. It holds
$$\sup_{(e_k)_{k\in I},(f_k)_{k\in I}} \sum_k |\langle Ae_k, f_k\rangle|^p < \infty,$$
where the supremum is over $(e_k)_{k\in I}$ and $(f_k)_{k\in I}$ orthonormal systems of $\mathcal{H}$.

The Schatten p-class of operators is denoted $\mathcal{B}_p(\mathcal{H})$.
It is a closed linear subspace of $\mathcal{K}(\mathcal{H})$. The quantity appearing in the points 2-3-4-5 above is the same, denoted $\|A\|_p^p$. $\|.\|_p$ is a norm on $\mathcal{B}_p(\mathcal{H})$ called Schatten p-norm. If $A \in \mathcal{B}_p(\mathcal{H})$, then $A^* \in \mathcal{B}_p(\mathcal{H})$ with $\|A^*\|_p = \|A\|_p$, and it holds for any $q \geq p$ that $A \in \mathcal{B}_q(\mathcal{H})$ with $\|A\|_p \geq \|A\|_q \geq \|A\|_{\mathrm{op}}$; $\mathcal{B}_p(\mathcal{H})$ is complete for this norm.

Proof. The equivalence of points 1 to 4 (and equality of the quantities defined therein) is a direct consequence of Propositions 2.21, 2.22, 2.23 for Hilbert-Schmidt operators, and of 2.25, remarking that since $|A|$ is self-adjoint, $\left\||A|^{\frac{p}{2}} e\right\|^2 = \langle |A|^p e, e\rangle$.
Concerning point 5, note that 5 $\Rightarrow$ 2 by choosing the "input and output" orthonormal systems of a singular value decomposition of A, so that $|\langle Ae_k, f_k\rangle|^p = \sigma_k^p$ for all k.
We now show the converse. Let $(e_k)_{k\in I}, (f_k)_{k\in I}$ be the orthonormal systems of a singular value decomposition of A as above, and $(\widetilde{e}_k)_{k\in I}, (\widetilde{f}_k)_{k\in I}$ arbitrary orthonormal systems of $\mathcal{H}$. We have
$$\left|\left\langle A\widetilde{e}_k, \widetilde{f}_k\right\rangle\right| = \left|\sum_\ell \sigma_\ell \langle \widetilde{e}_k, e_\ell\rangle \langle f_\ell, \widetilde{f}_k\rangle\right| \leq \sum_\ell \sigma_\ell W_{k,\ell},$$

where

$$W_{k,\ell} := \Big|\langle e_\ell, \widetilde{e}_k\rangle\langle f_\ell, \widetilde{f}_k\rangle\Big|.$$

Observe that by the Cauchy-Schwarz inequality, followed by Bessel's, it holds

$$0 \le W_{k,\bullet} := \sum_\ell W_{k,\ell} \le \Big(\sum_\ell |\langle e_\ell, \widetilde{e}_k\rangle|^2\Big)^{\frac12}\Big(\sum_\ell \big|\langle f_\ell, \widetilde{f}_k\rangle\big|^2\Big)^{\frac12} \le \|\widetilde{e}_k\|\Big\|\widetilde{f}_k\Big\| = 1,$$

and similarly $0 \le W_{\bullet,\ell} := \sum_k W_{k,\ell} \le 1$. Therefore, by Jensen's inequality:

$$\begin{aligned}\sum_k \Big|\Big\langle A\widetilde{e}_k, \widetilde{f}_k\Big\rangle\Big|^p &\le \sum_k \Big(\sum_\ell \sigma_\ell W_{k,\ell}\Big)^p \\ &\le \sum_k \sum_\ell \sigma_\ell^p W_{k,\ell} W_{k,\bullet}^{p-1} \\ &\le \sum_\ell \sigma_\ell^p W_{\bullet,\ell} \\ &\le \sum_\ell \sigma_\ell^p.\end{aligned}$$

This argument also shows that the quantities appearing in points 2 and 5 are identical.
The variational characterization of point 5 allows us to establish the triangle inequality (note that the other characterizations are not so nice for this since they involve $|A|$, not A, and we are not sure what to do with $|A+B|$), namely for any bases $(e_k)_{k\in I}$, $(f_k)_{k\in I}$:

$$\begin{aligned}\Big(\sum_k |\langle A+Be_k, f_k\rangle|^p\Big)^{\frac1p} &\le \Big(\sum_k \Big(|\langle Ae_k, f_k\rangle| + |\langle Be_k, f_k\rangle|\Big)^p\Big)^{\frac1p} \\ &\le \Big(\sum_k |\langle Ae_k, f_k\rangle|^p\Big)^{\frac1p} + \Big(\sum_k |\langle Be_k, f_k\rangle|^p\Big)^{\frac1p} \\ &\le \|A\|_p + \|B\|_p.\end{aligned}$$

The announced norm inequalities follow directly from point 2, and the closedness/completeness property a similar argument as in the proof of Theorem 2.22, using the characterization of point 3 and Fatou's lemma. □

2.7 Trace-class operators

Proposition/Definition 2.27. For any $A \in \mathcal{B}_1(\mathcal{H})$ and any basis $(u_\ell)_{\ell\in I}$, the sum $\sum_{\ell\in I}\langle Au_\ell, u_\ell\rangle$ is a Hilbert sum (it converges absolutely) and its value is independent of the choice of basis.
This quantity is called trace of A and denoted $\mathrm{Tr}(A)$.
For this reason an operator in $\mathcal{B}_1(\mathcal{H})$ is also called "trace-class" and $\|A\|_1$ sometimes "trace norm" (note that $\|A\|_1 \neq \mathrm{Tr}(A)$ in general, though!)

Proof. As usual we start with a svd of A, $A = \sum_{k\in I} \sigma_k f_k \otimes e_k^*$. Then for any orthonormal basis $(u_\ell)_{\ell\in I}$, it holds

$$\begin{aligned}
\sum_{\ell\in I} |\langle Au_\ell, u_\ell\rangle| &= \sum_{\ell\in I} \left| \sum_{k\in I} \sigma_k \langle u_\ell, e_k\rangle\langle f_k, u_\ell\rangle \right| \\
&\leq \sum_{\ell\in I}\sum_{k\in I} \sigma_k |\langle u_\ell, e_k\rangle\langle f_k, u_\ell\rangle| \\
&= \sum_{k\in I} \sigma_k \sum_{\ell\in I} |\langle u_\ell, e_k\rangle\langle f_k, u_\ell\rangle| \\
&\leq \sum_{k\in I} \sigma_k \left(\sum_{\ell\in I} |\langle u_\ell, e_k\rangle|^2 \right)^{\frac{1}{2}} \left(\sum_{\ell\in I} |\langle f_k, u_\ell\rangle|^2 \right)^{\frac{1}{2}} \\
&= \sum_{k\in I} \sigma_k = \|A\|_1 < \infty,
\end{aligned}$$

thus all the sums involved in the above chain of inequalities converge (absolutely). Since the first sum converges absolutely, we can write

$$\begin{aligned}
\sum_{\ell\in I} \langle Au_\ell, u_\ell\rangle &= \sum_{\ell\in I}\sum_{k\in I} \sigma_k \langle u_\ell, e_k\rangle\langle f_k, u_\ell\rangle \\
&= \sum_{k\in I} \sigma_k \sum_{\ell\in I} \langle f_k, u_\ell\rangle \overline{\langle e_k, u_\ell\rangle} \\
&= \sum_{k\in I} \sigma_k \langle f_k, e_k\rangle;
\end{aligned}$$

hence the value of the sum is independent of the choice of basis. □

Proposition 2.28. *The following properties hold:*

1. *The trace is a linear functional on $\mathcal{B}_1(\mathcal{H})$.*
2. *If $A \in \mathcal{B}_1(\mathcal{H})$ then $|\mathrm{Tr}(A)| \leq \|A\|_1$.*
3. *If $A \in \mathcal{B}_1(\mathcal{H})$ then $\mathrm{Tr}(A^*) = \overline{\mathrm{Tr}(A)}$.*
4. *If $A \in \mathcal{B}_1(\mathcal{H})$ is self-adjoint, then $\mathrm{Tr}(A) = \sum_{k\in J} \lambda_k$, where $(\lambda_k)_{k\in J}$ are the eigenvalues of A (with multiplicity).*
5. *(Consequence of the previous point) if $A \in \mathcal{B}_1(\mathcal{H})$ is self-adjoint positive, then $\mathrm{Tr}(A) \geq 0$.*
6. *If $A \in \mathcal{B}_p(\mathcal{H})$ for $p \in [1, \infty)$, then $\|A\|_p = \mathrm{Tr}(|A|^p)^{\frac{1}{p}}$.*
7. *If A, B are Hilbert-Schmidt operators, then $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ and $\langle A, B\rangle_2 = \mathrm{Tr}(B^*A)$.*

8. *If $A \in \mathcal{B}_1(\mathcal{H})$ and $B \in \mathcal{B}(\mathcal{H})$ then AB and BA are both trace-class and* $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

Proof. We only prove the two last points, as the previous ones are straightforward (possibly reusing the explicit expression found in the proof of Proposition 2.27).
If A, B are Hilbert-Schmidt operators, we first establish that $C := AB$ is in $\mathcal{B}_1(\mathcal{H})$. Let $C = W|C|$ be the polar decomposition of C, this entails that $|C| = W^*C$, because W^*W is the identity on the range of $|C|$. Thus $|C| = (W^*A)B$ is a product of two Hilbert-Schmidt operators (remember from Prop. 2.23 that $A' := (W^*A)$ is Hilbert-Schmidt as soon as W is bounded and A Hilbert-Schmidt).
Now from (2.10) for any basis $(e_k)_{k\in I}$:

$$\sum_k \langle |AB| e_k, e_k\rangle = \sum_k \langle A'Be_k, e_k\rangle = \sum_k \langle Be_k, (A')^* e_k\rangle = \langle B, (A')^*\rangle_2,$$

we know that the right-hand side is a Hilbert sum, so the left-hand side too, implying $AB \in \mathcal{B}_1(\mathcal{H})$.

We can therefore write from (2.10) that for any basis $(e_k)_{k\in I}$:

$$\langle A, B\rangle_2 = \sum_k \langle Ae_k, Be_k\rangle = \sum_k \langle B^*Ae_k, e_k\rangle = \mathrm{Tr}(B^*A).$$

Using (2.10) again, we have

$$\mathrm{Tr}(AB) = \langle B, A^*\rangle_2 = \sum_{k,\ell} \langle Be_k, e_\ell\rangle \overline{\langle A^*e_k, e_\ell\rangle} = \sum_{k,\ell} \langle Be_k, e_\ell\rangle \langle Ae_\ell, e_k\rangle, \tag{2.12}$$

we first check that the double sum is absolutely convergent:

$$\sum_{k,\ell} |\langle Be_k, e_\ell\rangle \langle Ae_\ell, e_k\rangle| \leq \left(\sum_{k,\ell} |\langle Be_k, e_\ell\rangle|^2\right)^{\frac{1}{2}} \left(\sum_{k,\ell} |\langle Ae_\ell, e_k\rangle|^2\right)^{\frac{1}{2}} \leq \|A\|_2 \|B\|_2,$$

where we have used characterization (2.9) of the HS norm. Since the double sum in expression (2.12) is symmetrical in A and B, it holds $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.
If $A \in \mathcal{B}_1(\mathcal{H})$ and $T \in \mathcal{B}(\mathcal{H})$, we can we write A as a product $A = BC$ of two Hilbert-Schmidt operators B, C (this is clear from a svd of A.) Furthermore, since T is bounded, CT and TB are also Hilbert-Schmidt (use Definition 2.21 of a Hilbert-Schmidt operator, or Proposition 2.32 in the next section), therefore using the previous point

$$\mathrm{Tr}(AT) = \mathrm{Tr}(B(CT)) = \mathrm{Tr}((CT)B) = \mathrm{Tr}(C(TB)) = \mathrm{Tr}((TB)C) = \mathrm{Tr}(BA).$$

□

We end this section with a definition that will be useful later.

Definition 2.29. If $A \in \mathcal{B}_1(\mathcal{H})$, $A \neq 0$, we call *intrinsic dimension* of A the quantity

$$\mathrm{intdim}(A) := \frac{\|A\|_1}{\|A\|_{\mathrm{op}}}.$$

(We define also $\mathrm{intdim}(0) = 0$.)

The interpretation of this quantity is that is measures over how many dimensions the spectrum of A is "mainly concentrated". Here are a few properties to get some intuition on this quantity.

Proposition 2.30.

- *If A is finite-rank, $A \neq 0$, then* $1 \leq \mathrm{intdim}(A) \leq \mathrm{rank}(A)$.
- *If A is an orthogonal projector onto a finite-dimensional subspace E, then* $\mathrm{intdim}(A) = \dim(E)$.
- *If $\|A\|_{\mathrm{op}} = 1$ and the singular values of A satisfy $\sigma_k(A) \leq k^{-\alpha}$, $\alpha > 1$, then* $\mathrm{intdim}(A) \leq \frac{\alpha}{\alpha-1}$.

Proof. For the last point, use the sum-integral comparison

$$\mathrm{Tr}(A) \leq \sum_{k \geq 1} k^{-\alpha} \leq 1 + \int_1^\infty x^{-\alpha} dx = \frac{\alpha}{\alpha - 1}.$$

□

2.8 Some operator inequalities

We begin with a different variational characterization of the Schatten p-norm (compare carefully with point 5 of Prop. 2.26).

Proposition 2.31. *Let $A \in \mathcal{K}(\mathcal{H})$. If $p \in [2, \infty)$, then*

$$\|A\|_p^p = \sup_{(u_k)_{k \in J}} \sum_k \|Au_k\|^p = \sup_{(u_k)_{k \in J}, (v_\ell)_{\ell \in J}} \sum_{k,\ell} |\langle Au_k, v_\ell \rangle|^p;$$

if $p \in [1, 2]$, then

$$\|A\|_p^p = \inf_{(u_k)_{k \in J}} \sum_k \|Au_k\|^p = \inf_{(u_k)_{k \in J}, (v_\ell)_{\ell \in J}} \sum_{k,\ell} |\langle Au_k, v_\ell \rangle|^p;$$

in each case the sup or inf is over (orthonormal) bases of $\mathcal{H}$.

Proof. Let $(\sigma_k, e_k, f_k)_{k\in I}$ be a svd of A, and let $(\widetilde{e}_k)_{k\in J}$ be a completion of the family $(e_k)_{k\in I}$ to a basis. Note that $\sum_k \|A\widetilde{e}_k\|^p = \sum_k \sigma_k^p = \|A\|_p^p$. On the other hand, for any basis $(u_k)_{k\in J}$ of $\mathcal{H}$, put $W_{k,\ell} := \langle u_k, e_\ell\rangle$ and $S_k := \sum_{k\in I}|W_{\ell,k}|^2 \leq \|u_\ell\|^2 = 1$, it holds:

$$\begin{aligned}
\sum_k \|Au_k\|^p &= \sum_k \Big(\Big\|\sum_\ell \sigma_\ell W_{k,\ell} f_\ell\Big\|^2\Big)^{\frac{p}{2}} \\
&= \sum_k \Big(\sum_\ell \sigma_\ell^2 |W_{k,\ell}|^2\Big)^{\frac{p}{2}} \\
&\underset{\geqq}{\leqq} \sum_{k,\ell} \sigma_\ell^p |W_{k,\ell}|^2 S_k^{\frac{p}{2}-1} \\
&\underset{\geqq}{\leqq} \sum_{k,\ell} \sigma_\ell^p |W_{k,\ell}|^2 \\
&= \sum_\ell \sigma_\ell^p,
\end{aligned}$$

where we have used Jensen's inequality for the convex (if $p \geq 2$) resp. concave (if $1 \leq p \leq 2$) function $x \mapsto x^{\frac{p}{2}}$ (with the inequality in different directions according to the case; it is an equality if $p = 2$); then $\sum_k |W_{k,\ell}|^2 = \|e_\ell\|^2 = 1$, since (u_k) is a basis. This establishes the first equality in both cases $p\underset{\geqq}{\leqq}2$.

Furthermore, for any bases $(u_k)_{k\in J}$, $(v_\ell)_{\ell\in J}$ of $\mathcal{H}$, it holds

$$\sum_k \|Au_k\|^p = \sum_k \Big(\sum_\ell |\langle Au_k, v_\ell\rangle|^2\Big)^{\frac{p}{2}} \underset{\leqq}{\geqq} \sum_{k,\ell} |\langle Au_k, v_\ell\rangle|^p,$$

where we have super- (if $p \geq 2$) resp sub-additivity (if $p \leq 2$) of the function $x \mapsto x^{\frac{p}{2}}$ (note that this is in the opposite direction as the previous display.) Again, if we take the input/output orthonormal families of the svd of A, both completed to a basis, we find $\|A\|_p^p$. □

Proposition 2.32. *If $A \in \mathcal{B}(\mathcal{H})$ and $B \in \mathcal{B}_p(\mathcal{H})$ ($p \in [1,\infty]$), then $AB \in \mathcal{B}_p(\mathcal{H})$ and $\|AB\|_p \leq \|A\|_{\mathrm{op}}\|B\|_p$.*

Similarly, $BA \in \mathcal{B}_p(\mathcal{H})$ and $\|BA\|_p \leq \|A\|_{\mathrm{op}}\|B\|_p$.

(So $\mathcal{B}_p(\mathcal{H})$ is an ideal of $\mathcal{B}(\mathcal{H})$.)

Proof. We assume $p < \infty$ as the case $p = \infty$ (i.e. $\|.\|_\infty = \|.\|_{\mathrm{op}}$) was handled before.
It holds for any orthonormal basis $(e_k)_{k\geq 1}$ of $\mathcal{H}$:

$$\sum_k \|ABe_k\|^p \leq \sum_k (\|A\|_{\mathrm{op}}\|Be_k\|)^p = \|A\|_{\mathrm{op}}^p \sum_k \|Be_k\|^p.$$

We now use the variational characterization of Proposition 2.31: if $p \leq 2$ we take an infimum over the basis on the left, then the right-hand side; if $p \geq 2$ we take a supremum over the right, then the left-hand side. In all cases we conclude to $AB \in \mathcal{B}_p(\mathcal{H})$ and

$\|AB\|_p \leq \|A\|_{\mathrm{op}}\|B\|_p$.
For the operator BA: we have

$$\|BA\|_p = \|A^*B^*\|_p \leq \|A^*\|_{\mathrm{op}}\|B^*\|_p = \|A\|_{\mathrm{op}}\|B\|_p;$$

note that the equality $\|A\|_p = \|A^*\|_p$ comes from $\mathrm{sv}(A) = \mathrm{sv}(A^*)$. □

Proposition 2.33 (Hölder's inequality for operators)**.** *Let $A \in \mathcal{B}_p(\mathcal{H})$ and $B \in \mathcal{B}_q(\mathcal{H})$ with $p^{-1} + q^{-1} = 1$. Then $AB \in \mathcal{B}_1(\mathcal{H})$, $BA \in \mathcal{B}_1(\mathcal{H})$, $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ and*

$$|\mathrm{Tr}(AB)| \leq \|A\|_p\|B\|_q. \tag{2.13}$$

Proof. As in the proof of Prop. 2.26, we first justify that $C = AB$ is trace-class. We start similarly by using the polar decomposition $C = W|C|$, implying $|C| = W^*C = A'B$, where $A' := W^*A \in \mathcal{B}_p(\mathcal{H})$, by Prop. 2.32. Furthermore, consider an svd of B, $B = \sum_k \sigma_k v_k \otimes u_k^*$. Thus, we can write

$$\begin{aligned}\left|\sum_k \langle |C|u_k, u_k\rangle\right| = \left|\sum_k \langle A'Bu_k, u_k\rangle\right| &= \left|\sum_k \sigma_k \langle A'v_k, u_k\rangle\right| \\ &\leq \left(\sum_k \sigma_k^q\right)^{\frac{1}{q}} \left(\sum_k |\langle A'v_k, u_k\rangle|^p\right)^{\frac{1}{p}} \\ &\leq \|B\|_q\|A'\|_p \leq \|B\|_q\|A\|_p\,,\end{aligned}$$

where we have used the (standard) Hölder's inequality for the first inequality, and point 5 of Proposition 2.26 for the second. This establishes $AB \in \mathcal{B}_1(\mathcal{H})$; we can now repeat the above computation with $C = AB$ instead of $|C| = A'B$ with the same final estimate, thus establishing (2.13).

Proving that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ in that context is annoying. If $A = \sum_k \mu_k f_k \otimes e_k^*$ is an svd of A, we can start as above, writing

$$\sum_k \langle ABu_k, u_k\rangle = \sum_k \sigma_k \langle Av_k, u_k\rangle = \sum_k \sum_\ell \sigma_k \mu_\ell \langle v_k, e_\ell\rangle \langle f_\ell, u_k\rangle. \tag{2.14}$$

It seems that the obtained expression is symmetric in the role of A, B and that we are done? Unfortunately, for this argument to be correct we have to establish that the double sum over k, ℓ is absolutely convergent (we know that for any fixed k, the sum over ℓ is absolutely convergent; that is not enough to establish that the double sum is.) Let us denote $W_{k,\ell} := |\langle v_k, e_\ell\rangle|$ and $W'_{k,\ell} := |\langle f_\ell, u_k\rangle|$. Assume $1 < p \leq 2 \leq q < \infty$. We want to

establish the convergence of

$$\begin{aligned}\sum_{k,\ell}\sigma_k\mu_\ell|\langle v_k,e_\ell\rangle||\langle f_\ell,u_k\rangle| &= \sum_{k,\ell}\sigma_k\mu_\ell W_{k,\ell}W'_{k,\ell}\\ &= \sum_{k,\ell}(\sigma_k W_{k,\ell}^{\frac{2}{q}})(\mu_\ell W_{k,\ell}^{1-\frac{2}{q}}W'_{k,\ell})\\ &\leq \left(\sum_{k,\ell}\sigma_k^q W_{k,\ell}^2\right)^{\frac{1}{q}}\left(\sum_{k,\ell}\mu_\ell^p W_{k,\ell}^{p(1-\frac{2}{q})}(W'_{k,\ell})^p\right)^{\frac{1}{p}}\\ &= \left(\sum_k\sigma_k^q\right)^{\frac{1}{q}}\left(\sum_\ell\mu_\ell^p\sum_k W_{k,\ell}^{p(1-\frac{2}{q})}(W'_{k,\ell})^p\right)^{\frac{1}{p}},\end{aligned}$$

where we have used Hölder's inequality, then $\sum_k W_{k,\ell}^2 \leq 1$. Finally, applying Hölder's inequality again:

$$\begin{aligned}\sum_k W_{k,\ell}^{p(1-\frac{2}{q})}(W'_{k,\ell})^p &\leq \left(\sum_k W_{k,\ell}^{p\frac{(1-\frac{2}{q})}{(1-\frac{p}{2})}}\right)^{1-\frac{p}{2}}\left(\sum_k(W'_{k,\ell})^2\right)^{\frac{p}{2}}\\ &= \left(\sum_k W_{k,\ell}^2\right)^{1-\frac{p}{2}}\left(\sum_k(W'_{k,\ell})^2\right)^{\frac{p}{2}} \leq 1.\end{aligned}$$

Thus (2.14) is absolutely convergent (we ended up also re-proving the trace-Hölder inequalities established before, in a more complicated way...) □

Proposition 2.34. *Let A, B be two self-adjoint Hilbert-Schmidt operators and $f : \mathbb{R} \to \mathbb{R}$ an L-Lipschitz function.*

Then $(f(A) - f(B))$ is Hilbert-Schmidt and it holds

$$\|f(A) - f(B)\|_2 \leq L\|A - B\|_2. \tag{2.15}$$

Important remark: one can wonder if (2.15) holds for other norms. It is *not* the case. In particular, it does **not** hold in general for the operator norm $\|.\|_{\text{op}}$: functions satisfying (2.15) for the operator norm are called "operator Lipschitz", and not every real-valued Lipschitz function is operator Lipschitz, even with a different constant.

Still, it is true that an inequality of this type holds (up to multiplicative constant C_p) for then Schatten p-norm with $p \in (1, \infty)$. This fact in this degree of generality had long been an open question and has been established only "recently" [25].

