Basic tail and concentration bounds (Part II)

Kaveh S. Nobari

Reading Sessions in High-Dimensional Statistics

Department of Mathematics and Statistics
Lancaster University

Contents

Motivation

- In the previous session we reviewed the bounds on sums of independent random variables that had been outlined by Wainwright (2019) and Vershynin (2018).
- In what follows we provide bounds on more general functions of random variables.
- A classical approach is based on martingale decomposition.

Consider the independent random variables $X_1, \cdots, X_n$ and consider a function $f(X) = f(X_1, \cdots, X_n)$ with the mapping $f : \mathbb{R}^n \to \mathbb{R}$. Suppose our goal is to obtain bounds on the deviations of f from its mean. To achieve this, let us consider the sequence of r.v.s given by $Y_0 = \mathbb{E}[f(X)]$, $Y_n = f(X)$, and

$$Y_k = \mathbb{E}[f(X) \mid X_1, \cdots, X_k] \quad k = 1, \cdots, n-1,$$

where Y_0 is a constant and the variables $Y_1, \cdots, Y_n$ tend to exhibit more fluctuations as they move along the sequence. Based on this intuition the martingale approach is based on the telescoping decomposition

$$f(X) - \mathbb{E}[f(X)] = Y_n - Y_0 = \sum_{i=1}^{n} \underbrace{Y_i - Y_{i-1}}_{D_i}$$

Thus, $f(X) - \mathbb{E}[f(X)]$ is expressed as the sum of increments $D_1, \cdots, D_n$. This is a specific example of a martingale sequence, most commonly referred to as Doob martingale, whereas $D_1, \cdots, D_n$ is a martingale difference sequence (MDS hereafter).

Example (Random Walk process): Let us consider a Random Walk process

$$x_t = x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \quad \text{for} \quad t = 1, \cdots, n$$

We know that $x_n = f(x_1, \cdots, x_{n-1})$, since using backward iteration, we may express the above expressions

$$\begin{aligned} x_n &= x_{n-1} + \varepsilon_n \\ &= x_{n-2} + \varepsilon_{n-1} + \varepsilon_n \\ &\vdots \\ x_n &= x_0 + \sum_{i=0}^{n-1} \varepsilon_{n-i} \end{aligned}$$

where x_0 is a constant. Hence,

$$\begin{aligned}\mathbb{E}[x_n] &= \mathbb{E}\left[x_0 + \sum_{i=0}^{n-1} \varepsilon_{n-i}\right] \\ &= x_0 + \sum_{i=1}^{n-1} \mathbb{E}[\varepsilon_{n-i}] \\ &= x_0\end{aligned}$$

Hence,

$$\begin{aligned}f(x) - \mathbb{E}[f(x)] &= x_n - \mathbb{E}[x_n] \\ &= x_n - x_0 \\ &= \sum_{i=1}^{n} \underbrace{X_i - X_{i-1}}_{\varepsilon_i}\end{aligned}$$

We now provide a quick recap of probability triples before providing definition for the next sections. For a quick recap see Williams (1991).

- A model for experiment involving randomness takes the form of a probability triple $(\Omega, \mathcal{F}, P)$.
- Ω is the sample space, where a point ω in Ω is a sample point.
- The σ-algebra $\mathcal{F}$ on Ω is called the family of events, so that an event is an element of $\mathcal{F}$, that is an $\mathcal{F}$-measurable subset of Ω.
- Finally, P is the probability measure on $(\Omega, \mathcal{F})$.

Example ($(\Omega, \mathcal{F})$ pairs): Toss a coin once and following Shreve (2004), let us define A_H as the set of all sequences beginning with Head or $H = \{\omega; \omega_1 = H\}$ and A_T as the set of all sequences beginning with Tail or $T = \{\omega; \omega_1 = T\}$. The sample space is thus,

$$\Omega = \{A_H, A_T\}$$

where the σ-field

$$\mathcal{F}_1 = \mathcal{P}(\Omega) = 2^{\Omega} := \{H, T, \emptyset, \Omega\}$$

is the σ-field spanned by one coin toss. Now we toss the coin twice. The sample space would then be

$$\Omega = \{A_{HH}, A_{TT}, A_{HT}, A_{TH}\}$$

and

$$\mathcal{F}_2 = \begin{Bmatrix} \Omega, \emptyset, A_H, A_T, A_{HH}, A_{TT}, A_{HT}, A_{TH}, A_{HH}^c, A_{TT}^c, A_{HT}^c, A_{TH}^c, \\ A_{HH} \cup A_{TH}, A_{HH} \cup A_{TT}, A_{HT} \cup A_{TH}, A_{HT} \cup A_{TT} \end{Bmatrix}$$

