# Courses

Daniele Zago

26 maggio 2022

# Indice

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**Classical asymptotic and nonasymptotic results**

Lecturer : Aaditya Ramdas

# 1 Classical limit theorems

In the following, we will always denote $S_n := X_1 + \cdots + X_n$.

**Theorem 1 (Strong Law of Large Numbers (SLLN))** *Let $X_i$ be iid with mean $\mu$. Then*

$$\frac{S_n}{n} \to \mu, \text{ almost surely.}$$

What's so strong about the strong law? The weak law states that $S_n/n \to \mu$ in probability, which is a weaker statement. So can we just forget about the weak law, and only study the strong law?! No, because the weak law actually holds under weaker assumptions. Despite what Wikipedia, Wolfram and other websites currently say, here is a more complete description of the WLLN.

**Theorem 2 (Weak Law of Large Numbers (WLLN) from Feller 1971, page 565)** *Let $X_i$ be iid with characteristic function $\phi$ and CDF $F$. Then the following three conditions are equivalent:*

1. *$\phi$ is differentiable at 0, and $\phi'(0) = i\mu$.*
2. *As $t \to \infty$, we have $t[1 - F(t) + F(-t)] \to 0$ and $\int_{-t}^{t} xF(dx) \to \mu$.*
3. *$S_n/n \to \mu$, in probability.*

**Example 3 (from Charles Geyer's lecture notes at UMN)** *Define a random variable via its CDF:*

$$F(t) = \begin{cases} 1 - \frac{\log 2}{t \log t}, t \geq 2 \\ 1/2, -2 \leq t \leq 2 \\ \frac{\log 2}{|t| \log |t|}, t \leq -2 \end{cases}$$

*Then one can show that its mean does not exist, and hence by Theorem 4(c) in Ferguson (A Course in Large Sample Theory, 1996), the SLLN does not hold. However, the above random variable is symmetric by construction, and condition (ii) above can be verified to hold with $\mu = 0$. Hence, the third condition (WLLN) holds.*

However these theorems do not provide a rate of convergence of sample averages to $\mu$, so this is not sufficient. However, the CLT provides an *asymptotic* rate of convergence, and hence an asymptotic confidence interval.

**Theorem 4 (Central Limit Theorem (CLT))** *Let $X_i$ be iid with mean $\mu$ and variance $\sigma^2$. Then, we have*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1), \text{ in distribution.}$$

Hence, if one divides $S_n$ by $n$ it is damped to zero, if one divides by $\sqrt{n}$, it can still be unbounded, and it is interesting to ask what function of $n$ one needs to divide by in order to get a nontrivial bound. It turns out that dividing $S_n$ by $n^{1/2+\epsilon}$, for any $\epsilon > 0$, results in a limit of zero. Hence, the "right" quantity has to be larger than $n^{1/2}$ but smaller than $n^{1/2+\epsilon}$ for any constant $\epsilon > 0$. However, we have:

**Theorem 5 (Law of the iterated logarithm (LIL))** *Let $X_i$ be a symmetric Rademacher ($\pm 1$ with equal probability). Then*

$$\limsup_{n\to\infty} \frac{|S_n|}{\sqrt{n \log\log n}} = \sqrt{2}, \text{ a.s.}$$

# 2 Nonasymptotic bounds

In order to get something non-asymptotic from the CLT, one needs strictly more than two moments, as exemplified by the following Berry-Esseen bound.

**Theorem 6 (Berry-Esseen theorem)** *Let $X_i$ be iid with mean $\mu$, variance $\sigma^2$, and $\mathbb{E}|X_i|^3 < \infty$. Then, we have*

$$\sup_t \left| P(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq t) - \Phi(t) \right| \leq \frac{C\mathbb{E}|X_i|^3}{\sigma^{3/2}\sqrt{n}}.$$

*where $0.41 \leq C \leq 0.4748$ is a universal constant.*

If we assume more, such as the random variable being bounded, or having a bounded MGF, we can prove many "tail" inequalities, such as Hoeffding's, Bennett's and Bernstein's inequalities. As an example, consider Hoeffding's inequality for bounded random variables.

**Theorem 7 (Hoeffding's inequality)** *Let $X_i$ be iid, mean 0, bounded in $[-a_i, a_i]$. Denoting $A^2 := \sum_{i=1}^n a_i^2/n$, then*

$$P(\frac{S_n}{n} > \epsilon) \leq \exp(-2n\epsilon^2/A^2).$$

There are also matrix extensions of these inequalities due to Ahlswede-Winter, Vershynin, Tropp and others.

# 3 The rest of this mini

The rest of this mini will be devoted to understanding one assumption, and one theorem. The assumption is a "canonical supermartingale assumption", and it is weaker than many standard nonparametric assumptions in the literature. The theorem is informally called the "mother of all exponential concentration inequalities", and it is stronger than many standard famous named theorems in the literaure.

The word "stronger" will become clearer in the rest of the course. For this, we introduce the A-B-C-D-E mnemonics. A: weaker assumptions, B: lower boundary, C: continuous time, D: higher dimensions, E: larger exponent. The meaning of those terms will become clearer later in the course.

Remarkably, we can even improve the aforementioned popular Hoeffding's inequality: it will hold under weaker dependence assumptions, have a lower boundary, have a continuous-time extension, extend to hold for matrices, and also have a tighter exponent (specifically holding when $A = \sum_{i=1}^n a_i^2/2n + \sum_{i=1}^n \mathbb{E}X_i^2/2n$).

We will encounter "self-normalized" ineqaulities for heavy-tailed distributions, concentration for matrices, and if there is time, concentration of continuous time processes, and martingales in smooth Banach spaces.

This course will require the use of (super)martingales, filtrations, convex analysis, and linear algebra. We will revise some of this background next, but most of it is assumed to be known (prerequisites).

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**Filtrations, stopping times, conditional expectations**

Lecturer : Aaditya Ramdas

# 1 Random walks

Let $X_1, X_2, \ldots$ be i.i.d. taking values in $\mathbb{R}$. Then $S_n = X_1 + \ldots + X_n$ is called a random walk. When $P(X_i = 1) = P(X_i = -1) = 1/2$, it's called a simple random walk.

**Theorem 1 (Durrett, Thm 4.1.2)** *For a random walk $(S_n)$ on $\mathbb{R}$, there are only four possibilities, one of which has probability one.*

1. $S_n = 0$ *for all* $n$.
2. $S_n \to \infty$.
3. $S_n \to -\infty$.
4. $-\infty = \liminf S_n < \limsup S_n = +\infty$.

Any nondegenerate symmetric random walk (meaning that $P(X_i = 0) < 1$), such as the simple random walk, will satisfy case (iv).

# 2 Filtrations and stopping times

The sigma-field $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ is the information known at time $n$, and the sequence of sigma-fields $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \ldots$ forms a "filtration" $(\mathcal{F}_n)$. A random variable $\tau$ taking values in $\{1, 2, \ldots\} \cup \{\infty\}$ is called a stopping time if for all $n \in \mathcal{N}$, we have $\{\tau = n\} \in \mathcal{F}_n$, meaning that we can decide whether to stop the process at time $n$ based only on the information known at time $n$. All constant times are stopping times, and if $S, T$ are stopping times, then $S \vee T$ and $S \wedge T$ are also stopping times. $\mathcal{F}_\tau = \sigma(X_1, \ldots, X_\tau)$ is the amount of information known at the stopping time $\tau$. A simple example is

$$\tau = \inf\{k : S_k \geq x\} \text{ for some fixed } x.$$

If $M, N$ are stopping times with $M \leq N$, then $\mathcal{F}_M \subseteq \mathcal{F}_N$, and if $Y_n \in \mathcal{F}_n$, then $Y_N \in \mathcal{F}_N$.

**Theorem 2 (Durrett, Thm 4.1.3)** *Let $X_1, X_2, \ldots$ be iid, and $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Let $N$ be a stopping time with $P(N < \infty) > 0$. Conditional on $\{N < \infty\}$, the random variables $\{X_{N+n}, n \geq 1\}$ is independent of $\mathcal{F}_N$ and has the same distribution as the original sequence.*

Note that for any fixed $n$, a random walk with integrable $X_i$ would satisfy $\mathbb{E}S_n = n\mathbb{E}X_1$, and a random walk with zero-mean and square-integrable increments would satisfy $\mathbb{E}S_n = n\mathbb{E}X_1^2$. Wald's identities extend these properties to stopping times with finite expectation.

**Theorem 3 (Wald's identities, Durrett Thm 4.1.5 and 4.1.6)** *Let $X_1, X_2, \ldots$ be iid and $N$ be a stopping time with $\mathbb{E}N < \infty$. If $\mathbb{E}|X_i| < \infty$, then $\mathbb{E}S_N = \mathbb{E}N\mathbb{E}X_1$. Further, if $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 < \infty$, then $\mathbb{E}S_N^2 = \mathbb{E}N\mathbb{E}X_1^2$.*

> Exercise 4.1.12. Let $X_1, X_2, \ldots$ be i.i.d. uniform on $(0, 1)$, let $S_n = X_1 + \cdots + X_n$ and let $T = \inf\{n : S_n > 1\}$. Show that $P(T > n) = 1/n!$, so $ET = e$ and $ES_T = e/2$.

# 3 Conditional Expectations

We start with a probability space $(\Omega, \mathcal{F}_0, P)$, a sigma-field $\mathcal{F} \subset \mathcal{F}_0$, and a random variable $X$ that is measurable with respect to the sigma-field $\mathcal{F}_0$, denoted $X \in \mathcal{F}_0$. If $X$ is integrable, meaning that $\mathbb{E}|X| = \int |X| dP < \infty$, recall that the conditional expectation of $X$ given $\mathcal{F}$, denoted $\mathbb{E}(X|\mathcal{F})$, is any $\mathcal{F}$-measurable random variable $Y$ such that for all $A \in \mathcal{F}$, we have $\int_A X dP = \int_A Y dP$. The conditional expectation is unique, in the sense that all "versions" that satisfy the above definition are equal almost surely. Quoting from Durrett, section 5.1:

> Intuitively, we think of $\mathcal{F}$ as describing the information we have at our disposal - for each event $A \in \mathcal{F}$, we know whether or not $A$ has occurred. $\mathbb{E}(X|\mathcal{F})$ is then our "best guess" of the value of X given the information we have.

If $X \in \mathcal{F}$ (perfect information), then $\mathbb{E}(X|\mathcal{F}) = X$, meaning that if $X$ is contained in the available information $\mathcal{F}$, then our best guess of $X$ is $X$ itself. If $\mathcal{F} = \varnothing$ (no information), then $\mathbb{E}(X|\mathcal{F}) = \mathbb{E}X$, and if $X$ is independent of $\mathcal{F}$ (useless information), then $\mathbb{E}(X|\mathcal{F}) = \mathbb{E}X$.

Conditional expectations are monotone and linear, meaning that if $X \leq Y$, then $\mathbb{E}(X|\mathcal{F}) \leq \mathbb{E}(Y|\mathcal{F})$, and also that $\mathbb{E}(aX + bY|\mathcal{F}) = a\mathbb{E}(X|\mathcal{F}) + b\mathbb{E}(Y|\mathcal{F})$. Further, if $\mathcal{F}_1 \subset \mathcal{F}_2$, then $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_1)|\mathcal{F}_2) = \mathbb{E}(X|\mathcal{F}_1)$, and also $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_2)|\mathcal{F}_1) = \mathbb{E}(X|\mathcal{F}_1)$.

If $X \in \mathcal{F}$ and $Y, XY$ are integrable, then $\mathbb{E}(XY|\mathcal{F}) = X\mathbb{E}(Y|\mathcal{F})$. Of course, a special case of this is that $\mathbb{E}(cX) = c\mathbb{E}X$ for any constant $c$. Define $\mathrm{Var}(X|\mathcal{F}) = \mathbb{E}(X^2|\mathcal{F}) - \mathbb{E}(X|\mathcal{F})^2$, so that $\mathrm{Var}(X) = \mathrm{Var}(\mathbb{E}(X|\mathcal{F})) + \mathbb{E}(\mathrm{Var}(X|F))$.

Geometric interpretation: Let $(\Omega, \mathcal{F}_0, P)$ be a probability space with $X \in \mathcal{F}_0$. If $\mathcal{F} \subset \mathcal{F}_0$ and $\mathbb{E}X^2 < \infty$, then $\mathbb{E}(X|\mathcal{F}) = \arg\min_{Y \in \mathcal{F}} \mathbb{E}(X - Y)^2$. In other words, $\mathbb{E}(X|\mathcal{F})$ is the projection of $X$ onto the closed subspace $\mathcal{L}_2(\mathcal{F}) = \{Y \in \mathcal{F} : \mathbb{E}Y^2 < \infty\}$ of the Hilbert space $L_2(\mathcal{F}_0)$.

# 4 Standard inequalities

Chebyshev's inequality:

$$P(|X| \geq a|\mathcal{F}) \leq \frac{\mathbb{E}(X^2|\mathcal{F})}{a^2}.$$

Jensen's inequality: if $\phi$ is convex, and $\mathbb{E}|X| < \infty, \mathbb{E}|\phi(X)| < \infty$, then

$$\phi(\mathbb{E}(X|\mathcal{F})) \leq \mathbb{E}(\phi(X)|\mathcal{F})$$

Cauchy-Schwarz inequality:

$$E(XY|\mathcal{G})^2 \leq E\left(X^2|\mathcal{G}\right) E\left(Y^2|\mathcal{G}\right)$$

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**Martingales, Ville and Doob**

Lecturer : Aaditya Ramdas

# 1 Martingales

Recall that a filtration $(\mathcal{F}_n)$ is a sequence of increasing sigma-fields.

**Definition 1** *A sequence $(S_n)$ is "adapted" to the filtration $(\mathcal{F}_n)$, if $S_n \in \mathcal{F}_n$ for all $n$. A sequence $(S_n)$ is "predictable" (with respect to the filtration $(\mathcal{F}_n)$) if $S_n \in \mathcal{F}_{n-1}$ for all $n$.*

If $(S_n)$ is integrable and adapted to $(\mathcal{F}_n)$, and it also satisfies $\mathbb{E}(S_{n+1}|\mathcal{F}_n) = S_n$, then $(S_n)$ is called a martingale (with respect to $(\mathcal{F}_n)$). If the equality above is replaced by $\leq$, it is called a supermartingale, and if replaced by $\geq$, it is a submartingale. If $n > m$, then we have $\mathbb{E}(S_n|\mathcal{F}_m) = \mathbb{E}(S_m)$ for martingales (and $\leq$ or $\geq$ for super-/sub-martingales).

If $(S_n)$ is a martingale wrt $(\mathcal{G}_n)$, and $\mathcal{F}_n = \sigma(S_1, \ldots, S_n)$, then we must have $\mathcal{G}_n \supset \mathcal{F}_n$ for all $n$ and that $(S_n)$ is a martingale wrt $(\mathcal{F}_n)$ as well. Further, $(\mathcal{F}_n)$ is the smallest filtration wrt which $(S_n)$ is adapted, and if the filtration is not mentioned, it is understood to be $\sigma(S_1, \ldots, S_n)$.

For example, the simple random walk $(S_n)$ is a martingale that is adapted to the "canonical" filtration $\sigma(S_1, \ldots, S_n) = \sigma(X_1, \ldots, X_n)$.

- If $(S_n)$ is a martingale, $\phi$ is a convex function, and $(\phi(S_n))$ is integrable, then it is a submartingale (by Jensen's inequality).
- If $(S_n)$ is a supermartingale and $N$ is a stopping time, then $(S_{N \wedge n})$ is a supermartingale.

Just as it is well known that a monotonically nondecreasing sequence of real numbers with upper bound a number $M$ converges to a limit which does not exceed $M$, we have the following stochastic analogue.

**Theorem 1 (Martingale convergence theorem, Durrett Thm 5.2.8)** *If $(S_n)$ is a martingale with $\sup_n \mathbb{E}S_n^+ < \infty$, then $S_n$ converges almost surely to an integrable limit $X$. As a corollary, if $(S_n)$ is a positive supermartingale, then $S_n$ converges to a limit $X$ with $\mathbb{E}X \leq \mathbb{E}S_0$.*

Any submartingale can also be decomposed into a martingale and a predictable increasing component.

**Theorem 2 (Doob's decomposition theorem, Durrett Thm 5.2.10)** *Any submartingale $(S_n)$ can be uniquely decomposed as $S_n = M_n + A_n$ where $(M_n)$ is a martingale, and $(A_n)$ is a predictable increasing sequence.*

There are variants of the above called Riesz's and Krickenberg's decompositions. Just as one can construct a filtration from a martingale, one can also do the reverse.

**Theorem 3 (Constructing a martingale from a filtration)** *Let $Z$ be integrable, $(\mathcal{F}_n)$ be a filtration, and define $M_n = \mathbb{E}(Z|\mathcal{F}_n)$. Then $(M_n)$ is a martingale (and moreover, it is a uniformly integrable martingale).*

This technique often allows us to use martingale methods even when there is no obvious martingale in plain sight, since one can construct it out of thin air.

# 2 Ville's and Doob's inequalities

The first of Doob's inequalities can be seen as a uniform generalization of Markov's inequality to submartingales.

**Theorem 4 (Doob's maximal inequality for submartingales, Durrett Thm 5.4.2)** *If $(S_n)$ is a submartingale, then for any $x > 0$, we have*

$$P(\max_{1\le n\le N} S_n^+ \ge x) \le \frac{\mathbb{E}(S_N^+)}{x}$$

If $S_n = \sum_{i=1}^n X_i$ is a random walk with $\mathbb{E}X_i = 0, \mathbb{E}X_i^2 = \sigma_i^2 < \infty$, then using the fact that $(S_n)$ is a martingale implies $(S_n^2)$ is a submartingale, we get Kolmogorov's maximal inequality, which can be interpreted as a uniform generalization of Chebyshev's inequality to martingales. Denoting $s_n^2 := \mathrm{Var}(S_n) = \sum_{i=1}^n \sigma_i^2$, we have

$$P(\max_{1\le n\le N} S_n \ge x) \le \frac{s_N^2}{x^2}.$$

For random walks like above, the process $(S_n^2 - s_n^2)$ is also a martingale (confirm for yourself). Using this fact, if we additionally had $|X_i| \le K$, then one may also prove that

$$P(\max_{1\le n\le N} S_n \le x) \le \frac{(x+K)^2}{s_N^2}.$$

Further, for *any* zero-mean, finite-variance martingale $(S_n)$, we have a uniform version of Upensky's inequality:

$$P(\max_{1\le n\le N} S_n \ge x) \le \frac{\mathrm{Var}(S_N)}{\mathrm{Var}(S_n) + x^2}$$

Ville's supermartingale maximal inequality is closely related to Doob's:

**Theorem 5 (Ville's maximal inequality for supermartingales)** *If $(S_n)$ is a nonnegative supermartingale, then for any $x > 0$, we have*

$$P(\sup_{n\in\mathbb{N}} S_n > x) \leq \frac{\mathbb{E}S_0}{x}.$$

# 3 Optional Stopping (also called Optional Sampling, different from *Optimal* Stopping)

**Theorem 6 (Supermartingale optional stopping, Durrett Thm 5.7.6)** *If $(S_n)$ is a nonnegative supermartingale, then for any stopping time $N \leq \infty$, we have*

$$\mathbb{E}S_N \leq \mathbb{E}S_0,$$

*recalling that $S_\infty = \lim_n S_n$ exists via the martingale convergence theorem.*

For martingales, equality does not hold above, because the above theorem permits unbounded stopping times. As an example, consider the simple random walk, and the stopping time $N = \inf n : S_n = 1$. Obviously $\mathbb{E}S_n = 0$ for any fixed $n$, but $\mathbb{E}S_N = 1$ by definition. The problem again is that $N$ is unbounded, and indeed $\mathbb{E}N = \infty$. Instead, we have the following:

**Theorem 7 (Doob's martingale optional sampling, Gut Corollary 7.1)** *If $(S_n)$ is a martingale, and $N$ is a bounded stopping time, i.e. $P(N \leq K) = 1$ for some constant $K$, then $\{S_N, S_K\}$ is a martingale, and specifically*

$$\mathbb{E}S_N = \mathbb{E}S_0 = \mathbb{E}S_K.$$

Bounded stopping times, in fact, characterize martingales, as claimed below.

**Theorem 8 (Gut Theorem 7.2)** *$(S_n)$ is a martingale if and only if $\mathbb{E}S_N =$ constant for every bounded stopping time $N$.*

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**Canonical supermartingale assumption**

Lecturer : Aaditya Ramdas

# 1 Canonical supermartingale assumption

Let $(S_t)_{t\in\mathcal{T}}$ and $(V_t)_{t\in\mathcal{T}}$ be two real-valued processes adapted to an underlying filtration $(\mathcal{F}_t)_{t\in\mathcal{T}\cup\{0\}}$, where either $\mathcal{T}=\mathbb{N}$ for discrete-time processes or $\mathcal{T}=(0,\infty)$ for continuous-time processes, and $V_t \geq 0$ a.s. for all $t\in\mathcal{T}$.

In continuous time, we assume $(\mathcal{F}_t)$ satisfies the "usual hypotheses", namely, that it is right-continuous and complete, and we assume $(S_t)$ and $(V_t)$ are càdlàg.

We think of $S_t$ as a summary statistic accumulating over time, while $V_t$ is an accumulated "variance" process which serves as a measure of *intrinsic time*, an appropriate quantity to control the deviations of $S_t$ from its expectation.

Broadly, the literature gives results for two situations: one in which the finite-dimensional distributions of $(S_t)$ are from a parametric family, and one in which they are not. When we say "parametric" and "nonparametric", we are referring to the structure of $(S_t)$. The simplest case is the scalar, parametric setting, when $S_t$ is a sum of i.i.d., real-valued, mean-zero random variables with known distribution $F$. We quantify the relationship between $S_t$ and $V_t$ by a real-valued function $\psi$ reminiscent of a cumulant generating function (CGF). In the i.i.d. scalar setting above, we take $V_t=t$ and let $\psi$ be the CGF of $F$. Our key assumption ensures that $S_t$ is unlikely to grow too quickly relative to intrinsic time $V_t$:

**Assumption 1** *Let $(S_t)_{t\in\mathcal{T}}$ and $(V_t)_{t\in\mathcal{T}}$ be two real-valued processes adapted to an underlying filtration $(\mathcal{F}_t)_{t\in\mathcal{T}}$ with $S_0=V_0=0$ and $V_t\geq 0$ a.s. for all $t$. Let $\psi$ be a real-valued function with domain $[0,\lambda_{\max})$. We assume, for each $\lambda\in[0,\lambda_{\max})$, there exists a supermartingale $(L_t(\lambda))_{t\in\mathcal{T}}$ with respect to $(\mathcal{F}_t)$ such that $\mathbb{E}L_0=\mathbb{E}L_0(\lambda)$ is constant for all $\lambda$, and such that $\exp\{\lambda S_t-\psi(\lambda)V_t\}\leq L_t(\lambda)$ a.s. for all $t\in\mathcal{T}$.*

In the scalar, parametric, i.i.d. setting, $\psi$ is the "cumulant generating function" (logarithm of the MGF) of the random variable, and $L_t(\lambda)$ just equals the martingale $\exp\{\lambda S_t-\psi(\lambda)t\}$ itself, so that the defining inequality of Assumption 1 is an equality.

