Concentration Inequalities

(Where the Wild Things Are)

ALT 2022 Mentoring Workshop
Clément Canonne

Picture by Maurice Sendak

What's a concentration inequality?

Say you have a random variable X, with expectation $\mathbb{E}[X]$.

Can we say that $X \approx \mathbb{E}[X]$ "most of the time"?

What's a concentration inequality?

Say you have a random variable X, with expectation $\mathbb{E}[X]$.

Can we say that $X \approx \mathbb{E}[X]$ "most of the time"?

(Unfortunately, no.)

What's a concentration inequality?

Say you have a random variable X, with expectation $\mathbb{E}[X]$.

Can we say that $X \approx \mathbb{E}[X]$ "most of the time"?

(Unfortunately, no.)

Still, *under some conditions*, yes.

What's a concentration inequality?

Say you have a random variable X, with expectation $\mathbb{E}[X]$.

Can we say that $X \approx \mathbb{E}[X]$ "most of the time"?

(Unfortunately, no.)

Still, *under some conditions*, yes.

Interlude: why do we care about that? 🤨

What's a concentration inequality?

First example: Markov's inequality.

If $X \geq 0$, then $\Pr[\, X \geq a \,] \leq \mathbb{E}[X]/a$.

What's a concentration inequality?

First example: Markov's inequality.

If $X \geq 0$, then $\Pr[\, X \geq a \,] \leq \mathbb{E}[X]/a$.

Proof. $\mathbb{E}[X] \geq \mathbb{E}[X \, I[\, X \geq a] \,] \geq \mathbb{E}[\, a \, I[\, X \geq a] \,] = a \Pr[X \geq a]$.

What's a concentration inequality?

First example: Markov's inequality.

If $X \geq 0$, then $\Pr[\, X \geq a \,] \leq \mathbb{E}[X]/a$.

Proof. $\mathbb{E}[X] \geq \mathbb{E}[X\, I[\, X \geq a]\,] \geq \mathbb{E}[\, a\, I[\, X \geq a]\,] = a \Pr[X \geq a]$.

(Check: where did we use $X \geq 0$?)

What's a concentration inequality?

Second example: Chebyshev's inequality.

$\Pr[\ |X - \mathbb{E}[X]| \geq a\ ] \leq \mathrm{Var}[X]/a^2.$

What's a concentration inequality?

Second example: Chebyshev's inequality.

$\Pr[\ |X - \mathbb{E}[X]| \geq a\ ] \leq \mathrm{Var}[X]/a^2.$

Proof. Apply Markov's inequality to $X' := (X - \mathbb{E}[X])^2$.

What's a concentration inequality?

Second example: Chebyshev's inequality.

$\Pr[\ |X - \mathbb{E}[X]| \geq a\ ] \leq \mathrm{Var}[X]/a^2$.

Proof. Apply Markov's inequality to $X' := (X - \mathbb{E}[X])^2$.

(Note: half or more of the results in my area rely on Chebyshev)

What's a concentration inequality?

Third example: Chernoff's inequality.

$\Pr[\ X \geq a\ ] \leq \inf_{t>0} \mathbb{E}[e^{tX}]/e^{ta}$.

What's a concentration inequality?

Third example: Chernoff's inequality.

$\Pr[\ X \geq a\ ] \leq \inf_{t>0} \mathbb{E}[e^{tX}]/e^{ta}$.

Proof. Apply Markov's inequality to $X' := e^{tX}$.

What's a concentration inequality?

Third example: Chernoff's inequality.

$\Pr[\ X \geq a\ ] \leq \inf_{t>0} \mathbb{E}[e^{tX}]/e^{ta}$.

Proof. Apply Markov's inequality to $X' := e^{tX}$.

(Note: this may not be the specific bounds you've seen, but that's usually how they are proven, after picking the right t and massaging the result)

Oh great, so everything's Markov? 🤷

Oh great, so everything's Markov?

No. But it's worth knowing those tricks.

Oh great, so everything's Markov?

No. But it's worth knowing those tricks. (Derive those things at least once, it builds "intuition." Then don't.)

Oh great, so everything's Markov?

