36-771 Fall 2018

Martingales 1 : Concentration inequalities

Martingale inequalities in Banach spaces

Lecturer : Aaditya Ramdas

1 Banach vs. Hilbert spaces

A Banach space $\mathcal{B}$ is a complete normed vector space. In terms of generality, it lies somewhere in between a metric space $\mathcal{M}$ (that has a metric, but no norm) and a Hilbert space $\mathcal{H}$ (that has an inner-product, and hence a norm, that in turn induces a metric). More formally, if a space is endowed with an inner-product $\langle\cdot,\cdot\rangle$, then it induces a norm $\|\cdot\|$ as $\|x\| = \sqrt{\langle x, x\rangle}$, and if a space is endowed with a norm, then it induces a metric $d(x, y) = \|x - y\|$. By "complete" normed vector space, one usually means that every Cauchy sequence (with respect to the norm) converges to a point that lies in the space. A metric space is called "separable" if it has a dense subset that is countable. A Hilbert space is separable iff it has a countable orthonormal basis.

When the underlying space is simply $\mathcal{C}^n$ or $\mathbb{R}^n$, any choice of norm $\|\cdot\|_p$ for $1 \leq p \leq \infty$ yields a Banach space, while only the choice $\|.\|_2$ leads to a Hilbert space. Similarly, if $(\mathcal{X}, \Omega, \mu)$ is a probability space, then the following space is a Banach space

$$\mathcal{L}^p(\mathcal{X}, \Omega, \mu) := \{f : \mathcal{X} \to \mathbb{C} \text{ such that } f \text{ is } \Omega\text{-measurable and } \int |f(x)|^p d\mu(x) < \infty\}$$

with norm $\|f\|_p := (\int |f(x)|^p d\mu(x))^{1/p}$ (with $f = g$ meaning that they are equal μ-a.e.). When $\mathcal{X} = \mathbb{R}$ or $\mathcal{X} = \mathcal{C}$ and μ is the Lebesgue measure, we sometimes just write

$$\mathcal{L}^p := \{f : \mathcal{C} \to \mathcal{C} \text{ such that } \int |f(x)|^p dx < \infty\}.$$

As another example, we write

$$\ell^\infty := \{(x_n)_{n\in\mathbb{N}} : \sup_n |x_n| < \infty\}$$

and the finite-dimensional variant as

$$\ell^\infty_d := \{(x_n)_{1\leq n\leq d} : \sup_{1\leq n\leq d} |x_n| < \infty\}.$$

Similarly, for matrices, the Frobenius norm induces a Hilbert space structure, but almost any of the other Schatten norms yield Banach spaces (the Schatten p-norm of a matrix is just the p-norm of its singular values).

2 Bounded/continuous linear operators

An operator A is linear if $A(\alpha x + \beta y) = \alpha A(x) + \beta A(y)$ for x, y in its domain. For linear operators, we denote $A(x)$ by just Ax for brevity, and is not to be confused with matrix-vector multiplication (which is nevertheless a useful special case). A linear operator $A : \mathcal{B} \to \mathcal{B}'$ is called bounded if

$$\|A\| := \sup_{x \in \mathcal{B}, x \neq 0} \frac{\|Ax\|_{\mathcal{B}'}}{\|x\|_{\mathcal{B}}} < \infty.$$

The above definition is then called the *operator norm* of A (it is the largest singular value for finite matrices, that is when $\mathcal{B}$ and $\mathcal{B}'$ are $\mathbb{R}^n$ and $\mathbb{R}^m$). Obviously, $\|A\|$ is the smallest number such that $\|Ax\|_{\mathcal{B}'} \leq \|A\|\|x\|_{\mathcal{B}}$.

The set of all such bounded linear operators $\mathcal{L}(\mathcal{B}, \mathcal{B}')$ is itself a Banach space with the above norm. Of course, if the domain of A is $D \subseteq \mathcal{B}$, the definition can be adjusted accordingly. A is said to be a continuous linear operator if $x_n \to x$ implies $Ax_n \to Ax$, meaning that

$$\text{if } \lim_{n \to \infty} \|x_n - x\|_{\mathcal{B}} = 0 \implies \lim_{n \to \infty} \|Ax_n - Ax\|_{\mathcal{B}'} = 0$$

For linear operators A, we have the following important fact:

$$A \text{ is continuous iff } A \text{ is bounded.}$$

3 Dual space

A linear functional on $\mathcal{B}$ is a linear operator $f : \mathcal{B} \to \mathbb{C}$ for which

$$\sup_{x \in \mathcal{B}, x \neq 0} \frac{|f(x)|}{\|x\|} < \infty.$$

The dual space $\mathcal{B}^*$ of a Banach space $\mathcal{B}$ is defined as the set of bounded linear functionals on $\mathcal{B}$. Clearly, $\mathcal{B}^*$ is itself a Banach space, and its norm is called the dual norm:

$$\|f\|_* := \sup_{x \in \mathcal{B}, x \neq 0} \frac{|f(x)|}{\|x\|}.$$

A reflexive Banach space is one such that $\mathcal{B}^{**} = \mathcal{B}$. Interestingly, ℓ^∞ is not reflexive, even though ℓ_p and ℓ_q are dual and reflexive whenever $1/p + 1/q = 1$ and $p, q \notin \{1, \infty\}$, and even though for d-dimensional sequences, ℓ_d^∞ is dual to ℓ_d^1.

