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 (μ_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 (μ_t); a corresponding assumption on the lower tails of (S_t) yields an upper confidence sequence for (μ_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 μ, the resulting confidence sequence sequentially estimates μ, but the setup requires neither independence nor a common mean. In general the estimand μ_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, μ_t may also be random, although none of our applications involve random μ_t.

To construct uniform boundaries u_α 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 ψ be a real-valued function with domain $[0, \lambda_{\max})$. We assume, for* *each* $\lambda \in [0, \lambda_{\max})$*, there exists a supermartingale* $(L_t(\lambda))_{t=0}^{\infty}$ *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*

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

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 λ exaggerate larger movements in S_t, and ψ 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 ψ 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-ψ uniform boundary *with crossing probability* α *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* ψ *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-ψ 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-ψ *uniform boundary with crossing probability* α*.*

Five particular ψ 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 $c = 1$.

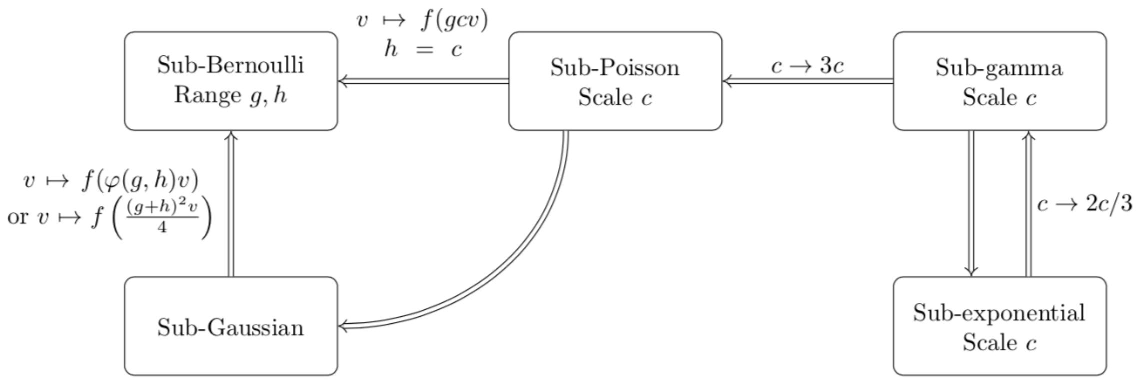

Figure 1: Schematic of relations among sub-ψ boundaries. Each arrow indicates that a sub-ψ boundary at the source node yields a sub-ψ boundary at the destination node with the modification indicated on the arrow. See Proposition 5 for a formal statement.

- $\psi_G(\lambda) := \lambda^2/(2(1-c\lambda))$ 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-ψ_G, and likewise for the other cases. The figure summarizes implications that hold among sub-ψ 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 ψ 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 α for scale c. Then $v \mapsto u_{k_1}(k_2 v)$ is a sub-ψ 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-ψ 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]* .