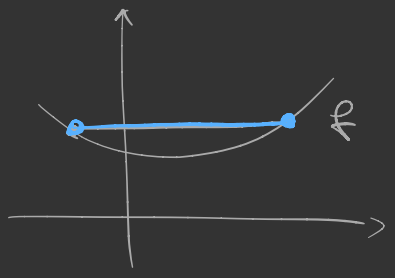
Convex function: connecting segment is always larger than the function itself, i.e.

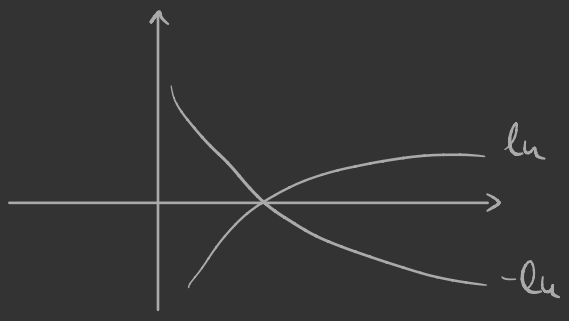


For $t \in [0,1]$:

$$f(tx + (1-t)y) \leq t \cdot f(x) + (1-t) f(y).$$

Examples: $t \mapsto t^2$, $t \mapsto \exp(t)$

$t \mapsto -\ln(t)$



For X a random variable. Then by Jensen's inequality we can obtain

$$-\ln(\mathbb{E}[X]) \leq \mathbb{E}[-\ln(X)].$$

3.2 Cramér's Theorem and Hoeffding's inequality

Connecting the law of large numbers and large deviation bounds.

Recall the moment generating function / Laplace transform:

$$M_X(\theta) = \mathbb{E}[\exp(\theta X)].$$

We define the cumulant generating function

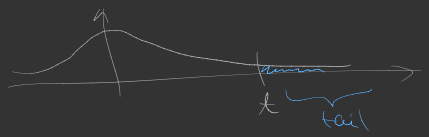
$$C_X(\theta) = \ln(\mathbb{E}[\exp(\theta X)]).$$

Theorem 3.6: Cramér's Theorem

Let $X_1, \dots, X_M$ be a sequence of independent random variables with cumulant generating functions C_{X_ℓ}, $\ell \in [M]$. Then for $t > 0$

$$\mathbb{P}\left(\sum_{\ell=1}^{M} X_\ell \geq t\right) \leq \exp\left(\inf_{\theta > 0} \left\{-\theta t + \sum_{\ell=1}^{M} C_{X_\ell}(\theta)\right\}\right).$$

probability to "produce a number in the tail"



Proof: Fix $\Theta > 0$. Then Markov inequality yields

$$\mathbb{P}\left(\sum_{k=1}^{M} X_k \geq t\right) = \mathbb{P}\left(\exp\left(\Theta \sum_{k=1}^{M} X_k\right) \geq \exp(\Theta t)\right)$$

$$\overset{\text{Markov}}{\leq} \frac{\mathbb{E}\left(\exp\left(\Theta \cdot \sum_{k=1}^{M} X_k\right)\right)}{\exp(\Theta t)}$$

$\exp(a+b) = \exp(a)\cdot\exp(b)$

$$= \frac{\mathbb{E}\left(\prod_{k=1}^{M} \exp(\Theta X_k)\right)}{\exp(\Theta t)}$$

independent

$$= \frac{\prod_{k=1}^{M} \mathbb{E}[\exp(\Theta X_k)]}{\exp(\Theta t)}$$

$C_{X_k}(\Theta) = \ln\left(\mathbb{E}[e^{\Theta X_k}]\right)$

$\Rightarrow \exp\left(C_{X_k}(\Theta)\right) = \mathbb{E}[e^{\Theta X_k}]$

$$= \frac{\prod_{k=1}^{M} \exp\left(C_{X_k}(\Theta)\right)}{\exp(\Theta t)}$$

$$= \exp\left(-\Theta t + \sum_{k=1}^{M} C_{X_k}(\Theta)\right).$$

Since this is true for every $\Theta > 0$, it is in particular for the infimum over $\Theta > 0$. ∎

Theorem 3.8: Hoeffding's inequality

Let $X_1, \dots, X_M$ be a sequence of indep. random variables, such that for $k \in [M]$: $\mathbb{E}X_k = 0$ and $|X_k| \leq B_k$ almost surely.

not really a restriction, because we can define $\tilde{X}_k = X_k - \mathbb{E}[X_k]$

means: $\mathbb{P}(|X_k| \leq B_k) = 1$, does not hold for Gaussian R.V.

