Notes 7 : Concentration inequalities

Math 733-734: Theory of Probability *Lecturer: Sebastien Roch*

References: [Roc, Sections 2.4].

Recall:

THM 7.1 (Markov's inequality) *Let X be a non-negative random variable. Then, for all $b > 0$,*

$$\mathbb{P}[X \geq b] \leq \frac{\mathbb{E}X}{b}. \tag{1}$$

Proof:

$$\mathbb{E}X \geq \mathbb{E}[X; X \geq b] \geq \mathbb{E}[b; X \geq b] = b\,\mathbb{P}[X \geq b].$$

■

THM 7.2 (Chebyshev's inequality) *Let X be a random variable with $\mathbb{E}X^2 < +\infty$. Then, for all $\beta > 0$,*

$$\mathbb{P}[|X - \mathbb{E}X| > \beta] \leq \frac{\mathrm{Var}[X]}{\beta^2}. \tag{2}$$

Proof: This follows immediately by applying (1) to $|X - \mathbb{E}X|^2$ with $b = \beta^2$. ■

We will also need the following lemma. In words, conditioning on an independent RV is equivalent to fixing its value. Note that independence plays a crucial role here. See Example 5.1.5 in [D] for a proof.

LEM 7.3 (Conditioning on an independent RV) *Suppose X and Y are independent. Let ϕ be a function with $\mathbb{E}|\phi(X,Y)| < +\infty$ and let $g(x) = \mathbb{E}(\phi(x,Y))$. Then,*

$$\mathbb{E}(\phi(X,Y)|X) = g(X).$$

1 Chernoff-Cramér method

Chebyshev's inequality (THM 7.2) gives a bound on the concentration around the mean of a square integrable random variable that is, in general, best possible. Indeed take X to be $\mu + b\sigma$ or $\mu - b\sigma$ with probability $(2b^2)^{-1}$ respectively, and μ

otherwise. Then $\mathbb{E}X = \mu$, $\mathrm{Var}X = \sigma^2$, and for $\beta = b\sigma$,

$$\mathbb{P}[|X - \mathbb{E}X| \geq \beta] = \mathbb{P}[|X - \mathbb{E}X| = \beta] = \frac{1}{b^2} = \frac{\mathrm{Var}X}{\beta^2}.$$

However, in many cases, much stronger bounds can be derived.

In this section we discuss the Chernoff-Cramér method, which produces *exponential* tail inequalities, provided the moment-generating function is finite in a neighborhood of 0.

DEF 7.4 (Moment-generating function) *The* moment-generating function *of X is the function*

$$M_X(s) = \mathbb{E}\left[e^{sX}\right],$$

defined for all $s \in \mathbb{R}$ where it is finite, which includes at least $s = 0$.

1.1 Tail bounds via the moment-generating function

We derive a general tail inequality first and then illustrate it on several standard cases.

Chernoff-Cramér bound Under a finite variance, squaring within Markov's inequality (THM 7.1) produces Chebyshev's inequality (THM 7.2). This "boosting" can be pushed further when stronger integrability conditions hold.

THM 7.5 (Chernoff-Cramér bound) *Assume X is a centered random variable such that $M_X(s) < +\infty$ for $s \in (-s_0, s_0)$ for some $s_0 > 0$. For any $\beta > 0$ and $s > 0$,*

$$\mathbb{P}[X \geq \beta] \leq \exp\left[-\left\{s\beta - \Psi_X(s)\right\}\right], \tag{3}$$

where

$$\Psi_X(s) = \log M_X(s),$$

is the cumulant-generating function of X.