Proof. Let $(e_k, \lambda_k)_{k\geq 1}$ and $(f_\ell, \mu_\ell)_{\ell\geq 1}$ be eigendecompositions of A and B, respectively.

Observe that in general, for an operator $M \in \text{HS}(\mathcal{H})$, since both $(e_k)_{k\geq 1}$ and $(f_\ell)_{\ell\geq 1}$ are Hilbert bases it holds

$$\|M\|_2^2 = \sum_k\|Me_k\|^2 = \sum_{k,\ell}|\langle Me_k, f_\ell\rangle|^2.$$

We apply this formula to $M = A - B$ to get

$$\begin{aligned}\|A - B\|_2^2 &= \sum_{k,\ell} |\langle (A-B)e_k, f_\ell\rangle|^2 \\ &= \sum_{k,\ell} |\langle Ae_k, f_\ell\rangle - \langle e_k, Bf_\ell\rangle|^2, \\ &= \sum_{k,\ell} |\lambda_k - \mu_\ell|^2 |\langle e_k, f_\ell\rangle|^2,\end{aligned}$$

where we have used that B is self-adjoint in the second equality.
By the same token,

$$\|f(A) - f(B)\|_2^2 = \sum_{k,\ell} |f(\lambda_k) - f(\mu_\ell)|^2 |\langle e_k, f_\ell\rangle|^2.$$

It is now obvious that we can use the Lipschitz property of f for each (k, ℓ) term to reach the claim. □

Proposition 2.35 (Cordes' inequality, [9], p.24)**.** *If $A, B \in \mathcal{B}(\mathcal{H})$ are positive self-adjoint operators and $s \in [0, 1]$, then*

$$\|A^s B^s\|_{\text{op}} \leq \|AB\|_{\text{op}}^s.$$

Short proof: see [11]. For an extension to unitary equivariant norms see [16].

3 Tools from concentration of measure

3.1 Random variables in Banach space

Proposition/Definition 3.1. Let $\mathcal{B}$ be a separable Banach space, with its Borel σ-algebra. A random variable X from a base probability space $(\Omega, \mathcal{F}, P)$ to $\mathcal{B}$ is Bochner integrable if $\mathbb{E}[\|X\|] < \infty$. In this case there is a well-defined expectation $\mathbb{E}[X] \in \mathcal{B}$ satisfying the following properties:

- $\|\mathbb{E}[X]\| \leq \mathbb{E}[\|X\|]$;
- Simple linearity: if X, Y are Bochner-integrable then $\mathbb{E}[\lambda X + Y] = \lambda \mathbb{E}[X] + \mathbb{E}[Y]$;
- Operator linearity: for any bounded linear operator A from $\mathcal{B}$ to a separable Banach space $\mathcal{B}'$ it holds that AX is Bochner-integrable in $\mathcal{B}'$ and

$$\mathbb{E}[AX] = A\mathbb{E}[X].$$

Proposition 3.2 (Positivity of expectation for operators)**.** *Let A be a Bochner-integrable random variable taking values in $\mathcal{B}(\mathcal{H})$ and such that A is a.s. self-adjoint positive. Then $\mathbb{E}[A]$ is self-adjoint positive.*

Proof. Since $A \mapsto A^*$ is a bounded linear operator on $\mathcal{B}(\mathcal{H})$, A^* is Bochner integrable as soon as A is, and it holds $\mathbb{E}[A]^* = \mathbb{E}[A^*] = \mathbb{E}[A]$ when A is a.s. self-adjoint, thus $\mathbb{E}[A]$ is self-adjoint. Furthermore, if A is a.s. positive, then for any $u \in \mathcal{H}$:

$$\langle u, \mathbb{E}[A]u\rangle = \mathbb{E}[\langle u, Au\rangle] \geq 0,$$

hence $\mathbb{E}[A]$ is positive. □

Proposition/Definition 3.3. Let X be random variable taking values in a separable Hilbert space $\mathcal{H}$, and assume $\mathbb{E}\left[\|X\|^2\right] < \infty$. Then $X \otimes X^*$ is Bochner-integrable in $\mathcal{B}(\mathcal{H})$ and we call $\Sigma = \mathbb{E}[X \otimes X^*]$ the second moment operator of X. It satisfies for all u, v in $\mathcal{H}$:

$$\mathbb{E}[\langle v, X\rangle\langle X, u\rangle] = \langle v, \Sigma u\rangle.$$

Proof. Recall that $\|u \otimes v^*\|_{\text{op}} = \|u\|\|v\|$. Thus $\|X \otimes X^*\|_{\text{op}} = \|X\|^2$ is integrable by assumption, therefore $X \otimes X^*$ is Bochner-integrable in the Banach space $\mathcal{B}(\mathcal{H})$.
Therefore, by operator linearity of the Bochner integral, it holds for any $u \in \mathcal{H}$ that $X \otimes X^* u = \langle u, X\rangle X$ is Bochner-integrable with $\mathbb{E}[\langle u, X\rangle X] = \mathbb{E}[(X \otimes X^*)u] = \Sigma u$, and further for any $v \in \mathcal{H}$ that $\langle X, u\rangle\langle v, X\rangle = \langle v, \langle u, X\rangle X\rangle$ so that

$$\mathbb{E}[\langle X, u\rangle\langle v, X\rangle] = \mathbb{E}[\langle v, \langle u, X\rangle X\rangle] = \langle v, \mathbb{E}[\langle u, X\rangle X]\rangle = \langle v, \Sigma u\rangle.$$

□

For Bochner integrable random variables, there is a well-defined notion of conditional expectation as in the real case and it satisfies the above properties as well.

We conclude this short overview with a useful property.

Proposition 3.4. *Let X, X' be two Bochner-integrable random variables taking values in a separable Hilbert space $\mathcal{H}$. If X, X' are independent, then $\langle X, X'\rangle$ is integrable and $\mathbb{E}[\langle X, X'\rangle] = \langle \mathbb{E}[X], \mathbb{E}[X']\rangle$.*

Proof. By assumption, the real-valued random variables $\|X\|$, $\|X'\|$ are integrable, and independent, so that their product is integrable. Since $|\langle X, X'\rangle| \leq \|X\|\|X'\|$, the variable $\langle X, X'\rangle$ is integrable.

To establish the independent scalar product formula, one possible approach is to use conditional expectations $\mathbb{E}[\langle X, X'\rangle | X']$ and apply the operator linearity property to it with the X'-measurable operator $A : v \mapsto \langle v, X'\rangle$, so that $\mathbb{E}[\langle X, X'\rangle | X'] = \langle \mathbb{E}[X], X'\rangle$ a.s.

Here is an argument without conditional expectations. Take a basis $(e_k)_{k\in I}$ of $\mathcal{H}$ and write $\langle X, X'\rangle = \sum_{k\in I} \langle X, e_k\rangle\overline{\langle X', e_k\rangle}$. Denote $Z_k = \langle X, e_k\rangle\overline{\langle X', e_k\rangle}$. Note that $(\omega, k) \in \Omega \times I \mapsto Z_k(\omega)$ is integrable on the product space $\Omega \times I$ with respect to $P \otimes \mu$, where μ is the counting measure on I; this follows from Tonelli's theorem, since $|Z_k| = |\langle X, e_k\rangle||\langle X', e_k\rangle|$ and $\sum_{k\in I} |Z_k| \leq \|X\|\|X'\|$, which is integrable (see above).

Furthermore, by the independent product formula in the complex case, it holds $\mathbb{E}[Z_k] = \mathbb{E}[\langle X, e_k\rangle]\mathbb{E}\Big[\overline{\langle X', e_k\rangle}\Big] = \langle \mathbb{E}[X], e_k\rangle\overline{\langle \mathbb{E}[X'], e_k\rangle}$ for all $k \in I$. As $(k, \omega) \to Z_k(\omega)$ is integrable on the product space, we can apply Fubini's theorem, obtaining

$$\mathbb{E}[\langle X, X'\rangle] = \mathbb{E}\left[\sum_{k\in I} Z_k\right] = \sum_{k\in I} \mathbb{E}[Z_k] = \sum_{k\in I} \langle \mathbb{E}[X], e_k\rangle\overline{\langle \mathbb{E}[X'], e_k\rangle} = \langle \mathbb{E}[X], \mathbb{E}[X']\rangle.$$

Note that we have used here that I is at most countable (since $\mathcal{H}$ is separable), because the Fubini-Tonelli theorem only holds for σ-finite measures. □

3.2 Hoeffding's inequality in Hilbert space

We first recall the Azuma-McDiarmid concentration theorem for "stable" functions of independent random variables, also called *Bounded difference inequality.*

Theorem 3.5 (Azuma-McDiarmid)**.** *Let $\mathcal{X}$ be a measurable space, and $f : \mathcal{X}^n \to \mathbb{R}$ a measurable function such that*

$$\forall i \in \{1, \ldots, n\},\ \ \forall (x_1, \ldots, x_n) \in \mathcal{X}^n,\ \ \forall x_i' \in \mathcal{X} : \\ |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq 2c_i, \quad \text{(Stab)}$$

for some positive constants $(c_1, \ldots, c_n)$.

Let $(X_1, \dots, X_n)$ be a independent family of random variables taking values in $\mathcal{X}$ (not necessarily identically distributed), then $f(X_1, \dots, X_n)$ is a sub-Gaussian variable with parameter $\sum_{i=1}^n c_i^2$, so that in particular

$$\mathbb{P}[f(X_1, \dots, X_n) > \mathbb{E}[f(X_1, \dots, X_n)] + t] \leq \exp - \left(\frac{t^2}{2 \sum_{i=1}^n c_i^2} \right). \tag{3.1}$$

(In particular, if all constants c_i are equal to c, the bound is $\exp(-t^2/(2nc^2))$.)

For a proof, see e.g. [17, Section 3.4] or [2, Section 6.2] or [4, Section 6.1].

Theorem 3.6 (Vectorial Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be i.i.d., Bochner-integrable random variables taking values in a Hilbert space $\mathcal{H}$, and having expectation 0.*
Assume $\|X_i\| \leq B$ a.s. for some constant B.
Then if $S_n := \sum_{i=1}^n X_i$, it holds for any $\delta \in (0, 1)$:

$$\mathbb{P}\left[\left\| \frac{1}{n} S_n \right\| \geq \frac{B}{\sqrt{n}} \left(1 + \sqrt{2 \log(\delta^{-1})} \right) \right] \leq \delta.$$

Proof. By the assumption $\|X_i\| \leq B$, we can assume that the variables X_i in fact take their values in the ball of $\mathcal{H}$ centered at the origin and of radius B. As a first step, we note that the function $F(X_1, \dots, X_n) := \left\| \frac{1}{n} S_n \right\|$ satisfies the condition (Stab) (with $c_i = B$ for all i) on this ball, by the triangle inequality: namely, if we replace X_i by X_i' in the sum S_n, denoting it $S_n^{(i)}$, it holds

$$\left| \left\| \frac{1}{n} S_n \right\| - \left\| \frac{1}{n} S_n^{(i)} \right\| \right| \leq \frac{1}{n} \left\| S_n - S_n^{(i)} \right\| = \frac{1}{n} \|X_i - X_i'\| \leq 2B.$$

Applying the Azuma-McDiarmid inequality, we get

$$\mathbb{P}\left[\left\| \frac{1}{n} S_n \right\| > \frac{1}{n} \mathbb{E}[\|S_n\|] + t \right] \leq \exp\left(-\frac{nt^2}{2B^2} \right). \tag{3.2}$$

In a Hilbert space, we have moreover due to Jensen's inequality:

$$\mathbb{E}[\|S_n\|] \leq \mathbb{E}\left[\|S_n\|^2 \right]^{\frac{1}{2}} \leq \mathbb{E}\left[\sum_{i,j=1}^n \langle X_i, X_j \rangle \right]^{\frac{1}{2}} = \sqrt{n} \mathbb{E}\left[\|X_1\|^2 \right]^{\frac{1}{2}} \leq B\sqrt{n},$$

since $\mathbb{E}[\langle X_i, X_j \rangle] = 0$ if $i \neq j$, using independence and $\mathbb{E}[X_i] = 0$. Combining this with (3.2) and taking $t = B\sqrt{2 \log(\delta^{-1})/n}$ yields the claim. □

3.2.1 (*) Extension to Banach space

Observe that equation (3.2) also holds more generally in a Banach space, since we have only used the triangle inequality for the bounded difference concentration inequality. Only the upper bound on $\mathbb{E}[\|S_n\|]$ used the Hilbertian structure. For Banach spaces, a convenient notion is that of *type*:

Definition 3.7. A Banach space $\mathcal{B}$ is said to be of (Rademacher) type $p \in [1,2]$ if there exists a constant $C > 0$ such that

$$\mathbb{E}_\varepsilon \left[\left\| \sum_{i=1}^n \varepsilon_i x_i \right\|^p \right] \leq C^p \sum_{i=1}^n \|x_i\|^p,$$

for all finite sequences $(x_1, \dots, x_n)$ of elements of $\mathcal{B}$, where $\varepsilon_1, \varepsilon_2, \dots$ is an infinite sequence of i.i.d. Rademacher random variables (random signs).
The best constant C so that the above holds is denoted $T_p(\mathcal{B})$.

Lemma 3.8. *Let $X_1, \dots, X_n$ be i.i.d., Bochner-integrable random variables taking values in a Banach space $\mathcal{B}$, and having expectation 0.*
Assume that $\mathcal{B}$ is a Banach space of type $p \in [1,2]$. Then if $S_n := \sum_{i=1}^n X_i$, it holds

$$\mathbb{E}[\|S_n\|] \leq 2T_p(\mathcal{B}) n^{\frac{1}{p}} \mathbb{E}[\|X_1\|^p]^{\frac{1}{p}}.$$

Proof. Denote $(X_1', \dots, X_n')$ an independent copy of $(X_1, \dots, X_n)$. Put $C = T_p(\mathcal{B})$. Then we have

$$\begin{aligned}
\mathbb{E}[\|S_n\|] &= \mathbb{E}\left[\left\| \sum_{i=1}^n (X_i - \mathbb{E}[X_i']) \right\| \right] \\
&\leq \mathbb{E}\left[\left\| \sum_{i=1}^n (X_i - X_i') \right\| \right] && (3.3) \\
&= \mathbb{E}\left[\left\| \sum_{i=1}^n \varepsilon_i (X_i - X_i') \right\| \right] && (3.4) \\
&\leq 2\mathbb{E}\left[\left\| \sum_{i=1}^n \varepsilon_i X_i \right\| \right] && (3.5) \\
&\leq 2\mathbb{E}\left[\left\| \sum_{i=1}^n \varepsilon_i X_i \right\|^p \right]^{\frac{1}{p}} && (3.6) \\
&\leq 2C \left(\sum_{i=1}^n \mathbb{E}[\|X_i\|^p] \right)^{\frac{1}{p}} && (3.7) \\
&= 2Cn^{\frac{1}{p}} \mathbb{E}[\|X_1\|^p]^{\frac{1}{p}}, && (3.8)
\end{aligned}$$

where (3.4) is due to invariance of the distribution of the vector $(X_i - X_i')_{1 \leq i \leq n}$ by sign-flipping, (3.5) is the triangle inequality, (3.6) is Jensen's inequality, and (3.7) is the definition of Rademacher type p. Concerning (3.3), we use the first property of Proposition 3.1 for the conditional expectation $\mathbb{E}[.|X_i]$. □

Once we plug this estimate into (3.2), we see that for Banach spaces of type $p < 2$, the situation is radically different than in Hilbert spaces, because the above bound on the

expectation of $\|n^{-1}S_n\|$ is of higher order $O(n^{-(1-1/p)})$ than the deviation term of order $O(n^{-\frac{1}{2}})$.

Still, he have the following facts:

- an L^p space for $p \in [1, \infty)$ is of type $\min(p, 2)$;
- for $p \in [1, \infty)$, the Schatten p-class $S_p(\mathcal{H})$ of operators is of the same type as L^p [31].

Hence, we have (as far as the order in n is concerned) a control of the same order in n (up to constants) as in a Hilbert space for L^p spaces and S_p spaces for $2 \leq p < \infty$. Unfortunately, $p = \infty$ is radically different as L^∞ spaces are of type 1. Note that the above arguments are useless in Banach spaces of type 1 (it does not give a better result than the triangle inequality; any Banach space is at least of type 1).

3.3 Bernstein's inequality in smooth Banach space

Source: [24].

Lemma 3.9. *Let $(X_i)_{i\in[\![n]\!]}$ be a sequence of Bochner integrable, independent random variables with values in a separable Banach space $\mathcal{B}$ with $\mathbb{E}[X_i] = 0$ for all i.*
Let $F : \mathcal{B} \to \mathbb{R}$ be a function with the following properties:

a) For all $x, u \in \mathcal{B}$, $u \neq 0$, the function $F_{x,u} : t \in [0, 1] \mapsto F(x + tu)$ is twice differentiable.

b) For all $x \in \mathcal{B}$, there exists an element $dF_x \in \mathcal{B}'$ such that $(F'_{x,u})_{|t=0} = dF_x(u)$ for all $u \neq 0$.

c) For all $x, u \in \mathcal{B}$, $u \neq 0$, it holds $F''_{x,u}(t) \leq H(u, t)F(x)$ for some function $H(u, t) \geq 0$.

Then denoting $S_n := \sum_{k=1}^n X_n$, it holds (a.s.)

$$\mathbb{E}_n[F(S_n)] \leq F(S_{n-1})\left(1 + \mathbb{E}\left[\int_0^1 H(X_n, t)(1 - t)dt\right]\right),$$

where $\mathbb{E}_n$ denotes expectation conditionally to $(X_1, \ldots, X_{n-1})$.

Proof. Using the assumptions and Taylor's expansion with integral remainder we get

$$\begin{aligned} F(S_n) = F(S_{n-1} + X_n) &= F_{S_{n-1},X_n}(1) \\ &= F(S_{n-1}) + dF_{S_{n-1}}(X_n) + \int_0^1 (1 - t)F''_{S_{n-1},X_n}dt \\ &\leq F(S_{n-1}) + dF_{S_{n-1}}(X_n) + F(S_{n-1})\int_0^1 (1 - t)H(X_n, t)dt. \end{aligned}$$

We now take expectation $\mathbb{E}_n$ with respect to X_n conditionally to S_{n-1}, due to independence it holds

$$\mathbb{E}_n\big[dF_{S_{n-1}}(X_n)\big] = dF_{S_{n-1}}(\mathbb{E}_n[X_n]) = 0,$$

and we obtain the claim. □

Theorem 3.10 (Pinelis)**.** *Let $(X_i)_{i\in[\![n]\!]}$ be a sequence of Bochner integrable, independent random variables with values in a separable Banach space $\mathcal{B}$ with $\mathbb{E}[X_i] = 0$ for all i. Denote $S_n = \sum_{i=1}^n X_n$.*
Assume $\Psi : \mathcal{B} \to \mathbb{R}_+$ is a function satisfying the following:

1. *$\Psi(0) = 0$ and Ψ is twice Fréchet differentiable;*
2. *$\|d\Psi_x\|_{\mathrm{op}} \leq 1$ for all $x \in \mathcal{B}$;*
3. *For a constant $D \geq 1$, it holds $\|d^2(\Psi^2)_x\|_{\mathrm{op}} \leq D^2$ for all $x \in \mathcal{B}$.*

Then for any $\lambda > 0$ such that $\mathbb{E}[\exp(\lambda\|X_k\|)]$ for all k, it holds for all $t > 0$:

$$\mathbb{P}[\Psi(S_n) > t] \leq 2\exp\left(-\lambda t + D^2\sum_{k=1}^n A_k\right), \tag{3.9}$$

where $A_k := \mathbb{E}[\pi(\lambda\|X_k\|)]$, with $\pi(u) = e^u - 1 - u$.

Proof. Since we want to bound deviations from 0 rather than from the expectation $\mathbb{E}[\Psi(S_n)]$, we introduce the symmetrized random variable $R\Psi(S_n)$, where R is a Rademacher variable (random sign) independent of S_n.

We start by the usual Chernov's method bound: for any $\lambda \geq 0$ and $t \geq 0$, and recalling $\Psi(\cdot) \geq 0$, it holds

$$\mathbb{P}[\Psi(S_n) > t] \leq 2\mathbb{P}[R\Psi(S_n) > t] \leq 2\frac{\mathbb{E}[\exp(\lambda R\Psi(S_n))]}{\exp(\lambda t)} = 2\frac{\mathbb{E}[\cosh(\lambda\Psi(S_n))]}{\exp(\lambda t)},$$

where we recall $\cosh(u) = \frac{1}{2}(e^u + e^{-u})$.

We now want to apply the principle of the previous lemma for the function $F(u) = \cosh(\lambda\Psi(u))$, so $F_{x,u}(t) = \cosh(\lambda(\Psi(x + tu))) = \cosh(\lambda G(t))$, where $G(t) = \Psi(x + tu)$. Assumption (a) of Lemma 3.9 is satisfied.

It comes

$$F'_{x,u}(t) = \lambda\sinh(\lambda G(t))G'(t) = \lambda\sinh(\lambda G(t))[d\Psi_{x+tu}](u),$$

which shows that Assumption (b) of Lemma 3.9 is satisfied (with $F_x = \sinh(\lambda G(0))d\Psi_x$); and

$$F''_{x,u}(t) = \lambda\sinh(\lambda G(t))G''(t) + \lambda^2\cosh(\lambda G(t))(G'(t))^2.$$

To upper bound the first term, we bound it by 0 if $G(t)$ and $G''(t)$ are of opposite signs, otherwise we use $|\sinh(u)| \leq u\cosh(u)$, thus

$$F''_{x,u}(t) \leq \begin{cases} (G(t)G''(t) + (G'(t))^2)\lambda^2\cosh(\lambda G(t)) & \text{if } G(t)G''(t) \geq 0; \\ (G'(t))^2)\lambda^2\cosh(\lambda G(t)) & \text{if } G(t)G''(t) < 0. \end{cases}$$

In the first case, we have

$$\big(G(t)G''(t) + (G'(t))^2\big) = \frac{1}{2}(G^2)''(t) = \frac{1}{2}(d^2(\Psi^2)_{x+tu})(u,u) \leq D^2\|u\|^2,$$

by Assumption 3, and in the second case, by Assumption 2:

$$(G'(t))^2 = (d\Psi_{x+tu}u)^2 \leq \|u\|^2 \leq D^2\|u\|^2,$$

since $D \geq 1$.

