Contents

1 My summary and notes

Summary and notes in rpub

2 Chapter 1: Concentration of sum of independent RV

High dimension probability in Data science

Chapter 1: Concentration of sum of independent RV

Author: Roman Vershynin *Learner: Weihao Li*

Proposition 2.1 (Tails of the normal distribution)**.** *Let $g \sim N(0,1)$. Then for all $t > 0$, we have*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \mathbb{P}\{g \geq t\} \leq \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

In particular, for $t \geq 1$ the tail is bounded by the density:

$$\mathbb{P}\{g \geq t\} \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

Theorem 2.2 (Berry-Esseen central limit theorem)**.**

$$Z_N := \frac{S_N - \mathbb{E}S_N}{\sqrt{\mathrm{Var}\,(S_N)}} = \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^{N} (X_i - \mu)$$

for every N and every $t \in \mathbb{R}$ we have

$$|\mathbb{P}\{Z_N \geq t\} - \mathbb{P}\{g \geq t\}| \leq \frac{\rho}{\sqrt{N}}$$

Here $\rho = \mathbb{E}\,|X_1 - \mu|^3 / \sigma^3$ and $g \sim N(0,1)$.

Theorem 2.3 (Hoeffding inequality)**.** *Let $X_1, \ldots, X_N$ be independent symmetrical Bernoulli random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for any $t \geq 0$, we have*

$$\mathbb{P}\left\{\sum_{i=1}^{N} a_i X_i \geq t\right\} \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right)$$

Theorem 2.4 (Hoeffding inequality for general bounded random variable)**.** *Let $X_1, \ldots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $t > 0$, we have*

$$\mathbb{P}\left\{\sum_{i=1}^{N} (X_i - \mathbb{E}X_i) \geq t\right\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N} (M_i - m_i)^2}\right)$$

Proposition 2.5 (subgaussian properties)**.**

$$\begin{gathered}
\mathbb{P}\{|X| \geq t\} \leq 2\exp\left(-ct^2/\|X\|_{\psi_2}^2\right) \quad \text{for all } t \geq 0 \\
\|X\|_{L^p} \leq C\|X\|_{\psi_2}\sqrt{p} \quad \text{for all } p \geq 1 \\
\mathbb{E}\exp\left(X^2/\|X\|_{\psi_2}^2\right) \leq 2 \\
\text{if } \mathbb{E}X = 0 \text{ then } \mathbb{E}\exp(\lambda X) \leq \exp\left(C\lambda^2\|X\|_{\psi_2}^2\right) \quad \text{for all } \lambda \in \mathbb{R}\ .
\end{gathered}$$

Proposition 2.6 (sum of independent sub-gaussian)**.**

$$\left\|\sum_{i=1}^{N} X_i\right\|_{\psi_2}^2 \leq C\sum_{i=1}^{N} \|X_i\|_{\psi_2}^2$$

Theorem 2.7 (General Hoeffding inequality 1)**.** *Let $X_1, \dots, X_N$ be independent, mean zero, sub-gaussian random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^{N}\|X_i\|_{\psi_2}^2}\right)$$

Theorem 2.8 (General Hoeffding inequality 2)**.** *Let $X_1, \dots, X_N$ be independent, mean zero, sub-gaussian random variables, and $a = (a_1, \dots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{ct^2}{K^2\|a\|_2^2}\right)$$

where $K = \max_i \|X_i\|_{\psi_2}$

Theorem 2.9 (Khintchine's inequality for $p \geq 2$)**.** *Let $X_1, \dots, X_N$ be independent sub-gaussian random variables with zero means and unit variances, and let $a = (a_1, \dots, a_N) \in \mathbb{R}^N$.*

$$\left(\sum_{i=1}^{N} a_i^2\right)^{1/2} \leq \left\|\sum_{i=1}^{N} a_i X_i\right\|_{L^p} \leq CK\sqrt{p}\left(\sum_{i=1}^{N} a_i^2\right)^{1/2}$$

where $K = \max_i \|X_i\|_{\psi_2}$ and C is an absolute constant.

Theorem 2.10 (Khintchine's inequality for $p = 1$)**.** *Same setting as 2.9*

$$c(K)\left(\sum_{i=1}^{N} a_i^2\right)^{1/2} \leq \left\|\sum_{i=1}^{N} a_i X_i\right\|_{L^1} \leq \left(\sum_{i=1}^{N} a_i^2\right)^{1/2}$$

Here $K = \max_i \|X_i\|_{\psi_2}$ and $c(K) > 0$ is a quantity which may depend only on K.

Definition 2.11 (sub-exponential norm)**.**

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}\exp(|X|/t) \leq 2\}$$

Lemma 2.12 (Centering)**.** *If X is a sub-gaussian random variable then $X - \mathbb{E}X$ is sub-gaussian, too, and*

$$\|X - \mathbb{E}X\|_{\psi_2} \leq C\|X\|_{\psi_2}$$
$$\|X - \mathbb{E}X\|_{\psi_1} \leq C\|X\|_{\psi_1}$$

where C is an absolute constant.

Lemma 2.13 (sub-exponential is sub-gaussian square)**.** *A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2$$

Lemma 2.14 (product of sub-gaussian is sub-exponential)**.** *Let X and Y be sub-gaussian random variables. Then XY is sub-exponential. Moreover,*

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$$

Theorem 2.15 (Bernstein's inequality 1)**.** *Let $X_1, \dots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{\sum_{i=1}^{N}\|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right)\right]$$

where $c > 0$ is an absolute constant.

Remark: The reason we have "min" here is: the bound for MGF does not hold for all lambda in case of subexponential

Theorem 2.16 (Bernstein's inequality 2)**.** *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right]$$

where $K = \max_i \|X_i\|_{\psi_1}$

Theorem 2.17 (Bernstein's inequality for bounded distributions)**.** *Let $X_1, \ldots, X_N$ be independent, mean zero random variables, such that $|X_i| \leq K$ all i. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right)$$

Here $\sigma^2 = \sum_{i=1}^{N} \mathbb{E}X_i^2$ is the variance of the sum.

Theorem 2.18 (bounded difference inequality)**.** *Theorem 2.9.1 (Bounded differences inequality). Let $X_1, \ldots, X_N$ be independent random variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a measurable function. Assume that the value of $f(x)$ can change by at most $c_i > 0$ under an arbitrary change of a single coordinate of $x \in \mathbb{R}^n$. Then, for any $t > 0$, we have*

$$\mathbb{P}\{f(X) - \mathbb{E}f(X) \geq t\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N} c_i^2}\right)$$

where $X = (X_1, \ldots, X_n)$

Theorem 2.19 (Bennett's inequality)**.** *Let $X_1, \ldots, X_N$ be independent random variables. Assume that $|X_i - \mathbb{E}X_i| \leq K$ almost surely for every i. Then, for any $t > 0$, we have*

$$\mathbb{P}\left\{\sum_{i=1}^{N} (X_i - \mathbb{E}X_i) \geq t\right\} \leq \exp\left(-\frac{\sigma^2}{K^2} h\left(\frac{Kt}{\sigma^2}\right)\right)$$

where $\sigma^2 = \sum_{i=1}^{N} \mathrm{Var}(X_i)$ is the variance of the sum, and $h(u) = (1+u)\log(1+u) - u$.

3 Chapter 2: Random vectors in high dimensions

High dimension probability in Data science

Chapter 2: Random vectors in high dimensions

Author: Roman Vershynin *Learner: Weihao Li*

Theorem 3.1 (Concentration of norm). *Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, sub-gaussian coordinates X_i that satisfy $\mathbb{E}X_i^2 = 1$. Then*

$$\| \|X\|_2 - \sqrt{n} \|_{\psi_2} \le CK^2$$

$$\mathbb{P}\left\{ \left| \|X\|_2 - \sqrt{n} \right| \ge t \right\} \le 2\exp\left(-\frac{ct^2}{K^4}\right) \quad \text{for all } t \ge 0$$

where $K = \max_i \|X_i\|_{\psi_2}$ and C is an absolute constant.

Lemma 3.2 (Characteristic of isotropy). *A random vector X in $\mathbb{R}^n$ is isotropic if and only if*

$$\mathbb{E}\langle X, x\rangle^2 = \|x\|_2^2 \quad \text{for all } x \in \mathbb{R}^n.$$

Lemma 3.3. *Let X be an isotropic random vector in $\mathbb{R}^n$. Then*

$$\mathbb{E}\|X\|_2^2 = n$$

Moreover, if X and Y are two independent isotropic random vectors in $\mathbb{R}^n$, then

$$\mathbb{E}\langle X, Y\rangle^2 = n$$

Lemma 3.4 (Normal and spherical distributions). *Let us represent $g \sim N(0, I_n)$ in polar form as*

$$g = r\theta$$

where $r = \|g\|_2$ is the length and $\theta = g/\|g\|_2$ is the direction of g. Then

(a) The length r and direction θ are independent random variables.

(b) The direction θ is uniformly distributed on the unit sphere S^{n-1}.

Lemma 3.5 (Sub-gaussian distributions with independent coordinates). *Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean zero, subgaussian coordinates X_i. Then X is a sub-gaussian random vector, and*

$$\|X\|_{\psi_2} \le C \max_{i \le n} \|X_i\|_{\psi_2}$$

Theorem 3.6 (Uniform distribution on the sphere is sub-gaussian). *Let X be a random vector uniformly distributed on the Euclidean sphere in $\mathbb{R}^n$ with center at the origin and radius $\sqrt{n}$:*

$$X \sim \ \mathit{Unif}\left(\sqrt{n}S^{n-1}\right)$$

Then X is sub-gaussian, and

$$\|X\|_{\psi_2} \le C$$

Remark: result also hold for Uniform distribution on the Euclidean ball Unif $(B(0, \sqrt{n}))$.