In matrix cases, $S_t$ will often not be a (super)martingale itself; instead there will be an auxiliary process $(Y_t)$ which is a matrix-valued martingale, and $S_t$ will be a scalar function of $Y_t$, for example $S_t=\gamma_{\max}(Y_t)$ when $Y_t$ is Hermitian, where $\gamma_{\max}(\cdot)$ denotes the maximum eigenvalue map. In such matrix cases, the process $\exp\{\lambda S_t-\psi(\lambda)V_t\}$ may not be a supermartingale

itself, but is majorized by one; in the scalar setting, by contrast, $\exp\{\lambda S_t - \psi(\lambda)V_t\}$ will be a supermartingale itself.

We remark also that it is important in Assumption 1 that $(V_t)$ is allowed to be adapted and not just predictable.

Even in nonparametric cases, $\psi$ will often still be a CGF of some distribution, though this is not required. However, our most interesting results require that $\psi$ satisfy certain properties which are true of CGFs for zero-mean random variables:

**Definition 1** *A real-valued function $\psi$ with domain $[0, \lambda_{\max})$ is called* CGF-like *if it is strictly convex and twice continuously differentiable with $\psi(0) = \psi'(0_+) = 0$ and also $\sup_{\lambda\in[0,\lambda_{\max})} \psi(\lambda) = \infty$. For such a function we write $\bar{b} = \bar{b}(\psi) := \sup_{\lambda\in[0,\lambda_{\max})} \psi'(\lambda) \in (0, \infty]$.*

We remark that in many cases $\lambda_{\max} = \infty$ and $\bar{b} = \infty$, but we allow finite values to handle a condition that arises later.

# 2 Sufficient conditions for Assumption 1

With the exception of martingales in Banach spaces, all discrete-time settings use $S_t = \gamma_{\max}(Y_t)$, where $(Y_t)_{t\in\mathcal{T}}$ is a martingale taking values in $\mathcal{H}^d$, the space of Hermitian, $d \times d$ matrices. Typically, setting $d = 1$ recovers the corresponding known scalar result exactly. We note also that our results for Hermitian matrices will extend directly to rectangular matrices $\mathcal{C}^{d_1\times d_2}$ using "Hermitian dilations".

In discrete time, the following general condition on $(Y_t)$ is sufficient to show that Assumption 1 holds; here the relation $A \preceq B$ denotes the semidefinite order, and $\Delta Y_t := Y_t - Y_{t-1}$ for any discrete-time process $(Y_t)_{t\in\mathcal{N}}$. We also give a version for continuous-time scalar processes which trivially implies Assumption 1, but which helps us avoid stating results twice in what follows. Below and throughout the paper we use $\mathbb{E}_t$ and $\mathcal{P}_t$ to denote expectation and probability conditioned on $\mathcal{F}_t$, respectively.

**Definition 2** *Let $\psi$ be a real-valued function with domain $[0, \lambda_{\max})$. We separate the definition of a* sub-$\psi$ *process into two cases.*

*(a) When $\mathcal{T} = \mathbb{N}$, an adapted, discrete-time, $\mathcal{H}^d$-valued process $(Y_t)_{t\in\mathbb{N}}$ is* sub-$\psi$ *with adapted, $\mathcal{H}^d$-valued, nondecreasing (in the semidefinite order)* self-normalizing process *$(U_t)_{t\in\mathcal{N}}$ and predictable, $\mathcal{H}^d$-valued, nondecreasing* variance process *$(W_t)_{t\in\mathbb{N}}$ if, for all $t \in \mathbb{N}$ and $\lambda \in [0, \lambda_{\max})$, we have*

$$\mathbb{E}_{t-1} \exp\{\lambda\Delta Y_t - \psi(\lambda)\Delta U_t\} \preceq \exp\{\psi(\lambda)\Delta W_t\}. \tag{1}$$

*If we say that $(Y_t)$ is sub-$\psi$ with self-normalizing process $(U_t)$ and do not specify a variance process $(W_t)$, then $(W_t)$ is understood to be identically zero. The analogous statement holds when we do not specify the self-normalizing process $(U_t)$. The latter is always true by convention in the continuous-time case below.*

*(b) When $\mathcal{T} = (0, \infty)$, an adapted, càdlàg, real-valued process $(Y_t)_{t\in(0,\infty)}$ is* sub-$\psi$ *with predictably measurable, càdlàg, real-valued, nondecreasing* variance process $(W_t)_{t\in(0,\infty)}$ *if, for all $0 \le s \le t < \infty$ and $\lambda \in [0, \lambda_{\max})$, we have*

$$\mathbb{E}_s \exp\{\lambda(Y_t - Y_s) - \psi(\lambda) \cdot (W_t - W_s)\} \le 1.$$

For a familiar example, suppose $\mathcal{T} = \mathbb{N}$, $d = 1$ and $(Y_t)$ has independent increments. Let $W_t = t$, $U_t \equiv 0$ and $\psi(\lambda) = \lambda^2/2$. Then (1) reduces to the usual definition of a 1-sub-Gaussian random variable (Boucheron, Lugosi, Massart). For a self-normalized example, let $(\Delta Y_t)$ be i.i.d. from any distribution symmetric about zero. Then, again letting $\psi(\lambda) = \lambda^2/2$, then de la Pena showed that $(Y_t)$ is sub-$\psi$ with self-normalizing process $U_t = \sum_{i=1}^t \Delta Y_i^2$.

The definition of sub-$\psi$ generalizes the standard notion of being sub-Gaussian or sub-gamma to permit a general function $\psi$ (Boucheron, Lugosi, Massart). The Cramér-Chernoff method typically begins with such an assumption, in the form $\mathbb{E}_{t-1} e^{\lambda \xi_t} \le e^{\psi(\lambda)\sigma_t^2}$ for $\sigma_t^2 \in \mathcal{F}_{t-1}$. Using the semidefinite order allows us to extend our results to $\mathcal{H}^d$-valued processes, following the methods of Tropp, and Oliveira. Using the adapted process $(U_t)$ in addition to the predictable process $(W_t)$ enables extensions to a variety of self-normalized bounds by de la Pena and others, for example yielding bounds on the deviation of a martingale in terms of its quadratic variation. This is the reason we call $(U_t)$ a "self-normalizing process".

In discrete time, the link between Definition 2 and Assumption 1 is the following lemma.

**Lemma 2** *Let $\mathcal{T} = \mathbb{N}$. If $(Y_t)_{t\in\mathbb{N}}$ is sub-$\psi$ with self-normalizing process $(U_t)_{t\in\mathbb{N}}$ and variance process $(W_t)_{t\in\mathcal{N}}$, then Assumption 1 is satisfied for $S_t = \gamma_{\max}(Y_t)$, $V_t = \gamma_{\max}(U_t + W_t)$, and $\psi$, with $\mathbb{E}L_0 = d$.*

The value $\mathbb{E}L_0 = d$, the ambient dimension, leads to a pre-factor of $d$ in all of our operator-norm matrix bounds. In cases when $\sup_{t\in\mathcal{T}} \operatorname{rank}(U_t + W_t) \le r < d$ a.s., the pre-factor $d$ in our bounds may be replaced by $r$.

We present five sub-$\psi$ cases: the sub-gamma case corresponding to Bernstein's inequality, the sub-Gaussian case in Hoeffding's inequality, the sub-Poisson case from Bennett's inequality, and the sub-exponential and sub-Bernoulli cases which are used in several other existing bounds.

1. We say $(Y_t)$ is *sub-gamma* with scale parameter $c$ when condition (1) holds for some suitable $(U_t)$ and $(W_t)$ using

$$\psi_G(\lambda) := \frac{\lambda^2}{2(1 - c\lambda)} \quad \text{for } 0 \le \lambda < \frac{1}{c} = \lambda_{\max}.$$

2. We say $(Y_t)$ is *sub-Gaussian* when condition (1) holds for some suitable $(U_t)$ and $(W_t)$ using

$$\psi_N(\lambda) := \lambda^2/2,$$

that is, when it is sub-gamma with scale parameter $c = 0$ (taking $\lambda_{\max} = \infty$).

3. We say $(Y_t)$ is *sub-Poisson* with scale parameter $c$ when condition (1) holds for some suitable $(U_t)$ and $(W_t)$ using

$$\psi_P(\lambda) := \frac{e^{c\lambda} - c\lambda - 1}{c^2}.$$

4. We say $(Y_t)$ is *sub-exponential* with scale parameter $c$ when condition (1) holds for some suitable $(U_t)$ and $(W_t)$ using

$$\psi_E(\lambda) := \frac{-\log(1 - c\lambda) - c\lambda}{c^2}, \quad \text{for } 0 \le \lambda < \frac{1}{c} = \lambda_{\max}.$$

Note this definition departs from the usage of sub-exponential in the literature, but we adopt it here for internal consistency.

5. We say $(Y_t)$ is *sub-Bernoulli* with range parameters $g, h > 0$ when condition (1) holds for some suitable $(U_t)$ and $(W_t)$ using

$$\psi_B(\lambda) := \log \frac{ge^{h\lambda} + he^{-g\lambda}}{g + h},$$

which is the cumulant generating function of a mean-zero random variable taking values $-g$ and $h$.

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**The mother theorem**

Lecturer : Aaditya Ramdas

# 1 The mother theorem

To state our main theorem on general exponential line-crossing inequalities, we will make use of the following transforms of $\psi$:

$$\psi^\star(u) := \sup_{\lambda \in [0, \lambda_{\max})} [\lambda u - \psi(\lambda)] \quad \text{(the Legendre-Fenchel transform)},$$

$$D(u) := \sup \left\{ \lambda \in [0, \lambda_{\max}) : \frac{\psi(\lambda)}{\lambda} \leq u \right\} \quad \text{(the "decay" transform), and}$$

$$\mathfrak{s}(u) := \frac{\psi(\psi^{\star\prime}(u))}{\psi^{\star\prime}(u)} \quad \text{(the "slope" transform)}.$$

In the definition of $D(u)$, we take the supremum of the empty set to equal zero instead of the usual $-\infty$. This case can arise in general, but not when $\psi$ is CGF-like. Note that $D(u)$ can also be infinite. We call $D(u)$ the "decay" transform because it determines the rate of exponential decay of the upcrossing probability bound in Theorem 1(a) below. We call $\mathfrak{s}(u)$ the "slope" transform because it gives the slope of the linear boundary in Theorem 1(b); this is defined only when $\psi$ is CGF-like.

Our main theorem has four parts, each of which facilitates comparisons with a particular related literature, as we discuss later.

**Theorem 1** *If the canonical supermartingale assumption (Assumption 1, previous lecture) holds, then*

*(a) For any $a, b \geq 0$, we have*

$$\mathcal{P}\{\exists t \in \mathcal{T} : S_t \geq a + bV_t\} \leq (\mathbb{E}L_0) \exp(-aD(b)).$$

*Additionally, whenever $\psi$ is CGF-like, the following three statements are equivalent to statement (a).*

*(b) For any $m > 0$ and $x \in [0, m\bar{b})$, we have*

$$\mathcal{P}\left(\exists t \in \mathcal{T} : S_t \geq x + \mathfrak{s}\left(\frac{x}{m}\right) \cdot (V_t - m)\right) \leq (\mathbb{E}L_0) \exp\left\{-m\psi^\star\left(\frac{x}{m}\right)\right\}.$$

*Furthermore, if the slope $\mathfrak{s}(x/m)$ were replaced by any other value, the probability bound on the right-hand side would need to be larger.*

*(c) For any $m \geq 0$ and $x \in [0, \bar{b})$, we have*

$$\mathcal{P}\left(\exists t \in \mathcal{T} : \frac{S_t}{V_t} \geq \mathfrak{s}(x) + \frac{m(x - \mathfrak{s}(x))}{V_t}\right) \leq (\mathbb{E}L_0) \exp\left\{-m\psi^\star(x)\right\}.$$

*Furthermore, if $\mathfrak{s}(x)$ were replaced by any other value, the probability bound on the right-hand side would need to be larger.*

*(d) For any $m \geq 0$, $x \geq 0$ and $b$ finite in $\left[0, \bar{b} \wedge \frac{x}{m}\right]$ (taking $\frac{0}{0} = \infty$), we have*

$$\mathcal{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } S_t \geq x + b(V_t - m)\right) \leq \begin{cases} (\mathbb{E}L_0) \exp\left\{-(x - bm)D(b)\right\}, & m = 0 \text{ or } \mathfrak{s}\left(\frac{x}{m}\right) \geq b \\ (\mathbb{E}L_0) \exp\left\{-m\psi^\star\left(\frac{x}{m}\right)\right\}, & m > 0 \text{ and } \mathfrak{s}\left(\frac{x}{m}\right) \leq \end{cases} \tag{1}$$

We later provide a straightforward proof of the above results that uses only Ville's maximal inequality for nonnegative supermartingales Ville (1939) and elementary convex analysis. We give here several remarks on the theorem, followed by three illustrative examples.

- It is useful to think of the parts of 1 as statements about the process $(V_t, S_t)$ or $(V_t, S_t/V_t)$ in $\mathbb{R}^2$. Many of our results are easily understood via this geometric intuition. 1 illustrates the following points.
  - 1(a) takes a given line $a + bV_t$ and bounds its $S_t$-upcrossing probability.
  - 1(b) takes a point $(m, x)$ in the $(V_t, S_t)$-plane and, out of the infinitely many lines passing through it, chooses the one which yields the tightest upper bound on the corresponding $S_t$-upcrossing probability.
  - 1(c) is like part (b), but instead of looking at $S_t$, we look at $S_t/V_t$, fix a point $(m, x)$ in the $(V_t, S_t/V_t)$-plane, and choose from among the infinitely many curves $b + a/V_t$ passing through it to minimize the probability bound.
  - The intuition for 1(d) is as follows. If we want to bound the upcrossing probability of the line $(x - bm) + bV_t$ on $\{V_t \geq m\}$, we can clearly obtain a conservative bound from 1(a) with $a = x - bm$. This yields the first case in (1). However, we can also apply 1(b) with the values $m, x$, obtaining a bound on the upcrossing probability for a line which passes through the point $(m, x)$ in the $(V_t, S_t)$-plane, and this line yields the minimum possible probability bound among all lines passing through $(m, x)$. If the slope of this line, $\mathfrak{s}(x/m)$, is less than $b$, then this optimal probability bound is conservative for the upcrossing probability over the original line $x + b(V_t - m)$ on $\{V_t \geq m\}$. This gives the second case in (1), which is guaranteed to be at least as small as the bound in the first case when $\mathfrak{s}(x/m) \leq b$.
- The purpose of excluding $\psi$ being CGF-like from Assumption 1 is to separate the truth of statement (a), which follows solely from the assumption, from its equivalence to (b), (c), and (d), which follows from $\psi$ being CGF-like.

- The factor $\mathbb{E}L_0$ will typically equal one when we have scalar observations, while in matrix cases it generally equals $d$, the dimension of the matrix observations. As mentioned earlier, in many cases $\lambda_{\max} = \infty$ and $\bar{b} = \infty$, but we allow finite values to handle some cases discussed later.

- 1 yields a uniform extension of many fixed-time or finite-horizon exponential bounds, losing nothing in going from a fixed-time to a uniform bound. We briefly revisit this property later, where we observe that the Dubins-Savage inequality does not possess this property.

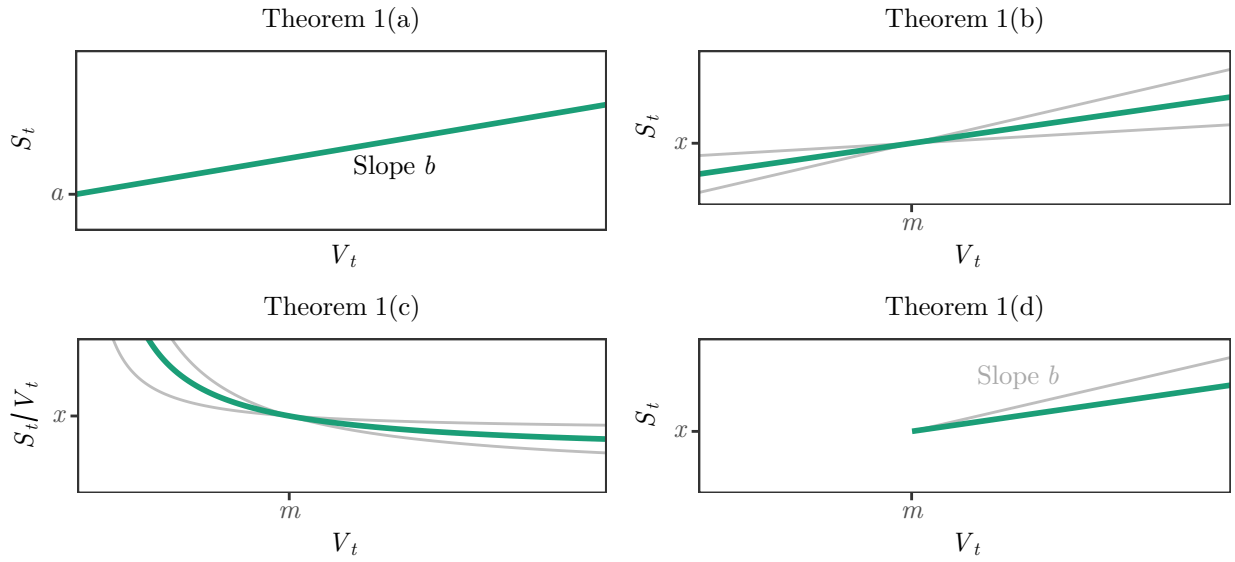


Figure 1: Illustration of the equivalent statements of 1, as described in the text.

# References

Ville, J. (1939), *Étude Critique de la Notion de Collectif.*, Gauthier-Villars, Paris.

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**Proof of the mother theorem**

Lecturer : Aaditya Ramdas

# 1 Proof of the mother theorem

Ville's maximal inequality for nonnegative supermartingales (Ville (1939); Durrett (2017), exercise 4.8.2), often attributed to Doob, is the foundation of all uniform bounds in this paper. It is an infinite-horizon uniform extension of Markov's inequality, asserting that a nonnegative supermartingale $(L_t)$ has probability at most $\mathbb{E}L_0/a$ of ever crossing level $a$: $\mathcal{P}(\exists t : L_t \geq a) \leq \mathbb{E}L_0/a$ for any $a > 0$. Applying this inequality to Assumption 1 gives, for any $\lambda \in (0, \lambda_{\max})$ and $z \in \mathbb{R}$,

$$\mathcal{P}(\exists t \in \mathcal{T} : \exp\{\lambda S_t - \psi(\lambda)V_t\} \geq e^z) \leq \mathcal{P}(\exists t \in \mathcal{T} : L_t \geq e^z) \leq (\mathbb{E}L_0)e^{-z}. \tag{1}$$

To derive Theorem 1(a) from (1), fix $a, b \geq 0$ and choose $\lambda \in [0, \lambda_{\max})$ such that $\psi(\lambda) \leq b\lambda$, supposing for the moment that some such value of $\lambda$ exists. Then

$$\begin{aligned}\mathcal{P}\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t\right) &= \mathcal{P}\left(\exists t \in \mathcal{T} : \exp\left\{\lambda S_t - b\lambda V_t\right\} \geq e^{a\lambda}\right) \\ &\leq \mathcal{P}\left(\exists t \in \mathcal{T} : \exp\left\{\lambda S_t - \psi(\lambda) V_t\right\} \geq e^{a\lambda}\right) \\ &\leq (\mathbb{E}L_0)e^{-a\lambda},\end{aligned}$$

applying (1) in the last step. This bound holds for all choices of $\lambda$ in the set $\{\lambda \in [0, \lambda_{\max}) : \psi(\lambda)/\lambda \leq b\}$, so to minimize the final bound, we take the supremum over this set, recovering the stated bound $(\mathbb{E}L_0)e^{-aD(b)}$ by the definition of $D(b)$. If no value $\lambda \in [0, \lambda_{\max})$ satisfies $\psi(\lambda) \leq b\lambda$, then $D(b) = 0$ by definition, so that the bound holds trivially. This shows that Assumption 1 implies Theorem 1(a).

To complete the proof we will show that the four parts of Theorem 1 are equivalent whenever $\psi$ is CGF-like. We require some simple facts about $\psi(\lambda)/\lambda$:

**Lemma 1** *Suppose $\psi$ is CGF-like with domain $[0, \lambda_{\max})$.*

1. *$\psi(\lambda)/\lambda < \psi'(\lambda)$ for all $\lambda \in (0, \lambda_{\max})$.*
2. *$\lambda \mapsto \psi(\lambda)/\lambda$ is continuous and strictly increasing on $\lambda > 0$.*
3. *$\inf_{\lambda \in (0, \lambda_{\max})} \psi(\lambda)/\lambda = 0$*
4. *$\sup_{\lambda \in (0, \lambda_{\max})} \psi(\lambda)/\lambda = \bar{b}$.*

5. *$\psi(D(b))/D(b) = b$ for any $b \in [0, \bar{b})$. That is, $D(b)$ is the inverse of $\psi(\lambda)/\lambda$.*

**Proof:** To see (i), write $\psi(\lambda) = \int_0^\lambda \psi'(t)\mathrm{d}t < \lambda\psi'(\lambda)$, where the inequality follows since $\psi$ is strictly convex so that $\psi'$ is strictly increasing. For (ii), the function is continuous because $\psi$ is continuous, and differentiating reveals it to be strictly increasing by part (i). L'Hôpital's rule implies (iii) along with the assumptions $\psi(\lambda) = \psi'(\lambda) = 0$, and implies (iv) along with the assumption $\sup_\lambda \psi(\lambda) = \infty$. Part (v) follows from the definition of $D(\cdot)$ and parts (ii), (iii) and (iv). ∎

We also repeatedly use the well-known fact about the Legendre-Fenchel transform that $\psi'^{-1}(u) = \psi^{\star\prime}(u)$, which follows by differentiating the identity $\psi^\star(u) = u\psi'^{-1}(u) - \psi(\psi'^{-1}(u))$.