Note that Assumption 2 implies that Ψ is Lipschitz; we use this to get the bound

$$\begin{aligned}\cosh(\lambda G(t)) &= \cosh(\lambda\Psi(x+tu)) \\ &\leq \cosh(\lambda\Psi(x) + \lambda t\|u\|)) \\ &\leq \cosh(\lambda\Psi(x))\exp(\lambda t\|u\|) \\ &= F(x)\exp(\lambda t\|u\|),\end{aligned}$$

where we have used $\cosh(a+b) \leq \cosh(a)\exp(b)$ for $b \geq 0$. Gathering the previous estimates we obtain (in all cases)

$$F''_{x,u}(t) \leq D^2\lambda^2\|u\|^2\exp(\lambda t\|u\|)F(x),$$

i.e. Assumption (c) of Lemma 3.9 is satisfied with $H(u,t) = D^2\lambda^2\|u\|^2\exp(\lambda t\|u\|)$. To apply the lemma all is left is to evaluate (using integration by parts)

$$\int_0^1 H(u,t)(1-t)dt = D^2\lambda^2\|u\|^2\int_0^1(1-t)\exp(\lambda t\|u\|)dt = D^2(\exp(\lambda\|u\|) - 1 - \lambda\|u\|).$$

Applying Lemma 3.9 recursively (starting from n backwards to 1), and using $\Psi(0) = 0$ for the last step, we thus get, with $A_k := \mathbb{E}[\exp(\lambda\|X_k\|) - 1 - \lambda\|X_k\|]$:

$$\mathbb{E}[\cosh(\lambda\Psi(S_n))] \leq \prod_{k=1}^n(1 + D^2A_k) \leq \exp\left(D^2\sum_{k=1}^n A_k\right),$$

leading to the announced conclusion. □

Corollary 3.11. *Under the same assumptions as Theorem 3.10, if for all $i = 1,\dots,n$, and for all integers $k \geq 2$:*

$$\mathbb{E}\Big[\|X_i\|^k\Big] \leq \frac{k!}{2D^2}\sigma^2M^{k-2}, \tag{3.10}$$

for some constants $M, \sigma > 0$,
then for all $t \geq 0$:

$$\mathbb{P}\left[\Psi\left(\frac{1}{n}S_n\right) > t\right] \leq 2\exp\left(-n\frac{\sigma^2}{M^2}h_1\left(t\frac{M}{\sigma^2}\right)\right),$$

where $h_1(u) := 1 + u - \sqrt{1+2u}$.
As a consequence, for any $x \geq 0$:

$$\mathbb{P}\left[\Psi\left(\frac{1}{n}S_n\right) > \sigma\sqrt{\frac{2x}{n}} + \frac{Mx}{n}\right] \leq 2\exp(-x).$$

Proof. Since $\pi(u) = e^u - 1 - u = \sum_{j\geq 2} \frac{u^j}{j!}$, it holds (using X_i/n instead of X_i to account for the sum normalization):

$$A_i = \mathbb{E}[\pi(\lambda\|X_i\|/n)] = \sum_{k\geq 2} \frac{\lambda^k}{k!n^k}\mathbb{E}\Big[\|X_i\|^k\Big] \leq \frac{\lambda^2\sigma^2}{2D^2n^2}\sum_{k\geq 2}(\lambda M n^{-1})^{k-2} = \frac{1}{D^2}\frac{\sigma^2 n^{-2}\lambda^2}{2(1-\lambda M n^{-1})},$$

for $\lambda \in [0, M^{-1})$. Plugging this into (3.9) yields

$$\mathbb{P}\left[\Psi\left(\frac{1}{n}S_n\right) > t\right] \leq 2\exp\left(-\lambda t + \frac{\sigma^2\lambda^2/n}{2(1-\lambda M/n)}\right),$$

and elementary computations show that

$$\sup_{\lambda\in(0,1/c)}\left(\lambda t - \frac{\lambda^2 v}{2(1-c\lambda)}\right) = \frac{v}{c^2}h_1\left(\frac{ct}{v}\right),$$

leading to the first claim. It can be checked that $h_1^{-1}(u) = u + \sqrt{2u}$, leading to the second claim. □

The primary application of Pinelis' concentration inequality is for Hilbert and (certain) Banach norms. However norms are not differentiable everywhere (in particular not at the origin). For this, we need the following result in order to slightly weaken the assumption on Ψ:

Definition 3.12. If $\mathcal{B}$ is a Banach space, we call a function $\Psi : \mathcal{B} \to \mathbb{R}_+$ $(2, D)$-smooth (for some constant $D \geq 1$) if it satisfies $\Psi(0) = 0$ and for any $x, u \in \mathcal{B}$:

$$|\Psi(x+u) - \Psi(u)| \leq \|u\|; \tag{3.11}$$

$$\Psi^2(x+u) - 2\Psi^2(x) + \Psi^2(x-u) \leq 2D\|u\|^2. \tag{3.12}$$

Proposition 3.13. *Theorem 3.10 and Corollary 3.11 also hold for any $(2, D)$-smooth function Ψ.*

Proof. We only sketch the proof. As a first step, a centered random variable X taking values in a separable Banach space $\mathcal{B}$ can be approximated by a sequence X_k such that X_k conververges to X in probability, X_k is centered and takes only a finite number of values. This can be seen as follows: put $\varepsilon = 1/k$, there exists a compact K_ε such that $\mathbb{P}[X \notin K_\varepsilon] \leq \varepsilon$ by tightness of a probability measure on a separable Banach space (Ulam's theorem). Cover K_ε by a finite number of closed balls of radius ε. Define $X_\varepsilon = \mathbb{E}[X|\mathcal{F}_\varepsilon]$, where $\mathcal{F}_\varepsilon$ is the (finite) sigma-algebra generated by those balls. Thus X_ε only takes a finite number of values, $\mathbb{E}[X_\varepsilon] = \mathbb{E}[X] = 0$, and $X_\varepsilon \xrightarrow{\mathcal{P}} X$ because $\mathbb{P}[\|X - X_\varepsilon\| > 2\varepsilon] \leq \mathbb{P}[X \notin K_\varepsilon] + \mathbb{P}[X \in K_\varepsilon; \|X - X_\varepsilon\| > 2\varepsilon] \leq \varepsilon$. The second event has probability 0 because on K_ε, X_ε is the conditional average of X on a partition piece of diameter less than 2ε, so $\|X - \mathbb{E}[X|\mathcal{F}_\varepsilon]\| \leq 2\varepsilon$.

As a second step, we establish that on a finite-dimensional Banach space $\widetilde{\mathcal{B}}$ (the one generated by the finite-numbered values of the approximant X_k for fixed k), there exists a sequence $\widetilde{\Psi}_n$ of functions $\widetilde{\mathcal{B}} \to \mathbb{R}_+$ such that $\widetilde{\Psi}_n(x) \to \Psi(x)$ for all $x \in \widetilde{\mathcal{B}}$, and functions $\widetilde{\Psi}_n$ satisfy (independently of n) conditions 1-2-3 of Theorem 3.10. Namely, let N be a centered Gaussian variable in $\widetilde{\mathcal{B}}$ (for instance, take any (finite) basis of $\widetilde{\mathcal{B}}$ and a standard normal with respect to that basis), then define

$$\Psi_\varepsilon(x) = \big(\mathbb{E}\big[\Psi^2(x - \varepsilon N)\big]\big)^{\frac{1}{2}}.$$

Then by properties of finite-dimensional convolution, since the Gaussian density is $\mathcal{C}^\infty$, it follows that Ψ_ε is $\mathcal{C}^\infty$, and since Ψ is Lipschitz, $\Psi_\varepsilon(x) \to \Psi(x)$ as $\varepsilon \to 0$.

Establishing condition 2 of Theorem 3.10 is not completely obvious, since we use a smoothing of Ψ^2 to define Ψ_ε, it does not follow straightforwardly from (3.11). However, we have:

$$\begin{aligned}
\big|d(\Psi_\varepsilon)_x^2(v)\big| &\leq \limsup_{t\to 0} \frac{1}{t}\mathbb{E}\big[\big|\Psi^2(x + tv - \varepsilon N) - \Psi^2(x - \varepsilon N)\big|\big] \\
&= \limsup_{t\to 0} \frac{1}{t}\mathbb{E}[|\Psi(x + tv - \varepsilon N) - \Psi(x - \varepsilon N)|(\Psi(x + tv - \varepsilon N) + \Psi(x - \varepsilon N))] \\
&\leq \|v\| \limsup_{t\to 0} \mathbb{E}[\Psi(x + tv - \varepsilon N) + \Psi(x - \varepsilon N)] \\
&= 2\|v\|\mathbb{E}[\Psi(x - \varepsilon N)]
\end{aligned}$$

where we have used Lipschitzness of Ψ, (3.11) for the second inequality. Thus:

$$\begin{aligned}
|d(\Psi_\varepsilon)_x(v)| &= \left|\frac{d(\Psi_\varepsilon)_x^2(v)}{2\Psi_\varepsilon(x)}\right| \\
&\leq \|v\| \frac{\mathbb{E}[\Psi(x - \varepsilon N)]}{\mathbb{E}[\Psi^2(x - \varepsilon N)]^{\frac{1}{2}}} \leq \|v\|.
\end{aligned}$$

This proves that $\|(d\Psi_\varepsilon)_x\|_{\mathrm{op}} \leq 1$ for any x.
Furthermore, since we know that Ψ_ε is $\mathcal{C}^\infty$, (3.12) implies by standard (Taylor expansion) arguments that $(d^2(\Psi_\varepsilon^2)_x)(v, v) \leq D\|v\|^2$, i.e. $\|d^2(\Psi_\varepsilon^2)_x\|_{\mathrm{op}} \leq D$ for any x. Thus points 1-2-3 of Theorem 3.10 hold for $\widetilde{\Psi}_k = \Psi_{1/k} - \Psi_{1/k}(0)$, and since $\widetilde{\Psi}_k \to \Psi$ pointwise the conclusion of the theorem holds for Ψ as well. □

A particular case of interest is for bounded random vectors. If X is a centered, bounded random vector in Hilbert space, with $\|X\| \leq M$ and $\mathbb{E}[X \otimes X^*] = \Sigma$ its covariance operator, then

$$\mathbb{E}\Big[\|X\|^k\Big] \leq M^{k-2}\mathbb{E}\big[\|X\|^2\big] = M^{k-2}\mathbb{E}[\mathrm{Tr}(X \otimes X^*)] = M^{k-2}\,\mathrm{Tr}(\Sigma),$$

so (3.10) is satisfied with $\sigma^2 = \mathrm{Tr}(\Sigma)$ (and $M = M$; and $D = 1$ for a Hilbert space).

As in the previous section, it is legitimate to ask: apart from a Hilbert space, are some of known spaces $(2, D)$-smooth? We have the following facts:

- any L^p space over a measure space is $(2, \sqrt{p-1})$-smooth for $p \geq 2$ [24] (in particular the ℓ_p norm on finite vectors or on sequences);
- using a closely related (and more standard) notion of smoothness (for a connection between the two notions see e.g. [14], eq. (9)), it was established by [31] that the p-Schatten classes share the same "modulus of smoothness" (up to constant factors) as an L^p space, for $1 \leq p < \infty$.

3.3.1 Discussion

TODO

3.4 Bernstein's inequality in operator norm

Sources: [32, 21]

As we have noted from the previous section, the vector-Hoeffding's inequality and Pinelis' inequality do not apply to L^∞ spaces, nor to $\mathcal{B}(\mathcal{H})$ with the operator norm. Because the latter space is fundamental, we now turn to concentration results on $\mathcal{B}(\mathcal{H})$. We will be adapting the "Matrix Bernstein inequalities" from [32] (see also M. Lerasle's notes [17], Chap. 4) to the operator setting. Note that we will only concentrate on self-adjoint operators in this section.

There are two issues with extending the arguments of Matrix Bernstein concentration in infinite-dimensional space:

- they rely on inequalities relating traces and functional calculus for matrices ("Trace inequalities"), in particular Lieb's theorem. One has to be somewhat careful that the traces exist for operators and that the arguments can be carried over.
- the most basic inequalities involve the matrix dimension, which will not be possible for operators over an infinite-dimensional Hilbert space. For this reason we will look at refined concentration inequalities using the *intrinsic* dimension rather than the ambient dimension.

Theorem 3.14 (Matrix Bernstein's inequality with intrinsic dimension)**.** *Let $X_1, \dots, X_n$ be independent random matrices of the same size with $\mathbb{E}[X_k] = 0$ and $\|X_k\|_{\mathrm{op}} \leq L$ for all k. Denote $S_n = \sum_{i=1}^n X_i$.*

Assume there are two matrices V_1, V_2 satisfying

$$V_1 \succeq \mathbb{E}[S_n S_n^*] = \sum_{i=1}^n \mathbb{E}[X_i X_i^*];$$

$$V_2 \succeq \mathbb{E}[S_n^* S_n] = \sum_{i=1}^n \mathbb{E}[X_i^* X_i];$$

(where for Hermitian matrices, $A \succeq B \Leftrightarrow (A-B)$ positive semidefinite).
Assume $\max \|V_1\|_{\mathrm{op}}, \|V_2\|_{\mathrm{op}} \leq \sigma^2$ *and let* $d := \frac{(\mathrm{Tr}\, V_1 + \mathrm{Tr}\, V_2)}{\sigma^2}$.
Then for $t \geq \sigma + L/3$, *it holds*

$$\mathbb{P}\Big[\|S_n\|_{\mathrm{op}} > t\Big] \leq 8d \exp\left(-\frac{t^2/2}{\sigma^2 + Lt/3}\right).$$

Lets us establish that this result extends to operators.

Corollary 3.15 (Operator Bernstein's inequality)**.** *Let* $X_1, \ldots, X_n$ *be independent, random elements of* $\mathcal{K}(\mathcal{H})$. *Assume* $\|X_k\|_{\mathrm{op}} \leq L$ *for all* k *(thus* $X_k, X_k X_k^*, X_k^* X_k$ *are Bochner-integrable in* $\mathcal{B}(\mathcal{H})$*) and that* $\mathbb{E}[X_k] = 0$.
Denote $S_n = \sum_{i=1}^n X_i$.
Assume that there exist two positive self-adjoint trace-class operators V_1, V_2 *such that*

$$V_1 \succeq \mathbb{E}[S_n S_n^*] = \sum_{i=1}^n \mathbb{E}[X_i X_i^*]; \tag{3.13}$$

$$V_2 \succeq \mathbb{E}[S_n^* S_n] = \sum_{i=1}^n \mathbb{E}[X_i^* X_i] \tag{3.14}$$

(where for self-adjoint operators $A \succeq B \Leftrightarrow (A-B)$ *positive operator).*
Assume $\max(\|V_1\|_{\mathrm{op}}, \|V_2\|_{\mathrm{op}}) \leq \sigma^2$ *and let* $d := \frac{(\mathrm{Tr}\, V_1 + \mathrm{Tr}\, V_2)}{\sigma^2}$.
Then for $t \geq \sigma + L/3$, *it holds*

$$\mathbb{P}\Big[\|S_n\|_{\mathrm{op}} > t\Big] \leq 8d \exp\left(-\frac{t^2/2}{\sigma^2 + Lt/3}\right).$$

Consequently, for any $\delta \in (0,1)$, *it holds with probability at least* $1-\delta$*:*

$$\Big\|\frac{1}{n} S_n\Big\|_{\mathrm{op}} \leq \sqrt{\frac{2\sigma^2 \beta}{n}} + \frac{2L\beta}{3n}, \qquad \beta = \log\big(8 \max(d,1) \delta^{-1}\big). \tag{3.15}$$

Remark: Since V_2 is trace-class, condition (3.14) implies by positivity of expectation that, in fact, X_i must be Hilbert-Schmidt operators a.s. We have not required this explicitly in the assumption, to avoid splitting hair about what space the X_is are Bochner integrable in. The condition $\|X_k\|_{\mathrm{op}} \leq L$ is enough to ensure Bochner integrability in $\mathcal{B}(\mathcal{H})$, but Bochner integrability in $\mathrm{HS}(\mathcal{H})$ is not formally required (note that the X_is may be unbounded in $\mathrm{HS}(\mathcal{H})$).

We use the same device as in the proof of Proposition 3.13, which we rewrite for convenience as a separate lemma:

Lemma 3.16. *Let* X *be a Bochner-integrable random variable taking values in the space* $\mathcal{K}(\mathcal{H}, \mathcal{H}')$ *of compact operators from* $\mathcal{H}$ *to* $\mathcal{H}'$, *with* $\mathbb{E}[X] = 0$. *Then there exists a sequence of random variables* $X^{(k)}$ *converging in probability to* X *for* $\|.\|_{\mathrm{op}}$, *i.e.*

$$\forall t > 0 : \lim_{k \to \infty} \mathbb{P}\Big[\big\|X - X^{(k)}\big\|_{\mathrm{op}} \geq t\Big] \to 0,$$

and such that:

- $\mathbb{E}\big[X^{(k)}\big] = 0$;
- *$X^{(k)}$ only takes a finite number of different values in a subspace $\mathcal{K}^{(k)}$ of operators, such that there exists two finite-dimensional subspaces $E^{(k)}, F^{(k)}$ of $\mathcal{H}$ and $\mathcal{H}'$ with*

$$\forall A \in \mathcal{K}^{(k)} : \mathrm{Ker}(A) \subseteq (E^{(k)})^{\perp};\ \ \mathrm{Ran}(A) \subseteq F^{(k)};$$

(in other words $P_{F^{(k)}} A P_{E^{(k)}} = A$ for all $A \in \mathcal{K}^{(k)}$, where $P_{E^{(k)}}, P_{F^{(k)}}$ are the orthogonal projections onto $E^{(k)}, F^{(k)}$).

- *If $\|X\|_{\mathrm{op}} \leq M$ a.s. for some constant M, then $\big\|X^{(k)}\big\|_{\mathrm{op}} \leq M$ a.s. as well.*
- *If XX^* and X^*X are Bochner integrable, then $\mathbb{E}\big[X^{(k)}(X^{(k)})^*\big] \preceq P_{F^{(k)}}\mathbb{E}[XX^*]P_{F^{(k)}}$, and $\mathbb{E}\big[(X^{(k)})^*X^{(k)}\big] \preceq P_{E^{(k)}}\mathbb{E}[X^*X]P_{E^{(k)}}$.*

Proof. For a fixed k, put $\varepsilon = 1/k$, there exists a compact K_ε such that $\mathbb{P}[X \notin K_\varepsilon] \leq \varepsilon$ by tightness of a probability measure on a separable Banach space. Cover K_ε by a finite number of closed balls of radius ε. Define $X_\varepsilon = \mathbb{E}[X|\mathcal{F}_\varepsilon]$, where $\mathcal{F}_\varepsilon$ is the (finite) sigma-algebra generated by those balls. Thus X_ε only takes a finite number of values, and $\mathbb{E}[X_\varepsilon] = \mathbb{E}[X] = 0$ by the properties of conditional expectation.
Let $(A_1, \ldots, A_{N_\varepsilon})$ be the finite set of values taken by X_ε. Since these are compact operators, there exists $(B_1, \ldots, B_{N_\varepsilon})$ such that B_i is finite-rank, $\|B_i - A_i\|_{\mathrm{op}} \leq \varepsilon$ for all $i \in \{1, \ldots, N_\varepsilon\}$. Let $E_\varepsilon = (\bigcap \mathrm{Ker}(B_i), i \leq N_\varepsilon)^\perp$ and $F_\varepsilon = [\mathrm{Ran}(B_i), i \leq N_\varepsilon]$. Then because $\mathrm{Ker}(B_i)$ is of finite codimension and $\mathrm{Ran}(A_i)$ is of finite dimension, E_ε and F_ε are of finite dimension; and we have $P_{F_\varepsilon} B_i P_{E_\varepsilon} = B_i$ for all $i \leq N_\varepsilon$.
Define now $\widetilde{X}_\varepsilon = P_{F_\varepsilon} X_\varepsilon P_{E_\varepsilon}$. Then $P_{F_\varepsilon}\widetilde{X}_\varepsilon P_{E_\varepsilon} = \widetilde{X}_\varepsilon$, $\mathbb{E}\big[\widetilde{X}_\varepsilon\big] = 0$ by linearity, and it holds

$$\mathbb{P}\Big[\big\|\widetilde{X}_\varepsilon - X\big\|_{\mathrm{op}} > 4\varepsilon\Big] \leq \mathbb{P}\Big[\big\|\widetilde{X}_\varepsilon - X_\varepsilon\big\|_{\mathrm{op}} > 2\varepsilon\Big] + \mathbb{P}\Big[\big\|X_\varepsilon - X\big\|_{\mathrm{op}} > 2\varepsilon\Big]. \tag{3.16}$$

Concerning the first term, we have

$$\begin{aligned}
\big\|\widetilde{X}_\varepsilon - X_\varepsilon\big\|_{\mathrm{op}} &\leq \sup_{i \leq N_\varepsilon} \big\|P_{F_\varepsilon} A_i P_{E_\varepsilon} - A_i\big\|_{\mathrm{op}} \\
&\leq \sup_{i \leq N_\varepsilon} \Big(\big\|P_{F_\varepsilon} A_i P_{E_\varepsilon} - B_i\big\|_{\mathrm{op}} + \|B_i - A_i\|_{\mathrm{op}}\Big) \\
&\leq \varepsilon + \sup_{i \leq N_\varepsilon} \Big(\big\|P_{F_\varepsilon}(A_i - B_i)P_{E_\varepsilon}\big\|_{\mathrm{op}}\Big) \\
&\leq \varepsilon + \sup_{i \leq N_\varepsilon} \Big(\big\|A_i - B_i\big\|_{\mathrm{op}}\Big) \\
&\leq 2\varepsilon,
\end{aligned}$$

hence the first probability is zero. Concerning the second term in (3.16):

$$\mathbb{P}\Big[\big\|X_\varepsilon - X\big\|_{\mathrm{op}} > 2\varepsilon\Big] \leq \mathbb{P}\Big[X \in K_\varepsilon; \big\|X_\varepsilon - X\big\|_{\mathrm{op}} > 2\varepsilon\Big] + \mathbb{P}[X \notin K_\varepsilon] \leq \varepsilon,$$

The first event above has probability 0 because on K_ε, X_ε is the conditional average of X on a partition piece of diameter less than 2ε, so $\|X - \mathbb{E}[X|\mathcal{F}_\varepsilon]\| \leq 2\varepsilon$. This proves that $X^{(k)} := \widetilde{X}_{1/k}$ converges in probability to X for $\|.\|_{\mathrm{op}}$.

Let us turn to the additional claims on boundedness and second moment. Since X_ε is defined as a conditional expectation of X, it inherits its boundedness property $\|X_\varepsilon\|_{\mathrm{op}} \leq L$, and it holds $\|\widetilde{X}_\varepsilon\|_{\mathrm{op}} = \|P_{F_\varepsilon} X_\varepsilon P_{E_\varepsilon}\|_{\mathrm{op}} \leq \|X_\varepsilon\|_{\mathrm{op}} \leq L$. Concerning the variance, first note that mimicking the usual argument for vector-valued variables, in general for an operator-valued random variable Z such that ZZ^* is Bochner integrable, it holds

$$\mathbb{E}[ZZ^*] - \mathbb{E}[Z]\mathbb{E}[Z]^* = \mathbb{E}[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^*] \succeq 0,$$

so $\mathbb{E}[ZZ^*] \succeq \mathbb{E}[Z]\mathbb{E}[Z]^*$; and this also holds for conditional expectations. Because X_ε is a conditional expectation of X, we therefore have $\mathbb{E}\big[X_\varepsilon X_\varepsilon^*\big] \preceq \mathbb{E}[XX^*]$.

Finally, we have $\widetilde{X}_\varepsilon = PX_\varepsilon P'$ for two orthogonal projectors P, P'.
In general, $A^*PA \preceq A^*A$. Namely, it holds for any u:

$$\langle A^*PPAu, u\rangle = \|PAu\|^2 \leq \|Au\|^2 = \langle A^*Au, u\rangle,$$

Similarly, it is easy to check that if $A \preceq B$, then $PAP \preceq PBP$. Hence $X_\varepsilon P' X_\varepsilon^* \preceq X_\varepsilon X_\varepsilon^*$, implying $\mathbb{E}\big[\widetilde{X}_\varepsilon \widetilde{X}_\varepsilon^*\big] \preceq P\mathbb{E}[XX^*]P$; the argument is similar to establish $\mathbb{E}\big[\widetilde{X}_\varepsilon^* \widetilde{X}_\varepsilon\big] \preceq P'\mathbb{E}[X^*X]P'$. □

Proof of Corollary 3.15. We use the construction of Lemma 3.16 for $X_1, \ldots, X_n$ (remember, these are compact operators), resulting in approximants $X_1^{(k)}, \ldots, X_n^{(k)}$. Since each $X_i^{(k)}$ only depends on X_i, the independence of X_is carries over to independence of $X_i^{(k)}$s (for fixed k). Furthermore, while in principle the finite dimensional spaces $K^{(k)}, E^{(k)}, F^{(k)}$ in the construction of Lemma 3.16 depend on i, obviously we can assume that they are common to all indices by replacing them by the respective linear span of their union over $i = 1, \ldots, n$. By the same token, since we apply Lemma 3.16 with $\mathcal{H} = \mathcal{H}'$, we can actually assume $E^{(k)} = F^{(k)}$; let us denote $P^{(k)}$ the orthogonal projector on $E^{(k)}$. Furthermore Lemma 3.16 guarantees $\|X_i^{(k)}\|_{\mathrm{op}} \leq L$ and (using the last point of the lemma) it holds

$$\sum_{i=1}^n \mathbb{E}\big[X_i^{(k)}\big(X_i^{(k)}\big)^*\big] \preceq P^{(k)} \sum_{i=1}^n \mathbb{E}[X_i X_i^*] P^{(k)} \preceq P^{(k)} V_1 P^{(k)} := V_1^{(k)}.$$

Observe that $\mathrm{Tr}(V_1^{(k)}) = \mathrm{Tr}(P^{(k)} V_1 P^{(k)}) \leq \|P^{(k)}\|_{\mathrm{op}}^2 \,\mathrm{Tr}\,|V_1| = \mathrm{Tr}\, V_1$ since V_1 is positive; and $\|V_1^{(k)}\|_{\mathrm{op}} = \|P^{(k)} V_1 P^{(k)}\|_{\mathrm{op}} \leq \|V_1\|_{\mathrm{op}} \leq \sigma^2$. Similar arguments hold to establish

$$\sum_{i=1}^n \mathbb{E}\big[\big(X_i^{(k)}\big)^* X_i^{(k)}\big] \preceq P^{(k)} V_2 P^{(k)} := V_2^{(k)},$$

with $\mathrm{Tr}(V_2^{(k)}) \leq \mathrm{Tr}(V_2)$ and $\|V_2^{(k)}\|_{\mathrm{op}} \leq \sigma^2$. To summarize, for fixed k the approximant variables $X_1^{(k)}, \ldots, X_n^{(k)}$ are independent, self-adjoint operators that are null on the orthogonal of the finite-dimensional subspace $E^{(k)}$, hence can be conceived as finite-dimensional

Hermitian matrices acting on $E^{(k)}$; this is also the case for $V_1^{(k)}, V_2^{(k)}$. We can thus apply Theorem 3.14 to these variables, resulting in

$$\mathbb{P}\Big[\big\|S_n^{(k)}\big\|_{\mathrm{op}} > t\Big] \leq 8d \exp\left(-\frac{t^2/2}{\sigma^2 + Lt/3}\right),$$

where $S_n^{(k)} = \sum_{i=1}^n X_i^{(k)}$. Since $X_i^{(k)}$ converges in probability to X_i, so does $S_n^{(k)}$ to S_n, yielding the claim. (As a crucial point, observe that the right-hand side of the above inequality does not depend on k, since for each k Theorem 3.14 is applied for the constants L, σ^2, d, independent of k.)