Theorem 3.7 (Projective limit theorem). *$X \sim \mathit{Unif}\left(\sqrt{n}S^{n-1}\right)$ then for any fixed unit vector x, we have $\langle X, x\rangle \to N(0,1)$ in distribution as $n \to \infty$.*

Theorem 3.8 (Grothendieck's inequality)**.** *Consider an $m \times n$ matrix (a_{ij}) of real numbers. Assume that, for any numbers $x_i, y_j \in \{-1, 1\}$, we have*

$$\left| \sum_{i,j} a_{ij} x_i y_j \right| \leq 1$$

Then, for any Hilbert space H and any vectors $u_i, v_j \in H$ satisfying $\|u_i\| = \|v_j\| = 1$, we have

$$\left| \sum_{i,j} a_{ij} \langle u_i, v_j \rangle \right| \leq K$$

where $K \leq 1.783$ is an absolute constant.

Definition 3.9 (semidefinite programme)**.** *A semidefinite program is an optimization problem of the following type:*

$$\text{maximize} \langle A, X \rangle : \quad X \succeq 0, \quad \langle B_i, X \rangle = b_i \text{ for } i = 1, \ldots, m$$

Here A and B_i are given $n \times n$ matrices and b_i are given real numbers.

Theorem 3.10. *Consider two optimization problem: for a given matrix $A \in \mathbb{R}^{n \times n}$*

$$\textit{INT(A)= maximize } \sum_{i,j=1}^{n} A_{ij} x_i x_j : \quad x_i = \pm 1 \text{ for } i = 1, \ldots, n$$

$$SDP(A) = \text{maximize} \langle A, X \rangle : \quad X \succeq 0, \quad X_{ii} = 1 \text{ for } i = 1, \ldots, n$$

Then

$$\text{INT}(A) \leq \text{SDP}(A) \leq 2K \cdot \text{INT}(A)$$

where $K \leq 1.783$ is the constant in Grothendieck's inequality.

Theorem 3.11 (Max-cut and SDP relaxation)**.** *Given graph G and adjacency matrix A*

$$\text{MAX} - \text{CUT}(G) = \frac{1}{4} \max \left\{ \sum_{i,j=1}^{n} A_{ij} (1 - x_i x_j) : x_i = \pm 1 \text{ for all } i \right\}$$

$$\text{SDP}(G) := \frac{1}{4} \max \left\{ \sum_{i,j=1}^{n} A_{ij} (1 - \langle X_i, X_j \rangle) : X_i \in \mathbb{R}^n, \|X_i\|_2 = 1 \text{ for all } i \right\}$$

Let $x = (x_i)$ be the result of a randomized rounding of the solution (X_i) of the semidefinite program, which means that

$$g \sim N(0, I_n)$$
$$x_i := \text{sign} \langle X_i, g \rangle, \quad i = 1, \ldots, n$$

Then we have

$$\mathbb{E} \, \text{CUT}(G, x) \geq 0.878 \, \text{SDP}(G) \geq 0.878 \; \textit{MAX-CUT}(G)$$

Lemma 3.12 (Grothendieck's identity)**.** *Consider a random vector $g \sim N(0, I_n)$. Then, for any fixed vectors $u, v \in S^{n-1}$, we have*

$$\mathbb{E} \, \text{sign} \langle g, u \rangle \, \text{sign} \langle g, v \rangle = \frac{2}{\pi} \arcsin \langle u, v \rangle$$

4 Chapter 3: Random matrices

High dimension probability in Data science

Chapter 3: Random matrices

Author: Roman Vershynin *Learner: Weihao Li*

Definition 4.1 (operator norm quadratic form)**.**

$$\|A\| = \max_{x \in S^{n-1}, y \in S^{m-1}} \langle Ax, y \rangle$$

Lemma 4.2 (Approximate isometry)**.** *Let A be an $m \times n$ matrix and $\delta > 0$. Suppose that*

$$\left\|A^\top A - I_n\right\| \leq \max\left(\delta, \delta^2\right).$$

Then

$$(1-\delta)\|x\|_2 \leq \|Ax\|_2 \leq (1+\delta)\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n.$$

Consequently, all singular values of A are between $1-\delta$ and $1+\delta$:

$$1 - \delta \leq s_n(A) \leq s_1(A) \leq 1 + \delta \ .$$

Definition 4.3 (packing number)**.** *A subset $\mathcal{N}$ of a metric space (T, d) is ε -separated if $d(x, y) > \varepsilon$ for all distinct points $x, y \in \mathcal{N}$. The largest possible cardinality of an ε -separated subset of a given set $K \subset T$ is called the packing number of K and is denoted $\mathcal{P}(K, d, \varepsilon)$.*

Lemma 4.4 (Nets from separated sets)**.** *Let $\mathcal{N}$ be a maximal [2] ε -separated subset of K. Then $\mathcal{N}$ is an ε-net of K.*

Lemma 4.5 (Equivalence of covering and packing numbers)**.** *For any set $K \subset T$ and any $\varepsilon > 0$, we have*

$$\mathcal{P}(K, d, 2\varepsilon) \leq \mathcal{N}(K, d, \varepsilon) \leq \mathcal{P}(K, d, \varepsilon)$$

Proposition 4.6 (Covering numbers and volume)**.** *Let K be a subset of $\mathbb{R}^n$ and $\varepsilon > 0$. Then*

$$\frac{|K|}{|\varepsilon B_2^n|} \leq \mathcal{N}(K, \varepsilon) \leq \mathcal{P}(K, \varepsilon) \leq \frac{|(K + (\varepsilon/2)B_2^n)|}{|(\varepsilon/2)B_2^n|}$$

Corollary 4.7 (Covering numbers of the Euclidean ball)**.**

$$\left(\frac{1}{\varepsilon}\right)^n \leq \mathcal{N}\left(B_2^n, \varepsilon\right) \leq \left(\frac{2}{\varepsilon} + 1\right)^n.$$

The same upper bound is true for the unit Euclidean sphere S^{n-1}.

Lemma 4.8 (Computing the operator norm on a net)**.** *Let A be an $m \times n$ matrix and $\varepsilon \in [0, 1)$. Then, for any ε-net $\mathcal{N}$ of the sphere S^{n-1}, we have*

$$\sup_{x \in \mathcal{N}} \|Ax\|_2 \leq \|A\| \leq \frac{1}{1-\varepsilon} \cdot \sup_{x \in \mathcal{N}} \|Ax\|_2$$

Theorem 4.9 (Norm of matrices with sub-gaussian entries)**.** *Let A be an $m\times$ n random matrix whose entries A_{ij} are independent, mean zero, sub-gaussian random variables. Then, for any $t > 0$ we have [6]*

$$\|A\| \leq CK(\sqrt{m} + \sqrt{n} + t)$$

with probability at least $1 - 2\exp\left(-t^2\right)$. Here $K = \max_{i,j} \|A_{ij}\|_{\psi_2}$.

Corollary 4.10 (Norm of symmetric matrices with sub-gaussian entries)**.** *Let A be an $n \times n$ symmetric random matrix whose entries A_{ij} on and above the diagonal are independent, mean zero, sub-gaussian random variables. Then, for any $t > 0$ we have*

$$\|A\| \leq CK(\sqrt{n} + t)$$

with probability at least $1 - 4\exp\left(-t^2\right)$. Here $K = \max_{i,j} \|A_{ij}\|_{\psi_2}$.

Theorem 4.11 (Weyl's inequality)**.** *For any symmetric matrices S and T with the same dimensions, we have*

$$\max_i |\lambda_i(S) - \lambda_i(T)| \leq \|S - T\|$$

Theorem 4.12 (Davis Kahan)**.** *Let S and T be symmetric matrices with the same dimensions. Fix i and assume that the i-th largest eigenvalue of S is well separated from the rest of the spectrum:*

$$\min_{j:j\neq i} |\lambda_i(S) - \lambda_j(S)| = \delta > 0$$

Then the angle between the eigenvectors of S and T corresponding to the i-th largest eigenvalues (as a number between 0 and $\pi/2$) satisfies

$$\sin \angle (v_i(S), v_i(T)) \leq \frac{2\|S - T\|}{\delta}$$

and

$$\exists \theta \in \{-1, 1\} : \quad \|v_i(S) - \theta v_i(T)\|_2 \leq \frac{2^{3/2}\|S - T\|}{\delta}$$

Theorem 4.13 (Two-sided bound on sub-gaussian matrices, not sharp)**.** *Let A be an $m \times n$ matrix whose rows A_i are independent, mean zero, sub-gaussian isotropic random vectors in $\mathbb{R}^n$. Then for any $t \geq 0$ we have*

$$\sqrt{m} - CK^2(\sqrt{n} + t) \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + CK^2(\sqrt{n} + t)$$

with probability at least $1 - 2\exp\left(-t^2\right)$. Here $K = \max_i \|A_i\|_{\psi_2}$.

Theorem 4.14 (covariance estimation)**.** *Let X be a sub-gaussian random vector in $\mathbb{R}^n$. More precisely, assume that there exists $K \geq 1$ such that*

$$\|\langle X, x\rangle\|_{\psi_2} \leq K\|\langle X, x\rangle\|_{L^2} \quad \text{for any } x \in \mathbb{R}^n.$$

Then, for every positive integer m, we have

$$\mathbb{E}\,\|\Sigma_m - \Sigma\| \leq CK^2\left(\sqrt{\frac{n}{m}} + \frac{n}{m}\right)\|\Sigma\|$$

OR

$$\|\Sigma_m - \Sigma\| \leq CK^2\left(\sqrt{\frac{n+u}{m}} + \frac{n+u}{m}\right)\|\Sigma\|$$

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

5 Chapter 4: Concentration without independence

High dimension probability in Data science

Chapter 4: Concentration without independence

Author: Roman Vershynin *Learner: Weihao Li*

Theorem 5.1 (Concentration of Lipschitz functions on the sphere)**.** *Consider a random vector* $X \sim \text{Unif}\left(\sqrt{n}S^{n-1}\right)$*, i.e. X is uniformly distributed on the Euclidean sphere of radius* $\sqrt{n}$*. Consider a Lipschitz function* $f : \sqrt{n}S^{n-1} \to \mathbb{R}$*. Then*

$$\|f(X) - \mathbb{E}f(X)\|_{\psi_2} \leq C\|f\|_{Lip}$$

$$\mathbb{P}\{|f(X) - \mathbb{E}f(X)| \geq t\} \leq 2\exp\left(-\frac{ct^2}{\|f\|_{\text{Lip}}^2}\right)$$