- $(a) \implies (b)$: Fix $m > 0$ and $x \in [0, m\bar{b})$. Any line with slope $b \in [0, x/m]$ and intercept $x - mb$ passes through the point $(m, x)$ in the $(V_t, S_t)$ plane, and part $(a)$ yields

$$\begin{aligned}\mathcal{P}\left(\exists t \in \mathcal{T} : S_t \geq x + b(V_t - m)\right) &\leq (\mathbb{E}L_0)\exp\left\{-(x - mb)D(b)\right\} \\ &= (\mathbb{E}L_0)\exp\left\{-m\left(\frac{x}{m}\cdot D(b) - \psi(D(b))\right)\right\}\end{aligned}$$

  using 1(v) in the second step. Now we choose the slope $b$ to minimize the probability bound. The unconstrained optimizer $b_\star$ satisfies $\psi'(D(b_\star)) = x/m$, and a solution is guaranteed to exist by our restriction on $x$. This solution is given by $D(b_\star) = \psi'^{-1}(x/m) = \psi^{\star\prime}(x/m)$. Hence $b_\star = \mathfrak{s}(x/m)$ using the definition of $\mathfrak{s}(\cdot)$. 1(i) shows $x/m = \psi'(D(b_\star)) > \psi(D(b_\star))/D(b_\star) = b_\star$, verifying that $b_\star$ is in the allowed range for part $(a)$. Identify the Legendre-Fenchel transformation $\psi^\star(x/m) = \sup_b[(x/m)D(b) - \psi(D(b))]$ to complete the proof of part $(b)$. The fact that this bound is exact for Brownian motion shows that, since we have optimized over the slope, the bound could not hold in general for any other slope.

- $(b) \implies (c)$: Fix $m \geq 0$ and $x \in [0, \bar{b})$ and observe that

$$\mathcal{P}\left(\exists t \in \mathcal{T} : \frac{S_t}{V_t} \geq \mathfrak{s}(x) + \frac{m(x - \mathfrak{s}(x))}{V_t}\right) = \mathcal{P}\left(\exists t \in \mathcal{T} : S_t \geq mx + \mathfrak{s}(x)\cdot(V_t - m)\right).$$

  Now applying part $(b)$ with values $m$ and $mx$ yields part $(c)$.

- $(c) \implies (a)$: Fix $a \geq 0$ and $b \in [0, \bar{b})$. Set $x = \psi'(D(b))$ and $m = a/(x - \mathfrak{s}(x))$. Recalling $\psi^{\star\prime} = \psi'^{-1}$ we see that $\mathfrak{s}(x) = \psi(D(b))/D(b) = b$. Now apply part $(c)$ to obtain

$$\begin{aligned}\mathcal{P}\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t\right) &\leq (\mathbb{E}L_0)\exp\left\{-a\cdot\frac{\psi^\star(x)}{x - \mathfrak{s}(x)}\right\} \\ &= (\mathbb{E}L_0)\exp\left\{-a\cdot\frac{\psi^\star(x)\cdot\psi^{\star\prime}(x)}{x\psi^{\star\prime}(x) - \psi(\psi^{\star\prime}(x))}\right\}.\end{aligned}$$

Recognizing the Legendre-Fenchel transform in the denominator of the final exponent, we see that the probability bound equals $(\mathbb{E}L_0)\exp\{-a\psi^{\star\prime}(x)\}$. Again using $\psi^{\star\prime}(x) = \psi'^{-1}(x) = D(b)$ yields $(a)$.

- $(a, b) \implies (d)$: Clearly $\{\exists t \in \mathcal{T} : V_t \geq m,\, S_t \geq x + b(V_t - m)\} \subseteq \{\exists t \in \mathcal{T} : S_t \geq x + b(V_t - m)\}$, and the probability of the latter event is upper bounded by $(EL_0)\exp\{-(x - bm)D(b)\}$ from part $(a)$; the intercept $x - bm$ is nonnegative by our restriction on $b$. However, if $m > 0$ and $b \geq \mathfrak{s}(x/m)$, then $\{\exists t \in \mathcal{T} : V_t \geq m,\, S_t \geq x + b(V_t - m)\} \subseteq \{\exists t \in \mathcal{T} : V_t \geq m,\, S_t \geq x + \mathfrak{s}(x/m)(V_t - m)\}$; that is, we may replace the slope $b$ by the smaller slope $\mathfrak{s}(x/m)$. The probability of the latter event is upper bounded by $(\mathbb{E}L_0)\exp\{-m\psi^\star(x/m)\}$ by part $(b)$.
- $(d) \implies (a)$: set $m = 0$ and $x = a$ to recover part $(a)$.

It is worth noting here that, unlike the proofs of Freedman (1975), Khan (2009), Tropp (2011), and Fan et al. (2015), we do not explicitly construct a stopping time in our proof. While an optional stopping argument is hidden within the proof of Ville's inequality, the underlying stopping time here is different from that in the aforementioned citations.

# References


Durrett, R. (2017), *Probability: Theory and Examples*, 5a edn.

Fan, X., Grama, I. & Liu, Q. (2015), 'Exponential inequalities for martingales with applications', *Electronic Journal of Probability* **20**(1), 1–22.

Freedman, D. A. (1975), 'On Tail Probabilities for Martingales', *The Annals of Probability* **3**(1), 100–118.

Khan, R. A. (2009), '$L_p$-Version of the Dubins–Savage Inequality and Some Exponential Inequalities', *Journal of Theoretical Probability* **22**(2), 348.

Tropp, J. A. (2011), 'Freedman's inequality for matrix martingales', *Electronic Communications in Probability* **16**, 262–270.

Ville, J. (1939), *Étude Critique de la Notion de Collectif.*, Gauthier-Villars, Paris.

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**Matrix exponential, Lieb's inequality, proof of connector lemma**

Lecturer : Aaditya Ramdas

# 1 Spectral decomposition of Hermitian matrices $\mathcal{H}_d$

They are a generalization of real-symmetric matrices to complex values: they satisfy the property that $A^* = A$, where $A^*$ is the conjugate-transpose of the matrix $A$. For the standard Euclidean inner-product, this implies that $\langle Ax, y\rangle = \langle x, Ay\rangle$.

As a result of the spectral theorem, Hermitian matrices can be diagonalized, and the eigenvalues are all real. Let $V_\lambda := \{v : Av = \lambda v\}$ be the subspace formed by (linearly independent) eigenvectors with eigenvalue $\lambda$, and let $P_\lambda$ denote the orthogonal projection onto this subspace. Then, one can write $A$ in terms of its spectral decomposition:

$$A = \sum_{i=1}^{d} \lambda_i P_{\lambda_i}.$$

# 2 Functions on matrices

The above spectral theorem is useful because one can extend functions over the reals to functions of Hermitian matrices as

$$f(A) = \sum_i f(\lambda_i) P_{\lambda_i}.$$

For any interval $I \subseteq \mathbb{R}$, a function $f : I \mapsto \mathbb{R}$ is operator monotone if $A \preceq B$ implies that $f(A) \preceq f(B)$, it is operator convex if $f(\lambda A + (1-\lambda)B) \preceq \lambda f(A) + (1-\lambda) f(B)$, and it is operator concave if $-f$ is operator convex).

**Theorem 1 (Lowner-Heinz)** *(a) For $-1 \le p \le 0$, the function $f(t) = -t^p$ is operator monotone and operator concave.*

*(b) For $0 \le p \le 1$, the function $f(t) = t^p$ is operator monotone and operator concave.*

*(c) For $1 \le p \le 2$, the function $f(t) = t^p$ is operator convex.*

*(d) The function $f(t) = \log t$ is operator monotone and operator concave, while $f(t) = t \log t$ is operator convex.*

Given any function $f$, the corresponding trace function is given by $A \mapsto Tr(f(A)) = \sum_j f(\lambda_j)$. The trace function preserves monotonicity and convexity:

**Lemma 2 (Trace-function preservation lemma)** *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be continuous. If the function $t \mapsto f(t)$ is monotone/convex/strictly-convex, then $A \mapsto Tr(f(A))$ is also monotone/convex/strictly-convex on $\mathcal{H}_d$.*

# 3 Matrix exponential

Unless otherwise mentioned, all matrices will be $d \times d$. Recall that $A \preceq B$ denotes the positive semidefinite (psd) ordering, and $\mathcal{S}_+$ denotes the psd cone. Define the exponential of a matrix as

$$\exp(A) := \sum_k \frac{A^k}{k!}.$$

The exponential of a matrix effectively exponentiates its eigenvalues; that is, if $A = PDP^{-1}$ is the eigenvalue decomposition of $A$, then we have:

$$\exp(A) = P\mathrm{diag}(\exp(d_i))P^{-1} = P\exp(D)P^{-1}.$$

Some properties include: $\exp(0) = I$, $\exp(A)^T = \exp(A^T)$, and $\exp(A^*) = \exp(A)^*$ (where $A^*$ is the conjugate transpose of $A$). If $X$ and $Y$ commute, that is $XY = YX$, then $\exp(X)\exp(Y) = \exp(X+Y)$. Hence $\exp(X)\exp(-X) = I$, and so the matrix exponential is always invertible. Jacobi's formula implies that

$$\det(\exp(A)) = \exp(Tr(A)).$$

The matrix exponential appears naturally in the solution of ODEs. Indeed, the solution to $\dot{y}(t) = Ay(t), y(0) = 0$ is given by $y(t) = \exp(At)y_0$.

The matrix exponential results in a psd matrix. While the trace-exponential is monotone and strictly convex, the matrix exponential is neither operator monotone nor operator convex.

# 4 Lieb and Golden-Thompson

**Theorem 3 (Lieb)** *For any fixed Hermitian matrix $H$, the function $A \mapsto Tr\exp(H + \log A)$ is concave on $\mathcal{S}_+$.*

**Theorem 4 (Golden-Thompson)** *For Hermitian matrices $A, B$, we have*

$$Tr(\exp(A+B)) \leq Tr(\exp(A)\exp(B)).$$

### 4.1 Proof of Connector Lemma

Suppose $(Y_t)$ is sub-$\psi$ with self-normalizing process $(U_t)$ and variance process $(W_t)$. Fixing $\lambda \in [0, \lambda_{\max})$, Lieb's theorem and Jensen's inequality together imply

$$\mathbb{E}_{t-1}\mathrm{Tr}\exp\{\lambda Y_t - \psi(\lambda)\cdot(U_t + W_t)\} \leq \mathrm{Tr}\exp\{\lambda Y_{t-1} - \psi(\lambda)\cdot(U_{t-1} + W_t) + \log \mathbb{E}_{t-1} e^{\lambda \Delta Y_t - \psi(\lambda)\cdot \Delta U_t}\}.$$

Now we apply the sub-$\psi$ property to the expectation, using the monotonicity of the matrix logarithm and trace exponential to obtain

$$\mathbb{E}_{t-1}\mathrm{Tr}\exp\{\lambda Y_t - \psi(\lambda)\cdot(U_t + W_t)\} \leq \mathrm{Tr}\exp\{\lambda Y_{t-1} - \psi(\lambda)\cdot(U_{t-1} + W_{t-1})\}.$$

This shows that the process $L_t := \mathrm{Tr}\exp\{\lambda Y_t - \psi(\lambda)\cdot(U_t + W_t)\}$ is a supermartingale, with $L_0 = d$. Next we show that $L_t \geq \exp\{\lambda\gamma_{\max}(Y_t) - \psi(\lambda)\gamma_{\max}(U_t + W_t)\}$ a.s. for all $t$, which is the canonical assumption. We repeat a short argument from Tropp (2012). First, by the monotonicity of the trace exponential,

$$\begin{aligned}\mathrm{Tr}\exp\{\lambda Y_t - \psi(\lambda)\cdot(U_t + W_t)\} &\geq \mathrm{Tr}\exp\{\lambda Y_t - \psi(\lambda)\gamma_{\max}(U_t + W_t) I_d\} \\ &\geq \gamma_{\max}(\exp\{\lambda Y_t - \psi(\lambda)\gamma_{\max}(U_t + W_t) I_d\}) =: B.\end{aligned}$$

using the fact that the trace of a positive semidefinite matrix is at least as large as its maximum eigenvalue. Then the spectral mapping property gives

$$B = \exp\{\gamma_{\max}(\lambda Y_t - \psi(\lambda)\gamma_{\max}(U_t + W_t) I_d).$$

Finally, we use the fact that $\gamma_{\max}(A - cI_d) = \gamma_{\max}(A) - c$ for any $A \in \mathcal{H}^d$ and $c \in \mathbb{R}$ to see that $B = \exp\{\lambda\gamma_{\max}(Y_t) - \psi(\lambda)\gamma_{\max}(U_t + W_t)\}$, completing the argument.

# References

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

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**The big reference table**

Lecturer : Aaditya Ramdas

# 1 What conditions imply sub-$\psi$?

In what follows, the matrix conditional variance is $\mathrm{Var}_t X := \mathbb{E}_t X^2 - (E_t X)^2$. We let $I_d$ denote the $d \times d$ identity matrix. For a process $(Y_t)_{t\in\mathcal{T}}$, let $[Y]_t$ denote the quadratic variation and $\langle Y\rangle_t$ the conditional quadratic variation; in discrete time, $[Y]_t := \sum_{i=1}^t \Delta Y_i^2$ and $\langle Y\rangle_t := \sum_{i=1}^t \mathbb{E}_{i-1}\Delta Y_i^2$. In the discrete time case, we have the following known results.

**Fact 1** *Let $(Y_t)_{t\in\mathcal{N}}$ be any $\mathcal{H}^d$-valued martingale.*

1. *(Scalar parametric) If $d = 1$ and $Y_t$ is a cumulative sum of i.i.d., real-valued random variables, each of which is mean zero with known cumulant generating function $\psi(\lambda)$ that is finite on $\lambda \in [0, \lambda_{\max})$, then $(Y_t)$ is sub-$\psi$ with variance process $W_t = t$.*

2. *(Bernoulli) If $-gI_d \preceq \Delta Y_t \preceq hI_d$ a.s. for all $t \in \mathcal{N}$, then $(Y_t)$ is sub-Bernoulli with variance process $W_t = tI_d$ and range parameters $g, h$ (Hoeffding 1963, Tropp 2012).*

3. *(Bennett) If $\Delta Y_t \preceq cI_d$ a.s. for all $t \in \mathcal{N}$ for some $c > 0$, then $(Y_t)$ is sub-Poisson with variance process $W_t = \langle Y\rangle_t$ and scale parameter $c$ (Bennett 1962, Hoeffding 1963, Tropp 2012).*

4. *(Bernstein) If $\mathbb{E}_{t-1}(\Delta Y_t)^k \preceq (k!/2)c^{k-2}\mathrm{Var}_{t-1}(\Delta Y_t)$ for all $t \in \mathcal{N}$ and $k = 2, 3, \ldots$, then $(Y_t)$ is sub-gamma with variance process $W_t = \langle Y\rangle_t$ and scale parameter $c$ (Bernstein 1927, Tropp 2012, Boucheron et al. 2013).*

5. *(Heavy on left) Let $T_a(y) := (y \wedge a) \vee -a$ for $a > 0$ denote the truncation of $y$. If $d = 1$ and*

$$\mathbb{E}_{t-1} T_a(\Delta Y_t) \le 0 \quad \text{for all } a > 0, t \in \mathcal{N}, \tag{1}$$

*then $(Y_t)$ is sub-Gaussian with self-normalizing process $U_t = [Y]_t$. A random variable satisfying (1) is called* heavy on left, *and $(Y_t)$ need not be a martingale in this case (Bercu & Touati 2008, Delyon 2015, Bercu et al. 2015). When $-\Delta Y_t$ satisfies (1) we say $\Delta Y_t$ is* heavy on right.

| | Condition | $\psi$ | $U_t$ | $W_t$ |
|---|---|---|---|---|
| *Discrete time* | | | | |
| Parametric ($d = 1$) | $\Delta Y_t \overset{\text{i.i.d.}}{\sim} F$ | $\log \mathbb{E} e^{\lambda \Delta Y_1}$ | | $t$ |
| Bernoulli | $-gI_d \preceq \Delta Y_t \preceq hI_d$ | $\psi_B$ | | $tI_d$ |
| Bennett | $\Delta Y_t \preceq cI_d$ | $\psi_P$ | | $\langle Y \rangle_t$ |
| Bernstein | $\mathbb{E}_{t-1}(\Delta Y_t)^k \preceq \frac{k!}{2} c^{k-2} \mathbb{E}_{t-1} \Delta Y_t^2$ | $\psi_G$ | | $\langle Y \rangle_t$ |
| Heavy on left | $\mathbb{E}_{t-1} T_a(\Delta Y_t) \leq 0$ for all $a > 0$ | $\psi_N$ | $[Y]_t$ | |
| Hoeffding I | $-G_t I_d \preceq \Delta Y_t \preceq H_t I_d$ | $\psi_N$ | | $\sum_{i=1}^t \left(\frac{G_i + H_i}{2}\right)^2 I_d$ |
| Symmetric | $\Delta Y_t \sim -\Delta Y_t \mid \mathcal{F}_{t-1}$ | $\psi_N$ | $[Y]_t$ | |
| Bounded below | $\Delta Y_t \succeq -cI_d$ | $\psi_E$ | $[Y]_t$ | |
| Self-normalized I | $\mathbb{E}_{t-1} \Delta Y_t^2 < \infty$ | $\psi_N$ | $[Y]_t/3$ | $2\langle Y \rangle_t/3$ |
| Self-normalized II | $\mathbb{E}_{t-1} \Delta Y_t^2 < \infty$ | $\psi_N$ | $[Y_+]_t/2$ | $\langle Y_- \rangle_t/2$ |
| Hoeffding II | $\Delta Y_t^2 \preceq A_t^2$ | $\psi_N$ | | $\sum_{i=1}^t A_i^2$ |
| Cubic self-normalized | $\mathbb{E}_{t-1} \lvert \Delta Y_t \rvert^3 < \infty$ | $\psi_G$ | $[Y]_t$ | $\sum_{i=1}^t \mathbb{E}_{i-1} \lvert \Delta Y_i \rvert^3$ |
| *Continuous time ($d = 1$)* | | | | |
| Lévy | $\mathbb{E} e^{\lambda Y_1} < \infty$ | $\log \mathbb{E} e^{\lambda Y_1}$ | | $t$ |
| Bennett | $\Delta Y_t \leq c$ | $\psi_P$ | | $\langle Y \rangle_t$ |
| Bernstein | $V_{m,t} \leq \frac{m!}{2} c^{m-2} W_t$ | $\psi_G$ | | $W_t$ |
| Continuous paths | $\Delta Y_t \equiv 0$ | $\psi_N$ | | $\langle Y \rangle_t$ |

Table 1: Summary of sufficient conditions for a martingale $(Y_t)$ to be sub-$\psi$ with the given self-normalizing and variance processes. See text for details of each case.

In addition, we give the following novel results for matrices by extending the corresponding scalar results. Here $[Y_+]_t := \sum_{i=1}^t \max(0, \Delta Y_i)^2$ and $\langle Y_- \rangle_t := \sum_{i=1}^t \mathbb{E}_{i-1} \min(0, \Delta Y_i)^2$, where the functions $\max(0, \cdot)$ and $\min(0, \cdot)$ extend to $^d$ by truncating the eigenvalues.

**Lemma 2** *Let $(Y_t)_{t \in \mathcal{N}}$ be any $\mathcal{H}^d$-valued martingale.*

1. *(Hoeffding I) If $-G_t I_d \preceq \Delta Y_t \preceq H_t I_d$ a.s. for all $t \in \mathcal{N}$ for some real-valued, pre-*

*dictable sequences* $(G_t)$ *and* $(H_t)$*, then* $(Y_t)$ *is sub-Gaussian with variance process* $W_t = \left[\sum_{i=1}^t (G_i + H_i)^2/4\right] I_d$*.*

2. *(Conditionally symmetric) If* $\Delta Y_t$ *and* $-\Delta Y_t$ *have the same distribution conditional on* $\mathcal{F}_{t-1}$ *for all* $t \in \mathcal{N}$*, then* $(Y_t)$ *is sub-Gaussian with self-normalizing process* $U_t = [Y]_t$*. In this case,* $(Y_t)$ *need not be a martingale, i.e., it need not be integrable.*

3. *(Bounded from below) If* $\Delta Y_t \succeq -cI_d$ *a.s. for all* $t \in \mathcal{N}$ *for some* $c > 0$*, then* $(Y_t)$ *is sub-exponential with self-normalizing process* $U_t = [Y]_t$ *and scale parameter* $c$*.*

4. *(General self-normalized I) If* $\mathbb{E}_{t-1}\Delta Y_t^2$ *is finite for all* $t \in \mathcal{N}$*, then* $(Y_t)$ *is sub-Gaussian with self-normalizing process* $U_t = [Y]_t/3$ *and variance process* $W_t = 2\langle Y \rangle_t /3$*.*

5. *(General self-normalized II) If* $\mathbb{E}_{t-1}\Delta Y_t^2$ *is finite for all* $t \in \mathcal{N}$*, then* $(Y_t)$ *is sub-Gaussian with self-normalizing process* $U_t = [Y_+]_t/2$ *and variance process* $W_t = \langle Y_- \rangle_t /2$*.*

6. *(Hoeffding II) If* $\Delta Y_t^2 \preceq A_t^2$ *a.s. for all* $t \in \mathcal{N}$ *for some* $\mathcal{H}^d$*-valued predictable sequence* $(A_t)$*, then* $(Y_t)$ *is sub-Gaussian with* $W_t = \sum_{i=1}^t A_i^2$*.*

7. *(Cubic self-normalized) If* $\mathbb{E}_{t-1}|\Delta Y_t|^3$ *is finite for all* $t \in \mathcal{N}$*, then* $(Y_t)$ *is sub-gamma with self-normalizing process* $U_t = [Y]_t$*, variance process* $W_t = \sum_{i=1}^t \mathbb{E}_{i-1}|\Delta Y_i|^3$*, and scale parameter* $c = 1/6$*.*

# References


Bennett, G. (1962), 'Probability Inequalities for the Sum of Independent Random Variables', *Journal of the American Statistical Association* **57**(297), 33–45.

Bercu, B., Delyon, B. & Rio, E. (2015), *Concentration Inequalities for Sums and Martingales*, Springer International Publishing, Cham.

Bercu, B. & Touati, A. (2008), 'Exponential inequalities for self-normalized martingales with applications', *The Annals of Applied Probability* **18**(5), 1848–1869.

Bernstein, S. (1927), *Theory of probability*, Gastehizdat Publishing House, Moscow.

Boucheron, S., Lugosi, G. & Massart, P. (2013), *Concentration inequalities: a nonasymptotic theory of independence*, 1st edn, Oxford University Press, Oxford.

Delyon, B. (2015), Exponential inequalities for dependent processes, Technical report.

Hoeffding, W. (1963), 'Probability Inequalities for Sums of Bounded Random Variables', *Journal of the American Statistical Association* **58**(301), 13–30.

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

36-771 Fall 2018

Martingales 1 : Concentration inequalities

# Improving Cramer-Chernoff & Freedman's, Hermitian dilation

Lecturer : Aaditya Ramdas

In the discrete-time, scalar setting, a simple sufficient condition for Assumption 1 is that

$$\mathbb{E}_{t-1} \exp\{\lambda \Delta S_t - \psi(\lambda) \Delta V_t\} \leq 1, \quad \forall t,$$

which is the standard assumption for a martingale-method Cramér-Chernoff inequality (McDiarmid 1998, Chung & Lu 2006, Boucheron et al. 2013). When $V_t$ is deterministic, the fixed-time Cramér-Chernoff method gives, for fixed $t$ and $x$,

$$\mathcal{P}(S_t \geq x) \leq \exp\{-V_t \psi^\star\left(\frac{x}{V_t}\right)\}, \tag{1}$$

so Theorem 1(b) is a uniform *extension* of the Cramér-Chernoff inequality, losing nothing at the fixed time $t$ [B; C or D]. A stopping time argument due to Freedman (1975) extends this to the uniform bound

$$\mathcal{P}(\exists t \in \mathcal{T} : S_t \geq x \text{ and } V_t \leq m) \leq \exp\{-m\psi^\star\left(\frac{x}{m}\right)\}.$$

When $V_t$ is deterministic, analogous uniform bounds follow from Doob's maximal inequality for submartingales, as in Hoeffding (1963, eq. 2.17). Theorem 1(b) strengthens this "Freedman-style" inequality [B; C or D], since it yields tighter bounds for all times $t$ such that $V_t < m$, and also extends the inequality to hold for all times $t$ with $V_t > m$, as illustrated by the figure.