It is evident that $\mathcal{F}_1 \subset \mathcal{F}_2$, and in fact we may generalize this for infinite independent coin tosses to

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \cdots$$

We now provide a general definition of a martingale sequence by first defining a filtration as follows

Filtration

Let $\{\mathcal{F}_i\}_{i=1}^{\infty}$ be a sequence of σ-fields that are nested, meaning that $\mathcal{F}_m \subseteq \mathcal{F}_n$ for $n \geq m$. Such a sequence is known as a filtration.

In the Doob martingale described earlier, the σ-field $\sigma(X_1, \cdots, X_m)$ is spanned by the first m variables $X_1, \cdots, X_m$ and plays the role of $\mathcal{F}_m$. Let $\{Y_i\}_{i=1}^{\infty}$ be a sequence of random variables such that Y_i is measurable with respect to the σ-field $\mathcal{F}_i$. We say that $\{Y_i\}_{i=1}^{\infty}$ is adapted to the filtration $\{\mathcal{F}_i\}_{i=1}^{\infty}$.

Martingale

Given a sequence $\{Y_i\}_{i=1}^{\infty}$ of r.v.s adapted to a filtration $\{\mathcal{F}_i\}_{i=1}^{\infty}$, the pair $\{(Y_i, \mathcal{F}_i)\}_{i=1}^{\infty}$ is a martingale if, for all $i \geq 1$

$$\mathbb{E}[|Y_i|] < \infty \quad \text{and} \quad \mathbb{E}[Y_{i+1} \mid \mathcal{F}_i] = Y_i.$$

Example (Partial sums as martingales)

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d random variables with mean μ, and define the partial sums $S_m := \sum_{i=1}^{m} X_i$. Define $\mathcal{F}_m = \sigma(X_1, \cdots, X_m)$, the r.v. S_m is measurable w.r.t to $\mathcal{F}_m$, and, we have

$$\begin{aligned}
\mathbb{E}[S_{m+1} \mid \mathcal{F}_m] &= \mathbb{E}[X_{m+1} + S_m \mid X_1, \cdots, X_m] \\
&= \mathbb{E}[X_{m+1} \mid X_1, \cdots, X_m] + \mathbb{E}[S_m \mid X_1, \cdots, X_m] \\
&= \mathbb{E}[X_{m+1}] + S_m = \mu + S_m.
\end{aligned}$$

A closely related concept is that of the martingale difference sequence, which is an adapted sequence $\{(D_i, \mathcal{F}_i)\}_{i=1}^{\infty}$ such that, for all $i \geq 1$,

$$\mathbb{E}[|D_i|] < \infty \quad \text{and} \quad \mathbb{E}[D_{i+1} \mid \mathcal{F}_i] = 0.$$

Difference sequences arise naturally from martingales. Given a martingale $\{(Y_i, \mathcal{F}_i)\}_{i=0}^{\infty}$, define $D_i = Y_i - Y_{i-1}$ for $i \geq 1$. We then have

$$\begin{aligned}\mathbb{E}[D_{i+1} \mid \mathcal{F}_i] &= \mathbb{E}[Y_{i+1} - Y_i \mid \mathcal{F}_i] \\ &= \mathbb{E}[Y_{i+1} \mid \mathcal{F}_i] - Y_i \\ &= Y_i - Y_i = 0\end{aligned}$$

using the martingale property and the fact that Y_i is measurable w.r.t to $\mathcal{F}_i$. Thus, for any martingale sequence $\{Y_i\}_{i=0}^{n}$, we have the telescoping decomposition.