Theorem 5.2 (Isoperimetric inequality on $\mathbb{R}^n$)**.** *Among all subsets* $A \subset \mathbb{R}^n$ *with given volume, the Euclidean balls have minimal area. Moreover, for any* $\varepsilon > 0$*, the Euclidean balls minimize the volume of the* ε*-neighborhood of A, defined as* [2]

$$A_\varepsilon := \{x \in \mathbb{R}^n : \exists y \in A \text{ such that } \|x - y\|_2 \leq \varepsilon\} = A + \varepsilon B_2^n$$

Theorem 5.3 (Isoperimetric inequality on the sphere)**.** *Let* $\varepsilon > 0$*. Then, among all sets* $A \subset S^{n-1}$ *with given area* $\sigma_{n-1}(A)$*, the spherical caps minimize the area of the neighborhood* $\sigma_{n-1}(A_\varepsilon)$*, where*

$$A_\varepsilon := \left\{x \in S^{n-1} : \exists y \in A \text{ such that } \|x - y\|_2 \leq \varepsilon\right\}$$

Lemma 5.4 (blow-up)**.** *Let A be a subset of the sphere* $\sqrt{n}S^{n-1}$*, and let* σ *denote the normalized area on that sphere. If* $\sigma(A) \geq 1/2$*, then, for every* $t \geq 0$*,*

$$\sigma(A_t) \geq 1 - 2\exp\left(-ct^2\right)$$

Lemma 5.5 (Concentration about expectation and median are equivalent)**.** *Consider a random variable Z with median M. Show that*

$$c\|Z - \mathbb{E}Z\|_{\psi_2} \leq \|Z - M\|_{\psi_2} \leq C\|Z - \mathbb{E}Z\|_{\psi_2}$$

Theorem 5.6 (Gaussian isoperimetric inequality)**.** *Let* $\varepsilon > 0$*. Then, among all sets* $A \subset \mathbb{R}^n$ *with fixed Gaussian measure* $\gamma_n(A)$*, the half spaces minimize the Gaussian measure of the neighborhood* $\gamma_n(A_\varepsilon)$*.*

Theorem 5.7 (Gaussian concentration)**.** *Consider a random vector* $X \sim N(0, I_n)$ *and a Lipschitz function* $f : \mathbb{R}^n \to \mathbb{R}$ *(with respect to the Euclidean metric). Then*

$$\|f(X) - \mathbb{E}f(X)\|_{\psi_2} \leq C\|f\|_{Lip}$$

Theorem 5.8 (Concentration on the Hamming cube)**.** *Consider a random vector* $X \sim$ *Unif* $\{0,1\}^n$*. (Thus, the coordinates of X are independent* $\text{Ber}(1/2)$ *random variables.) Consider a function* $f : \{0,1\}^n \to \mathbb{R}$*. Then*

$$\|f(X) - \mathbb{E}f(X)\|_{\psi_2} \leq \frac{C\|f\|_{Lip}}{\sqrt{n}}$$

Theorem 5.9 (Concentration on the continuous cube)**.** *Consider a random vector* $X \sim \text{Unif}([0,1]^n)$*. (Thus, the coordinates of X are independent random variables uniformly distributed on* $[0,1]$*.) Consider a Lipschitz function* $f : [0,1]^n \to \mathbb{R}$*. (The Lipschitz norm is with respect to the Euclidean distance.) Then*

$$\|f(X) - \mathbb{E}f(X)\|_{\psi_2} \leq C\|f\|_{Lip}$$

Theorem 5.10 (Concentration on the Euclidean ball). *Consider the random vector* $X \sim \text{Unif}\left(\sqrt{n}B_2^n\right)$. *Consider a Lipschitz function* $f : \sqrt{n}B_2^n \to \mathbb{R}$. *(The Lipschitz norm is with respect to the Euclidean distance.) Then*

$$\|f(X) - \mathbb{E}f(X)\|_{\psi_2} \le C\|f\|_{Lip}$$

Theorem 5.11 (Concentration of concave density). *Consider a random vector* X *in* $\mathbb{R}^n$ *whose density has the form* $f(x) = e^{-U(x)}$ *for some function* $U : \mathbb{R}^n \to \mathbb{R}$. *Assume there exists* $\kappa > 0$ *such that* [12]
$\text{Hess}\, U(x) \succeq \kappa I_n \quad$ *for all* $x \in \mathbb{R}^n$ *Then any Lipschitz function* $f : \mathbb{R}^n \to \mathbb{R}$ *satisfies*

$$\|f(X) - \mathbb{E}f(X)\|_{\psi_2} \le \frac{C\|f\|_{Lip}}{\sqrt{\kappa}}$$

Theorem 5.12 (Talagrand's concentration inequality). *Consider a random vector* $X = (X_1, \dots, X_n)$ *whose coordinates are independent and satisfy*

$$|X_i| \le 1 \quad \textit{almost surely.}$$

Then for any convex Lipschitz function $f : [0,1]^n \to \mathbb{R}$

$$\|f(X) - \mathbb{E}f(X)\|_{\psi_2} \le \frac{C\|f\|_{Lip}}{\sqrt{\kappa}}$$

Remark: In particular, Talagrand's concentration inequality holds for any norm on $\mathbb{R}^n$.

Theorem 5.13 (Johnson-Lindenstrauss Lemma). *Let* $\mathcal{X}$ *be a set of* N *points in* $\mathbb{R}^n$ *and* $\varepsilon > 0$. *Assume that*

$$m \ge \left(C/\varepsilon^2\right) \log N$$

Consider a random m *-dimensional subspace* E *in* $\mathbb{R}^n$ *uniformly distributed in* $G_{n,m}$. *Denote the orthogonal projection onto* E *by* P. *Then, with probability at least* $1 - 2\exp\left(-c\varepsilon^2 m\right)$, *the scaled projection*

$$Q := \sqrt{\frac{n}{m}}P$$

is an approximate isometry on $\mathcal{X}$ *:*

$$(1-\varepsilon)\|x-y\|_2 \le \|Qx - Qy\|_2 \le (1+\varepsilon)\|x-y\|_2 \quad \textit{for all } x, y \in \mathcal{X}$$

Remark: Let A *be an* $m \times n$ *random matrix whose rows are independent, mean zero, subgaussian isotropic random vectors in* $\mathbb{R}^n$. *Show that the conclusion of JohnsonLindenstrauss lemma holds for* $Q = (1/\sqrt{n})A$.

Lemma 5.14 (Random projection). *Let* P *be a projection in* $\mathbb{R}^n$ *onto a random* m *-dimensional subspace uniformly distributed in* $G_{n,m}$. *Let* $z \in \mathbb{R}^n$ *be a (fixed) point and* $\varepsilon > 0$. *Then:*
(a)

$$\left(\mathbb{E}\|Pz\|_2^2\right)^{1/2} = \sqrt{\frac{m}{n}}\|z\|_2$$

(b) With probability at least $1 - 2\exp\left(-c\varepsilon^2 m\right)$, *we have*

$$(1-\varepsilon)\sqrt{\frac{m}{n}}\|z\|_2 \le \|Pz\|_2 \le (1+\varepsilon)\sqrt{\frac{m}{n}}\|z\|_2$$

Theorem 5.15 (Matrix Bernstein's inequality). *Let* $X_1, \dots, X_N$ *be independent, mean zero,* $n \times n$ *symmetric random matrices, such that* $\|X_i\| \le K$ *almost surely for all* i. *Then, for every* $t \ge 0$, *we have*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^N X_i\right\| \ge t\right\} \le 2n\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right)$$

Here $\sigma^2 = \left\|\sum_{i=1}^N \mathbb{E}X_i^2\right\|$ *is the norm of the matrix variance of the sum. In particular, we can express this bound as the mixture of sub-gaussian and sub-exponential tail, just like in the scalar Bernstein's inequality:*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^N X_i\right\| \geq t\right\} \leq 2n \exp\left[-c \cdot \min\left(\frac{t^2}{\sigma^2}, \frac{t}{K}\right)\right]$$

Remark: $\max(a,b) \asymp a+b$

For rectangular matrix $m \times n$*,*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^N X_i\right\| \geq t\right\} \leq 2(m+n) \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right)$$

where

$$\sigma^2 = \max\left(\left\|\sum_{i=1}^N \mathbb{E}X_i^\top X_i\right\|, \left\|\sum_{i=1}^N \mathbb{E}X_i X_i^\top\right\|\right)$$

Theorem 5.16 (Golden-Thompson inequality)**.** *For any* $n \times n$ *symmetric matrices* A *and* B*, we have*

$$\operatorname{tr}\left(e^{A+B}\right) \leq \operatorname{tr}\left(e^A e^B\right)$$

Unfortunately, Goldon-Thpmpson inequality does not hold for three or more matrices: in general, the inequality $\operatorname{tr}\left(e^{A+B+C}\right) \leq \operatorname{tr}\left(e^A e^B e^C\right)$ *may fail.*

Theorem 5.17 (Lieb's inequality)**.** *Let* H *be an* $n \times n$ *symmetric matrix. Define the function on matrices*

$$f(X) := \operatorname{tr} \exp(H + \log X)$$

Then f *is concave on the space on positive definite* $n \times n$ *symmetric matrices.*

Theorem 5.18 (Lieb's inequality for random matrices)**.** *Let* H *be a fixed* $n \times n$ *symmetric matrix and* Z *be a random* $n \times n$ *symmetric matrix. Then*

$$\mathbb{E} \operatorname{tr} \exp(H+Z) \leq \operatorname{tr} \exp\left(H + \log \mathbb{E}e^Z\right)$$

Lemma 5.19 (Moment generating function)**.** *Let* X *be an* $n \times n$ *symmetric mean zero random matrix such that* $\|X\| \leq K$ *almost surely. Then*

$$\mathbb{E} \exp(\lambda X) \preceq \exp\left(g(\lambda)\mathbb{E}X^2\right) \qquad \text{where} \qquad g(\lambda) = \frac{\lambda^2/2}{1 - |\lambda|K/3}$$

provided that $|\lambda| < 3/K$*.*

Theorem 5.20 (Matrix Bernstein's inequality: expectation)**.** *Let* $X_1, \ldots, X_N$ *be independent, mean zero,* $n \times n$ *symmetric random matrices, such that* $\|X_i\| \leq K$ *almost surely for all* i*. Deduce from Bernstein's inequality that*

$$\mathbb{E}\left\|\sum_{i=1}^N X_i\right\| \lesssim \left\|\sum_{i=1}^N \mathbb{E}X_i^2\right\|^{1/2} \sqrt{1+\log n} + K(1+\log n).$$

