Subgaussian and subexponential r.v.s
High Dim Probability & Linear Algebra
for ML and Sig Proc

Namrata Vaswani

Iowa State University
2-norm of a subGaussian vector is close to $\sqrt{n}$ w.h.p.:

**Theorem (Concen of norm of a subG vector)**

Let $X \in \mathbb{R}^n$ be a r. vector with independent entries $X_i$ with $\mathbb{E}[X_i^2] = 1$. Let $K = \max_i \|X_i\|_{\psi_2}$. Then $\|X\| - \sqrt{n}$ is a sub-G r.v. with sub-G norm at most $K^2$. Equivalently,

$$\Pr(|\|X\| - \sqrt{n}| \geq t) \leq 2 \exp(-ct^2/K^4)$$

**Proof:**

1. For a subG r.v. with $\mathbb{E}[Z^2] = 1$, $K_Z \geq 1$
   - Reason: using $1 + x \leq e^x$, with $x = Z^2/K_Z^2$. $\mathbb{E}[1 + Z^2/K^2] \leq \mathbb{E}[e^{Z^2/K^2}]$ which implies $1 + 1/K^2 \leq \mathbb{E}[e^{Z^2/K^2}]$. By subG property, $\mathbb{E}[e^{Z^2/K^2}] \leq 2$ and this gives $K \geq 1$.

2. Consider $\frac{1}{n}\|X\|^2 - 1 = \frac{1}{n}\sum_i (X_i^2 - 1)$. By the properties from earlier, $X_i^2 - 1$ are independent, zero mean, sub-expo r.v.s with $K_{\text{expo}} \leq CK^2$. So we can apply the sub-expo Bernstein inequality to conclude that

$$\Pr(|\frac{1}{n}\|X\|^2 - 1| \geq u) \leq 2 \exp\left(-c \frac{n}{K^4} \min(u^2, u)\right)$$

(the above also used $K \geq 1$).
Use $|z - 1| \geq \delta$ implies $|z^2 - 1| \geq \max(\delta, \delta^2)$ and the fact that $A \Rightarrow B$ implies $\Pr(A) \leq \Pr(B)$ to conclude that

$$\Pr\left(\left|\frac{1}{\sqrt{n}}\|X\| - 1\right| \geq \delta\right) \leq \Pr\left(\left|\frac{1}{n}\|X\|^2 - 1\right| \geq \max(\delta, \delta^2)\right) \leq 2 \exp\left(-c \frac{n}{K^4} \delta^2\right)$$

(used: for $u = \max(\delta, \delta^2)$, $\min(u^2, u) = \delta^2$).

Set $\delta = t/\sqrt{n}$ to conclude that

$$\Pr(\||X\| - \sqrt{n}| \geq t) \leq 2 \exp\left(-c \frac{1}{K^4} t^2\right)$$

When working with random vectors, we generally subtract mean first to get zero-mean random vectors.

Isotropic random vectors: $X \in \mathbb{R}^n$ is isotropic if

$$\mathbb{E}[XX^\top] = I_n$$

Properties of isotropic $X$

$\mathbb{E}[(a^\top X)^2] = \|a\|^2$ for all $a \in \mathbb{R}^n$ (this is equivalent to the definition)
High-dimensional random vectors III

- $\mathbb{E}[\|X\|^2] = n$

- $X, Y$ independent and isotropic, then $\mathbb{E}[(X'Y)^2] = n$
  - Implication of this and concentration of norm result (Remark 3.2.5): can argue that if $X, Y$ are indep., then $\frac{X}{\|X\|}, \frac{Y}{\|Y\|}$ are almost orthogonal, i.e. their inner product is of order $1/\sqrt{n}$.
  - TBD: quantify above claim, it is not quantified in the book.

- Examples of isotropic r. vectors:
  - i.i.d symmetric Bernoulli;
  - standard Gaussian vector;
  - any “product” distribution (coordinates of $X$ are independent) with zero mean and unit variance;
  - coordinate distribution ($X$ equally likely to be $\sqrt{n}e_i, i = 1, 2, \ldots, n$; recall $e_i$ is the $i$-th column of $I$
  - $X \sim Unif(\sqrt{n}S^{n-1})$: this is isotropic but coordinates are not independent (proof is not obvious, TBD);
  - unif distrib on frames

- Sub-Gaussian random vector
Definition: X is a sub-G vector iff $a'X$ is sub-G for all $a \in \mathbb{R}^n$. Sub-G norm of X is

$$\|X\|_{\psi 2} := \sup_{a \in S^{n-1}} \|a'X\|_{\psi 2}$$

Sub-G with independent coordinates $X = (X_1, X_2, ... X_n)'$ with $X_i$’s independent sub-G: then

$$\|X\|_{\psi 2} \leq C \max_{i=1,2,...,n} \|X_i\|_{\psi 2}$$

Spherical distribution is sub-Gaussian: $Z \sim \text{Unif}(\sqrt{n}S^{n-1})$ is sub-G with subG norm at most $C$. Proof:

1. Use the following property: For a standard Gaussian random vector, $X$, i.e., $X \sim \mathcal{N}(0, I)$

$$\theta := \frac{X}{\|X\|} \sim \text{Unif}(S^{n-1})$$

Also, $\|X\|, \theta$ are independent.