Proof: Exponentiating within Markov's inequality gives, for any $\beta > 0$ and $s > 0$,

$$\mathbb{P}[X \geq \beta] = \mathbb{P}[e^{sX} \geq e^{s\beta}] \leq \frac{M_X(s)}{e^{s\beta}} = \exp\left[-\left\{s\beta - \Psi_X(s)\right\}\right].$$

■

EX 7.6 (Gaussian random variables) *Let $X \sim N(0,\nu)$ where $\nu > 0$ is the variance. By Chebyshev's inequality (THM 7.2)*

$$\mathbb{P}[|X| \geq \beta] \leq \frac{\nu}{\beta^2}. \tag{4}$$

On the other hand, note that

$$\begin{aligned} M_X(s) &= \int_{-\infty}^{+\infty} e^{sx} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{x^2}{2\nu}}\,\mathrm{d}x \\ &= \int_{-\infty}^{+\infty} e^{\frac{s^2\nu}{2}} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{(x-s\nu)^2}{2\nu}}\,\mathrm{d}x \\ &= \exp\left(\frac{s^2\nu}{2}\right), \end{aligned}$$

so that straightforward calculus gives for $\beta > 0$

$$\sup_{s>0}(s\beta - s^2\nu/2) = \frac{\beta^2}{2\nu}, \tag{5}$$

achieved at $s_\beta = \beta/\nu$. Plugging this into (3) *leads for $\beta > 0$ to the bound*

$$\mathbb{P}[X \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right). \tag{6}$$

By symmetry

$$\mathbb{P}[|X| \geq \beta] \leq 2\exp\left(-\frac{\beta^2}{2\nu}\right).$$

which is much sharper than Chebyshev's inequality for large β—compare to (4)*.*

As a second illustration, we consider simple random walk on $\mathbb{Z}$.

THM 7.7 (Chernoff bound for simple random walk on $\mathbb{Z}$) *Let $Z_1, \ldots, Z_n$ be independent $\{-1,1\}$-valued random variables with $\mathbb{P}[Z_i = 1] = \mathbb{P}[Z_i = -1] = 1/2$. Let $S_n = \sum_{i\leq n} Z_i$. Then, for any $\beta > 0$,*

$$\mathbb{P}[S_n \geq \beta] \leq e^{-\beta^2/2n}.$$

Proof: The moment-generating function of Z_1 can be bounded as follows

$$M_{Z_1}(s) = \frac{e^s + e^{-s}}{2} = \sum_{j\geq 0} \frac{s^{2j}}{(2j)!} \leq \sum_{j\geq 0} \frac{(s^2/2)^j}{j!} = e^{s^2/2}. \tag{7}$$

Taking $s = \beta/n$ in the Chernoff-Cramér bound (3), we get

$$\begin{aligned}\mathbb{P}[S_n \geq \beta] &\leq \exp\left(-s\beta + n\Psi_{Z_1}(s)\right)\\ &\leq \exp\left(-s\beta + ns^2/2\right)\\ &= e^{-\beta^2/2n},\end{aligned}$$

which concludes the proof. ■

For any $0 < s < s_0$, the Chernoff-Cramér bound (3) already leads to an exponential concentration bound on X. A better bound may be obtained by optimizing the choice of s. The Chernoff-Cramér method is particularly useful for sums of independent random variables as the moment-generating function then factorizes. Let

$$\Psi_X^*(\beta) = \sup_{s\in\mathbb{R}_+} (s\beta - \Psi_X(s)),$$

be the *Fenchel-Legendre dual* of the cumulant-generating function of X.

THM 7.8 (Chernoff-Cramér method for sums of IID RVs) *Let $S_n = \sum_{i\leq n} X_i$, where the X_is are i.i.d. centered random variables. Assume $M_{X_1}(s) < +\infty$ on $s \in (-s_0, s_0)$ for some $s_0 > 0$. For any $\beta > 0$,*

$$\mathbb{P}[S_n \geq \beta] \leq \exp\left(-n\Psi_{X_1}^*\left(\frac{\beta}{n}\right)\right). \tag{8}$$

In particular, in the large deviations regime, i.e., when $\beta = bn$ for some $b > 0$, we have

$$-\limsup_n \frac{1}{n}\log\mathbb{P}[S_n \geq bn] \geq \Psi_{X_1}^*(b). \tag{9}$$

