Concentration Inequalities 101

Clément Lalanne

Notation

Sample. $S = (X_1, \ldots, X_n)$ drawn i.i.d. from $\mathbb{P}$, so $S \sim \mathbb{P}^{\otimes n}$. Measures are capital blackboard letters, densities are lowercase: p, p_θ.

Subscripts. Every expectation, variance and supremum carries what it is taken over: $\mathbb{E}_{X\sim\mathbb{P}}$, $\mathrm{Var}_{X\sim\mathbb{P}}$, $\mathbb{E}_{S\sim\mathbb{P}^{\otimes n}}$. When the law is clear we abbreviate to $\mathbb{E}_X$, Var_X, $\mathbb{E}_S$, and this is the only abbreviation used.

Sample dependence. Anything random and sample dependent carries the sample: $\widehat{R}_S(f)$, $\hat{f}_S$.

Centering. Unless said otherwise X is centered, $\mathbb{E}_X[X] = 0$, and $S_n = \sum_{i\leq n} X_i$.

Part I. Markov and the exponential method

Markov's inequality

Theorem. Let $Z \geq 0$ almost surely and $t > 0$. Then

$$\mathbb{P}_Z(Z \geq t) \leq \frac{\mathbb{E}_Z[Z]}{t}.$$

Proof.

$$\begin{aligned} t\mathbf{1}\{Z \geq t\} &\leq Z && \text{pointwise, using } Z \geq 0 \\ t\mathbb{P}_Z(Z \geq t) = \mathbb{E}_Z[t\mathbf{1}\{Z \geq t\}] &\leq \mathbb{E}_Z[Z] && \text{monotonicity of } \mathbb{E}_Z \end{aligned}$$

Remark. Only nonnegativity is used, and the bound is attained: if Z equals t with probability q and 0 otherwise, both sides equal q. Every improvement below comes from feeding Markov a better variable, never from improving Markov.

Markov is the mother of the others

Proposition. Let $\phi : \mathbb{R} \to [0, \infty)$ be nondecreasing with $\phi(t) > 0$. Then $\{Z \geq t\} \subseteq \{\phi(Z) \geq \phi(t)\}$, so Markov applied to $\phi(Z)$ gives

$$\mathbb{P}_Z(Z \geq t) \leq \frac{\mathbb{E}_Z[\phi(Z)]}{\phi(t)}.$$

Changing the function changes the bound. $\phi(z) = z^2$ applied to $|Z - \mathbb{E}_Z[Z]|$ gives Chebyshev. $\phi(z) = z^p$ applied to $|Z|$ gives the moment method. $\phi(z) = e^{\lambda z}$ with $\lambda > 0$ gives the Chernoff method, and with it Hoeffding, Bernstein and Bennett.

Where the work goes. The left side is fixed. All the difficulty is in bounding $\mathbb{E}_Z[\phi(Z)]$, and the three regimes of this lecture are three ways of bounding it.

Chebyshev, and why it is not enough

Corollary. With $\sigma^2 = \mathrm{Var}_X(X)$ finite,

$$\mathbb{P}_X(|X - \mathbb{E}_X[X]| \geq t) \leq \frac{\sigma^2}{t^2},$$

by Markov applied to $\phi(z) = z^2$ and $Z = |X - \mathbb{E}_X[X]|$.

Why it is weak for sums. For S_n a sum of n i.i.d. centered variables, $\mathrm{Var}_S(S_n) = n\sigma^2$, so Chebyshev gives $\mathbb{P}_S(|S_n| \geq t) \leq n\sigma^2/t^2$. At $t = u\sqrt{n}\sigma$ this is u^{-2}, a bound that decays polynomially in u while the central limit theorem suggests $e^{-u^2/2}$. Two moments buy a polynomial rate. Exponential rates require exponential moments.

