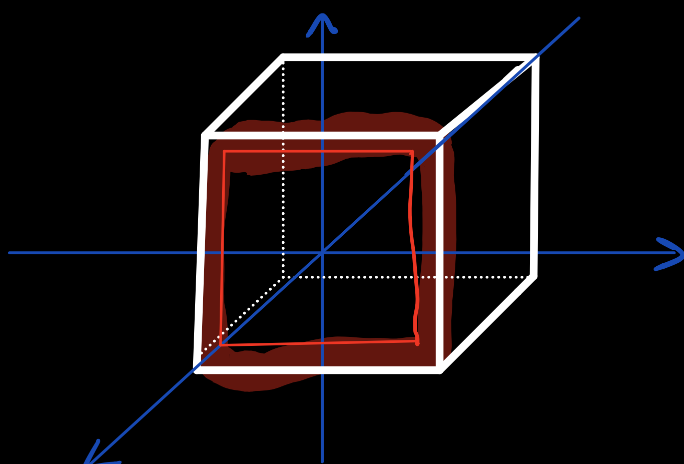
Concentration inequalities for the sums of random matrices, and their applications

- What is "concentration of measure"?

  Consider a unit cube $D_R = \{ (x_1, \dots, x_d) : |x_j| \le R, j = 1, \dots, d \}$

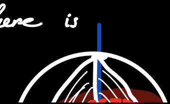


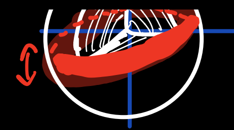
Then $\text{Volume}(D_R) = (2R)^d$

$\text{Volume}(D_{(1-\varepsilon)R}) = (2(1-\varepsilon)R)^d$

$$\Rightarrow \frac{\text{Vol}(D_{(1-\varepsilon)R})}{\text{Vol}(D_R)} = (1-\varepsilon)^d \to 0 \text{ as } d \to \infty.$$

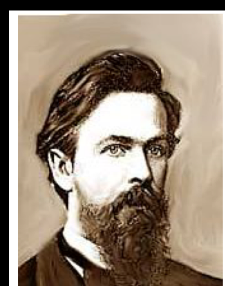
Exercise: (a) show that the "volume" of a high-dimensional ball concentrates in a thin "shell" near its boundary

(b) show that the surface area of a high-dimensional sphere is concentrated in a thin "band" around the equator: 



- <u>Probability theory</u>: let X be a random variable such that $P(X \geq 0) = 1$. Then $P(X \geq t) \leq \frac{\mathbb{E}X}{t}$ – <u>Markov's inequality</u>

<u>Proof</u>: $\mathbb{E}X = \mathbb{E}\, X \cdot I\{X \leq t\} + \mathbb{E}X \cdot I\{X \geq t\} \geq \mathbb{E}X \cdot I\{X \geq t\}$
$\geq t\ \mathbb{E}\, I\{X \geq t\}$



Andrey Markov
(1856 - 1922)

- Now let X be a random variable such that $\mathbb{E}X^2 < \infty$.

Take $Y := |X - \mathbb{E}X|$. Then

$$P(|X - \mathbb{E}X| \geq t) = P(|X - \mathbb{E}X|^2 \geq t^2) \leq \frac{Var(X)}{t^2}$$

- Chebyshev's inequality

Pafnuty Chebyshev
(1821 - 1894)

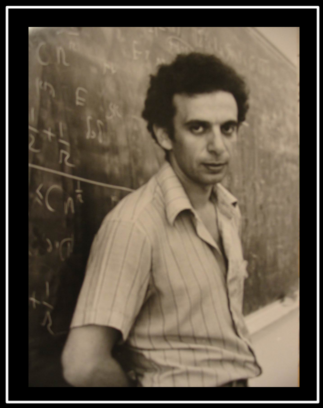
<u>Exercise</u>: (a) show that Chebyshev's inequality can not be uniformly improved, that is, for any t, find a random variable X such that
$P(|X - \mathbb{E}X| \geq t \cdot Var(X)) = \frac{1}{t^2}$

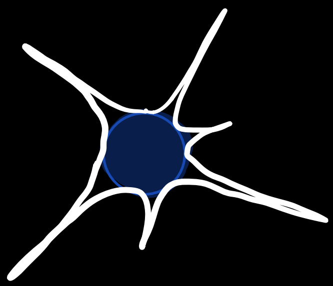
(b) Prove the Chebyshev - Cantelli inequality:

$$P(X - \mathbb{E}X \geq t) \leq \frac{Var(Y)}{Var(Y)+t^2}$$

- The term "concentration of measure" was put forward by Vitali Milman in the 1970s