Tropp (2011, 2012) extends the scalar Cramér-Chernoff approach to random matrices via control of the matrix moment-generating function, giving matrix analogues of Hoeffding's, Bennett's, Bernstein's and Freedman's inequalities. Following this approach, Theorem 1 gives corresponding strengthened versions of these inequalities for matrix-valued processes [B].

We summarize explicit results for special cases below. Recall the definitions of $\mathfrak{s}_P, \psi_P^\star, \mathfrak{s}_G, \psi_G^\star$ from earlier.

**Corollary 1** *Let $\mathcal{T} = \mathbb{N}$ and $(Y_t)_{t\in\mathbb{N}}$ be an adapted, $\mathcal{H}^d$-valued martingale, or let $\mathcal{T} = (0,\infty)$ and $(Y_t)_{t\in(0,\infty)}$ be an adapted, real-valued local martingale. Let $S_t := \gamma_{\max}(Y_t)$.*

*(a) When $\mathcal{T} = \mathbb{N}$, suppose $\Delta Y_t^2 \preceq A_t^2$ a.s. for all $t$ for some $\mathcal{H}^d$-valued predictable sequence $(A_t)$, and let either $V_t := \frac{1}{2}\gamma_{\max}\left(\langle Y\rangle_t + \sum_{i=1}^t A_i^2\right)$ or $V_t := \gamma_{\max}\left(\sum_{i=1}^t A_i^2\right)$. Then for*

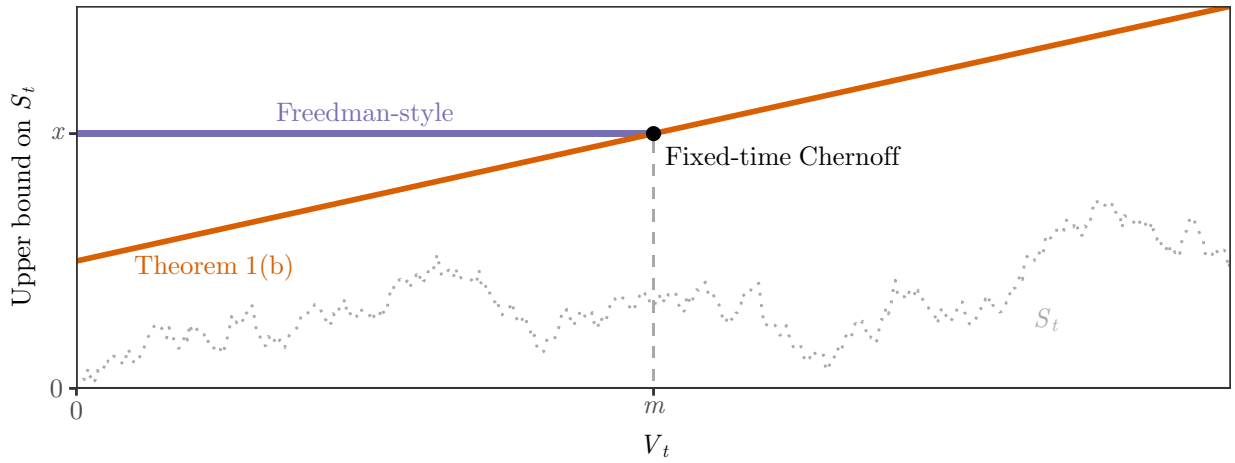


Figure 1: Comparison of (i) fixed-time Cramér-Chernoff bound, which bounds the deviations of $S_m$ at a fixed time $m$; (ii) "Freedman-style" constant uniform bound, which bounds the deviations of $S_t$ for all $t$ such that $V_t \leq m$, with a constant boundary equal in value to the fixed-time Cramér-Chernoff bound; and (iii) linear uniform bound from Theorem 1, which bounds the deviations of $S_t$ for all $1 \leq n < \infty$, with a boundary growing linearly in $V_t$. Each bound gives the same tail probability and thus implies the preceding one.

*any $x, m > 0$, we have*

$$\mathcal{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \frac{x}{2m}(V_t - m)\right) \leq d \exp\{-\frac{x^2}{2m}\}.$$

*This strengthens Hoeffding's inequality (Hoeffding 1963) [A,B,D] and its matrix analogues in Tropp (2012, Theorem 7.1) [B,E] and Mackey et al. (2014, Corollary 4.2) [A,B].*

*(b) Suppose $\gamma_{\max}(\Delta Y_t) \leq c$ a.s. for all $t$ for some constant $c$, and let $V_t := \gamma_{\max}(\langle Y \rangle_t)$. Then for any $x, m > 0$, we have*

$$\mathcal{P}\left(\exists t \in \mathcal{T} : S_t \geq x + \mathfrak{s}_P\left(\frac{x}{m}\right) \cdot (V_t - m)\right) \leq d \exp\{-m\psi_P^\star\left(\frac{x}{m}\right)\} \leq d \exp\{-\frac{x^2}{2(m + cx/3)}\}.$$

*This strengthens Bennett's and Freedman's inequalities (Bennett 1962, Freedman 1975) [B; C or D] for scalars and the corresponding matrix bounds from Tropp (2011, 2012) [B].*

*(c) Suppose $(Y_t)$ is sub-gamma with self-normalizing process $(U_t)$, variance process $(W_t)$ and scale parameter $c$, and let $V_t := \gamma_{\max}(U_t + W_t)$. Then for any $x, m > 0$, we have*

$$\mathcal{P}\left(\exists t \in \mathcal{T} : S_t \geq x + \mathfrak{s}_G\left(\frac{x}{m}\right) \cdot (V_t - m)\right) \leq d \exp\{-m\psi_G^\star\left(\frac{x}{m}\right)\} \leq d \exp\{-\frac{x^2}{2(m + cx)}\}.$$

*This strengthens Bernstein's inequality (Bernstein 1927) [B; C or D], along with the matrix Bernstein inequality (Tropp 2012) [B].*

The first setting of $V_t$ in case (a) follows from the bound $[Y_+]_t \preceq \sum_{i=1}^t A_i^2$, and further upper bounding $\langle Y_- \rangle_t \preceq \sum_{i=1}^t A_i^2$ yields the second setting of $V_t$. As is well known, 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 $\psi_P^\star \leq \psi_N^\star$, so neither subsumes the other. Additionally, the Hoeffding-style bound requires two-sided boundedness of increments while the Bennett-style bound requires only an upper bound on the deviations of increments above their expectations. It is also worth remarking that $\psi_P^\star(u) \geq \frac{u}{2c} arcsinh\left(\frac{cu}{2}\right)$, so the Bennett-style inequality in part (b) is an improvement on the inequality of Prokhorov (1959) for sums of independent random variables, as noted by Hoeffding (1963), as well as its extension to martingales in de la Peña (1999).

As an example of the Hermitian dilation technique, we give a bound for rectangular matrix Gaussian and Rademacher series, following Tropp (2012); here $\|A\|_{op}$ denotes the largest singular value of $A$. The proof will be given later.

**Corollary 2** *Let $\mathcal{T} = \mathbb{N}$, consider a sequence $(B_t)_{t\in\mathbb{N}}$ of fixed matrices with dimension $d_1 \times d_2$, and let $(\epsilon_t)_{t\in\mathbb{N}}$ be a sequence of independent standard normal or Rademacher variables. Let $S_t := \|\sum_{i=1}^t \epsilon_i B_i\|_{op}$ and $V_t := \max\{\|\sum_{i=1}^t B_i B_i^\star\|_{op}, \|\sum_{i=1}^t B_i^\star B_i\|_{op}\}$. Then for any $x, m > 0$, we have*

$$\mathcal{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \frac{x}{2m}(V_t - m)\right) \leq (d_1 + d_2)\exp\{-\frac{x^2}{2m}\}.$$

*This strengthens Corollary 4.2 of Tropp (2012) [B].*

**Proof:** Define the $\mathcal{H}^{d_1+d_2}$-valued process $(Y_t)$ using the dilation of $B_t$:

$$\Delta Y_t := \epsilon_t \begin{pmatrix} 0 & B_t \\ B_t^\star & 0 \end{pmatrix}.$$

Since the dilation operation is linear and preserves spectral information, $\gamma_{\max}(Y_t) = \|\sum_{i=1}^t \epsilon_i B_i\|_{op}$ (Tropp 2012, Eq. 2.12). Furthermore, since each $B_i$ is fixed and $\epsilon_i$ is 1-sub-Gaussian, $(Y_t)$ is sub-Gaussian with variance process

$$W_t = \sum_{i=1}^t \begin{pmatrix} B_i B_i^\star & 0 \\ 0 & B_i^\star B_i \end{pmatrix},$$

which has $\|W_t\|_{op} = \max\{\|\sum_{i=1}^t B_i B_i^\star\|_{op}, \|\sum_{i=1}^t B_i^\star B_i\|_{op}\}$ (Tropp 2012, Lemma 4.3). The result now follows the connector lemma and Theorem 1(b) applied to $(Y_t)$ and $(W_t)$. ∎

# References


Bennett, G. (1962), 'Probability Inequalities for the Sum of Independent Random Variables', *Journal of the American Statistical Association* **57**(297), 33–45.

Bernstein, S. (1927), *Theory of probability*, Gastehizdat Publishing House, Moscow.

Boucheron, S., Lugosi, G. & Massart, P. (2013), *Concentration inequalities: a nonasymptotic theory of independence*, 1st edn, Oxford University Press, Oxford.

Chung, F. & Lu, L. (2006), 'Concentration inequalities and martingale inequalities: a survey', *Internet Mathematics* **3**(1), 79–127.

de la Peña, V. H. (1999), 'A General Class of Exponential Inequalities for Martingales and Ratios', *The Annals of Probability* **27**(1), 537–564.

Freedman, D. A. (1975), 'On Tail Probabilities for Martingales', *The Annals of Probability* **3**(1), 100–118.

Hoeffding, W. (1963), 'Probability Inequalities for Sums of Bounded Random Variables', *Journal of the American Statistical Association* **58**(301), 13–30.

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B. & Tropp, J. A. (2014), 'Matrix concentration inequalities via the method of exchangeable pairs', *The Annals of Probability* **42**(3), 906–945.

McDiarmid, C. (1998), Concentration, *in* M. Habib, C. McDiarmid, J. Ramirez-Alfonsin & B. Reed, eds, 'Probabilistic Methods for Algorithmic Discrete Mathematics', Spinger, New York, pp. 195–248.

Prokhorov, Y. V. (1959), 'An Extremal Problem in Probability Theory', *Theory of Probability & Its Applications* **4**(2), 201–203.

Tropp, J. A. (2011), 'Freedman's inequality for matrix martingales', *Electronic Communications in Probability* **16**, 262–270.

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

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 $\mu$-a.e.). When $\mathcal{X} = \mathbb{R}$ or $\mathcal{X} = \mathcal{C}$ and $\mu$ 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

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

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, $\ell^\infty$ is not reflexive, even though $\ell_p$ and $\ell_q$ are dual and reflexive whenever $1/p + 1/q = 1$ and $p, q \notin \{1, \infty\}$, and even though for $d$-dimensional sequences, $\ell_d^\infty$ is dual to $\ell_d^1$.

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

$$\text{the gradient } \nabla f \text{ of function } f \text{ is an element of the dual space.}$$

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 $(q, \lambda)$-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

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

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

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

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 $(q, \lambda)$-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.

36-771 Fall 2018

Martingales 1 : Concentration inequalities

# Improving continuous time martingale concentration

Lecturer : Aaditya Ramdas

We start with a probability space $(\Omega, \mathcal{F}, P)$. In addition to that, consider a filtration $(\mathcal{F}_t)$ meaning a sequence of sigma-algebras such that $s \leq t$ implies $\mathcal{F}_s \subseteq \mathcal{F}_t$. One typically assumes that the resulting filtered complete probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ satisfies the "usual" hypotheses (from Protter):

- $\mathcal{F}_0$ contains all the $P$-null sets of $\mathcal{F}$.
- The filtration is right-continuous, meaning $\mathcal{F}_t = \bigcap_{u>t} \mathcal{F}_u$.

A random variable $T$ is called a stopping time if $\{T \leq t\} \in \mathcal{F}_t$ for all $0 \leq t \leq \infty$. Because the filtration is right continuous, we also have that the event $\{T < t\} \in \mathcal{F}_t$ for every $0 \leq t \leq \infty$ if and only if $T$ is a stopping time.

A stochastic process $(X_t)$ is a collection of $\mathbb{R}$-valued (or $\mathbb{R}^d$-valued) random variables. $(X_t)$ is called adapted if $X_t \in \mathcal{F}_t$ for every $t$. There are two different concepts of equality of two stochastic processes:

- $(X_t)$ and $(Y_t)$ are modifications if

$$\text{for each } t, \text{ we have } X_t = Y_t \text{ almost surely.}$$

- $(X_t)$ and $(Y_t)$ are indistinguishable if

$$\text{almost surely, we have } X_t = Y_t \text{ for each } t.$$

The second is much stronger.

A stochastic process is called cadlag, if it almost surely has sample paths that are right-continuous with left limits. If $(X_t)$ and $(Y_t)$ are modifications that have right continuous paths almost surely, then they are indistinguishable. Hence for cadlag processes, the above two concepts are identical.

A continuous time process $(M_t)$ is a martingale with respect to filtration $(\mathcal{F}_t)$ if for any $0 \leq s \leq t$, we have $\mathbb{E}[M_t|\mathcal{F}_s] = M_s$. Every martingale has a right continuous modification that is cadlag, and without further mention, we will always assume we are dealing with this version of the martingale.

# 1 Brownian Motion, Poisson process, Levy process

An $(\mathcal{F}_t)$ adapted process $(W_t)$ is a Brownian motion if

- $W_0 = 0$ a.s.
- Independent increments: for all $0 \leq s \leq t$, $W_t - W_s$ is independent of $\mathcal{F}_s$.
- Stationary increments: $W_t - W_s \sim N(0, t-s)$.
- The sample paths are continous a.s.

**Fact 1** *$W_t$ and $W_t^2 - t$ are martingales. Also for any $\lambda > 0$, $\exp(\lambda W_t - \frac{\lambda^2}{2} t)$ is a martingale.*

An $(\mathcal{F}_t)$ adapted process $(N_t)$ is a Poisson process if the last two properties are

- Stationary increments: $N_t - N_s \sim Poi(\mu(t-s))$ for some $\mu > 0$.
- The sample paths are continuous in probability:

$$\forall \epsilon > 0, t \geq 0, \text{ we have } \lim_{h \to 0} P(|N_{t+h} - N_t| > \epsilon) = 0.$$

**Fact 2** *$N_t - \mu t$ and $(N_t - \mu t)^2 - \mu t$ are both martingales. Also for any $\lambda > 0$, we have $\exp(\lambda N_t - \mu t(e^\lambda - 1)) = \exp(\lambda(N_t - \mu t) - \mu t(e^\lambda - \lambda - 1))$ is a martingale.*

These are both examples of Levy processes, which are all characterized by $X_0 = 0$, independent increments, stationary increments, and continuity in probability.

**Fact 3** *If a Levy process $(Y_t)$ satisfies $\mathbb{E}\exp(\lambda Y_1) < \infty$ then Assumption 1 is satisfied with $S_t = Y_t - \mathbb{E}[Y_t]$ and $\psi(\lambda) = \log \mathbb{E}\exp(\lambda Y_1)$ and $V(t) = t$. (the above two facts are special cases of this one.)*

Levy processes are "infinitely divisible" meaning that for any integer $n$ and time $t$, the law of $X_t$ matches the law of the sum of $n$ iid random variables $X_{t/n}, X_{t2/n} - X_{t/n}, ....$

Some other examples include the Gamma process, the compound poisson process and "stable" processes like the Cauchy process. In fact by the Levy-Khinchtine representation theorem, a Levy process can be uniquely defined by three components $(a, \sigma, \Pi)$ with correspond respectively to a drift term, a brownian motion variance term, and a measure defining a compound Poisson process.

## 2 Local properties

For a continous stochastic process $(X_t)$, a property $\pi$ is said to hold locally if there exists a sequence of stopping times $(T_n)_{n\in\mathbb{N}}$ increasing to infinity a.s. such that for every $n$, the stopped process $(X_{T_n\wedge t}1_{T_n>0})$ has property $\pi$.

This allows us to differentiate between a square-integrable "local martingale", and a "locally square integrable" martingale. Every martingale is a local martingale. Every bounded local martingale is a martingale. Every local martingale bounded from below is a supermartingale. In general a local martingale is not a martingale, because its expectation can be distorted by large values of small probability.

Local martingales form a very important class of processes in the theory of stochastic calculus. This is because the local martingale property is preserved by the stochastic integral, but the martingale property is not. Further, all continuous local martingales are time-changes of brownian motions.

## 3 Variation of a function/process

For $T > 0$ let $\pi$ be a partition of $[0, T]$: $0 = t_0^\pi < t_1^\pi < \cdots < t_k^\pi = T$, and define $\|\pi\| := \max_{1\le i\le k} t_i^\pi - t_{i-1}^\pi$. For a function $f : [0, T] \to \mathbb{R}$, and $p \ge 1$, define the notion of $p$-variation of $f$ on $[0, T]$ as

$$V^{(p)}(f) := \lim_{\|\pi\|\to 0, k\to\infty} \sum_{i=1}^{k} |f(t_i^\pi) - f(t_{i-1}^\pi)|^p$$

provided the limit exists. $p = 1$ is called the total variation on $[0, t]$, and $p = 2$ is called the quadratic variation of $f$ on $[0, T]$.

Analogously replacing the function $f$ with a stochastic process $(X_t)$, one defines the total variation process and quadratic variation process $([X]_t)$, provided the limit exists in the sense of convergence in probability. The total variation process of any nonzero continuous $(M_t)$ is a.s. infinite on any interval.

Confusingly, there is another notion of $p$-variation defined in functional analysis, defined for functions $f : \mathbb{R} \to (M, d)$ where the latter is a metric space. This notion takes a supremum over finite partitions $\pi$ (though they are finite, they can be arbitrarily fine).

$$\|f\|_{p-var} = \sup_\pi \left( \sum_{t_k\in\pi} [d(f(t_k) - f(t_{k-1}))]^p \right)^{1/p}$$

When $p = 1$, this is also called total variation, and the class of functions where this is finite, is called the class of bounded variation.

The 2-variation could be much larger than the quadratic variation. For eg: for a brownian motion, its quadratic variation is $t$, while its p-variation is infinite for $p \le 2$.

# 4 Doob-Meyer decomposition

In discrete time, $\langle M \rangle_t$ can be defined as the unique predictable and increasing process such that $M_t^2 - \langle M \rangle_t$ is a martingale.

Consider a locally square-integrable martingale $(M_t)$ wrt filtration $(\mathcal{F}_t)$, meaning that $\mathbb{E}M_t^2 < \infty$ for all $t \geq 0$ and $M_t$. (Recall that if $M_t$ is a martingale then $M_t^2$ is a submartingale.) According to the Doob-Meyer decomposition theorem, there exists a unique nondecreasing stochastic process $(A_t)$ adapted to $(\mathcal{F}_t)$, starting at 0 with right-continuous paths, such that $(M_t^2 - A_t)$ is a martingale wrt $(\mathcal{F}_t)$. Then $A_t$ is the quadratic variation $[M]_t$ of $M_t$.

There is also a unique nondecreasing right-continuous predictable process $(A_t)$ such that $(M_t^2 - A_t)$ is a martingale, and this $A_t$ is denoted by $\langle M \rangle_t$.

**Fact 4** *For continuous local martingales we also have* $[M]_t = \langle M \rangle_t$.

For a Brownian motion or Wiener process $W_t$, we have $\langle W \rangle_t = [W]_t = t$ almost surely, so we say that the BM accumulates quadratic variation at rate one per unit time. (in stochastic calculus, we write $dW_t dW_t = dt$, and also $dW_t dt = 0$ and $dtdt = 0$.) In fact, this property is characteristic of a BM:

**Fact 5** *If* $(M_t)$ *is a martingale with continuous paths and* $(M_t^2 - t)$ *is a martingale, then* $(M_t)$ *is a BM.*

# 5 Concentration inequalities

While many of the earlier results already generalize results known in discrete time to new results for continuous-time martingales [C], here we summarize a few more useful bounds explicitly for continuous-time processes which are corollaries of the mother theorem.

**Corollary 6** *Let* $(S_t)_{t \in (0,\infty)}$ *be a real-valued process.*

*(a) If* $(S_t)$ *is a locally square-integrable martingale with a.s. continuous paths, then*

$$\mathcal{P}\left(\exists t \in (0,\infty) : S_t \geq a + b \langle S \rangle_t\right) \leq \exp\{-2ab\}.$$

*If* $\langle S \rangle_t \uparrow \infty$ *as* $t \uparrow \infty$*, then the probability inequality may be replaced with an equality. This recovers as a special case the standard line-crossing probability for Brownian motion (e.g., Durrett 2017, Exercise 7.5.2).*

*(b) If $(S_t)$ is a local martingale with $\Delta S_t \leq c$ for all $t$, then*

$$\mathcal{P}\left(\exists t \in (0,\infty) : S_t \geq x + \mathfrak{s}_P\left(\frac{x}{m}\right) \cdot (\langle S \rangle_t - m)\right) \leq d\exp\{-m\psi_P^\star\left(\frac{x}{m}\right)\} \leq d\exp\{-\frac{x^2}{2(m+cx/3)}\}. \tag{1}$$

*This strengthens Appendix B, Inequality 1 of Shorack & Wellner (1986) [B].*

*(c) If $(S_t)$ is any locally square-integrable martingale satisfying the Bernstein condition of (see the Big Table lecture) for some predictable process $(W_t)$, then*

$$\mathcal{P}\left(\exists t \in (0,\infty) : S_t \geq x + \mathfrak{s}_G\left(\frac{x}{m}\right) \cdot (V_t - m)\right) \leq d\exp\{-m\psi_G^\star\left(\frac{x}{m}\right)\} \leq d\exp\{-\frac{x^2}{2(m+cx)}\}.$$

*This strengthens Lemma 2.2 of van de Geer (1995) [B,E].*

Clearly, statement (b) applies to centered Poisson processes with $c = 1$, but this can be inferred directly from the fact that it is a Levy process. The point of statement (b) is that any local martingale with bounded jumps obeys this inequality, and so concentrates like a centered Poisson process in this sense.