The Cramer-Chernoff method

Definition. The log-Laplace transform, or log-MGF, of X is $\psi_X(\lambda) = \log \mathbb{E}_X[e^{\lambda X}] \in (-\infty, +\infty]$, defined on $D_X = \{\lambda : \psi_X(\lambda) < \infty\}$.

Theorem. For every $t > 0$,

$$\mathbb{P}_X(X \geq t) \leq e^{-\lambda t}\mathbb{E}_X[e^{\lambda X}] = e^{-(\lambda t - \psi_X(\lambda))} \quad \text{Markov with } \phi(z) = e^{\lambda z},\ \lambda > 0$$

$$\mathbb{P}_X(X \geq t) \leq \exp(-\sup_{\lambda > 0}\{\lambda t - \psi_X(\lambda)\}) \quad \text{the left side does not depend on } \lambda$$

The free parameter λ is the whole point: we get one bound per λ and keep the best.

The log-MGF, four facts

Proposition. On the interior of D_X, with $\mathbb{E}_X[X] = 0$:

1. ψ_X is convex and $\psi_X(0) = 0$.

Proof. $\psi_X(0) = \log 1 = 0$; for $\lambda = \alpha\lambda_1 + (1-\alpha)\lambda_2$ with $\alpha \in [0,1]$, Holder with exponents $1/\alpha, 1/(1-\alpha)$ gives $\mathbb{E}_X[e^{\lambda X}] \le \mathbb{E}_X[e^{\lambda_1 X}]^\alpha \mathbb{E}_X[e^{\lambda_2 X}]^{1-\alpha}$.

2. Exponential tilting: let $d\mathbb{P}_\lambda = e^{\lambda x - \psi_X(\lambda)} d\mathbb{P}$. Then $\psi'_X(\lambda) = \mathbb{E}_{X\sim\mathbb{P}_\lambda}[X]$ and $\psi''_X(\lambda) = \mathrm{Var}_{X\sim\mathbb{P}_\lambda}(X)$.

Proof. Differentiate under the integral, legitimate on the interior of D_X: $\psi'_X(\lambda) = \mathbb{E}_X[Xe^{\lambda X}]/\mathbb{E}_X[e^{\lambda X}] = \mathbb{E}_{X\sim\mathbb{P}_\lambda}[X]$, and one more derivative turns the quotient into $\mathbb{E}_{\mathbb{P}_\lambda}[X^2] - \mathbb{E}_{\mathbb{P}_\lambda}[X]^2$.

3. Consequently $\psi'_X(0) = \mathbb{E}_X[X] = 0$ and $\psi''_X(0) = \mathrm{Var}_X(X) = \sigma^2$, so $\psi_X(\lambda) = \frac{\sigma^2\lambda^2}{2} + o(\lambda^2)$ near 0.

Proof. Fact 2 at $\lambda = 0$, where $\mathbb{P}_0 = \mathbb{P}$; then Taylor at 0 with $\psi_X(0) = \psi'_X(0) = 0$.

4. Tensorization: if $X_1, \ldots, X_n$ are independent then $\psi_{S_n} = \sum_{i\le n} \psi_{X_i}$.

Proof. $\mathbb{E}_S[e^{\lambda S_n}] = \prod_{i\le n} \mathbb{E}_{X_i}[e^{\lambda X_i}]$ by independence.

The Cramer-Legendre transform

Definition. The Cramer transform of X is $\psi_X^*(t) = \sup_{\lambda>0}\{\lambda t - \psi_X(\lambda)\}$, the Legendre transform of ψ_X restricted to $\lambda > 0$. The Chernoff bound reads $\mathbb{P}_X(X \geq t) \leq e^{-\psi_X^*(t)}$.

Properties. ψ_X^* is convex and nonnegative as a supremum of affine functions vanishing at $\lambda = 0$; it is nondecreasing on $t \geq 0$; and for centered X it satisfies $\psi_X^*(t) = \frac{t^2}{2\sigma^2} + o(t^2)$ as $t \to 0$, the dual statement of fact 3.