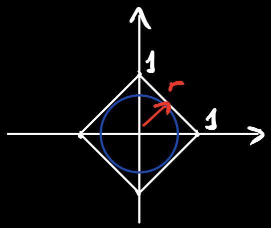
- Milman studied the geometry of high-dimensional convex bodies: according to Milman, they look like this:





Example: ℓ_1-ball in $\mathbb{R}^d$

$$B^d_{\ell_1}(R) = \{ (x_1, .., x_d) : \sum_{j=1}^{d} |x_j| \leq R \}$$



$$Vol(B^d_{\ell_1}(1)) = \frac{2^d}{d!} \text{ (proof below)}$$

$r = ?$

$\sum_{j=1}^{d} x_j^2 \longrightarrow$ minimize subject to $x_j \geq 0$, $\sum_{j=1}^{d} x_j = 1$.

$$x_j = \frac{1}{d} \quad \forall j \quad \Rightarrow \left(\sum_{j=1}^{d} x_j^2 \right)^{1/2} = r = \frac{1}{\sqrt{d}}$$

Consider random vector $X = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$ uniformly distributed in $B^d_{\ell_1}(1)$. Then

$$P(X_1 \le t) = \int_{-1}^{t} \int_{\substack{x_2,\dots,x_d : \\ \sum_{j=2}^{d} |x_j| \le 1 - |x_1|}} dx_2 \dots dx_d \cdot dx_1 \cdot \frac{1}{Vol\left(B^d_{\ell_1}(1)\right)}$$

$$= \int_{-1}^{t} Vol\left(B^{d-1}_{\ell_1}(1-|x_1|)\right) dx_1 \cdot \frac{1}{Vol\left(B^d_{\ell_1}(1)\right)}$$

$$= \underbrace{\int_{-1}^{t} (1-|x_1|)^{d-1} \cdot dx_1}_{\frac{2}{d}} \cdot \frac{Vol\left(B^{d-1}_{\ell_1}(1)\right)}{Vol\left(B^d_{\ell_1}(1)\right)}$$

Set $t = 1 \Rightarrow P(X_1 \le 1) = 1$

$$\Rightarrow \frac{Vol\left(B^{d-1}_{\ell_1}(1)\right)}{Vol\left(B^d_{\ell_1}(1)\right)} = \frac{d}{2}$$

$$\Rightarrow p_{X_1}(t) = \frac{d}{2}\left(1 - |t|\right)^{d-1}$$

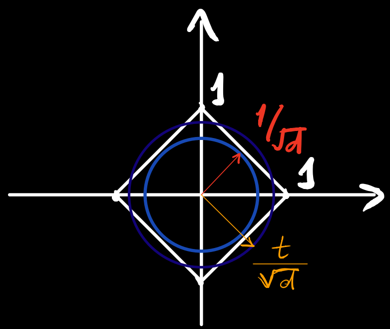
$$\Rightarrow \mathbb{E} X_1^2 = (\text{Exercise}) = \frac{2}{(d+1)(d+2)}$$

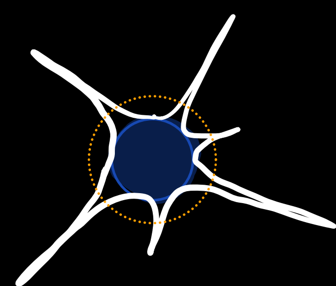
$$\Rightarrow \mathbb{E} \|X\|_2^2 = \frac{2d}{(d+1)(d+2)}$$

$$\Rightarrow \mathbb{E} \|X\|_2 \le \left(\mathbb{E} \|X\|_2^2\right)^{1/2} = \sqrt{\frac{2}{d+2}} \sqrt{\frac{d}{d+1}} \sim \sqrt{\frac{2}{d}}$$

$$P\left(\|X\|_2 \geq \frac{t}{\sqrt{d}}\right) \leq \frac{\mathbb{E}\|X\|_2^2}{t^2}\, d = \frac{2d^2}{t^2(d+1)(d+2)}$$

$$\sim \frac{2}{t^2} \text{ as } d \to \infty.$$





- What if we have a collection of random variables $\{X_t\}_{t\in T}$, and we want to estimate $P\left(\sup_{t\in T} X_t \geq z\right)$?

