Basic Concentration Properties of Real-Valued Distributions

Odalric-Ambrym Maillard

BASIC CONCENTRATION PROPERTIES OF REAL-VALUED DISTRIBUTIONS

ODALRIC-AMBRYM MAILLARD

Inria Lille - Nord Europe
SequeL team
odalricambrym.maillard@inria.fr

In this note we introduce and discuss a few concentration tools for the study of concentration inequalities on the real line. After recalling versions of the Chernoff method, we move to concentration inequalities for predictable processes. We especially focus on bounds that enable to handle the sum of real-valued random variables, where the number of summands is itself a random stopping time, and target fully explicit and empirical bounds. We then discuss some important other tools, such as the Laplace method and the transportation lemma.

Keywords: Concentration of measure, Statistics.

Contents

1 Markov Inequality and the Chernoff method

In this section, we start by introducing the celebrated Markov's inequality, and show how this seemingly weak result leads to some of the most powerful tool in statistics: the Chernoff method, and the Laplace transform.

Lemma 1 (Markov's inequality) *For any measurable real-valued random variable that is almost surely non-negative, then it holds for all $\varepsilon > 0$ that*

$$\mathbb{P}(X \geqslant \varepsilon) \leqslant \frac{\mathbb{E}[X]}{\varepsilon} .$$

Proof of Lemma 1:

The proof uses the following straightforward decomposition:

$$X = X\mathbb{I}\{X \geqslant \varepsilon\} + X\mathbb{I}\{X < \varepsilon\}$$

Now since X is almost surely non-negative, it holds almost surely that $X\mathbb{I}\{X < \varepsilon\} \geqslant 0$, and thus $X \geqslant \varepsilon\mathbb{I}\{X \geqslant \varepsilon\}$. We conclude by taking expectations on both sides (which is valid since $\mathbb{E}[X] < \infty$), and deduce that $\mathbb{E}[X] \geqslant \varepsilon\mathbb{P}[X \geqslant \varepsilon]$. □

1.1 A first consequence

We can apply this result immediately to real-valued random variables by remarking that for any random variable distributed according to ν which we note $X \sim \nu$) and $\lambda \in \mathbb{R}$, the random variable $\exp(\lambda X)$ is non-negative. Thus if we now define the domain of ν by $\mathcal{D}_\nu = \{\lambda : \mathbb{E}[\exp(\lambda X)] < \infty\}$, we deduce by application of Markov's inequality that for all $t > 0$,

$$\begin{aligned} \forall \lambda \in \mathbb{R}_\star^+ \cap \mathcal{D}_\nu \quad \mathbb{P}(X \geqslant t) &= \mathbb{P}(\exp(\lambda X) \geqslant \exp(\lambda t)) \\ &\leqslant \exp(-\lambda t)\mathbb{E}[\exp(\lambda X)]\,. \qquad (1) \\ \forall \lambda \in \mathbb{R}_\star^- \cap \mathcal{D}_\nu \quad \mathbb{P}(X \leqslant t) &= \mathbb{P}(\exp(\lambda X) \geqslant \exp(\lambda t)) \\ &\leqslant \exp(-\lambda t)\mathbb{E}[\exp(\lambda X)]\,. \qquad (2) \end{aligned}$$

In this construction, the exp transform may seem arbitrary, and one could indeed use more general transforms. The benefit of using other transforms will be discussed later. Currently, we explore what happens with the exp case. One first immediate result is the following:

Lemma 2 (Chernoff's rule) *Let $X \sim \nu$ be a real-valued random variable. Then*

$$\log \mathbb{E}\exp(X) \leqslant 0\,, \quad \textit{implies} \quad \forall \delta \in (0,1], \quad \mathbb{P}\Big(X \geqslant \ln(1/\delta)\Big) \leqslant \delta\,.$$

The proof is immediate by considering $t = \ln(1/\delta)$ and $\lambda = 1$ in (1).

1.2 Two complementary results

Now one can consider two complementary points of view: The first one is to fix the value of t in (1) and (2) and minimize the probability level (the term on the right-hand side of the inequality). The second one is to fix the value of the probability level, and optimize the value of t. This leads to the following lemmas.

Lemma 3 (Cramer-Chernoff) *Let $X \sim \nu$ be a real-valued random variable. Let us introduce the log-Laplace transform and its Legendre transform:*

$$\begin{aligned} \forall \lambda \in \mathbb{R}, \quad \varphi_\nu(\lambda) &= \log \mathbb{E}[\exp(\lambda X)], \\ \forall t \in \mathbb{R}, \quad \varphi_\nu^\star(t) &= \sup_{\lambda \in \mathbb{R}} \Big(\lambda t - \varphi_\nu(\lambda)\Big)\,, \end{aligned}$$

and let $\mathcal{D}_\nu = \{\lambda \in \mathbb{R} : \varphi_\nu(\lambda) < \infty\}$.
If $\mathcal{D}_\nu \cap \mathbb{R}_\star^+ \neq \emptyset$, then $\mathbb{E}[X] < \infty$ and for all $t \geqslant \mathbb{E}[X]$

$$\log \mathbb{P}(X \geqslant t) \leqslant -\varphi_\nu^\star(t)\,.$$

Likewise, if $\mathcal{D}_\nu \cap \mathbb{R}_\star^- \neq \emptyset$, $\mathbb{E}[X] > -\infty$ and for all $t \leqslant \mathbb{E}[X]$,

$$\log \mathbb{P}(X \leqslant t) \leqslant -\varphi_\nu^\star(t)\,.$$

Remark 1 *The log-Laplace transform φ_ν is also called known as the cumulant generative function.*

Proof of Lemma 3:

First, note that $\{\lambda \in \mathbb{R} : \mathbb{E}[\exp(\lambda X)] < \infty\}$ coincides with $\{\lambda \in \mathbb{R} : \varphi_\nu(\lambda) < \infty\}$. Using equations (1) and (2), it holds:

$$\begin{aligned} \mathbb{P}(X \geqslant t) &\leqslant \inf_{\lambda \in \mathbb{R}_\star^+ \cap \mathcal{D}_\nu} \exp(-\lambda t + \log \mathbb{E}[\exp(\lambda X)]) \\ \mathbb{P}(X \leqslant t) &\leqslant \inf_{\lambda \in \mathbb{R}_\star^- \cap \mathcal{D}_\nu} \exp(-\lambda t + \log \mathbb{E}[\exp(\lambda X)]) \end{aligned}$$

The Legendre transform $\varphi_\nu^\star$ of the log-Laplace function φ_ν unifies these two cases. Indeed, a striking property of $\varphi_\nu^\star$ is that if $\lambda \in \mathcal{D}_\nu$ for some $\lambda > 0$, then $\mathbb{E}[X] < \infty$. This can be seen by Jensen's inequality applied to the function $\ln$: Indeed it holds $\lambda \mathbb{E}[X] = \mathbb{E}[\ln \exp(\lambda X)] \leqslant \varphi_\nu(\lambda)$. Further, for all $t \geqslant \mathbb{E}[X]$, it holds

$$\varphi_\nu^\star(t) = \sup_{\lambda \in \mathbb{R}^+ \cap \mathcal{D}_\nu} \Big(\lambda t - \varphi_\nu(\lambda) \Big).$$

Note that this also applies if $\mathbb{E}[X] = -\infty$. Likewise, if $\lambda \in \mathcal{D}_\nu$ for some $\lambda < 0$ then $\mathbb{E}[X] > -\infty$ and for all $t \leqslant \mathbb{E}[X]$, it holds

$$\varphi_\nu^\star(t) = \sup_{\lambda \in \mathbb{R}^- \cap \mathcal{D}_\nu} \Big(\lambda t - \varphi_\nu(\lambda) \Big).$$

□

Alternatively, the second point of view is to fix the confidence level $\delta \in (0, 1]$, and then to solve the equation $\exp(-\lambda t)\mathbb{E}[\exp(\lambda X)] = \delta$ in $t = t(\delta)$. We then optimize over t. This leads to:

Lemma 4 (Alternative Cramer-Chernoff) *Let $X \sim \nu$ be a real-valued random variable and let $\mathcal{D}_\nu = \{\lambda \in \mathbb{R} : \log \mathbb{E} \exp(\lambda X) < \infty\}$. It holds,*

$$\mathbb{P}\Big[X \geqslant \inf_{\lambda \in \mathcal{D}_\nu \cap \mathbb{R}_\star^+} \Big\{ \frac{1}{\lambda} \log \mathbb{E}[\exp(\lambda X)] + \frac{\log(1/\delta)}{\lambda} \Big\} \Big] \leqslant \delta \tag{3}$$

$$\mathbb{P}\Big[X \leqslant \sup_{\lambda \in (-\mathcal{D}_\nu) \cap \mathbb{R}_\star^+} \Big\{ -\frac{1}{\lambda} \log \mathbb{E}[\exp(-\lambda X)] - \frac{\log(1/\delta)}{\lambda} \Big\} \Big] \leqslant \delta \,. \tag{4}$$