# References


Durrett, R. (2017), *Probability: Theory and Examples*, 5a edn.

Shorack, G. R. & Wellner, J. A. (1986), *Empirical processes with applications to statistics*, Wiley, New York.

van de Geer, S. (1995), 'Exponential Inequalities for Martingales, with Application to Maximum Likelihood Estimation for Counting Processes', *The Annals of Statistics* **23**(5), 1779–1801.

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**Improving de la Peña's self-normalized inequalities**

Lecturer : Aaditya Ramdas

# 1 Self-normalized uniform bounds

De la Peña (1999) and de la Peña et al. (2004, 2007, 2009) give a variety of sufficient conditions for Assumption 1 to hold with equality in the scalar case in both discrete- and continuous-time settings. They formulate their bounds for ratios involving $S_t$ in the numerator and $V_t$ in the denominator, as in Theorem 1(c), and they often specify initial-time conditions, as in Theorem 1(d). In this section we draw some direct comparisons between Theorem 1 and their results. As a first example, consider the boundary of Theorem 1(c) for the ratio $S_t/V_t$, strictly decreasing towards the asymptotic level $\mathfrak{s}(x)$. In particular, at time $V_t = m$ the boundary equals $x$, so Theorem 1(c) strengthens various theorems of de la Peña (1999) and de la Peña et al. (2007) which use a constant boundary after time $V_t = m$ [B; C or D]. The figure below illustrates the relationship between the boundary of Theorem 1(c) and those of de la Peña et al. As before, we give explicit results for special cases.

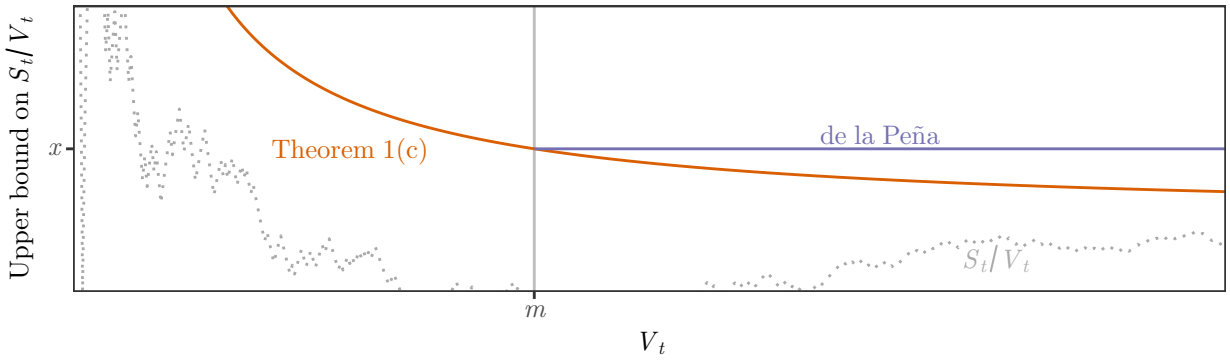


Figure 1: Comparing our decreasing boundary from Theorem 1(c) to the constant boundaries of de la Peña (1999).

**Corollary 1** *Let $\mathcal{T} = \mathcal{N}$ and $(Y_t)_{t\in\mathcal{N}}$ be an adapted, $\mathcal{H}^d$-valued process, or let $\mathcal{T} = (0,\infty)$ and $(Y_t)_{t\in(0,\infty)}$ be an adapted, real-valued process. Suppose $(Y_t)$ is sub-gamma with self-normalizing process $(U_t)$, variance process $(W_t)$ and scale parameter $c$, and let $S_t := \gamma_{\max}(Y_t)$, $V_t := \gamma_{\max}(U_t + W_t)$. Then for any $x, m \geq 0$, we have*

$$\mathcal{P}\left(\exists t \in \mathcal{T} : \frac{S_t}{V_t} \geq \mathfrak{s}_G(x)\left(1 + \frac{m\sqrt{1+2cx}}{V_t}\right)\right) \leq d\exp\{-m\psi_G^\star(x)\} \leq d\exp\{-\frac{mx^2}{2(1+cx)}\}.$$

*This strengthens the final statement of Theorem 1.2B of de la Peña (1999) [B; C or D]. In the sub-Gaussian case (obtained as $c \to 0$), the above bound simplifies to:*

$$\mathcal{P}\left(\exists t \in \mathcal{T} : \frac{S_t}{V_t + m} \geq x\right) \leq d \exp\{-2mx^2\}.$$

*This strengthens Theorem 2.1 of de la Peña et al. (2007) and Theorem 6.1 of de la Peña (1999) [B, C or D].*

More generally, when we normalize by $\alpha + \beta V_t$ and include an initial time condition $V_t \geq m$, Theorem 1(d) becomes the following:

**Corollary 2** *If Assumption 1 holds for some real-valued processes $(S_t)_{t\in\mathcal{T}}$ and $(V_t)_{t\in\mathcal{T}}$, then*

$$\mathcal{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } \frac{S_t}{\alpha + \beta V_t} \geq x\right) \leq \begin{cases} (\mathbb{E}L_0) \exp\{-\alpha x D(\beta x)\}, & \beta x \leq \mathfrak{s}\left(\frac{x(\alpha+\beta m)}{m}\right) \\ (\mathbb{E}L_0) \exp\{-m\psi^\star\left(\frac{x(\alpha+\beta m)}{m}\right)\}, & \beta x \geq \mathfrak{s}\left(\frac{x(\alpha+\beta m)}{m}\right) \end{cases}$$
$$\leq (\mathbb{E}L_0) \exp\{-m\psi^\star(\beta x) - \alpha x \psi^{\star\prime}(\beta x)\}.$$

*For the sub-Gaussian case, let $\mathcal{T} = \mathcal{N}$ and $(Y_t)_{t\in\mathcal{N}}$ be an adapted, $\mathcal{H}^d$-valued process, or let $\mathcal{T} = (0, \infty)$ and $(Y_t)_{t\in(0,\infty)}$ be an adapted, real-valued process. Suppose $(Y_t)$ is sub-Gaussian with self-normalizing process $(U_t)$ and variance process $(W_t)$, and let $S_t := \gamma_{\max}(Y_t)$, $V_t := \gamma_{\max}(U_t + W_t)$. Then for any $\alpha, \beta, m \geq 0$, we have*

$$\mathcal{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } \frac{S_t}{\alpha + \beta V_t} \geq x\right) \leq \exp\{-x^2\left(2\alpha\beta + \frac{(\beta m - \alpha)^2}{2m}\mathbf{1}\{\alpha \leq \beta m\}\right)\}.$$

*This improves the final statement in Theorem 6.2 of de la Peña (1999) [B; C or D; E].*

A defining feature of self-normalized bounds is that they involve an intrinsic time process $(V_t)$ constructed with the squared observations themselves rather than just conditional variances or constants. Such normalization can be found in common statistical procedures such as the $t$-test. Furthermore, it allows for Gaussian-like concentration while reducing or eliminating moment conditions.

**Corollary 3** *Suppose $\mathcal{T} = \mathcal{N}$ and $(Y_t)_{t\in\mathcal{N}}$ is an $\mathcal{H}^d$-valued martingale with $\mathbb{E}Y_t^2 < \infty$ for all $t \in \mathcal{N}$, and let $S_t := \gamma_{\max}(Y_t)$ and either $V_t := \frac{1}{2}\gamma_{\max}([Y_+]_t + \langle Y_-\rangle_t)$ or $V_t := \frac{1}{3}\gamma_{\max}([Y]_t + 2\langle Y\rangle_t)$. Then for any $x, m \geq 0$, we have*

$$\mathcal{P}\left(\exists t \in \mathcal{N} : \frac{S_t}{V_t + m} \geq x\right) \leq d \exp\{-2mx^2\}.$$

*This strengthens the third statement in Theorem 4 of Delyon (2009) [B,D], Theorem 2.1 of Bercu & Touati (2008) [B,D,E], and an implicit self-normalized bound of Mackey et al. (2014, Corollary 4.2) [B].*

The above corollary is remarkable for the fact that it gives Gaussian-like concentration with only the existence of second moments for the increments. If the increments have conditionally symmetric distributions, one may instead achieve Gaussian-like concentration without existence of any moments, as illustrated by the following example.

**Example 4 (Cauchy increments)** *Let $(\Delta S_t)_{t\in\mathcal{N}}$ be i.i.d. standard Cauchy random variables. Since the distribution of $\Delta S_t$ is symmetric about zero, we earlier proved that $(S_t)$ is sub-Gaussian with variance process $V_t = [S]_t$. Hence our corollary yields for any $m, x \geq 0$,*

$$\mathcal{P}\left(\exists t \in \mathcal{N} : \frac{S_t}{[S]_t + m} \geq x\right) \leq \exp\{-2mx^2\}.$$

The above result is new to the best of our knowledge, and we are not aware of other ways to prove it. For another example, we give a self-normalized bound involving third rather than second moments:

**Corollary 5** *Suppose $\mathcal{T} = \mathcal{N}$ and $(Y_t)_{t\in\mathcal{N}}$ is an $\mathcal{H}^d$-valued martingale with $\mathbb{E}|Y_t|^3 < \infty$ for all $t \in \mathcal{N}$, and let $S_t := \gamma_{\max}(Y_t)$ and $V_t := \gamma_{\max}([Y]_t + \sum_{i=1}^t \mathbb{E}_{i-1}(\Delta Y_i)_-^3)$. Then for any $x, m \geq 0$, we have*

$$\mathcal{P}\left(\exists t \in \mathcal{N} : S_t \geq x + \mathfrak{s}_G\left(\frac{x}{m}\right) \cdot (V_t - m)\right) \leq d\exp\{-m\psi_G^\star\left(\frac{x}{m}\right)\} \leq d\exp\{-\frac{x^2}{2(m + x/6)}\}, \tag{1}$$

*where $\mathfrak{s}_G$ and $\psi_G^\star$ use $c = 1/6$. This is a uniform alternative to Corollary 2.2 of Fan et al. (2015) [B,D].*

Note the exponent in (1) is different from that in Fan et al. (2015), and neither strictly dominates the other. Also note that, unlike the classical Bernstein bound, neither of the above two bounds assume existence of moments of all orders.

# References


Bercu, B. & Touati, A. (2008), 'Exponential inequalities for self-normalized martingales with applications', *The Annals of Applied Probability* **18**(5), 1848–1869.

de la Peña, V. H. (1999), 'A General Class of Exponential Inequalities for Martingales and Ratios', *The Annals of Probability* **27**(1), 537–564.

de la Peña, V. H., Klass, M. J. & Lai, T. L. (2004), 'Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws', *The Annals of Probability* **32**(3), 1902–1933.

de la Peña, V. H., Klass, M. J. & Lai, T. L. (2007), 'Pseudo-maximization and self-normalized processes', *Probability Surveys* **4**, 172–192.

de la Peña, V. H., Lai, T. L. & Shao, Q.-M. (2009), *Self-normalized processes: limit theory and statistical applications*, Springer, Berlin.

Delyon, B. (2009), 'Exponential inequalities for sums of weakly dependent variables', *Electronic Journal of Probability* **14**, 752–779.

Fan, X., Grama, I. & Liu, Q. (2015), 'Exponential inequalities for martingales with applications', *Electronic Journal of Probability* **20**(1), 1–22.

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B. & Tropp, J. A. (2014), 'Matrix concentration inequalities via the method of exchangeable pairs', *The Annals of Probability* **42**(3), 906–945.

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**Proof of Ville's inequality**

Lecturer : Aaditya Ramdas

# 1 Proof of Ville's inequality

**Theorem 1 (Ville (1939))** *For any nonnegative supermartingale $(L_t)$ and any $x > 1$, define the (possibly infinite) stopping time*

$$N := \inf\{t \geq 1 : L_t \geq x\}$$

*and denote the expected overshoot when $L_t$ surpasses $x$ as*

$$o = \mathbb{E}\left[\frac{L_N}{x} | N < \infty\right] \geq 1.$$

*Then,*

$$\Pr(\exists t : L_t \geq x) \leq \frac{\mathbb{E}L_0}{ox} \overset{(i)}{\leq} \frac{\mathbb{E}L_0}{x}.$$

**Proof:** Using the optional stopping theorem and the supermartingale convergence theorem (to establish existance of $L_\infty$), we have the following chain of inequalities:

$$\begin{aligned}
\mathbb{E}L_0 &\overset{(ii)}{\geq} \mathbb{E}L_N \\
&= \mathbb{E}(L_N | N < \infty)P(N < \infty) + \mathbb{E}(L_\infty | N = \infty)P(N = \infty) \\
&\geq \mathbb{E}(L_N | N < \infty)P(N < \infty) \\
&= oxP(N < \infty),
\end{aligned}$$

immediately proving the theorem. ■

For nonnegative martingales, the inequality $(ii)$ is actually an equality. For continuous-time supermartingales with continuous paths, we have $o = 1$, making inequality $(i)$ into an equality. In fact, for continous-time martingales with continuous paths, Ville's inequality holds with equality.

# References

Ville, J. (1939), *Étude Critique de la Notion de Collectif.*, Gauthier-Villars, Paris.

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**Confidence sequences**

Lecturer : Aaditya Ramdas

# 1 Definition of a confidence sequence

It has become standard practice for organizations with online presence to run large-scale randomized experiments, or A/B tests, to improve product performance and user experience. Such experiments are inherently sequential: visitors arrive in a stream and outcomes are typically observed quickly relative to the duration of the test. Results are often monitored continuously using inferential methods that assume a fixed sample, despite the well-known problem that such monitoring can inflate Type I error substantially (Armitage et al. 1969, Berman et al. 2018). Furthermore, most A/B tests are run with little formal planning and very fluid decision-making, as compared with clinical trials or industrial quality control, the traditional applications of sequential analysis.

In this mini, we present methods for deriving *confidence sequences* as a flexible tool for inference in sequential experiments (Darling & Robbins 1967*a*, Lai 1984, Jennison & Turnbull 1989). A confidence sequence is a sequence of confidence sets $(\mathrm{CI}_t)_{t=1}^{\infty}$, typically intervals $\mathrm{CI}_t = (L_t, U_t) \subseteq \mathbb{R}$, satisfying a uniform coverage guarantee: after observing the $t^{\text{th}}$ unit, we calculate an updated confidence set $\mathrm{CI}_t$ for the unknown quantity of interest $\theta_t$, with the coverage property

$$\mathcal{P}(\forall t \geq 1 : \theta_t \in \mathrm{CI}_t) \geq 1 - \alpha. \tag{1}$$

With only a uniform lower bound $(L_t)$ on $\theta_t \in \mathbb{R}$, i.e., if $U_t \equiv \infty$, we have a *lower confidence sequence*. Likewise, if $L_t \equiv -\infty$ we have an *upper confidence sequence* given by the uniform upper bound $(U_t)$. We will build upon the general framework for uniform exponential concentration introduced in the previous mini (Howard et al. 2018), which means our techniques apply to a wide variety of situations: scalar, matrix and Banach-space-valued observations, with possibly unbounded support; self-normalized bounds applicable to observations satisfying very weak moment or symmetry conditions; and continuous-time scalar martingales. Some bounds will yield closed-form confidence sequences, while others give a method for numerical computation of tighter intervals. Both methods allow for flexible control of the "shape" of the confidence sequence, that is, how the sequence of intervals shrink in width over time. As a simple example, given a sequence of observations from a 1-sub-Gaussian distribution whose mean we would like to track, we may choose any $\eta > 1$ and an increasing function $h : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ with $\sum_{k=0}^{\infty} 1/h(k) = 1$, to obtain a confidence sequence of the form

$$\frac{S_t}{t} \pm \frac{\eta^{1/4} + \eta^{-1/4}}{\sqrt{2}} \sqrt{\frac{\log h(\log_\eta t) + \log(2/\alpha)}{t}}. \tag{2}$$

Our theorems will generalize and sharpen related methods from Darling & Robbins (1967*b*, 1968), Jamieson et al. (2014), Kaufmann et al. (2014), Balsubramani (2014), Zhao et al. (2016). Our confidence sequences possess the following properties:

(P1) **Non-asymptotic and nonparametric**: our confidence sequences offer provable coverage for all sample sizes, without exact distributional assumptions or asymptotic approximations.

(P2) **Unbounded sample size**: our methods do not require a final sample size to be chosen ahead of time. They may be tuned for a planned sample size, but always permit additional sampling.

(P3) **Arbitrary stopping rules**: we make no assumptions on the stopping rule used by an experimenter to decide when to end the experiment, or when to act on certain inferences.

These properties give us strong guarantees and broad applicability. An experimenter may always choose to gather more samples, and may stop at any time according to any rule, even one not formally defined, and the resulting inferential guarantees hold under the stated assumptions without any approximations. Of course, this flexibility comes with a cost: our intervals are wider than those that rely on asymptotics, and without assuming a rigid stopping rule, we cannot explicitly correct for selective bias introduced by adaptive stopping. The typical, fixed-sample confidence intervals derived from the central limit theorem do not satisfy any of these properties, and accommodating any one property necessitates wider intervals. It is remarkable that we can accommodate all three and incur a cost of less than doubling the interval width—the discrete mixture bound stays within a factor of two of the fixed-sample central limit theorem bounds over five orders of magnitude in time. Our work gives another example of gaining flexibility and robustness by "doubling" uncertainty estimates, an observation made recently in multiple testing by Katsevich & Ramdas (2018), and a theme more broadly explored by Meng (2018). It may seem that the definition (1) of a confidence sequence is stronger than necessary to achieve these properties, but as we show below, it is equivalent to a definition in terms of arbitrary, unbounded stopping times. It is therefore reasonable to say that any procedure satisfying these three properties will satisfy a guarantee similar to (1).

We will later demonstrate two applications in sequential estimation. First, under a randomization inference model in the Neyman-Rubin potential outcomes framework, we give a tight *empirical variance* confidence sequence for Bernoulli treatment assignment. This method sequentially estimates the variance of the underlying process and uses it to generate a valid confidence sequence, giving a non-asymptotic, sequential analogue of the $t$-test. Such a confidence sequence follows from a general empirical variance confidence sequences for bounded observations. Second, we give asymptotic and non-asymptotic iterated logarithm bounds for the operator norm of a matrix martingale and demonstrate their application to sequential covariance matrix estimation.

**Lemma 1** *Let $(A_t)_{t=1}^{\infty}$ be an adapted sequence of events in some filtered probability space and let $A_\infty := \limsup_{t\to\infty} A_t$. The following are equivalent:*

1. $\mathcal{P}\left(\bigcup_{t=1}^{\infty} A_t\right) \leq \alpha$.
2. $\mathcal{P}(A_T) \leq \alpha$ *for all random times $T$, possibly infinite and not necessarily stopping times.*
3. $\mathcal{P}(A_\tau) \leq \alpha$ *for all stopping times $\tau$, possibly infinite.*

**Proof:** The implication $(a) \implies (b)$ follows from

$$A_T = \left(\bigcup_{t=1}^{\infty} A_t \cap \{T = t\}\right) \cup [A_\infty \cap \{T = \infty\}] \subseteq \bigcup_{t=1}^{\infty} A_t. \tag{3}$$

It is clear that $(b) \implies (c)$. For $(c) \implies (a)$, take $\tau = \inf\{t \in \mathcal{N} : A_t \text{ occurs}\}$, so that $A_\tau = \bigcup_{t=1}^{\infty} A_t$. ∎

# References


Armitage, P., McPherson, C. K. & Rowe, B. C. (1969), 'Repeated Significance Tests on Accumulating Data', *Journal of the Royal Statistical Society. Series A (General)* **132**(2), 235–244.

Balsubramani, A. (2014), 'Sharp Finite-Time Iterated-Logarithm Martingale Concentration', *arXiv:1405.2639 [cs, math, stat]* .

Berman, R., Pekelis, L., Scott, A. & Van den Bulte, C. (2018), p-Hacking and False Discovery in A/B Testing, SSRN Scholarly Paper ID 3204791, Social Science Research Network, Rochester, NY.

Darling, D. A. & Robbins, H. (1967*a*), 'Confidence Sequences for Mean, Variance, and Median', *Proceedings of the National Academy of Sciences* **58**(1), 66–68.

Darling, D. A. & Robbins, H. (1967*b*), 'Iterated Logarithm Inequalities', *Proceedings of the National Academy of Sciences* **57**(5), 1188–1192.

Darling, D. A. & Robbins, H. (1968), 'Some Further Remarks on Inequalities for Sample Sums', *Proceedings of the National Academy of Sciences* **60**(4), 1175–1182.

Howard, S. R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2018), 'Exponential line-crossing inequalities', *arXiv:1808.03204 [math]* .

Jamieson, K., Malloy, M., Nowak, R. & Bubeck, S. (2014), lil' UCB: An Optimal Exploration Algorithm for Multi-Armed Bandits, *in* 'Proceedings of The 27th Conference on Learning Theory', Vol. 35 of *Proceedings of Machine Learning Research*, pp. 423–439.

Jennison, C. & Turnbull, B. W. (1989), 'Interim Analyses: The Repeated Confidence Interval Approach', *Journal of the Royal Statistical Society. Series B (Methodological)* **51**(3), 305–361.

Katsevich, E. & Ramdas, A. (2018), 'Towards "simultaneous selective inference": post-hoc bounds on the false discovery proportion', *arXiv:1803.06790 [math, stat]* .

Kaufmann, E., Cappé, O. & Garivier, A. (2014), 'On the Complexity of Best Arm Identification in Multi-Armed Bandit Models', *arXiv:1407.4443 [cs, stat]* .

Lai, T. L. (1984), 'Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: a sequential approach', *Communications in Statistics - Theory and Methods* **13**(19), 2355–2368.

Meng, X.-L. (2018), 'Double Your Variance, Dirtify Your Bayes, Devour Your Pufferfish, and Draw Your Kidstogram'.

Zhao, S., Zhou, E., Sabharwal, A. & Ermon, S. (2016), Adaptive Concentration Inequalities for Sequential Decision Problems, *in* '30th Conference on Neural Information Processing Systems (NIPS 2016)', Barcelona, Spain.