As a matter of notation,

$$\text{for } f \in \mathcal{B}^* \text{ and } x \in B, \text{ we write } \langle f, x \rangle := f(x),$$

but this is not to be confused with the usual inner-product in which both elements are from a Hilbert space. By definition of the operator norm of f, which is the dual norm of

$\| \cdot \|$, we have $|f(x)| = |\langle f, x\rangle| \leq \|f\|_* \|x\|$, which can be interpreted as a version of Holder's inequality. When equality holds, f, x are called "aligned", and when it equals zero, f, x are called "orthogonal", and this is how in Banach spaces one defines the orthogonal complement $U^\perp \in \mathcal{B}^*$ of a set $U \in \mathcal{B}$.

All Hilbert spaces are self-dual, meaning that its dual space is isomorphic to itself (Riesz representation theorem). Every finite dimensional Hilbert space with dimension n is isomorphic to $\mathbb{C}^n$ (the set of n-dimensional complex vectors with Euclidean inner product). If $\mathcal{H}$ is infinite-dimensional and separable, then it is isomorphic to the set of square summable sequences $\ell^2 := \{(x_n)_{n\in\mathbb{N}} : \sum_{n\in\mathbb{N}} x_n^2 < \infty\}$ endowed with the inner-product $\langle (x_n), (y_n)\rangle = \sum_n \bar{x}_n y_n$. Further, for every infinite-dimensional Hilbert space W, there is a linear operator $W : \mathcal{H} \to \mathcal{H}$ that is defined everywhere but is not bounded. The adjoint A^* of an operator $A : \mathcal{H} \to \mathcal{H}'$ is defined as follows: A^*y is the unique vector such that $\langle Ax, y\rangle = \langle x, A^*y\rangle$.

4 Derivatives

Bounded linear operators are used to extend the concept of derivatives to Banach spaces. A map $f : \mathcal{B} \to \mathcal{B}'$ is said to be Fréchet differentiable at x if there exists a bounded linear operator $A : \mathcal{B} \to \mathcal{B}'$ such that

$$\lim_{h\to 0} \frac{\|f(x+h) - f(x) - Ah\|_{\mathcal{B}'}}{\|h\|_{\mathcal{B}}} = 0.$$

If such an A exists, then it is unique and we write $Df(x) := Ax$. When $\mathcal{B}' = \mathbb{R}$ and f is a function, then $\nabla f := Df$ is a bounded linear functional, we hence we say that

the gradient ∇f of function f is an element of the dual space.

A map f is called Gateaux differentiable at x if has directional derivatives for every direction $u \in \mathcal{B}$, that is if there exists a function $g : \mathcal{B} \to \mathcal{B}'$ such that

$$g(u) = \lim_{h\to 0} \frac{f(x+hu) - f(x)}{u}.$$

Fréchet differentiability implies Gateau differentiability but not vice versa (like over the reals, existence of directional derivatives at x does not imply differentiability at x).

5 Convexity and Smoothness

Using the above notions of differentiability and mnemonic for dot-product in Banach spaces, we are now prepared to define convex and smooth functions on Banach spaces. A function $f : \mathcal{B} \to \mathbb{R}$ is said to be (q, λ)-uniformly convex with respect to the norm $\| \cdot \|$ if

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\lambda t(1-t)}{q}\|x - y\|^q$$

for all $x, y \in \mathcal{B}$ (or in the relative interior of the domain of f) and $t \in (0,1)$. It is a fact that a convex function is differentiable almost everywhere (except at a countable number of points). Hence, an equivalent definition is to require

$$f(y) \geq f(x) + \langle y - x, \nabla f(x) \rangle + \frac{\lambda}{q} \|x - y\|^q,$$

or even

$$\|\nabla f(x) - \nabla f(y)\|_* \geq \lambda \|x - y\|$$

Strong convexity is simply uniform convexity with $q = 2$.