Proof of Lemma 4:

Solving $\exp(-\lambda t)\mathbb{E}[\exp(\lambda X)] = \delta$ for $\delta \in (0, 1]$ and $\lambda \neq 0$, we obtain he following equivalence

$$\begin{aligned} -\lambda t + \log \mathbb{E}[\exp(\lambda X)] &= \log(\delta) \\ \lambda t &= -\log(\delta) + \log \mathbb{E}[\exp(\lambda X)] \\ t &= \frac{1}{\lambda} \log(1/\delta) + \frac{1}{\lambda} \log \mathbb{E}[\exp(\lambda X)] \,. \end{aligned}$$

Thus, we deduce from (1) and (2) that

$$\begin{aligned} \forall \lambda > 0 \quad \mathbb{P}\Big[X \geqslant \frac{1}{\lambda} \log(1/\delta) + \frac{1}{\lambda} \log \mathbb{E}[\exp(\lambda X)] \Big] &\leqslant \delta \\ \forall \lambda > 0 \quad \mathbb{P}\Big[X \leqslant -\frac{1}{\lambda} \log(1/\delta) - \frac{1}{\lambda} \log \mathbb{E}[\exp(-\lambda X)] \Big] &\leqslant \delta \,. \end{aligned}$$

□

The rescaled Laplace transform $\lambda \to \frac{1}{\lambda} \log \mathbb{E}[\exp(\lambda X)]$ is sometimes called the *entropic risk* measure. Note that Lemma 3 and 4 involve slightly different quantities, depending on whether we focus on the probability level δ or the threshold on X.

1.3 The illustrative case of sub-Gaussian random variables

An immediate corollary is the following:

Corollary 1 (Sub-Gaussian Concentration Inequality) *Let $\{X_i\}_{i\leqslant n}$ be independent R-sub-Gaussian random variables with mean μ, that is such that*

$$\forall \lambda \in \mathbb{R}, \quad \log \mathbb{E} \exp\Big(\lambda(X_i - \mu)\Big) \leqslant \frac{1}{2}\lambda^2 R^2 \,.$$

Then,

$$\forall \delta \in (0,1) \qquad \mathbb{P}\Big[\sum_{i=1}^n (X_i - \mu) \geqslant \sqrt{2R^2 n \log(1/\delta)}\Big] \leqslant \delta$$

Remark 2 *This corollary naturally applies to Gaussian random variables with variance σ^2, in which case $R = \sigma$. It also applies to bounded random variable. Indeed random variables $\{X_i\}_{i\leqslant n}$ bounded in $[0,1]$ are $1/2$-sub-Gaussian. This can be understood intuitively by remarking that distributions with the highest variance on $[0,1]$ are Bernoulli, and that the variance of a Bernoulli with parameter $\theta \in [0,1]$ is $\theta(1-\theta) \leqslant 1/4$, thus resulting in $R^2 = 1/4$. This is proved more formally via Hoeffding's lemma.*

Proof of Corollary 1:

Indeed, it holds that

$$\begin{aligned}
\frac{1}{\lambda}\log \mathbb{E}\big[\exp(\lambda \sum_{i=1}^n (X_i - \mu))\big] &= \frac{1}{\lambda}\log \mathbb{E}\Big[\prod_{i=1}^n \exp(\lambda(X_i - \mu))\Big] \\
&\overset{(a)}{=} \frac{1}{\lambda}\log \prod_{i=1}^n \mathbb{E}\Big[\exp(\lambda(X_i - \mu))\Big] \\
&= \frac{1}{\lambda}\sum_{i=1}^n \log \mathbb{E}\Big[\exp(\lambda(X_i - \mu))\Big] \\
&\overset{(b)}{\leqslant} \frac{n}{2}\lambda R^2 \,,
\end{aligned}$$

where (a) is by independence, and (b) holds by using the sub-Gaussian assumption. We deduce by Lemma 4 that

$$\begin{aligned}
&\mathbb{P}\Big[\sum_{i=1}^n (X_i - \mu) \geqslant \inf_{\lambda \in \mathcal{D}_\nu \cap \mathbb{R}_\star^+} \Big\{\lambda R^2 n/2 + \frac{\log(1/\delta)}{\lambda}\Big\}\Big] \\
&\overset{(a)}{\leqslant} \mathbb{P}\Big[\sum_{i=1}^n (X_i - \mu) \geqslant \inf_{\lambda \in \mathcal{D}_\nu \cap \mathbb{R}_\star^+} \Big\{\frac{1}{\lambda}\log \mathbb{E}\Big[\exp\Big(\lambda \sum_{i=1}^n (X_i - \mu)\Big)\Big] + \frac{\log(1/\delta)}{\lambda}\Big\}\Big] \\
&\leqslant \delta \,,
\end{aligned}$$

where in (a), we used that $x < y$ implies $\mathbb{P}(X \geqslant y) \leqslant \mathbb{P}(X \geqslant x)$.

Now we note that $\mathcal{D}_\nu = \mathbb{R}$ by the sub-Gaussian assumption, where ν is the distribution of $\sum_{i=1}^n X_i$. We conclude by noticing that $\lambda_\delta = \sqrt{\frac{2\log(1/\delta)}{R^2 n}}$ achieves the minimum in

$$\inf_{\lambda \in \mathbb{R}_\star^+} \Big\{\lambda R^2 n/2 + \frac{\log(1/\delta)}{\lambda}\Big\} = \sqrt{2R^2 n \log(1/\delta)} \,. \qquad \square$$

2 Concentration inequalities for predictable processes

In practice, it is often desirable to control not only a random variable such as an empirical mean at a single time step n, but also at multiple time steps $n = 1, \ldots$. The naive approach to do so is by controlling the concentration at each different time step and then use a union-bound to deduce the final bound.

However, this is generally sub-optimal as the empirical mean at time n and at time $n+1$ are close to each other and correlated. We study here two powerful methods that enable to improve on this naive strategy.

2.1 Doob's maximal inequalities

We start by recalling two standard and important inequalities that can be found in most introductory textbooks on statistics:

Lemma 5 (Doob's maximal inequality for non-negative sub-martingale) *Let $\{W_t\}_{t\in\mathbb{N}}$ be a non-negative sub-martingale with respect to the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$, that is*

$$\forall t \in \mathbb{N}, \qquad \mathbb{E}[W_{t+1}|\mathcal{F}_t] \geqslant W_t \,, \text{ and } W_t \geqslant 0 \,.$$

Then, for all $p \geqslant 1$ and $\varepsilon > 0$, it holds for all $T \in \mathbb{N}$

$$\mathbb{P}\left(\max_{0\leqslant t\leqslant T} W_t \geqslant \varepsilon \right) \leqslant \frac{\mathbb{E}[W_T^p]}{\varepsilon^p} \,.$$

Lemma 6 (Doob's maximal inequality for non-negative super-martingale) *Let $\{W_t\}_{t\in\mathbb{N}}$ be a non-negative super-martingale with respect to the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$, that is*

$$\forall t \in \mathbb{N}, \qquad \mathbb{E}[W_{t+1}|\mathcal{F}_t] \leqslant W_t \,, \text{ and } W_t \geqslant 0 \,.$$

Then, for all $p \geqslant 1$ and $\varepsilon > 0$, it holds for all $T \in \mathbb{N}$

$$\mathbb{P}\left(\max_{0\leqslant t\leqslant T} W_t \geqslant \varepsilon \right) \leqslant \frac{\mathbb{E}[W_0^p]}{\varepsilon^p} \,.$$

In particular, if $\mathbb{E}[W_0] \leqslant 1$, then for all $T \in \mathbb{N}$, $\mathbb{P}\left(\max_{0\leqslant t\leqslant T} W_t \geqslant \varepsilon \right) \leqslant \varepsilon^{-1}$.

2.2 The peeling technique for random stopping times

In this section, we provide a powerful result that is useful when dealing with generic real-valued distributions. We say a process generating a sequence of random variables $\{Z_i\}_{i=1}^{\infty}$ is predictable if there exists a filtration $\mathcal{H} = (\mathcal{H}_n)_{n\in\mathbb{N}}$ ("filtration of the past") such that Z_n is $\mathcal{H}_n$-measurable for all n. We say a random variable N is a random stopping time for $\mathcal{H}$ if $\forall m \in \mathbb{N}, \{N \leqslant m\}$ is $\mathcal{H}_{m-1}$-measurable.

