Probability in High Dimension

Based on the course *Probability in High Dimension* (048716) by Nir Weinberger, Technion, 2020.

Main external sources:

1. Non-asymptotic Statistics / Rakhlin & Rigollet (Lectures 1-8)
2. High Dimensional Probability / R. Vershynin, 2019, California (Chapters 3,4,7)
3. High Dimensional Statistics / Rigollet & Hutter (Pages 54-58)

Summarized by Ido Greenberg in 2020.

Contents

Summary 2

Introduction 5

SubGaussian random variables 5

SubExponential random variables 6

Random vectors in high dimensions 7

Maximum inequalities 7

Nets, covering numbers and packing numbers 8

Linear regression 9

Gaussian processes 11

Notation comments:

- We use log and ln notations alternately, both refer to the same natural logarithm.
- $\lesssim$ means "$\leq$, up to a constant".

Summary

Introduction

1. **Tail bounds**: bounds on $P(|X - E[X]| > t)$.
    a. For n i.i.d normal variables – the tail bound on the mean decays exponentially in n.
    b. This **can't be exploited through Central Limit Theorem**, which is only true up to $O\left(1/\sqrt{n}\right)$.
    c. ***Chebyshev inequality*** ($VarX = \sigma^2$): $P(|X - E[X]| \geq t\sigma) \leq 1/t^2$
    d. ***Chernoff bound*** ($\forall s > 0$): $P(X \geq t) \leq e^{-st}E[e^{sX}]$
2. A covariance matrix $\Sigma \in R^{d\times d}$ can be estimated using $n \to \infty$ samples if d is constant.
    a. In general, $n = O(d^2)$ samples are required.
    b. If $d \approx n$ $(d/n \to \gamma \wedge n \to \infty)$ then $\hat{\Sigma} \nrightarrow \Sigma$ – not entry-wise nor in spectrum.

Tail bounds on subGaussian & subExponential variables

3. ***SubGaussian variables*** ($\boldsymbol{subG}\left(\boldsymbol{\sigma^2}\right)$): lighter tail than a Gaussian (e.g. finite-support distribution).
    a. Equivalent definitions: $E[e^{sX}] \leq e^{\frac{s^2\sigma^2}{2}}$, $P(|X| \geq t) \leq 2e^{-c_2t^2}$, $||X||_p \leq c_3\sqrt{p}$.
    b. ***Orlicz $\boldsymbol{\psi_2}$-norm*** – $||X||_{\psi_2} = \inf\left\{t > 0 \middle| E\left[e^{|X|^2/t^2}\right] \leq 2\right\}$ – is finite iff $X \sim subG$.
4. ***Hoeffding inequality*** – for independent $\{X_i\}$ and for any $t > 0$:
    a. Bounded supports $[a_i, b_i]$: $P(|\bar{x} - E[\bar{x}]| \geq t) \leq e^{-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$
    b. $subG\left(\sigma_i^2\right)$: $P(|\bar{x} - E[\bar{x}]| \geq t) \leq e^{-\frac{n^2t^2}{2\sum_{i=1}^n \sigma_i^2}}$
5. ***SubExponential variables*** ($\boldsymbol{subE(\lambda)}$): lighter tail than an exponential variable (e.g. χ^2).
    a. Equivalent definitions: $E[e^{sX}] \leq e^{s^2\lambda^2}$, $P(|X| > t) \leq 2e^{-c_2t}$, $||X||_p \leq c_3p$.
    b. ***Orlicz $\boldsymbol{\psi_1}$-norm*** – $||X||_{\psi_1} = \inf\left\{t > 0 \middle| E\left[e^{|X|/t}\right] \leq 2\right\}$ – is finite iff $X \sim subE$.
6. ***Bernstein inequality***: $P(|\bar{X}| > t) \leq 2e^{-\frac{nt^2}{2(\bar{\sigma}^2 + Bt)}}$
    a. Assuming *Bernstein condition* ($E[|X|^k] \leq \frac{1}{2}Var(X)k!\,B^{k-2}$), which is equivalent to subE.

Random vectors in high dimensions

7. Concentration of norm: $\left\| \, ||X||_2 - \sqrt{d} \, \right\|_{\psi_2} \leq CK^2$ $\left(K^2 \coloneqq \max_i ||X_i||_{\psi_2}^2\right)$.
    a. For $d \gg 1$, **most of the density is concentrated around the sphere** rather than the origin!
8. ***Isotropic random vector***: $\Sigma(X) \coloneqq EXX^\top = I_d$ (e.g. Gaussian, symmetric Bernoulli, spherical).
    a. Equivalently, expected projection in any direction is 1: $\forall x \in R^d: E\langle X, x\rangle^2 = ||x||_2^2$.
9. ***SubGaussian vectors***: $||X||_{\Psi_2} \coloneqq \sup_{x \in S^{d-1}} ||\langle X, x\rangle^2||_{\Psi_2} < \infty$ (subG variable in any direction).
    a. SubG coordinates form a subG vector.
    b. SubG vectors are nearly-orthogonal: $|X^\top Y| \leq 1/\sqrt{d}$ w.h.p.

Maximum, covering & packing

10. For general (dependent) variables: $E\left[\max_i X_i\right] \leq N^{1/p} \max_i ||X||_p$ (using Jensen inequality).
11. For subG: $E\left[\max_{1\leq i\leq N} X_i\right] \leq \sigma\sqrt{2\log N}$ and $P\left(\max_i X_i > t\right) \leq Ne^{-\frac{t^2}{2\sigma^2}}$.

a. Linear combinations of subG: $E\left[\max_{\theta\in P}\theta^\top X\right] \le \sigma\sqrt{2\ln N}$ (if P is a convex polytope).

b. Projections of subG in different directions: $E\left[\max_{\theta\in B_2}\theta^\top X\right] \le 4\sigma\sqrt{d}$ (B_2 is the unit ball).