A function f is said to be $(2, L)$-strongly smooth if it is everywhere differentiable and

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\|$$

for all $x, y \in \mathcal{B}$ or equivalently if

$$f(y) \leq f(x) + \langle y - x, \nabla f(x) \rangle + \frac{L}{2} \|x - y\|^2,$$

or even

$$f(tx + (1-t)y) \geq t f(x) + (1-t) f(y) - \frac{Lt(1-t)}{2} \|x - y\|^2.$$

Recall that the Legendre-Fenchel dual $f^* : \mathcal{B}^* \to \mathbb{R}$ of a function $f : \mathcal{B} \to \mathbb{R}$ is defined as

$$f^*(u) = \sup_{x \in \mathcal{B}} \langle u, x \rangle - f(x),$$

where the supremum can be taken over the domain of f if it has a restricted domain.

As a consequence of both convex duality and Banach space duality, we have $f^{**} = f$ iff f is closed and convex, and for such a function

$$f \text{ is } (2, \lambda)\text{-strongly-convex wrt } \|\cdot\| \text{ iff } f^* \text{ is } (2, 1/\lambda)\text{-strongly-smooth wrt } \|\cdot\|_*.$$

As examples, $f(w) := 1/2\|w\|_q^2$ for $w \in \mathbb{R}^d$ is $(2, q-1)$-strongly convex w.r.t. $\|\cdot\|_q$ for $q \in (1, 2]$. An analogous result holds for matrices due to a complex-analysis proof by Ball et al. (1994). (Recall that the Schatten q-norm of a matrix $A \in \mathbb{R}^{m \times n}$ is the q-norm of the singular values of A, denoted as $\|A\|_{S(q)} = \|\sigma(A)\|_q$.) For $q \in (1, 2]$, the function $\frac{1}{2}\|\sigma(A)\|_q^2$ is $(2, q-1)$-strongly-convex with respect to the $\|\sigma(X)\|_q$ norm. For $q = 1$, the following result is known (Kakade, Shalev-Shwartz, Tiwari): defining $q' = \frac{\ln d}{\ln d - 1}$, we have that $\frac{1}{2}\|w\|_{q'}^2$ is $(2, 1/(3 \ln d))$-strongly convex wrt $\|\cdot\|_1$, with an analogous result holding for Schatten norms using $d = \min\{m, n\}$.

6 Martingale type and co-type of a Banach space

Let (Z_t) be a martingale difference sequence (mds) taking values in a Banach space $\mathcal{B}$, meaning that $(\sum_{i=1}^t Z_i)$ is a martingale.

A Banach space $\mathcal{B}$ is said to be of martingale type p if for any $n \geq 1$ and any mds (Z_t), we have

$$\mathbb{E}\|\sum_{i=1}^n Z_i\| \leq C(\mathbb{E}\sum_{i=1}^n \|Z_i\|^p)^{1/p}$$

for some constant $C > 0$. Defining $p^* := \sup\{p : \mathcal{B} \text{ has martingale type } p\}$, we say p^* is the best martingale type of $\mathcal{B}$. It is a fact that

$\mathcal{B}$ has martingale type p iff $\mathcal{B}^*$ has martingale co-type q, where $1/p + 1/q = 1$,

where $\mathcal{B}^*$ having co-type q means that for any $n \geq 1$ and any mds $(Y_t) \in \mathcal{B}^*$, we have

$$(\mathbb{E}\sum_{i=1}^n \|Z_i\|^q)^{1/q} \leq C\mathbb{E}\|\sum_{i=1}^n Z_i\|$$

Theorem 1 (Pisier) *A Banach space $\mathcal{B}^*$ has martingale co-type q iff there exists a (q, λ)-uniformly convex function on $\mathcal{B}^*$ for some $\lambda > 0$. As an important corollary for $q = 2$, a Banach space $\mathcal{B}$ has martingale type 2 iff there exists a $(2, L)$ strongly smooth function on $\mathcal{B}$ for some $L > 0$.*

There are other equivalent definitions of a strongly smooth functions, as we shall see in the next section.

7 Concentration for $(2, D)$-strongly smooth functions

The applications presented thus far allow us to uniformly bound the operator norm deviations of a sequence of random Hermitian matrices in $\mathcal{C}^{d\times d}$. A different approach is due to Pinelis (1992, 1994). For this section, let $(Y_t)_{t\in\mathcal{N}}$ be a martingale with respect to $(\mathcal{F}_t)$ taking values in a separable Banach space $(\mathcal{X}, \|\cdot\|)$. We can use Pinelis's device to uniformly bound the process $(\Psi(Y_t))$ for any function $\Psi : \mathcal{X} \to \mathbb{R}$ which satisfies the following smoothness property:

Definition 1 (Pinelis 1994) *A function $\Psi : \mathcal{X} \to \mathbb{R}$ is called* $(2, D)$-smooth *for some $D > 0$ if, for all $x, v \in \mathcal{X}$, we have*

$$\Psi(0) = 0 \tag{1a}$$

$$|\Psi(x + v) - \Psi(x)| \leq \|v\| \tag{1b}$$

$$\Psi^2(x + v) - 2\Psi^2(x) + \Psi^2(x - v) \leq 2D^2\|v\|^2. \tag{1c}$$

A Banach space is called $(2, D)$-smooth if its norm is $(2, D)$-smooth; in such a space we may take $\Psi(\cdot) = \|\cdot\|$ to uniformly bound the deviations of a martingale. In this case, observe that property (1a) is part of the definition of a norm, property (1b) is the triangle inequality, and property (1c) can be seen to hold with $D = 1$ for the norm induced by the inner product in any Hilbert space, regardless of the (possibly infinite) dimensionality of the space. Note also that setting $x = 0$ shows that $D \geq 1$ whenever $\Psi(\cdot) = \|\cdot\|$.