Lemma 7 (Concentration inequality for predictable processes) *Let $\{Z_i\}_{i=1}^{\infty}$ be a sequence of random variables generated by a predictable process. Let $\varphi : \mathbb{R} \to \mathbb{R}^+$ be a convex upper-envelope of the cumulant generative function of the conditional distributions with $\varphi(0) = 0$, and $\varphi_\star$ its Legendre-Fenchel transform, that is:*

$$\begin{aligned} \forall \lambda \in \mathcal{D}, \forall i, \qquad & \ln \mathbb{E}\Big[\exp\Big(\lambda Z_i\Big) \Big| \mathcal{H}_{i-1} \Big] \leqslant \varphi(\lambda)\,, \\ \forall x \in \mathbb{R} \qquad & \varphi_\star(x) = \sup_{\lambda \in \mathbb{R}} \big(\lambda x - \varphi(\lambda)\big), \end{aligned}$$

where $\mathcal{D} = \{\lambda \in \mathbb{R} : \forall i, \ln \mathbb{E}\Big[\exp\Big(\lambda Z_i\Big) \Big| \mathcal{H}_{i-1} \Big] < \infty\}$. Assume that $\mathcal{D}$ contains an open neighborhood of 0. Then, $\forall c \in \mathbb{R}^+$, there exists a unique x_c such that for all i, $x_c > \mathbb{E}\Big[Z_i \Big| \mathcal{H}_{i-1}\Big]$, and $\varphi_\star(x_c) = c$, and a unique x'_c such that for all i, $x'_c < \mathbb{E}\Big[Z_i \Big| \mathcal{H}_{i-1}\Big]$ and $\varphi_\star(x'_c) = c$. We define $\varphi^{-1}_{\star,+} : c \mapsto x_c$, $\varphi^{-1}_{\star,-} : c \mapsto x'_c$. Then $\varphi^{-1}_{\star,+}$ is not decreasing and $\varphi^{-1}_{\star,-}$ is not increasing.
Let N_n be a random stopping time (for the filtration generated by $\{Z_i\}_{i=1}^{\infty}$) a.s. bounded by n. Then for all $\alpha \in (1, n]$, and $\delta \in (0, 1)$,

$$\begin{aligned} \mathbb{P}\left[\frac{1}{N_n} \sum_{i=1}^{N_n} Z_i \geqslant \varphi^{-1}_{\star,+}\left(\frac{\alpha}{N_n} \ln\left(\left\lceil \frac{\ln(n)}{\ln(\alpha)} \right\rceil \frac{1}{\delta} \right) \right) \right] &\leqslant \delta \\ \mathbb{P}\left[\frac{1}{N_n} \sum_{i=1}^{N_n} Z_i \leqslant \varphi^{-1}_{\star,-}\left(\frac{\alpha}{N_n} \ln\left(\left\lceil \frac{\ln(n)}{\ln(\alpha)} \right\rceil \frac{1}{\delta} \right) \right) \right] &\leqslant \delta \end{aligned}$$

In particular, one can take α to be the minimal solution to $\ln(\alpha) e^{1/\ln(\alpha)} = \ln(n)/\delta$.
Now, if N is a (possibly unbounded) random stopping time for the filtration generated by $\{Z_i\}_{i=1}^{\infty}$, it holds for all deterministic $\alpha > 1$ and $\delta \in (0, 1)$,

$$\begin{aligned} \mathbb{P}\left[\frac{1}{N} \sum_{i=1}^{N} Z_i \geqslant \varphi^{-1}_{\star,+}\left(\frac{\alpha}{N} \ln\left[\frac{\ln(N)\ln(\alpha N)}{\delta \ln^2(\alpha)} \right] \right) \right] &\leqslant \delta \\ \mathbb{P}\left[\frac{1}{N} \sum_{i=1}^{N} Z_i \leqslant \varphi^{-1}_{\star,-}\left(\frac{\alpha}{N} \ln\left[\frac{\ln(N)\ln(\alpha N)}{\delta \ln^2(\alpha)} \right] \right) \right] &\leqslant \delta \end{aligned}$$

Proof of Lemma 7:

First, one easily derives the following properties, from properties of the Legendre-Fenchel transform.

- $\varphi_\star(0) = 0$, $\varphi_\star(x) \overset{x \to +\infty}{\to} \infty$, $\varphi_\star$ is convex, increasing on $\mathbb{R}^+$.
- $\forall x$ such that $\varphi_\star(x) < \infty$, there exists a unique $\lambda_x \in \mathcal{D}_\nu$ such that $\varphi_\star(x) = \lambda_x x - \varphi(\lambda_x)$.
- $\forall c \in \mathbb{R}^+$, there exists a unique $x_c > \mathbb{E}[Z]$ such that $\varphi_\star(x_c) = c$. We write it $\varphi^{-1}_{\star,+}(c)$. $\varphi^{-1}_{\star,+}$ is not decreasing.

---

1. A peeling argument We start with a peeling argument. Let us choose some $\eta > 0$ and define $t_k = (1+\eta)^k$, for $k = 0, \dots, K$, with $K = \lceil \frac{\ln(n)}{\ln(1+\eta)} \rceil$ (thus $n \leqslant t_K$).

Let $\varepsilon_t \in \mathbb{R}^+$ be a sequence that is non-increasing in t.

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big)\\
&\quad\leqslant\ \mathbb{P}\Big(\bigcup_{k=1}^{K}\{t_{k-1} < N_n \leqslant t_k\}\cap\{\sum_{i=1}^{N_n} Z_i \geqslant N_n\varepsilon_{N_n}\}\Big)\\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\exists t\in(t_{k-1},t_k] : \sum_{i=1}^{t} Z_i \geqslant t\varepsilon_t\Big)
\end{aligned}$$

Let $\lambda_k > 0$, for $k = 1,\dots,K$.

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big)\\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\exists t\in(t_{k-1},t_k] : \sum_{i=1}^{t} Z_i \geqslant t\varepsilon_t\Big)\\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\exists t\in(t_{k-1},t_k] : \exp\Big(\lambda_k\Big(\sum_{i=1}^{t} Z_i\Big)\Big) \geqslant \exp(\lambda_k t\varepsilon_t)\Big)\\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\exists t\in(t_{k-1},t_k] : \underbrace{\exp\Big(\lambda_k\Big(\sum_{i=1}^{t} Z_i\Big) - t\varphi(\lambda_k)\Big)}_{W_{k,t}}\\
&\qquad\qquad\qquad\qquad\geqslant \exp\Big(t\big(\lambda_k\varepsilon_t - \varphi(\lambda_k)\big)\Big)\Big)\\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\exists t\in(t_{k-1},t_k] : W_{k,t} \geqslant \exp\Big(t\big(\lambda_k\varepsilon_{t_k} - \varphi(\lambda_k)\big)\Big)\Big).
\end{aligned}$$

Since $\varepsilon_{t_k} > 0$, we can choose a $\lambda_k > 0$ such that $\varphi^\star(\varepsilon_{t_k}) = \lambda_k\varepsilon_{t_k} - \varphi(\lambda_k)$.

---

2. Doob's maximal inequality At this, point, we show that the sequence $\{W_{k,t}\}_t$ is a non-negative super-martingale, where $W_{k,t} = \exp\Big(\lambda_k\big(\sum_{i=1}^{t} Z_i\big) - t\varphi(\lambda_k)\Big)$. Indeed, note that:

$$\begin{aligned}
\mathbb{E}[W_{k,t+1}|\mathcal{F}_t] &= W_{k,t}\mathbb{E}[\exp\big(\lambda_k Z_{t+1}\big)|\mathcal{F}_t]\exp(-\varphi(\lambda_k))\\
&\leqslant W_{k,t}\,.
\end{aligned}$$

Thus, using that $t_{k-1} \geqslant t_k/(1+\eta)$, we find

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big)\\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\exists t\in(t_{k-1},t_k] W_{k,t} \geqslant \exp\Big(t\varphi^\star(\varepsilon_{t_k})\Big)\Big)\\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\max_{t\in(t_{k-1},t_k]} W_{k,t} \geqslant \exp\Big(\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\Big)\\
&\overset{(a)}{\leqslant}\ \sum_{k=1}^{K}\exp\Big(-\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big),
\end{aligned}$$

where (a) holds by application of Doob's maximal inequality for non-negative super-martingales, using that $\max_{t\in(t_{k-1},t_k]} W_{k,t} \leqslant \max_{t\in(0,t_k]} W_{k,t}$ and $W_{k,0} \leqslant 1$.