No. But it's worth knowing those tricks. (Derive those things at least once, it builds "intuition." Then don't.)

Read books about them, especially those discussing "why" we get those bounds, when we can expect them, and what are the general ideas.

Oh great, so everything's Markov?

No. But it's worth knowing those tricks. (Derive those things at least once, it builds "intuition." Then don't.)

Read books about them, especially those discussing "why" we get those bounds, when we can expect them, and what are the general ideas.

E.g., *"subgaussian random variables have gaussian-like tail bounds." "The square of a subgaussian is subexponential."*

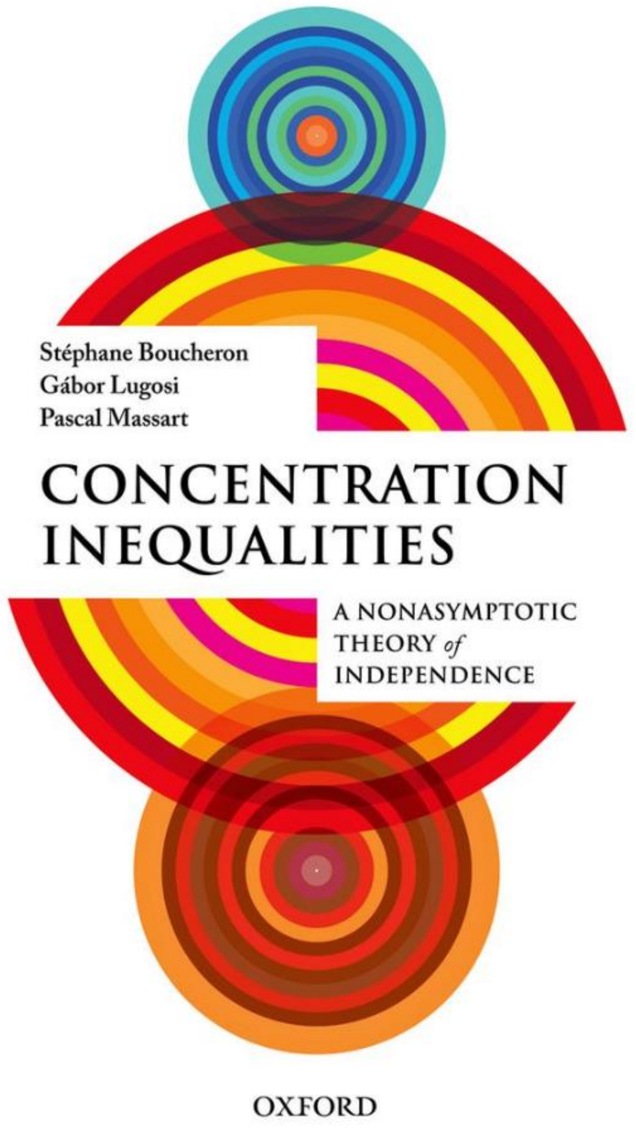
Oh great, so everything's Markov?

Markov, Chebyshev, Chernoff, Hoeffding, Bennett, Bernstein, McDiarmid, Azuma, McNugget: make sure you have seen those names (and others, maybe), *vaguely* know what they are about, but honestly, no point in learning the exact statements. You have Google and books.

(Also, again, *most* of the time you'll probably only need one or two of those.)