Theorem 5.21 (Matrix Hoeffding's inequality)**.** *Let* $\varepsilon_1, \ldots, \varepsilon_n$ *be independent symmetric Bernoulli random variables and let* $A_1, \ldots, A_N$ *be symmetric* $n \times n$ *matrices (deterministic). Prove that, for any* $t \geq 0$*, we have*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^N \varepsilon_i A_i\right\| \geq t\right\} \leq 2n \exp\left(-t^2/2\sigma^2\right)$$

where $\sigma^2 = \left\|\sum_{i=1}^N A_i^2\right\|$*.*

Theorem 5.22 (Matrix Khintchine's inequality)**.** *Let $\varepsilon_1, \dots, \varepsilon_N$ be independent symmetric Bernoulli random variables and let $A_1, \dots, A_N$ be symmetric $n \times n$ matrices (deterministic).*

$$\mathbb{E} \left\| \sum_{i=1}^N \varepsilon_i A_i \right\| \le C\sqrt{1+\log n} \left\| \sum_{i=1}^N A_i^2 \right\|^{1/2}$$

More generally, prove that for every $p \in [1, \infty)$ we have

$$\left(\mathbb{E} \left\| \sum_{i=1}^N \varepsilon_i A_i \right\|^p \right)^{1/p} \le C\sqrt{p+\log n} \left\| \sum_{i=1}^N A_i^2 \right\|^{1/2}$$

Theorem 5.23 (General covariance estimation)**.** *Let X be a random vector in $\mathbb{R}^n, n \ge 2$. Assume that for some $K \ge 1$,*

$$\|X\|_2 \le K \left(\mathbb{E}\|X\|_2^2 \right)^{1/2} \qquad \text{almost surely}$$

Then, for every positive integer m, we have

$$\mathbb{E} \|\Sigma_m - \Sigma\| \le C \left(\sqrt{\frac{K^2 n \log n}{m}} + \frac{K^2 n \log n}{m} \right) \|\Sigma\|$$

Theorem 5.24 (Low dimension covariance estimation)**.** *Intrinsic dimension $r = \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|}$*

$$\mathbb{E} \|\Sigma_m - \Sigma\| \le C \left(\sqrt{\frac{K^2 r \log n}{m}} + \frac{K^2 r \log n}{m} \right) \|\Sigma\|$$

In particular, this stronger bound implies that a sample of size

$$m \asymp \varepsilon^{-2} r \log n$$

is sufficient to estimate the covariance matrix.

Tail bound:

$$\|\Sigma_m - \Sigma\| \le C \left(\sqrt{\frac{K^2 r(\log n + u)}{m}} + \frac{K^2 r(\log n + u)}{m} \right) \|\Sigma\|$$

with probability at least $1 - 2e^{-u}$. Here $r = \operatorname{tr}(\Sigma)/\|\Sigma\| \le n$.

6 Chapter 5: Quadratic forms, symmetrization and contraction

High dimension probability in Data science

Chapter 5: Quadratic forms, symmetrization and contraction

Author: Roman Vershynin *Learner: Weihao Li*

Lemma 6.1. *Let Y and Z be independent random variables such that $\mathbb{E}Z = 0$. Then, for every convex function F, one has*

$$\mathbb{E}F(Y) \leq \mathbb{E}F(Y+Z)$$

Theorem 6.2 (Decoupling)**.** *Let A be an $n \times n$, diagonal-free matrix (i.e. the diagonal entries of A equal zero). Let $X = (X_1, \ldots, X_n)$ be a random vector with independent mean zero coordinates X_i. Then, for every convex function $F : \mathbb{R} \to \mathbb{R}$, one has*

$$\mathbb{E}F\left(X^\top AX\right) \leq \mathbb{E}F\left(4X^\top AX'\right)$$

where X' is an independent copy of X.

Stronger version: for any square matrix $A = (a_{ij})$ we have

$$\mathbb{E}F\left(\sum_{i,j:i\neq j} a_{ij}X_iX_j\right) \leq \mathbb{E}F\left(4\sum_{i,j} a_{ij}X_iX_j'\right)$$

Theorem 6.3 (Hanson-Wright inequality)**.** *Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean zero, sub-gaussian coordinates. Let A be an $n \times n$ matrix. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|X^\top AX - \mathbb{E}X^\top AX\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|}\right)\right]$$

where $K = \max_i \|X_i\|_{\psi_2}$

Lemma 6.4 (MGF of Gaussian chaos)**.** *Let $X, X' \sim N(0, I_n)$ be independent and let $A = (a_{ij})$ be an $n \times n$ matrix. Then*

$$\mathbb{E}\exp\left(\lambda X^\top AX'\right) \leq \exp\left(C\lambda^2\|A\|_F^2\right)$$

for all λ satisfying $|\lambda| \leq c/\|A\|$.

Lemma 6.5 (Comparison)**.** *Consider independent, mean zero, sub-gaussian random vectors X, X' in $\mathbb{R}^n$ with $\|X\|_{\psi_2} \leq K$ and $\|X'\|_{\psi_2} \leq K$. Consider also independent random vectors $g, g' \sim N(0, I_n)$. Let A be an $n \times n$ matrix. Then*

$$\mathbb{E}\exp\left(\lambda X^\top AX'\right) \leq \mathbb{E}\exp\left(CK^2\lambda g^\top Ag'\right)$$

for any $\lambda \in \mathbb{R}$.

Theorem 6.6 (Higher-dimensional Hanson-Wright inequality)**.** *Let $X_1, \ldots, X_n$ be independent, mean zero, sub-gaussian random vectors in $\mathbb{R}^d$. Let $A = (a_{ij})$ be an $n \times n$ matrix. Prove that for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i,j:i\neq j}^{n} a_{ij}\langle X_i, X_j\rangle\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^4d\|A\|_F^2}, \frac{t}{K^2\|A\|}\right)\right]$$

where $K = \max_i \|X_i\|_{\psi_2}$.

Theorem 6.7 (Concentration of random vectors). *Let B be an $m \times n$ matrix, and let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean zero, unit variance, sub-gaussian coordinates. Then*

$$\| \|BX\|_2 - \|B\|_F \|_{\psi_2} \leq CK^2 \|B\|$$

where $K = \max_i \|X_i\|_{\psi_2}$

Lemma 6.8 (Symmetrization). *Let $X_1, \dots, X_N$ be independent, mean zero random vectors in a normed space. Then*

$$\frac{1}{2}\mathbb{E}\left\|\sum_{i=1}^N \varepsilon_i X_i\right\| \leq \mathbb{E}\left\|\sum_{i=1}^N X_i\right\| \leq 2\mathbb{E}\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|.$$

Lemma 6.9 (Symmetrization general). *Let $F : \mathbb{R}_+ \to \mathbb{R}$ be an increasing, convex function. Show that the same inequalities in Lemma 6.8 hold if the norm $\|\cdot\|$ is replaced with $F(\|\cdot\|)$, namely*

$$\mathbb{E}F\left(\frac{1}{2}\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|\right) \leq \mathbb{E}F\left(\left\|\sum_{i=1}^N X_i\right\|\right) \leq \mathbb{E}F\left(2\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|\right)$$

Theorem 6.10 (Norms of random matrices with non-i.i.d. entries). *Let A be an $n \times n$ symmetric random matrix whose entries on and above the diagonal are independent, mean zero random variables. Then*

$$\mathbb{E}\|A\| \leq C\sqrt{\log n} \cdot \mathbb{E}\max_i \|A_i\|_2$$

where A_i denote the rows of A.

Remark: do not confuse with norm for gaussian matrix, this is a general result, hence not sharp.

For rectangular matrix:

$$\mathbb{E}\|A\| \leq C\sqrt{\log(m+n)}\left(\mathbb{E}\max_i \|A_i\|_2 + \mathbb{E}\max_j \left\|A^j\right\|_2\right)$$

where A_i and A^j denote the rows and columns of A, respectively.

Theorem 6.11 (Matrix completion). $\mathrm{Rank}(X) = r, r \ll n, p = m/n^2$

$$Y_{ij} := \delta_{ij} X_{ij} \quad \text{where} \quad \delta_{ij} \sim \mathrm{Ber}(p) \text{ are independent.}$$

we are shown m entries of X on average. Let $\hat{X}$ be a best rank r approximation to $p^{-1}Y$. Then

$$\mathbb{E}\frac{1}{n}\|\hat{X} - X\|_F \leq C\sqrt{\frac{rn\log n}{m}}\|X\|_\infty$$

as long as $m \geq n \log n$. Here $\|X\|_\infty = \max_{i,j} |X_{ij}|$ is the maximum magnitude of the entries of X.