3. Parameter tuning for bounded N_n Now, let us choose ε_t such that $t\varphi_\star(\varepsilon_t) = c > 1$ is a constant, that is $\varepsilon_t = \varphi_{\star,+}^{-1}(c/t)$ (non increasing with t). Thus, we get for all $\eta \in (0, n-1)$:

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big) &\leqslant \sum_{k=1}^{\lceil \frac{\ln(n)}{\ln(1+\eta)}\rceil} \exp\Big(-\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\\ &\leqslant \lceil\frac{\ln(n)}{\ln(1+\eta)}\rceil \exp\Big(-\frac{c}{1+\eta}\Big),\end{aligned}$$

which suggest to set $c = (1+\eta)\ln\Big(\lceil\frac{\ln(n)}{\ln(1+\eta)}\rceil\frac{1}{\delta}\Big)$. We thus obtain for all $\eta \in [0, n-1]$,

$$\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varphi_{\star,+}^{-1}\Big[\frac{1+\eta}{N_n}\ln\Big(\Big\lceil\frac{\ln(n)}{\ln(1+\eta)}\Big\rceil\frac{1}{\delta}\Big)\Big]\Big) \leqslant \delta\,.$$

Then, it makes sense to find the minimum value of $f : x \to x\ln(\frac{a}{\ln(x)})$, for $x > 1$. An optimal point $x_\star > 1$ satisfies

$$f'(x) = \ln\Big(\frac{a}{\ln(x)}\Big) + x\frac{-(1/x)/\ln^2(x)}{1/\ln(x)} = \ln\Big(\frac{a}{\ln(x)}\Big) - \frac{1}{\ln(x)} = 0\,,$$

that is $x_\star$ satisfies $a = \ln(x_\star)e^{1/\ln(x_\star)}$. We may thus choose the (slightly suboptimal) minimal value x that satisfies $\ln(x)e^{1/\ln(x)} = \ln(n)/\delta$.

4. Parameter tuning for unbounded N
We restart from

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i \geqslant \varepsilon_N\Big) \leqslant \sum_{k=1}^{K}\exp\Big(-\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big),$$

where $t_k = (1+\eta)^k$ and $K = \lceil\frac{\ln(n)}{\ln(1+\eta)}\rceil$, and choose a different tuning for ε_t in order to handle an infinite sum (with $K = \infty$). Let us choose ε_t that satisfies $t\varphi_\star(\varepsilon_t) = c(t)$, where $c(t)$ is chosen such that

$$\sum_{k=1}^{\infty}\exp\Big(-\frac{c(t_k)}{1+\eta}\Big) < \infty\,.$$

Choosing $c(t) = (1+\eta)\ln\Big(\frac{\ln(t)}{\delta\ln(1+\eta)}[\frac{\ln(t)}{\ln(1+\eta)}+1]\Big)$, it comes $c(t_k) = (1+\eta)\ln(k(k+1)\delta)$ and thus

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i \geqslant \varepsilon_N\Big) \leqslant \sum_{k=1}^{\infty}\frac{\delta}{k(k+1)} = \delta\,,$$

With this choice, we thus deduce

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i \geqslant \varphi_{\star,+}^{-1}\Big(\frac{(1+\eta)}{N}\ln\Big(\frac{\ln(N)\ln(N(1+\eta))}{\delta\ln^2(1+\eta)}\Big)\Big) \leqslant \delta\,.$$

5. Reverse bounds. We now provide a similar result for the reverse bound. Let $\varepsilon_t \in \mathbb{R}$ be a sequence that is non-decreasing with t, and $\lambda_k > 0$, for $k = 1, \ldots, K$. Then

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i \leqslant \varepsilon_N\Big) &\leqslant \sum_{k=1}^{K}\mathbb{P}\Big(\exists t \in (t_{k-1}, t_k]\exp\Big(-\lambda_k\Big(\sum_{i=1}^{t} Z_i\Big) - t\varphi(-\lambda_k)\Big)\\ &\qquad \geqslant \exp\Big(t(-\lambda_k\varepsilon_t - \varphi(-\lambda_k))\Big)\Big)\\ &\leqslant \sum_{k=1}^{K}\mathbb{P}\Big(\exists t \in (t_{k-1}, t_k], W_{k,t} \geqslant \exp\Big(t(-\lambda_k\varepsilon_{t_k} - \varphi(-\lambda_k))\Big)\Big)\end{aligned}$$

If $\varepsilon_{t_k} < \mathbb{E}[Z_{t_k}]$, we can choose $\lambda_k = \lambda_{\varepsilon_{t_k}} > 0$ such that $\varphi^\star(\varepsilon_{t_k}) = -\lambda_k \varepsilon_{t_k} - \varphi(-\lambda_k) \geqslant 0$. Thus, using that $t_{k-1} > t_k/(1+\eta)$, it comes

$$\begin{aligned}
\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^N Z_i \leqslant \varepsilon_N\Big) &\leqslant \sum_{k=1}^K \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k], W_{k,t} \geqslant \exp\Big(t\varphi^\star(\varepsilon_{t_k})\Big)\Big)\\
&\leqslant \sum_{k=1}^K \mathbb{P}\Big(\max_{t\in(t_{k-1},t_k]} W_{k,t} \geqslant \exp\Big(\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\Big)\\
&\leqslant \sum_{k=1}^K \exp\Big(\frac{-t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)
\end{aligned}$$

Now, let us choose $\varepsilon_t < \mathbb{E}[Z_t]$ such that $t\varphi_\star(\varepsilon_t) = c > 1$, that is $\varepsilon_t = \varphi_{\star,-}^{-1}(c/t)$ (non decreasing with t). For $\eta = 1/(c-1)$ and $c = \ln(e/\delta)$, we obtain

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^N Z_i \leqslant \varphi_{\star,-}^{-1}(\ln(e/\delta)/N)\Big) \leqslant \lceil \ln(n)\ln(e/\delta)\rceil \delta\,.$$

□

Improvement In some situations, it is possible to refine the previous result

Corollary 2 (Improved concentration inequality for predictable processes) *Under the same setting as Lemma 7, let* $h_n(x) = \log\lceil\frac{\log(n)}{\log(1/x)}\rceil$, *and* $h_{n,\star}$ *its Legendre-Fenchel transform. Finally let* $c = h_{n,\star,+}^{-1}(\log(1/\delta))$.
If it holds that $\sup_{x\in(1/n,1)} cx - h_n(x) = h_{n,\star}(c)$, *then,*

$$\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varphi_{\star,+}^{-1}\left(\frac{h_{n,\star,+}^{-1}(\log(1/\delta))}{N_n}\right)\Big) \leqslant \delta\,.$$

Note that the restrictive condition that $\sup_{x\in(1/n,1)} cx - h_n(x) = h_{n,\star}(c)$ is on whether the maximum of $cx - h_n(x)$ is reached by a point x in the set $(1/n, 1)$.

Proof of Corollary 2:

Indeed, following the proof of Lemma 7, it suffices to refine the last step:

$$\begin{aligned}
&\inf_{\eta\in(0,n-1)} \lceil\frac{\log(n)}{\log(1+\eta)}\rceil \exp\Big(-\frac{c}{1+\eta}\Big)\\
&= \inf_{x\in(1/n,1)} \exp\Big(-cx + \log\lceil\frac{\log(n)}{\log(1/x)}\rceil\Big)\\
&= \exp\Big(-\Big(\sup_{x\in(1/n,1)} cx - \log\lceil\frac{\log(n)}{\log(1/x)}\rceil\Big)\Big)\\
&= \exp\Big(-h_{n,\star}(c)\Big) = \delta.
\end{aligned}$$

□

2.3 Birge-Massart concentration

We conclude this section by applying Lemma 7 to the concentration of the quadratic sum of a noise term ξ_i. We believe that this illustrates the power of this method. Assume that the noise terms are strongly sub-Gaussian in the sense that

$$\forall\lambda \in \mathcal{D}_\nu, \forall i \quad \log\mathbb{E}[\exp(\lambda\xi_i^2)|\mathcal{H}_{i-1}] \leqslant \varphi(\lambda)$$

where $\varphi(\lambda) = -\frac{1}{2}\log(1-2\lambda R^2)$. Note that this is the cumulant generative function of the square of a centered Gaussian. Then we can prove the following result:

Lemma 8 (Birge-Massart concentration for predictable process) *Assume that N_n is a random stopping time that satisfies $N_n \leqslant n$ almost surely, then it holds for all $\alpha > 1$*

$$\mathbb{P}\Big[\frac{1}{N_n}\sum_{i=1}^{N_n}\xi_i^2 \geqslant R^2 + 2R^2\sqrt{\frac{2\alpha}{N_n}\ln\Big(\Big\lceil\frac{\ln(n)}{\ln(\alpha)}\Big\rceil\frac{1}{\delta}\Big)} + \frac{2\alpha R^2}{N_n}\ln\Big(\Big\lceil\frac{\ln(n)}{\ln(\alpha)}\Big\rceil\frac{1}{\delta}\Big)\Big] \leqslant \delta$$

$$\mathbb{P}\Big[\frac{1}{N_n}\sum_{i=1}^{N_n}\xi_i^2 \leqslant R^2 - 2R^2\sqrt{\frac{\alpha}{N_n}\ln\Big(\Big\lceil\frac{\ln(n)}{\ln(\alpha)}\Big\rceil\frac{1}{\delta}\Big)}\Big] \leqslant \delta$$