  <u>Example</u>: let $Y \in \mathbb{R}^{d\times d}$ be a random matrix, and

  $X_v := \|Yv\|_2$, $\|v\|_2 = 1$.

  Then $\sup_{v : \|v\|_2 = 1} X_v = ?$ $\|Y\|$ – operator (spectral) norm of Y.

- <u>More general problem</u>: let $X_1, \dots, X_n$ be independent random elements with values in S, and $f : S^d \to \mathbb{R}$. Find bounds for $P\left(|f(X_1,\dots,X_n) - \mathbb{E}f(X_1,\dots,X_n)| > t\right)$.

It turns out that the following "bounded difference" property yields interesting results: assume that for all j,

$$\sup_{x_1,\dots,x_j,x_j',\dots,x_n} |f(x_1,\dots,x_j,\dots,x_n) - f(x_1,\dots,x_j',\dots,x_n)| \le c_j$$

__Example__ Assume that $Y \in \mathbb{R}^{d\times d}$ has bounded entries, and

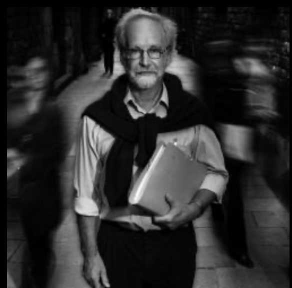
$$f(Y) = f(Y_{11},\dots,Y_{dd}) = \sup_{\|v\|_2=1} \|Yv\|_2 .$$

The bounded difference property holds with $c_j = ?$

[let Y and Y' differ in (i,j)th coordinate. Then $\big| \|Y\| - \|Y'\| \big| \le \|Y - Y'\| = M$ where M is the sup-norm bound for Y_{ij}].

__Theorem__ (McDiarmid's inequality)
Let $X_1,\dots,X_n \in S$ be independent, and assume that f has bounded differences. Then $\forall t \ge 0$,

$$P(|f(X_1,\dots,X_n) - \mathbb{E}f| \ge t) \le 2e^{-\frac{2t^2}{\sum_1^n c_j^2}}$$



Colin McDiarmid

__Proof__ (of a slightly weaker bound)

"Martingale decomposition" technique:

Let $Y_1,\dots,Y_n$ be an independent copy of $X_1,\dots,X_n$, and set

$$V_j := \mathbb{E}_Y\left[f(X_1,\dots,X_j,Y_{j+1},\dots,Y_n) - f(X_1,\dots,X_{j-1},Y_j,\dots,Y_n)\right]$$

Note that (a) $|V_j| \le c_j$, $\mathbb{E} V_j = 0$

(b) $f(X_1, \dots, X_n) - \mathbb{E} f = \sum_{j=1}^{n} V_j$

Next, $P(f(X_1, \dots, X_n) - \mathbb{E} f \geq t) = P(\lambda(f - \mathbb{E} f) \geq \lambda t)$

$$= P\left(e^{\lambda(f - \mathbb{E} f)} \geq e^{\lambda t}\right) \leq \mathbb{E}\, e^{\lambda(f - \mathbb{E} f)} \cdot e^{-\lambda t}$$

$$\mathbb{E}\, e^{\lambda(f - \mathbb{E} f)} = \mathbb{E}\, e^{\lambda \sum_{j=1}^{n} V_j} = \mathbb{E}\, \mathbb{E}\left[e^{\lambda \sum_{j=1}^{n-1} V_j + \lambda V_n} \,\middle|\, X_1, \dots, X_{n-1}\right]$$

$$= \mathbb{E}\, e^{\lambda \sum_{j=1}^{n-1} V_j} \underbrace{\mathbb{E}\, e^{\lambda V_n}}_{\leq \exp\left(\frac{\lambda^2 c_n^2}{2}\right)} \leq \dots \leq \mathbb{E}\, e^{\frac{\lambda^2}{2} \sum_1^n c_j^2}$$

$G(\lambda) = \frac{\lambda^2}{2} \sum_{j=1}^{n} c_j^2 - \lambda t \longrightarrow$ minimize over $\lambda > 0$, $\lambda_* = \dfrac{t}{\sum_1^n c_j^2}$