Corollary 2 *Consider a martingale $(Y_t)_{t\in\mathcal{N}}$ taking values in a separable Banach space $(\mathcal{X}, \|\cdot\|)$. Let the function $\Psi : \mathcal{X} \to \mathbb{R}$ be $(2, D)$-smooth and define $D_\star := 1 \vee D$.*

1. *Suppose $\|\Delta Y_t\| \leq c_t$ a.s. for all $t \in \mathcal{N}$ for some constants $(c_t)_{t\in\mathcal{N}}$, and let $V_t := \sum_{i=1}^t c_i^2$. Then for any $x, m > 0$, we have*

$$\mathcal{P}\left(\exists t \in \mathcal{N} : \Psi(Y_t) \geq x + \frac{D_\star^2 x}{2m}(V_t - m)\right) \leq 2\exp\{-\frac{x^2}{2D_\star^2 m}\}. \tag{2}$$

   This strengthens Theorem 3.5 from Pinelis (1994) [B].

2. *Suppose $\|\Delta Y_t\| \leq c$ a.s. for all $t \in \mathcal{N}$ for some constant c, and let $V_t := \sum_{i=1}^t \mathbb{E}_{i-1}\|\Delta Y_i\|^2$. Then for any $x, m > 0$, we have*

$$\begin{aligned}\mathcal{P}\left(\exists t \in \mathcal{N} : \Psi(Y_t) \geq x + D_\star^2 \mathfrak{s}_P\left(\frac{x}{m}\right) \cdot (V_t - m)\right) &\leq 2\exp\{-D_\star^2 m \psi_P^\star\left(\frac{x}{D_\star^2 m}\right)\} \\ &\leq 2\exp\{-\frac{x^2}{2(D_\star^2 m + cx/3)}\}. \end{aligned} \tag{3}$$

   This strengthens Theorem 3.4 from Pinelis (1994) [B].

As before, the Hoeffding-style bound in part (a) and the Bennett-style bound in part (b) are not directly comparable: V_t may be smaller in part (b), but the exponent is also smaller.

We briefly highlight some of the strengths and limitations of this approach. Since the Euclidean l_2-norm is induced by the standard inner product in $\mathbb{R}^d$, the above corollary gives a dimension-free uniform bound on the l_2-norm deviations of a vector-valued martingale in $\mathbb{R}^d$ which exactly matches the form for scalars. Compare this to bounds based on the operator norm of a Hermitian dilation: the bound of Tropp (2012) includes dimension dependence [B,E] while the bound of Minsker (2017, Corollary 4.1) incurs an extra constant factor of 14 [B,E]. Our bounds extend to martingales taking values in sequence space $\{(a_i)_{i\in\mathcal{N}} : \sum_i |a_i|^2 < \infty\}$ or function space $L^2[0, 1]$, and we may instead use the l_p norm, $p \geq 2$, in which case $D = \sqrt{p-1}$. These cases follow from Pinelis (1994, Proposition 2.1).

Similarly, the above corollary gives dimension-free uniform bounds for the Frobenius norm deviations of a matrix-valued martingale. This extends to martingales taking values in a space of Hilbert-Schmidt operators on a separable Hilbert space, with deviations bounded in the Hilbert-Schmidt norm; compare Minsker (2017, S3.2), which gives operator-norm bounds. The method of the above corollary does not extend directly to operator-norm bounds because the operator norm is not $(2, D)$-smooth for any D.

References

Minsker, S. (2017), 'On Some Extensions of Bernstein's Inequality for Self-adjoint Operators', *Statistics and Probability Letters* **127**, 111–119.

Pinelis, I. (1992), An Approach to Inequalities for the Distributions of Infinite-Dimensional Martingales, *in* 'Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference', Birkhäuser, Boston, MA, pp. 128–134.

Pinelis, I. (1994), 'Optimum Bounds for the Distributions of Martingales in Banach Spaces', *The Annals of Probability* **22**(4), 1679–1706.

Tropp, J. A. (2012), 'User-friendly tail bounds for sums of random matrices', *Foundations of Computational Mathematics* **12**(4), 389–434.