Further, for a random stopping time N, then it holds for all $\alpha > 1$,

$$\mathbb{P}\Big[\frac{1}{N}\sum_{i=1}^{N}\xi_i^2 \geqslant R^2 + 2R^2\sqrt{\frac{2\alpha}{N}\ln\Big[\frac{\ln(N)\ln(\alpha N)}{\delta\ln^2(\alpha)}\Big]} + \frac{2\alpha R^2}{N}\ln\Big[\frac{\ln(N)\ln(\alpha N)}{\delta\ln^2(\alpha)}\Big]\Big] \leqslant \delta$$

$$\mathbb{P}\Big[\frac{1}{N}\sum_{i=1}^{N}\xi_i^2 \leqslant R^2 - 2R^2\sqrt{\frac{\alpha}{N}\ln\Big[\frac{\ln(N)\ln(\alpha N)}{\delta\ln^2(\alpha)}\Big]}\Big] \leqslant \delta$$

Proof of Lemma 8:

According to Lemma 7 applied to $Z_i = \xi_i^2$, all we have to do is to compute an upper bound on the quantity $\varphi_{\star,+}^{-1}(c)$, first for the value $c = \frac{\ln(e/\delta)}{N_n}$, then for $c = \frac{\ln(e/\delta)}{N}(1 + \frac{2}{\ln(1/\delta)}\ln\Big(\frac{\pi\ln(N)\ln(1/\delta)}{6^{1/2}(1+\ln(1/\delta))}\Big)) \leqslant \frac{\ln(e/\delta)}{N}(1 + c_N/\ln(1/\delta))$. We proceed in the following way. First, the envelope function is given by

$$\varphi(\lambda) = -\frac{1}{2}\ln(1-2\lambda R^2) \leqslant \frac{\lambda R^2}{1-2\lambda R^2}\,.$$

for $\lambda \in (0, \frac{1}{2R^2})$. Let $x > R^2$. It holds that $\varphi^\star(x) \geqslant \sup_\lambda[\lambda x - \frac{\lambda R^2}{1-2\lambda R^2}]$. Solving this optimization by differentiating over λ, the supremum is reached for $\lambda = (1-\frac{R}{\sqrt{x}})\frac{1}{2R^2} \in (0, \frac{1}{2R^2})$, with corresponding value given by

$$\begin{aligned}\tilde{\varphi}^\star(x) &= (1-\frac{R}{\sqrt{x}})\frac{x}{2R^2} - (1-\frac{R}{\sqrt{x}})\frac{\sqrt{x}}{2R} \\ &= \frac{x}{2R^2} - \frac{\sqrt{x}}{R} + \frac{1}{2}\,.\end{aligned}$$

Now, for $c > 0$, it is easily checked that $\tilde{\varphi}^\star(x) = c$ holds for $x_c = R^2(1+\sqrt{2c})^2$. As a result, we deduce that $\varphi_{\star,+}^{-1}(c) \leqslant R^2(1+\sqrt{2c})^2 = R^2 + 2R^2c + 2R^2\sqrt{2c}$.

Now, for the reverse inequality, we have to compute a lower bound on the quantity $\varphi_{\star,-}^{-1}(c)$, first for $c = \frac{\ln(e/\delta)}{N_n}$, then for $c = \frac{\ln(e/\delta)}{N}(1+\frac{2}{\ln(1/\delta)}\ln\Big(\frac{\pi\ln(N)\ln(1/\delta)}{6^{1/2}(1+\ln(1/\delta))}\Big)) \leqslant \frac{\ln(e/\delta)}{N}(1+c_N/\ln(1/\delta))$. We proceed in the following way. First, the envelope function is given for $\lambda > 0$ by

$$\varphi(-\lambda) = -\frac{1}{2}\ln(1+2\lambda R^2) \geqslant -\frac{\lambda R^2}{1+\lambda R^2}\,.$$

Thus, for $0 < x < R^2$ it holds $\varphi^\star(x) \geqslant \sup_{\lambda>0}[-\lambda x + \frac{\lambda R^2}{1+\lambda R^2}] = 1 + \sup_{\lambda>0}[-\lambda x - \frac{1}{1+\lambda R^2}]$. Solving this optimization by differentiating over λ, the supremum is reached for $\lambda = \frac{1}{R^2}(\frac{R}{\sqrt{x}} - 1) > 0$ with corresponding value given by

$$\begin{aligned}\tilde{\varphi}^\star(x) &= 1 - \frac{x}{R^2}(\frac{R}{\sqrt{x}} - 1) - \frac{\sqrt{x}}{R} \\ &= \frac{x}{R^2} - 2R\frac{\sqrt{x}}{R} + 1\,.\end{aligned}$$

Now, for $c > 0$, it is easily checked that $\tilde{\varphi}^\star(x) = c$ holds for $x_c = R^2(1-\sqrt{c})^2$, and $x_c < R^2$ if $c < 1$. As a result, we deduce that if $c \in (0,1)$, then $\varphi^{-1}_{\star,-}(c) \geqslant R^2 - 2R^2\sqrt{c} + R^2 c$. On the other hand, for all $c > 0$, choosing $\lambda = \frac{1}{R^2}\sqrt{c}$, and using the inequality $\frac{1}{1+v} \geqslant 1 - v$ for $v > 0$, then

$$\begin{aligned}\varphi^\star(x) &\geqslant -\frac{x}{R^2}\sqrt{c} + 1 - \frac{1}{1+\sqrt{c}} = \sqrt{c}\big(-\frac{x}{R^2} + \frac{1}{1+\sqrt{c}}\big) \\ &\geqslant \tilde{\varphi}^\star(x) \stackrel{\text{def}}{=} \sqrt{c}\big(-\frac{x}{R^2} + 1 - \sqrt{c}\big)\end{aligned}$$

Thus, $\tilde{\varphi}^\star(x) = c$ for $x_c = R^2 - 2R^2\sqrt{c} < R^2$. As a result, we deduce that if $c > 0$, then $\varphi^{-1}_{\star,-}(c) \geqslant R^2 - 2R^2\sqrt{c}$. □

3 Uniform bounds and the Laplace method

In this section, we present another very powerful tool, that is the Laplace method (method of mixtures for sub-Gaussian random variables). We provide the illustrative following result here, for real-valued randome variables. The result however extends naturally to dimension d, and even, to some extent, to infinite dimension.

Lemma 9 (Uniform confidence intervals) *Let $Y_1, \dots Y_t$ be a sequence of t i.i.d. real-valued random variables bounded in $[0,1]$, with mean μ. Let $\mu_t = \frac{1}{t}\sum_{s=1}^t Y_s$ be the empirical mean estimate. Then, for all $\delta \in (0,1)$, it holds*

$$\begin{aligned}\mathbb{P}\left(\exists t \in \mathbb{N}, \quad \mu_t - \mu \geqslant \sqrt{\Big(1+\frac{1}{t}\Big)\frac{\ln\big(\sqrt{t+1}/\delta\big)}{2t}}\right) &\leqslant \delta \\ \mathbb{P}\left(\exists t \in \mathbb{N}, \quad \mu - \mu_t \geqslant \sqrt{\Big(1+\frac{1}{t}\Big)\frac{\ln\big(\sqrt{t+1}/\delta\big)}{2t}}\right) &\leqslant \delta\,.\end{aligned}$$

Proof of Lemma 9:

The first result is Hoeffding's inequality for i.i.d. bounded random variables. For the second one, we introduce for a fixed $\delta \in [0,1]$ the random variable

$$\tau = \min\left\{t \in \mathbb{N} : \mu_t - \mu \geqslant \sqrt{\Big(1+\frac{1}{t}\Big)\frac{\ln\big(\sqrt{1+t}/\delta\big)}{2t}}\right\}.$$

This quantity is a random stopping time for the filtration $\mathcal{F} = (\mathcal{F}_t)_t$, where $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$, since $\{\tau \leqslant m\}$ is $\mathcal{F}_m$-measurable for all m. We want to show that $\mathbb{P}(\tau < \infty) \leqslant \delta$. To this end, for any λ, and t, we introduce the following quantity

$$M_t^\lambda = \exp\left(\sum_{s=1}^t (\lambda(Y_s - \mu) - \frac{\lambda^2}{8})\right).$$

By the i.i.d. bounded assumption, the random variables are $1/2$-sub-Gaussian and it is immediate to show that $\{M_t^\lambda\}_{t\in\mathbb{N}}$ is a non-negative super-martingale that satisfies $\ln \mathbb{E}[M_t^\lambda] \leqslant 0$ for all t. It then follows that $M_\infty^\lambda = \lim_{t\to\infty} M_t^\lambda$ is almost surely well-defined and so, M_τ^λ as well. Further, let us introduce the stopped version $Q_t^\lambda = M^\lambda_{\min\{\tau,t\}}$. An application of Fatou's lemma shows that $\mathbb{E}[M_\tau^\lambda] = \mathbb{E}[\liminf_{t\to\infty} Q_t^\lambda] \leqslant \liminf_{t\to\infty} \mathbb{E}[Q_t^\lambda] \leqslant 1$. Thus, $\mathbb{E}[M_\tau^\lambda] \leqslant 1$.