Then:

$$\mathbb{P}\left(\sum_{k=1}^{M} X_k \geq t\right) \leq \exp\left(-\frac{t^2}{2\cdot\sum_{k=1}^{M} B_k^2}\right).$$

easy function with known quantities

Concept of "almost surely": emphasise that there might be realizations, which contradict the statement but these occasions are so isolated that they "don't count".

As a consequence (using $-X_k$):

$$\mathbb{P}\left(\sum_{k=1}^{M} |X_k| \geq t\right) \leq 2 \cdot \exp\left(-\frac{t^2}{2\sum_{k=1}^{M} B_k^2}\right) \xrightarrow{t \to \infty} 0$$

Remark: For $t > 0$: $\mathbb{P}(|X| \geq t) = \mathbb{P}(X \geq y) + \mathbb{P}(X \leq -y)$ (distinct events)

Remark: Using basic transformations, we get that (in particular using $t = \delta \cdot M$ and $\left|\frac{1}{M}\sum_{k=1}^{M}(X_k - \mathbb{E}[X_k])\right| \leq \frac{1}{M}\sum_{k=1}^{M}|X_k - \mathbb{E}[X_k]|$ a.s.)

$$\mathbb{P}\left(\left|\frac{1}{M}\sum_{k=1}^{M}(X_k - \mathbb{E}[X_k])\right| < \underbrace{\frac{\delta \cdot M}{M}}_{=\delta}\right) \geq 1 - 2\exp\left(-\frac{M^2\delta^2}{2\sum_{k=1}^{M} B_k^2}\right)$$

(≥ : complement)

If the X_k's are identically distributed, i.e. $X_k \sim X_1$, then $\mathbb{E}X_k = \mathbb{E}X$ and $B_k = B$. Then we can simplify

$$\mathbb{P}\left(\left|\left(\frac{1}{M}\sum_{k=1}^{M} X_k\right) - \mathbb{E}X\right| \leq \delta\right) \geq \underbrace{1 - 2\exp\left(-\frac{M\delta^2}{2B^2}\right)}_{\text{"overwhelming probability"}}.$$

$M \to \infty \longrightarrow 0$.

gives an first understanding how many samples M we need to estimate the mean value $\Leftarrow$

which gives us a **Law of Large Numbers** (LLN).

Intuition about the tail of R.V.'s:

Tails are telling us, how far away a realization is from its expected value.

If our RV is bounded almost surely by B, we have a density, which is compactly supported by $[-B, B]$.

$\rightsquigarrow \mathbb{E}[X] \in [-B, B]$

$\rightsquigarrow X - \mathbb{E}[X] \leq 2B$ a.s.

For RV's for which the tails decay very fast, we need less samples to get a good estimate for the mean value.

Vorlesung 16

16.06.2020

Proof for thm 3.8: $\mathbb{P}\left(\sum_{k=1}^{M} X_k \geq t \right) \leq \exp\left(- \frac{t^2}{2 \cdot \sum_{k=1}^{M} B_k^2} \right)$.

By assumption, $X_k \in [-B_k, B_k]$.

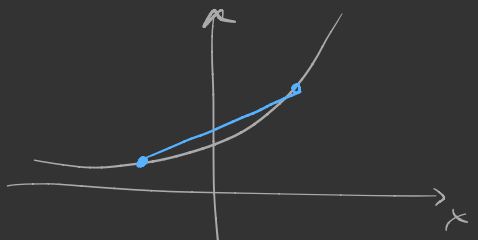
We can write

$$X_k = t(-B_k) + (1-t)B_k \quad \text{for } t \in [0,1].$$

with $t = \frac{B_k - X_k}{2B_k}$.