<u>Lemma</u> Let X be a random variable such that $\mathbb{E} X = 0$ and $P(a \leq X \leq b) = 1$. Then $\mathbb{E}\, e^{\lambda X} \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right)$

<u>Proof</u>: (Exercise)

(i) write $X = \frac{b - X}{b - a} \cdot a + \frac{X - a}{b - a} \cdot b$, and deduce that $e^{\lambda X} \leq \frac{b - X}{b - a} e^{\lambda a} + \frac{X - a}{b - a} e^{\lambda b}$,

$\mathbb{E}\, e^{\lambda X} \leq ?$

(ii) let $p = \frac{b}{b - a}$, $h = \lambda(b - a)$,

$F(h) = ph + \log(1 - p + pe^{-h})$, and show that

$F'(0) = 0$, $F''(h) \leq \frac{1}{4}$ so that $F(h) \leq \frac{h^2}{8}$.

Deduce the result from (i) and (ii).

Exercise: let $X_1, \dots, X_n \in \mathbb{R}^d$ be iid random vectors s.t. $\mathbb{E}X_1 = 0$, $\|X_1\| \le M$ with probability 1, and let $\hat{\Sigma}_n := \frac{1}{n}\sum_{j=1}^{n} X_j X_j^T$ be the sample covariance matrix.

Verify the bounded difference property for

$$f(X_1, \dots, X_n) := \|\hat{\Sigma}_n - \Sigma\|, \text{ where } \Sigma = \mathbb{E}(X_1 X_1^T)$$

- McDiarmid's inequality only controls the deviation of $f(X_1, \dots, X_n)$ around $\mathbb{E} f(X_1, \dots, X_n)$. But how can one estimate $\mathbb{E} f(X_1, \dots, X_n)$?
This is a separate (difficult) question.

- McDiarmid's inequality does not take into account the variance $\mathrm{Var}(f(X_1, \dots, X_n))$, hence it often yields the bounds that can be improved; e.g., for $|X_j| \le M$ and $f(x_1, \dots, x_n) = \frac{1}{n}\sum_{j=1}^{n} x_j$, we get

$$P\left(\left|\frac{1}{n}\sum_{j=1}^{n} X_j - \mathbb{E}X\right| \ge \sqrt{\frac{t}{n}}\right) \le 2e^{-\frac{t}{2M^2}}.$$

Exercise: compare this with the Chernoff bound for Bernoulli random variables ($P(X=1) = p$, $P(X=0) = 1-p$) for p close to 0 or 1. $\left[P\left(\frac{\sum_1^n X_j}{n} - p \ge \delta p\right) \le e^{-\frac{\delta^2 np}{3}} \text{ - Chernoff Bound}\right]$.

- Dependence on the variance is improved by the celebrated Talagrand's concentration inequality.

Matrix concentration inequalities

- Goal: given a sequence $X_1, \dots, X_n$ of independent random matrices, estimate $$P\left(\left\|\sum_{j=1}^{n} X_j - \mathbb{E}\left(\sum_{j=1}^{n} X_j\right)\right\| \geq t\right),$$ where $\|\cdot\|$ is the operator norm, in terms of known quantities.

- Idea: instead of viewing $\left\|\sum_{j=1}^{n} X_j - \mathbb{E}\left(\sum_{j=1}^{n} X_j\right)\right\|$ as a supremum of a stochastic process
$$v \to \left\langle \sum_{j=1}^{n} X_j - \mathbb{E}\left(\sum_{j=1}^{n} X_j\right), v \right\rangle, \quad \|v\|_2 = 1,$$
view it as a norm of a matrix.

- Recall that, given a random variable Z such that $P(Z \geq 0) = 1$,
$$\mathbb{E} Z = \int_0^\infty P(Z \geq t)\, dt.$$
Assume that we can find the desired bound. Then
$$\mathbb{E}\left\|\sum_{j=1}^{n} X_j - \mathbb{E}\left(\sum_{j=1}^{n} X_j\right)\right\| = \int_0^\infty P\left(\left\|\sum_{j=1}^{n} X_j - \mathbb{E}\left(\sum_{j=1}^{n} X_j\right)\right\| \geq t\right) dt$$
$\Rightarrow$ we get a bound for the expectation!