12. For subE: $E\left[\max_i X_i\right] \le \max_i ||X_i||_{\psi_1} \log N$.

13. **Maximum over infinite set** – $E\left[\max_{x\in K} f(x)\right]$ – can be approximated by a grid on K if f is Lipschitz.

a. ***Covering number*** ($\boldsymbol{N(K,\epsilon)}$): min size of ***$\epsilon$-net*** (that covers K with ϵ-balls).

b. ***Packing number*** ($\boldsymbol{P(K,\epsilon)}$): max size of ***$\epsilon$-separated set*** (that fits into K with ϵ-distances).

c. Covering vs. packing: $P(K,2\epsilon) \le N(K,\epsilon) \le P(K,\epsilon)$

d. Covering vs. volume: $\forall K \subset R^d$: $\frac{|K|}{|\epsilon B_2^d|} \le N(K,\epsilon) \le P(K,\epsilon) \le \frac{\left|K+\frac{\epsilon}{2}B_2^d\right|}{\left|\frac{\epsilon}{2}B_2^d\right|}$ (B_2^d=unit ball)

i. In particular, the covering number of the unit ball itself is exponential in d.

ii. For a random matrix A with independent $subG(\sigma^2)$ entries:

$$\sigma\left(\sqrt{n}+\sqrt{m}\right) \lesssim E\left[||A||\right] \lesssim \sigma\left(\sqrt{n}+\sqrt{m}\right)$$

where the upper bound considers $||A|| = \max_{x\in B_2^n}||Ax||$ and uses the bounds above.

Linear regression

14. ***Linear regression***: estimate $\theta \in R^d$ from $X \in R^{n\times d}, Y \in R^n$ where $Y = X\theta^* + \epsilon$ ($\epsilon_i \sim subG(\sigma^2)$).

a. ***Least squares estimator***: $\theta^{LS} := \underset{\theta\in R^d}{\operatorname{argmin}}|Y - X\theta|_2^2 = (X^\top X)^{-1}X^\top Y$.

i. ***Ridge regression***: $\hat\theta := \underset{\theta\in R^d}{\operatorname{argmin}}|Y - X\theta|_2^2 + \lambda||\theta||_2^2 = (X^\top X + \lambda I)^{-1}X^\top Y$

b. $\boldsymbol{E\left[MSE\left(X\theta^{LS}\right)\right] \lesssim \frac{\sigma^2\cdot rank(X^\top X)}{n}}$, and w.p. $1-\delta$, $\boldsymbol{MSE\left(X\theta^{LS}\right) \lesssim \frac{\sigma^2}{n}\log\frac{1}{\delta} + \frac{\sigma^2\cdot rank(X^\top X)}{n}}$.

i. Note: MSE = [in-sample square error of $Y_{clean} := X\theta^*$] = denoising error.

15. **Constrained LS** ($1-\delta$ bounds): $\theta^* \in \boldsymbol{B_1} \Rightarrow MSE \lesssim \sigma\sqrt{\frac{\log\frac{2d}{\delta}}{n}}$, $\theta^* \in \boldsymbol{B_0}(s) \Rightarrow MSE \lesssim \frac{\sigma^2 s}{n} log\frac{d}{s\delta}$.

a. B_0 is an explicit sparsity constraint, but is not convex (since L_0 is not a norm) and thus hard to optimize. However, if we replace knowing s by knowing σ, then we can obtain a similar result easily using the ***hard threshold estimator*** $\hat\theta_j := \begin{cases} Y_j & if\ |Y_j| > 2\tau \\ 0 & else \end{cases}$ if the dataset X is orthogonal. Without orthogonality, we would need ***Bayes Information Criterion*** (***BIC***), which is also computationally hard. Alternatively, B_1 (***Lasso***) can be seen as a convex (thus easier to solve) approximation, and also encourages sparsity. Lasso assumption also reduces the error bounds significantly if $d \gtrsim \sqrt{n}$. Under orthogonality, Lasso is solved by a ***soft threshold estimator***.

16. ***Misspecified linear regression***: $Y = f(X) + \epsilon$ with $f(X) \ne X\theta^*$ ($\theta \in K$).

a. ***Oracle inequality*** for $K = R^d, B_1$: $\boldsymbol{E\left[MSE\left(X\hat\theta\right)\right] \le MSE_{oracle} + err_{est}}$ where $MSE_{oracle} := \inf_{\theta\in K} MSE(X\theta)$ and err_{est} is the noise – the estimation error under the corresponding linear model (as specified above).

Gaussian processes

17. ***Gaussian process***: random process $\{X_t\}_{t\in T}$ where $\{X_t\}_{t\in T_0}$ is Gaussian for any finite subset T_0.

18. **Slepian inequality**: centered Gaussian processes with $EX_t^2 = EY_t^2$ and $E(X_t - X_s)^2 \leq E(Y_t - Y_s)^2$, satisfy $\forall \rho \in R$: $P\left[\sup_{t \in T} X_t \geq \rho\right] \leq P\left[\sup_{t \in T} Y_t \geq \rho\right]$ and in particular $\boldsymbol{E}\left[\sup_{\boldsymbol{t} \in \boldsymbol{T}} \boldsymbol{X_t}\right] \leq \boldsymbol{E}\left[\sup_{\boldsymbol{t} \in \boldsymbol{T}} \boldsymbol{Y_t}\right]$.