Theorem 6.12 (Contraction Principle). *Let $\varepsilon_1, \varepsilon_2, \varepsilon_3, \dots$ a sequence of independent symmetric Bernoulli random variables, $x_1, \dots, x_N$ be (deterministic) vectors in some normed space, and let $a = (a_1, \dots, a_n) \in \mathbb{R}^n$. Then*

$$\mathbb{E}\left\|\sum_{i=1}^N a_i \varepsilon_i x_i\right\| \leq \|a\|_\infty \cdot \mathbb{E}\left\|\sum_{i=1}^N \varepsilon_i x_i\right\|$$

Lemma 6.13 (Symmetrization with Gaussians). *Let $X_1, \dots, X_N$ be independent, mean zero random vectors in a normed space. Let $g_1, \dots, g_N \sim N(0,1)$ be independent Gaussian random variables, which are also independent of X_i. Then*

$$\frac{c}{\sqrt{\log N}}\mathbb{E}\left\|\sum_{i=1}^N g_i X_i\right\| \leq \mathbb{E}\left\|\sum_{i=1}^N X_i\right\| \leq 3\mathbb{E}\left\|\sum_{i=1}^N g_i X_i\right\|.$$

Lemma 6.14 (Talagrand's contraction principle)**.** *Consider a bounded subset $T \subset \mathbb{R}^n$, and let $\varepsilon_1, \ldots, \varepsilon_n$ be independent symmetric Bernoulli random variables. Let $\phi_i : \mathbb{R} \to \mathbb{R}$ be contractions, i.e. Lipschitz functions with $\|\phi_i\|_{Lip} \leq 1$. Then*

$$\mathbb{E} \sup_{t \in T} \sum_{i=1}^{n} \varepsilon_i \phi_i (t_i) \leq \mathbb{E} \sup_{t \in T} \sum_{i=1}^{n} \varepsilon_i t_i$$

Lemma 6.15 (Gaussian contraction principle)**.** *Consider a bounded subset $T \subset \mathbb{R}^n$, and let $g_1, \ldots, g_n$ be independent $N(0,1)$. Let $\phi_i : \mathbb{R} \to \mathbb{R}$ be contractions, i.e. Lipschitz functions with $\|\phi_i\|_{Lip} \leq 1$. Then*

$$\mathbb{E} \sup_{t \in T} \sum_{i=1}^{n} g_i \phi_i (t_i) \leq \mathbb{E} \sup_{t \in T} \sum_{i=1}^{n} g_i t_i$$

7 Chapter 6: Random processes

High dimension probability in Data science

Chapter 6: Random processes

Author: Roman Vershynin *Learner: Weihao Li*

Definition 7.1. *Increments of the random process are defined as*

$$d(t,s) := \|X_t - X_s\|_{L^2} = \left(\mathbb{E}\left(X_t - X_s\right)^2\right)^{1/2}, \quad t,s \in T$$

Lemma 7.2 (Symmetrization for random processes)**.** *Let $X_1(t),\dots,X_N(t)$ be N independent, mean zero random processes indexed by points $t \in T$. Let $\varepsilon_1,\dots,\varepsilon_N$ be independent symmetric Bernoulli random variables. Prove that*

$$\frac{1}{2}\mathbb{E}\sup_{t\in T}\sum_{i=1}^{N}\varepsilon_i X_i(t) \le \mathbb{E}\sup_{t\in T}\sum_{i=1}^{N} X_i(t) \le 2\mathbb{E}\sup_{t\in T}\sum_{i=1}^{N}\varepsilon_i X_i(t)$$

Theorem 7.3 (Slepian's inequality)**.** Let $(X_t)_{t\in T}$ *and* $(Y_t)_{t\in T}$ *be two mean zero Gaussian processes. Assume that for all $t,s \in T$, we have*

$$\mathbb{E}X_t^2 = \mathbb{E}Y_t^2 \quad \text{and} \quad \mathbb{E}\left(X_t - X_s\right)^2 \le \mathbb{E}\left(Y_t - Y_s\right)^2$$

Then for every $\tau \in \mathbb{R}$ we have

$$\mathbb{P}\left\{\sup_{t\in T} X_t \ge \tau\right\} \le \mathbb{P}\left\{\sup_{t\in T} Y_t \ge \tau\right\}$$

Consequently,

$$\mathbb{E}\sup_{t\in T} X_t \le \mathbb{E}\sup_{t\in T} Y_t$$

Whenever the tail comparison inequality (7.3) holds, we say that the random variable X is stochastically dominated by the random variable Y.

Lemma 7.4 (Stein identity)**.**

- *If $X \sim N\left(0,\sigma^2\right)$, for any differentiable function $f:\mathbb{R}\to\mathbb{R}$ we have*

$$\mathbb{E}Xf(X) = \sigma^2\mathbb{E}f'(X)$$

- *Multivariate stein identity: Let $X \sim N(0,\Sigma)$. Then for any differentiable function $f:\mathbb{R}^n\to\mathbb{R}$ we have*

$$\mathbb{E}Xf(X) = \Sigma\cdot\mathbb{E}\nabla f(X)$$

$$\iff \mathbb{E}X_i f(X) = \sum_{j=1}^{n}\Sigma_{ij}\mathbb{E}\frac{\partial f}{\partial x_j}(X), \quad i = 1,\dots,n$$

Lemma 7.5 (Gaussian interpolation)**.** *Consider two independent Gaussian random vectors $X \sim N\left(0,\Sigma^X\right)$ and $Y \sim N\left(0,\Sigma^Y\right)$. Define the interpolation Gaussian vector*

$$Z(u) := \sqrt{u}X + \sqrt{1-u}Y, \quad u \in [0,1]$$

Then for any twice-differentiable function $f:\mathbb{R}^n\to\mathbb{R}$, we have

$$\frac{d}{du}\mathbb{E}f(Z(u)) = \frac{1}{2}\sum_{i,j=1}^{n}\left(\Sigma_{ij}^X - \Sigma_{ij}^Y\right)\mathbb{E}\left[\frac{\partial^2 f}{\partial x_i\partial x_j}(Z(u))\right]$$

Lemma 7.6 (Slepian's inequality, functional form)**.** *Consider two mean zero Gaussian random vectors X and Y in $\mathbb{R}^n$. Assume that for all $i, j = 1, \dots, n$, we have*

$$\mathbb{E}X_i^2 = \mathbb{E}Y_i^2 \quad \text{and} \quad \mathbb{E}\left(X_i - X_j\right)^2 \le \mathbb{E}\left(Y_i - Y_j\right)^2$$

Consider a twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \ge 0 \quad \text{for all } i \ne j$$

Then

$$\mathbb{E}f(X) \ge \mathbb{E}f(Y)$$

Theorem 7.7 (Sudakov-Fernique's inequality)**.** *Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two mean zero Gaussian processes. Assume that for all $t, s \in T$, we have*

$$\mathbb{E}\left(X_t - X_s\right)^2 \le \mathbb{E}\left(Y_t - Y_s\right)^2$$

Then

$$\mathbb{E}\sup_{t\in T} X_t \le \mathbb{E}\sup_{t\in T} Y_t$$

Theorem 7.8 (Gordon's inequality)**.** *Let $(X_{ut})_{u\in U, t\in T}$ and $Y = (Y_{ut})_{u\in U, t\in T}$ be two mean zero Gaussian processes indexed by pairs of points (u, t) in a product set $U \times T$. Assume that we have*

$$\mathbb{E}X_{ut}^2 = \mathbb{E}Y_{ut}^2, \quad \mathbb{E}\left(X_{ut} - X_{us}\right)^2 \le \mathbb{E}\left(Y_{ut} - Y_{us}\right)^2 \quad \text{for all } u, t, s$$

$\mathbb{E}\left(X_{ut} - X_{vs}\right)^2 \ge \mathbb{E}\left(Y_{ut} - Y_{vs}\right)^2$ *for all $u \ne v$ and all t, s Then for every $\tau \ge 0$ we have*

$$\mathbb{P}\left\{\inf_{u\in U}\sup_{t\in T} X_{ut} \ge \tau\right\} \le \mathbb{P}\left\{\inf_{u\in U}\sup_{t\in T} Y_{ut} \ge \tau\right\}$$

Consequently,

$$\mathbb{E}\inf_{u\in U}\sup_{t\in T} X_{ut} \le \mathbb{E}\inf_{u\in U}\sup_{t\in T} Y_{ut}$$

Theorem 7.9 (Norms of Gaussian random matrices)**.** *Let A be an $m \times n$ matrix with independent $N(0, 1)$ entries. Then*

$$\mathbb{E}\|A\| \le \sqrt{m} + \sqrt{n}$$
$$\mathbb{P}\{\|A\| \ge \sqrt{m} + \sqrt{n} + t\} \le 2\exp\left(-ct^2\right)$$
$$\mathbb{E}s_n(A) \ge \sqrt{m} - \sqrt{n}$$

Lemma 7.10 (Symmetric random matrix)**.** *Gaussian orthogonal ensemble (GOE): diagonal entries are independent $N(0, 2)$ random variables.*

$$\mathbb{E}\|A\| \le 2\sqrt{n}$$

tail bound

$$\mathbb{P}\{\|A\| \ge 2\sqrt{n} + t\} \le 2\exp\left(-ct^2\right)$$

Theorem 7.11 (Sudakov's minoration inequality)**.** *Let $(X_t)_{t\in T}$ be a mean zero Gaussian process. Then, for any $\varepsilon \ge 0$, we have*

$$\mathbb{E}\sup_{t\in T} X_t \ge c\varepsilon\sqrt{\log \mathcal{N}(T, d, \varepsilon)}$$

where $d(t, s) := \|X_t - X_s\|_{L^2} = \left(\mathbb{E}\left(X_t - X_s\right)^2\right)^{1/2}$.

Corollary 7.12 (Sudakov's minoration inequality in $\mathbb{R}^n$)**.** *Let $T \subset \mathbb{R}^n$. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{E} \sup_{t \in T} \langle g, t \rangle \geq c\varepsilon \sqrt{\log \mathcal{N}(T, \varepsilon)}$$

Here $\mathcal{N}(T, \varepsilon)$ is the covering number of T by Euclidean balls − the smallest number of Euclidean balls with radii ε and centers in T that cover T.

Theorem 7.13. *$x_1, \ldots, x_N$ denote the vertices of P*

$$\mathbb{E} \sup_{t \in P} \langle g, t \rangle = \mathbb{E} \sup_{i \leq N} \langle g, x_i \rangle$$

The equality here follows since the maximum of the linear function on the convex set P is attained at an extreme point, i.e. at a vertex of P.