Proof: Observe that

$$\Psi_{S_n}^*(\beta) = \sup_{s>0}(s\beta - n\Psi_{X_1}(s)) = \sup_{s>0} n\left(s\left(\frac{\beta}{n}\right) - \Psi_{X_1}(s)\right) = n\Psi_{X_1}^*\left(\frac{\beta}{n}\right),$$

and optimize over s in (3). ■

1.2 Binomial case

Let Z be a binomial random variable with parameters n and p. Recall that Z is a sum of i.i.d. indicators $Y_1, \ldots, Y_n$ and, letting $X_i = Y_i - p$ and $S_n = Z - np$,

$$\Psi_{X_1}(s) = \log\left(pe^s + (1-p)\right) - ps.$$

For $b \in (0, 1-p)$, letting $a = b + p$, direct calculation gives

$$\begin{aligned}\Psi^*_{X_1}(b) &= \sup_{s>0}(sb - (\log[pe^s + (1-p)] - ps)) \\ &= (1-a)\log\frac{1-a}{1-p} + a\log\frac{a}{p} =: D(a\|p), \qquad (10)\end{aligned}$$

achieved at $s_b = \log\frac{(1-p)a}{p(1-a)}$. The function $D(a\|p)$ in (10) is the so-called *Kullback-Leibler divergence* or *relative entropy* of Bernoulli variables with parameters a and p respectively. By (8) for $\beta > 0$

$$\mathbb{P}[Z \geq np + \beta] \leq \exp\left(-n\,D\left(p + \beta/n\|p\right)\right).$$

Applying the same argument to $Z' = n - Z$ gives a bound in the other direction.

In the large deviations regime, it can be shown that the previous bound is tight in the sense that

$$-\frac{1}{n}\log\mathbb{P}[Z \geq np + bn] \to D\left(p + b\|p\right) =: I^{\mathrm{Bin}}_{n,p}(b),$$

as $n \to +\infty$. See e.g. [Dur10, Section 2.6].

2 Sub-Gaussian random variables

The bounds described above were obtained by computing the moment-generating function explicitly. This is seldom possible. In this section, we give some important examples of concentration inequalities derived from the Chernoff-Cramér method for broad classes of random variables under natural conditions on their distributions.

Sub-Gaussian random variables We say that a centered random variable X is *sub-Gaussian with variance factor* $\nu > 0$ if for all $s \in \mathbb{R}$

$$\Psi_X(s) \leq \frac{s^2\nu}{2},$$

which is denoted by $X \in \mathcal{G}(\nu)$. Note that the r.h.s. is the cumulant-generating function of a $\mathrm{N}(0,\nu)$. By the Chernoff-Cramér method and (5) it follows immediately that

$$\mathbb{P}[X \leq -\beta] \vee \mathbb{P}[X \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right), \qquad (11)$$

where we used that $X \in \mathcal{G}(\nu)$ implies $-X \in \mathcal{G}(\nu)$. When considering linear combinations of independent sub-Gaussian random variables, we get the following.

THM 7.9 (General Hoeffding inequality) *Let $X_1, \ldots, X_n$ be independent centered random variables where, for each i, $X_i \in \mathcal{G}(\nu_i)$ with $0 < \nu_i < +\infty$ and let $(a_1, \ldots, a_n) \in \mathbb{R}^n$. Let $S_n = \sum_{i \leq n} a_i X_i$. Then $S_n \in \mathcal{G}(\sum_{i=1}^n a_i^2 \nu_i)$. In particular, for all $\beta > 0$,*

$$\mathbb{P}\left[S_n \geq \beta\right] \leq \exp\left(-\frac{\beta^2}{2\sum_{i=1}^n a_i^2 \nu_i}\right).$$

Proof: By independence,

$$\Psi_{S_n}(s) = \sum_{i \leq n} \Psi_{a_i X_i}(s) = \sum_{i \leq n} \Psi_{X_i}(s a_i) \leq \sum_{i \leq n} \frac{(s a_i)^2 \nu_i}{2} = \frac{s^2 \sum_{i \leq n} a_i^2 \nu_i}{2}.$$

■

EX 7.10 *The proof of THM 7.7 shows that uniforms on $\{-1, +1\}$ are sub-Gaussian with variance factor 1.*

Hoeffding's inequality For bounded random variables, the previous inequality gives the following useful bound.