Proof. Each $t \mapsto \lambda t - \psi_X(\lambda)$ is affine, hence convex, and increasing since $\lambda > 0$; a pointwise supremum inherits both, so ψ_X^* is convex and nondecreasing on $t \geq 0$. Letting $\lambda \to 0^+$ in the supremum gives $\psi_X^*(t) \geq 0$, as $\psi_X(0) = 0$. For the expansion, substitute $\psi_X(\lambda) = \frac{\sigma^2\lambda^2}{2} + o(\lambda^2)$: the leading part is maximised at $\lambda^\star = t/\sigma^2 \to 0$ with value $\frac{t^2}{2\sigma^2}$, and the remainder at $\lambda^\star$ is $o(t^2)$.

Remark. When ψ_X is differentiable and strictly convex, the supremum is attained at the $\lambda^\star$ solving $\psi_X'(\lambda^\star) = t$: the optimal tilt is the one that moves the mean of the tilted law exactly onto t. Bounding a deviation and reweighting the law until the deviation becomes typical are the same operation.

Empirical means. The transform is built for $\frac{1}{n}S_n$. For i.i.d. X_i, tensorization gives $\psi_{S_n} = n\psi_X$, so $\psi_{S_n}^*(nt) = n\psi_X^*(t)$ and $\mathbb{P}_S(\frac{1}{n}S_n \geq t) = \mathbb{P}_S(S_n \geq nt) \leq e^{-n\psi_X^*(t)}$.

Why the exponent is the right object

Theorem (Cramer, 1938). For $X_1, \ldots, X_n$ i.i.d. with ψ_X finite in a neighbourhood of 0,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_S\left(\tfrac{1}{n} S_n \geq t\right) = -\psi_X^*(t), \quad t > 0.$$

The upper bound is exactly the Chernoff bound of the previous slide, applied to S_n and divided by n. Proof of the lower bound omitted.

Consequence. The Chernoff method is exact at the exponential scale. Nothing is lost by working with ψ^*, and any weakness in what follows comes from bounding ψ crudely, not from the method.

Part II. The sub-gaussian regime

Sub-gaussian variables

Definition. A centered X is sub-gaussian with variance proxy σ^2, written $X \in \mathcal{G}(\sigma^2)$, if

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

Consequence. By Chernoff and the same bound applied to $-X$,

$$\mathbb{P}_X(X \geq t) \leq e^{-t^2/(2\sigma^2)}, \quad \mathbb{P}_X(|X| \geq t) \leq 2e^{-t^2/(2\sigma^2)} \quad \text{for all } t > 0.$$

The defining bound holds for every λ, so the optimal $\lambda^\star = t/\sigma^2$ is always admissible and the tail is gaussian at every scale. Remember this sentence: the entire difference with the next regime is that there $\lambda^\star$ can be forbidden.

Equivalent definitions

Theorem. For centered X the following are equivalent up to absolute constants, meaning each K_i is bounded by a universal multiple of each other K_j:

1. $\mathbb{P}_X(|X| \geq t) \leq 2e^{-t^2/K_1^2}$ for all $t > 0$.
2. $(\mathbb{E}_X|X|^p)^{1/p} \leq K_2\sqrt{p}$ for all $p \geq 1$.
3. $\mathbb{E}_X[e^{X^2/K_3^2}] \leq 2$.
4. $\psi_X(\lambda) \leq K_4^2\lambda^2/2$ for all $\lambda \in \mathbb{R}$.

The proof is the cycle $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4 \Rightarrow 1$ on the next three slides.

Remark. Item 2 is the one to remember for Part V: sub-gaussian means moments growing like $\sqrt{p}$.

Equivalent definitions, proof: $1 \Rightarrow 2 \Rightarrow 3$

All constants are absolute; $a \lesssim b$ means $a \leq Cb$ for a universal C.