Proposition 7.14 (Gaussian width)**.**

- *$w(T)$ is finite if and only if T is bounded.*
- *Gaussian width is invariant under affine unitary transformations. Thus, for every orthogonal matrix U and any vector y, we have*
$$w(UT + y) = w(T)$$
- *Gaussian width is invariant under taking convex hulls. Thus,*
$$w(\mathrm{conv}(T)) = w(T)$$
- *Gaussian width respects Minkowski addition of sets and scaling. Thus, for $T, S \subset \mathbb{R}^n$ and $a \in \mathbb{R}$ we have*
$$w(T + S) = w(T) + w(S); \quad w(aT) = |a| w(T)$$
- *We have*
$$w(T) = \frac{1}{2} w(T - T) = \frac{1}{2} \mathbb{E} \sup_{x, y \in T} \langle g, x - y \rangle$$
- *(Gaussian width and diameter). We have*
$$\frac{1}{\sqrt{2\pi}} \cdot \mathrm{diam}(T) \leq w(T) \leq \frac{\sqrt{n}}{2} \cdot \mathrm{diam}(T)$$
- *Gaussian width under linear transformations*
- $$w(AT) \leq \|A\| w(T)$$

Definition 7.15 (Spherical width)**.** *The spherical width of a subset $T \subset R^n$ is defined as*

$$w_s(T) := \mathbb{E} \sup_{x \in T} \langle \theta, x \rangle \quad \text{where} \quad \theta \sim \mathrm{Unif}\left(S^{n-1}\right)$$

Lemma 7.16 (Gaussian vs. spherical widths)**.** *We have*

$$(\sqrt{n} - C) w_s(T) \leq w(T) \leq (\sqrt{n} + C) w_s(T)$$

Definition 7.17 (Squared gaussian width)**.**

$$h(T)^2 := \mathbb{E} \sup_{t \in T} \langle g, t \rangle^2, \quad \text{where} \quad g \sim N(0, I_n)$$

Lemma 7.18.

$$w(T-T) \le h(T-T) \le w(T-T) + C_1 \operatorname{diam}(T) \le Cw(T-T)$$

In particular, we have

$$2w(T) \le h(T-T) \le 2Cw(T)$$

Definition 7.19 (stable dimension)**.**

$$d(T) := \frac{h(T-T)^2}{\operatorname{diam}(T)^2} \asymp \frac{w(T)^2}{\operatorname{diam}(T)^2}$$

The stable dimension is always bounded by the algebraic dimension: $d(T) \le \dim(T)$

Definition 7.20 (Stable rank)**.**

$$r(A) := \frac{\|A\|_F^2}{\|A\|^2}$$

Theorem 7.21 (Sizes of random projections of sets)**.** *Consider a bounded set* $T \subset \mathbb{R}^n$*. Let* P *be a projection in* $\mathbb{R}^n$ *onto a random* m *-dimensional subspace* $E \sim \operatorname{Unif}(G_{n,m})$*. Then, with probability at least* $1 - 2e^{-m}$*, we have*

$$\operatorname{diam}(PT) \le C\left[w_s(T) + \sqrt{\frac{m}{n}}\operatorname{diam}(T)\right].$$

Optimal:

$$\mathbb{E}\operatorname{diam}(PT) \ge c\left[w_s(T) + \sqrt{\frac{m}{n}}\operatorname{diam}(T)\right]$$

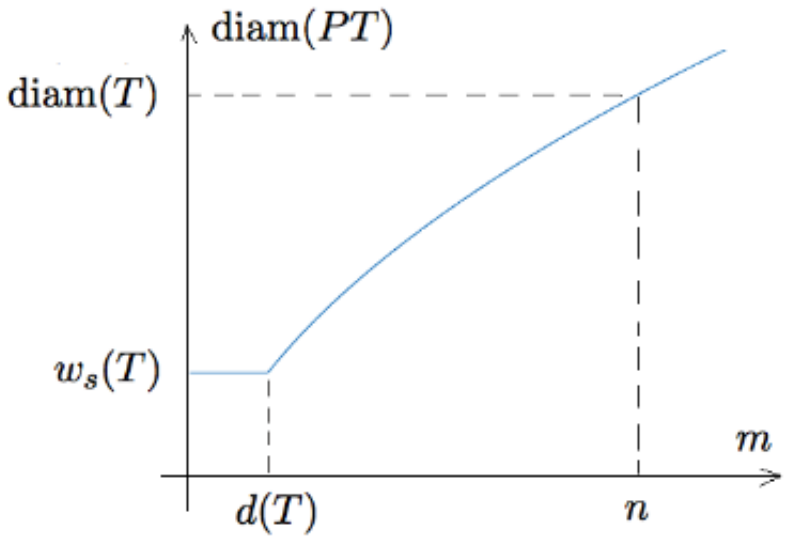

Figure 1: The diameter of a random m -dimensional projection of a set T as a function of m.

8 Chapter 7: Chaining

High dimension probability in Data science

Chapter 7: Chaining

Author: Roman Vershynin *Learner: Weihao Li*

Definition 8.1 (Sub-gaussian increment). *Consider a random process $(X_t)_{t\in T}$ on a metric space (T,d). We say that the process has sub-gaussian increments if there exists $K\geq 0$ such that*

$$\|X_t - X_s\|_{\psi_2} \leq Kd(t,s) \quad \text{for all } t,s\in T$$

where $d(t,s) := \|X_t - X_s\|_{L^2}, \quad t,s\in T$

Theorem 8.2 (Dudley's integral inequality). *Let $(X_t)_{t\in T}$ be a mean zero random process on a metric space (T,d) with sub-gaussian increments defined above. Then*

$$\mathbb{E}\sup_{t\in T} X_t \leq CK\int_0^\infty \sqrt{\log \mathcal{N}(T,d,\varepsilon)}d\varepsilon$$

Remark:

$$\mathbb{E}\sup_{t\in T} X_t \leq CK\int_0^{\mathrm{diam}(T)} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}d\varepsilon$$

Theorem 8.3 (Discrete Dudley's inequality). *Let $(X_t)_{t\in T}$ be a mean zero random process on a metric space (T,d) with sub-gaussian increments defined above. Then*

$$\mathbb{E}\sup_{t\in T} X_t \leq CK\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log \mathcal{N}\left(T,d,2^{-k}\right)}$$

Theorem 8.4 (Dudley's integral inequality: tail bound). *Let $(X_t)_{t\in T}$ be a random process on a metric space (T,d) with sub-gaussian increments as in (8.1). Then, for every $u\geq 0$, the event*

$$\sup_{t,s\in T}|X_t - X_s| \leq CK\left[\int_0^\infty \sqrt{\log \mathcal{N}(T,d,\varepsilon)}d\varepsilon + u\cdot \mathrm{diam}(T)\right]$$

holds with probability at least $1-2\exp\left(-u^2\right)$.

Theorem 8.5 (Dudley's inequality for sets in $\mathbb{R}^n$). *For any set $T\subset R^n$, we have*

$$w(T) \leq C\int_0^\infty \sqrt{\log \mathcal{N}(T,\varepsilon)}d\varepsilon$$

For example:

$$w\left(B_2^n\right) \leq C\int_0^1 \sqrt{n\log\frac{3}{\varepsilon}}d\varepsilon \leq C_1\sqrt{n}$$

Theorem 8.6 (Uniform law of large number). $\mathcal{F} := \{f:[0,1]\to\mathbb{R}, \|f\|_{Lip} \leq L\}$. *Let $X, X_1, X_2, \ldots, X_n$ be i.i.d. random variables taking values in* $[0,1]$. *Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f\left(X_i\right) - \mathbb{E}f(X)\right| \leq \frac{CL}{\sqrt{n}}$$

Definition 8.7 (VC dimension). *Consider a class $\mathcal{F}$ of Boolean functions on some domain Ω. We say that a subset $\Lambda \subseteq \Omega$ is shattered by $\mathcal{F}$ if any function $g : \Lambda \to \{0, 1\}$ can be obtained by restricting some function $f \in \mathcal{F}$ onto Λ. The VC dimension of $\mathcal{F}$, denoted $\mathrm{vc}(\mathcal{F})$, is the largest cardinality [1] of a subset $\Lambda \subseteq \Omega$ shattered by $\mathcal{F}$.*

Lemma 8.8 (Pajor's Lemma). *Let $\mathcal{F}$ be a class of Boolean functions on a finite set Ω. Then*

$$|\mathcal{F}| \leq |\ \{\Lambda \subseteq \Omega : \Lambda \text{ is shattered by } \mathcal{F}\}$$

We include the empty set $\Lambda = \emptyset$ in the counting on the right side.

Theorem 8.9 (Sauer-Shelah Lemma). *Let $\mathcal{F}$ be a class of Boolean functions on an n -point set Ω. Then*

$$|\mathcal{F}| \leq \sum_{k=0}^{d} \left(\begin{array}{c} n \\ k \end{array} \right) \leq \left(\frac{en}{d}\right)^d$$

where $d = \mathrm{vc}(\mathcal{F})$.

Definition 8.10 (distance on probability measure μ).

$$d(f, g) = \|f - g\|_{L^2(\mu)} = \left(\int_\Omega |f - g|^2 d\mu \right)^{1/2}, \quad f, g \in \mathcal{F}$$

Theorem 8.11 (Covering numbers via VC dimension). *Let $\mathcal{F}$ be a class of Boolean functions on a probability space (Ω, Σ, μ). Then, for every $\varepsilon \in (0, 1)$, we have*

$$\mathcal{N}\left(\mathcal{F}, L^2(\mu), \varepsilon\right) \leq \left(\frac{2}{\varepsilon}\right)^{Cd}$$

Lemma 8.12 (Dimension reduction). *Let $\mathcal{F}$ be a class of N Boolean functions on a probability space (Ω, Σ, μ). Assume that all functions in $\mathcal{F}$ are ε -separated, that is*

$$\|f - g\|_{L^2(\mu)} > \varepsilon \text{ for all distinct } f, g \in \mathcal{F} \ .$$

Then there exist a number $n \leq C\varepsilon^{-4} \log N$ and an n -point subset $\Omega_n \subset \Omega$ such that the uniform probability measure μ_n on Ω_n satisfies

$$\|f - g\|_{L^2(\mu_n)} > \frac{\varepsilon}{2} \quad \text{for all distinct } f, g \in \mathcal{F}.$$

Theorem 8.13 (Empirical processes via VC dimension). *Let $\mathcal{F}$ be a class of Boolean functions on a probability space (Ω, Σ, μ) with finite VC dimension $\mathrm{vc}(\mathcal{F}) \geq 1$. Let $X, X_1, X_2, \ldots, X_n$ be independent random points in Ω distributed according to the law μ. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f\left(X_i\right) - \mathbb{E} f(X) \right| \leq C \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}$$

Lemma 8.14 (Symmetrization for empirical processes). *Let $\mathcal{F}$ be a class of functions on a probability space (Ω, Σ, μ). Let $X, X_1, X_2, \ldots, X_n$ be random points in Ω distributed according to the law μ. Prove that*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f\left(X_i\right) - \mathbb{E} f(X) \right| \leq 2\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f\left(X_i\right) \right|$$

where $\varepsilon_1, \varepsilon_2, \ldots$ are independent symmetric Bernoulli random variables (which are also independent of $X_1, X_2, \ldots$).