2. Use this property to conclude that we can express $Z$ as

$$Z = \sqrt{n}G/\|G\|$$

where $G \sim \mathcal{N}(0, I)$. 

Namrata Vaswani (Iowa State U.)  High Dim Prob & Lin Alg for ML 5 / 16
To prove that $Z$ is sub-$G$, we need to prove that $a'Z$ is sub-$G$ for all $a \in \mathbb{R}^n$.

1. Rotation invariance property of $G$ implies that $a'G = e_1' U_a G = \tilde{G}_1$ where $\tilde{G} = U_a' G \sim \mathcal{N}(0, I)$ too and $\|\tilde{G}\| = \|G\|$. Here $U_a$ is an orthonormal matrix with first column $a/\|a\|$.

2. Thus, w.l.o.g., $a'Z = \sqrt{n} \tilde{G}_1/\|\tilde{G}_1\|$ and we need to bound $\Pr(\sqrt{n} \tilde{G}_1/\|\tilde{G}_1\| \geq u)$.

3. Apply concentration of norm result on $\|\tilde{G}\|$ with $t = \sqrt{n}/2$ to conclude that

$$\Pr(\|\tilde{G}\| \geq \sqrt{n}/2) \geq 1 - 2\exp(-cn)$$

(follows since $K$ for a standard Gaussian vector is a constant).

4. Using total probability with $Ev, Ev^c$,

$$\Pr(\sqrt{n} \tilde{G}_1/\|\tilde{G}_1\| \geq u) \leq \Pr(\sqrt{n} \tilde{G}_1/\|\tilde{G}_1\| \geq u \text{ and } Ev) + \Pr( Ev^c)$$

$$\leq \Pr( \tilde{G}_1 \geq u/2 \text{ and } Ev) + 2\exp(-cn)$$

$$\leq \Pr( \tilde{G}_1 \geq u/2) + 2\exp(-cn)$$

$$\leq 2\exp(-u^2/8) + 2\exp(-cn) \leq 4\exp(-u^2/8)$$

Reason for last bound:
If $u < \sqrt{n}$, then first term dominates and we can conclude that $Z$ is sub-$G$.
If $u \geq \sqrt{n}$, then $\Pr(\sqrt{n} \tilde{G}_1/\|\tilde{G}_1\| \geq u) = 0$ since $\tilde{G}_1 \leq \|\tilde{G}_1\|$.
Epsilon-net I

Epsilon net is a finite set of points that is used to “cover” a compact set in a metric space by using balls of radius $\epsilon$. More precisely, it is a set of finite points so that any point on the compact set is within $\epsilon$ distance of some point in the epsilon-net.

1. Definition for $N_\epsilon$ that covers $S^{n-1}$ in Euclidean distance: $N_\epsilon \subset S^{n-1}$ is an $\epsilon$-net of $S^{n-1}$ if for any $x \in S^{n-1}$, there exists a $\bar{x} \in N_\epsilon$ s.t. $\|x - \bar{x}\| \leq \epsilon$.

2. Bound size of epsilon-net: can use volume arguments to show that we can find an $\epsilon$-net that covers $S^{n-1}$ with cardinality

$$|N_\epsilon| \leq (1 + 2/\epsilon)^n$$

3. Use to bound $\|A\|$ by using $\|A\| = \max_{x \in S^{n-1}} \|Ax\|$: Suppose $x$ is the point on the sphere that achieves the above max. By definition, there exists an $\bar{x}(x)$ in the net s.t. $\|\bar{x} - x\| \leq \epsilon$. Thus

$$\|A\| = \|Ax\| = \|A(\bar{x} + x - \bar{x})\| \leq \|A\bar{x}\| + \|A\|\|x - \bar{x}\| \leq \|A\bar{x}\| + \|A\|\epsilon$$

So

$$(1 - \epsilon)\|A\| \leq \|A\bar{x}\| \leq \max_{\bar{x} \in N_\epsilon} \|A\bar{x}\|$$

and hence

$$\|A\| \leq \frac{1}{1 - \epsilon} \max_{\bar{x} \in N_\epsilon} \|A\bar{x}\|$$
4 Use to bound $\sigma_{\min}(A)$ by using $\sigma_{\min}(A) = \min_{x \in S^{n-1}} \|Ax\|$:
proceed as above; this bound uses the bound on $\|A\|$ from above.