To get to the form (3.15) with explicit deviation expression at probability at most δ, notice that the above bound holds *a fortiori* when replacing d by $\widetilde{d} = \max(d, 1)$; thus $\beta = \log(8\widetilde{d}\delta^{-1}) \geq \log(8) \geq 2$, so that the condition $t \geq \sigma + L/3$ is satisfied with t given by the right-hand side expression in (3.15), and the rest is standard computation.

□

To prove Theorem 3.14, we concentrate on the Hermitian (=self-adjoint) case; for an extension to the general case see [32]. We want to use a "matrix version" of Chernov's method, but a critical point is that it does **not** hold that $\exp(A+B) = \exp(A)\exp(B)$ in general (unless A and B commute). One can think of it the following way: $\exp(A+B) = \exp(B+A)$, but it would seem strange that $\exp(A)$ and $\exp(B)$ commute if A and B don't (and indeed, this is not true). We will actually be interested in the trace of the exponential, and it turns out that $\mathrm{Tr}\exp(A+B) \leq \mathrm{Tr}(\exp(A)\exp(B))$ (Golden-Thompson inequality) but unfortunately this does not extend to more than 2 matrices.

It turns out that a convenient central tool for the development of the "Matrix Chernov's method" is the following:

Theorem 3.17 (Lieb's theorem)**.** *Let H be an Hermitian matrix with dimension d, then the function*

$$A \mapsto \mathrm{Tr}\exp(H + \log(A)) \tag{3.17}$$

is concave on the the cone $\mathcal{C}_d$ of $d \times d$ positive-definite matrices to $\mathbb{R}$.

For a proof, see e.g. [32]. From this, we deduce the principal device underlying the Matrix Chernov's method:

Proposition 3.18. *Let $X_1, \ldots, X_n$ be random Hermitian matrices of the same dimension. Then it holds for any real λ:*

$$\mathbb{E}\left[\mathrm{Tr}\exp\left(\lambda \sum_{i=1}^n X_i\right)\right] \leq \mathrm{Tr}\exp\left(\sum_{i=1}^n \log \mathbb{E}[\exp \lambda X_i]\right).$$

Proof. We take successively expectation with respect to $X_1, \ldots, X_n$. Assume that after $k-1$ steps, we have established

$$\mathbb{E}\left[\mathrm{Tr}\exp\left(\lambda \sum_{i=1}^n X_i\right)\Big| X_k, \ldots, X_n\right] \leq \mathrm{Tr}\exp\left(\sum_{i=1}^{k-1} \xi_i + \lambda \sum_{i=k}^n X_i\right), \tag{3.18}$$

where $\xi_i := \log \mathbb{E}[\exp \lambda X_i]$. Putting $H_k = \sum_{i=1}^{k-1} \xi_i + \lambda \sum_{i=k+1}^{n} X_i$ and $A_k = \exp \lambda X_k$, we use Jensen's inequality and the concavity property of the function defined in (3.17) to obtain, when taking expectation with respect to X_k:

$$\mathbb{E}[\operatorname{Tr} \exp(H_k + \log(\exp \lambda X_k))|X_{k+1}, \ldots, X_n] \leq \operatorname{Tr} \exp(H_k + \log(\mathbb{E}[\exp \lambda X_k]));$$

combining with (3.18), and replacing the value of H_k we get (3.18) for $k \leftarrow (k+1)$. We conclude by a straightforward recursion. □

In order to establish Theorem 3.14, we will need the following properties on Hermitian matrix functional calculus (defined in the same way as operator functional calculus, see Section 2.4):

Proposition 3.19. *Let A, B be Hermitian matrices. We denote $A \preceq B$, resp. $A \prec B$ iff $B - A$ is positive semidefinite, resp. positive definite.*

1. *If f is nondecreasing on the union of the spectra of A and B, and $A \preceq B$, then $\operatorname{Tr} f(A) \leq \operatorname{Tr} f(B)$.*
2. *If $f \leq g$ on the spectrum of A, then $f(A) \preceq g(A)$.*

Proof. For the first point, use the Courant-Fisher "max-min" theorem: for a Hermitian matrix A, denote $\lambda_i(A)$ the i-th largest eigenvalue of A (counted with multiplicity), then it holds

$$\lambda_i(A) = \max_{V:\dim(V)=1} \min_{u \in V, \|u\|=1} \langle u, Au \rangle,$$

where the maximum runs over linear subspaces of dimension i. From this it follows that

$$\begin{aligned} A \preceq B &\Rightarrow \forall i \;\; \lambda_i(A) \leq \lambda_i(B) \\ &\Rightarrow \forall i \;\; \lambda_i(f(A)) = f(\lambda_i(A)) \leq f(\lambda_i(B)) = \lambda_i(f(B)) \\ &\Rightarrow \operatorname{Tr}(f(A)) \leq \operatorname{Tr}(f(B)). \end{aligned}$$

(note that we have used the fact that f is nondecreasing to justify *each* of the relations in the second implication).

For the second point, since $(g - f)$ is a nonnegative function, we have (see e.g. points (a) and (f) of Prop. 2.19) $g(A) - f(A) = (g - f)(A) \succeq 0$. □

Lemma 3.20. *If X is a random Hermitian matrix such that its spectrum is upper bounded by L and $\mathbb{E}[X] = 0$, then, for $\lambda \in [0, 3/L]$:*

$$\log \mathbb{E}[\exp \lambda X] \preceq \frac{\lambda^2/2}{1 - \lambda L/3} \mathbb{E}\big[X^2\big].$$

Proof. We start with (re)defining the real function $\pi(x) := \exp(x) - x - 1$ and $f(x) := \pi(x)/x^2$. By inspection of its series expansion, it holds that $f(x)$ is nondecreasing and that $f(x) \leq g(x) := \frac{1}{2(1-x/3)}$ for $x \in [0, 3]$. Thus for $x \leq L$ and $\lambda \in [0, 3/L]$:

$$\pi(\lambda x) \leq \lambda^2 x^2 g(\lambda L).$$

Using point 2 of Proposition 3.19, it follows that for a Hermitian matrix X such that its spectrum is upper bounded by L, it holds

$$\pi(\lambda X) \preceq \lambda^2 X^2 g(\lambda L).$$

Taking the expectation (and using positivity of expectation) yields

$$\mathbb{E}[\pi(\lambda X)] = \mathbb{E}[\exp(\lambda X)] - \boldsymbol{I} \preceq \frac{\lambda^2/2}{1-\lambda L/3}\mathbb{E}\big[X^2\big].$$

Using $\log(u) \leq u - 1$ for $u > 0$ and Point 2 of the proposition again, we get

$$\log \mathbb{E}[\exp(\lambda X)] \preceq \mathbb{E}[\exp(\lambda X)] - \boldsymbol{I} \preceq \frac{\lambda^2/2}{1-\lambda L/3}\mathbb{E}\big[X^2\big].$$

□

We can now turn to the proof of Theorem 3.14.

Proof of Theorem 3.14. Applying Lemma 3.20 repeatedly, we get for $\lambda \in [0, L/3]$:

$$\sum_{i=1}^n \log \mathbb{E}[\exp \lambda X_i] \preceq \frac{\lambda^2/2}{1-\lambda L/3} \sum_{i=1}^n \mathbb{E}\big[X_i^2\big] \preceq g(\lambda)V,$$

where $g(\lambda) = \frac{\lambda^2/2}{1-\lambda L/3}$. Using point 1 of Proposition 3.19, it comes

$$\operatorname{Tr} \exp\left(\sum_{i=1}^n \log \mathbb{E}[\exp \lambda X_i]\right) \leq \operatorname{Tr} \exp(g(\lambda)V).$$

Combining the above with Proposition 3.18, we obtain a bound on the "matrix Laplace transform"

$$\operatorname{Tr} \mathbb{E}[\exp \lambda S_n] \leq \operatorname{Tr} \exp(g(\lambda)V). \tag{3.19}$$

First, we relate the left-hand side of (3.19) to the probability of deviation of $\lambda_{\max}(S_n)$, where $\lambda_{\max}$ denotes largest eigenvalue. In general, if π is a nondecreasing function from $\mathbb{R}$ to $\mathbb{R}_+$, it holds:

$$\mathbb{P}[\lambda_{\max}(S_n) > t] \leq \mathbb{P}[\pi(\lambda_{\max}(S_n)) > \pi(t)] \leq \frac{\mathbb{E}[\pi(\lambda_{\max}(S_n))]}{\pi(t)} \leq \frac{\operatorname{Tr} \mathbb{E}[\pi(S_n)]}{\pi(t)}.$$

A crucial idea (or technical trick...) to obtain the right dependence in the effective dimension is to apply the above "Matrix Markov" inequality not directly to the exponential function, but to the function (again!) $\pi(t) = \exp(\lambda t) - \lambda t - 1$ (with $\lambda > 0$), use $\mathbb{E}[S_n] = 0$ and combine with (3.19) to obtain

$$\mathbb{P}[\lambda_{\max}(S_n) > t] \leq \frac{\operatorname{Tr} \mathbb{E}[\exp(\lambda S_n) - \boldsymbol{I}]}{\exp(\lambda t) - \lambda t - 1} \leq \frac{\operatorname{Tr}(\exp(g(\lambda)V) - \boldsymbol{I})}{\exp(\lambda t) - \lambda t - 1}. \tag{3.20}$$

Second, we further bound the above right-hand side. For this, denote $\widetilde{\pi}(t) = \exp(t) - 1$; we note that since $\widetilde{\pi}$ is convex with $\widetilde{\pi}(0) = 0$, it holds $\widetilde{\pi}(t) \leq \frac{t}{M}\widetilde{\pi}(M)$ for $t \leq M$. Since $\|V\|_{\mathrm{op}} \leq \sigma^2$, we deduce by points 1-2 of Proposition 3.19 and recalling $d = 2\operatorname{Tr} V/\sigma^2$, that

$$\operatorname{Tr}(\exp(g(\lambda)V) - \boldsymbol{I}) \leq \frac{\operatorname{Tr}(g(\lambda)V)}{\sigma^2 g(\lambda)}\widetilde{\pi}(\sigma^2 g(\lambda)) \leq d\exp(\sigma^2 g(\lambda)). \tag{3.21}$$

Thus, combining with (3.20), we finally arrive at the estimate

$$\mathbb{P}[\lambda_{\max}(S_n) > t] \leq d\frac{\exp(\sigma^2 g(\lambda))}{\exp(\lambda t) - \lambda t - 1}. \tag{3.22}$$

The rest is just somewhat tedious estimates. We rewrite the expression in the above bound as

$$\frac{\exp(\sigma^2 g(\lambda))}{\exp(\lambda t) - \lambda t - 1} = \left(\frac{\exp(\lambda t)}{\exp(\lambda t) - \lambda t - 1}\right)\exp(\sigma^2 g(\lambda) - \lambda t) \leq \left(1 + \frac{3}{\lambda^2 t^2}\right)\exp(\sigma^2 g(\lambda) - \lambda t),$$

where one used $\frac{e^u}{e^u - u - 1} \leq 1 + \frac{3}{u^2}$ for $u \geq 0$, the proof of which is left to the reader. Next, we choose $\lambda = t/(\sigma^2 + Lt/3)$. If $t \geq \sigma + L/3$, it can be checked (again, left to the reader) that with this choice of λ it holds $\left(1 + \frac{3}{\lambda^2 t^2}\right) \leq 4$, and substituting the value of λ into $g(\lambda)$ in $\exp(\sigma^2 g(\lambda) - \lambda t)$ yields

$$\mathbb{P}[\lambda_{\max}(S_n) > t] \leq 4d\exp\left(-\frac{t^2/2}{\sigma^2 + Lt/3}\right);$$

we conclude by a union bound with a similar control for $\mathbb{P}[\lambda_{\max}(-S) > t]$. □

4 Spectral regularization methods

Many sources on this theme, including [1, 6, 3, 20, 19, 26]...

4.1 Setting

In this chapter, we will consider a problem of linear regression with random design where the covariate X lies in a Hilbert space, of the form

$$Y = \langle f^*, X \rangle + \xi, \tag{4.1}$$

where f^* is an unknown element of an Hilbert space $\mathcal{H}$, and X is a random variable taking values in $\mathcal{H}$.
This model is relevant in particular for Functional Data Analysis (FDA) and reproducing kernel Hilbert space (rkHs) methods. We will come back to the latter particular setting later, but for now we will focus on the abstract model (4.1) and assume we "observe" the data in Hilbert space, putting computational feasibility aside.

We will consider the following simple distributional assumptions:

Assumption 4.1 (Distribution assumptions).

We observe n i.i.d. data points $(X_i, Y_i)_{i \in [\![n]\!]}$ following the model (4.1). The unknown joint distribution on $\mathcal{H} \times \mathbb{R}$ is denoted P. Its marginal on $\mathcal{H}$ (X-marginal) is denoted ρ.

The covariate is bounded: $\|X\| \leq \kappa$ (ρ-a.s.)

The noise ξ satisfies $\mathbb{E}[\xi|X] = 0$.

The ouput variable Y is bounded: $|Y| \leq M$ (P-a.s.).

Note that the latter point implies $|\langle X, f^* \rangle| \leq M$ as well and thus $|\xi| \leq 2M$, P-a.s.

These assumptions are quite restrictive and can be significantly weakened in the literature with the price of a more refined analysis. We will study this setting here for simplicity.

In the analysis to come, a lot of constants will depend on the parameters appearing in the assumptions (such as κ and M abovel there will be more later.) To avoid a cumbersome tracking of the effect of the constants, we will often use the notation $C_\blacktriangle$ to denote a number implicitly depending on "less important" parameters in the assumptions. For this section $C_\blacktriangle$ will be a positive number only depending on (κ, M). *Note that the value of $C_\blacktriangle$ might change in different contexts and even change from line to line!*

We first need to introduce some notation which will be the infinite-dimensional analogue of quantities appearing in traditional linear regression. For this we will need to introduce the Hilbert space $L^2(\mathcal{H}, \rho)$ whose norm (and scalar product) we will denote as $\|\cdot\|_\rho$ resp $\langle \cdot, \cdot \rangle_\rho$.

Proposition/Definition 4.2. Let ρ be a distribution on the Hilbert space $\mathcal{H}$ such that $\mathbb{E}\big[\|X\|^2\big] < \infty$ (this is implied in particular by Assumption 4.1). Denote S the "population evaluation" operator

$$S: \quad \mathcal{H} \to L^2(\mathcal{H}, \rho), \qquad f \mapsto \langle f, \cdot \rangle = [x \mapsto \langle f, x \rangle].$$

This is a Hilbert-Schmidt operator, and its adjoint is given by

$$S^*: \quad L^2(\mathcal{H}, \rho) \to \mathcal{H}, \qquad g \mapsto \mathbb{E}[g(X)X].$$

Finally, it holds $S^*S = \mathbb{E}[X \otimes X^*] = \Sigma$, and Σ is a trace-class operator.

Proof. Let $(e_i)_{i \in I}$ be an orthonormal basis of $\mathcal{H}$. Then we have

$$\sum_i \|Se_i\|^2 = \sum_i \mathbb{E}\big[|\langle e_i, X \rangle|^2\big] = \mathbb{E}\big[\|X\|^2\big] < \infty,$$

establishing that S is Hilbert-Schmidt. To determine its adjoint, we notice

$$\langle Sf, g \rangle_\rho = \mathbb{E}[\langle f, X \rangle \overline{g}(X)] = \mathbb{E}[\langle f, Xg(X) \rangle] = \langle f, \mathbb{E}[Xg(X)] \rangle,$$

etablishing the announced formula for S^*. Observe that $\mathbb{E}[\|Xg(X)\|] \leq \mathbb{E}\big[\|X\|^2\big]^{\frac{1}{2}} \mathbb{E}\big[|g(X)|^2\big]^{\frac{1}{2}} < \infty$, since g is an element of $L^2(\mathcal{H}, \rho)$; this proves that $Xg(X)$ is Bochner-integrable.

Finally, $\Sigma = S^*S$ is a trace-class operator as product of two Hilbert-Schmidt operators, and we identify

$$S^*Sf = \mathbb{E}[X(Sf)(X)] = \mathbb{E}[X\langle X, f \rangle] = \mathbb{E}[X \otimes X^*]f,$$

establishing the announced formula. Observe that $X \otimes X^*$ is Bochner-integrable as an element of the Banach space of trace-class operators $\mathcal{B}_1(\mathcal{H})$, since $\|X \otimes X^*\|_1 = \|X\|^2$, which is integrable. □

We need empirical analogues of the above operators. For this we mimic the above construction with $L^2(\mathcal{H}, \widehat{\rho})$ instead of $L^2(\mathcal{H}, \rho)$, where $\widehat{\rho}$ is the empirical distribution associated to the sample $(X_1, \ldots, X_n)$. We actually identify $L^2(\mathcal{H}, \widehat{\rho})$ with $\mathbb{R}^n$ with the scalar product $\langle u, v \rangle_n := \frac{1}{n} \sum_{i \in [\![n]\!]} u_i v_i$. (This identication is not quite correct if some of the values X_i repeat)

Proposition/Definition 4.3. Conditional to $(X_1, \ldots, X_n)$ denote $\widehat{S}$ the "sample evaluation" operator

$$\widehat{S}: \quad \mathcal{H} \to (\mathbb{R}^n, \langle \cdot, \cdot \rangle_n), \qquad f \mapsto (\langle f, X_1 \rangle, \ldots, \langle f, X_n \rangle).$$

Its adjoint is given by

$$\widehat{S}^*: \quad (\mathbb{R}^n, \langle \cdot, \cdot \rangle_n) \to \mathcal{H}, \qquad (a_1, \ldots, a_n) \mapsto \frac{1}{n} \sum_{i \in [\![n]\!]} a_i X_i.$$

Finally, it holds $\widehat{S}^*\widehat{S} = \frac{1}{n} \sum_{i \in [\![n]\!]} X_i \otimes X_i^* = \widehat{\Sigma}$.

Proof. Just note

$$\left\langle \widehat{S}f, u \right\rangle_n = \frac{1}{n} \sum_{i \in [\![n]\!]} \langle f, X_i \rangle u_i = \left\langle f, \frac{1}{n} \sum_{i \in [\![n]\!]} u_i X_i \right\rangle,$$

and

$$\widehat{S}^* S f = \frac{1}{n} \sum_{i \in [\![n]\!]} X_i \langle f, X_i \rangle = \left(\frac{1}{n} \sum_{i \in [\![n]\!]} X_i \otimes X_i^* \right) f.$$

□

Finally, let us define the excess risk we want to control for an estimator $\widehat{f}$ of f^*. We will focus here on the quadratic prediction risk $R(\widehat{f}) = \mathbb{E}\Big[(\langle \widehat{f}, X \rangle - Y)^2\Big]$, where the expectation is over a new, independent example (X, Y) drawn from P. Consequently, the excess risk with respect to the optimal prediction f^* can be rewritten as

$$\begin{aligned} R(\widehat{f}) - R(f^*) &= \mathbb{E}\Big[(\langle \widehat{f}, X \rangle - Y)^2 - (\langle f^*, X \rangle - Y)^2\Big] \\ &= \mathbb{E}\Big[(\langle \widehat{f} - f^*, X \rangle - \xi)^2 - \xi^2\Big] \\ &= \mathbb{E}\Big[(\langle \widehat{f} - f^*, X \rangle)^2\Big] \\ &= \left\langle S(\widehat{f} - f^*), S(\widehat{f} - f^*) \right\rangle_\rho \\ &= \left\langle (\widehat{f} - f^*), \Sigma(\widehat{f} - f^*) \right\rangle_{\mathcal{H}} \\ &= \left\| \Sigma^{\frac{1}{2}} (\widehat{f} - f^*) \right\|_{\mathcal{H}}^2 \end{aligned} \tag{4.2}$$

4.2 Probabilistic inequalities

In analogy with the finite-dimensional ordinary least squares estimator $\widehat{\theta} = \left(\frac{1}{n} \boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \left(\frac{1}{n} \boldsymbol{X}^T \boldsymbol{Y}\right) = \widehat{\Sigma}^{-1} \frac{1}{n} \boldsymbol{X}^T \boldsymbol{Y}$ (where $\boldsymbol{X}$ is the (n, d) design matrix whose rows are the X_i, and $\boldsymbol{Y}$ is the column vector $(Y_1, \ldots, Y_n)^T$), note that the operator $\widehat{S}$ is the Hilbert space analogue of $\boldsymbol{X}$, so that we will consider etimates of the form

$$\widehat{f}_\lambda = F_\lambda(\widehat{\Sigma}) \widehat{S}^* \boldsymbol{Y},$$

where $F_\lambda(.)$ is a suitable "regularized inverse" driven by a regularization parameter $\lambda > 0$.

Under model (4.1), we have $\boldsymbol{Y} = \widehat{S} f^* + \boldsymbol{\xi}$, where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$, so

$$\widehat{f}_\lambda = F_\lambda(\widehat{\Sigma}) \widehat{S}^* (\widehat{S} f^* + \boldsymbol{\xi}) = F_\lambda(\widehat{\Sigma}) \widehat{\Sigma} f^* + F_\lambda(\widehat{\Sigma}) (\widehat{S}^* \boldsymbol{\xi}). \tag{4.3}$$

In view of the above, two quantities we wish to have control on are $\widehat{\Sigma}$ and $\widehat{S}^* \boldsymbol{\xi}$.