THM 7.11 (Hoeffding's inequality) *Let $X_1, \ldots, X_n$ be independent random variables where, for each i, X_i takes values in $[a_i, b_i]$ with $-\infty < a_i \leq b_i < +\infty$. Let $S_n = \sum_{i \leq n} X_i$. For all $\beta > 0$,*

$$\mathbb{P}[S_n - \mathbb{E}S_n \geq \beta] \leq \exp\left(-\frac{2\beta^2}{\sum_{i \leq n}(b_i - a_i)^2}\right).$$

By Theorem 7.9, it suffices to show that $X_i - \mathbb{E}X_i \in \mathcal{G}(\nu_i)$ with $\nu_i = \frac{1}{4}(b_i - a_i)^2$. We first give a quick proof of a weaker bound that uses a trick called *symmetrization*. (Note: This proof uses conditioning, which will be covered later in the course, and therefore can be skipped.) Suppose the X_is are centered and satisfy $|X_i| \leq c_i$ for some $c_i > 0$. Let X_i' be an independent copy of X_i and let Z_i be an independent uniform random variable in $\{-1, 1\}$. For any s, by Jensen's inequality and (7),

$$\begin{aligned}
\mathbb{E}\left[e^{sX_i}\right] &= \mathbb{E}\left[e^{s\mathbb{E}[X_i - X_i' \,|\, X_i]}\right] \\
&\le \mathbb{E}\left[\mathbb{E}\left[e^{s(X_i - X_i')} \,\middle|\, X_i\right]\right] \\
&= \mathbb{E}\left[e^{s(X_i - X_i')}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e^{s(X_i - X_i')} \,\middle|\, Z_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e^{sZ_i(X_i - X_i')} \,\middle|\, Z_i\right]\right] \\
&= \mathbb{E}\left[e^{sZ_i(X_i - X_i')}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e^{sZ_i(X_i - X_i')} \,\middle|\, X_i - X_i'\right]\right] \\
&\le \mathbb{E}\left[e^{(s(X_i - X_i'))^2/2}\right] \\
&\le e^{-4c_i^2 s^2/2},
\end{aligned}$$

where on the fifth line we used the fact that $X_i - X_i'$ is identically distributed to $-(X_i - X_i')$ and that Z_i is independent of both X_i and X_i' (together with LEM 7.3). That is, X_i is sub-Gaussian with variance factor $4c_i^2$. By THM 7.9, S_n is sub-Gaussian with variance factor $\sum_{i\le n} 4c_i^2$ and

$$\mathbb{P}[S_n \ge t] \le \exp\left(-\frac{t^2}{8\sum_{i\le n} c_i^2}\right).$$

Proof:[Proof of THM 7.11] As pointed out above, it suffices to show that $X_i - \mathbb{E}X_i$ is sub-Gaussian with variance factor $\frac{1}{4}(b_i - a_i)^2$. This is the content of Hoeffding's lemma. First an observation:

LEM 7.12 (Variance of bounded random variables) *For any random variable Z taking values in $[a, b]$ with $-\infty < a \le b < +\infty$, we have*

$$\mathrm{Var}[Z] \le \frac{1}{4}(b - a)^2.$$

Proof: Indeed

$$\left|Z - \frac{a+b}{2}\right| \le \frac{b-a}{2},$$

and

$$\mathrm{Var}[Z] = \mathrm{Var}\left[Z - \frac{a+b}{2}\right] \le \mathbb{E}\left[\left(Z - \frac{a+b}{2}\right)^2\right] \le \left(\frac{b-a}{2}\right)^2.$$

■

LEM 7.13 (Hoeffding's lemma) *Let X be a random variable taking values in $[a,b]$ for $-\infty < a \leq b < +\infty$. Then $X - \mathbb{E}X \in \mathcal{G}\left(\frac{1}{4}(b-a)^2\right)$.*