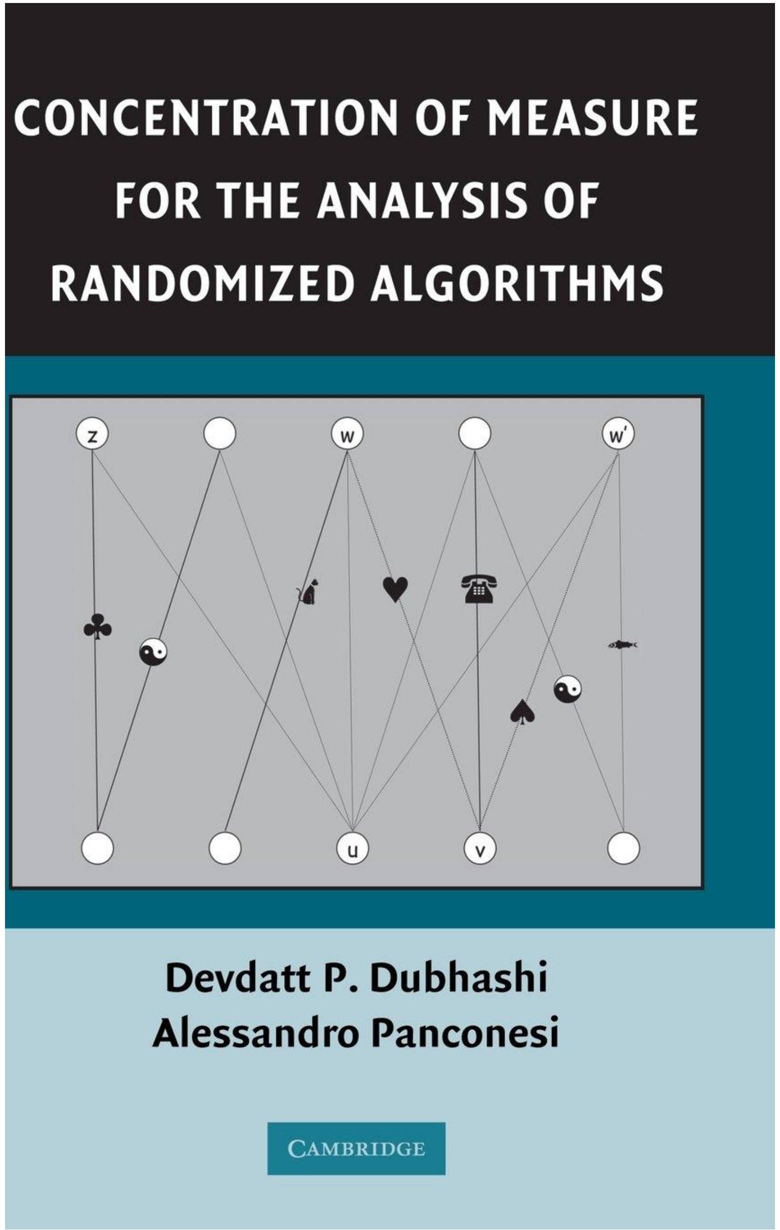
Where the Wild Things Are

Boucheron, Stéphane; **Lugosi, Gábor**; **Massart, Pascal**. Concentration inequalities. A nonasymptotic theory of independence. *Oxford University Press, Oxford,* 2013.



Where More Wild Things Are

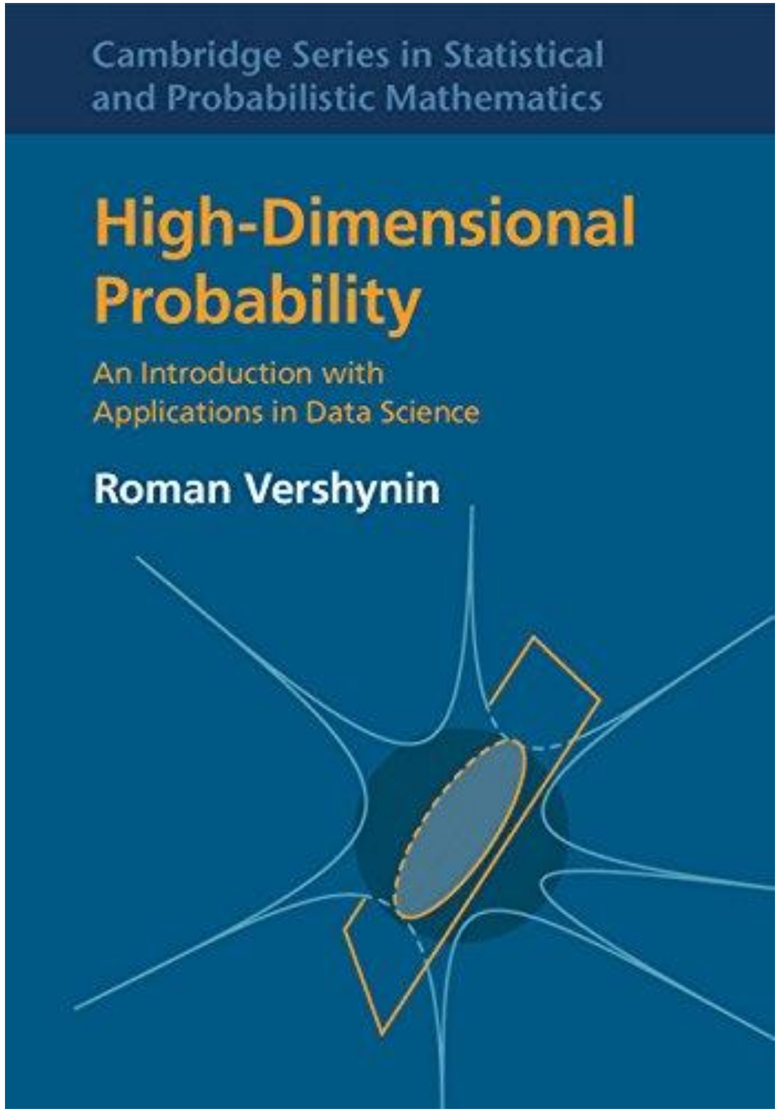
Dubhashi, Devdatt P.; **Panconesi, Alessandro**. Concentration of measure for the analysis of randomized algorithms. *Cambridge University Press, Cambridge,* 2009.