Telescoping decomposition

Let $\{D_i\}_{i=1}^{\infty}$ be a MDS. Then for any martingale sequence $\{Y_i\}_{i=0}^{\infty}$, we have the telescoping decomposition

$$Y_n - Y_0 = \sum_{i=1}^{n} D_i$$

Example (Doob construction)

Consider the sequence on independent random variables $X_1, \cdots, X_n$, recall the sequence $Y_k = \mathbb{E}[f(X) \mid X_1, \cdots, X_k]$ previously defined, and suppose that $\mathbb{E}[|f(X)|] < \infty$. We claim that $Y_0, \cdots, Y_n$ is a martingale w.r.t to $X_1, \cdots, X_n$. We have

$$\mathbb{E}[|Y_k|] = \mathbb{E}[|\mathbb{E}[f(X) \mid X_1, \cdots, X_k]|].$$

From Jensen's inequality, we have

$$\mathbb{E}[|\mathbb{E}[f(X) \mid X_1, \cdots, X_k]|] \leq \mathbb{E}[|f(X)|] < \infty.$$

From the 2nd property of martingales, we have

$$\mathbb{E}[Y_{k+1} \mid X_1^k] = \mathbb{E}[\mathbb{E}[f(X) \mid X_1^{k+1}] \mid X_1^k] = \mathbb{E}[f(X) \mid X_1^k] = Y_k$$

We now turn to the derivation of concentration inequalities for martingales, either

1) as bounds for the difference $Y_n - Y_0$; or
2) as bounds for the sum $\sum_{i=1}^n D_i$ of the associated MDS.

We begin by stating and proving a general Bernstein-type bound for a MDS, based on imposing a sub-exponential condition on the martingale differences. To do so, we adopt the standard approach of controlling the mgf of $\sum_{i=1}^n D_i$ and then applying the Chernoff bound.

Let $\{(D_i, \mathcal{F}_i)\}_{i=1}^\infty$ be a MDS and suppose that $\mathbb{E}[\exp(\lambda D_i) \mid \mathcal{F}_{i-1}] \leq \exp\left(\frac{\lambda^2 \nu_i^2}{2}\right)$ a.s. for any $|\lambda| < \frac{1}{\alpha_i}$

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n D_i\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n D_i\right) \,\middle|\, \mathcal{F}_{n-1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp(\lambda D_n) \exp\left(\lambda \sum_{i=1}^{n-1} D_i\right) \,\middle|\, \mathcal{F}_{n-1}\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1} D_i\right)\mathbb{E}\left[\exp\left(\lambda D_n\right)\,\middle|\,\mathcal{F}_{n-1}\right]\right] \\
&\leq \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1} D_i\right)\right]\exp\left(\frac{\lambda^2\nu_n^2}{2}\right)
\end{aligned}
$$

we may iterate this procedure again for $\mathbb{E}[\exp(\lambda\sum_{i=1}^{n-1} D_i)]$ and we'd obtain,

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1} D_i\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-1} D_i\right)\,\middle|\,\mathcal{F}_{n-2}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda D_{n-1}\right)\exp\left(\lambda\sum_{i=1}^{n-2} D_i\right)\,\middle|\,\mathcal{F}_{n-2}\right]\right] \\
&= \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n-2} D_i\right)\mathbb{E}\left[\exp\left(\lambda D_{n-1}\right)\,\middle|\,\mathcal{F}_{n-2}\right]\right]
\end{aligned}
$$

$$\leq \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-2} D_i\right)\right] \exp\left(\frac{\lambda^2 \nu_{n-1}^2}{2}\right)$$

Continuously iterating this process yields,

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} D_i\right)\right] \leq \exp\left(\frac{\lambda^2 \sum_{i=1}^{n} \nu_i^2}{2}\right),$$

valid for all $|\lambda| < \frac{1}{\alpha^*}$. Hence, by definition, it can be concluded that $\sum_{i=1}^{n} D_i$ is sub-exponential with parameters $(\sqrt{\sum_{i=1}^{n} \nu_i^2}, \alpha^*)$. The tail bounds can be derived by using the Chernoff-type approach as before. In other words we are interested in

$$P\left[\sum_{i=1}^{n} D_i \geq t\right] = P\left[\exp\left(\lambda \sum_{i=1}^{n} D_i\right) \geq \exp(\lambda t)\right] \leq \frac{\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} D_i\right)\right]}{\exp(\lambda t)}$$

where from the definition of sub-exponential variables and the earlier results, we know that

$$P\left[\exp\left(\lambda\sum_{i=1}^{n}D_i\right)\geq\exp\left(\lambda t\right)\right]\leq\exp\left(\frac{\lambda^2\sum_{i=1}^{n}\nu_i^2}{2}-\lambda t\right),\quad\forall\lambda\in\left[0,\frac{1}{\alpha^*}\right)$$

where the Chernoff optimisation problem is

$$\log P\left[\sum_{i=1}^{n}D_i\geq t\right]\leq\inf_{\lambda\in[0,\alpha_*^{-1}]}\left\{\underbrace{\frac{\lambda^2\sum_{i=1}^{n}\nu_i^2}{2}-\lambda t}_{g(\lambda,t)}\right\}.$$