$1 \Rightarrow 2$. For $p \geq 2$,

$$\begin{aligned} \mathbb{E}_X|X|^p &= \int_0^\infty p t^{p-1} \mathbb{P}_X(|X| \geq t) dt \leq 2p \int_0^\infty t^{p-1} e^{-t^2/K_1^2} dt && \text{item 1} \\ &= p K_1^p \Gamma(\tfrac{p}{2}) \leq p K_1^p (p/2)^{p/2} && \Gamma(x) \leq x^x \ (x \geq 1) \end{aligned}$$

so $\|X\|_p \leq p^{1/p} K_1 \sqrt{p/2} \lesssim K_1 \sqrt{p}$. The range $1 \leq p < 2$ follows because $p \mapsto \|X\|_p$ is nondecreasing. Hence $K_2 \lesssim K_1$.

$2 \Rightarrow 3$. Now, by item 2 at $p = 2k$ and since $k! \geq (k/e)^k$:

$$\mathbb{E}_X[e^{X^2/K^2}] = \sum_{k \geq 0} \frac{\mathbb{E}_X[X^{2k}]}{k! K^{2k}} \leq 1 + \sum_{k \geq 1} \frac{(K_2 \sqrt{2k})^{2k}}{(k/e)^k K^{2k}} = 1 + \sum_{k \geq 1} (2eK_2^2/K^2)^k.$$

Take $K^2 = 4eK_2^2$: the series is $\sum_{k \geq 1} 2^{-k} = 1$, so $\mathbb{E}_X[e^{X^2/K^2}] \leq 2$ and $K_3 \leq 2\sqrt{e} K_2$.

Equivalent definitions, proof: $3 \Rightarrow 4$

$3 \Rightarrow 4$. From $\lambda x \leq \frac{x^2}{K_3^2} + \frac{\lambda^2 K_3^2}{4}$,

$$\mathbb{E}_X[e^{\lambda X}] \leq e^{\lambda^2 K_3^2/4} \mathbb{E}_X[e^{X^2/K_3^2}] \leq 2e^{\lambda^2 K_3^2/4} \quad (\text{all } \lambda). \quad (*)$$

The factor 2 is removed for small λ by centering. Using $e^u \leq u + e^{u^2}$, then $\mathbb{E}_X[X] = 0$, then Jensen with $\theta = \lambda^2 K_3^2 \in [0, 1]$,

$$\mathbb{E}_X[e^{\lambda X}] \leq \mathbb{E}_X[e^{\lambda^2 X^2}] = \mathbb{E}_X[(e^{X^2/K_3^2})^\theta] \leq 2^\theta \leq e^{\lambda^2 K_3^2} \quad (|\lambda| \leq 1/K_3).$$

For $|\lambda| > 1/K_3$, $(*)$ gives $\psi_X(\lambda) \leq \log 2 + \frac{\lambda^2 K_3^2}{4} \leq \lambda^2 K_3^2$. Either way $\psi_X(\lambda) \leq \frac{(\sqrt{2} K_3)^2 \lambda^2}{2}$, so $K_4 \leq \sqrt{2} K_3$.

Equivalent definitions, proof: $4 \Rightarrow 1$

$4 \Rightarrow 1$. Chernoff, exactly as on the sub-gaussian slide: minimize $e^{-\lambda t + K_4^2 \lambda^2/2}$ at $\lambda = t/K_4^2$, then repeat for $-X$,

$$\mathbb{P}_X(|X| \geq t) \leq 2e^{-t^2/(2K_4^2)}, \quad K_1 \leq \sqrt{2}K_4.$$

Finally $K_1 \lesssim K_4 \lesssim K_3 \lesssim K_2 \lesssim K_1$. $\blacksquare$

Hoeffding's lemma

Lemma. If X is centered and $X \in [a, b]$ almost surely, then $X \in \mathcal{G}((b-a)^2/4)$, that is $\psi_X(\lambda) \leq \lambda^2(b-a)^2/8$ for all λ.