<u>Preliminaries:</u> let $M \in \mathbb{C}^{d \times d}$ be a matrix with complex entries. Then

(a) M is called <u>self-adjoint</u> if

$$\langle Mx, y \rangle = \langle x, My \rangle \quad \forall x, y$$

$$\Longleftrightarrow \quad \overline{M}^T = M^* = M$$

(b) $\text{Trace}(M) = tr(M) = \sum_{j=1}^{d} M_{jj} = \langle \text{exercise} \rangle = \sum_{j=1}^{d} \langle Me_j, e_j \rangle$ for <u>any</u> ONB basis $\{e_1, .., e_d\}$.

$$\| M \| = \sup_{v \neq 0} \frac{\langle Mv, v \rangle}{\|v\|^2} = \max\left(\lambda_{max}(M), -\lambda_{min}(M) \right)$$

(c) If M is self-adjoint, then

$$M = Q \Lambda Q^*,$$

where Q is a unitary matrix $(QQ^* = I_d)$

and $\Lambda = diag(\lambda_1, .., \lambda_d) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix}$, <u>$\lambda_j$'s $\in \mathbb{R}$</u>

(d) A self-adjoint matrix M is called non-negative definite if $\langle v, Mv \rangle \geq 0 \quad \forall v \in \mathbb{C}^d$.

We write that $M \succcurlyeq 0$.

Similarly, $M_1 \succcurlyeq M_2 \Longleftrightarrow M_1 - M_2 \succcurlyeq 0$.

(e) If $M_1 \succcurlyeq M_2$ then $A^* M_1 A \succcurlyeq A^* M_2 A$ for any $A \in \mathbb{C}^{d \times d}$. Moreover, $tr\, M_1 \geq tr\, M_2$

<u>Functions of self-adjoint matrices</u>:

(a) Let $f: I \to \mathbb{R}$ be a function, and assume that $\lambda_j(M) \in I \subseteq \mathbb{R}$, $j = 1, .., d$, where $M = Q \Lambda Q^*$.

Then $f(M) = Q f(\Lambda) Q^*$ and

$$f(\Lambda) = \begin{pmatrix} f(\lambda_1) & & 0 \\ & \ddots & \end{pmatrix}$$

$$\Big(0 \qquad \ddots \; f(\lambda_d) \Big)$$

(b) "Transfer rule": if $f(x) \geq g(x)$ $\forall x$, then $f(M) \succeq g(M)$ for any self-adjoint $M \in \mathbb{C}^{d \times d}$.

(c) Assume that $M_1 \preceq M_2$ and f is a monotone function.
Then $\operatorname{tr} f(M_1) \leq \operatorname{tr} f(M_2)$

<u>Example</u>
Consider the function $f(x) = e^x$.
Then $M_1 \preceq M_2 \Rightarrow \operatorname{tr} e^{M_1} \leq \operatorname{tr} e^{M_2}$

But it is <u>not true</u> that $e^{M_1} \preceq e^{M_2}$! (find an example)

(d) <u>Operator monotone functions</u>
f is operator monotone if $A \preceq B \Rightarrow f(A) \preceq f(B)$

$f : \mathbb{R} \to \mathbb{R}$ is monotone $\not\Rightarrow$ f is operator monotone!

<u>Example</u>: $f(x) = x^2$, $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, $B = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$

<u>However</u>, log is operator monotone:
$$0 \prec A \preceq B \Rightarrow \log(A) \preceq \log(B)$$

<u>Proof</u>: (i) $\displaystyle \log A = \int_0^\infty \left(\frac{1}{1+u} I - (A + uI)^{-1} \right) du$

Indeed, $\displaystyle \log a = \int_0^\infty \left(\frac{1}{1+u} - \frac{1}{a+u} \right) du$

(ii) $0 \prec A \preceq B \Rightarrow -(A+uI)^{-1} \preceq -(B+uI)^{-1}$ $\forall u \geq 0$

Indeed, $A+uI \preceq B+uI$

$$\Rightarrow \quad (B+uI)^{-1/2}(A+uI)(B+uI)^{-1/2} \preceq I$$

$$\Rightarrow \quad (B+uI)^{1/2}(A+uI)^{-1}(B+uI)^{1/2} \succeq I$$

$$\Rightarrow \quad (A+uI)^{-1} \succeq (B+uI)^{-1}$$

(iii) <u>Exercise</u>: Show that $\log(A) \preceq \log(B)$ using (i) and (ii).