36-771 Fall 2018

Martingales 1 : Concentration inequalities

# The inadequacy of linear boundaries

Lecturer : Aaditya Ramdas

Given a sequence of observations $(X_t)_{t=1}^{\infty}$, suppose we wish to estimate the average conditional expectation $\mu_t := t^{-1}\sum_{i=1}^{t}\mathbb{E}_{i-1}X_i$ at each time $t$ using the sample mean $t^{-1}\sum_{i=1}^{t}X_i$. Let $S_t = \sum_{i=1}^{t}(X_i - \mathbb{E}_{i-1}X_i)$, the zero-mean deviation of our sample sum from its estimand at time $t$. Suppose we can construct a uniform upper tail bound $u_\alpha(\cdot)$ satisfying

$$\mathcal{P}\big(\exists t \geq 1 : S_t \geq u_\alpha(V_t)\big) \leq \alpha \tag{1}$$

for some *intrinsic time* process $(V_t)_{t=1}^{\infty}$, an appropriate quantity to measure the deviations of $(S_t)$. This uniform upper bound on the centered sum $(S_t)$ yields a lower confidence sequence for $(\mu_t)$ with radius $u_\alpha(V_t)/t$:

$$\mathcal{P}\left(\forall t \geq 1 : \frac{1}{t}\sum_{i=1}^{t}X_i - \frac{u_\alpha(V_t)}{t} \leq \mu_t\right) \geq 1-\alpha. \tag{2}$$

Note that an assumption on the upper tail of $(S_t)$ yields a lower confidence sequence for $(\mu_t)$; a corresponding assumption on the lower tails of $(S_t)$ yields an upper confidence sequence for $(\mu_t)$. In this paper we formally focus on upper tail bounds, from which lower tail bounds can be derived by examining $(-S_t)$ in place of $(S_t)$. In general, the left and right tails of $(S_t)$ may behave differently and require different sets of assumptions, so that our upper and lower confidence sequences may have different forms. Regardless, we can always combine an upper confidence sequence with a lower confidence sequence using a union bound to obtain a two-sided confidence sequence.

Under the typical assumption that the $(X_t)$ are independent with common mean $\mu$, the resulting confidence sequence sequentially estimates $\mu$, but the setup requires neither independence nor a common mean. In general the estimand $\mu_t$ may be changing at each time $t$; we will see an application to causal inference in which this changing estimand makes a great deal of practical sense. In principle, $\mu_t$ may also be random, although none of our applications involve random $\mu_t$.

To construct uniform boundaries $u_\alpha$ satisfying inequality (1), we build upon the following general assumption:

**Assumption 1 (Howard et al. 2018, Assumption 1)** *Let $(S_t)_{t=0}^{\infty}$ and $(V_t)_{t=0}^{\infty}$ be two real-valued processes adapted to an underlying filtration $(\mathcal{F}_t)_{t=0}^{\infty}$ with $S_0 = V_0 = 0$ and $V_t \geq 0$ a.s. for all $t$. Let $\psi$ be a real-valued function with domain $[0, \lambda_{\max})$. We assume, for*

*each* $\lambda \in [0, \lambda_{\max})$*, there exists a supermartingale* $(L_t(\lambda))_{t=0}^{\infty}$ *with respect to* $(\mathcal{F}_t)$ *such that* $\mathbb{E}L_0 := \mathbb{E}L_0(\lambda)$ *is constant for all* $\lambda$*, and such that*

$$\exp\{\lambda S_t - \psi(\lambda) V_t\} \leq L_t(\lambda) \text{ a.s. for all } t.$$

Intuitively, the process $\exp\{\lambda S_t - \psi(\lambda)V_t\}$ measures how quickly $S_t$ has grown relative to intrinsic time $V_t$. Larger values of $\lambda$ exaggerate larger movements in $S_t$, and $\psi$ captures how much we must correspondingly exaggerate $V_t$. It is related to the heavy-tailedness of $S_t$ and the reader may think of it as a cumulant-generating function (CGF). We will organize our presentation of uniform boundaries according to the $\psi$ function used in the above assumption, based on the following definition:

**Definition 1** *Given a function* $\psi : [0, \lambda_{\max}) \to \mathbb{R}$*, we call a function* $u : \mathbb{R}_{\geq 0} \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ *as a* sub-$\psi$ uniform boundary *with crossing probability* $\alpha$ *if the inequality*

$$\mathcal{P}(\exists t \geq 1 : S_t \geq u(V_t, \mathbb{E}L_0)) \leq \alpha \tag{3}$$

*holds whenever* $(S_t)$*,* $(V_t)$ *and* $\psi$ *satisfy Assumption 1.*

For clarity, we will omit the dependence of $u$ on $\mathbb{E}L_0$ from our notation in what follows.

The simplest uniform boundaries are linear: $u(v) = a + bv$ for some $a, b > 0$. As seen below, all such linear boundaries are sub-$\psi$ uniform boundaries. We partially restate this result from Howard et al. (2018) as a lemma:

**Lemma 2 (Howard et al. 2018, Theorem 1)** *For any* $\lambda \in [0, \lambda_{\max})$ *and* $\alpha \in (0, 1)$*, the boundary*

$$u(v) := \frac{\log(\mathbb{E}L_0/\alpha)}{\lambda} + \frac{\psi(\lambda)}{\lambda} \cdot v \tag{4}$$

*is a sub-*$\psi$ *uniform boundary with crossing probability* $\alpha$*.*

Five particular $\psi$ functions play important roles in our development:

- $\psi_B(\lambda) := \log\left(\frac{ge^{h\lambda} + he^{-g\lambda}}{g+h}\right)$, the CGF of a centered random variable with support on just two points $-g$ and $h$ for some $g, h > 0$.
- $\psi_N(\lambda) := \lambda^2/2$, the CGF of a standard Gaussian random variable.
- $\psi_P(\lambda) := c^{-2}(e^{c\lambda} - c\lambda - 1)$ for some scale parameter $c > 0$, which is the CGF of a centered Poisson random variable with rate one when $c = 1$.
- $\psi_E(\lambda) := c^{-2}(-\log(1 - c\lambda) - c\lambda)$ on $\lambda < 1/c$ for some scale parameter $c > 0$, which is the CGF of a centered exponential random variable with rate one when $c = 1$.

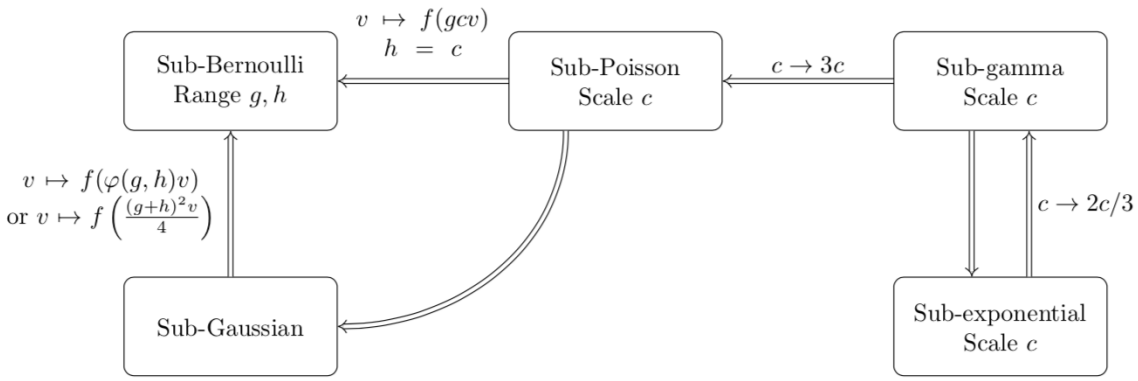


Figure 1: Schematic of relations among sub-$\psi$ boundaries. Each arrow indicates that a sub-$\psi$ boundary at the source node yields a sub-$\psi$ boundary at the destination node with the modification indicated on the arrow. See Proposition 5 for a formal statement.

- $\psi_G(\lambda) := \lambda^2/(2(1-c\lambda))$ on $\lambda < 1/c$ (taking $1/0 = \infty$) for some scale parameter $c \geq 0$; this is not the CGF of a gamma random variable, but is rather a convenient upper bound which also includes the sub-Gaussian case at $c = 0$ and permits analytically tractable results presented below. Our terminology follows that of Boucheron et al. (2013).

When we speak of a *sub-gamma* uniform boundary, we mean that it is sub-$\psi_G$, and likewise for the other cases. The figure summarizes implications that hold among sub-$\psi$ uniform boundaries. It shows, in particular, that a sub-gamma or sub-exponential uniform boundary also yields a sub-Poisson, sub-Gaussian or sub-Bernoulli uniform boundary. Indeed, sub-gamma and sub-exponential uniform bounds are universal in a certain sense:

**Proposition 5** *Suppose $\psi$ is twice continuously differentiable and $\psi(0) = \psi'(0_+) = 0$. Suppose, for each $c > 0$, $u_c(v)$ is a sub-gamma or sub-exponential uniform boundary with crossing probability $\alpha$ for scale $c$. Then $v \mapsto u_{k_1}(k_2 v)$ is a sub-$\psi$ uniform boundary for some constants $k_1, k_2 > 0$.*

While the above lemma provides a versatile building block, the linear growth of the boundary may be undesirable. Indeed, from a concentration point of view, the typical deviations of $S_t$ tend to be only $O(\sqrt{V_t})$ while the aforementioned boundary grows like $O(V_t)$, so the bound will rapidly become loose for large $t$. From a confidence sequence point of view, the confidence radius will be $O(V_t/t)$, and $V_t/t$ typically does not approach zero as $t \uparrow \infty$, so the confidence sequence width will not shrink towards zero. In other words, we cannot achieve arbitrary estimation precision with arbitrarily large samples. We address this problem later, building upon the above lemma to construct *curved* sub-$\psi$ uniform boundaries.

# References

Boucheron, S., Lugosi, G. & Massart, P. (2013), *Concentration inequalities: a nonasymptotic theory of independence*, 1st edn, Oxford University Press, Oxford.

Howard, S. R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2018), 'Exponential line-crossing inequalities', *arXiv:1808.03204 [math]* .

36-771 Fall 2018

Martingales 1 : Concentration inequalities

# Stitching for subGamma/subGaussian boundaries

Lecturer : Aaditya Ramdas

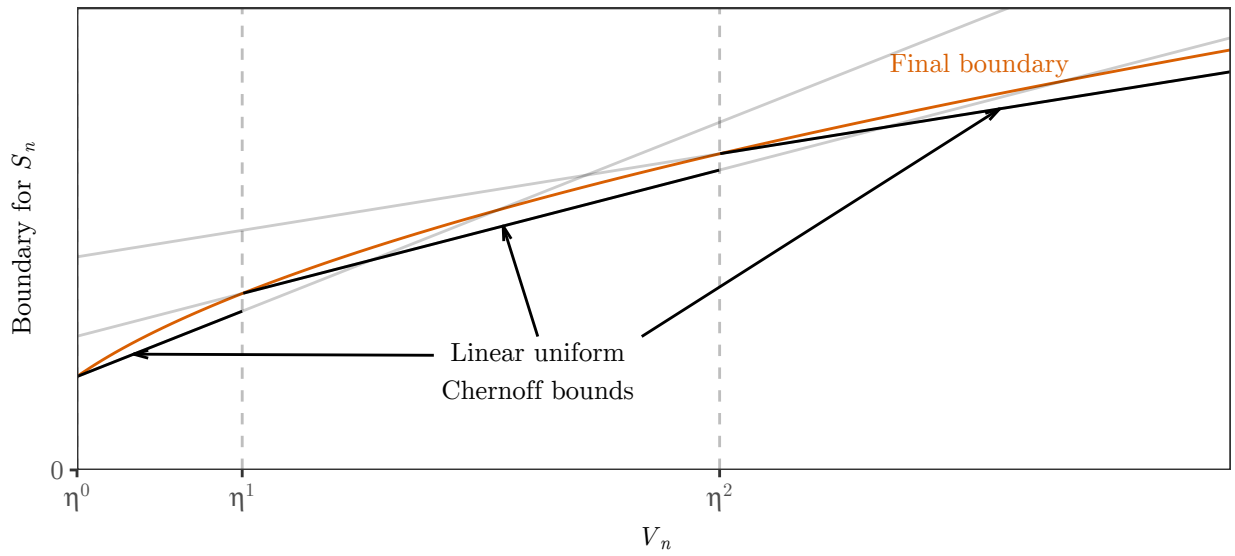


Figure 1: Illustration of stitching together linear boundaries to construct a curved boundary. We break time into geometrically-spaced epochs $\eta^k \leq V_t < e^{k+1}$, construct a linear uniform bound using Lemma 1 (previous lecture) optimized for each epoch, and take a union bound over all crossing events. The final boundary is a smooth analytical upper bound to the piecewise linear bound.

## 1 Analytical bounds: the stitching method

The idea behind Theorem 1is to divide intrinsic time into geometrically spaced epochs, $\eta^k \leq V_t < \eta^{k+1}$ for some $\eta > 1$. We construct a linear boundary within each epoch using Lemma 1 (previous lecture) and take a union bound over crossing events of the different boundaries. The resulting, piecewise-linear boundary may then be upper bounded by a smooth, concave function. Figure 1 illustrates the construction.

The boundary shape is determined by choosing the function $h$ and setting the nominal crossing probability in the $k^{\text{th}}$ epoch to equal $\alpha/h(k)$. Then Theorem 1gives a curved boundary which grows at a rate $\mathcal{O}(\sqrt{V_t \log h(\log_\eta V_t)})$ as $V_t \uparrow \infty$. The more slowly $h(k)$ grows as $k \uparrow \infty$, the more slowly the resulting boundary will grow as $V_t \uparrow \infty$. A simple choice is exponential growth, $h(k) = \eta^{sk}/(1-\eta^{-s})$ for some $s > 1$, yielding $\mathcal{S}_\alpha(v) = \mathcal{O}(\sqrt{v \log v})$.

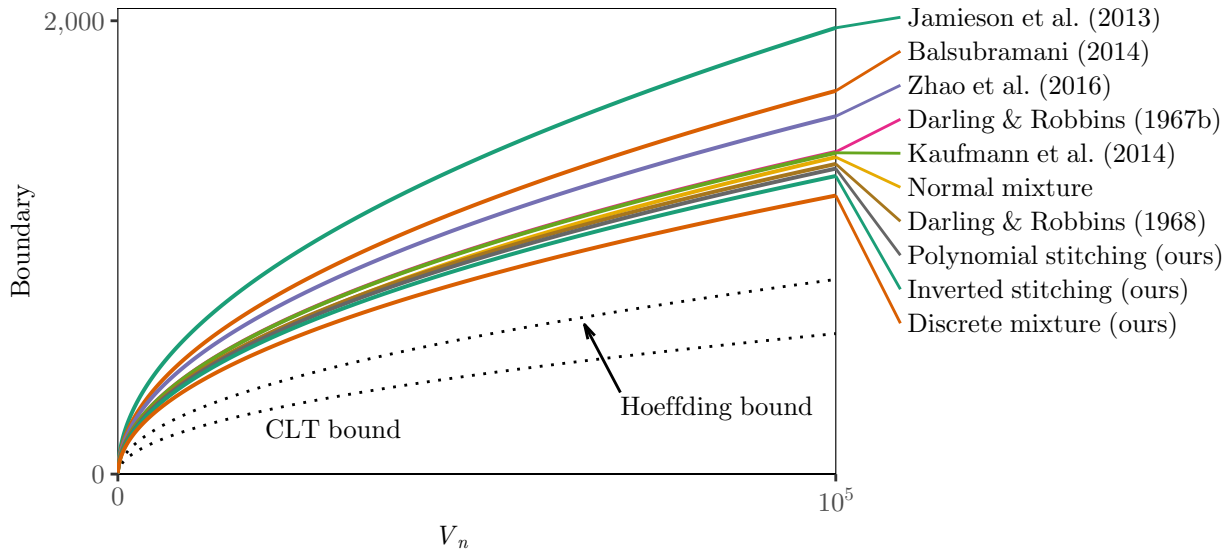


Figure 2: Finite LIL bounds for independent 1-sub-Gaussian observations, $\alpha = 0.025$. The dotted lines show fixed-sample Hoeffding bound $\sqrt{2t \log \alpha^{-1}}$, which is nonasymptotically pointwise valid but not uniformly valid, and the fixed-sample CLT bound $z_{1-\alpha}\sqrt{t}$ which is asymptotically pointwise valid. Polynomial stitching uses Theorem 1with $\eta = 2.04$ and $h(k) = (k+1)^{1.4}\zeta(1.4)$. The inverted stitching boundary is $1.7\sqrt{V_t(\log(1 + \log V_t) + 3.5)}$, using the inverted stitching theorem (later class) with $\eta = 2.99$, $v_{\max} = 10^{20}$, and error rate $0.815\alpha$ to account for finite horizon. Discrete mixture uses the discrete mixture theorem (later class) with $f(\lambda) \propto 1/\lambda \log^{1.4}(1/\lambda), \eta = 1.1$, and $\lambda_{\max} = 4$. The normal mixture bound (later class) uses $\rho = 0.129$. See Howard et al. (2018*b*) for details.

## 1.1 Polynomial stitching and finite LIL bounds

Recall that we used $\zeta(s) = \sum_{k=1}^{\infty} s^{-k}$ to represent the Riemann zeta function. Choosing $h(k) = (k+1)^s\zeta(s)$ for some $s > 1$ in Theorem 1yields $\mathcal{S}_\alpha(v) \sim \sqrt{(2+\delta)v \log\log v)}$, where we may attain any $\delta > 0$ by taking $\eta$ and $s$ sufficiently close to one, coming arbitrarily close to the lower bound furnished by the classical LIL. Uniform bounds achieving this iterated logarithm growth rate are known as *finite LIL bounds*. One may substitute a series converging yet more slowly; for example, $h(k) \propto (k+2)\log^s(k+2)$ for $s > 1$ yields

$$\log h(\log_\eta V_t) = \log\log_\eta(\eta^2 V_t) + s\log\log\log_\eta(\eta^2 V_t) + \log\left(\frac{\log^{1-s}(3/2)}{s-1}\right), \qquad (1)$$

matching related analysis in Darling & Robbins (1967), Robbins & Siegmund (1969), Robbins (1970), and Balsubramani (2014). In practice, the bound (1) appears to behave like bound (**??**) with worse constants. However, the fact that the stitching approach can recover key theoretical results like these gives some indication of its power. Figure 2 compares our

polynomial stitching bound for 1-sub-Gaussian increments to a variety of bounds from the literature; our bound shows a slight improvement. We also include a numerically-computed discrete mixture bound with a mixture distribution roughly corresponding to $h(k) \propto (k+1)^{1.4}$, as described later. This acts as a lower bound and shows that not too much is lost by the approximations involved in the stitching construction.

## 1.2 Why do we get tighter finite LIL bounds than past work?

The idea of taking a union bound over geometrically spaced epochs is standard in the proof of the classical law of the iterated logarithm (Durrett 2017, Theorem 8.5.1). The idea has been extended to finite-time bounds by Darling & Robbins (1967), Jamieson et al. (2014), Kaufmann et al. (2014), and Zhao et al. (2016), usually when the observations are independent and sub-Gaussian. Of course, Theorem 1generalizes these constructions much beyond the independent sub-Gaussian case, but it also achieves tighter constants for the sub-Gaussian setting. Here, we briefly discuss how the improved constants arise.

Both Jamieson et al. (2014) and Zhao et al. (2016) construct a constant boundary rather than a linear increasing boundary over each epoch. They apply Doob's maximal inequality for submartingales (Durrett 2017, Theorem 4.4.2), as in Hoeffding (1963, eq. 2.17), to obtain boundaries similar to that of Freedman (1975). As illustrated in Howard et al. (2018*a*, Figure 2), the linear bounds from Lemma 1 (previous lecture) are stronger than corresponding Freedman-style bounds, and the additional flexibility yields tighter constants.

Both Darling & Robbins (1967) and Kaufmann et al. (2014) use linear boundaries within each epoch analogous to those of Lemma 1 (previous lecture). Both methods share a great deal in common with ours, and Darling & Robbins give consideration to general cumulant-generating functions. Recall from Lemma 1 (previous lecture) that such linear boundaries may be chosen to optimize for some fixed time $V_t = m$. Our method chooses the linear boundary within each epoch to be optimal at the geometric center of the epoch, i.e., at $V_t = \eta^{k+1/2}$, so that at both epoch endpoints the boundary will be equally "loose", that is, equal multiples of $\sqrt{V_t}$. Darling & Robbins choose the boundaries to be tangent at the start of the epoch, hence their boundary is looser than ours at the end of the epoch. Kaufmann et al. choose the boundary as we do, but appear to incur more looseness in the subsequent inequalities used to construct a smooth upper bound.

# References


Balsubramani, A. (2014), 'Sharp Finite-Time Iterated-Logarithm Martingale Concentration', *arXiv:1405.2639 [cs, math, stat]* .

Darling, D. A. & Robbins, H. (1967), 'Iterated Logarithm Inequalities', *Proceedings of the National Academy of Sciences* **57**(5), 1188–1192.

Durrett, R. (2017), *Probability: Theory and Examples*, 5a edn.

Freedman, D. A. (1975), 'On Tail Probabilities for Martingales', *The Annals of Probability* **3**(1), 100–118.

Hoeffding, W. (1963), 'Probability Inequalities for Sums of Bounded Random Variables', *Journal of the American Statistical Association* **58**(301), 13–30.

Howard, S. R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2018*a*), 'Exponential line-crossing inequalities', *arXiv:1808.03204 [math]* .

Howard, S. R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2018*b*), 'Uniform, nonparametric, non-asymptotic confidence sequences', *arXiv:1808.08240 [math]* .

Jamieson, K., Malloy, M., Nowak, R. & Bubeck, S. (2014), lil' UCB: An Optimal Exploration Algorithm for Multi-Armed Bandits, *in* 'Proceedings of The 27th Conference on Learning Theory', Vol. 35 of *Proceedings of Machine Learning Research*, pp. 423–439.

Kaufmann, E., Cappé, O. & Garivier, A. (2014), 'On the Complexity of Best Arm Identification in Multi-Armed Bandit Models', *arXiv:1407.4443 [cs, stat]* .

Robbins, H. (1970), 'Statistical Methods Related to the Law of the Iterated Logarithm', *The Annals of Mathematical Statistics* **41**(5), 1397–1409.

Robbins, H. & Siegmund, D. (1969), 'Probability Distributions Related to the Law of the Iterated Logarithm', *Proceedings of the National Academy of Sciences* **62**(1), 11–13.

Zhao, S., Zhou, E., Sabharwal, A. & Ermon, S. (2016), Adaptive Concentration Inequalities for Sequential Decision Problems, *in* '30th Conference on Neural Information Processing Systems (NIPS 2016)', Barcelona, Spain.

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**Universality of sub-Gamma boundaries**

Lecturer : Aaditya Ramdas

# Universality of sub-Gamma bounds

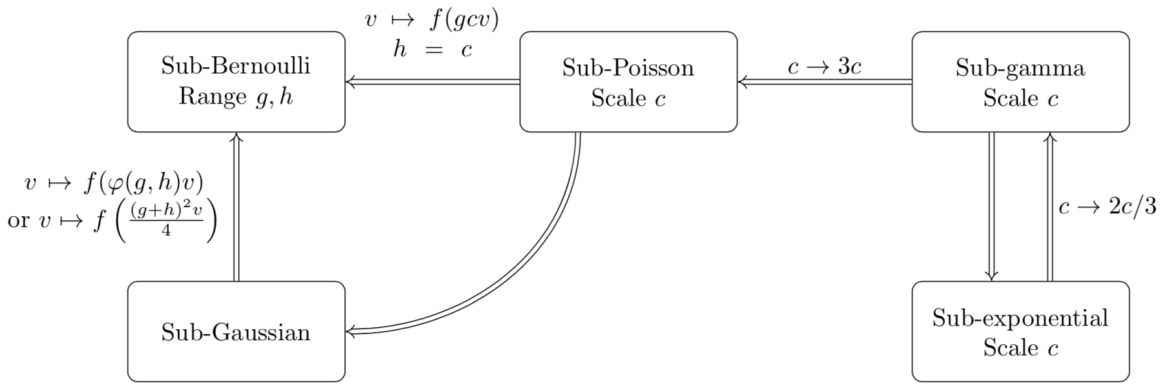


Figure 1: Schematic of relations among sub-$\psi$ boundaries. Each arrow indicates that a sub-$\psi$ boundary at the source node yields a sub-$\psi$ boundary at the destination node with the modification indicated on the arrow.