Theorem 8.15 (Glivenko-Cantelli Theorem: non-asymptotic)**.** *Let $X_1, \ldots, X_n$ be independent random variables with common cumulative distribution function F. Then*

$$\mathbb{E} \left\| F_n - F \right\|_\infty = \mathbb{E} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq \frac{C}{\sqrt{n}}$$

Definition 8.16. *Ideally, we would like to find a function f^* from the hypothesis space $\mathcal{F}$ which would minimize the risk $R(f) = \mathbb{E}(f(X) - T(X))^2$, that is*

$$f^* := \arg\min_{f \in \mathcal{F}} R(f)$$

The empirical risk for a function $f : \Omega \to \mathbb{R}$ is defined as

$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} \left(f\left(X_i\right) - T\left(X_i\right) \right)^2 .$$

$$f_n^* := \arg\min_{f \in \mathcal{F}} R_n(f)$$

The main question is: how large is the excess risk:

$$R\left(f_n^*\right) - R\left(f^*\right)$$

Theorem 8.17 (Excess risk via VC dimension)**.** *Assume that the target T is a Boolean function, and the hypothesis space $\mathcal{F}$ is a class of Boolean functions with finite VC dimension* $\operatorname{vc}(\mathcal{F}) \geq 1$*. Then*

$$\mathbb{E} R\left(f_n^*\right) \leq R\left(f^*\right) + C\sqrt{\frac{\operatorname{vc}(\mathcal{F})}{n}}$$

Lemma 8.18 (Excess risk via uniform deviations)**.** *We have*

$$R\left(f_n^*\right) - R\left(f^*\right) \leq 2 \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$$

pointwise.

Theorem 8.19 (Learning in the class of Lipschitz functions)**.** *Consider*

$$\mathcal{F} := \{ f : [0,1] \to \mathbb{R}, \|f\|_{Lip} \leq L \}$$

and a target function $T : [0,1] \to [0,1]$.

- *random process $X_f := R_n(f) - R(f)$ has sub-gaussian increment*

$$\|X_f - X_g\|_{\psi_2} \leq \frac{CL}{\sqrt{n}} \|f - g\|_\infty \quad \text{for all } f, g \in \mathcal{F}$$

- $$\mathbb{E} \sup_{f \in \mathcal{F}} |R_n(f) - R(f)| \leq \frac{C(L+1)}{\sqrt{n}}$$

- $$R\left(f_n^*\right) - R\left(f^*\right) \leq \frac{C(L+1)}{\sqrt{n}}$$

Definition 8.20. $(T_k)_{k=0}^{\infty}$ *is called an admissible sequence if*

$$|T_0| = 1, \quad |T_k| \leq 2^{2^k}, \quad k = 1, 2, \ldots$$

Definition 8.21 (Talagrand's γ_2 functional)**.** *Let (T, d) be a metric space. The γ_2 functional of T is defined as*

$$\gamma_2(T, d) = \inf_{(T_k)} \sup_{t \in T} \sum_{k=0}^{\infty} 2^{k/2} d\left(t, T_k\right)$$

where the infimum is with respect to all admissible sequences.

Theorem 8.22 (Generic chaining bound)**.** *Let $(X_t)_{t \in T}$ be a mean zero random process on a metric space (T, d) with sub-gaussian increments . Then*

$$\mathbb{E} \sup_{t \in T} X_t \leq CK\gamma_2(T, d)$$

Theorem 8.23 (Generic chaining: tail bound)**.** Let $(X_t)_{t \in T}$ *be a random process on a metric space (T, d) with sub-gaussian increments. Then, for every $u \geq 0$, the event*

$$\sup_{t,s \in T} |X_t - X_s| \leq CK \left[\gamma_2(T, d) + u \cdot \operatorname{diam}(T)\right]$$

holds with probability at least $1 - 2\exp\left(-u^2\right)$.

Theorem 8.24 (Talagrand's majorizing measure theorem)**.** *Let $(X_t)_{t \in T}$ be a mean zero Gaussian process on a set T. Consider the canonical metric defined on T by (7.13), i.e. $d(t, s) = \|X_t - X_s\|_{L^2}$. Then*

$$c \cdot \gamma_2(T, d) \leq \mathbb{E} \sup_{t \in T} X_t \leq C \cdot \gamma_2(T, d)$$

Corollary 8.25 (Talagrand's comparison inequality)**.** *Let $(X_t)_{t \in T}$ be a mean zero random process on a set T and let $(Y_t)_{t \in T}$ be a mean zero Gaussian process. Assume that for all $t, s \in T$, we have*

$$\|X_t - X_s\|_{\psi_2} \leq K \|Y_t - Y_s\|_{L^2}$$

Then

$$\mathbb{E} \sup_{t \in T} X_t \leq CK\mathbb{E} \sup_{t \in T} Y_t$$

Corollary 8.26 (Talagrand's comparison inequality: geometric form)**.** Let $(X_x)_{x \in T}$ *be a mean zero random process on a subset $T \subset \mathbb{R}^n$. Assume that for all $x, y \in T$, we have*

$$\|X_x - X_y\|_{\psi_2} \leq K\|x - y\|_2$$

Then

$$\mathbb{E} \sup_{x \in T} X_x \leq CKw(T)$$

Theorem 8.27 (Sub-gaussian Chevet's inequality)**.** *Let A be an $m \times n$ random matrix whose entries A_{ij} are independent, mean zero, sub-gaussian random variables. Let $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$ be arbitrary bounded sets. Then*

$$\mathbb{E} \sup_{x \in T, y \in S} \langle Ax, y \rangle \leq CK[w(T) \operatorname{rad}(S) + w(S) \operatorname{rad}(T)]$$

where $K = \max_{ij} \|A_{ij}\|_{\psi_2}$.

9 Chapter 8: Deviations of random matrices and geometric consequences

High dimension probability in Data science

Ch 8: Deviations of random matrices and geometric consequences

Author: Roman Vershynin *Learner: Weihao Li*

Theorem 9.1 (Matrix deviation inequality)**.** *Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Then for any subset $T \subset \mathbb{R}^n$, we have*

$$\mathbb{E} \sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right| \le CK^2 \gamma(T)$$

Here $\gamma(T)$ is the Gaussian complexity, and $K = \max_i \|A_i\|_{\psi_2}$

Theorem 9.2 (Sub-gaussian increments)**.** *Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Then the random process*

$$X_x := \|Ax\|_2 - \sqrt{m}\|x\|_2$$

has sub-gaussian increments, namely

$$\|X_x - X_y\|_{\psi_2} \le CK^2 \|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

Here $K = \max_i \|A_i\|_{\psi_2}$

Lemma 9.3. *Let $x, y \in S^{n-1}$. Then*

$$\| \|Ax\|_2 - \|Ay\|_2 \|_{\psi_2} \le CK^2 \|x - y\|_2$$

Proposition 9.4 (Sizes of random projections of sets)**.** *Consider a bounded set $T \subset \mathbb{R}^n$. Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Then the scaled matrix*

$$P := \frac{1}{\sqrt{n}} A$$

(a "sub-gaussian projection") satisfies

$$\mathbb{E} \operatorname{diam}(PT) \le \sqrt{\frac{m}{n}} \operatorname{diam}(T) + CK^2 w_s(T)$$

Theorem 9.5 (Covariance estimation for lower-dimensional distributions)**.** *Let X be a sub-gaussian random vector in $\mathbb{R}^n$. More precisely, assume that there exists $K \ge 1$ such that*

$$\|\langle X, x \rangle\|_{\psi_2} \le K \|\langle X, x \rangle\|_{L^2} \quad \text{for any } x \in \mathbb{R}^n.$$

Then, for every positive integer m, we have

$$\mathbb{E} \|\Sigma_m - \Sigma\| \le CK^4 \left(\sqrt{\frac{r}{m}} + \frac{r}{m} \right) \|\Sigma\|$$

where $r = \operatorname{tr}(\Sigma)/\|\Sigma\|$ is the stable rank of $\Sigma^{1/2}$.

$$\|\Sigma_m - \Sigma\| \le CK^4 \left(\sqrt{\frac{r+u}{m}} + \frac{r+u}{m} \right) \|\Sigma\|$$

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

Proposition 9.6 (Additive Johnson-Lindenstrauss Lemma)**.** *Consider a set $\mathcal{X} \subset \mathbb{R}^n$. Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Then, with high probability (say, 0.99), the scaled matrix*

$$Q := \frac{1}{\sqrt{m}} A$$

satisfies

$$\|x - y\|_2 - \delta \le \|Qx - Qy\|_2 \le \|x - y\|_2 + \delta \quad \text{for all } x, y \in \mathcal{X}$$

where

$$\delta = \frac{CK^2 w(\mathcal{X})}{\sqrt{m}}$$

and $K = \max_i \|A_i\|_{\psi_2}$.

Theorem 9.7 (M^* bound)**.** *Consider a set $T \subset \mathbb{R}^n$. Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Then the random subspace $E = \ker A$ satisfies*

$$\mathbb{E} \operatorname{diam}(T \cap E) \le \frac{CK^2 w(T)}{\sqrt{m}}$$

where $K = \max_i \|A_i\|_{\psi_2}$.

Corollary 9.8 (Affine sections)**.**

$$\mathbb{E} \max_{z \in \mathbb{R}^n} \operatorname{diam}(T \cap E_z) \le \frac{CK^2 w(T)}{\sqrt{m}}$$

where $E_z = z + \ker A$.

Theorem 9.9 (Escape theorem)**.** *Consider a set $T \subset S^{n-1}$. Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. If*

$$m \ge CK^4 w(T)^2$$

then the random subspace $E = \ker A$ satisfies

$$T \cap E = \emptyset$$

with probability at least $1 - 2\exp\left(-\mathrm{cm}/\mathrm{K}^4\right)$. Here $K = \max_i \|A_i\|_{\psi_2}$.