To complete the proof, it remains to compute for each $t\geq 0$, the quantity $g^*(t):=\inf_{\lambda\in[0,\alpha^{-1})}g(\lambda,t)$, where using the same unconstrained optimisation approach as for the sub-Gaussian variables, we'd obtain $\lambda_{opt}=\frac{t}{\sum_{i=1}^{n}\nu_i^2}$ as the unconstrained minimum of the function $g(.,t)$, which yields the minimum $-\frac{t^2}{2\sum_{i=1}^{n}\nu_i^2}$.

Recall the constraint $0 \leq \lambda < \frac{1}{\alpha^*}$. This implies that the unconstrained optimal λ_{opt} must be between $0 \leq \frac{t}{\sum_{i=1}^n \nu_i^2} < \frac{1}{\alpha^*}$, which implies that in the interval $0 \leq t < \frac{\sum_{i=1}^n \nu_i^2}{\alpha_*}$, the unconstrained optimum corresponds to the constrained optimum.
Otherwise for $t \geq \frac{\sum_{i=1}^n \nu_i^2}{\alpha_*^2}$, considering that the function $g(., t) = \frac{\lambda^2 \sum_{i=1}^n \nu_i^2}{2} - \lambda t$ is monotonically decreasing, in the interval $[0, \lambda_{opt})$, the constrained minimum is obtained at the boundary - i.e. $\lambda^{\#} = \frac{1}{\alpha}$, which leads to the minimum

$$g^*(t) = g(\lambda^{\#}, t) = -\frac{t}{\alpha^*} + \frac{1}{2\alpha^*} \frac{\sum_{i=1}^n \nu_i^2}{\alpha^*} \leq -\frac{t}{2\alpha^*}$$

where this inequality used the fact that $\frac{\sum_{i=1}^n \nu_i^2}{\alpha} \leq t$, which leads to the following

Concentration inequalities for MDS

Let $\{(D_i, \mathcal{F}_i)\}_{i=1}^{\infty}$ be a martingale difference sequence and suppose that $\mathbb{E}[\exp(\lambda D_i) \mid \mathcal{F}_{i-1}] \leq \frac{\lambda^2 \nu_i^2}{2}$ a.s. for any $|\lambda| < \frac{1}{\alpha}$. Then the following hold

- The sum $\sum_{i=1}^n D_i$ is sub-exponential with parameters $\left(\sqrt{\sum_{i=1}^n \nu_i^2}, \alpha^*\right)$, where $\alpha^* := \max_{i=1,\cdots,n} \alpha_i$.
- The sum satisfies the concentration inequality

$$P\left[\left|\sum_{i=1}^n D_i\right| \geq t\right] \leq \begin{cases} 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \nu_i^2}\right), & 0 \leq t \leq \frac{\sum_{i=1}^n \nu_i^2}{\alpha^*} \\ 2\exp\left(-\frac{t}{2\alpha^*}\right), & t > \frac{\sum_{i=1}^n \nu_i^2}{\alpha^*} \end{cases}$$

For the concentration inequalities to be useful in practice, we must isolate sufficient easily checkable conditions for the differences D_i to be a.s. sub-exponential (or sub-Gaussian when $\alpha = 0$). As mentioned earlier, bounded r.v.s are sub-Gaussian, which leads to the following corollary

Azuma-Hoeffding

Let $\{(D_i, \mathcal{F}_i)\}_{i=1}^{\infty}$ be a MDS for which there are constants $\{(a_i, b_i)\}_{i=1}^{n}$ such that $D_i \in [a_i, b_i]$ a.s. for all $k = 1, \cdots, n$. Then for all $t \geq 0$

$$P\left[\left|\sum_{i=1}^{n} D_i\right| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Proof: All that needs showing is that the $\mathbb{E}[\exp(\lambda D_i \mid \mathcal{F}_{i-1})] \leq \exp\left(\frac{\lambda^2(b_i-a_i)^2}{8}\right)$ a.s. for each $i = 1, \cdots, n$. But since $D_i \in [a_i, b_i]$ a.s., the conditioned variables $(D_i \mid \mathcal{F}_{i-1})$ also belongs to this interval a.s.