The next step is to introduce the auxiliary variable $\Lambda = \mathcal{N}(0,4)$, independent of all other variables, and study the quantity $M_t = \mathbb{E}[M_t^\Lambda | \mathcal{F}_\infty]$. Note that the standard deviation of Λ is $(1/2)^{-1}$ due to the fact we consider $1/2$-sub-Gaussian random variables. We immediately get $\mathbb{E}[M_\tau] = \mathbb{E}[M_\tau^\Lambda | \Lambda]] \leqslant 1$. For

convenience, let $S_t = t(\mu_t - \mu)$. By construction of M_t, we have

$$\begin{aligned} M_t &= \frac{1}{\sqrt{8\pi}} \int_{\mathbb{R}} \exp\left(\lambda S_t - \frac{\lambda^2 t}{8} - \frac{\lambda^2}{8}\right) d\lambda \\ &= \frac{1}{\sqrt{8\pi}} \int_{\mathbb{R}} \exp\Big(- \Big[\lambda\sqrt{\frac{t+1}{8}} - \frac{\sqrt{2}S_t}{\sqrt{t+1}}\Big]^2 + \frac{2S_t^2}{t+1}\Big) d\lambda \\ &= \exp\left(\frac{2S_t^2}{t+1}\right) \frac{1}{\sqrt{8\pi}} \int_{\mathbb{R}} \exp\Big(- \lambda^2 \frac{t+1}{8}\Big) d\lambda \\ &= \exp\left(\frac{2S_t^2}{t+1}\right) \frac{\sqrt{8\pi/(t+1)}}{\sqrt{8\pi}} \,. \end{aligned}$$

Thus, we deduce that

$$S_t = \sqrt{\frac{t+1}{2} \ln\Big(\sqrt{t+1} M_t\Big)} \,.$$

We conclude by applying a simple Markov inequality:

$$\mathbb{P}\Bigg(\tau(\mu_\tau - \mu) \geqslant \sqrt{\frac{\tau+1}{2} \ln\Big(\sqrt{\tau+1}/\delta\Big)}\Bigg) = \mathbb{P}(M_\tau \geqslant 1/\delta) \leqslant \mathbb{E}[M_\tau]\delta \,.$$

□

Proceeding with the steps, more generally we obtain the following result for sums of sub-Gaussian random variables.

Lemma 10 *Let $Y_1, \dots Y_t$ be a sequence of t independent real-valued random variables where for each $s \leqslant t$, Y_s has mean μ_s and is σ_s-sub-Gaussian. Then for all $\delta \in (0,1)$, it holds*

$$\begin{aligned} \mathbb{P}\Bigg(\exists t \in \mathbb{N}, \quad \sum_{s=1}^t (Y_s - \mu_s) \geqslant \sqrt{2\sum_{s=1}^t \sigma_s^2 \Big(1 + \frac{1}{t}\Big) \ln\big(\sqrt{t+1}/\delta\big)}\Bigg) &\leqslant \delta \\ \mathbb{P}\Bigg(\exists t \in \mathbb{N}, \quad \sum_{s=1}^t (\mu_s - Y_s) \geqslant \sqrt{2\sum_{s=1}^t \sigma_s^2 \Big(1 + \frac{1}{t}\Big) \ln\big(\sqrt{t+1}/\delta\big)}\Bigg) &\leqslant \delta \,. \end{aligned}$$

4 Some other applications

We provide below some other applications of the basic concentration inequalities we derived earlier that we believe provide interesting insights.

4.1 Change of measure and code-length theory

For arbitrary random variable X admitting a finite cumulant generative function around 0, one has the properties that

$$\mathbb{P}\Big[X \geqslant \inf_{\lambda > 0} \Big\{ \frac{1}{\lambda} \log \mathbb{E} \exp(\lambda X) + \frac{\log(1/\delta)}{\lambda} \Big\}\Big] \leqslant \delta \tag{5}$$

$$\mathbb{P}\Big[X \leqslant \sup_{\lambda > 0} \Big\{ -\frac{1}{\lambda} \log \mathbb{E} \exp(-\lambda X) - \frac{\log(1/\delta)}{\lambda} \Big\}\Big] \leqslant \delta \,. \tag{6}$$

Importantly, note that in equations (5) and (6), the random variable X can be chosen to be a sum of any sequence of variables, with arbitrary dependency.

Now, let us consider some space $\mathcal{X}$ and two non-foreign distributions $P, Q \in \mathcal{M}(\mathcal{X})$ with density p, q. Then we have that

$$\mathbb{E}_P\left[\frac{q(X)}{p(X)}\right] = \int_{\mathcal{X}} q(x) dx = 1 \,.$$

Thus, we deduce from this simple change of measure that we have

$$\log \mathbb{E}_P \exp\Big(\log(q(X)) - \log(p(X)) \Big) = 0\,,$$

and in particular, this is less than 0, so that we deduce by Markov's inequality that for all $\delta \in [0,1]$ then

$$\mathbb{P}_P\Big[-\log(q(X)) \leqslant -\log(p(X)) - \log(1/\delta) \Big] \leqslant \delta\,,$$

which is precisely the core inequality of compression theory (using $\delta = e^{-K}$ for some number of bits K).

4.2 Chernoff Importance Sampling

Another way to use the previous construction is to consider importance sampling[1]. In many application of Importane sampling, people start by considering they want to estimate $\mathbb{E}_Q[f(X)]$ but we have samples $X_1, \dots, X_n$ sampled from P. The classical way is to reweight the values by P, namely to form

$$f_n^{P\to Q} = \frac{1}{n}\sum_{i=1}^n f(X_i)\frac{Q(X_i)}{P(X_i)} = \frac{1}{n}\sum_{i=1}^n \tilde{f}^{P\to Q}(X_i)$$

Since indeed $\mathbb{E}_P[f_n^{P\to Q}] = \mathbb{E}_Q[\frac{1}{n}\sum_{i=1}^n f(X_i)]$.

However, in many applications, one do not really care about $\mathbb{E}_Q[f(X)]$ but rather about the deviations of $\frac{1}{n}\sum_{i=1}^n f(X_i)$ around its mean, which is a very different questions. For that purpose, the estimate $f_n^{P\to Q}$ can turn out to be very bad since it classically suffers from a high variance.

A number of techniques have been suggested to reduce this variance. Here we directly tackle the control of the tail of $\frac{1}{n}\sum_{i=1}^n f(X_i)$, which classically requires a control of its log Laplace transform. More precisely:

Lemma 11 (Chernoff Importance Sampling) *Let P be a distribution on $\mathcal{X}$ and $X_1, \dots, X_n \overset{i.i.d.}{\sim} P$. Let $f : \mathcal{X} \to \mathbb{R}$ be some function taking real values. Let Q be another distribution on $\mathcal{X}$ such that the real random variable $f(X)$, for $X \sim Q$, is known to be R-sub-Gaussian. Let $\delta \in [0,1]$ be given, and form the quantity*

$$\tilde{f}_\delta^{P\to Q}(X_i) = f(X_i) + R\sqrt{\frac{n}{2\log(1/\delta)}}\log\left(\frac{Q(X_i)}{P(X_i)}\right).$$

Then, it holds for this precise δ and number of observations n,

$$\mathbb{P}_P\left[\frac{1}{n}\sum_{i=1}^n \tilde{f}_\delta^{P\to Q}(X_i) - \mathbb{E}_Q[f] \geqslant R\sqrt{\frac{2\log(1/\delta)}{n}}\right] \leqslant \delta\,.$$

This result shows that even though we do not directly have access to samples coming from Q, it is possible, knowing some concentration properties of Q, to reshape the empirical estimate of the mean in order to have a good control of the tails.