Let us start with an application of the simple vectorial Hoeffding's inequality:

Proposition 4.4. *Under Assumption 4.1, for $\delta \in (0,1)$ denote $L_\delta := 1 + \log \delta^{-1}$, it holds with probability $1-\delta$:*

$$\left\|\widehat{S}^*\boldsymbol{\xi}\right\| \leq \frac{4M\kappa\sqrt{L_\delta}}{\sqrt{n}} \leq C_{\blacktriangle}\sqrt{\frac{L_\delta}{n}}. \tag{4.4}$$

Also, with probability $1-\delta$:

$$\left\|\widehat{\Sigma} - \Sigma\right\|_{\mathrm{op}} \leq \left\|\widehat{\Sigma} - \Sigma\right\|_2 \leq \frac{4\kappa^2\sqrt{L_\delta}}{\sqrt{n}} \leq C_{\blacktriangle}\sqrt{\frac{L_\delta}{n}}. \tag{4.5}$$

Proof. Note that $\widehat{S}^*\boldsymbol{\xi} = \frac{1}{n}\sum_{i=1}^m \xi_i X_i$, $\mathbb{E}[\xi_i X_i] = 0$ and we have $\|\xi_i X_i\| \leq 2M\kappa$. Applying Hoeffding's inequality in Hilbert space (Theorem 3.6) yields the first claim. (We have used the inequality $(1+\sqrt{2a}) \leq 2\sqrt{1+a}$ with $a = \log\delta^{-1}$ to simplify the expression.)

For the second, we recall that $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i \otimes X_i^*$, $\mathbb{E}[X_i \otimes X_i^*] = \Sigma$, and $\|X_i \otimes X_i^*\|_2 = \|X_i\|^2 \leq \kappa^2$, so that $\|\Sigma\|_2 \leq \mathbb{E}[\|X_i \otimes X_i^*\|_2] \leq \kappa^2$ as well. Applying Hoeffding's inequality in the Hilbert space $\mathrm{HS}(\mathcal{H})$ yields the second claim. □

We turn to applications of Bernstein's inequality. To better exploit it, we will consider a "warped" version of the quantities of interest. The following quantity will play an important role:

Definition 4.5. In the context of Assumption 4.1, introduce and denote for $\lambda > 0$:

$$\Sigma_\lambda := (\Sigma + \lambda I),$$

and

$$\mathcal{N}(\lambda) := \mathrm{Tr}\big(\Sigma\Sigma_\lambda^{-1}\big) = \sum_{k\geq 1} \frac{\lambda_k}{\lambda_k + \lambda}, \tag{4.6}$$

where $(\lambda_k)_{k\geq 1}$ is the sequence of eigenvalues of Σ (with multiplicity).

(Observe that $\mathcal{N}(\lambda)$ is well-defined for any $\lambda > 0$ since Σ is trace-class.)

Proposition 4.6. *Under Assumption 4.1, for $\delta \in (0,1)$, $\lambda > 0$, denoting $L_\delta = 1 + \log\delta^{-1}$ each of the following events hold with probability $1-\delta$:*

$$\left\|\Sigma_\lambda^{-\frac{1}{2}}\widehat{S}^*\boldsymbol{\xi}\right\| \leq 2\sqrt{2}M\sqrt{\frac{L_\delta\mathcal{N}(\lambda)}{n}} + \frac{2M\kappa L_\delta}{n\sqrt{\lambda}} \leq C_{\blacktriangle}\left(\sqrt{\frac{L_\delta\mathcal{N}(\lambda)}{n}} + \frac{L_\delta}{n\sqrt{\lambda}}\right); \tag{4.7}$$

$$\left\|\Sigma_\lambda^{-\frac{1}{2}}\big(\widehat{\Sigma} - \Sigma\big)\right\|_2 \leq \kappa\sqrt{\frac{2\mathcal{N}(\lambda)L_\delta}{n}} + \frac{2L_\delta\kappa^2}{n\sqrt{\lambda}} \leq C_{\blacktriangle}\left(\sqrt{\frac{L_\delta\mathcal{N}(\lambda)}{n}} + \frac{L_\delta}{n\sqrt{\lambda}}\right); \tag{4.8}$$

and, provided $\lambda \leq \|\Sigma\|_{\mathrm{op}}$:

$$\left\|\Sigma_\lambda^{-\frac{1}{2}}\big(\widehat{\Sigma} - \Sigma\big)\Sigma_\lambda^{-\frac{1}{2}}\right\|_{\mathrm{op}} \leq \kappa\sqrt{\frac{2(\log(16\mathcal{N}(\lambda)) + L_\delta)}{\lambda n}} + \frac{2\kappa^2(\log(16\mathcal{N}(\lambda)) + L_\delta)}{3\lambda n}. \tag{4.9}$$

$$\leq C_{\blacktriangle}\left(\sqrt{\frac{\log\mathcal{N}(\lambda) + L_\delta}{\lambda n}} + \frac{\log\mathcal{N}(\lambda) + L_\delta}{\lambda n}\right). \tag{4.10}$$

Proof. For the first two cases, we will apply Pinelis' inequality in a Hilbert space (Cor. 3.11 with $\Psi(x) = \|x\|$ in a Hilbert space). For the first one, note that

$$\Sigma_\lambda^{-\frac{1}{2}} \widehat{S}^* \boldsymbol{\xi} = \frac{1}{n} \sum_{i \in [\![n]\!]} Z_i, \qquad Z_i := \Sigma_\lambda^{-\frac{1}{2}} X_i \xi_i.$$

Since $\left\|\Sigma_\lambda^{-\frac{1}{2}}\right\|_{\mathrm{op}} \leq \lambda^{-\frac{1}{2}}$, we have $\|Z_i\| \leq 2M\kappa/\sqrt{\lambda}$. Moreover, $\mathbb{E}[Z_i] = 0$ and

$$\begin{aligned}
\mathbb{E}\big[\|Z_i\|^2\big] &\leq 4M^2 \mathbb{E}\Big[\left\|\Sigma_\lambda^{-\frac{1}{2}} X_i\right\|^2\Big] \\
&= 4M^2 \mathbb{E}\Big[\operatorname{Tr}\left(\Sigma_\lambda^{-\frac{1}{2}} X_i \otimes (\Sigma_\lambda^{-\frac{1}{2}} X_i)^*\right)\Big] \\
&= 4M^2 \mathbb{E}\Big[\operatorname{Tr}\left(\Sigma_\lambda^{-\frac{1}{2}} (X_i \otimes X_i^*) \Sigma_\lambda^{-\frac{1}{2}}\right)\Big] \\
&= 4M^2 \operatorname{Tr}(\Sigma_\lambda^{-\frac{1}{2}} \mathbb{E}[X_i \otimes X_i^*] \Sigma_\lambda^{-\frac{1}{2}}) \\
&= 4M^2 \mathcal{N}(\lambda).
\end{aligned}$$

For the second one, we have $\Sigma_\lambda^{-\frac{1}{2}}(\widehat{\Sigma} - \Sigma) = \frac{1}{n}\sum_{i \in [\![n]\!]} A_i$, with $A_i = \Sigma_\lambda^{-\frac{1}{2}}((X_i \otimes X_i^*) - \mathbb{E}[X_i \otimes X_i^*])$. It holds $\mathbb{E}[A_i] = 0$ and

$$\left\|\Sigma_\lambda^{-\frac{1}{2}}(X_i \otimes X_i^*)\right\|_2 \leq \left\|\Sigma_\lambda^{-\frac{1}{2}}\right\|_{\mathrm{op}} \|X_i \otimes X_i^*\|_2 \leq \lambda^{-\frac{1}{2}} \kappa^2,$$

so that $\|A_i\|_2 \leq 2\lambda^{-\frac{1}{2}}\kappa^2$; and, due to $\mathbb{E}\big[\|Z - \mathbb{E}[Z]\|^2\big] = \mathbb{E}[\|Z^2\|] - \|\mathbb{E}[Z]\|^2 \leq \mathbb{E}\big[\|Z\|^2\big]$ for a Hilbert norm:

$$\begin{aligned}
\mathbb{E}\big[\|A_i\|_2^2\big] &\leq \mathbb{E}\left[\left\|\Sigma_\lambda^{-\frac{1}{2}}(X_i \otimes X_i^*)\right\|_2^2\right] \\
&= \mathbb{E}\big[\operatorname{Tr}\big((X_i \otimes X_i^*)\Sigma_\lambda^{-1}(X_i \otimes X_i^*)\big)\big] \\
&\leq \mathbb{E}\Big[\|X_i \otimes X_i^*\|_{\mathrm{op}} \operatorname{Tr}\big(\Sigma_\lambda^{-1}(X_i \otimes X_i^*)\big)\Big] \\
&\leq \kappa^2 \mathcal{N}(\lambda).
\end{aligned}$$

For the last claim, we will apply the operator Bernstein's inequality (Theorem 3.15). The estimates are similar to the above, now we consider a sum of i.i.d. self-adjoint random operators having the form $B_i := \Sigma_\lambda^{-\frac{1}{2}}(X_i \otimes X_i^* - \mathbb{E}[X_i \otimes X_i^*])\Sigma_\lambda^{-\frac{1}{2}}$. Using $\langle u, X_i \otimes X_i^* u\rangle \in [0, \kappa^2]$ for any unit vector u and linearity of expectation, it holds $\|B_i\|_{\mathrm{op}} \leq \kappa^2/\lambda$, and due to $\mathbb{E}[(M - E(M))^2] \preceq \mathbb{E}[M^2]$ for a self-adjoint operator-valued random variable M (s.t. M^2 is Bocher-integrable), it holds

$$\mathbb{E}\big[B_i^2\big] \preceq \Sigma_\lambda^{-\frac{1}{2}} \mathbb{E}\big[(X_i \otimes X_i^*)\Sigma_\lambda^{-1}(X_i \otimes X_i^*)\big]\Sigma_\lambda^{-\frac{1}{2}},$$

Observe that in general, $(u \otimes v)M(w \otimes x) = \langle v, Mw\rangle u \otimes x$. Therefore

$$(X_i \otimes X_i^*)\Sigma_\lambda^{-1}(X_i \otimes X_i^*) = \big\langle X_i, \Sigma_\lambda^{-1} X_i\big\rangle X_i \otimes X_i^* \preceq \frac{\kappa^2}{\lambda}(X_i \otimes X_i^*).$$

Using this into the previous display, and positivity of expectation, we obtain

$$\mathbb{E}\big[B_i^2\big] \preceq \frac{\kappa^2}{\lambda}\Sigma_\lambda^{-\frac{1}{2}}\Sigma\Sigma_\lambda^{-\frac{1}{2}} := V.$$

It holds

$$\|V\|_{\mathrm{op}} \le \frac{\kappa^2}{\lambda}\Big\|\Sigma_\lambda^{-\frac{1}{2}}\Sigma\Sigma_\lambda^{-\frac{1}{2}}\Big\|_{\mathrm{op}} \le \frac{\kappa^2}{\lambda} =: \sigma^2,$$

where the notation σ^2 is in order to link with the notation of Theorem 3.15. Furthermore, we have for the (proxy) intrinsic dimension d of V appearing in that theorem:

$$d = \frac{2\operatorname{Tr}(V)}{\sigma^2} = 2\operatorname{Tr}\Big(\Sigma_\lambda^{-\frac{1}{2}}\Sigma\Sigma_\lambda^{-\frac{1}{2}}\Big) = 2\mathcal{N}(\lambda).$$

Note that if we denote $\lambda_1 = \|\Sigma\|_{\mathrm{op}}$, then $\big\|\Sigma_\lambda^{-\frac{1}{2}}\Sigma\Sigma_\lambda^{-\frac{1}{2}}\big\|_{\mathrm{op}} = \lambda_1/(\lambda_1+\lambda) \ge 1/2$ provided $\lambda \le \lambda_1$, ensuring $\mathcal{N}(\lambda) \ge 1/2$, so $d \ge 1$ and we can simplify $\max(d,1) = d$ in (3.15). This also ensures $\log(\mathcal{N}(\lambda)) + L_\delta \ge 1 - \log(2) > 0$ allowing the simplification in (4.10). □

The following corollary of (4.9) is extremely important and useful.

Corollary 4.7. *Under Assumption 4.1, for $\lambda \in (0, \|\Sigma\|_{\mathrm{op}})$, and $\delta \in (0,1)$, provided*

$$n \ge C_\blacktriangle A\frac{\log(\mathcal{N}(\lambda)) + L_\delta}{\lambda} \tag{4.11}$$

for some $A \ge 1$, then with probability at least $(1-\delta)$ it holds simultaneously

$$\Big\|\Sigma_\lambda^{-\frac{1}{2}}\widehat{\Sigma}_\lambda^{\frac{1}{2}}\Big\|_{\mathrm{op}}^2 \le 1 + \frac{C_\blacktriangle}{\sqrt{A}}; \tag{4.12}$$

$$\Big\|\widehat{\Sigma}_\lambda^{-\frac{1}{2}}\Sigma_\lambda^{\frac{1}{2}}\Big\|_{\mathrm{op}}^2 \le 1 + \frac{C_\blacktriangle}{\sqrt{A}}. \tag{4.13}$$

Proof. Provided $C_\blacktriangle$ is chosen large enough in condition (4.11), we have from (4.9) (with probability at least $1-\delta$):

$$\Big\|\Sigma_\lambda^{-\frac{1}{2}}\big(\widehat{\Sigma}-\Sigma\big)\Sigma_\lambda^{-\frac{1}{2}}\Big\|_{\mathrm{op}} \le \frac{C'_\blacktriangle}{\sqrt{A}}.$$

(we can assume $C'_\blacktriangle \le \frac{1}{2}$ provided $C_\blacktriangle$ is chosen large enough in condition (4.11)).
This immediately implies the first claim, since

$$\Big\|\Sigma_\lambda^{-\frac{1}{2}}\widehat{\Sigma}_\lambda^{\frac{1}{2}}\Big\|_{\mathrm{op}}^2 = \Big\|\Sigma_\lambda^{-\frac{1}{2}}\widehat{\Sigma}_\lambda\Sigma_\lambda^{-\frac{1}{2}}\Big\|_{\mathrm{op}} = \Big\|\Sigma_\lambda^{-\frac{1}{2}}(\widehat{\Sigma}-\Sigma)\Sigma_\lambda^{-\frac{1}{2}} + I\Big\|_{\mathrm{op}} \le 1 + \frac{C'_\blacktriangle}{\sqrt{A}}.$$

For the second claim, we have

$$\left\|\widehat{\Sigma}_\lambda^{-\frac{1}{2}}\Sigma_\lambda^{\frac{1}{2}}\right\|_{\mathrm{op}}^2 = \left\|\Sigma_\lambda^{\frac{1}{2}}\widehat{\Sigma}_\lambda^{-1}\Sigma_\lambda^{\frac{1}{2}}\right\|_{\mathrm{op}} = \left\|\left(\Sigma_\lambda^{-\frac{1}{2}}\widehat{\Sigma}_\lambda\Sigma_\lambda^{-\frac{1}{2}}\right)^{-1}\right\|_{\mathrm{op}} = \left\|\left(I + \Sigma_\lambda^{-\frac{1}{2}}(\widehat{\Sigma}-\Sigma)\Sigma_\lambda^{-\frac{1}{2}}\right)^{-1}\right\|_{\mathrm{op}}$$
$$\leq \left(1 - \frac{C'_\blacktriangle}{\sqrt{A}}\right)^{-1} = 1 + \frac{C''_\blacktriangle}{\sqrt{A}},$$

where we have used $C'_\blacktriangle/\sqrt{A} \leq \frac{1}{2}$. The inequality used above $\|(I-B)^{-1}\|_{\mathrm{op}} \leq (1-\|B\|_{\mathrm{op}})^{-1}$ for $\|B\|_{\mathrm{op}} < 1$ can be checked by a variety of means, either by returning to the definition of the operator norm or by using the Neumann operator series $(1+B)^{-1} = \sum_{k\geq 0} B^k$. □

4.3 Analysis of spectral regularization methods

As announced earlier, we will study estimates of the form

$$\widehat{f}_\lambda = F_\lambda(\widehat{\Sigma})\widehat{S}^*\boldsymbol{Y}, \tag{4.14}$$

where $F_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ is a "regularized inverse" function depending on a regularization parameter $\lambda > 0$.

We will study the statistical properties of this type of algorithms under somewhat "generic" conditions for the family F_λ. These conditions are meant to allow for a large variety of different methods and algorithms in practice. We defer precise examples to a later section.

Assumption 4.8. The family of functions $F_\lambda : [0,\kappa^2] \to \mathbb{R}$ defined for $\lambda \in [0,\kappa^2]$, is said to be a regularization (or filter) function of qualification $q > 0$ if there exist positive constants D, E, such that for all $\lambda \in [0,\kappa^2]$ and $t \in [0,\kappa^2]$, it holds:

$$|F_\lambda(t)| \leq E\min(\lambda^{-1}, t^{-1}); \tag{4.15}$$
$$|1 - tF_\lambda(t)| \leq D\left(\frac{\lambda}{t}\right)^q. \tag{4.16}$$

The following useful estimates are direct consequences:

Lemma 4.9. *Under Assumption 4.8, the following holds true for all $\lambda \in [0,\kappa^2]$ and $t \in [0,\kappa^2]$:*

$$\text{for all } \beta \in [0,1]: \qquad |F_\lambda(t)|t^\beta \leq E\lambda^{\beta-1}; \tag{4.17}$$
$$\text{for all } \gamma \in [0,q]: \qquad |1 - tF_\lambda(t)|t^\gamma \leq D'\lambda^\gamma, \tag{4.18}$$

where $D' = \max(D, 1+E)$.

Proof. For any $\lambda \in [0,\kappa^2]$ and $t \in [0,\kappa^2]$, and $\beta \in [0,1]$, it holds using (4.15):

$$|F_\lambda(t)|t^\beta = |F_\lambda(t)|^{1-\beta}|tF_\lambda(t)|^\beta \leq (E/\lambda)^{1-\beta}E^\beta \leq E\lambda^{\beta-1}.$$

Furthermore, for any $\gamma \in [0, q]$, using (4.15) and (4.16):

$$|1 - tF_\lambda(t)|t^\gamma = (|1 - tF_\lambda(t)|t^q)^{\frac{\gamma}{q}}|1 - tF_\lambda(t)|^{1-\frac{\gamma}{q}} \leq D^{\frac{\gamma}{q}}(1+E)^{1-\frac{\gamma}{q}}\lambda^\gamma.$$

□

The second type of assumption we will make concerns the "regularity" of the target function f^*, expressed in the "scale" of the second moment operator Σ.

Assumption 4.10. Under the notation of Assumption 4.1, we say that the target f^* has a *Hölder source regularity condition* of order $r \geq 0$ if it can be written under the form

$$f^* = \Sigma^r g_0, \tag{4.19}$$

for some $g_0 \in \mathcal{H}$.

Observe that since Σ is not invertible, the image of Σ is not $\mathcal{H}$, and thus this condition is not trivial; not every element of $\mathcal{H}$ has a non-trivial ($r > 0$) source regularity condition. The higher r, the most restrictive the condition, and hence the higher "regularity" the target function has. Note that there exist more general source conditions in the literature that use functions of Σ different from (fractional) powers, but the Hölder source condition (using powers of Σ) is the most classical, and is the only type we will consider.

From now on, the "generic number" $C_\blacktriangle$ will be allowed to depend on κ, M, the constants (E, D) from Assumption 4.8 as well as $(r, \|g_0\|)$ from Assumption 4.10. In fact, it is probably more enlightening to say that $C_\blacktriangle$ indicates a factor that does *not* depend on n, λ or δ. Remember that the value of $C_\blacktriangle$ might change from one line to the other.

Proposition 4.11. *Suppose granted Assumption 4.1, regularization function Assumption 4.8 with qualification q, and Hölder source Assumption 4.19 of order r, such that $q \geq r + \frac{1}{2}$.*
For any $\lambda \in [0, \|\Sigma\|_{\mathrm{op}}]$, $\delta \in (0,1)$ and n such that condition (4.11) holds, it holds with probability at least $1 - \delta$:

$$\left\|\Sigma^{\frac{1}{2}}(\widehat{f}_\lambda - f^*)\right\|_{\mathcal{H}} \leq C_\blacktriangle \left(\sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \lambda^{r+\frac{1}{2}}\right)\sqrt{L_\delta}. \tag{4.20}$$

Corollary 4.12. *Under the same assumptions as Proposition 4.11, assume additionally that the ordered eigenvalues of Σ satisfy*

$$\lambda_k(\Sigma) \leq ck^{-\alpha}, \tag{4.21}$$

for some constants $c > 0$ and $\alpha > 1$. Put $\beta := \alpha\big(r + \frac{1}{2}\big)$. Then, choosing the regularization constant

$$\lambda_n = n^{-\frac{\alpha}{2\beta+1}}, \tag{4.22}$$

for fixed $\delta \in (0,1)$, for n big enough it holds with probability at least $1 - \delta$:

$$R(\widehat{f}_{\lambda_n}) - R(f^*) \leq C_\blacktriangle n^{-\frac{2\beta}{2\beta+1}}, \tag{4.23}$$

where $C_\blacktriangle$ depends on (c, α) in addition to the constants appearing in the other assumptions.

Proof. Let us derive a rough estimate of the effective dimension $\mathcal{N}(\lambda)$ in this case. Denote $k_\lambda^* = \min_{k\geq 1:\lambda_k\leq\lambda}$. Then $\lambda_k/(\lambda_k+\lambda) \leq \frac{1}{2}$ for $k \leq k_\lambda^*$; and the assumption (4.21) implies that $k_\lambda^* \leq (\lambda/c)^{-\alpha^{-1}}$. Thus

$$\begin{aligned}
\mathcal{N}(\lambda) = \sum_{k\geq 1} \frac{\lambda_k}{\lambda+\lambda_k} &= \sum_{1\leq k<k_\lambda^*} \frac{\lambda_k}{\lambda_k+\lambda} + \sum_{k\geq k_\lambda^*} \frac{\lambda_k}{\lambda_k+\lambda} \\
&\leq \frac{1}{2}k_\lambda^* + \lambda^{-1}c\sum_{k\geq k_\lambda^*} k^{-\alpha} \\
&\leq \frac{1}{2}k_\lambda^* + c\lambda^{-1}\int_{t\geq k_\lambda^*} t^{-\alpha}dt \\
&\leq \frac{1}{2}k_\lambda^* + \frac{1}{\alpha-1}\lambda^{-1}(k_\lambda^*)^{1-\alpha} \\
&\leq C(c,\alpha)\lambda^{-\alpha^{-1}}.
\end{aligned}$$

It can then be checked that the choice (4.22) for λ_n, which balances the two terms for the obtained risk bound (4.20), leads to (4.23). □

Proof of Prop. 4.11. We will assume for this proof that the probabilistic inequalities of Proposition 4.6 are satisfied, as well as those of Corollary 4.7. By the assumptions made, the required conditions for Corollary 4.7, namely $\lambda \leq \|\Sigma\|_{\mathrm{op}} \leq \kappa^2$ and (4.11) are satisfied. Note that we also implicitly use a union bound to get simultaneously the controls of Proposition 4.6, but this amounts to replace L_δ by $L_{\delta/c} \leq C_\blacktriangle L_\delta$ for a finite number of events c to apply the union bound over, and this can be included in the numerical constants.