A reader who is familiar with Howard et al. (2018) will note that the arrows in the above figure are reversed with respect to Figure 3 in their paper. Indeed, since any sub-Bernoulli process is also sub-Gaussian, it follows that any sub-Gaussian uniform boundary is also a sub-Bernoulli uniform boundary, and so on.

The above figure summarizes implications that hold among sub-$\psi$ uniform boundaries. It shows, in particular, that a sub-gamma or sub-exponential uniform boundary also yields a sub-Poisson, sub-Gaussian or sub-Bernoulli uniform boundary. Indeed, sub-gamma and sub-exponential uniform bounds are universal in a certain sense:

**Proposition 1** *Suppose $\psi$ is twice continuously differentiable and $\psi(0) = \psi'(0_+) = 0$. Suppose, for each $c > 0$, $u_c(v)$ is a sub-gamma or sub-exponential uniform boundary with crossing probability $\alpha$ for scale $c$. Then $v \mapsto u_{k_1}(k_2 v)$ is a sub-$\psi$ uniform boundary for some constants $k_1, k_2 > 0$.*

**Proof:** Suppose, for each $c > 0$, $u_c$ is a sub-gamma uniform boundary for scale $c$. Applying Taylor's theorem to $\psi$ at the origin, we have $\psi(x) = \left[\frac{\psi''(0_+)}{2} + h(x)\right] x^2$ where $h(x) \to 0$ as

$x \downarrow 0$. Choose $x_0 > 0$ small enough so that $\psi(x) \leq \psi''(0_+)x^2$ for all $0 \leq x \leq x_0$. Then, setting $c = k_1 := 1/x_0$ in $\psi_G$, and using that fact that $\psi_G \geq \psi_N$, we have $\psi(x) \leq k_2\psi_G(x)$ for all $0 \leq x \leq 1/c$ where $k_2 := 2\psi''(0_+)$. We conclude that, if $(S_t)$ and $(V_t)$ satisfy the canonical Assumption 1 for $\psi$, then $(S_t)$ and $(k_2V_t)$ satisfy Assumption 1 for $\psi_G$. This implies $\mathcal{P}(\exists t \geq 1 : S_t \geq u_{k_1}(k_2V_t)) \leq \alpha$, which is the desired conclusion. The same argument holds if $u_c$ is a sub-exponential uniform boundary, replacing $\psi_G$ with $\psi_E$. ■

The following proposition formalizes the relationships illustrated in the above figure, and follows directly from Proposition 3 of Howard et al. (2018).

**Proposition 2** *Let $u : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a sub-$\psi$ uniform boundary with crossing probability $\alpha$ (we omit the dependence on $\mathbb{E}L_0$, as elsewhere).*

1. *If $u$ is a sub-Gaussian uniform boundary, then $v \mapsto u(\varphi(g,h)v)$ is a sub-Bernoulli uniform boundary with crossing probability $\alpha$ for range parameters $g, h$, where*

$$\varphi(g,h) := \begin{cases} \frac{h^2-g^2}{2\log(h/g)}, & g < h \\ gh, & g \geq h. \end{cases} \tag{3}$$

2. *If $u$ is a sub-Gaussian uniform boundary, then $v \mapsto u((g+h)^2v/4)$ is a sub-Bernoulli uniform boundary with crossing probability $\alpha$ for range parameters $g, h$.*

3. *If $u$ is a sub-Poisson uniform boundary for scale $c$, then $v \mapsto u(gcv)$ is a sub-Bernoulli uniform boundary with crossing probability $\alpha$ for range parameters $g, c$.*

4. *If $u$ is a sub-Poisson uniform boundary for scale $c$, then it is also a sub-Gaussian uniform boundary with crossing probability $\alpha$.*

5. *If $u$ is a sub-gamma uniform boundary for scale $c$, then it is also a sub-Poisson uniform boundary with crossing probability $\alpha$ for scale $3c$.*

6. *If $u$ is a sub-gamma uniform boundary for scale $c$, then it is also a sub-exponential uniform boundary with crossing probability $\alpha$ for scale $c$.*

7. *If $u$ is a sub-exponential uniform boundary for scale $c$, then it is also a sub-gamma uniform boundary with crossing probability $\alpha$ for scale $2c/3$.*

# References

Howard, S. R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2018), 'Exponential line-crossing inequalities', *arXiv:1808.03204 [math]* .

36-771 Fall 2018

Martingales 1 : Concentration inequalities

**Conjugate mixtures**

Lecturer : Aaditya Ramdas

# 1 Conjugate mixtures

Let $f$ be a probability density on $\mathbb{R}$. For appropriate choices of $f$ and $\psi$, the integral $\int \exp\{\lambda S_t - \psi(\lambda)V_t\}f(\lambda)d\lambda$ will be analytically tractable. Since, under Assumption 1, this mixture process is upper bounded by a mixture supermartingale $\int L_t(\lambda)f(\lambda)d\lambda$, such mixtures yield closed-form or efficiently computable curved boundaries. This approach is known as the method of mixtures, one of the most widely-studied techniques for constructing uniform bounds (Ville 1939, Wald 1945, Darling & Robbins 1968, Robbins 1970, Robbins & Siegmund 1969, 1970, Lai 1976). With the exception of the sub-Gaussian case, most prior work on the method of mixtures has focused on parametric settings. We instead derive a variety of nonparametric uniform boundaries using this approach.

Unlike the stitching bound of two lectures ago, which involves a small amount of looseness in the analytical approximations, mixture boundaries are unimprovable in a sense we make precise later. We present both one-sided and two-sided boundaries. Each conjugate mixture boundary includes a tuning parameter $\rho$ which controls the sample size for which the boundary is optimzed. Such tuning is critical in practice, as we explain later.

In the sub-Gaussian case, the following boundary is well-known (Robbins 1970, example 2).

**Proposition 1 (Two-sided normal mixture)** *Suppose $(S_t)$ and $(V_t)$ satisfy Assumption 1 with $\psi = \psi_N$ and $\lambda_{\max} = \infty$, and suppose the same holds for $(-S_t)$. Fix $\alpha \in (0,1)$ and $\rho > 0$, and define*

$$f(v) := \sqrt{(v+\rho)\log\left(\frac{(\mathbb{E}L_0)^2(v+\rho)}{\alpha^2\rho}\right)}. \tag{2}$$

*Then $\mathcal{P}(\forall t \geq 1 : |S_t| < \mathrm{NM}_2(V_t)) \geq 1-\alpha$.*

When only a one-sided sub-Gaussian assumption holds, the normal mixture can still be used to obtain a sub-Gaussian uniform boundary.

When tails are heavier than Gaussian, the normal mixture boundary is not applicable, However, the follow sub-exponential mixture boundary based a gamma mixing density is universally applicable, as described in the previous lecture. Below we make use of the regularized lower incomplete gamma function $\gamma(a,x) := (\int_0^x u^{a-1}e^{-u}\underset{\cdot}{u})/\Gamma(a)$, available in standard statistical software packages.

**Theorem 1 (Gamma-exponential mixture)** *Fix $c > 0, \rho > 0$ and define*

$$\mathrm{GE}_\alpha(v) := \inf\{s \geq 0 : m(s,v) \geq \frac{\mathbb{E}L_0}{\alpha}\}, \tag{3}$$

$$\textit{where} \quad m(s,v) := \frac{\left(\frac{\rho}{c^2}\right)^{\frac{\rho}{c^2}}}{\Gamma\left(\frac{\rho}{c^2}\right)\gamma\left(\frac{\rho}{c^2}, \frac{\rho}{c^2}\right)} \frac{\Gamma\left(\frac{v+\rho}{c^2}\right)\gamma\left(\frac{v+\rho}{c^2}, \frac{cs+v+\rho}{c^2}\right)}{\left(\frac{cs+v+\rho}{c^2}\right)^{\frac{v+\rho}{c^2}}} \exp\{\frac{cs+v}{c^2}\}. \tag{4}$$

*Then $\mathrm{GE}_\alpha$ is a sub-exponential uniform boundary with crossing probability $\alpha$ for scale $c$.*

When a sub-exponential condition applies to $(-S_t)$ as well, we may apply these boundaries to both tails and take a union bound, obtaining a two-sided confidence sequence.

# 2 More examples

The basic idea behind the method of mixtures is as follows. If $(S_t)$, $(V_t)$, and $\psi(\lambda)$ satisfy Assumption 1, and for any probability distribution $F$ on $\mathbb{R}_{\geq 0}$, we have, for all $t$,

$$\int \exp\{\lambda S_t - \psi(\lambda)V_t\}dF(\lambda) \leq \int L_t(\lambda)dF(\lambda), \tag{5}$$

and the right-hand side is a nonnegative supermartingale with initial expectation $\mathbb{E}L_0$. So defining

$$\mathcal{M}_\alpha(v) := \inf\{s \in \mathbb{R} : \int \exp\{\lambda s - \psi(\lambda)v\}dF(\lambda) \geq \frac{\mathbb{E}L_0}{\alpha}\}, \tag{6}$$

and invoking Ville's maximal inequality for nonnegative supermartingales, we have the following basic result:

**Lemma 2** *$\mathcal{M}_\alpha$ is a sub-$\psi$ uniform boundary with crossing probability $\alpha$.*

We suppress the dependence of $\mathcal{M}_\alpha$ on $\psi$, $F$ and $\mathbb{E}L_0$ for notational simplicity, as we did with $\mathcal{S}_\alpha$. With $F$ a point mass at $\lambda$ we recover the linear uniform bounds.

In the sub-Gaussian case, we can take the mixture distribution $F$ to be half-normal over the positive reals. The integral in (6) can be evaluated explicitly, yielding the mixture boundary

$$\mathrm{NM}_\alpha(v) = \inf\{s \in \mathbb{R} : \sqrt{\frac{4\rho}{V_t+\rho}}\exp\{\frac{s^2}{2(V_t+\rho)}\}\Phi\left(\frac{s}{\sqrt{V_t+\rho}}\right) \geq \frac{\mathbb{E}L_0}{\alpha}\}. \tag{7}$$

This is easily evaluated to high precision by numerical root finding. Alternatively, we have the following tight analytical upper bound:

$$\mathrm{NM}(v) \leq \widetilde{\mathrm{NM}}_\alpha(v) := \sqrt{2(v+\rho)\log\left(\frac{\mathbb{E}L_0}{2\alpha}\sqrt{\frac{v+\rho}{\rho}}+1\right)}. \tag{8}$$

**Proposition 9 (One-sided normal mixture)** *For any $\alpha \in (0,1)$ and $\rho > 0$, the boundaries $\mathrm{NM}_\alpha$ and $\widetilde{\mathrm{NM}}_\alpha$ are sub-Gaussian uniform boundaries with crossing probability $\alpha$.*

Many of our definitions and results have focused on one-sided uniform bounds, which yield one-sided (upper or lower) confidence sequences. Such one-sided bounds can always be combined via a union bound to form a two-sided confidence sequence, and for typical values of $\alpha$ used in statistical practice, such a union bound is hardly wasteful, as the intersection of the two error events will have very small probability. In the method of mixtures, however, it is sometimes convenient to derive a two-sided bound directly using a mixture distribution $F$ with support on both positive and negative values of $\lambda$.

In the sub-Bernoulli case, we first rewrite the exponential process $\exp\{\lambda S_t - \psi_B(\lambda)V_t\}$ in terms of the transformed parameter $p = (1+e^{-(g+h)\lambda})^{-1}$. This is motivated by the transform from the canonical parameter to the mean parameter of a Bernoulli family, but keep in mind that we make no parametric assumption here, these are merely analytical manipulations. Then a truncated Beta distribution on $p \in [g/(g+h), 1]$ yields the one-sided Beta-Binomial uniform boundary. Below, $B_x(a,b) = \int_0^x p^{a-1}(1-p)^{b-1}dp$ denotes the incomplete Beta function, whose implementation is readily available in statistical software packages.

**Proposition 10 (One-sided Beta-Binomial mixture)** *Fix any $g, h > 0$, $\alpha \in (0,1)$, and $\rho > gh$, let $m = \rho/gh - 1$ and define*

$$f_{g,h}(v) \;:=\; \inf\{s \geq 0 : m_{g,h}(s,v) \geq \frac{\mathbb{E}L_0}{\alpha}\}, \tag{11}$$

$$\textit{where} \quad m_{g,h}(s,v) \;:=\; \frac{(g+h)^v}{g^{\frac{gv+s}{g+h}} h^{\frac{hv-s}{g+h}}} \frac{B_{h/(g+h)}\left(\frac{h(m+v)-s}{g+h}, \frac{g(m+v)+s}{g+h}\right)}{B_1\{\frac{mg}{g+h}, \frac{mh}{g+h}\}}. \tag{12}$$

*Then $f_{g,h}$ is a sub-Bernoulli uniform boundary with crossing probability $\alpha$ and range $g, h$.*

Sub-Bernoulli conditions typically follow from the assumption that centered observations are $[-g, h]$-bounded. In such a case, the following two-sided bound may be preferable. Simpler versions of this boundary have long been studied i.i.d. Bernoulli sampling (Ville 1939, Robbins 1970, Lai 1976, Shafer et al. 2011).

**Proposition 13 (Two-sided Beta-Binomial mixture)** *Suppose $(S_t)$ and $(V_t)$ satisfy Assumption 1 with $\psi = \psi_B$ for range $g, h$ and $\lambda_{\max} = \infty$, and suppose the same holds for $(-S_t)$ with range $h, g$. Fix any $\rho > gh$, let $m = \rho/gh - 1$ and define*

$$f_{g,h}(v) \;:=\; \inf\{s \geq 0 : m_{g,h}(s,v) \geq \frac{\mathbb{E}L_0}{\alpha}\}, \tag{14}$$

$$\textit{where} \quad m_{g,h}(s,v) \;:=\; \frac{(g+h)^v}{g^{\frac{gv+s}{g+h}} h^{\frac{hv-s}{g+h}}} \frac{B_1\left(\frac{g(m+v)+s}{g+h}, \frac{h(m+v)-s}{g+h}\right)}{B_1\left(\frac{mg}{g+h}, \frac{mh}{g+h}\right)}. \tag{15}$$

*Then $\mathcal{P}(\forall t \geq 1 : -f_{h,g}(V_t) < S_t < f_{g,h}(V_t)) \geq 1 - \alpha$.*

The Beta mixing density is chosen so that the corresponding mixture on $\lambda$ is approximately mean zero with precision $\rho$, making the boundary comparable to a normal mixture bound with the same precision. This allows the user to choose $\rho$ following the same logic as for the normal mixture boundary, as described later, and indeed this is true by construction for all of our conjugate mixture boundaries.

The gamma-exponential mixture is the result of evaluating the mixture integral (6) with mixture density

$$\frac{dF}{d\lambda} = \frac{1}{\gamma(\rho/c^2, \rho/c^2)} \frac{(\rho/c)^{\rho/c^2}}{\Gamma(\rho/c^2)} (c^{-1} - \lambda)^{\rho/c^2 - 1} e^{-\rho(c^{-1}-\lambda)/c}. \tag{16}$$

This is a gamma distribution with shape $\rho/c^2$ and scale $\rho/c$ applied to the transformed parameter $u = c^{-1} - \lambda$, truncated to the support $[0, c^{-1}]$. The distribution has mean zero and variance equal to $1/\rho$, making it comparable to the normal mixture distribution used above. As $\rho \to \infty$, the gamma mixture distribution converges to a normal distribution and concentrates about $\lambda = 0$, the regime in which $\psi_E(\lambda) \sim \psi_N(\lambda)$, which gives some intuition for why the gamma mixture recovers the normal mixture when $\rho \gg c^2$. Like the normal mixture, the gamma mixture is unimprovable and is effective in practice.

A similar mixture boundary holds in the sub-Poisson case:

**Proposition 17 (Gamma-Poisson mixture)** *Fix $c > 0, \rho > 0$ and define*

$$\mathrm{GP}_\alpha(v) := \inf\{s \geq 0 : m(s, v) \geq \frac{\mathbb{E}L_0}{\alpha}\}, \tag{18}$$

$$\textit{where} \quad m(s, v) := \frac{\left(\frac{\rho}{c^2}\right)^{\frac{\rho}{c^2}}}{\Gamma\left(\frac{\rho}{c^2}\right) \gamma\left(\frac{\rho}{c^2}, \frac{\rho}{c^2}\right)} \frac{\Gamma\left(\frac{cs+v+\rho}{c^2}\right) \gamma\left(\frac{cs+v+\rho}{c^2}, \frac{v+\rho}{c^2}\right)}{\left(\frac{v+\rho}{c^2}\right)^{\frac{cs+v+\rho}{c^2}}} \exp\{\frac{v}{c^2}\}. \tag{19}$$

*Then $\mathrm{GP}_\alpha$ is a sub-Poisson uniform boundary with crossing probability $\alpha$ for scale $c$.*

# References


Darling, D. A. & Robbins, H. (1968), 'Some Further Remarks on Inequalities for Sample Sums', *Proceedings of the National Academy of Sciences* **60**(4), 1175–1182.

Lai, T. L. (1976), 'On Confidence Sequences', *The Annals of Statistics* **4**(2), 265–280.

Robbins, H. (1970), 'Statistical Methods Related to the Law of the Iterated Logarithm', *The Annals of Mathematical Statistics* **41**(5), 1397–1409.

Robbins, H. & Siegmund, D. (1969), 'Probability Distributions Related to the Law of the Iterated Logarithm', *Proceedings of the National Academy of Sciences* **62**(1), 11–13.

Robbins, H. & Siegmund, D. (1970), 'Boundary Crossing Probabilities for the Wiener Process and Sample Sums', *The Annals of Mathematical Statistics* **41**(5), 1410–1429.

Shafer, G., Shen, A., Vereshchagin, N. & Vovk, V. (2011), 'Test Martingales, Bayes Factors and p-Values', *Statistical Science* **26**(1), 84–101.

Ville, J. (1939), *Étude Critique de la Notion de Collectif.*, Gauthier-Villars, Paris.

Wald, A. (1945), 'Sequential Tests of Statistical Hypotheses', *Annals of Mathematical Statistics* **16**(2), 117–186.

36-771 Fall 2018

Martingales 2 : Sequential Analysis

**Discrete mixtures and inverted stitching**

Lecturer : Aaditya Ramdas

# 1 Numerical bounds using discrete mixtures.

In applied use, there is often no need for an explicit closed-form expression so long as the bound can be easily computed numerically. Our discrete mixture method gives a straightforward and efficient technique for numerical computation of curved boundaries whenever Assumption 1 is satisfied. It permits arbitrary mixture densities and thus can produce boundaries growing at the asymptotically-optimal $\mathcal{O}(V_t \log\log V_t)$ rate.

Recall that the shape of the stitching bound was determined by the user-specified function $h$. For the discrete mixture bound, one instead specifies a distribution $F$. We then discretize $F$ using a series of support points $\lambda_k$, geometrically spaced according to successive powers of some $\eta > 1$, and an associated set of weights $w_k$:

$$\lambda_k := \frac{\lambda_{\max}}{\eta^{k+1/2}} \quad \text{and} \quad w_k := \frac{\lambda_{\max}(\eta-1) f(\lambda_k \sqrt{\eta})}{\eta^{k+1}} \quad \text{for} \quad k = 1, 2, \ldots. \tag{1}$$

With the above definitions in place, we have a discrete mixture bound as follows.

**Theorem 1 (Discrete mixture bound)** *Fix $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ and $\alpha \in (0, 1)$. Employing any continuous distribution $F$ with density $f$ that is nonincreasing and positive on a nonempty interval $(0, \lambda_{\max}]$, if we define*

$$\widetilde{\mathcal{M}}_\alpha(v) := \inf\{s \in \mathbb{R} : \sum_{k=0}^{\infty} w_k \exp\{\lambda_k s - \psi(\lambda_k) v\} \geq \frac{\mathbb{E}L_0}{\alpha}\}, \tag{2}$$

*then $\widetilde{\mathcal{M}}_\alpha$ is a sub-$\psi$ uniform boundary with crossing probability $\alpha$.*

We suppress the dependence of $\widetilde{\mathcal{M}}_\alpha$ on $F$, $\mathbb{E}L_0$, $\lambda_{\max}$ and $\eta$ for notational simplicity. Though the above theorem is a straightforward consequence of the method of mixtures, our choice of discretization makes it effective, broadly applicable, and easy to compute numerically.

To see heuristically why the exponentially-spaced grid $\lambda_k = \mathcal{O}(\eta^{-k})$ makes sense, observe that the integrand $\exp\{\lambda s - \lambda^2 v/2\}$ is a scaled normal density in $\lambda$ with mean $s/v$ and standard deviation $1/\sqrt{v}$. In the regime relevant to our curved boundaries, $s$ is of order $\sqrt{v}$, ignoring logarithmic factors. Hence the integrand at time $v$ has both center and spread of order $1/\sqrt{v}$, so as $v \to \infty$, the relevant scale of the integrand shrinks. With the grid

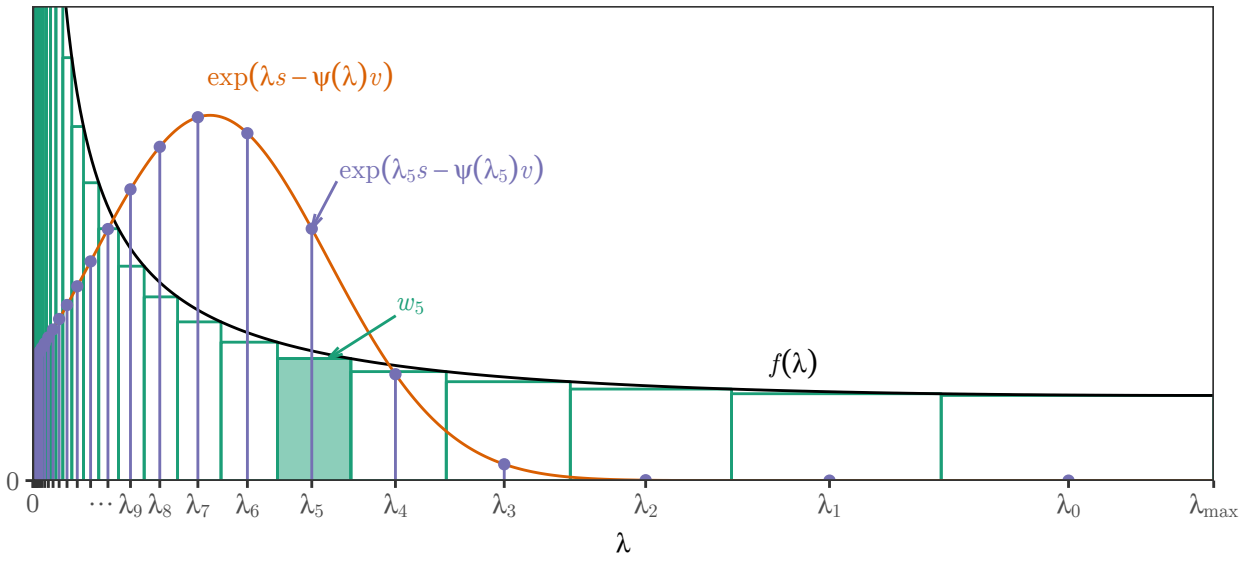


Figure 1: Illustration of the discrete mixture method. Mixture density $f(\lambda)$ is discretized on a grid $(\lambda_k)_{k=0}^{\infty}$ which gets finer as $\lambda \downarrow 0$. Resulting discrete mixture weights are represented by areas within green bars. Integrand $\exp\{\lambda s - \psi(\lambda)v\}$ is evaluated at grid points $\lambda_k$, illustrated by purple points. Multiplying one integrand evaluation $\exp\{\lambda_k s - \psi(\lambda_k)v\}$ by the corresponding weight $w_k$ gives one term of the sum (2).