Proof.

$$\begin{aligned}
\mathrm{Var}_{Y\sim\mathbb{Q}}(Y) \leq \mathbb{E}_{Y\sim\mathbb{Q}}[(Y - \tfrac{a+b}{2})^2] \leq \tfrac{(b-a)^2}{4} &\quad \text{any } \mathbb{Q} \text{ supported in } [a,b] \\
\psi''_X(\lambda) = \mathrm{Var}_{X\sim\mathbb{P}_\lambda}(X) \leq \tfrac{(b-a)^2}{4} &\quad \mathbb{P}_\lambda \text{ is supported in } [a,b] \text{ too} \\
\psi_X(\lambda) = \tfrac{\lambda^2}{2}\psi''_X(\theta) \leq \tfrac{\lambda^2(b-a)^2}{8} &\quad \text{Taylor, } \psi_X(0) = \psi'_X(0) = 0
\end{aligned}$$

Hoeffding's inequality

Theorem. Let $X_1, \ldots, X_n$ be independent with $X_i \in [a_i, b_i]$ almost surely, and $S_n = \sum_{i \le n}(X_i - \mathbb{E}_{X_i}[X_i])$. Then for all $t > 0$

$$\mathbb{P}_S(S_n \ge t) \le \exp\left(-\frac{2t^2}{\sum_{i \le n}(b_i - a_i)^2}\right).$$

Proof. By tensorization and Hoeffding's lemma, $\psi_{S_n}(\lambda) = \sum_i \psi_{X_i - \mathbb{E}[X_i]}(\lambda) \le \frac{\lambda^2}{8}\sum_i (b_i - a_i)^2$, so $S_n \in \mathcal{G}(\sigma^2)$ with $\sigma^2 = \frac{1}{4}\sum_i (b_i - a_i)^2$, and the sub-gaussian tail bound gives $\exp(-t^2/(2\sigma^2))$, which is the stated bound.

Special case. For $X_i \in [0, 1]$ i.i.d. and the empirical mean $\bar{X}_S = \frac{1}{n}\sum_i X_i$, we get $\mathbb{P}_S(|\bar{X}_S - \mathbb{E}_X[X]| \ge t) \le 2e^{-2nt^2}$, hence a deviation of order $\sqrt{\log(2/\delta)/(2n)}$ at confidence $1 - \delta$.

Operations on sub-gaussian variables

Recall $\|X\|_{\psi_2} = \inf\{K > 0 : \mathbb{E}_X[e^{X^2/K^2}] \leq 2\}$, the Orlicz norm.

Scaling. $X \in \mathcal{G}(\sigma^2)$ implies $cX \in \mathcal{G}(c^2\sigma^2)$, since $\psi_{cX}(\lambda) = \psi_X(c\lambda)$.

Independent sums. If $X_i \in \mathcal{G}(\sigma_i^2)$ are independent then $\sum_i X_i \in \mathcal{G}(\sum_i \sigma_i^2)$, since tensorization gives $\psi_{\sum_i X_i}(\lambda) = \sum_i \psi_{X_i}(\lambda) \leq (\sum_i \sigma_i^2)\lambda^2/2$. Variance proxies add exactly as variances do.

Arbitrary sums. Without independence, $\|X + Y\|_{\psi_2} \leq \|X\|_{\psi_2} + \|Y\|_{\psi_2}$, so the sub-gaussian variables form a normed vector space. The price of dropping independence is that proxies add before squaring, not after.

Centering. $\|X - \mathbb{E}_X[X]\|_{\psi_2} \leq C\|X\|_{\psi_2}$ for an absolute constant C, so nothing is lost by assuming centering.

Gaussian case. $X \sim \mathcal{N}(0, \sigma^2)$ satisfies $\psi_X(\lambda) = \sigma^2\lambda^2/2$ with equality, so the class is nonempty and the definition is sharp.

The two Orlicz-norm inequalities, **arbitrary sums** and **centering**, are proved on the next slide.