Introduction

1. Mean estimation of i.i.d variables:
   - a. ***Law of Large Numbers*** (LLN): $\bar{X}_n \to E[X_1]$ in prob. / a.s. (depending on finite moments).
   - b. ***Central Limit Theorem*** (CLT): $\sqrt{n}(\bar{X}_n - E[X_1\,]) \to N\big(0, Var(X_1)\big)$ in distribution.
   - c. Error estimation – *quadratic risk*: $E[(\bar{X}_n - E[X_1\,])^2] = Var(X_1)/n$.
2. **Tail bounds** (*concentration inequalities*): bounds on $P(|X - E[X]| > t)$.
   - a. ***Mills ratio inequality***: $X \sim N(\mu, \sigma^2)$ ➔ $\forall t > 0$: $\boldsymbol{P(|X - \mu| \geq t\sigma) \leq \sqrt{\frac{2}{\pi}}\frac{e^{-t^2/2}}{t}}$.
   - b. ***Berry-Esseen***: $\left|\boldsymbol{P\left(\sqrt{n}\frac{|\bar{X}_n - E[X_1]|}{\sqrt{Var(X_1)}} > t\right) - P(|Z| > t)}\right| \leq \boldsymbol{C/\sqrt{n}}$.
     - i. $Z \sim N(0,1)$ and $C = E|X_1 - EX_1|^3/\sigma^3$.
     - ii. The bound is tight (can be demonstrated using i.i.d Bernoulli variables).

   ➔ **The Gaussian's exponential tail-bound can't be exploited through CLT approximation**.
   - c. More advanced tail bounds (e.g. Hoeffding) are discussed below.
3. **Covariance matrix estimation** of n i.i.d variables in R^d:
   - a. Sample cov. matrix $\hat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$ converges to $\Sigma := E[X_1 X_1^\top]$ element-wise as $\frac{n}{d} \to \infty$.
     - i. In the general case we need $n = O(d^2)$ for estimation of Σ.
   - b. **High-dimensional asymptotics** – $d, n \to \infty,\ \ d/n \to \gamma \in (0,1]$; example for $\Sigma = I_d$:
     - i. $eigs(\hat{\Sigma})$ converge to *Marcenko-Pastur distribution* (a wide dist. maximized at γ).
     - ii. Specifically for top-eig: $\forall t > 0: P\left(\lambda_{max}(\hat{\Sigma}) > 1 + C\left(\sqrt{\boldsymbol{d/n}} + t + \sqrt{t}\right)\right) \leq e^{-nt}$.

SubGaussian random variables

4. ***Moment Generating Function*** (MGF) of X: $M(s) := E[e^{sX}]$ $\quad\big(\partial^k M/\partial s^k\,|_{s=0} = E[X^k]\big)$
5. ***Orlicz norm***: $\quad ||X||_\psi := \inf\left\{t > 0 \middle| E\left[\psi\left(\frac{|X|}{t}\right)\right] \leq 1\right\}$
   - a. ψ is convex, increasing, and $\psi(0) = 0,\ \lim\limits_{x\to\infty} \psi(x) = \infty$.
   - b. Examples:
     - i. $\psi = x^p \ \Rightarrow\ ||X||_\psi = ||X||_{L_p}$
     - ii. $\psi_1 = e^x - 1 \ \Rightarrow\ ||X||_{\psi_1} = \inf\{t > 0 | E[e^{|X|/t}] \leq 2\}$
     - iii. $\psi_2 = e^{x^2} - 1 \ \Rightarrow\ ||X||_{\psi_2} = \inf\{t > 0 | E[e^{|X|^2/t^2}] \leq 2\}$
   - c. $||\cdot||_{\psi_1}$ vs. $||\cdot||_{\psi_2}$: For any random variables X, Y:
     - i. $||X^2||_{\psi_1} = ||X||_{\psi_2}^2$
     - ii. $||XY||_{\psi_1} \leq ||X||_{\psi_2} ||Y||_{\psi_2}$
6. ***SubGaussian R.V.*** with variance-proxy σ^2 ($\boldsymbol{subG(\sigma^2)}$): $\boldsymbol{E[X] = 0} \ \wedge\ \boldsymbol{\forall s \in R: E[e^{sX}] \leq e^{\frac{s^2\sigma^2}{2}}}$
   - a. $X \sim subG(\sigma^2) \ \Leftrightarrow\ ||X||_{\psi_2} \leq c\sigma$ $\qquad$ (for $E[X] = 0$)
7. If $E[X] = 0 \wedge Var(X) = 1$, the following are **equivalent**:
   - a. $\forall s \in R$: $\qquad \boldsymbol{E[e^{sX}] \leq e^{c_1 s^2}}$
   - b. $\forall t \geq 0$: $\qquad \boldsymbol{P(|X| \geq t) \leq 2e^{-c_2 t^2}}$
   - c. $\forall p \in N$: $\qquad \boldsymbol{||X||_p = (E|X|^p)^{1/p} \leq c_3\sqrt{p}}$

d. $\forall s \in (0, c_4')$: $\boldsymbol{E\left[e^{sX^2}\right] \leq e^{c_4 s}}$

8. Example: X has a bounded support ➔ $X \sim subG$.
9. ***Hoeffding inequality*** – for independent $\{X_i\}$ and for any $t > 0$:

a. Bounded supports $[a_i, b_i]$: $P(|\bar{x} - E[\bar{x}]| \geq t) \leq e^{-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$

b. $subG(\sigma_i^2)$: $P(|\bar{x} - E[\bar{x}]| \geq t) \leq e^{-\frac{n^2t^2}{2\sum_{i=1}^n \sigma_i^2}}$

c. Reference – ***Chebyshev inequality*** ($VarX = \sigma^2$): $P(|X - E[X]| \geq t\sigma) \leq 1/t^2$

d. Reference – ***Chernoff bound***: $P(X \geq t) \leq \inf_{s>0} e^{-st} E[e^{sX}]$

10. Example – **median of means**:

a. Boosting – independent binary classifiers with $P(C_i\ is\ correct) \geq \frac{1}{2} + \delta$:

using Hoeffding, $P(voting\ is\ wrong) = P(\bar{X} < 1/2) \leq e^{-2n\delta^2}$.

b. $\mu^* :=$ median over k estimators $\{\hat{\mu}_i\}_{i=1}^k$, each based on n samples:

i. $n \geq \frac{4\sigma^2}{\epsilon^2} \Rightarrow P(|\hat{\mu}_i - \mu| \leq \epsilon) > \frac{3}{4}$ (Chebyshev)

ii. In addition $k \geq 8 \ln \frac{1}{\delta} \Rightarrow P(|\mu^* - \mu| \leq \epsilon) \geq 1 - \delta$ (Boosting)

c. Mean has best expected accuracy, but med.-of-means is stable in the sense of tail bound.