We recall the starting decomposition 4.3 coming from model (4.1) and the definition of the estimator:

$$\widehat{f}_\lambda = F_\lambda(\widehat{\Sigma})\widehat{S}^*(\widehat{S}f^* + \boldsymbol{\xi}) = F_\lambda(\widehat{\Sigma})\widehat{\Sigma}f^* + F_\lambda(\widehat{\Sigma})(\widehat{S}^*\boldsymbol{\xi}), \tag{4.24}$$

hence the quantity we want to analyze for the control of the excess risk (4.2) is

$$\Sigma^{\frac{1}{2}}(\widehat{f}_\lambda - f^*) = \Sigma^{\frac{1}{2}}(F_\lambda(\widehat{\Sigma})\widehat{\Sigma} - I)f^* + \Sigma^{\frac{1}{2}}F_\lambda(\widehat{\Sigma})(\widehat{S}^*\boldsymbol{\xi}). \tag{4.25}$$

We will control the two terms above, starting with the second one, "noise", term. It holds

$$\left\|\Sigma^{\frac{1}{2}}F_\lambda(\widehat{\Sigma})(\widehat{S}^*\boldsymbol{\xi})\right\| \leq \left\|\Sigma^{\frac{1}{2}}\Sigma_\lambda^{-\frac{1}{2}}\right\|_{\mathrm{op}}\left\|\Sigma_\lambda^{\frac{1}{2}}\widehat{\Sigma}_\lambda^{-\frac{1}{2}}\right\|_{\mathrm{op}}\left\|\widehat{\Sigma}_\lambda^{\frac{1}{2}}F_\lambda(\widehat{\Sigma})\widehat{\Sigma}_\lambda^{\frac{1}{2}}\right\|_{\mathrm{op}}\left\|\widehat{\Sigma}_\lambda^{-\frac{1}{2}}\Sigma_\lambda^{\frac{1}{2}}\right\|_{\mathrm{op}}\left\|\Sigma_\lambda^{-\frac{1}{2}}\widehat{S}^*\boldsymbol{\xi}\right\|.$$

The first factor is bounded by 1. The second and fourth factors are bounded by a number $C_\blacktriangle$ (with high probability) due to Corollary 4.7. The last factor is bounded using (4.7). As for the the third factor, it holds

$$\begin{aligned}
\left\|\widehat{\Sigma}_\lambda^{\frac{1}{2}}F_\lambda(\widehat{\Sigma})\widehat{\Sigma}_\lambda^{\frac{1}{2}}\right\|_{\mathrm{op}} &\leq \sup_{t\in[0,\kappa^2]}|F_\lambda(t)|(t+\lambda) \\
&\leq 2E,
\end{aligned}$$

using (4.15). In the end, we get

$$\left\|\Sigma^{\frac{1}{2}} F_\lambda(\widehat{\Sigma})(\widehat{S}^*\boldsymbol{\xi})\right\| \le C_\blacktriangle\left(\sqrt{\frac{L_\delta \mathcal{N}(\lambda)}{n}} + \frac{L_\delta}{n\sqrt{\lambda}}\right) \le C'_\blacktriangle\sqrt{\frac{L_\delta\mathcal{N}(\lambda)}{n}}. \tag{4.26}$$

For the last inequality, we used condition (4.11) which implies (noting that $\lambda \le \|\Sigma\|_{\mathrm{op}}$ implies $\mathcal{N}(\lambda) \ge \frac{1}{2}$):

$$n \ge C_\blacktriangle \frac{\log(2\mathcal{N}(\lambda)) + L_\delta}{\lambda} \ge C_\blacktriangle \frac{L_\delta}{\lambda},$$

so that

$$\frac{L_\delta}{n\sqrt{\lambda}} \le C_\blacktriangle \frac{\sqrt{L_\delta}}{\sqrt{n}} \le C'_\blacktriangle \sqrt{\frac{L_\delta \mathcal{N}(\delta)}{n}}.$$

Let us turn to the first, "approximation", term in (4.25). We use the assumed source condition (4.19) and start similarly as above; we denote $R_\lambda(t) := (tF_\lambda(t) - 1)$:

$$\left\|\Sigma^{\frac{1}{2}}(F_\lambda(\widehat{\Sigma})\widehat{\Sigma} - I)f^*\right\| \le \left\|\Sigma^{\frac{1}{2}}\Sigma_\lambda^{-\frac{1}{2}}\right\|_{\mathrm{op}}\left\|\Sigma_\lambda^{\frac{1}{2}}\widehat{\Sigma}_\lambda^{-\frac{1}{2}}\right\|_{\mathrm{op}}\left\|\widehat{\Sigma}_\lambda^{\frac{1}{2}} R_\lambda(\widehat{\Sigma})\Sigma^r\right\|\|g_0\| \le C_\blacktriangle\left\|\widehat{\Sigma}^{\frac{1}{2}} R_\lambda(\widehat{\Sigma})\Sigma^r\right\|_{\mathrm{op}}. \tag{4.27}$$

We will distinguish two cases: first, if $r \le \frac{1}{2}$, we will use the Cordes inequality (Proposition 2.35), namely $\|A^sB^s\|_{\mathrm{op}} \le \|AB\|_{\mathrm{op}}^s$ if A,B, are self-adjoint and $s \in [0,1]$ to obtain

$$\left\|\widehat{\Sigma}_\lambda^{\frac{1}{2}} R_\lambda(\widehat{\Sigma})\Sigma^r\right\|_{\mathrm{op}} \le \left\|\widehat{\Sigma}_\lambda^{\frac{1}{2}} R_\lambda(\widehat{\Sigma})\widehat{\Sigma}_\lambda^r\right\|_{\mathrm{op}}\left\|\widehat{\Sigma}_\lambda^{-\frac{1}{2}}\Sigma_\lambda^{\frac{1}{2}}\right\|_{\mathrm{op}}^{2r}\left\|\Sigma_\lambda^{-r}\Sigma^r\right\|_{\mathrm{op}}$$

The last factor is bounded by 1 as before, the second by a number $C_\blacktriangle$ (with high probability) due to Corollary 4.7, and the first by

$$\begin{aligned}\left\|\widehat{\Sigma}_\lambda^{\frac{1}{2}} R_\lambda(\widehat{\Sigma})\widehat{\Sigma}_\lambda^r\right\|_{\mathrm{op}} &\le \sup_{t\in[0,\kappa^2]}\left(R_\lambda(t)(t+\lambda)^{r+\frac{1}{2}}\right)\\ &\le \sup_{t\in[0,\kappa^2]}\left(R_\lambda(t)(t^{r+\frac{1}{2}} + \lambda^{r+\frac{1}{2}})\right)\\ &\le 2D'\lambda^{r+\frac{1}{2}},\end{aligned}$$

where we have used $r + \frac{1}{2} \le 1$ for the second inequality, and property (4.18) for the last (under the assumption $r + \frac{1}{2} \le q$).

If $r \ge \frac{1}{2}$, we modify the argument:

$$\begin{aligned}\left\|\widehat{\Sigma}_\lambda^{\frac{1}{2}} R_\lambda(\widehat{\Sigma})\Sigma^r\right\|_{\mathrm{op}} &\le \left\|\widehat{\Sigma}_\lambda^{\frac{1}{2}} R_\lambda(\widehat{\Sigma})\Sigma_\lambda^r\right\|_{\mathrm{op}}\underbrace{\left\|\Sigma_\lambda^{-r}\Sigma^r\right\|}_{\le 1}\\ &\le \left\|\widehat{\Sigma}_\lambda^{\frac{1}{2}} R_\lambda(\widehat{\Sigma})\widehat{\Sigma}_\lambda^r\right\|_{\mathrm{op}} + \left\|\widehat{\Sigma}_\lambda^{\frac{1}{2}} R_\lambda(\widehat{\Sigma})\right\|_{\mathrm{op}}\left\|\Sigma_\lambda^r - \widehat{\Sigma}_\lambda^r\right\|_{\mathrm{op}}\\ &\le 2D'\lambda^{r+\frac{1}{2}} + 2D'\lambda^{\frac{1}{2}}\frac{C_\blacktriangle r(2\kappa)^{2(r-1)_+}(2\kappa^2)\sqrt{L_\delta}}{\lambda^{(1-r)_+}\sqrt{n}}.\end{aligned}$$

To justify the last inequality, we use the HS-norm control (4.5) together with the Hilbert-Schmidt Lipschitz perturbation inequality (2.15), for the function $\varphi_r : x \mapsto x^r$ on the interval $[\lambda, 2\kappa^2]$ containing the spectrum of both Σ_λ and $\widehat{\Sigma}_\lambda$ (remember we assume $\lambda \leq \kappa^2$). On this interval the function φ_r is $r\lambda^{r-1}$-Lipschitz if $r \leq 1$, and $r(2\kappa)^{r-1}$-Lipschitz if $r \geq 1$. Summing up the last computations into (4.27) and wrapping various factors into the generic constant, we get

$$\left\|\Sigma^{\frac{1}{2}}(F_\lambda(\widehat{\Sigma})\widehat{\Sigma} - I)f^*\right\| \leq C_\blacktriangle\left(\lambda^{r+\frac{1}{2}} + \frac{\mathbf{1}\{r \geq \frac{1}{2}\}\sqrt{\lambda}\sqrt{L_\delta}}{\lambda^{(1-r)_+}\sqrt{n}}\right). \tag{4.28}$$

Plugging in (4.26) and (4.28) into (4.25), we thus obtain the risk bound holding with high probability (using $L_\delta \geq 1$ to pull it out as a factor up to changes in the front constant):

$$\left\|\Sigma^{\frac{1}{2}}(\widehat{f} - f^*)\right\|_{\mathcal{H}} \leq C_\blacktriangle\left(\sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \lambda^{r+\frac{1}{2}} + \frac{\mathbf{1}\{r \geq \frac{1}{2}\}\sqrt{\lambda}}{\lambda^{(1-r)_+}\sqrt{n}}\right)\sqrt{L_\delta}. \tag{4.29}$$

Let us finally clean up the above expression by noticing that the third term is upper bounded by the first up to a $C_\blacktriangle$-factor, since, for $r \geq \frac{1}{2}$:

$$\frac{\sqrt{\lambda}}{\lambda^{(1-r)_+}} = \lambda^{\min\left(\frac{1}{2}, r-\frac{1}{2}\right)} \leq \max(1, \kappa) \leq \max(1, \kappa)\sqrt{2\mathcal{N}(\lambda)}.$$

This finally implies the announced estimate (4.20). □

4.4 Examples

In this section we give a few examples of classical regularization functions and check that they satisfy the conditions of Assumption 4.8. Most of these examples come from the theory of inverse problems [10].

Spectral cut-off. The spectral cut-off (or truncated singular value decomposition, TSVD) regularization function is given by $F_\lambda(t) = \mathbf{1}\{t \geq \lambda\}/t$. In words, once applied to a self-adjoint operator, this regularization function projects this operator onto the sum of eigenspaces for eigenvalues less than λ, and takes its Moore-Penrose pseudoinverse. It is immediate to check that $F_\lambda(t) \leq t^{-1}$, $F_\lambda(t) \leq \lambda^{-1}$, and $|1 - tF_\lambda(t)| = \mathbf{1}\{t < \lambda\} \leq (\lambda/t)^q$, for any $q \geq 0$. Therefore, the conditions of Assumption 4.8 are satisfied for $E = D = 1$ and any $q > 0$ we can say that this regularization has "infinite qualification".

Having infinite qualification sounds like a very desirable property, since it can adapt to an arbitrarily regular source condition. However, eigendecomposition truncation is difficult in practice since it requires to compute the eigendecomposition of $\widehat{\Sigma}$. Furthermore, in practice somewhat more "smooth" regularization functions turn out to have better behavior.

Ridge regression/Tikhonov regularization. The ridge regression regularizer, also known as Tikhonov regularization, is given by $F_\lambda(t) = (t + \lambda)^{-1}$. It is easy to check that:

$$\frac{1}{\lambda + t} \leq \min\left(\frac{1}{\lambda}, \frac{1}{t}\right); \qquad \left|1 - \frac{t}{\lambda + t}\right| = \frac{\lambda}{\lambda + t} \leq \frac{\lambda}{t},$$

hence the conditions of Assumption 4.8 are satisfied for $E = D = 1$ and qualification $q = 1$. On the other hand, it can be checked that qualification higher than 1 does not hold.

Iterated ridge regression/Tikhonov. To compensate for the limited qualification of the standard ridge regression, it can be proposed to iterate it by applying it (with the the same regularization parameter λ) recursively to the residuals. The following formulas can be easily shown by recursion for m-times iteration:

$$F_\lambda^{(m)}(t) = \sum_{i=1}^{m} \frac{\lambda^{i-1}}{(\lambda+t)^i} = \frac{1}{t}\left(1 - \frac{\lambda^m}{(\lambda+t)^m}\right)$$
$$(\text{residuals}) R_\lambda^{(m)}(t) = 1 - tF_\lambda^{(m)}(t) = \frac{\lambda^m}{(\lambda+t)^m}.$$

It is easy to check that:

$$\sum_{i=1}^{m} \frac{\lambda^{i-1}}{(\lambda+t)^i} \le \frac{m}{\lambda+t} \le m \min\left(\frac{1}{\lambda}, \frac{1}{t}\right); \qquad \left(\frac{\lambda}{\lambda+t}\right)^m \le \left(\frac{\lambda}{t}\right)^m,$$

hence the conditions of Assumption 4.8 are satisfied for $E = m$, $D = 1$ and qualification $q = m$.

Gradient descent/Landweber iteration. Consider the gradient method based on the quadratic loss objective function

$$\widehat{L}(f) = \frac{1}{n}\sum_{i=1}^{n}(\langle f, X_i\rangle - Y_i)^2 = \left\|\widehat{S}f - \boldsymbol{Y}\right\|_n^2,$$

with the gradient

$$\nabla_f \widehat{L} = 2\widehat{S}^*(\widehat{S}f - \boldsymbol{Y}) = 2\Big(\widehat{\Sigma}f - \widehat{S}^*\boldsymbol{Y}\Big).$$

Thus, if the estimate after k gradient iterations (with fixed stepsize $\eta/2$) has the form $\widehat{f}_k = F_k(\widehat{\Sigma})\widehat{S}^*\boldsymbol{Y}$, the next gradient descent step is $-\eta(\widehat{\Sigma}F_k(\widehat{\Sigma}) - I)\widehat{S}^*\boldsymbol{Y}$. Therefore, the k-th step of gradient descent take the form of a regularization function $F_k(t)$ satisfying the recursion

$$F_{k+1}(t) = F_k(t) + \eta(1 - tF_k(t)) = \eta + (1 - \eta t)F_k(t),$$

after unfolding the recursion we get

$$F_k(t) = \eta\sum_{\ell=0}^{k-1}(1-\eta t)^\ell = \frac{1}{t}\big(1 - (1-\eta t)^k\big).$$

If $t \in [0, \kappa^2]$ and $\eta \in (0, \kappa^{-2})$ then $\eta t < 1$ and we have $F_k(t) \le 1/t$ and, using $(1-x)^k \ge 1 - kx$ for $x \le 1$ by convexity,

$$F_k(t) \le \frac{1}{t}k\eta t \le k\eta.$$

Thus, if we define the equivalent regularization parameter $\lambda_k := (\eta k)^{-1}$, the first part of Assumption 4.8 is satisfied (for $E = 1$) and it holds for any $q > 0$:

$$|1 - tF_k(t)| = (1 - \eta t)^k \leq \exp(-k\eta t) \leq c_q \left(\frac{1}{k\eta t}\right)^q = c_q \left(\frac{\lambda}{t}\right)^q,$$

where $c_q = (q/e)^q$. For the last inequality we used the elementary fact that $\max_{u>0}\left(\log u - \frac{u}{q}\right) = \log q - 1$, so that $\exp(-u) \leq (q/e)^q u^{-q}$. Hence the second part of Assumption 4.8 is satisfied for any $q > 0$ with $D = c_q$.

To summarize, gradient descent (for the quadratic risk) with fixed stepsize acts a regulazation if is *stopped early* at step k, provided the stopping time is chosen in accordance with the target function's source regularity.

5 Reproducing kernel methods

5.1 Reproducing kernel Hilbert spaces

We quickly review some important facts about the reproducing kernel Hilbert space methodology in data science. Initially considered in the context of spline methods in the 1970s [33], for which what we now call "Kernel ridge regression" was already introduced, they enjoyed an important resurgence in the 2000s due to their versatility for applications of machine learning methods, in particular Support Vector Machines. While they have been outperformed by modern Deep Learning methods, they remain an important tool to understand and analyze machine learning methods (see e.g. works on the "Neural Tangent Kernel" [15])

A (reproducing) kernel Hilbert space (rkHs) can be defined in several equivalent ways (see [2], Chap. 5, for instance).

Definition 5.1 (kernel Hilbert space, abstract version)**.** Given a base space $\mathcal{X}$, a kernel Hilbert space over $\mathcal{X}$ is a $\mathbb{R}$- or $\mathbb{C}$-Hilbert space together with a "feature" mapping $\Phi : \mathcal{X} \to \mathcal{H}$. The associated *kernel* is the function k defined as

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \text{ or } \mathbb{C} \ : \qquad (x, y) \mapsto \langle \Phi(x), \Phi(y) \rangle. \tag{5.1}$$

Proposition/Definition 5.2 (rkHs, functional space version)**.** Given a base space $\mathcal{X}$, a reproducing kernel Hilbert space over $\mathcal{X}$ is $\mathbb{R}$- or $\mathbb{C}$-Hilbert space whose elements are $\mathbb{R}$- or $\mathbb{C}$-valued functions on $\mathcal{X}$, together with a "feature" mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such that the following property holds:

$$\forall f \in \mathcal{H}, x \in \mathcal{X} : \qquad f(x) = \langle f, \Phi(x) \rangle. \tag{5.2}$$

The associated *kernel* is the function k defined as (5.1).
As a consequence of (5.2), it can be checked that Φ must in fact be the mapping

$$\forall x \in \mathcal{X} : \qquad \Phi(x) = k(x, \cdot) = (y \mapsto k(x, y)), \tag{5.3}$$

implying in particular that all functions $k(x, \cdot)$ must belong to $\mathcal{H}$.

The proof of (5.3) is left to the reader. The property (5.2) is called "self-reproducing property" because once applied to the functions $\Phi(x) = k(x, \cdot)$ and $\Phi(y) = k(y, \cdot)$ it yields

$$\langle k(x, \cdot), k(y, \cdot) \rangle = k(x, y).$$

In this sense the term "reproducing" only makes sense in the functional space version, thus when referring to a rKHs one always implicitly assumes the functional space.

The following equivalent definition of the functional space version is sometimes useful:

Property 5.3. Let $\mathcal{H}$ be a $\mathcal{X}$ is $\mathbb{R}$- or $\mathbb{C}$-Hilbert space whose elements are $\mathbb{R}$- or $\mathbb{C}$-valued functions on $\mathcal{X}$, and such that for any $x \in \mathcal{X}$, the (linear) evaluation functional at point x: $f \in \mathcal{H} \mapsto f(x)$ is continuous. Then $\mathcal{H}$ is a rkHs (functional space version).

Proof. By Riesz' theorem, since for any $x \in \mathcal{X}$ the evaluation mapping $f \in \mathcal{H} \mapsto f(x)$ is continuous, there exists an element $\Phi(x)$ such that $f(x) = \langle f, \Phi(x) \rangle$, i.e. (5.2) is satisfied. □

Obviously, a rkHs (functional space version) is a rkHs (abstract version). However, the functional space version is canonical in the sense given by the following theorem.

Theorem 5.4 (Characterization theorem)**.** *If $\mathcal{H}$ is a rkHs over $\mathcal{X}$, the associated kernel function k is of positive semidefinite (psd) type, meaning that for any $n \in \mathbb{N}$, $(x_1, \ldots, x_n) \in \mathcal{H}^n$, the "kernel Gram matrix" G given by $G_{ij} = k(x_i, x_j)$ is Hermitian (=self-adjoint).*

Conversely, if k is a kernel function of psd type on $\mathcal{X}$, then there exists a rkHs (functional version) $\mathcal{H}_k$ on $\mathcal{X}$ with kernel k. It is the completion of the pre-Hilbert space of functions $\mathcal{H}_{\text{pre}} = \text{Span}\{k(x, .), x \in \mathcal{X}\}$ with the inner product

$$\left\langle \sum_i \alpha_i k(x_i, .), \sum_j \beta_j k(x_j, .) \right\rangle = \sum_{i,j} \alpha_i \bar{\beta}_j k(x_i, x_j).$$

This Hilbert space $\mathcal{H}_k$ is the canonical (functional space) rkHs associated with the psd kernel k. For any rkHs (abstract version) $\mathcal{H}_\circ$ on $\mathcal{X}$ with feature mapping $\Phi_\circ$ and kernel k, the mapping $\xi : u \mapsto (x \mapsto \langle u, \Phi_\circ(x) \rangle_\circ)$ maps $\mathcal{H}_\circ$ onto $\mathcal{H}_k$ and satisfies $\xi \circ \Phi_\circ(x) = k(x, \cdot)$ i.e. $\xi \circ \Phi_\circ$ is the canonical feature mapping from $\mathcal{X}$ to $\mathcal{H}_k$.

For a proof see e.g. [30, 2].

Observe that if $\mathcal{H}$ is a Hilbert space, we can see its dual $\mathcal{H}^*$ as a rkHs on $\mathcal{H}$, with the canonical mapping $f \in \mathcal{H} \mapsto f^* \in \mathcal{H}^*$ and the "linear" kernel $k(f, g) = \langle f, g \rangle$. This somewhat convoluted way to present a Hilbert space is merely to remark that the case of the linear regression model with covariate in a Hilbert space (4.1) considered in the previous chapter can be cast into the framework considered here.

To come back to our purposes, given a rkHs $\mathcal{H}$ with kernel k, feature mapping Φ on $\mathcal{X}$, and data $(X_i, Y_i)_{i \in [\![n]\!]}$ taking values in $\mathcal{X} \times \mathbb{R}$, we will intend to apply the results of the previous chapter to the mapped data $(\widetilde{X}_i, Y_i) \in \mathcal{H} \times \mathbb{R}$, where $\widetilde{X} = \Phi(X)$.

The following property is immediate and useful: it relates boundedness of the kernel to boundedness of the Hilbert-valued covariate $\widetilde{X}$:

Lemma 5.5. *If k is a psd kernel on a space $\mathcal{X}$, and if $\sup_{x \in \mathcal{X}} k(x, x) = \kappa^2 < \infty$, then for any rkHs with kernel k over $\mathcal{X}$ with feature mapping Φ, it holds $\|\Phi(x)\| \leq \kappa$ for any $x \in \mathcal{X}$. Furthermore, it holds $|k(x, y)| \leq \kappa^2$ for any $(x, y) \in \mathcal{X}^2$.*

The proof is almost tautological and left to the reader (for the second part, use the Cauchy-Schwarz inequality).

Measurability assumptions.

Up until this point there was no structural assumption on the space $\mathcal{X}$ whatsoever. Since we'll consider probability distributions on $\mathcal{X}$, we will assume from now on that $\mathcal{X}$ is a measurable space. We'll need basic measurability properties to ensure that expectations involving the kernel are well-defined. In the sequel we will always make the following assumption without futher mention:

Assumption 5.6 (Assumption [**M**]). The base space $\mathcal{X}$ is a measurable space, $\mathcal{H}$ is a **separable** rkHs with kernel k on $\mathcal{X}$, and for any $y \in \mathcal{X}$, the function $x \mapsto k(x, y)$ is measurable $\mathcal{X} \to \mathbb{R}$ or $\mathbb{C}$.

We have the following consequences:

Proposition 5.7. *Under Assumption* [**M**]*, it holds that:*

1. *Every function $f \in \mathcal{H}$ is measurable.*
2. *The mapping $x \mapsto \Phi(x) = k(x, \cdot) \in \mathcal{H}$ is measurable.*

Proof. For the first point, we can use that $\mathcal{H}$ is the closure of $\mathcal{H}_{pre} := \mathrm{Span}\{k(x, \cdot), x \in \mathcal{X}\}$, which entails that any function in $\mathcal{H}$ is a pointwise limit of elements of $\mathcal{H}_{pre}$, left as an exercise (or see [2], Thm. 5.10). Since the functions $k(x, \cdot)$ are measurable, so are their linear combinations, and pointwise limits thereof.

Thus, for any $f \in \mathcal{H}$, the function $x \mapsto \langle f, \Phi(x)\rangle = f(x)$ is measurable. This means that the mapping $\Phi : \mathcal{X} \to \mathcal{H}$ is *weakly measurable*. By Pettis' theorem [insert ref here!], because $\mathcal{H}$ is *separable*, it entails that Φ is measurable. □

Most often, $\mathcal{X}$ will be a topological space with its Borel sigma-algebra. It is not completely obvious to ensure that $\mathcal{H}$ is separable in general, but it is granted if $\mathcal{X}$ is a second-countable topological space and k is continuous [add a ref here].