Operations on sub-gaussian variables, proof

Both facts use only the convexity of $z \mapsto e^{z^2}$, applied to the norm $\|\cdot\|_{\psi_2}$ recalled on the previous slide.

Arbitrary sums. Let $a = \|X\|_{\psi_2}, b = \|Y\|_{\psi_2}$. Since $\frac{X+Y}{a+b} = \frac{a}{a+b} \cdot \frac{X}{a} + \frac{b}{a+b} \cdot \frac{Y}{b}$ is a convex combination,

$$\mathbb{E}_{X,Y}[e^{(X+Y)^2/(a+b)^2}] \leq \tfrac{a}{a+b}\mathbb{E}_X[e^{X^2/a^2}] + \tfrac{b}{a+b}\mathbb{E}_Y[e^{Y^2/b^2}] \leq 2,$$

so $\|X+Y\|_{\psi_2} \leq a+b$.

Centering. A constant c has $\|c\|_{\psi_2} = |c|/\sqrt{\ln 2}$, so by the triangle inequality

$$\|X - \mathbb{E}_X[X]\|_{\psi_2} \leq \|X\|_{\psi_2} + \frac{\mathbb{E}_X|X|}{\sqrt{\ln 2}} \lesssim \|X\|_{\psi_2},$$

the last step because $\mathbb{E}_X|X| \lesssim \|X\|_{\psi_2}$, which is item 2 of the equivalence at $p = 1$.

Maximum of sub-gaussian variables

Proposition. Let $X_1, \ldots, X_n$ each be in $\mathcal{G}(\sigma^2)$, with no independence assumed. Then $\mathbb{E}_S[\max_{i \le n} X_i] \le \sigma\sqrt{2\log n}$.

Proof. For $\lambda > 0$,

$$e^{\lambda \mathbb{E}_S[\max_i X_i]} \le \mathbb{E}_S[e^{\lambda \max_i X_i}] \le \sum_{i \le n} \mathbb{E}_{X_i}[e^{\lambda X_i}] \le n e^{\lambda^2 \sigma^2/2} \quad \text{Jensen, then max} \le \text{sum}$$

$$\mathbb{E}_S[\max_i X_i] \le \frac{\log n}{\lambda} + \frac{\lambda \sigma^2}{2} = \sigma\sqrt{2\log n} \quad \text{at } \lambda = \sqrt{2\log n}/\sigma$$

Remark. The $\sqrt{\log n}$ is the correct order, attained by n independent gaussians. This is the seed of chaining and of every uniform bound over a finite class.

Beyond sums: bounded differences

Definition. $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the bounded differences property with constants $c_1, \ldots, c_n$ if for every i, every $x \in \mathcal{X}^n$ and every $x_i' \in \mathcal{X}$,

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \le c_i.$$

Changing one coordinate moves f by at most c_i, whatever the other coordinates are.

Why this generalizes Hoeffding. For $f(x) = \sum_i x_i$ with $x_i \in [a_i, b_i]$ the property holds with $c_i = b_i - a_i$. But it also covers functions that are not sums at all: a supremum over a class, the largest eigenvalue of a random matrix, the length of a longest common subsequence.

Closure under averaging. If f has bounded differences (c_i), then for any k so does

$$(x_1, \ldots, x_k) \longmapsto \mathbb{E}[f(x_1, \ldots, x_k, X_{k+1}, \ldots, X_n)],$$

with the same constants, since $|f(\ldots, x_j, \ldots) - f(\ldots, x_j', \ldots)| \le c_j$ holds pointwise and survives the expectation. This is the only extra fact the proof of McDiarmid needs.