SubExponential random variables

11. ***SubExponential R.V.*** with param. $\lambda > 0$ ($\boldsymbol{subE(\lambda)}$): $\boldsymbol{E[X] = 0 \ \wedge \ \forall |s| \leq \frac{1}{\lambda}: E[e^{sX}] \leq e^{s^2\lambda^2}}$

a. $X \sim subE(1) \Rightarrow ||X||_{\psi_1} \leq c_1$ and $||X||_{\psi_1} \leq 1 \Rightarrow X \sim subE(c_2)$ (for $E[X] = 0$)

i. Note: $||X||_{\psi_1}^2$ cannot be bounded by $C \cdot Var(X)$ (see e.g. $X_n = \begin{cases} \pm 1 & w.p. 1/n \\ 0\ w.p. 1 - 2/n \end{cases}$).

b. Example: $\boldsymbol{\chi^2 \sim subE}$ (chi-square).

12. If $E[X] = 0$, the following are **equivalent**:

a. $\exists c_1 > 0, \forall |s| \leq \frac{1}{c_1}$: $\boldsymbol{E[e^{sX}] \leq e^{s^2\lambda^2}}$

b. $\exists c_2 > 0, \forall t > 0$: $\boldsymbol{P(|X| > t) \leq 2e^{-c_2 t}}$

c. $\exists c_3 > 0, \forall p \in N$: $\boldsymbol{||X||_p = (E|X|^p)^{1/p} \leq c_3 p}$

13. ***Bernstein inequality***: $\{X_i\}_{i=1}^n$ independent & $subE$ ➔ $\boldsymbol{P(|\bar{X}| > t) \leq 2e^{-Cn \cdot \min\left(t^2/\bar{\sigma}^2, t/\sigma_{max}\right)}}$

a. $\bar{\sigma}^2 := \frac{1}{n}\sum ||X_i||_{\psi_1}^2$, $\sigma_{max} := \max_i ||X_i||_{\psi_1}$.

b. By choosing $P = 1 - \delta$ and solving for t: $\boldsymbol{|\bar{X}| \leq C\left[\frac{\sigma_{max}}{n} \log\frac{2}{\delta} + \frac{\bar{\sigma}}{\sqrt{n}}\sqrt{\log\frac{2}{\delta}}\right]}$

14. ***Bernstein inequality*** – more standard form:

a. ***Bernstein condition*** ($\boldsymbol{BC(b)}$, $b > 0$): $\forall k \in N$: $E[|X|^k] \leq 0.5 \cdot Var(X) k!\, b^{k-2}$

b. $\{X_i\}_{i=1}^n$ independent, centered & follow $BC(b_i)$ ➔ $\boldsymbol{P(|\bar{X}| > t) \leq 2e^{-\frac{nt^2}{2(\bar{\sigma}^2 + b_{max} t)}}}$

i. $\bar{\sigma}^2 := mean\big(Var(X_i)\big)$, $b_{max} := \max_i b_i$.

c. By choosing $P = 1 - \delta$ and solving for t: $\boldsymbol{|\bar{X}| \leq C\left[\frac{b_{max}}{n} \log\frac{1}{\delta} + \frac{\bar{\sigma}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}\right]}$

d. Example: $\boldsymbol{|X| \leq B} \Rightarrow BC(B/3)$; better than Hoeffding if 1st term dominates ($\sigma < B/\sqrt{n}$)

15. Example – **classifier accuracy estimation**: in case where the empirical error is 0% over n samples, C.L.T yields $p_{err} \lesssim 1/\sqrt{n}$ but Bernstein can yield $p_{err} \lesssim 1/n$.

Random vectors in high dimensions

16. **Concentration of norm**:

$$\{X_i\}_{i=1}^{d} \in R^d \text{ independent \& subGaussian with } EX_i^2 = 1 \;\Rightarrow\; \left\| \left||\boldsymbol{X}|\right|_2 - \sqrt{\boldsymbol{d}} \right\|_{\boldsymbol{\psi_2}} \le \boldsymbol{CK^2}$$

   a. $K^2 \coloneqq \max_i \left||X_i|\right|_{\psi_2}^2$.
   b. More specifically, $\sqrt{d} - CK^2 \le E\left||X|\right|_2 \le \sqrt{d}$ and $Var\left(\left||X|\right|_2\right) \le CK^2$.
   c. I.e. in HD, most of the density is concentrated around the sphere rather than the origin.
   d. Intuition: the area S^{d-1} is $\propto r^{d-1}$ ➔ the volume around it gets dominant as d increases.

17. HD Gaussian is isotropic and concentrated around the norm ➔ spherical.
   a. Its *typical set* (points whose prob. density doesn't go to 0) is a sphere.

18. ***Isotropic random vector***: $\Sigma(X) \coloneqq EXX^{\top} = I_d \in R^{d\times d}$
   a. Examples: Gaussian; symmetric Bernoulli ($Unif\left(\{-,1,1\}^d\right)$); Spherical ($Unif\left(\sqrt{d}S^{d-1}\right)$).
   b. X isotropic $\Leftrightarrow$ $\forall x \in R^d : E\langle X, x\rangle^2 = \left||x|\right|_2^2$.
      i. I.e. the **expected projection of *X* in any direction is 1**.
   c. X isotropic ➔ $E\left||X|\right|_2^2 = d;$ X, Y isotropic & independent ➔ $E\langle X, Y\rangle^2 = d$.