Proof of Lemma 11 :

Indeed, we want to control $\log \mathbb{E}_Q \exp[\lambda \sum_{i=1}^n (f(X_i) - \mathbb{E}_Q[f])]$. This can be written using a change of measure argument by

$$\log \mathbb{E}_P\left[\exp\big[\lambda\sum_{i=1}^n (f(X_i) - \mathbb{E}_Q[f])\big]\frac{Q(X_i)}{P(X_i)}\right] \quad = \quad \log \mathbb{E}_P \exp\left[\lambda\left[\sum_{i=1}^n f(X_i) + \frac{1}{\lambda}\log\left(\frac{Q(X_i)}{P(X_i)}\right) - \mathbb{E}_Q[f]\right]\right]$$

At this point, let us consider that the target distribution $f(X)$ for $X \sim Q$ is known to be, say R-sub-Gaussian. In this case, we know that what matters is to control the Legendre-Fenchel dual potential

[1] In the litterature, Chernoff Importance sampling refers to a different approach. However, there is no reason not to call the scheme considered in this section otherwise. An alternative name may be Chernoff-Laplace importance sampling.

function

$$\sup_{\lambda\in\mathbb{R}}\left(\lambda n\varepsilon - \log \mathbb{E}_Q \exp\left[\lambda\sum_{i=1}^n (f(X_i)-\mathbb{E}_Q[f])\right]\right) \;\geqslant\; \sup_{\lambda\in\mathbb{R}}\left(\lambda n\varepsilon - R^2\lambda^2 n/2\right) = n\varepsilon^2/2R^2,$$

where the equality its obtained for the value $\lambda^\star = \varepsilon/R^2$. This suggests to make the choice

$$\tilde{f}(X_i) = f(X_i) + \frac{R^2}{\varepsilon}\log\left(\frac{Q(X_i)}{P(X_i)}\right),$$

for some $\varepsilon > 0$. Indeed, one obtains that

$$\log \mathbb{E}_P \exp\left[\lambda^\star \sum_{i=1}^n (\tilde{f}(X_i)-\mathbb{E}_Q[f])\right] = \log \mathbb{E}_Q \exp\left[\lambda^\star \sum_{i=1}^n (f(X_i)-\mathbb{E}_Q[f])\right]$$

and thus for this specific value of ε, it holds

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \tilde{f}(X_i) \geqslant \mathbb{E}_Q[f]+\varepsilon\right] \leqslant \exp\left(-\frac{n\varepsilon^2}{2R^2}\right).$$

Alternatively, choosing $\varepsilon = R\sqrt{2\log(1/\delta)/n}$ for some $\delta\in[0,1]$, we obtain that with probability higher than $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^n \tilde{f}(X_i) - \mathbb{E}_Q[f] < R\sqrt{\frac{2\log(1/\delta)}{n}},$$

where

$$\tilde{f}(X_i) = f(X_i) + R\sqrt{\frac{n}{2\log(1/\delta)}}\log\left(\frac{Q(X_i)}{P(X_i)}\right).$$

Likewise, writing the reverse quantity

$$\log \mathbb{E}_P\left[\exp\left[\lambda\sum_{i=1}^n (\mathbb{E}_Q[f]-f(X_i))\right]\frac{Q(X_i)}{P(X_i)}\right] \;=\; \log \mathbb{E}_P \exp\left[\lambda\left[\sum_{i=1}^n \mathbb{E}_Q[f] + \frac{1}{\lambda}\log\left(\frac{Q(X_i)}{P(X_i)}\right) - f(X_i)\right]\right]$$

suggests to make the choice

$$\tilde{f}_-(X_i) = f(X_i) - \frac{R^2}{\varepsilon}\log\left(\frac{Q(X_i)}{P(X_i)}\right),$$

Indeed, one obtains that

$$\log \mathbb{E}_P \exp\left[\lambda^\star \sum_{i=1}^n (\mathbb{E}_Q[f]-\tilde{f}(X_i))\right] = \log \mathbb{E}_Q \exp\left[\lambda^\star \sum_{i=1}^n (\mathbb{E}_Q[f]-f(X_i))\right]$$

and then for the same ε as before, we get

$$\mathbb{P}_P\left[\frac{1}{n}\sum_{i=1}^n \mathbb{E}_Q[f] - \tilde{f}_{-,\delta}^{P\to Q}(X_i) \geqslant R\sqrt{\frac{2\log(1/\delta)}{n}}\right] \leqslant \delta.$$

□

The previous argument can be generalized, by following the construction of Lemma 7. This leads to the following result, that enables to transfer the concentration of measure from a single target distribution to a set of distributions that have been previously sampled. We present this version of the result without considering the random stopping time for simplicity of exposure.

Lemma 12 (Concentration inequality for predictable Processes under Covariate-Shift) *Let $\{X_i\}_{i=1}^{\infty}$ be a predictable sequence of random variables for a filtration $\mathcal{H}$, with known distributions $\{P_i\}_{i=1}^{\infty}$ such that $X_i|\mathcal{H}_{i-1} \sim P_i$. Let Q be a probability measure and $\varphi : \mathbb{R} \to \mathbb{R}^+$ be a convex upper-envelope of the cumulant generative function corresponding to Q, with $\varphi(0) = 0$, and $\varphi_\star$ its Legendre-Fenchel transform, that is:*

$$\begin{aligned} \forall \lambda \in \mathcal{D}, \qquad & \ln \mathbb{E}_Q\Big[\exp\Big(\lambda(X - \mathbb{E}_Q[X])\Big)\Big] \leqslant \varphi(\lambda)\,, \\ \forall x \in \mathbb{R} \qquad & \varphi_\star(x) = \sup_{\lambda \in \mathbb{R}} \big(\lambda x - \varphi(\lambda)\big), \end{aligned}$$

where $\mathcal{D} = \{\lambda \in \mathbb{R} : \ln \mathbb{E}\Big[\exp\Big(\lambda X\Big)\Big] < \infty\}$. Assume that $\mathcal{D}$ contains an open neighborhood of 0. Let $\delta \in [0,1]$ be given, and form the quantities

$$\begin{aligned} X_i^{+,P_i\to Q} &= X_i + \frac{1}{\lambda^+} \ln\left(\frac{Q(X_i)}{P_i(X_i)}\right) \qquad \textit{where} \quad \lambda^+ = \operatorname*{Argmax}_{\lambda \in \mathbb{R}^+} \Big[\lambda \varphi_{\star,+}^{-1}(\ln(1/\delta)/n) - \varphi(\lambda)\Big] \\ X_i^{-,P_i\to Q} &= X_i + \frac{1}{\lambda^-} \ln\left(\frac{Q(X_i)}{P_i(X_i)}\right) \qquad \textit{where} \quad \lambda^- = \operatorname*{Argmax}_{\lambda \in \mathbb{R}^-} \Big[\lambda \varphi_{\star,-}^{-1}(\ln(1/\delta)/n) - \varphi(\lambda)\Big] \end{aligned}$$

Now, for this specific choice of δ and a deterministic number of observations n, it holds

$$\begin{aligned} \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_{i,\delta}^{+,P_i\to Q} - \mathbb{E}_Q[X] \geqslant \varphi_{\star,+}^{-1}\left(\frac{\ln(1/\delta)}{n}\right)\right] &\leqslant \delta \\ \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_{i,\delta}^{-,P_i\to Q} - \mathbb{E}_Q[X] \leqslant \varphi_{\star,-}^{-1}\left(\frac{\ln(1/\delta)}{n}\right)\right] &\leqslant \delta \end{aligned}$$

Proof of Lemma 12:

1. Change of measure Let us introduce for convenience the notation $\tilde{Z}_i = X_i^{-,P_i\to Q} - \mathbb{E}_Q[X]$. Then, we want to control, for some $\varepsilon_n > 0$ to be defined later

$$\begin{aligned} \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \tilde{Z}_i \geqslant \varepsilon_n\Big) &\leqslant \mathbb{P}\Big(\exp\Big(\lambda\Big(\sum_{i=1}^n \tilde{Z}_i\Big)\Big) \geqslant \exp(\lambda n \varepsilon_n)\Big) \\ &\leqslant \mathbb{P}\Big(\underbrace{\exp\Big(\lambda\Big(\sum_{i=1}^n \tilde{Z}_i\Big) - n\varphi(\lambda_n)\Big)}_{W_n} \geqslant \exp\Big(n\big(\lambda_n\varepsilon_n - \varphi(\lambda_n)\big)\Big)\Big) \\ &\leqslant \mathbb{P}\Big(W_n \geqslant \exp\Big(n\big(\lambda_n\varepsilon_n - \varphi(\lambda_n)\big)\Big)\Big). \end{aligned}$$

Since $\varepsilon_n > 0$, we can choose a $\lambda_n > 0$ such that $\varphi^\star(\varepsilon_n) = \lambda_n\varepsilon_n - \varphi(\lambda_n)$.