McDiarmid's inequality

Theorem. If $X_1, \ldots, X_n$ are independent and f satisfies bounded differences with constants c_i, then $f(S) - \mathbb{E}_S[f(S)] \in \mathcal{G}(\frac{1}{4}\sum_i c_i^2)$, hence

$$\mathbb{P}_S(f(S) - \mathbb{E}_S[f(S)] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i \leq n} c_i^2}\right).$$

Proof, setup. Put $g(S) = f(S) - \mathbb{E}_S[f(S)]$ and $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$, with $\mathcal{F}_0$ trivial. The Doob martingale $D_i = \mathbb{E}_S[f(S) \mid \mathcal{F}_i]$ runs from $D_0 = \mathbb{E}_S[f(S)]$ to $D_n = f(S)$, so its increments $V_i = D_i - D_{i-1}$ satisfy

$$g(S) = \sum_{i \leq n} V_i, \quad \mathbb{E}_S[V_i \mid \mathcal{F}_{i-1}] = 0.$$

McDiarmid's inequality, proof

1. Each V_i has conditional range $\le c_i$. By independence, integrating $X_{i+1}, \dots, X_n$ against their laws,

$$V_i = h_i(X_{1:i}) - \mathbb{E}_{X_i'}[h_i(X_{1:i-1}, X_i')], \quad h_i(x_{1:i}) := \mathbb{E}[f(x_{1:i}, X_{i+1:n})].$$

By the closure property, $x \mapsto h_i(X_{1:i-1}, x)$ still has oscillation $\le c_i$, and the second term is $\mathcal{F}_{i-1}$-measurable, so given $\mathcal{F}_{i-1}$ the variable V_i lies in an interval of length $\le c_i$.

2. Conditional Hoeffding. Given $\mathcal{F}_{i-1}$, V_i is centered in an interval of length $\le c_i$, so Hoeffding's lemma applied to its conditional law give $\mathbb{E}_S[e^{\lambda V_i} \mid \mathcal{F}_{i-1}] \le e^{\lambda^2 c_i^2/8}, \quad \lambda \in \mathbb{R}$.

3. Peel off one factor at a time.

$$\begin{aligned}
\mathbb{E}_S[e^{\lambda \sum_{i\le n} V_i}] &= \mathbb{E}_S[e^{\lambda \sum_{i\le n-1} V_i}\mathbb{E}_S[e^{\lambda V_n} \mid \mathcal{F}_{n-1}]] && e^{\lambda \sum_{i\le n-1} V_i} \text{ is } \mathcal{F}_{n-1}\text{-measurable} \\
&\le e^{\lambda^2 c_n^2/8}\mathbb{E}_S[e^{\lambda \sum_{i\le n-1} V_i}] && \text{step 2} \\
&\le \cdots \le \prod_{i\le n} e^{\lambda^2 c_i^2/8} && \text{repeat for } i = n-1, \dots, 1
\end{aligned}$$

So $\psi_{g(S)}(\lambda) \le \frac{\lambda^2}{8}\sum_i c_i^2$, i.e. $g(S) \in \mathcal{G}(\frac{1}{4}\sum_i c_i^2)$; the sub-gaussian tail bound gives $\exp(-2t^2/\sum_i c_i^2)$. ■

Application: uniform deviation

Let $\mathcal{F}$ be a class of functions with values in $[0, 1]$, $\widehat{R}_S(f) = \frac{1}{n}\sum_{i\le n} f(X_i)$ and $R(f) = \mathbb{E}_{X\sim\mathbb{P}}[f(X)]$. Put

$$Z(S) = \sup_{f\in\mathcal{F}} |\widehat{R}_S(f) - R(f)|.$$

Changing one sample point moves each $\widehat{R}_S(f)$ by at most $1/n$, hence moves the supremum by at most $1/n$, so Z has bounded differences with $c_i = 1/n$ and McDiarmid gives

$$\mathbb{P}_S(Z(S) \ge \mathbb{E}_S[Z(S)] + t) \le e^{-2nt^2}.$$

Remark. Concentration has reduced a uniform statement to computing one number, $\mathbb{E}_S[Z(S)]$, which is where Rademacher complexity or chaining then enters. This division of labour is the standard architecture of a generalization bound.