$\lambda_k = \mathcal{O}(\eta^{-k})$ we have $\lambda_k - \lambda_{k+1} = \mathcal{O}(\lambda_k)$, ensuring that the resolution of the grid around the peak of the integrand matches the scale of the integrand as $v \to \infty$.

The choice of $\lambda_{\max}$ depends on the minimum value of $V_t$ relevant to inference: making $\lambda_{\max}$ larger will make the resulting bound tighter over smaller values of $V_t$ at the cost of a looser bound for all larger values of $V_t$. In practice, for $\psi = \psi_G$, setting $\lambda_{\max} = [c + \sqrt{v_{\min}/2 \log \alpha^{-1}}]^{-1}$ will ensure the bound is tight for $V_t \geq v_{\min}$. Furthermore, in practice the sum can be truncated after $k_{\max} = \lceil \log_\eta(\lambda_{\max}[c + \sqrt{5v/\log \alpha^{-1}}]) \rceil$ terms.

To illustrate the accuracy of the discrete mixture, we compare it to the one-sided normal mixture bound. By using the same half-normal mixing density from last class, and setting $\eta = 1.05$, $\lambda_{\max} = 100$, we may evaluate a corresponding discrete mixture bound $\widetilde{\mathcal{M}}_\alpha$. With $\rho = 14.3$, $\alpha = 0.05$ and $\mathbb{E}L_0 = 1$, numerical calculations show that

$$\sup_{1 \leq t \leq 10^6} \frac{\widetilde{\mathcal{M}}_\alpha(t)}{\mathrm{NM}_\alpha(t)} \leq 1.004, \tag{3}$$

suggesting that the discrete mixture theorem gives an excellent conservative approximation to the corresponding continuous mixture boundary to over a large practical range. Of course, when a closed form is available as in the normal mixture, one should use it in practice. But an exact closed form integral is rarely available as it is here, and substantial looseness often accompanies closed-form approximations which provably maintain crossing probability

guarantees. In such cases, unless a closed form is required, the discrete mixture method is preferable.

## 2 Inverted stitching for arbitrary boundaries.

In the discrete mixture method, we choose a mixture distribution $F$ and the machinery yields a boundary $\widetilde{\mathcal{M}}_\alpha$. Likewise, in the stitching construction from a few lectures ago, we choose an error decay function $h$ and the machinery yields a boundary $\mathcal{S}_\alpha$. In this section we invert the procedure: we choose a boundary function $g(v)$ and numerically compute an upper bound on its $S_t$-upcrossing probability using a stitching-like construction. For simplicity we restrict to the sub-Gaussian case; we are currently working on extending this idea beyond sub-Gaussianity.

**Theorem 2** *For any nonnegative, strictly concave function $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $v_{\max} > 1$, the function*

$$u(v) := \begin{cases} g(1 \vee v), & v \leq v_{\max}, \\ \infty, & \textit{otherwise} \end{cases} \tag{4}$$

*is a sub-Gaussian uniform boundary with crossing probability*

$$(\mathbb{E}L_0) \inf_{\eta > 1} \sum_{k=0}^{\lceil \log_\eta v_{\max} \rceil} \exp\{-\frac{2(g(\eta^{k+1}) - g(\eta^k))(\eta g(\eta^k) - g(\eta^{k+1}))}{\eta^k(\eta-1)^2}\}. \tag{5}$$

The proof follows a straightforward idea. We break time into epochs $\eta^k \leq V_t < \eta^{k+1}$. Within each epoch we consider the linear boundary passing through the points $(\eta^k, g(\eta^k))$ and $(\eta^{k+1}, g(\eta^{k+1}))$. This line lies below $g(V_t)$ throughout the epoch, and its crossing probability is determined by its slope and intercept as in the mother theorem. Taking a union bound over epochs yields the result.

A similar idea was considered by Darling & Robbins (1968), using a mixture integral approximation instead of an epoch-based construction to derive closed-form bounds. Inverted stitching requires numerical summation but yields tighter bounds with fewer assumptions. As an example, the above theorem with $\eta = 2.99$ shows that

$$\mathcal{P}\left(\exists t : 1 \leq V_t \leq 10^{20} \text{ and } S_t \geq 1.7\sqrt{V_t(\log\log(eV_t) + 3.46)}\right) \leq 0.025. \tag{6}$$

## References

Darling, D. A. & Robbins, H. (1968), 'Some Further Remarks on Inequalities for Sample Sums', *Proceedings of the National Academy of Sciences* **60**(4), 1175–1182.

36-771 Fall 2018

Martingales 2 : Sequential Analysis

# Sequential testing, always valid p-values

Lecturer : Aaditya Ramdas

We have organized our presentation around confidence sequences and their closely related uniform concentration bounds. We have emphasized confidence sequences due to our belief that they offer a useful "user interface" for sequential inference. However, our methods may alternatively be viewed as sequential hypothesis tests or always-valid p-values processes (Johari et al. 2015). Indeed, a slew of related definitions from the literature are equivalent or dual to one another. Here we briefly discuss these connections, building upon the definitions and dualities of Johari et al. (2015). Recall Lemma 1 from lecture 15, which gives equivalent formulations of certain common definitions in sequential testing.

First, let us mention that our definition of confidence sequence based on Darling & Robbins (1967*a*) and Lai (1984), differs from that Johari et al. (2015), who require that $\mathcal{P}(\theta_\tau \in \mathrm{CI}_\tau) \geq 1-\alpha$ for all stopping times $\tau$. They allow $\tau = \infty$ by defining $\mathrm{CI}_\infty := \liminf_{t\to\infty} \mathrm{CI}_t$. By taking $A_t := \{\theta_t \notin \mathrm{CI}_t\}$ in Lemma 1 from lecture 14, we see that the distinction is immaterial, and furthermore that we could equivalently define confidence sequences in terms of arbitrary random times, not necessarily stopping times. This generalizes Proposition 1 of Zhao et al. (2016).

As an alternative to confidence sequences, Johari et al. (2015) define an *always-valid p-value process* for some null hypothesis $H_0$ as an adapted, $[0,1]$-valued sequence $(p_t)_{t=1}^\infty$ satisfying $\mathcal{P}_0(p_\tau \leq \alpha) \leq \alpha$ for all stopping times $\tau$, where $\mathcal{P}_0$ denotes probability under the null $H_0$. Taking $A_t := \{p_t \leq \alpha\}$ in Lemma 1 from lecture 14 shows that we may replace this definition with an equivalent one over all random times, not necessarily stopping times, or with the uniform condition $\mathcal{P}_0(\exists t \in \mathcal{N} : p_t \leq \alpha) \leq \alpha$. By analogy to the usual dual construction between fixed-sample p-values and confidence intervals[1], one can see that confidence sequences are dual to always-valid p-values, and both are dual to sequential hypothesis tests, as defined by a stopping time and a binary random variable indicating rejection (Johari et al. 2015, Proposition 5). In particular, for the null $H_0 : \theta = \theta^\star$, if $(\mathrm{CI}_t)$ is a $(1-\alpha)$-confidence sequence for $\theta$, it is clear that a test which stops and rejects the null as soon as $\theta^\star \notin \mathrm{CI}_t$ controls type I error: $\mathcal{P}_0(\text{reject } H_0) = \mathcal{P}_0(\exists t \in \mathcal{N} : \theta^\star \notin \mathrm{CI}_t) \leq \alpha$. Typically, then, a confidence sequence based on any of the curved uniform bounds in this paper with radius $u(v) = o(v)$ will yield a *test of power one* (Darling & Robbins 1967*b*, Robbins 1970). In particular, for a confidence sequence with limits $\bar{X}_t \pm u(V_t)$, it is sufficient that $\bar{X}_t$ converges a.s. to $\theta$

[1]Indeed, if $(\mathrm{CI}_t^\alpha)$ is a $(1-\alpha)$-level confidence sequence for some constant parameter $\theta$, for each $\alpha \in (0,1)$, then $p_t := \inf\{\alpha \in (0,1) : \theta^\star \notin \mathrm{CI}_t^\alpha\}$ gives an always-valid p-value process for the null hypothesis $H_0 : \theta = \theta^\star$. Conversely, if $(p_t^{\theta^\star})$ is an always-valid p-value process for the null hypothesis $H_0 : \theta = \theta^\star$, for each $\theta^\star$ in some domain $\Theta$, then $\mathrm{CI}_t := \{\theta^\star \in \Theta : p_t^{\theta^\star} > \alpha\}$ gives a $(1-\alpha)$-level confidence sequence for $\theta$.

and $\limsup_{t\to\infty} V_t/t < \infty$ a.s., conditions that will typically hold. These conditions imply that the radius of the confidence sequence, $u(V_t)/t$, approaches zero, while the center $\bar{X}_t$ is eventually bounded away from $\theta^\star$ whenever $\theta \neq \theta^\star$, so that the confidence sequence will eventually exclude $\theta^\star$ with probability one.

In the one-parameter exponential family case, the exponential process $\exp\{\lambda S_t(\mu) - t\psi_\mu(t)\}$ is exactly the likelihood ratio for testing $H_0 : \theta = \theta(\mu)$ against $H_1 : \theta = \theta(\mu) + \lambda$. When using a mixture uniform boundary, a sequential test which rejects as soon as the confidence sequence excludes $\mu^\star$ can be seen as equivalently rejecting as soon as either of the mixture likelihood ratios $\int \exp\{\lambda S_t - \psi_{\mu^\star}(\lambda)t\}\underset{\cdot}{\mathrm{F}}(\lambda)$ or $\int \exp\{-\lambda S_t - \psi_{\mu^\star}(-\lambda)t\}\underset{\cdot}{\mathrm{F}}(\lambda)$ exceeds $2/\alpha$. Thus a sequential hypothesis test built upon a mixture-based confidence sequence is equivalent to a mixture sequential probability ratio test (Robbins 1970) in the parametric setting. As we have discussed, stitching bounds can also be viewed as approximations to certain mixture bounds, so that hypothesis tests based on stitching bounds are also approximations to mixture SPRTs. Importantly, the confidence sequences defined in this paper are natural nonparametric generalizations of the mixture SPRT, recovering various mixture SPRTs in the parametric cases.

# References


Darling, D. A. & Robbins, H. (1967*a*), 'Confidence Sequences for Mean, Variance, and Median', *Proceedings of the National Academy of Sciences* **58**(1), 66–68.

Darling, D. A. & Robbins, H. (1967*b*), 'Iterated Logarithm Inequalities', *Proceedings of the National Academy of Sciences* **57**(5), 1188–1192.

Johari, R., Pekelis, L. & Walsh, D. J. (2015), 'Always valid inference: Bringing sequential analysis to A/B testing', *arXiv preprint arXiv:1512.04922* .

Lai, T. L. (1984), 'Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: a sequential approach', *Communications in Statistics - Theory and Methods* **13**(19), 2355–2368.

Robbins, H. (1970), 'Statistical Methods Related to the Law of the Iterated Logarithm', *The Annals of Mathematical Statistics* **41**(5), 1397–1409.

Zhao, S., Zhou, E., Sabharwal, A. & Ermon, S. (2016), Adaptive Concentration Inequalities for Sequential Decision Problems, *in* '30th Conference on Neural Information Processing Systems (NIPS 2016)', Barcelona, Spain.

# 36-771 : Martingales 1 (Concentration inequalities)
# 36-772 : Martingales 2 (Sequential analysis)

2 minis (Aug 28 to Oct 18, Oct 23 to Dec 6), 6 credits each
Fall 2018, Syllabus

August 25, 2018

## 1 Basic Course Information

**Instructor** Aaditya Ramdas, `aramdas@stat.cmu.edu` [Office hours: on demand]
**Teaching Assistants**: there are no TAs for this course.

**Time:** Tue, Thu 10:30-11:50am
**Location**: Wean 4625
**Exceptions**: There will likely be no class on Sep 13 (Thu), Nov 1 (Thu) due to instructor travel, Oct 16 due to university rules, and on Nov 20 (Tue), Nov 22 (Thu) due to Thanksgiving. Hence there will be 14 lectures in Mini 1, and 11 lectures in Mini 2 (Mini 1 is longer than Mini 2 by design). See the CMU academic calendar.

**Website** See http://www.stat.cmu.edu/∼aramdas/martingales18 for basic course material.
**Announcements** All announcements will be made on Canvas.

**Participants** This course is intended for advanced PhD students with strong mathematical background.
**Prerequisites** The first mini is a prerequisite for the second mini. There are no formal prerequisites for the first mini, but non-PhD students must email the instructor for permission to enroll in the course. Students are expected to have completed at least one intermediate statistics course (like 10-705 at CMU), and preferably an advanced statistics course. Students must be familiar with

1. basic concentration inequalities (such as Markov, Chebysheff, Hoeffding)
2. basics of martingales (filtrations, stopping times, Doob's maximal inequalities and optional stopping theorems)
3. basics of testing and estimation (probability ratio tests, risk, type-1 error, power)
4. basics of convex analysis (strict convexity, Legendre-Fenchel transform)

**Textbook** There is no single textbook we will follow. Some related textbooks include:

- Large deviations techniques and applications, by Dembo and Zeitouni
- Stopped random walks: limit theorems and applications, by Gut
- Self-normalized processes, by de la Peña, Lai and Shao
- Concentration inequalities for sums and martingales, by Bercu, Delyon and Rio
- Concentration inequalities: a nonasymptotic theory of independence, by Boucheron, Lugosi and Massart

The presentation in Mini 1 will broadly follow a recent paper: Exponential line-crossing inequalities. Other relevant papers will be distributed as necessary.

## 2 Course Description

**36-771: Martingales 1 (Concentration inequalities)** Martingales are a central topic in statistics, but are even more relevant today due to modern applications to sequential learning and decision making problems. The first mini will present a unified derivation of a wide-variety of new and old concentration inequalities for martingales. We will prove inequalities for scalars and matrices, that hold under a wide variety of nonparametric assumptions.

Note to those who have already taken Advanced Probability or Advanced Statistics, you will also learn about (a) self-normalized exponential concentration inequalities for heavy-tailed distributions like those with only two moments, and even for those with no moments (like the Cauchy), (b) concentration of continuous time processes like Brownian motions (and possibly, Poisson processes, Levy processes), (c) concentration of martingales in smooth Banach spaces (useful for vector and matrix concentration).

**36-772: Martingales 2 (Sequential analysis)** The second mini will focus on deriving guarantees for a variety of important problems in sequential analysis using the tools developed in the first mini, as well as new tools such as uniform nonasymptotic versions of the law of the iterated logarithm for scalars and matrices. Applications include sequential analogs of the t-test that are valid without a Gaussian assumption, best-arm identification in multi-armed bandits, average treatment effect estimation in sequential clinical trials, sequential covariance matrix estimation, and other such problems.

## 3 Graded Components

**Mini 1** The grade will be based on one long homework (40%) whose questions will be progressively released, and a course project (10% for proposal, 50% for final report). Homework 1 will be due on Oct 5 (Fri). The project proposal (approximately one page) is due on Sep 13 (Thu), and the final report (approximately five pages) due on Oct 16 (Tue, when class is cancelled due to university policy). The course summary, reflection and open problems discussion will occur on the last day of the mini, Oct 18.

**Mini 2** 40% of the grade will be based on one long homework whose questions will be progressively released. For the rest, students will preferably develop a significant extension of the course project in Mini 1 (50% for final report, 10% for a short in-class presentation). Homework 2 will be due on Nov 15 (before Thanksgiving week). The in-class presentations will occur on Dec 4 (Tue). The project final report (approximately ten pages) will be due on Dec 5 (Wed). The course summary, reflection and open problems discussion will occur on the last day of classes, Dec 6.

**Projects** There are a wide variety of options available for course projects. Examples include:

- (Low risk) You can survey an area of the literature (covered in a textbook, or a set of advanced papers) that is related to the course, and is complementary to what is covered in class.
- (Medium-low risk) You can create a set of graphs, plots, or interactive figures, which allow the user to visualize several of the concentration inequalities and/or applications covered in the course. For inspiration, check out distill.pub, and specifically, a paper on why momentum works.
- (Medium-high risk) You can apply the contents of the class to your own research problem, for example by improving the guarantees you had achieved by using older tools, or by extending the analysis of your problem to hold for new settings.
- (High risk) If you are mathematically very mature, and want to work on a new research problem in this area from scratch, talk to the instructor privately in person.

Other ideas for course projects are also welcome. Grades will ultimately be awarded based on the instructor's judgment of the amount of work completed in the project, with some amount of subjective discounting for the risk of the project taken up.

# 4 Learning Objectives

**Mini 1** Upon successful completion of the first mini, the student will be able to

- Identify mathematical expressions that correspond to typical martingales and supermartingales.
- Quote Ville's maximal inequality for supermartingales, and explain it using a figure.
- Verify that "the central assumption" of the course is satisfied under various nonparametric conditions.
- State and apply "the mother of all exponential concentration inequalities", the main theorem of the course.
- Map an "exponential line-crossing inequality" onto a traditional concentration inequality.
- Assess which of two concentration inequalities is more general, using the A-B-C-D-E method.

**Mini 2** Upon successful completion of the second mini, the student will be able to

- Describe the difference between a confidence interval and a "confidence sequence".
- Explain how to derive an "exponential curve-crossing inequality" from a line-crossing inequality.
- Translate a nonasymptotic law-of-iterated-logarithm concentration inequality into a confidence sequence.
- Derive variants of the sequential probability ratio test using the supermartingale technique.
- Assess the pros and cons of the standard t-test, compared to the LIL-t-test.

# 5 Approximate Schedule (will change adaptively)

**Mini 1** (14 lectures)

- Intro: Concentration of measure, Doob's & Ville's inequalities, optional stopping, examples (1-2 lectures).
- The central assumption and sufficient conditions: sub-Gaussian, sub-exponential, sub-Poisson, etc (2 lectures).
- The main theorem: four equivalent statements (2 lectures).
- Generalizations of Hoeffding, Bennett, Bernstein, Freedman, etc (2 lectures).
- Heavy-tails, and self-normalized inequalities (2 lectures).
- Strengthening standard matrix concentration inequalities (2 lectures).
- Continuous-time processes and/or Banach-space concentration (1-2 lectures).
- Summary, reflection and open directions (1 lecture).

**Mini 2** (11 lectures)

- Intro: uniform, nonparametric, nonasymptotic confidence sequences, inadequacy of line-crossing (2 lectures).
- Stitching & mixing : from exponential line- to curve-crossing inequalities (2 lectures).
- Inverted stitching : numerically estimating curve-crossing probabilities (1 lecture).
- Application 1: a nonasymptotic, nonparametric, uniform sequential t-test (1 lecture).
- Application 2: multi-armed bandits, best-arm identification, adaptive hypothesis testing (1 lecture).
- Application 3: sequential covariance matrix estimation (1 lecture).
- Application 4: sequential average treatement effect estimation (1 lecture).
- Project presentations (1 lectures).
- Summary, reflection and open directions (1 lecture).

# 6 Course policies

## 6.1 Attendance

It is not compulsory, but recommended and expected. Every research study on this topic that I have read concludes that academic performance is negatively affected by not showing up to class. To encourage attendance, subtle hints for exam questions will be dropped from time to time.

## 6.2 Collaboration

Discussion of class material is heavily encouraged. Additionally,

- After submission of a homework, discussion of answers is encouraged.
- Before submission of a homework, reasonable verbal discussion of homeworks is allowed. An example of unreasonable verbal discussion: one person reciting formulae orally while another one writes them down. Written discussion (in any form) is permitted in groups smaller than 3 (or in rare exceptions 4) students.
- No matter what discussions have taken place, every homework and cheat sheet and mini-project and self-test (in its entirety) must be written up or coded up alone.

## 6.3 Academic Integrity

I have a zero tolerance policy for violation of class policies. If you are in any doubt whether a form of collaboration or obtaining solutions is permitted, please clarify it with me before proceeding.

- For each question on each homework, collaborators for that question must be acknowledged. Copying solutions from the internet is explicitly disallowed. You may search for material to help you understand a concept better, but be sure to create your own final solution. If you happen to use results from Wikipedia or textbooks, you must cite the source and are expected to completely understand the result you are citing. However, it is disallowed to copy solutions to exercises from elsewhere on the internet, like other courses or papers. When quoting text from a textbook, paper or website, use the `\begin{quote}` option in Latex.
- Any deviation from the rules will be dealt with according to the severity of the case. For example: evidence of written discussion in a larger group than 3-4 will result in points earned for that question becoming zero for all those relevant students; blindly copying one solution from someone else or online will result in the maximum points that can be earned for that homework becoming zero (maximum eligible grade becomes B); repeat occurrences will result in a failing grade for the course.
- In line with university policy, all instances of cheating/plagiarism will be reported to your academic advisor and the dean of student affairs. See the university policy on academic integrity.

## 6.4 Use of Mobile Devices and Laptops in Class

These are allowed but not encouraged. Learning research shows that unexpected noises or movement automatically divert and capture people's attention, meaning that you are affecting everyone's learning experience. For this reason, I ask you turn off your mobile devices and close your laptops during class. If you must use your laptop or mobile, make sure you are sitting at the back of the class.

## 6.5 Late Assignments

Every student is allowed a total of 2 late days per mini. Beyond that, the maximum earnable points for that assignment will drop by 20% per day.

# 7 Additional information

## 7.1 Accommodations for Students with Disabilities

If you have a disability and are registered with the Office of Disability Resources, I encourage you to use their online system to notify me of your accommodations and discuss your needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

## 7.2 Statement of Support for Students' Health & Well-being

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit http://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

If you or someone you know is feeling suicidal or in danger of self-harm, call someone immediately, day or night (CaPS: 412-268-2922, Resolve Crisis Network: 888-796-8226). If the situation is life threatening, call the police (On-campus CMU Police: 412-268-2323, Off-campus Police: 911).