Oh, look, More Wild Things

Vershynin, Roman. High-dimensional probability. An introduction with applications in data science. *Cambridge University Press, Cambridge,* 2018.

https://www.math.uci.edu/~rvershyn/papers/HDP-book/ (free copy)



Wild Things on YouTube

"Concentration inequalities" by Aditya Gopalan and Himanshu Tyagi

https://youtube.com/playlist?list=PLgMDNELGJ1CZp3yTR9r5utkisB8i92_dZ

Prof. Aditya Gopalan

Prof. Himanshu Tyagi

If You Want More Wild Things

Anything by David Pollard (also, sometimes, fun to read).

http://www.stat.yale.edu/~pollard/Books/Mini/Basic.pdf

Chapter 2

A few good inequalities

SECTION 2.1 *introduces exponential tail bounds.*
SECTION 2.2 *introduces the method for bounding tail probabilities using moment generating functions.*
SECTION 2.3 *describes analogs of the usual $\mathcal{L}^p$ norms, which have proved useful in empirical process theory.*
SECTION 2.4 *discusses subgaussianity, a most useful extension of the gaussianity property.*
SECTION 2.5 *discusses the Bennett inequality for sums of independent random variables that are bounded above by a constant.*
SECTION 2.6 *shows how the Bennett inequality implies one form of the Bernstein inequality, then discusses extensions to unbounded summands.*
SECTION 2.7 *describes two ways to capture the idea of tails that decrease like those of the exponential distribution.*
SECTION 2.8 *illustrates the ideas from the previous Section by deriving a subgaussian/subexponential tail bound for quadratic forms in independent subgaussian random variables.*

Some nuggets

- Markov is only for non-negative stuff! 👀
- Chebyshev usually good enough for constant-probability statements

  Things usually are around a few standard deviations

  Only requires pairwise independence 🎲
- Hoeffding/Chernoff what you need most of the rest

  Fancy stuff is fun, but "basic" often works
- Bernstein, Bennett: both subgaussian (near the mean) and subexponential (far tails) parts: "Poisson-like" bounds 🐟
- Sanity checks: e.g., Gaussians first!

Some wilder nuggets*

* Beginning to get my metaphors mixed up here.

Negative association

What if I have a sum of non-independent things? 😱

Negative association

What if I have a sum of non-independent things?

> A property of negatively associated random variables that is very useful in applications to the analysis of algorithms is that one can apply the Chernoff–Hoeffding(CH) bounds to give tail estimates on their sum; in effect, for purposes of stochastic bounds on the sum, one can treat the variables as if they were independent.