10 Chapter 9: Sparse Recovery

High dimension probability in Data science

Chapter 9: Sparse Recovery

Author: Roman Vershynin *Learner: Weihao Li*

Theorem 10.1. *Truth $y = Ax, \quad x \in T$, optimization find $x' : y = Ax', \quad x' \in T$. Suppose the rows A_i of A are independent, isotropic and subgaussian random vectors. Then any solution $\widehat{x}$ of the optimization satisfies*

$$\mathbb{E}\|\widehat{x} - x\|_2 \leq \frac{CK^2 w(T)}{\sqrt{m}}$$

where $K = \max_i \|A_i\|_{\psi_2}$

Corollary 10.2 (Sparse recovery: guarantees)**.** *Consider optimization*

$$\text{Find } x' : y = Ax', \quad \|x'\|_1 \leq \sqrt{s}$$

Assume the unknown s-sparse signal $x \in \mathbb{R}^n$ satisfies $\|x\|_2 \leq 1$. Then x can be approximately recovered from the random measurement vector $y = Ax$ by a solution $\widehat{x}$ of the optimization. The recovery error satisfies

$$\mathbb{E}\|\widehat{x} - x\|_2 \leq CK^2\sqrt{\frac{s \log n}{m}}$$

Theorem 10.3 (Exact sparse recovery)**.** *Consider optimization program*

$$\text{minimize } \|x'\|_1 \text{ s.t. } y = Ax' \tag{1}$$

Suppose the rows A_i of A are independent, isotropic and sub-gaussian random vectors, and let $K := \max_i \|A_i\|_{\psi_2}$. Then the following happens with probability at least $1 - 2\exp\left(-\mathrm{cm}/K^4\right)$.

Assume an unknown signal $x \in \mathbb{R}^n$ is s-sparse and the number of measurements m satisfies

$$m \geq CK^4 s \log n$$

Then a solution $\widehat{x}$ of the program is exact, i.e.

$$\widehat{x} = x.$$

Definition 10.4 (RIP)**.** *An $m \times n$ matrix A satisfies the restricted isometry property (RIP) with parameters α, β and s if the inequality*

$$\alpha\|v\|_2 \leq \|Av\|_2 \leq \beta\|v\|_2$$

holds for all vectors $v \in \mathbb{R}^n$ such that $\|v\|_0 \leq s$

Theorem 10.5 (RIP implies exact recovery)**.** *Suppose an $m \times n$ matrix A satisfies RIP with some parameters α, β and $(1 + \lambda)s$, where $\lambda > (\beta/\alpha)^2$. Then every s-sparse vector $x \in \mathbb{R}^n$ can be recovered exactly by solving the program (1), i.e. the solution satisfies*

$$\widehat{x} = x.$$

Theorem 10.6 (Random matrices satisfy RIP)**.** *Consider an $m \times n$ matrix A whose rows A_i of A are independent, isotropic and sub-gaussian random vectors, and let $K := \max_i \|A_i\|_{\psi_2}$. Assume that*

$$m \geq CK^4 s \log(en/s)$$

Then, with probability at least $1 - 2\exp\left(-\mathrm{cm}/K^4\right)$, the random matrix A satisfies RIP with parameters $\alpha = 0.9\sqrt{m}, \beta = 1.1\sqrt{m}$ and s.

Theorem 10.7 (Performance of Lasso)**.** *Linear regression setting $Y = X\beta + w$ Consider the Lasso program*

$$\text{minimize } \|y - Ax'\|_2 \text{ s.t. } \|x'\|_1 \leq R \tag{2}$$

Suppose the rows A_i of A are independent, isotropic and sub-gaussian random vectors, and let $K := \max_i \|A_i\|_{\psi_2}$. Then the following happens with probability at least $1 - 2\exp(-s\log n)$.

Assume an unknown signal $x \in \mathbb{R}^n$ is s-sparse and the number of measurements m satisfies

$$m \geq CK^4 s \log n$$

Then a solution $\widehat{x}$ of the program (2) with $R := \|x\|_1$ is accurate, namely

$$\|\widehat{x} - x\|_2 \leq C\sigma\sqrt{\frac{s\log n}{m}},$$

where $\sigma = \|w\|_{L_2}/\sqrt{m}$.

11 Chapter 10: Dvoretzky-Milman's Theorem

High dimension probability in Data science

Chapter 10: Dvoretzky-Milman's Theorem

Author: Roman Vershynin *Learner: Weihao Li*

Definition 11.1. *Let V be a vector space. A function $f : V \to \mathbb{R}$ is called positive-homogeneous if $f(\alpha x) = \alpha f(x)$ for all $\alpha \geq 0$ and $x \in V$.*

The function f is called subadditive if

$$f(x + y) \leq f(x) + f(y) \text{ for all } x, y \in V$$

Theorem 11.2 (General matrix deviation inequality)**.** *Let A be an $m \times n$ Gaussian random matrix with i.i.d. $N(0, 1)$ entries. Let $f : \mathbb{R}^m \to \mathbb{R}$ be a positive-homogeneous and subadditive function, and let $b \in \mathbb{R}$ be such that*

$$f(x) \leq b\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n$$

Then for any subset $T \subset \mathbb{R}^m$, we have

$$\mathbb{E} \sup_{x \in T} |f(Ax) - \mathbb{E}f(Ax)| \leq Cb\gamma(T)$$

Here $\gamma(T)$ is the Gaussian complexity.

Lemma 11.3 (Sub-gaussian increments)**.** *Let A be an $m \times n$ Gaussian random matrix with i.i.d. $N(0, 1)$ entries, and let $f : \mathbb{R}^m \to \mathbb{R}$ be a positive homogenous and subadditive function satisfying (11.3). Then the random process*

$$X_x := f(Ax) - \mathbb{E}f(Ax)$$

has sub-gaussian increments with respect to the Euclidean norm, namely

$$\|X_x - X_y\|_{\psi_2} \leq Cb\|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

Corollary 11.4 (Johnson-Lindenstrauss Lemma for ℓ_1 norm)**.** *Let $\mathcal{X}$ be a set of N points in $\mathbb{R}^n$, let A be an $m \times n$ Gaussian matrix with i.i.d. $N(0, 1)$ entries, and let $\varepsilon \in (0, 1)$. Suppose that*

$$m \geq C(\varepsilon) \log N$$

With high probability the matrix $Q := \sqrt{\pi/2} \cdot m^{-1} A$ satisfies

$$(1 - \varepsilon)\|x - y\|_2 \leq \|Qx - Qy\|_1 \leq (1 + \varepsilon)\|x - y\|_2 \text{ for all } x, y \in \mathcal{X}$$

Corollary 11.5 (Johnson-Lindenstrauss Lemma for ℓ_∞ norm)**.** *Let $\mathcal{X}$ be a set of N points in $\mathbb{R}^n$, let A be an $m \times n$ Gaussian matrix with i.i.d. $N(0, 1)$ entries, and let $\varepsilon \in (0, 1)$. Suppose that*

$$m \geq N^{C(\varepsilon)}$$

With high probability the matrix $Q := \sqrt{\pi/2} \cdot m^{-1} A$ satisfies

$$(1 - \varepsilon)\|x - y\|_2 \leq \|Qx - Qy\|_1 \leq (1 + \varepsilon)\|x - y\|_2 \text{ for all } x, y \in \mathcal{X}$$

With high probability the matrix $Q := C(\log m)^{-1/2} A$, for some appropriate constant C, satisfies

$$(1 - \varepsilon)\|x - y\|_2 \leq \|Qx - Qy\|_\infty \leq (1 + \varepsilon)\|x - y\|_2 \quad \text{for all } x, y \in \mathcal{X}$$

Note that in this case $m \geq N$, so Q gives an almost isometric embedding (rather than a projection) of the set $\mathcal{X}$ into ℓ_∞.

Theorem 11.6 (General Chevet's inequality)**.** *Let A be an $m \times n$ Gaussian random matrix with i.i.d. $N(0,1)$ entries. Let $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$ be arbitrary bounded sets. Then*

$$\mathbb{E} \sup_{x \in T} \left| \sup_{y \in S} \langle Ax, y \rangle - w(S) \|x\|_2 \right| \leq C\gamma(T) \operatorname{rad}(S)$$

Theorem 11.7 (Random projections of sets)**.** *Let A be an $m \times n$ Gaussian random matrix with i.i.d. $N(0,1)$ entries, and $T \subset \mathbb{R}^n$ be a bounded set. Then the following holds with probability at least 0.99 :*

$$r_- B_2^m \subset \operatorname{conv}(AT) \subset r_+ B_2^m$$

where

$$r_\pm := w(T) \pm C\sqrt{m} \operatorname{rad}(T)$$

Theorem 11.8 (voretzky-Milman's theorem: Gaussian form)**.** *Let A be an $m \times n$ Gaussian random matrix with i.i.d. $N(0,1)$ entries, $T \subset \mathbb{R}^n$ be a bounded set, and let $\varepsilon \in (0,1)$. Suppose*

$$m \leq c\varepsilon^2 d(T)$$

where $d(T)$ is the stable dimension of T introduced in Section 7.6. Then with probability at least 0.99, we have

$$(1-\varepsilon)B \subset \operatorname{conv}(AT) \subset (1+\varepsilon)B$$

where B is a Euclidean ball with radius $w(T)$.

Corollary 11.9 (Gaussian cloud)**.** *Consider a Gaussian cloud of n points in $\mathbb{R}^m$, which is formed by i.i.d. random vectors $g_1, \ldots, g_n \sim N(0, I_m)$. Suppose that*

$$n \geq \exp(Cm)$$

with large enough absolute constant C. Show that with high probability, the convex hull the Gaussian cloud is approximately a Euclidean ball with radius $\sim \sqrt{\log n}$.

A random projection of a set T in $\mathbb{R}^n$ onto an m dimensional subspace approximately preserves the geometry of T if $m \gtrsim d(T)$. For smaller m, the projected set PT becomes approximately a round ball of diameter $\sim w_s(T)$, and its size does not shrink with m.