2. Doob's maximal inequality At this, point, we show that the sequence $\{W_n\}_n$ is a non-negative super-martingale, where $W_n = \exp\Big(\lambda_n\big(\sum_{i=1}^n \tilde{Z}_i\big) - t\varphi(\lambda_n)\Big)$. Indeed, note that:

$$\begin{aligned} \mathbb{E}[W_{t+1}|\mathcal{F}_t] &= W_t \mathbb{E}[\exp\big(\lambda_n \tilde{Z}_{t+1}\big)|\mathcal{F}_t]\exp(-\varphi(\lambda_n)) \\ &= W_t \mathbb{E}\Big[\exp\big(\lambda_n(X_{t+1} + \frac{1}{\lambda_n}\ln\Big(\frac{Q(X_{t+1})}{P_{t+1}(X_{t+1})}\Big) - \mathbb{E}_Q[X])\big)|\mathcal{F}_t\Big]\exp(-\varphi(\lambda_n)) \\ &= W_t \mathbb{E}_Q\Big[\exp\big(\lambda_n(X - \mathbb{E}_Q[X])\big)\Big]\exp(-\varphi(\lambda_n)) \\ &\leqslant W_t\,. \end{aligned}$$

We thus find

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \tilde{Z}_i \geqslant \varepsilon_n\Big) &\leqslant \mathbb{P}\Big(W_n \geqslant \exp\Big(n\varphi^\star(\varepsilon_n)\Big)\Big)\\ &\overset{(a)}{\leqslant} \exp\Big(-n\varphi^\star(\varepsilon_n)\Big),\end{aligned}$$

where (a) holds by application of Doob's maximal inequality for non-negative super-martingales, using that $\mathbb{E}[W_0] \leqslant 1$.

---

3. Parameter tuning Now, let us choose ε_n such that $n\varphi_\star(\varepsilon_n) = c > 1$ is a constant, that is $\varepsilon_n = \varphi_{\star,+}^{-1}(c/n)$. Thus, we get for all $\eta \in (0, n-1)$:

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \tilde{Z}_i \geqslant \varepsilon_n\Big) \leqslant \exp\Big(-n\varphi^\star(\varepsilon_n)\Big) \leqslant \exp(-c)$$

We conclude by choosing $c = \log(1/\delta) > 1$.

---

5. Reverse bounds. Now for $\varepsilon_n < \mathbb{E}_Q[X]$, we can follow the same steps, but choose $\lambda_n > 0$ such that $\varphi^\star(\varepsilon_n) = -\lambda_n, \varepsilon_n - \varphi(-\lambda_n) \geqslant 0$. □

4.3 Transportation lemma

We conclude this section with a powerful result known as the transportation lemma.

Lemma 13 *For any function f, let us introduce $\varphi_f : \lambda \mapsto \log \mathbb{E}_P \exp(\lambda(f(X) - \mathbb{E}_P[f]))$. Whenever φ_f is defined on some possibly unbounded interval $0 \in I$, define its dual $\varphi_{\star,f}(x) = \sup_{\lambda\in I} \lambda x - \varphi_f(\lambda)$. Then it holds*

$$\begin{aligned}\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\leqslant \varphi_{+,f}^{-1}(KL(Q,P)) \quad \textit{where } \varphi_{+,f}^{-1}(t) = \inf\{x \geqslant 0 : \varphi_{\star,f}(x) > t\}\\ \forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\geqslant \varphi_{-,f}^{-1}(KL(Q,P)) \quad \textit{where } \varphi_{-,f}^{-1}(t) = \sup\{x \leqslant 0 : \varphi_{\star,f}(x) > t\}.\end{aligned}$$

Proof :

Let us recall the fundamental equality

$$\forall \lambda \in \mathbb{R}, \log \mathbb{E}_P \exp(\lambda(X - \mathbb{E}_P[X]) = \sup_{Q\ll P}\Big[\lambda\Big(\mathbb{E}_Q[X] - \mathbb{E}_P[X]\Big) - KL(Q,P)\Big].$$

In particular, we obtain on the one hand that

$$\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \leqslant \min_{\lambda\in\mathbb{R}^+} \frac{\varphi_f(\lambda) + KL(Q,P)}{\lambda}$$

Since $\varphi_f(0) = 0$, then the right hand side quantity is non-negative. Let us call it u. Then, we note that for any t such that $u \geqslant t \geqslant 0$, then by construction of u, it holds $KL(Q,P) \geqslant \varphi_{\star,f}(t)$. Thus, $\{t \geqslant 0 : \varphi_{f,\star}(t) \geqslant KL(Q,P)\} = (u, \infty)$ and thus $u = \varphi_{+,f}^{-1}(KL(Q,P))$.

On the other hand, it holds

$$\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \geqslant \max_{\lambda\in\mathbb{R}^-} \frac{\varphi_f(\lambda) + KL(Q,P)}{\lambda}$$

Since $\varphi(0) = 0$, then the right hand side quantity is non-positive. Let us call it v. Then, we note that for any t such that $v \leqslant t \leqslant 0$, then by construction of v, it holds $KL(Q,P) \geqslant \varphi_{\star,f}(t)$.Thus, $\{t \leqslant 0 : \varphi_{\star,f}(t) \geqslant KL(Q,P)\} = (-\infty, v)$ and thus $v = \varphi_{-,f}^{-1}(KL(Q,P))$.

□

Corollary 3 *Assume that f is such that $\mathbb{V}_P[f]$ and $\mathbb{S}(f) = \max_x f(x) - \min_x f(x)$ are finite. Then it holds*

$$\begin{aligned}\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\leqslant \sqrt{2\mathbb{V}_P[f]KL(Q,P)} + \frac{2\mathbb{S}(f)}{3}KL(Q,P)\,,\\ \forall Q \ll P, \quad \mathbb{E}_P[f] - \mathbb{E}_Q[f] &\leqslant \sqrt{2\mathbb{V}_P[f]KL(Q,P)}\,.\end{aligned}$$

In particular, this shows that it is enough to control the Kullback-Leibler divergence between a distribution and its empirical counter part in order to derive immediately a concentration result for the empirical mean of virtually any function (with finite variance and span).

Proof :

Indeed, by a standard Bernstein argument, it holds

$$\begin{aligned}\forall \lambda \in [0, \frac{3}{\mathbb{S}(f)}), \quad \varphi_f(\lambda) &\leqslant \frac{\mathbb{V}_P[f]}{2}\frac{\lambda^2}{1-\frac{\mathbb{S}(f)\lambda}{3}}\\ \forall x \geqslant 0, \quad \varphi_{\star,f}(x) &\geqslant \frac{x^2}{2(\mathbb{V}_P[f] + \frac{\mathbb{S}(f)}{3}x)}\end{aligned}$$

Then, a direct computation shows that

$$\varphi_{+,f}^{-1}(t) \leqslant \frac{\mathbb{S}(f)}{3}t + \sqrt{2t\mathbb{V}_P[f] + \Big(\frac{\mathbb{S}(f)}{3}t\Big)^2}\,.$$

$$\varphi_{-,f}^{-1}(t) \geqslant \frac{\mathbb{S}(f)}{3}t - \sqrt{2t\mathbb{V}_P[f] + \Big(\frac{\mathbb{S}(f)}{3}t\Big)^2}\,.$$

Combining these two bounds, we obtain that

$$\begin{aligned}\mathbb{E}_Q[f] - \mathbb{E}_P[f] &\leqslant \sqrt{2\mathbb{V}_P[f]KL(Q,P) + \Big(\frac{\mathbb{S}(f)}{3}\Big)^2 KL(Q,P)^2} + \frac{\mathbb{S}(f)}{3}KL(Q,P)\\ \mathbb{E}_P[f] - \mathbb{E}_Q[f] &\leqslant \sqrt{2\mathbb{V}_P[f]KL(Q,P) + \Big(\frac{\mathbb{S}(f)}{3}\Big)^2 KL(Q,P)^2} - \frac{\mathbb{S}(f)}{3}KL(Q,P)\,.\end{aligned}$$

□

Conclusion

In this short note, we have provided a few basic results for concentration of real-valued predictable processes. There are much more results out there regarding concentration of measure. This includes Sanov's concentration inequalities for the empirical distribution, concentration results for the variance leading to empirical Bernstein bounds, concentration of the cumulative distribution function, of the conditional value at risk, concentration inequalities for the median instead of the mean, or in terms of (inverse) Information projection. Beyond the real-valued random variables, one can consider concentration for vector-valued or matrix-valued martingales, and look at concentration in terms of various norms, for instance based on the Wasserstein or total variation distance.