Dubhashi, Devdatt; **Ranjan, Desh**. Balls and bins: a study in negative dependence. *Random Structures Algorithms* **13** (1998), no. 2, 99--124.

https://doi.org/10.7146/brics.v3i25.20006

Decoupling

Say you are considering quadratic forms (or worse) in $X_1, \ldots, X_n$:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} X_i X_j$$

Terms: not independent, not negatively associated... Ouch.

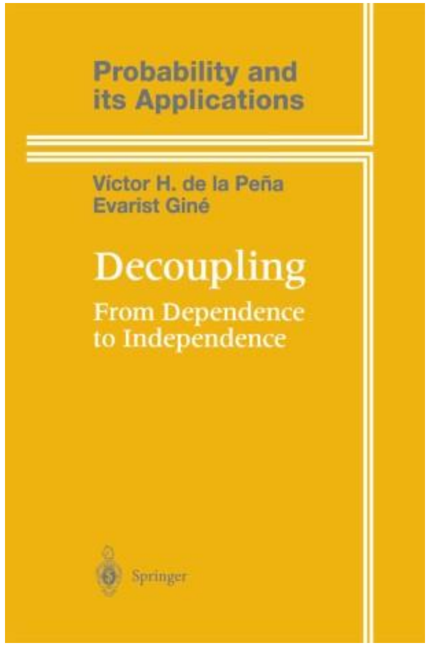
Decoupling

Can we just replace this with

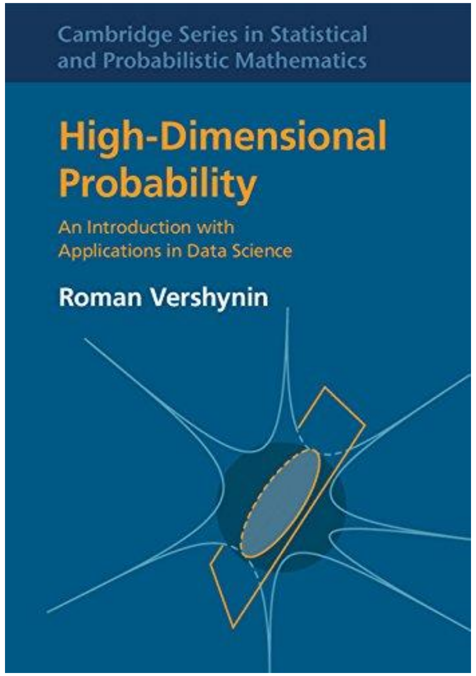
$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} X_i Y_j$$

Where the X_i's and Y_js **are** independent?

Decoupling



Basically, yes.



Theorem 6.1.1 (Decoupling). *Let A be an $n \times n$, diagonal-free matrix (i.e. the diagonal entries of A equal zero). Let $X = (X_1, \ldots, X_n)$ be a random vector with independent mean zero coordinates X_i. Then, for every convex function $F : \mathbb{R} \to \mathbb{R}$, one has*

$$\mathbb{E}\, F(X^\mathsf{T} A X) \le \mathbb{E}\, F(4 X^\mathsf{T} A X') \tag{6.3}$$

where X' is an independent copy of X.

Anticoncentration!

What is I wanted to argue things are not too tightly concentrated around the expectation?

Anticoncentration!

What is I wanted to argue things are not too tightly concentrated around the expectation?

E.g., *"things will be* $\Omega(1)$ *standard dev away with proba* $\Omega(1)$*."*

Anticoncentration!

What is I wanted to argue things are not too tightly concentrated around the expectation?

E.g., *"things will be* $\Omega(1)$ *standard dev away with proba* $\Omega(1)$*."* True for Gaussians ✅

Anticoncentration!

What is I wanted to argue things are not too tightly concentrated around the expectation?

One-trick-wild-poney: Paley–Zygmund. Simple, but involves higher moments.

Anticoncentration!

One-trick-wild-poney: Paley–Zygmund. 🐎

https://en.wikipedia.org/wiki/Paley%E2%80%93Zygmund_inequality

Example: if X has mean 0,

$$\Pr\left[|X| > \sqrt{\frac{1}{2}\mathrm{Var}[X]}\right] \geq \frac{1}{4} \cdot \frac{\mathbb{E}[X^2]^2}{\mathbb{E}[X^4]}$$

That's all (for me)