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 ψ reminiscent of a cumulant generating function (CGF). In the i.i.d. scalar setting above, we take $V_t=t$ and let ψ 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 ψ 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 λ, 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, ψ 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, ψ will often still be a CGF of some distribution, though this is not required. However, our most interesting results require that ψ satisfy certain properties which are true of CGFs for zero-mean random variables:

Definition 1 *A real-valued function ψ 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 ψ be a real-valued function with domain $[0, \lambda_{\max})$. We separate the definition of a* sub-ψ *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-ψ *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-ψ 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-ψ *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 (ΔY_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-ψ with self-normalizing process $U_t = \sum_{i=1}^t \Delta Y_i^2$.

The definition of sub-ψ generalizes the standard notion of being sub-Gaussian or sub-gamma to permit a general function ψ (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-ψ 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 ψ, 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-ψ 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

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

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.