5 Use to bound $\|A\|$ by using $\|A\| = \max_{x \in S^{n-1}, y \in S^{m-1}} y'Ax$. In some proofs, the above
norm definition is needed. One can show that

$$
\|A\| \leq \frac{1}{1 - 2\epsilon} \max_{\bar{x} \in \mathcal{N}_\epsilon(S^{n-1}), \bar{y} \in \mathcal{N}_\epsilon(S^{m-1})} \bar{y}'A\bar{x}
$$
Bound on min and max singular values of an $m \times n$ matrix with independent isotropic sub-Gaussian rows.

**Theorem (Sub-Gaussian rows matrix)**

Let $A$ be an $m \times n$ matrix whose rows, $A^i$, are independent, zero-mean, sub-G, isotropic r.vectors. Let $K = \max_i \|A^i\|_{\psi^2}$. Then, for a large enough numerical constant $C$,

$$\sqrt{m} - CK^2(\sqrt{n} + t) \leq s_i(A) \leq \sqrt{m} + CK^2(\sqrt{n} + t)$$

w.p. at least $1 - 2 \exp(-t^2)$. Here $s_i(A)$ is the $i$-th singular value of $A$.

Claim: The bounds of the theorem will hold if we can instead prove that

$$\|\frac{1}{m} A^\prime A - I\| \leq K^2 \max(\delta, \delta^2), \quad \delta = \frac{\sqrt{n} + t}{\sqrt{m}}$$

(this claim follows using the simple algebra fact that $\max(|z - 1|, |z - 1|^2) \leq |z^2 - 1|$)

Bounding $\|\frac{1}{m} A^\prime A - I\|$:
Approximation: use the following results for epsilon-nets: for a symmetric $M$,

$$\|M\| := \max_{x \in S^{n-1}} |x'Ax| \leq \frac{1}{1 - 2\epsilon} \max_{x \in \mathcal{N}_\epsilon} |x'Ax|$$

where $\mathcal{N}_\epsilon \subset S^{n-1}$ is an epsilon-net on $S^{n-1}$. By the covering number bound, we can find a $1/4$-net for which

$$|\mathcal{N}_\epsilon| \leq (1 + 2/\epsilon)^n$$

Using these with $\epsilon = 1/4$ and simplifying,

$$\|\frac{1}{m} A'A - I\| \leq 2 \max_{x \in \mathcal{N}_{1/4}} |\frac{1}{m} \|Ax\|^2 - 1|$$

and

$$|\mathcal{N}_{1/4}| \leq 9^n$$
Concentration: for a fixed $x \in \mathcal{N}_{1/4} \subset S^{n-1}$: Since the rows $A^i$ are isotropic (implies $\mathbb{E}[(x'A^i)^2] = 1$), sub-G, independent, with sub-G norm at most $K$, 

$$\frac{1}{m} \|Ax\|^2 - 1 = \frac{1}{m} \sum_{i=1}^{m} ((x'A^i)^2 - 1),$$

is a sum of $m$ independent, zero-mean, sub-expo r.v.s with sub-expo norm at most $CK^2/m$. We can apply sub-expo Bernstein ineq to conclude that 

$$\Pr(|\frac{1}{m} \|Ax\|^2 - 1| \geq \epsilon /2) \leq 2 \exp(-cm \min(\epsilon^2 / K^4, \epsilon / K^2))$$

Use $\epsilon = K^2 \max(\delta, \delta^2)$ with $\delta = C(\sqrt{n} + t) / \sqrt{m}$ to get

$$\Pr(|\frac{1}{m} \|Ax\|^2 - 1| \geq K^2 \max(\delta, \delta^2)) \leq 2 \exp(-cm\delta^2) \leq 2 \exp(-cC^2(n + t^2))$$
Union bound: over all $x \in N_{1/4} \subset S^{n-1}$ gives:

$$\Pr(\max_{x \in N_{1/4}} \frac{1}{m} \|Ax\|^2 - 1 \geq K^2 \max(\delta, \delta^2)) \leq 9^n 2 \exp(-cC^2(n + t^2)) \leq \exp(-t^2)$$

by choosing $C$ large enough.

By combining this with the Approximation step, (1) holds w.p. $\geq 1 - \exp(-t^2)$.

Implication of the theorem: if $m \geq CK^2 n$, then the min singular value of $A/\sqrt{m}$ is at least a constant $c < 1$ and the max singular value is at most a constant $C > 1$, thus the condition number is a constant.