Proof: Note first that $X - \mathbb{E}X \in [a - \mathbb{E}X, b - \mathbb{E}X]$ and $\frac{1}{4}((b-\mathbb{E}X)-(a-\mathbb{E}X))^2 = \frac{1}{4}(b-a)^2$. So w.l.o.g. we assume $\mathbb{E}X = 0$. Because X is bounded, $M_X(s)$ is finite for all $s \in \mathbb{R}$. From standard results on moment-generating functions (e.g., [Bil95, Section 21]; see also [Dur10, Theorem A.5.1]), for any $k \in \mathbb{Z}$,

$$M_X^{(k)}(s) = \mathbb{E}\left[X^k e^{sX}\right].$$

Hence

$$\Psi_X(0) = \log M_X(0) = 0, \qquad \Psi_X'(0) = \frac{M_X'(0)}{M_X(0)} = \mathbb{E}X = 0,$$

and by a Taylor expansion

$$\Psi_X(s) = \Psi_X(0) + s\Psi_X'(0) + \frac{s^2}{2}\Psi_X''(s^*) = \frac{s^2}{2}\Psi_X''(s^*),$$

for some $s^* \in [0,s]$. Therefore it suffices to show that for all s

$$\Psi_X''(s) \leq \frac{1}{4}(b-a)^2. \tag{12}$$

Note that

$$\begin{aligned}
\Psi_X''(s) &= \frac{M_X''(s)}{M_X(s)} - \left(\frac{M_X'(s)}{M_X(s)}\right)^2 \\
&= \frac{1}{M_X(s)}\mathbb{E}\left[X^2 e^{sX}\right] - \left(\frac{1}{M_X(s)}\mathbb{E}\left[X e^{sX}\right]\right)^2 \\
&= \mathbb{E}\left[X^2 \frac{e^{sX}}{M_X(s)}\right] - \left(\mathbb{E}\left[X \frac{e^{sX}}{M_X(s)}\right]\right)^2.
\end{aligned}$$

The trick to conclude is to notice that $\frac{e^{sx}}{M_X(s)}$ defines a density (i.e., a Radon-Nikodym derivative) on $[a,b]$ with respect to the law of X. The variance under this density—the last line above—is less than $\frac{1}{4}(b-a)^2$ by LEM 7.12. This establishes (12) and concludes the proof. ■

■

3 Epsilon-net arguments

Exponential tail inequalities are useful, among other things, to study the deviations (or expectations) of suprema of random variables. When the supremum is over an infinite index set, one way to proceed is to apply a tail inequality to a sufficiently dense finite subset of the index set, and then extend the resulting bound by continuity. This is referred to as an *ε-net argument* and it is easier to understand on an example. In this section, we show how to use an ε-net argument to bound the spectral norm of a random matrix with independent, sub-Gaussian entries.

ε-nets But, first, a few general definitions. To elaborate on the last point, which is known as an *ε-net argument*, we make the following definition.

DEF 7.14 (ε-net) *Let S be a subset of a metric space (M, ρ) and let $\varepsilon > 0$. The collection of points $N \subseteq S$ is called an* ε-net *of S if all pairs of points in N are at distance greater than ε and N is maximal by inclusion in S. In particular for all $z \in S$, $\inf_{y \in N} \rho(z, y) \leq \varepsilon$. The* covering number *of S, denoted by $\mathcal{N}(S, \rho, \varepsilon)$, is the smallest cardinality of an ε-net of S.*

The definition of an ε-net immediately suggests an algorithm for constructing one. Start with $N = \emptyset$ and successively add a point to N at distance at least ε from all other previous points until that is not possible to do so anymore. (Provided S is compact, this procedure will terminate after a finite number of steps.)