5.2 Kernel operators in reproducing kernel Hilbert spaces

As a first step, we generalize the model and different operators introduced in the previous chapter in the framework of data mapped into a (functional space) rkHs. For this

Proposition 5.8. *Let $\mathcal{H}$ be a rkHs over $\mathcal{X}$ with kernel k and canonical feature mapping $\Phi(x) = k(x, \cdot)$. Let ρ be a probability distribution over $\mathcal{X}$ such that $\mathbb{E}_\rho[k(X, X)] < \infty$ (for instance, this is the case under the assumption $\sup_{x\in\mathcal{X}} k(x, x) = \kappa^2 < \infty$).*

Then we generalize the different operators appearing in Propositions 4.2 and 4.3 as follows:

- *The population evaluation operator S maps an element $f \in \mathcal{H}$ to itself, as an element of $L^2(\mathcal{X}, \rho)$. (In this sense it is a "change of geometry" operator). It is a Hilbert-Schmidt operator.*
- *The adjoint S^* of S is given by the kernel integral operator*

$$f \in L^2(\mathcal{X}, \rho) \mapsto \left(t \mapsto \int_{\mathcal{X}} k(x, t) f(x) d\rho(x) = \mathbb{E}_\rho[f(X)k(X, t)]\right) \in \mathcal{H}. \tag{5.4}$$

- *The operator $S^*S = \mathbb{E}[k(x, \cdot) \otimes k(x, \cdot)^*]$, which is a trace-class operator, is also given by the kernel integral operator* (5.4)*, but as an operator from $\mathcal{H}$ to $\mathcal{H}$; the operator SS^* is again given by the same formula, but as an operator from $L^2(\mathcal{X}, \rho)$ to itself.*

- *The sample evaluation operator $\widehat{S}$ is given by*

$$\widehat{S}: \quad \mathcal{H} \to (\mathbb{R}^n, \langle \cdot, \cdot \rangle_n), \qquad f \mapsto (f(X_1), \dots, f(X_n)).$$

- *Its adjoint is given by*

$$\widehat{S}^*: \quad (\mathbb{R}^n, \langle \cdot, \cdot \rangle_n) \to \mathcal{H}, \qquad (a_1, \dots, a_n) \mapsto \frac{1}{n} \sum_{i \in [\![n]\!]} a_i k(X_i, \cdot). \tag{5.5}$$

- *It holds $\widehat{S}^* \widehat{S} = \widehat{\Sigma} = \frac{1}{n} \sum_{k=1}^n k(X_i, \cdot) \otimes k(X_i, \cdot)^*$, while $\widehat{S} \widehat{S}^* : \mathbb{R}^n \to \mathbb{R}^n$ is the normalized Gram kernel matrix $\widehat{G}$ given by*

$$\widehat{G}_{i,j} = \frac{1}{n} k(X_i, X_j), \qquad (i,j) \in [\![n]\!]^2.$$

Proof. For the first point, we must check that S is a well-defined, Hilbert-Schmidt operator. First, under the assumption, any element $f \in \mathcal{H}$ is (as a function over $\mathcal{X}$) squared-integrable with respect to ρ, since by the reproducing property and Cauchy-Schwarz:

$$\mathbb{E}_\rho\big[|f(x)|^2\big] = \mathbb{E}_\rho\big[|\langle f, k(x,\cdot)\rangle_{\mathcal{H}}|^2\big] \leq \mathbb{E}_\rho\big[\|f\|_{\mathcal{H}}^2 \|k(x,\cdot)\|_{\mathcal{H}}^2\big] = \|f\|_{\mathcal{H}}^2 \mathbb{E}_\rho[k(x,x)] < \infty.$$

If $(e_i)_{i \in I}$ is a basis of $\mathcal{H}$, then

$$\sum_{i \in I} \|Se_i\|_{L^2(\mathcal{X},\rho)}^2 = \sum_{i \in I} \mathbb{E}_\rho\big[|e_i(x)|^2\big] = \sum_{i \in I} \mathbb{E}_\rho\big[|\langle e_i, k(x,\cdot)\rangle_{\mathcal{H}}|^2\big] = \mathbb{E}_\rho\big[\|k(x,\cdot)\|^2\big] = \mathbb{E}_\rho[k(x,x)] < \infty,$$

establishing that S is Hilbert-Schmidt.

For the second point, we first check that for $f \in L^2(\mathcal{X}, \rho)$, the variable $Z(X) = f(X)k(X,\cdot)$ is Bochner-integrable in $\mathcal{H}$ (it does take its values in $\mathcal{H}$, since for fixed X it is a multiple of $k(X,\cdot) \in \mathcal{H}$.) We have by the Cauchy-Schwarz inequality:

$$\begin{aligned} \mathbb{E}_\rho[\|f(X)k(X,\cdot)\|] = \mathbb{E}_\rho[|f(X)|\|k(X,\cdot)\|] &\leq \big(\mathbb{E}_\rho\big[|f(X)|^2\big]\mathbb{E}_\rho\big[\|k(X,\cdot)\|^2\big]\big)^{\frac{1}{2}} \\ &= \|f\|_{L^2(\mathcal{X},\rho)} \mathbb{E}_\rho[k(X,X)]^{\frac{1}{2}} < \infty, \end{aligned}$$

establishing that $Z(X)$ is Bochner-integrable; we can then write:

$$\langle Sf, g \rangle_{L^2(\mathcal{X},\rho)} = \mathbb{E}_\rho[f(X)\overline{g}(X)] = \mathbb{E}_\rho[\langle f, k(X,\cdot)\rangle_{\mathcal{H}} \overline{g}(X)] = \langle f, \mathbb{E}_\rho[k(X,\cdot)g(X)]\rangle_{\mathcal{H}},$$

leading to the announced formula for S^*. The rest is left to the reader. □

Again, the "abstract" setting of the previous chapter can be recovered if we assume directly random data taking values in a Hilbert space $\mathcal{H}$, linear kernel and the rkHs given by the dual $\mathcal{H}^*$.

Conversely, if we have a rkHs over $\mathcal{X}$ with a feature mapping $\widetilde{X} = \Phi(X)$, we can "forget" the original covariate space $\mathcal{X}$ and its rkHs and see the problem in terms only in

terms of $\widetilde{X}$ and the setting of the previous chapter. From the point of view of the statistical analysis, it does not change the arguments. But the rkHs view is richer in the sense that it describes the model in terms of functions the original covariate X which is more concrete than its mapped version $\widetilde{X}$.

In particular, the linear regression in Hilbert space model (4.1) becomes, in the rkHs setting and in terms of the original covariate $X \in \mathcal{X}$:

$$Y = f^*(X) + \xi, \tag{5.6}$$

where f^* is an element of the rkHs $\mathcal{H}$.

5.3 Spectral regularization in a rkHs regression setting

Let us consider the model (5.6), which is (4.1) "incarnated" in a rkHs setting. Remember we analyzed last chapter spectral regularization methods of the form 4.14

$$\widehat{f_\lambda} = F_\lambda(\widehat{\Sigma})\widehat{S}^*\boldsymbol{Y}. \tag{5.7}$$

Computing this "abstract" estimator seems impossible in practice since we apparently need to manipulate infinite-dimensional vectors and operators in a Hilbert space. However, thanks to the shift formula (2.8) we can rewrite the above as

$$\widehat{f_\lambda} = F_\lambda(\widehat{S}^*\widehat{S})\widehat{S}^*\boldsymbol{Y} = \widehat{S}^*F_\lambda(\widehat{S}^*\widehat{S})\boldsymbol{Y} = \widehat{S}^*F_\lambda(\widehat{G})\boldsymbol{Y}. \tag{5.8}$$

The benefit of this rewriting is that since $\widehat{G}$ is a finite $n \times n$ matrix, and $\boldsymbol{Y}$ an n-dimensional vector, we can (at least in principle) numerically compute the n-vector of coefficients $\widehat{\alpha} = F_\lambda(\widehat{G})\boldsymbol{Y}$. Using (5.5), the estimated function is then

$$\widehat{f_\lambda} = \frac{1}{n}\sum_{i\in[\![n]\!]} \widehat{\alpha}_i k(X_i, \cdot).$$

Hence, given the vector of coefficients $\widehat{\alpha}$ and continued access to the training points (X_i), we can compute easily the prediction $\widehat{f_\lambda}(x)$ at any test point x.

To sum up, we can use the equivalent representation (5.8) for actual numerical computation of the estimated function at any point, while we use the "abstract" representation (5.7) for the statistical analysis of the estimator, for which the entirety of the arguments from the previous chapter apply.

A worked example. As a an additional step towards interpretation of the statistical results in this setting, let us look at the kind of "regularity" that the source assumption 4.10 entails. We will be looking at a deliberately simplified illustrative example.

Let $\mathcal{X}$ be the interval $[0, 2\pi]$ (seen as the unit circle) and assume the covariate distribution ρ is the uniform distribution on $\mathcal{X}$. Furthermore, assume we consider a kernel k on $\mathcal{X}$ of the form $k(x, y) = F(x - y)$, where F is a function $\mathcal{X} \to \mathbb{R}$. We assume F can be written as a Fourier series

$$F(t) = a_0 + \sum_{k\geq 0} 2a_k \cos(kt),$$

with positive, summable coefficients $(a_k)_{k\in\mathbb{N}}$ (the factor 2 for $k \geq 1$ is introduced to simplify things later on).

Let us first justify that k is a spd kernel. We have

$$k(x,y) = F(x-y) = a_0 + \sum_{k\geq 1} 2a_k(\cos(kx)\cos(ky) + \sin(kx)\sin(ky)) = \langle \Phi(x), \Phi(y)\rangle_{\ell_2},$$

where $\Phi(x) = (\sqrt{a_0}, \sqrt{2a_1}\cos(x), \sqrt{2a_1}\sin(x), \ldots, \sqrt{2a_k}\cos(kx), \sqrt{2a_k}\sin(kx), \ldots) \in \ell_2(\mathbb{N})$. Through this explicit representation in the (real) Hilbert space $\ell_2(\mathbb{N})$, via Definition 5.1 it is indeed checked that k is a spd kernel on $\mathcal{X}$.

Moreover, via Proposition 5.8 and in particular formula (5.4), we see that if f is a function on $\mathcal{X}$ with Fourier expansion

$$f(x) = f_0 + \sum_{k\geq 1}(f_k^c \cos(kx) + f_k^s \sin(kx)), \tag{5.9}$$

then

$$S^*f(x) = a_0 f_0 + \sum_{k\geq 1}(a_k f_k^c \cos(kx) + a_k f_k^s \sin(kx)).$$

In particular, we see that $S^*(x \mapsto \cos(kx)) = a_k(x \mapsto \cos(kx))$, thus all trigonometric functions are elements of $\mathcal{H}$, are in fact eigenfunctions of the operator $\Sigma = S^*S$ with corresponding eigenvalues $(\lambda_k = a_k)_{k\geq 0}$.

Due to the fact that $S(S^*S)^{-\frac{1}{2}}$ is a partial isometry (this holds in general, as seen from the singular value decomposition), whose range is dense in $L^2([0,2\pi])$, we can deduce that if f is a function with Fourier decomposition (5.9), then f is an element of $\mathcal{H}$ iff $\sum_{k\geq 0} a_k^{-1}((f_k^c)^2 + (f_k^s)^2) < \infty$, and in fact we have

$$\|f\|_{\mathcal{H}}^2 = a_0^{-1} f_0^2 + \sum_{k\geq 1} a_k^{-1}((f_k^c)^2 + (f_k^s)^2).$$

From this, we understand precisely the nature of the functions in the rkHs, which have (very) roughly speaking half the regularity of the kernel function F. For instance, if $a_k = \mathcal{O}(k^{-\alpha})$, then the kernel function is in the Sobolev space H^s for all $s < \alpha - \frac{1}{2}$, i.e. it is "almost" $\alpha - \frac{1}{2}$ weakly differentiable (possibly in a fractional derivative sense), while the rkHs is made of functions that are (at least) $\alpha/2$ times weakly differentiable (i.e. belong to the Sobolev space $H^{\alpha/2}$).

By the same type of arguments, we find that if f is in the range of Σ^r then

$$\left\|\Sigma^{-r} f\right\|_{\mathcal{H}}^2 = a_0^{-(1+2r)} f_0^2 + \sum_{k\geq 1} a_k^{-(1+2r)}((f_k^c)^2 + (f_k^s)^2),$$

so the functions satifying the Hölder source condition of order r are exactly those such that $\sum_{k\geq 1} a_k^{-(1+2r)}((f_k^c)^2 + (f_k^s)^2) < \infty$. For instance, if $a_k \propto k^{-\alpha}$, a function satisfying a source condition of order r iff it is in the Sobolev space H^β with $\beta = \alpha(r + \frac{1}{2})$, i.e.

β times (fractional, weakly) differentiable. It is notable that this is only this "intrinsic" regularity parameter of the target function that drives the convergence rate of the statistical analysis (4.23). In other words, if we used a different kernel function with a different power spectral decay of order α' of the associated kernel integral operator, we would get the same convergence rate for target functions of Sobolev regularity β because they would satisfy a different source regularity condition of order r' for this kernel, leading to the same convergence rate only depending on β. (In fact the convergence rate appearing in (4.23) can be shown to be statistically minimax optimal for functions of that intrinsic regularity). This argument holds provided $\alpha' < 2\beta$ (since we need $r' \geq 0$) and the qualification of the methods is large enough to cover source regularity r'.

A conclusion from this example is that it could be preferable to choose a less regular kernel and use a regularization method with large qualification, because it can adapt to target functions more regular than the kernel via the source condition, while the converse is not guaranteed: it is not clear if using a smooth kernel can adapt to irregular functions (of smoothness less than half of that of the kernel). There are actually results in that direction, but they are more difficult and require more assumptions.

5.4 Kernel mean embeddings of distributions

For a non-technical overview of the topic and its applications, see eg. [22].

Let $\mathcal{H}$ be a rkHs over the space $\mathcal{X}$. For any probability distribution $\mathbb{P}$ on $\mathcal{X}$, and $X \sim \mathbb{P}$, the founding idea of kernel mean embedding (KME) is to consider the expectation of $\Phi(X)$ (if it exists) and use this as a "representation" of the probability distribution $\boldsymbol{p}$.

Proposition/Definition 5.9 (Kernel Mean Embedding (KME))**.** Let $\mathcal{H}$ be a rkHs with kernel k over the space $\mathcal{X}$. Let $\mathcal{P}_k$ be the set of probability distributions $\mathbb{P}$ over $\mathcal{X}$ such that $\mathbb{E}_{X\sim\mathbb{P}}\left[\sqrt{k(X,X)}\right] < \infty$. The following mapping, called kernel mean embedding mapping, also denoted by Φ, is well-defined:

$$\Phi : \qquad \mathcal{P}_k \to \mathcal{H} : \qquad \mathbb{P} \mapsto \Phi(\mathbb{P}) := \mathbb{E}_{X\sim\mathbb{P}}[\Phi(X)].$$

Proof. Measurability of Φ is ensured via implicit assumption **[M]** and Proposition 5.7. Since $\|\Phi(X)\| = \sqrt{k(X,X)}$, the random variable $\Phi(X)$ with $X \sim \mathbb{P}$ is Bochner-integrable as soon as $\mathbb{P} \in \mathcal{P}_k$. □

Remarks.

- for a bounded kernel, the KME is defined for any probability distribution on $\mathcal{X}$.
- the use of the same symbol Φ for the usual kernel feature mapping and the KME makes sense if we consider the KME as an "extension" of the kernel mapping, in the sense that the KME coincides the usual kernel feature mapping on Dirac-delta probability distributions.

- the KME mapping is affine in the sense that $\Phi(t\mathbb{P}+(1-t)\mathbb{Q}) = t\Phi(\mathbb{P})+(1-t)\Phi(\mathbb{Q})$. In fact by linearity it can be extended to any finite measure $\mu = a\mathbb{P} - b\mathbb{Q}$ for $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_k$.

The KME mapping gives rise to a pseudometric on $\mathcal{P}_k$ called *maximum mean discrepancy* (MMD):

Definition 5.10 (maximum mean discrepancy (MMD))**.** Under the same setting and notation as Def. 5.9, we define for $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_k$ the pseudometric

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) = \|\Phi(\mathbb{P}) - \Phi(\mathbb{Q})\|. \tag{5.10}$$

Proposition 5.11. *The MMD is an* integral probability (pseudo)metric *over $\mathcal{P}_k$:*

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) = \sup_{\substack{f \in \mathcal{H} \\ \|f\|=1}} (\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{X\sim\mathbb{Q}}[f(X)]). \tag{5.11}$$

Also, we have the formula

$$\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}[k(X, X') + k(Y, Y') - 2k(X, Y)], \tag{5.12}$$

where $(X, X', Y, Y') \sim \mathbb{P}^{\otimes 2} \otimes \mathbb{Q}^{\otimes 2}$.

Proof. It holds

$$\|\Phi(\mathbb{P}) - \Phi(\mathbb{Q})\| = \sup_{\substack{f \in \mathcal{H} \\ \|f\|=1}} \langle f, \Phi(\mathbb{P}) - \Phi(\mathbb{Q})\rangle;$$

we can then use the definition of KME, exchange expectation and scalar product by Bochner integrability, and use the reproducting property of the rkHs.
Furthermore, we have

$$\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \langle \Phi(\mathbb{P}), \Phi(\mathbb{P})\rangle + \langle \Phi(\mathbb{Q}), \Phi(\mathbb{Q})\rangle - 2\langle \Phi(\mathbb{P}), \Phi(\mathbb{Q})\rangle.$$

If we look at the first term, it can be rewritten

$$\begin{aligned}\langle \Phi(\mathbb{P}), \Phi(\mathbb{P})\rangle = \langle \mathbb{E}_{X\sim P}[\Phi(X)], \mathbb{E}_{X\sim P}[\Phi(X')]\rangle &= \mathbb{E}_{(X,X')\sim\mathbb{P}^{\otimes 2}}[\langle \Phi(X), \Phi(X')\rangle] \\ &= \mathbb{E}_{(X,X')\sim\mathbb{P}^{\otimes 2}}[k(X, X')];\end{aligned}$$

the other terms are handled in the same way. □

The KME/MMD (or rather its estimation from data) can be used for a variety of purposes, the most well-known is for testing equality of distributions [13]. More precisely, if we have $\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) > 0$, the obviously $\mathbb{P} \neq \mathbb{Q}$; by estimating MMD_k from data with a confidence interval, we can therefore construct a test for $\mathbb{P} = \mathbb{Q}$.

Assume we have two independent, i.i.d samples $(X_i)_{i\in[\![n]\!]}$ and $(Y_i)_{i\in[\![m]\!]}$ of $\mathbb{P}$ and $\mathbb{Q}$, with $m, n \geq 2$. Then the following is an unbiased estimator of $\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q})$:

$$\widehat{\mathrm{MMD}}_k^2(\mathbb{P}, \mathbb{Q}) := \\ \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} k(X_i, X_j) + \frac{2}{m(m-1)} \sum_{1\leq i<j\leq m} k(Y_i, Y_j) - \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} k(X_i, Y_j) \tag{5.13}$$

We can bound the estimation error roughly through the following bound:

Proposition 5.12. *Assume the kernel k is bounded by κ^2. Then for $\delta \in (0,1)$, with probability at least $1-\delta$ it holds*

$$\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) \geq \widehat{\mathrm{MMD}}_k^2(\mathbb{P},\mathbb{Q}) - 8\kappa^2\sqrt{\frac{\log \delta^{-1}}{\min(m,n)}}. \tag{5.14}$$

Proof. Apply the bounded difference inequality (Thm. 3.5) to the function

$$f((x_i)_{i\in[\![n]\!]},(y_j)_{j\in[\![m]\!]}) = \mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) - \widehat{\mathrm{MMD}}_k^2(\mathbb{P},\mathbb{Q}),$$

which is applied to a family of $(n+m)$ independent random variables. Changing arbitarily the value of x_i results in a modification of at most $(n-1)$ terms in the first sum of (5.13), and m terms in the last sum. Since each term is bounded in absolute value by κ^2, overall the value of the function changes at most by $8\kappa^2/n = 2c_i$, arguing similarly for changing a value of y_j it follows $\sum_{i=1}^{n+m} c_i^2 = 16\kappa^4(n^{-1}+m^{-1}) \leq 32\kappa^4 \min(m,n)^{-1}$, leading to the claim. □

The bound (5.14) gives us the criterion to reject the null hypothesis $\mathbb{P}=\mathbb{Q}$ if $\widehat{\mathrm{MMD}}_k^2(\mathbb{P},\mathbb{Q}) \geq 8\kappa^2\sqrt{(\log\delta^{-1}/\min(m,n))}$. In practice, this rejection threshold is not very sharp since it is based on a non-asymptotic inequality, using only a uniform upper bound. Sharper rejection thresholds based on U-statistics asymptotics can be proposed. Furthermore, computing $\widehat{\mathrm{MMD}}$ is quadratic in $\max(m,n)$, which can be a problem for large datasets in practice. A "linearized" version consists in considering

$$\widehat{\mathrm{MMDL}}_k := \frac{1}{\ell}\sum_{i=1}^{\ell}\big(k(X_{2i},X_{2i+1}) + k(Y_{2i},Y_{2i+1}) - k(X_{2i},Y_{2i}) - k(X_{2i+1},Y_{2i+1})\big), \tag{5.15}$$

where $\ell = \lfloor \frac{\min(m,n)}{2} \rfloor$, when m,n are both "large". This statistic has linear computation time, is an unbiased estimate of $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q})$, furthermore since is is a sum of ℓ independent terms one can apply Hoeffding's, Bernstein's, or even better in practice an asymptotic CLT approximation (estimating the variance separately) in order to get a rejection region. This method (or variations thereof) is used in practice for large datasets.

From (5.14) it is seen that (assuming bounded kernel) the test is consistent, i.e. rejects with probability tending to 1 as $m,n\to\infty$ if $\mathbb{P}\neq\mathbb{Q}$, *provided* $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q})>0$. We have *universal consistency* if $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q})>0$ for any $\mathbb{P}\neq\mathbb{Q}$, which in turn is implied if the KME mapping $\Phi(\mathbb{P})$ is injective. If this is true k is called a *characteristic kernel*. For a detailed analysis of this property in relation with other properties of kernels, see for instance [29].

Hilbert independence test

The previous KME/MMD principle can be put to use in order to construct an *independence test* of two variables (X,Y) on a domain $\mathcal{X}\times\mathcal{Y}$. For this let k be an spd kernel on $\mathcal{X}\times\mathcal{Y}$ and define

$$\mathrm{HSIC}_k(X,Y) = \mathrm{MMD}_k(\mathbb{P}_{XY},\mathbb{P}_X\otimes\mathbb{P}_Y), \tag{5.16}$$

where $\mathbb{P}_{XY}$ denotes the joint distribution of (X, Y) and $\mathbb{P}_X, \mathbb{P}_Y$ its marginals. Obviously $\mathrm{HSIC}(X, Y) > 0$ implies that (X, Y) are not independent, so again we can use an estimation of this criterion in order to test the null hyopthesis of independence.

Usually the kernel k on $\mathcal{X} \times \mathcal{Y}$ is taken as the product $k_X(x, x')k_Y(y, y')$ of two spd kernels on $\mathcal{X}$ and $\mathcal{Y}$. In this case the formula (5.12) becomes

$$\mathrm{HSIC}(X, Y) = \mathbb{E}[k_X(X, X')k_Y(Y, Y') + k_X(X'', X''')k_Y(Y'', Y''') - 2k_X(X, X'')k_Y(Y, Y'')],$$

where $(X, Y), (X', Y')$ are independent draws from $\mathbb{P}_{XY}$ and $(X'', Y''), (X''', Y''')$ are independent draws from the product distribution $\mathbb{P}_X \otimes \mathbb{P}_Y$.

Given an i.i.d. sample $(X_i, Y_i)_{i \in [\![n]\!]}$ of size $n > 3$ from $\mathbb{P}_{XY}$, the first term above can be estimated without bias via

$$\frac{1}{n(n-1)} \sum_{i \neq j} k_X(X_i, X_j)k_Y(Y_i, Y_j),$$

the second via

$$\frac{1}{n(n-1)(n-2)(n-3)} \sum_{i \neq j \neq k \neq \ell} k_X(X_i, X_j)k_X(Y_k, Y_\ell),$$

the third via

$$\frac{2}{n(n-1)(n-2)} \sum_{i \neq j \neq k} k_X(X_i, X_j)k_X(Y_i, Y_k).$$

Again, in practice the sums can be reduced so that they run only over distinct groups of indices in order to reduce computation load (also allowing an asymptotic CLT approximation), at the price of increased variability.

5.5 Kernel PCA

TODO!