19. ***SubGaussian random vector***: $\left||X|\right|_{\psi_2} \coloneqq \sup_{x\in S^{d-1}} \left||\langle X, x\rangle^2|\right|_{\psi_2} < \infty$
   a. $\left||X|\right|_{\psi_2}$ is the *subGaussian norm*.
   b. Intuitively, **1D subGaussian in any direction**.
   c. Examples: Gaussian; symmetric Bernoulli; Spherical.
   d. $\{X_i\}_{i=1}^{d}$ are independent centered 1D subGaussians ➔ they are a subGaussian vector with $\left||X|\right|_{\psi_2} \le C \cdot \max_i \left||X_i|\right|_{\psi_2}$.
      i. If $\{X_i\}_{i=1}^{d}$ are dependent they're still a subG vector but w/o the norm bound.
   e. Near-orthogonality: X, Y are d-dimensional independent isotropic subG vectors ➔ $|X^{\top}Y| \le \frac{1}{\sqrt{d}}$ w.h.p. (with high probability).

Maximum inequalities

20. Motivation: how far a random walk would get (max over time steps); how wrong could $\hat{\theta}$ be (max over different ground truths θ)…

21. $\{X_i\}_{i=1}^{N}$ (possibly dependent), $p \in N$: $E\left[\max_i X_i\right] \le N^{1/p} \max_i \left||X|\right|_p$
   a. Proved by Jensen inequality: $E\left[(\max|X_i|^p)^{1/p}\right] \le (E[\max|X_i|^p])^{1/p}$
   b. $X_i \sim subG(\sigma^2) \wedge p \coloneqq \log N$ ➔ $E\left[\max_i X_i\right] \le e \cdot c \cdot \sigma \cdot \sqrt{\log N}$
      i. Proved by property (c) of subG variables.

22. More accurate theorem: $X_i \sim subG(\sigma^2)$ ➔ $\boldsymbol{E\left[\max_i X_i\right] \le \sigma\sqrt{2\log N}}$
   a. Version with absolute value: $E\left[\max_i |X_i|\right] \le \sigma\sqrt{2\log 2N}$

b. For subE: $X_i \sim subE$ ➔ $\boldsymbol{E}\left[\max_i \boldsymbol{X_i}\right] \le \max_i \left|\left|\boldsymbol{X_i}\right|\right|_{\psi_1} \log \boldsymbol{N}$.

23. $X_i \sim subG(\sigma^2)$ ➔ $\boldsymbol{P}\left(\max_i \boldsymbol{X_i} > \boldsymbol{t}\right) \le \boldsymbol{N}e^{-\frac{t^2}{2\sigma^2}}$.

24. ***Convex polytope*** P: the convex hull of a finite set of points in R^d (all convex combinations).

    a. $V(P)$ is that finite set of points (the vertices of the polytope).

    b. Max over convex polytope is max over the vertices: $\forall x \in R^d$: $\max_{\boldsymbol{\theta} \in \boldsymbol{P}} \boldsymbol{\theta}^\top \boldsymbol{x} = \max_{\boldsymbol{\theta} \in \boldsymbol{V(P)}} \boldsymbol{\theta}^\top \boldsymbol{x}$

    c. If $[|V(P)| = N] \wedge \left[\forall v_i \in V(P): v_i^\top X \sim subG(\sigma^2)\right]$ then:

        i. $\boldsymbol{E}\left[\max_{\boldsymbol{\theta} \in \boldsymbol{P}} \boldsymbol{\theta}^\top \boldsymbol{X}\right] \le \boldsymbol{\sigma}\sqrt{\boldsymbol{2}\ln \boldsymbol{N}}$ (abs: $E\left[\max_{\theta \in P}|\theta^\top X|\right] \le \sigma\sqrt{2\ln 2N}$)

        ii. $\boldsymbol{P}\left(\max_{\boldsymbol{\theta} \in \boldsymbol{P}} \boldsymbol{\theta}^\top \boldsymbol{X} > \boldsymbol{t}\right) \le \boldsymbol{N}e^{-\frac{t^2}{2\sigma^2}}$ (abs: $P\left(\max_{\theta \in P}|\theta^\top X| > t\right) \le 2Ne^{-\frac{t^2}{2\sigma^2}}$)

    d. Example: $V \coloneqq \boldsymbol{B_1} = \boldsymbol{L_1}$ **ball** $= \{x \in R^d \mid \left|\left|x\right|\right|_1 \le 1\}$: C.P. with $N = \left|\{\pm e_i\}_{i=1}^d\right| = 2d$.

25. **Euclidean ball**: $\boldsymbol{B_2} = \boldsymbol{L_2}$ **ball** $= \{x \in R^d \mid \left|\left|x\right|\right|_2 \le 1\}$.

    a. Not a polytope, but contained within one: $B_2 \subset \sqrt{d}B_1$.

        ➔ $X_i \sim subG(\sigma^2) \Rightarrow \boldsymbol{E}\left[\max_{\boldsymbol{\theta} \in \boldsymbol{B_2}} \boldsymbol{\theta}^\top \boldsymbol{X}\right] \le \sqrt{d}E\left[\max_{\theta \in B_1}|\theta^\top X|\right] \le \sigma\sqrt{\boldsymbol{2d}\ln \boldsymbol{2d}}$

    b. It can be shown that in fact, $E\left[\max_{\theta \in B_2} \theta^\top X\right] \le 4\sigma\sqrt{d}$ (without the log).

        i. Also, w.p. $1-\delta$: $\max_{\theta \in B_2} \theta^\top X \le 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log 1/\delta}$.