EX 7.15 (Sphere in $\mathbb{R}^k$) *Let $\mathbb{S}^{k-1}$ be the sphere of radius 1 centered around the origin in $\mathbb{R}^k$ with the Euclidean metric. Let $0 < \varepsilon < 1$.*

CLAIM 7.16

$$\mathcal{N}(S, \rho, \varepsilon) \leq \left(\frac{3}{\varepsilon}\right)^k$$

Let N be any ε-net of S. We claim that indeed $|N| \leq (3/\varepsilon)^k$. The balls of radius $\varepsilon/2$ around points in N, $\{\mathbb{B}^k(x_i, \varepsilon/2) \,:\, x_i \in N\}$, satisfy two properties:

1. *They are pairwise disjoint: if $z \in \mathbb{B}^k(x_i, \varepsilon/2) \cap \mathbb{B}^k(x_j, \varepsilon/2)$, then $\|x_i - x_j\|_2 \leq \|x_i - z\|_2 + \|x_j - z\|_2 \leq \varepsilon$, a contradiction.*

2. *They are included in the ball of radius $3/2$ around the origin: if $z \in \mathbb{B}^k(x_i, \varepsilon/2)$, then $\|z\|_2 \leq \|z - x_i\|_2 + \|x_i\| \leq \varepsilon/2 + 1 \leq 3/2$.*

The volume of a ball of radius is $\varepsilon/2$ is $\frac{\pi^{k/2}(\varepsilon/2)^k}{\Gamma(k/2+1)}$ and that of a ball of radius $3/2$ is $\frac{\pi^{k/2}(3/2)^k}{\Gamma(k/2+1)}$. Dividing one by the other proves the claim.

Spectral norm of a random matrix For a $m \times n$ matrix $A \in \mathbb{R}^{m\times n}$, recall that the spectral norm is defined as

$$\|A\| := \sup_{\mathbf{x}\in\mathbb{R}^n\setminus\{0\}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\substack{\mathbf{x}\in\mathbb{S}^{n-1}\\ \mathbf{y}\in\mathbb{S}^{m-1}}} \langle A\mathbf{x}, \mathbf{y}\rangle, \tag{13}$$

where $\mathbb{S}^{n-1}$ is the sphere of radius 1 around the origin in $\mathbb{R}^n$. (To see the equality above, note that Cauchy-Schwarz implies $\langle A\mathbf{x}, \mathbf{y}\rangle \leq \|A\mathbf{x}\|_2\|\mathbf{y}\|_2$ and that one can take $\mathbf{y} = A\mathbf{x}/\|A\mathbf{x}\|_2$ for any $\mathbf{x}$ such that $A\mathbf{x} \neq 0$ in the rightmost expression.)

THM 7.17 *Let $A \in \mathbb{R}^{m\times n}$ be a random matrix whose entries are centered, independent and sub-Gaussian with variance factor ν. Then there exist a constant $0 < C < +\infty$ such that, for all $t > 0$,*

$$\|A\| \leq C\sqrt{\nu}(\sqrt{m} + \sqrt{n} + t),$$

with probability at least $1 - e^{-t^2}$.

EX 7.18 *Without the independence assumption, the spectral norm can be much larger. Say $A \in \mathbb{R}^{n\times n}$ is all-$(+1)$ or all-(-1) with equal probability. Then considering the all-$(+1)$ vector (and the orthogonal complement) shows that $\|A\| = n$.*

Proof:[of THM 7.17] Fix $\varepsilon = 1/4$. By CLAIM 7.16, there is an ε-net N (respectively M) of $\mathbb{S}^{n-1}$ (respectively $\mathbb{S}^{m-1}$) with $|N| \leq 12^n$ (respectively $|M| \leq 12^m$). We proceed in two steps:

1. We first apply the general Hoeffding inequality (THM 7.9) to control the deviations of the supremum in (13) *restricted to N and M*.

2. We then extend the bound to the full supremum by continuity.

Formally, the result follows from the following two lemmas.