6 Acceleration methods

In this section we will consider different approaches to speed up the numerical computation of procedures seen previously, such as the spectral regularization procedures (4.14). Usually, this is done at the price of some approximation, and it is of interest to analyze if this does not jeopardize the statistical guarantee on the obtained estimator. Note that a particularly important computational bottleneck is the computation of the regularized inverse $F_\lambda(\widehat{G})$. Even if, for some regularization methods, it consists of forward matrix multiplications, we still have to manipulate a $n \times n$ matrix. When the datasize n is large, it can be problematic or time-costly. We would like to find ways to alleviate this point in particular.

6.1 Parallelizing: divide and average

References: [34, 23, 19] (between many others)

A common situation is to have a large number m of computers available. Distributing across machines is easy when operations have to be done in parallel and on different parts of the data. However it is not obvious a priori how to identify a parallelization opportunity in the estimator (4.14) and in particular, as mentioned above, concerning the computation of $F_\lambda(\widehat{G})$.

An apparently naive approach is simply to divide the data into m blocks of approximately equal size $N = n/m$ (we will assume that n is a multiple of m to simplify, but obviously this is not required), compute independently the estimators $\widehat{f}_\lambda^{(1)}, \dots, \widehat{f}_\lambda^{(m)}$ on the separate blocks the data on the m separate machines, and take a simple average of their output. Not only is the computational burden parallelized accross machines, but the computations for each $\widehat{f}_\lambda^{(k)}$ can be simpler since they rely on less data. For example, if computing exactly the estimator $\widehat{f}_\lambda$ has a superlinear cost (in floating point operations, flops) $O(n^\gamma)$ with $\gamma \geq 1$, then the total cost for the divide-and-average approach is only $O(m(n/m)^\gamma)$, giving an operation complexity gain factor of $O(m^{\gamma-1})$ (and a time complexity gain factor of $O(m^\gamma)$ since the machines run in parallel).

Of course, the important question is whether this approach preserves the statistical convergence rate shown in Section 4. Surprisingly enough, it is actually the case (whithin certain limits). The following is an analogue of Proposition 4.11 in the distributed setting.

Proposition 6.1. *Suppose granted Assumption 4.1, regularization function Assumption 4.8 with qualification q, and Hölder source Assumption 4.19 of order r, such that $q \geq r + \frac{1}{2}$. For any $\lambda \in [0, \|\Sigma\|_{\mathrm{op}}]$, consider the "distribute-and-average" estimator*

$$\widetilde{f}_\lambda = \frac{1}{m} \sum_{k \in [\![m]\!]} \widehat{f}_\lambda^{(k)},$$

where $\widehat{f}_\lambda^{(k)}$ are given by (4.14) applied to each of the m subsamples of size n/m.

Assume that it holds

$$m \leq C_{\blacktriangle} n \frac{\lambda}{\log(2\mathcal{N}(\lambda))}, \tag{6.1}$$

and, if $r > \frac{1}{2}$*:*

$$m \leq \mathcal{N}(\lambda)\lambda^{-\min(1,2r-1)}. \tag{6.2}$$

Then, for n *large enough, it holds with probability at least* $1 - 2/n$*:*

$$\left\|\Sigma^{\frac{1}{2}}(\widehat{f}_\lambda - f^*)\right\|_{\mathcal{H}} \leq C_{\blacktriangle}\left(\sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \lambda^{r+\frac{1}{2}}\right)\log n. \tag{6.3}$$

This means that we can attain the same stastistical convergence rates as in Corollary 4.12 (up to logarithmic factor in n) in the distributed setting, with the same choice of regularization parameter given by (4.22), provided the constraints (6.1) and (6.2) are satisfied. Concretely, in the same setting as in Corollary 4.12, the first constraints reads as

$$m \leq C_{\blacktriangle} n^{1-\frac{\alpha}{2\beta+1}},$$

and the second constraint as (for $r \in [\frac{1}{2}, 1]$, i.e. $\beta \in [\alpha, 3\alpha/2]$) :

$$m \leq C_{\blacktriangle} n^{1-\frac{1+\alpha}{2\beta+1}},$$

and for $r \geq 1$, i.e. $\beta \geq 3\alpha/2$:

$$m \leq C_{\blacktriangle} n^{1-\frac{2(\beta-\alpha)}{2\beta+1}}.$$

While the complete interpretation is not obvious (also the above bounds are not claimed to be optimal), one can check that there is always a non-trivial possibility to distribute, i.e. m can be chosen as a a non-trivial power of n, whithin some limits.

Proof. The key is to start again from the decomposition (4.25). We restate here for each of the estimators $\widehat{f}_\lambda^{(k)}$ (indicating with the index (k) that the empirical quantities only depend on subsample k):

$$\begin{aligned}\Sigma^{\frac{1}{2}}(\widehat{f}_\lambda^{(k)} - f^*) &= \Sigma^{\frac{1}{2}}(F_\lambda(\widehat{\Sigma}_{(k)})\widehat{\Sigma}_{(k)} - I)f^* + \Sigma^{\frac{1}{2}}F_\lambda(\widehat{\Sigma}_{(k)})(\widehat{S}^*_{(k)}\boldsymbol{\xi}_{(k)}) \\ &= (I)_{(k)} + (II)_{(k)}. \end{aligned} \tag{6.4}$$

We note that the "bias" term $(I)_{(k)}$has non-zero expectation which is the same accross machines, while the "noise" term $(II)_{(k)}$ has expectation zero and has independent realizations across machines; we can therefore hope for a noise reduction effect for term $(II)_{(k)}$ when averaging accross machines. On the other hand, this effect won't apply to term $(I)_{(k)}$, which should therefore be uniformly small accross machines.

We will need to reiterate the arguments used in the proof of Proposition 4.11 separately on each of the m machines each dealing with an independent data set of size N. For this, as before we will assume that the probabilistic inequalities of Proposition 4.6 are satisfied,

as well as those of Corollary 4.7, simulaneously for all machines and data subsets. We therefore use an union bound over machines, which amounts to say that the statements we make will hold with probability $1 - m\delta$ rather than $1 - \delta$. (Thus, for statements with probability $1 - \delta$ we will need to replace L_δ by $L_{\delta/m} \leq C_\blacktriangle L_\delta + \log(m)$, which we will do at the end.)

Furthermore, remember that requirements for these inequalities to hold is $\lambda \in (0, \|\Sigma\|_{\mathrm{op}})$ (which will be satisfied for n large enough, so we won't discuss it in more detail), and more importantly, condition (4.11) for a data sample of size N:

$$N \geq C_\blacktriangle A \frac{\log(2\mathcal{N}(\lambda)) + L_\delta}{\lambda}. \tag{6.5}$$

Let us start with term $(I)_{(k)}$; we recall the control obtained in (4.28) for a data sample of size N:

$$\left\|(I)_{(k)}\right\| = \left\|\Sigma^{\frac{1}{2}}(F_\lambda(\widehat{\Sigma}_{(k)})\widehat{\Sigma}_{(k)} - I)f^*\right\| \leq C_\blacktriangle \left(\lambda^{r+\frac{1}{2}} + \frac{\mathbf{1}\{r \geq \frac{1}{2}\}\sqrt{\lambda}\sqrt{L_\delta}}{\lambda^{(1-r)_+}\sqrt{N}}\right), \tag{6.6}$$

which will hold (with probability $1 - m\delta$) for all machines simultaneously (for the $\widehat{\Sigma}_{(k)}$ corresponding to their respective data subsamples, $k \in [\![m]\!]$).

For the "noise" term $(II)_{(k)}$, under the same condition (6.5), for any individual $k \in [\![m]\!]$ (indicating the subsample) we have with probability $1 - \delta$ the control (4.26), which we rewrite here:

$$\left\|(II)_{(k)}\right\| = \left\|\Sigma^{\frac{1}{2}}F_\lambda(\widehat{\Sigma}_{(k)})(\widehat{S}^*_{(k)}\boldsymbol{\xi})\right\| \leq C_\blacktriangle \left(\sqrt{\frac{L_\delta \mathcal{N}(\lambda)}{N}} + \frac{L_\delta}{N\sqrt{\lambda}}\right) \leq C'_\blacktriangle \sqrt{\frac{L_\delta \mathcal{N}(\lambda)}{N}}. \tag{6.7}$$

Let us denote the Hilbert-valued random variables

$$U_{(k)} := (II)_{(k)} = \Sigma^{\frac{1}{2}}F_\lambda(\widehat{\Sigma}_{(k)})(\widehat{S}^*_{(k)}\boldsymbol{\xi}_{(k)}), \qquad k \in [\![m]\!].$$

Note that the variables $U_{(k)}$ are independent (since $U_{(k)}$ only depends on the subsample k, and these subsamples are independent), and bounded in norm by $B := C'_\blacktriangle \sqrt{\frac{L_\delta \mathcal{N}(\lambda)}{N}}$ with high probability $1-\delta$ taken individually, or simultaneously with probability $1-m\delta$. We will therefore apply Hoeffding's inequality, using the following trick: consider the "truncated" random variables:

$$\widetilde{U}_{(k)} := \begin{cases} U_{(k)}, & \text{if } \left\|U_{(k)}\right\| \leq B; \\ 0, & \text{if } \left\|U_{(k)}\right\| > B. \end{cases}$$

We will apply Hoeffding's inequality to the family $(\widetilde{U}_{(k)})_{k\in[\![m]\!]}$ and argue that with high probability their sum coincides with that of the $(U_{(k)})_{k\in[\![m]\!]}$.

First, an annoyance point is that the modified variables $\widetilde{U}_{(k)}$ are not guaranteed to be centered like the $U_{(k)}$s were. Let us therefore roughly upper bound this discrepancy: since

$\sup_{t\in[0,\kappa^2]}|F_\lambda(t)| \leq E/\lambda$ and $\widehat{S}^*_{(k)}\boldsymbol{\xi}_k$ is an average of variables bounded in norm by $2M\kappa$ (see e.g. proof of Proposition 4.4), the following rough upper bound holds (always):

$$\big\|U_{(k)}\big\| \leq \frac{C_\blacktriangle}{\lambda},$$

and therefore

$$\begin{aligned}\Big\|\mathbb{E}\big[\widetilde{U}_{(k)}\big]\Big\| = \Big\|\mathbb{E}\big[\widetilde{U}_{(k)}\big] - U_{(k)}\Big\| &= \Big\|\mathbb{E}\big[(\widetilde{U}_{(k)} - U_{(k)})\mathbf{1}\big\{\widetilde{U}_{(k)} \neq U_{(k)}\big\}\big]\Big\| \\ &\leq \frac{C_\blacktriangle}{\lambda}\mathbb{P}\big[\widetilde{U}_{(k)} \neq U_{(k)}\big] \\ &= \frac{C_\blacktriangle}{\lambda}\mathbb{P}\big[\|U\|_{(k)} > B\big] \\ &\leq C_\blacktriangle\frac{\delta}{\lambda}.\end{aligned}$$

We will assume in the sequel the condition

$$\delta \leq \frac{\lambda}{\sqrt{n}}, \tag{6.8}$$

which implies in particular from the above that $\Big\|\mathbb{E}\big[\widetilde{U}_{(k)}\big]\Big\| \leq C_\blacktriangle B$ (using the definition of B to check this).

Applying the vectorial Hoeffding's inequality to the centered variables $\overline{U}_{(k)} := \widetilde{U}_{(k)} - \mathbb{E}\big[\widetilde{U}_{(k)}\big]$, $k \in [\![m]\!]$, independent and bounded in norm by $C_\blacktriangle B$, therefore yields that for any $\eta \in (0,1)$, with probability at least $1-\eta$ it holds

$$\left\|\frac{1}{m}\sum_{k\in[\![m]\!]}\overline{U}_{(k)}\right\| \leq \frac{C_\blacktriangle B\sqrt{L_\eta}}{\sqrt{m}},$$

entailing

$$\begin{aligned}\left\|\frac{1}{m}\sum_{k\in[\![m]\!]}\widetilde{U}_{(k)}\right\| &\leq C_\blacktriangle\left(\frac{\delta}{\lambda} + \frac{C_\blacktriangle B\sqrt{L_\eta}}{\sqrt{m}}\right) \\ &\leq C_\blacktriangle\sqrt{\frac{\mathcal{N}(\lambda)L_\delta L_\eta}{n}},\end{aligned}$$

where we have used the definition of B, condition (6.8), and $n = Nm$.

Finally, the latter bound also holds for $\left\|\frac{1}{m}\sum_{k\in[\![m]\!]}U_{(k)}\right\|$ with probability $1-\eta-m\delta$, since it holds $\widetilde{U}_{(k)} = U_{(k)}$ for all $k \in [\![m]\!]$ with probability at least $1-m\delta$.

Summing up: plugging in this estimate as well as the estimate (6.6) into (6.4), we obtain

$$\left\|\Sigma^{\frac{1}{2}}\left(\frac{1}{m}\sum_{k\in[\![m]\!]}f_\lambda^{(k)} - f^*\right)\right\| \leq C_\blacktriangle\left(\lambda^{r+\frac{1}{2}} + \frac{\mathbf{1}\big\{r\geq\frac{1}{2}\big\}\sqrt{\lambda}}{\lambda^{(1-r)_+}\sqrt{N}} + \sqrt{\frac{\mathcal{N}(\lambda)}{n}}\right)\sqrt{L_\delta L_\eta}, \tag{6.9}$$

with probability $1-\eta-m\delta$, and provided that conditions (6.5) and (6.8) hold. This is to be compared to the bound (4.29) that we obtained for the "single machine" analysis.

Let us analyze these conditions: first, condition (6.5) implies in particular $\lambda \geq C_{\blacktriangle}/\sqrt{N} \geq C_{\blacktriangle}/\sqrt{n}$. Hence, (6.8) is ensured if (6.5) is and if $\delta \leq C_{\blacktriangle} n^{-3/2}$. This is reasonable since it will only result in a logarithmic factor (coming from L_δ). We choose henceforth $\delta = n^{-2}$, so that $m\delta \leq 1/n$; and $\eta = n^{-1}$.

As for condition (6.5) itself, due to $N = n/m$ is is implied by the sufficient condition on the number of subsamples/machines m:

$$m \leq C_{\blacktriangle} n \frac{\lambda}{\log(2\mathcal{N}(\lambda))}.$$

Finally, as in the analysis of (4.29), we would like to be able to wrap the second term into the third one. This is the case (i.e. the second term is smaller), again using $N = n/m$, if

$$m \leq \mathcal{N}(\lambda)\lambda^{-\min(1,2r-1)};$$

remember that this constraint is only relevant if $r \geq \frac{1}{2}$. □

6.2 Nyström methods

Sources: [12, 28, 18] (between many others)

The Nyström approximation method consists, roughly speaking, in approximating the kernel covariance and/or the kernel Gram matrix by a lower rank matrix obtained by subsampling points.

We expose it here for kernel ridge regression (KRR). Remember that the kernel ridge regression algorithm using a rkHs $\mathcal{H}$ with kernel k over the input space $\mathcal{X}$ can be seen as the solution of the following minimization problem:

$$\begin{aligned}\widehat{f}_\lambda &= \operatorname*{Arg\,Min}_{f\in\mathcal{H}} \left(\frac{1}{n}\sum_{i\in[\![n]\!]} (f(X_i)-Y_i)^2 + \lambda\|f\|_{\mathcal{H}}^2 \right) \\ &= \operatorname*{Arg\,Min}_{f\in\mathcal{H}} \Big(\big\|\widehat{S}f - \boldsymbol{Y}\big\|_n^2 + \lambda\|f\|_{\mathcal{H}}^2 \Big), \end{aligned} \tag{6.10}$$

whose explicit solution is

$$\widehat{f}_\lambda = (\widehat{S}^*\widehat{S} + \lambda I)^{-1}\widehat{S}^*\boldsymbol{Y} = \widehat{S}^*(\widehat{S}\widehat{S}^* + \lambda I)^{-1}\boldsymbol{Y};$$

we recall that the latter form means that the solution can be written as

$$\widehat{f}_\lambda = \widehat{S}\widehat{\boldsymbol{\alpha}} = \sum_{i\in[\![n]\!]} \widehat{\alpha}_i k(X_i,\cdot); \qquad \text{with } \widehat{\boldsymbol{\alpha}} = (\widehat{G}+\lambda I)^{-1}\boldsymbol{Y},$$

which is also known as a form of the so-called representer theorem (see for instance [2] for details).

The idea of Nyström-based methods is to approximate the above expansion of f by a reduced expansion on a subset points of size m, i.e. of the form

$$\widetilde{f}_\lambda = \sum_{i\in I} \widetilde{\alpha}_i k(X_i, \cdot), \tag{6.11}$$

for a subset of indices I with $|I| = m$.

Since the original KRR estimator is obtained by minimization of objective (6.10) over $f \in \mathcal{H}$, or, equivalently because of the above representation, over $f \in \mathcal{H}_n = \mathrm{Span} k(X_i, \cdot), i \in [\![n]\!]$, the natural approach when considering elements with the reduced expansion (6.11), i.e. $f \in \mathcal{H}_I = \mathrm{Span} k(X_i, \cdot), i \in I$ is to minimize the same objective under this constraint.

Proposition 6.2. *Let $I \subseteq [\![m]\!]$ be a subset of indices and $\mathcal{H}_I = \mathrm{Span}\{k(X_i, \cdot), i \in I\}$. Then the solution of*

$$\operatorname*{Arg\,Min}_{f\in\mathcal{H}_I} \Big(\big\| \widehat{S} f - \boldsymbol{Y} \big\|_n^2 + \lambda \|f\|_{\mathcal{H}}^2 \Big) \tag{6.12}$$

is given by

$$\widetilde{f}_\lambda = \frac{1}{n} \sum_{i\in I} \widetilde{\alpha}_i k(X_i, \cdot), \qquad \textit{with } \widetilde{\boldsymbol{\alpha}} = (\widehat{G}_{I,[\![n]\!]} \widehat{G}^t_{I,[\![n]\!]} + \lambda \widehat{G}_{I,I})^{-1} \widehat{G}_{I,[\![n]\!]} \boldsymbol{Y}, \tag{6.13}$$

where $\widehat{G}_{I,J}$ denotes the submatrix of the normalized Gram kernel $\widehat{G}$ corresponding to indices sets I and J.

Proof. We can write explicitly, for $f = \sum_{i\in I} \alpha_i k(X_i, \cdot)$, that the vector of evaluation of f at points $(X_1, \ldots, X_n)$ is given by $n\widehat{G}_{[\![n]\!],I}\boldsymbol{\alpha}$ (the factor n because $\widehat{G}$ is the normalized Gram matrix). Furthermore, by properties of a rkHs, it holds $\|f\|_{\mathcal{H}}^2 = \sum_{i,j\in I} \alpha_i \alpha_j k(X_i, X_j) = n\boldsymbol{\alpha}^t \widehat{G}_{I,I} \boldsymbol{\alpha}$.

Thus, (6.12) is rewritten as the minimization of

$$\Big\| n\widehat{G}_{[\![n]\!],I}\boldsymbol{\alpha} - \boldsymbol{Y} \Big\|_n^2 + n\lambda \boldsymbol{\alpha}^t \widehat{G}_{I,I} \boldsymbol{\alpha} = \boldsymbol{\alpha}^t \Big(n\widehat{G}_{I,[\![n]\!]} \widehat{G}^t_{I,[\![n]\!]} + n\lambda \widehat{G}_{I,I} \Big) \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^t \widehat{G}_{I,[\![n]\!]} \boldsymbol{Y} + \|\boldsymbol{Y}\|_n^2.$$

Standard formulas for quadratic optimization and some bookkeeping yield (6.13). □

Note the interesting fact that to compute the Nyström approximate solution, it is not necessary to compute the full kernel Gram matrix $\widehat{G}$, only the submatrix $\widehat{G}_{I,[\![n]\!]}$. Furthermore, the costly step of matrix inversion only concerns a (m, m) matrix instead of a (n, n) one, thus significantly reducing computation.

Several strategies can be proposed for selection of the subset I:

- Uniform sampling (with or without replacement) of m indices whithin $[\![n]\!]$;
- *Leverage score* sampling, where the indices are sampled with weights proportional to so-called leverage scores,

$$\ell_\lambda(i) := \Big(\widehat{G} \big(\widehat{G} + \lambda I \big)^{-1} \Big)_{ii}.$$

A theoretical analysis (see [28, 18]) shows that under some lower bound of the subsample size m depending on the problem parameters (source condition, intrinsic dimension) but still allow $m \ll n$, the statistical convergence rate obtained in Corollary 4.12 can be preserved for the Nyström approximated estimator $\widetilde{f}_\lambda$. The use of the leverage score sampling allows further reduction of the subsample size m, however there is a chicken-and-egg problem in that exact computation of these scores itselfs in principle require inversion of the (n, n) matrix we were trying to avoid! To alleviate this, several approaches to approximate the leverage scores have been proposed, see for instance [27].

References

[1] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

[2] Gilles Blanchard. Mathematics for artificial intelligence I. (M1 Lecture notes), 2022.

[3] Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

[4] Stéphane Boucheron, Gabor Lugósi, and Pascal Massart. *Concentration Inequalities: a nonasymptotic theory of independence*. Oxford University Press, 2013.

[5] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer Nature, New York, NY, 2010.

[6] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

[7] John B Conway. *A Course in Functional Analysis*, volume 96 of *Graduate Texts in Mathematics*. Springer, 1985. [Disponible en version électronique à la bibliothèque du LMO].

[8] John B. Conway. *A Course in Operator Theory*, volume 21 of *Graduate studies in mathematics*. American Mathematical Society, 2000. [Disponible en version électronique à la bibliothèque du LMO].

[9] Heinz Otto Cordes. *Spectral theory of linear differential operators and comparison algebras*, volume 76. Cambridge University Press, 1987.

[10] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996.

[11] Takayuki Furuta. Norm inequalities equivalent to Loewner-Heinz theorem. *Reviews in Mathematical Physics*, 1(1):135–137, 1989.

[12] Alex Gittens and Michael Mahoney. Revisiting the Nyström method for improved large-scale machine learning. In *International Conference on Machine Learning*, pages 567–575. PMLR, 2013.

[13] Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[14] Milen Ivanov and Stanimir Troyanski. Uniformly smooth renorming of Banach spaces with modulus of convexity of power type 2. *Journal of Functional Analysis*, 237(2):373–390, 2006.

[15] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[16] Fuad Kittaneh. Norm inequalities for fractional powers of positive operators. *Letters in mathematical physics*, 27:279–285, 1993.

[17] Matthieu Lerasle. Lectures on high-dimensional probability. (M2 Lecture notes), 2023.

[18] Jian Li, Yong Liu, and Weiping Wang. Optimal convergence rates for agnostic Nyström kernel learning. In *International Conference on Machine Learning*, pages 19811–19836. PMLR, 2023.

[19] Junhong Lin and Volkan Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *Journal of Machine Learning Research*, 21(147):1–63, 2020.

[20] Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

[21] Stanislav Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017.

[22] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: a review and beyond. *Foundations and Trends in Machine Learning*, 10, 2017.

[23] Nicole Mücke and Gilles Blanchard. Parallelizing spectrally regularized kernel algorithms. *Journal of Machine Learning Research*, 19(30):1–29, 2018.

[24] Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

[25] Denis Potapov and Fedor Sukochev. Operator-Lipschitz functions in Schatten-von Neumann classes. *Acta Mathematica*, 207(2):375 – 389, 2011.

[26] Abhishake Rastogi and Sivananthan Sampath. Optimal rates for the regularized learning algorithms under general source condition. *Frontiers in Applied Mathematics and Statistics*, 3:3, 2017.

[27] Alessandro Rudi, Daniele Calandriello, Luigi Carratino, and Lorenzo Rosasco. On fast leverage score sampling and optimal learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[28] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in neural information processing systems*, 28, 2015.

[29] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

[30] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

[31] Nicole Tomczak-Jaegermann. The moduli of smoothness and convexity and the Rademacher averages of the trace classes S_p $(1 \leq p < \infty)$. *Studia Mathematica*, 50(2):163–182, 1974.

[32] Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[33] G. Wahba. *Spline Models for Observational Data*, volume 59. SIAM CBMS-NSF Series in Applied Mathematics, 1990.

[34] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression. In *Conference on learning theory*, pages 592–617. PMLR, 2013.