Nets, covering numbers and packing numbers

26. Goal: find $\boldsymbol{E}\left[\max_{\boldsymbol{x} \in \boldsymbol{D}} \boldsymbol{f}(\boldsymbol{x})\right]$ (random function over a domain K / R.V.'s indexed by continuous K).

    a. Intuition: if f is Lipschitz, this can be approximated by a finite "ϵ-grid" (with accuracy$\sim\epsilon$); and expected max over a finite grid is solved above (with accuracy$\sim \log N$, $N \sim \epsilon^{-\dim D}$).

        i. Note: optimized bound requires to solve the tradeoff over ϵ.

27. ***Covering number*** ($\boldsymbol{N(K,d,\epsilon)}$): minimal size (cardinality) of an ϵ-net of K.

    a. ***ϵ-net***: $N \subseteq K$ that covers K up to ϵ-balls: $\forall x \in K, \exists x_0 \in N: d(x,x_0) \le \epsilon$.

    b. By permitting ball-centers outside of K, we may save only a little:

    $$N^{ext}(K,d,\epsilon) \le N(K,d,\epsilon) \le N^{ext}(K,d,\epsilon/2)$$

    c. If outside points are not permitted, we have anomalies (up to factor 2 in ϵ) such as $N\left([0,1]\setminus\{\frac{1}{2}\}, \epsilon = \frac{1}{2}\right) = 2 > 1 = \left|\left\{\frac{1}{2}\right\}\right| = N\left([0,1], \epsilon = \frac{1}{2}\right)$ even though $[0,1]\setminus\{\frac{1}{2}\} \subset [0,1]$.

28. **Packing number** ($\boldsymbol{P(K,d,\epsilon)}$): maximal size (cardinality) of an ϵ-separated subset.

    a. **ϵ-separated set**: $N \subseteq K$ with: $\forall x \neq y \in N: d(x,y) > \epsilon$.

    b. In a normed space (but not in every metric space), P is the largest number of disjoint closed $\epsilon/2$-balls whose centers are in K.

29. **A maximal ϵ-separated set is an ϵ-net.**

    a. "Maximal" – in the sense that any additional point would ruin the separation.

    b. Proof: otherwise there exists an additional point that is far from any point of N.

    c. Accordingly, an ϵ-net can be constructed by repeatedly choosing new "ϵ-far" points until no such points exist. This process is guaranteed to converge if K is compact.

30. Covering vs. packing: $\boldsymbol{P(K,d,2\epsilon) \le N(K,d,\epsilon) \le P(K,d,\epsilon)}$

    a. The right inequality is trivial since the separated set of size P is itself also a net.

31. Covering vs. volume: $\forall K \subset \boldsymbol{R^n}$: $\frac{|\boldsymbol{K}|}{|\boldsymbol{\epsilon B_2^n}|} \le \boldsymbol{N(K,\epsilon)} \le \boldsymbol{P(K,\epsilon)} \le \frac{|\boldsymbol{K+(\epsilon/2)B_2^n}|}{|\boldsymbol{(\epsilon/2)B_2^n}|}$
    a. $|\cdot|$ = volume; B_2^n = Euclidean unit ball; sum of sets = Minkowski sum (all sums of pairs).
    b. The left side is $|K|$ divided by the volume of an ϵ-ball. The right side is the volume of all the points "$\epsilon/2$-close" to K, divided by the volume of an $\epsilon/2$-ball.
    c. In particular for $K = B_2^n$ = Euclidean unit ball, the covering $\boldsymbol{N(B_2^n,\epsilon)}$ **is exponential in $\boldsymbol{n}$.**

$$\left(\frac{1}{\epsilon}\right)^n \le N(B_2^n,\epsilon) \le \left(\frac{2}{\epsilon}+1\right)^n \le \left(\frac{3}{\epsilon}\right)^n \quad \text{(last inequality is for } \epsilon \le 1\text{)}$$

32. Example – operator norm of a random matrix $A \in R^{n\times m}$:
    a. $||A|| \coloneqq \max_{x\in R^n, ||x||_2=1} ||Ax||_2 = \max_{x\in B_2^n} \max_{y \in B_2^m} y^\top Ax$ (vector norm equals its max projection).
    b. A tight bound for the expected norm of a random A with independent $subG(\sigma^2)$ entries:

$$\sigma\left(\sqrt{n}+\sqrt{m}\right) \lesssim E\left[||A||\right] \le 2\sigma\sqrt{2(n+m)\log 9} \lesssim \sigma\left(\sqrt{n}+\sqrt{m}\right)$$

        i. The "9" comes from the covering num. $N(B_2^n, 1/4\ )$ using the upper bound above.