Bound on expected value: using the above result and the integral identity applied to $Z = \|A'A - mI\|$, 

$$\mathbb{E}[\frac{1}{m} A'A - I\|] \leq CK^2(\sqrt{n/m} + (n/m))$$
Proof [skip]: above result and $\max(a, b) < a + b$ tells us that $\Pr(Z < CK^2\sqrt{mn} + n + \sqrt{mt} + t^2) \geq 1 - \exp(-t^2)$. Let $u_0 = CK^2(\sqrt{mn} + n)$. Thus, using integral identity applied to $Z = \|A'A - mI\|$, 

$$
\mathbb{E}[Z] \leq u_0 + \int_{\tau = u_0}^{\infty} \Pr(Z > \tau) d\tau \\
= u_0 + \int_{t=0}^{\infty} \Pr(Z > u_0 + \sqrt{mt} + t^2)\sqrt{mdt} + \Pr(Z > u_0 + \sqrt{mt} + t^2)2tdt \\
\leq u_0 + \int_{t=0}^{\infty} \exp(-t^2)\sqrt{mdt} + \int_{t=0}^{\infty} \exp(-t^2)2tdt \\
\leq u_0 + 3
$$

Second row used $\tau = u_0 + \sqrt{mt} + t^2$ so that $d\tau = \sqrt{mdt} + 2tdt$; third row used Theorem conclusion; last row follows by simple integration by parts. Since $u_0 = CK^2(\sqrt{mn} + n)$, for $n, m$ large enough, $u_0 + 3 < 1.1u_0$. Thus, $\mathbb{E}[Z] \leq 1.1u_0$ and so $\mathbb{E}[Z/m] \leq 1.1u_0/m$, i.e., 

$$
\mathbb{E}[\frac{1}{m} \|A'A - I\|] \leq CK^2(\sqrt{n/m} + (n/m))
$$
We can also use the Theorem and integral identity to show that
\[ \sqrt{m} - CK^2 \sqrt{n} \leq \mathbb{E}[s_n(A)], \text{ and } \mathbb{E}[s_1(A)] \leq \sqrt{m} + CK^2 \sqrt{n} \]

Can obtain an easy extension for the non-isotropic case as well.

Matrix Bernstein:

**Theorem (Matrix Bernstein)**

Let \( X_1, X_2, \ldots, X_m \) be independent, zero-mean, \( d_1 \times d_2 \) matrices with \( \|X_i\| \leq L \) for all \( i = 1, 2, \ldots, m \). Define the “variance parameter” of the sum
\[ v := \max \left( \| \sum_i X_i X_i' \|, \| \sum_i X_i' X_i \| \right) \]

Then
\[ \Pr(\| \sum_{i=1}^m X_i \| \geq t) \leq (d_1 + d_2) \exp \left( -c \frac{t^2}{v + Lt/3} \right) \leq \exp(-c \min(\frac{t^2}{v}, \frac{t}{L})) \]

For symmetric matrices \( X_i \) of size \( nxn \), \( v = \| \sum_i X_i^2 \|, d_1 = d_2 = n \).
For nonzero mean matrices, the above bound, along with Weyl’s inequality, implies that, w.p. $\geq 1 - \exp(-c \min(\frac{t^2}{\nu}, \frac{t}{L}))$,

$$s_{\min}\left(\sum_{i=1}^{m} \mathbb{E}[X_i]\right) - t \leq s_{\min}\left(\sum_{i=1}^{m} X_i\right) \leq s_{\max}\left(\sum_{i=1}^{m} X_i\right) \leq s_{\max}\left(\sum_{i=1}^{m} \mathbb{E}[X_i]\right) + t$$

i.e. the min and max singular values of the sum are close to those of the expected values w.h.p.

**Proof:** TBD; first prove the result for sums of symmetric matrices, then extend to any general matrices using the dilation trick; for symmetric matrices: bound the MGF of $\lambda_{\max}(\sum_{i} X_i)$ conditioned on $X_1, X_2, X_{m-1}$, using Leib’s inequality or Gold-Thompson inequality; followed by averaging over $X_{m-1}$ and repeating the steps. Do this one at a time, eventually get a bound on the MGF. Not straightforward. See Vershynin book Sec 5.4 or the original reference “User-friendly tail bounds for sums of random matrices” by Joel Tropp.
Matrix Bernstein vs subGaussian rows’ result: Matrix Bernstein bounds the norm of the sum of bounded, independent, zero-mean random matrices. For matrices with nonzero means, it helps to provide a lower bound on the min singular value and upper bound on the max singular value of the matrix sum. Comparison with the previous result: $A' A$ can also be interpreted as the sum of rank-one matrices $A' A = \sum_i (A^i) (A^i)'$. Previous result is for a sum of rank-one matrices, $(A^i) (A^i)'$ with $A^i$ being sub-Gaussian. Matrix Bernstein applies to this setting if the $A^i$ are bounded.