LEM 7.19 *Let N and M be as above. For C large enough, for all $t > 0$,*

$$\mathbb{P}\left[\max_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \geq \frac{1}{2}C\sqrt{\nu}(\sqrt{m} + \sqrt{n} + t)\right] \leq e^{-t^2}.$$

LEM 7.20 *For any ε-nets N and M of $\mathbb{S}^{n-1}$ and $\mathbb{S}^{m-1}$ respectively, the following inequalities hold*

$$\sup_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \leq \|A\| \leq \frac{1}{1-2\varepsilon} \sup_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle.$$

Proof:[Proof of LEM 7.19] Note that the quantity $\langle A\mathbf{x}, \mathbf{y}\rangle$ is a linear combination of independent random variables:

$$\langle A\mathbf{x}, \mathbf{y}\rangle = \sum_{i,j} x_i y_j A_{ij}.$$

By the general Hoeffding inequality (THM 7.9), $\langle A\mathbf{x}, \mathbf{y}\rangle$ is sub-Gaussian with variance factor

$$\sum_{i,j} (x_i y_j)^2 \, \nu = \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2 \, \nu = \nu,$$

for all $\mathbf{x} \in N$ and $\mathbf{y} \in M$. In particular, for all $\beta > 0$,

$$\mathbb{P}\left[\langle A\mathbf{x}, \mathbf{y}\rangle \geq \beta\right] \leq \exp\left(-\frac{\beta^2}{2\nu}\right).$$

Hence, by a union bound over N and M,

$$\begin{aligned}
&\mathbb{P}\left[\max_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \geq \frac{1}{2} C\sqrt{\nu}(\sqrt{m} + \sqrt{n} + t)\right] \\
&\qquad \leq \sum_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \mathbb{P}\left[\langle A\mathbf{x}, \mathbf{y}\rangle \geq \frac{1}{2} C\sqrt{\nu}(\sqrt{m} + \sqrt{n} + t)\right] \\
&\qquad \leq |N||M| \exp\left(-\frac{1}{2\nu}\left\{\frac{1}{2} C\sqrt{\nu}(\sqrt{m} + \sqrt{n} + t)\right\}^2\right) \\
&\qquad \leq 12^{n+m} \exp\left(-\frac{C^2}{8}\left\{m + n + t^2)\right\}\right) \\
&\qquad \leq e^{-t^2},
\end{aligned}$$

for $C^2/8 = \ln 12 \geq 1$, where in the third inequality we ignored all cross-products since they are non-negative. ■

Proof:[Proof of LEM 7.20] The first inequality is immediate. For the second inequality, we will use the following observation

$$\langle A\mathbf{x}, \mathbf{y}\rangle - \langle A\mathbf{x}_0, \mathbf{y}_0\rangle = \langle A\mathbf{x}, \mathbf{y} - \mathbf{y}_0\rangle + \langle A(\mathbf{x} - \mathbf{x}_0), \mathbf{y}_0\rangle. \qquad (14)$$

Fix $x \in \mathbb{S}^{n-1}$ and $y \in \mathbb{S}^{m-1}$ such that $\langle A\mathbf{x}, \mathbf{y}\rangle = \|A\|$, and let $\mathbf{x}_0 \in N$ and $\mathbf{y}_0 \in M$ such that

$$\|\mathbf{x} - \mathbf{x}_0\|_2 \leq \varepsilon \qquad \text{and} \qquad \|\mathbf{y} - \mathbf{y}_0\|_2 \leq \varepsilon.$$

Then (14), Cauchy-Schwarz and the definition of the spectral norm imply

$$\|A\| - \langle A\mathbf{x}_0, \mathbf{y}_0\rangle \leq \|A\|\|\mathbf{x}\|_2\|\mathbf{y}-\mathbf{y}_0\|_2 + \|A\|\|\mathbf{x}-\mathbf{x}_0\|_2\|\mathbf{y}_0\|_2 \leq 2\varepsilon\|A\|.$$

Rearranging gives the claim. ■

Putting the two lemmas together concludes the proof of THM 7.17. ■

References

[Bil95] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1995.

[Dur10] Rick Durrett. *Probability: theory and examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010.

[Roc] Sebastien Roch. Modern Discrete Probability: An Essential Toolkit. *Book in preparation*. http://www.math.wisc.edu/ roch/mdp/.