Bounded differences property

Given vectors $x, x' \in \mathbb{R}^n$ and an index $k \in \{1, 2, \cdots, n\}$, define the vector $\{x^{\setminus k} \in \mathbb{R}^n\}$ via

$$x^{\setminus k} := (x_1, x_2, \cdots, x_{k-1}, x'_k, x_{k+1}, \cdots, x_n)'.$$

We say that $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference property with parameters $(L_1, \cdots, L_n)$ if, for each $k = 1, 2, \cdots, n$,

$$|f(x) - f(x^{\setminus k})| \leq L_k \quad \forall x, x' \in \mathbb{R}^n$$

Bounded differences inequality

Suppose that f satisfies the bounded difference property with parameters $(L_1, \cdots, L_n)$ and that the random vector $X = (X_1, X_2, \cdots, X_n)'$ has independent components. Then

$$P[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n L_k^2}\right), \quad \forall t \geq 0$$

Example

Say we have the bounded r.v.s $X_i \in [a, b]$ almost surely, and consider the function $f(x_1, \cdots, x_n) = \sum_{i=1}^{n}(x_i - \mu_i)$, where $\mu_i = \mathbb{E}[X_i]$ is the mean of the i^{th} rv. For any index $l \in \{1, \cdots, n\}$, we have

$$\begin{aligned} |f(x) - f(x^{\backslash k})| &= |(x_k - \mu_k) - (x_k' - \mu_k)| \\ &= |x_k - x_k'| \leq b - a \end{aligned}$$

which shows that f satisfies the bounded difference inequality in each coordinate with parameter $L = b - a$. Consequently, from the bounded inequality it follows

$$P\left[|\sum_{i=1}^{n}(x_i - \mu_i)| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right)$$

which is classical Hoeffding bound for independent r.v.s.

Consider a Gaussian random variable $X \sim N(0, I_n)$ and a function $f : \mathbb{R}^n \to \mathbb{R}$. When does the random vector $f(X)$ concentrate about its mean, i.e.,

$$f(X) \approx \mathbb{E}f(X)$$

with high probability?

In the case of linear functions f this question is easy, where $f(X)$ has a normal distribution, and it concentrates around its mean well. However, we must also consider the the case of non-linear functions $f(X)$ of random vectors X. We cannot expect to have good concentration for completely arbitrary f. However, if f does not oscillate too wildly, we might expect concentration. The concept of Lipschitz functions will help us to rule out functions that have wild oscillations.

L-Lipschitz functions

We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz w.r.t the Euclidean norm $\|.\|_2$ if

$$|f(x) - f(y)| \leq L\|x - y\|_2, \quad \forall x, y \in \mathbb{R}^n$$

In other words, Lipschitz functions may not blow up distance between points too much. The following guarantees that any such function is sub-Gaussian with parameter at most L:

Let $(X_1, \cdots, X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz w.r.t the Euclidean norm. Then the variable $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, and hence

$$P[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{t^2}{2L^2}\right), \quad \forall t \geq 0$$

The earlier result is of great importance, as it guarantees that any L-Lipschitz function of a standard Gaussian random vector, regardless of the dimension, exhibits concentration like a scalar Gaussian variable with variance L^2.

Any Lipschitz function is differentiable almost everywhere and the Lipschitz property further guarantees $\|\nabla f(x)\|_2 \leq L$ for all $x \in \mathbb{R}^n$. Therefore, to prove the earlier results, we first begin by providing the following Lemma:

Lemma

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then for any convex function $\phi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}_X[\phi(f(X) - \mathbb{E}(f(X)))] \leq \mathbb{E}_{X,Y}\left[\phi\left(\frac{\pi}{2}\langle \nabla f(X), Y\rangle\right)\right]$$

where $X, Y \sim N(0, I_n)$ are standard multivariate and independent.

Proof: For any fixed $\lambda \in \mathbb{R}$ applying the inequality in above Lemma to the convex function $f : t \to \exp(\lambda t)$ yields

$$\mathbb{E}_X[\exp(\lambda\{f(X) - \mathbb{E}[f(X)]\})] \leq \mathbb{E}_{X,Y}\left[\exp\left(\frac{\pi}{2}\langle \nabla f(X), Y\rangle\right)\right]$$

References

Shreve, S. E. (2004). *Stochastic calculus for finance II: Continuous-time models*, volume 11. Springer Science & Business Media.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Williams, D. (1991). *Probability with martingales*. Cambridge university press.