Linear regression

33. *Regression*: $Y_i = f(X_i) + \epsilon_i \qquad \left(1 \le i \le n,\ \ X_i \in R^d,\ \ E[\epsilon_i] = 0\right)$
    a. Goal: estimate $\mu \coloneqq \left(f(x_1),\dots,f(x_n)\right)^\top \in R^n$ given data $\{(x_i,y_i)\}_{i=1}^n$ (i.e. from $\mu+\epsilon$).
    b. Metric to minimize: $MSE\left(\hat{f}_n\right) \coloneqq \frac{1}{n}\sum_{i=1}^n \left(\hat{f}_n(x_i) - f(x_i)\right)^2 = \frac{1}{n}|\hat{\mu}_n - \mu|_2^2$
34. ***Linear regression***: $\mu = X\theta^*$, i.e. $\boldsymbol{Y = X\theta^* + \epsilon} \quad \left(X \in R^{n\times d},\ \ \theta^* \in R^d,\ \ Y,\epsilon \in R^n\right)$
    a. $\boldsymbol{MSE} \coloneqq \frac{1}{n}\left|X\hat{\theta} - X\theta^*\right|_2^2 = \left(\hat{\theta}-\theta^*\right)^\top \frac{X^\top X}{n}\left(\hat{\theta}-\theta^*\right)$
        i. In-sample error in $Y_{clean} \coloneqq X\theta^*$ ➔ denoising error.
            1. Note: unlike $\left|\hat{\theta}-\theta^*\right|_2^2$, the defined MSE is not invariant to X's scale.
        ii. If X rows are orthogonal, then $MSE \propto \left|\hat{\theta}-\theta^*\right|_2^2$.
35. ***Least squares estimator*** – $\boldsymbol{\theta^{LS} \in \operatorname*{argmin}_{\theta\in R^d} |Y - X\theta|_2^2}$:
    a. Theorem: $\boldsymbol{\theta^{LS} = (X^\top X)^{-1}X^\top Y}$ (proven by zeroizing the derivative explicitly)
    b. If $X^\top X \in R^{d\times d}$ is not invertible (e.g. $d > n$), we can use ***Moore-Penrose pseudoinverse***, which chooses $\theta = \theta_1 + \theta_2$ whose projection on the kernel of $X^\top X$ is $\theta_2 = 0$.
    c. ***Ridge regression***: $\hat{\theta} \coloneqq \operatorname*{argmin}_{\theta\in R^d}|Y - X\theta|_2^2 + \boldsymbol{\lambda}||\boldsymbol{\theta}||_{\boldsymbol{2}}^{\boldsymbol{2}} = (X^\top X + \boldsymbol{\lambda I})^{-1}X^\top Y$
36. **Non-asymptotic LS analysis** – for $Y = X\theta^* + \epsilon_i$ with **independent noises** $\boldsymbol{\epsilon_i \sim subG\left(\sigma^2\right)}$:
    a. Unconstrained case ($\boldsymbol{\theta^* \in R^d}$): $\boldsymbol{E\left[MSE\left(X\theta^{LS}\right)\right] \lesssim \frac{\sigma^2}{n}r}$
        i. $r \coloneqq rank(X^\top X) \le d,n$
        ii. Also w.p. $1-\delta$ (by bounding over the unit ball): $\boldsymbol{MSE\left(X\theta^{LS}\right) \lesssim \frac{\sigma^2}{n}\left(\log\frac{1}{\delta} + r\right)}$
        iii. In this case, the proof can rely explicitly on the closed-form solution. Alternatively, we can use the definition of θ^{LS} to have $||Y - X\theta^{LS}||_2^2 \le ||Y - X\theta^*||_2^2 = ||\epsilon||_2^2$ (along with $Y - X\theta^* = X(\theta^{LS} - \theta^*) + \epsilon$).
    b. Constrained: $\boldsymbol{\theta^* \in B_1}$ and $\forall 1 \le j \le d$: $\left|\left|X_j\right|\right|_2 \le \sqrt{n}$: $\boldsymbol{E\left[MSE\left(X\hat{\theta}\right)\right] \lesssim \sigma\sqrt{\frac{\log 2d}{n}}}$

i. W.p. $1-\delta$: $MSE\left(X\hat{\theta}\right) \lesssim \sigma\sqrt{\frac{\log\frac{2d}{\delta}}{n}}$

ii. Note that by exploiting the L1-constraint (and assuming it's true…) we changed the error from $\propto \frac{d}{n}$ to $\propto \sqrt{\frac{\log d}{n}}$, which is better if $d \gtrsim \sqrt{n}$ (otherwise the general bounds are not better than the unconstrained case).

iii. The proof uses a max over B1's vertices for the bounding.

iv. This is ***Lasso Regression*** (the L1-constraint is equivalent to L1-penalty up to choice of parameters). While L0 derives pure sparsity, L1 is the closest one which is an actual norm (and thus derives a convex domain). It is also known to yield a sparse estimator (but also to decrease the entries of the estimator).

c. $\boldsymbol{\theta}^* \in \boldsymbol{B_0(s)}$ (L0-ball = up to s nonzero entries = sparse regression): $\boldsymbol{E \leq \frac{\sigma^2 s}{n} log\frac{ed}{s}}$

i. W.p. $1-\delta$: $MSE\left(X\hat{\theta}\right) \lesssim \frac{\sigma^2 s}{n} log\frac{d}{s\delta}$

ii. The proof assumes known s and uses a max over the $\binom{d}{2s} \leq \left(\frac{ed}{2s}\right)^{2s}$ possible choices of s coordinates of $\hat{\theta}$ and s of θ^* ($2s$ disjoint coordinates in worst case).

iii. Since B_0 is not convex it is **not straight-forward to find the minimizer**, and thus the actual error is [minimizer's error] + [computation/algorithm error].

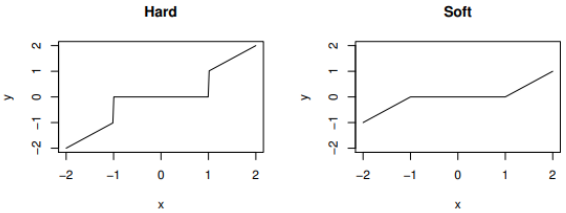
d. $\boldsymbol{\theta}^* \in \boldsymbol{B_0(s)}$ with unknown s, *orthogonal design* $X^\top X = nI_d$, and known σ:

i. W.p. $1-\delta$: $MSE \lesssim \frac{\sigma^2 s}{n} log\frac{2d}{\delta}$

ii. The proof relies on ***Gaussian Sequence Model*** (***GSM***): multiplying the linear regression model $Y = X\theta^* + \epsilon$ by $X^\top/n$, we get $\tilde{Y} \coloneqq \frac{X^\top Y}{n} = \theta^* + \frac{\epsilon}{n}$ (using orthogonality of X). This simplification is called *direct (observation) model* (in contrast to the inverse problem where we had to invert X). We also use $MSE = \left|\hat{\theta}-\theta^*\right|_2^2$ (by orthogonality).

iii. Unlike the case of known s, this problem does not have exponential complexity (d-choose-2s) – the estimator is simply a ***hard threshold***: $\hat{\theta}_j \coloneqq \begin{cases} Y_j & if\ \left|Y_j\right| > 2\tau \\ 0 & else \end{cases}$, for $\boldsymbol{\tau} \coloneqq \boldsymbol{\sigma}\sqrt{\frac{\boldsymbol{2\log(2d/\delta)}}{\boldsymbol{n}}}$.

iv. The hard threshold estimator is also the optimizer of the MSE with regularization term $4\tau^2\left|\left|\theta\right|\right|_0$. By replacing this with $4\tau^2\left|\left|\theta\right|\right|_1$ (Lasso), the optimizer becomes the ***soft threshold*** estimator$\hat{\theta}_j \coloneqq \begin{cases} Y_j - sign\left(Y_j\right)2\tau & if\ \left|Y_j\right| > 2\tau \\ 0 & else \end{cases}$.



Hard threshold vs. soft threshold

v. In the more general case – without the orthogonal design $X^\top X = nI_d$ – the regularized form $\left|\left|Y - X\theta\right|\right|_2^2 + 4\tau^2\left|\left|\theta\right|\right|_0$ can still be used, named ***Bayes Information Criterion*** (***BIC***). In this case, by setting $\tau^2 \propto \frac{\sigma^2 \log d}{n}$, we have (w.p. $1-\delta$): $MSE \lesssim \frac{||\theta^*||_0 \sigma^2 \log ed/\delta}{n}$.

37. ***Misspecified linear model***: $Y = f(X) + \epsilon,\ f(X) \approx X\theta^*$ (not equal this time). For $\theta \in R^d$ and $\theta \in B_1$ the ***oracle inequality*** shows that the approximation error (***oracle error***, since that would be the error of an all-knowing oracle with a linear model) is at most added to the noise: $\boldsymbol{E}\left[\boldsymbol{MSE}\left(\boldsymbol{X\hat{\theta}}\right)\right] \le \inf_{\boldsymbol{\theta \in K}} \boldsymbol{MSE}(\boldsymbol{X\theta}) + \boldsymbol{err}$, where $err = C\frac{\sigma^2 rank(X^\top X)}{n}$ for $K = R^d$ and $err = C\sigma\sqrt{\log(d)/n}$ for $K = B_1$ (same as in the corresponding linear models above).

Gaussian processes

38. ***Random process***: random variables $\{X_t\}_{t \in T}$ (for some set T).

a. ***Gaussian process***: $\{X_t\}_{t \in T_0}$ is a Gaussian vector for any finite subset $T_0 \subset T$.

i. Equivalently, any finite linear combination $\sum a_t X_t$ is a Gaussian variable.

b. We assume $EX_t = 0$ and denote the covariance Σ.

39. ***Increments***: $d(t,s) \coloneqq E[(X_t - X_s)^2]$ (e.g. for Brownian motion: $d(t,s) = \sqrt{|t-s|}$)

40. ***Canonical Gaussian process***: $X_t \coloneqq g_0^\top t$, where $t \in T \subset R^n,\ g_0 \sim N(0, I_n)$.

a. The increments are $d(t,s) = \left|\left|t-s\right|\right|_2^2$ (by direct calculation).

41. A *uniform control* on a process: a bound on $E\left[\sup_{t \in T} X_t\right]$.

a. E.g. for Brownian motion: $E\left[\sup_{t \le t_0} X_t\right] = \sqrt{\frac{2t_0}{\pi}}$ $(t_0 \ge 0)$

42. ***Slepian inequality***: let $\{X_t\}, \{Y_t\}$ be centered Gaussian processes with $\forall t, s: EX_t^2 = EY_t^2$ and $\boldsymbol{E}(\boldsymbol{X_t} - \boldsymbol{X_s})^2 \le \boldsymbol{E}(\boldsymbol{Y_t} - \boldsymbol{Y_s})^2$. **Then** $\forall \rho \in R: P\left[\sup_{t \in T} X_t \ge \rho\right] \le P\left[\sup_{t \in T} Y_t \ge \rho\right]$ (and thus also $\boldsymbol{E}\left[\sup_{\boldsymbol{t \in T}} \boldsymbol{X_t}\right] \le \boldsymbol{E}\left[\sup_{\boldsymbol{t \in T}} \boldsymbol{Y_t}\right]$).

a. The proof is based on *Gaussian interpolation*, where we define $Z(u) \coloneqq \sqrt{u}X + \sqrt{1-u}Y$. Note that if X, Y are independent, then $\Sigma\left(Z(u)\right) = u\Sigma(X) + (1-u)\Sigma(Y)$. By defining $f(X) \coloneqq \chi_{\left\{\max_t X_t < \rho\right\}}$ and showing that $f\left(Z(u)\right)$ is non-positive in u, we have $f(X) = f\left(Z(1)\right) \le f\left(Z(0)\right) = f(Y)$ as required.

b. This is a private case of a more general technique named ***coupling argument***: the inequality is independent of the dependence between X, Y (since only one of them appears in each side), so we may construct their joint distribution in any convenient way without losing generality (as long as the marginal distributions are conserved). In this case we assume that they are independent.

c. The monotony is proven using *Gaussian integration by parts* (AKA *Stein identity*): for $X \sim N(0, \Sigma)$ and differentiable $f: R^n \to R$, we have $\boldsymbol{\Sigma} \cdot \boldsymbol{E}[\boldsymbol{\nabla f}(\boldsymbol{X})] = \boldsymbol{E}[\boldsymbol{Xf}(\boldsymbol{X})]$.