Consider

$$f(x) = \exp(\theta x), \quad \theta > 0.$$



Then f is convex, hence, by definition, we have for $t \in [0,1]$

$$f(x) = f(t \cdot x_1 + (1-t) x_2)$$
$$\leq t \cdot f(x_1) + (1-t) x_2.$$

Combining this we get that

$$f(X_k) = \exp(\theta X_k) = f(t(-B_k) + (1-t)B_k)$$
$$\leq t \cdot f(-B_k) + (1-t) f(B_k)$$
$$= \frac{B_k - X_k}{2B_k} \exp(-\theta B_k) + \frac{B_k + X_k}{2B_k} \exp(\theta B_k).$$

Using $\mathbb{E}[X_k] = 0$, we have

$$\mathbb{E}[\exp(\theta X_k)] \leq \mathbb{E}\left[\frac{B_k - X_k}{2B_k} \exp(-\theta B_k) \right]$$

$$+ \mathbb{E}\left[\frac{B_k + X_k}{2B_k} \exp(\theta B_k) \right]$$

with $\mathbb{E}\left[\frac{B_k \pm X_k}{2B_k} \right] \overset{\text{lin.}}{=} \frac{B_k \pm \mathbb{E}[X_k]}{2B_k} = \frac{B_k}{2B_k} = \frac{1}{2}$.

($\mathbb{E}[X_k] = 0$)

$$\Rightarrow \mathbb{E}[\exp(\theta X_\ell)] \leq \frac{1}{2} e^{-\theta B_\ell} + \frac{1}{2} e^{\theta B_\ell}.$$

Using the series def. of e^x, we have

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

$$\Rightarrow \mathbb{E}[\exp(\theta X_\ell)] \leq \frac{1}{2} \sum_{k=0}^{\infty} \frac{(-\theta B_\ell)^k}{k!} + \frac{1}{2} \sum_{k=0}^{\infty} \frac{(\theta B_\ell)^k}{k!}$$

cancel terms with odd exponent, $2\times$ for even exponent $=$

$$\sum_{h=0}^{\infty} \frac{(\theta B_\ell)^{2h}}{(2h)!}$$

$(2h)! \geq 2^h \cdot h!$ $\leq$

$$\sum_{h=0}^{\infty} \frac{(\theta B_\ell)^{2h}}{2^h \cdot h!} = \sum_{h=0}^{\infty} \left(\frac{(\theta B_\ell)^2}{2}\right)^h \cdot \frac{1}{h!}$$

series def. $=$

$$\exp\left(\frac{(\theta B_\ell)^2}{2}\right).$$

$$C_{X_\ell}(\theta) = \ln \mathbb{E}[\exp(\theta X_\ell)] \leq \ln\left(\exp\left(\frac{(\theta B_\ell)^2}{2}\right)\right) = \frac{(\theta B_\ell)^2}{2}.$$

(ln mon. function)

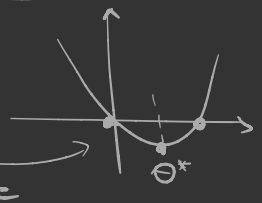
Cramer

$$\mathbb{P}\left(\sum_{\ell=1}^{M} X_\ell \geq t\right) \leq \exp\left(\inf_{\theta > 0} \left\{-\theta \cdot t + \sum_{\ell=1}^{M} C_{X_\ell}(\theta)\right\}\right)$$

exp increasing function $\leq$

$$\exp\left(\inf_{\theta > 0} \left\{-\theta \cdot t + \frac{\theta^2}{2} \cdot \sum_{\ell=1}^{M} B_\ell^2\right\}\right)$$

parabola in θ



want to minimize

We know by basic calculations, that

$$\theta^* = \frac{t}{\sum_{\ell=1}^{M} B_\ell^2}$$

$$\Rightarrow \mathbb{P}\left(\sum_{\ell=1}^{M} X_\ell \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{\ell=1}^{M} B_\ell^2}\right).$$

$$\mathbb{E}[X_k] = 0 = \mathbb{E}[-X_k] \quad \text{and} \quad |-X_k| = |X_k| \leq B_k.$$

$$\Rightarrow \quad \mathbb{P}\Big(\sum_{k=1}^{n} (-X_k) \geq t\Big) \leq \exp\Big(-\frac{t^2}{2\sum_{k=1}^{n} B_k^2}\Big).$$

$$\sum_{k=1}^{n} (-X_k) = -\sum_{k=1}^{n} X_k$$

$$\Rightarrow \quad \mathbb{P}\Big(\Big|\sum_{k=1}^{n} X_k\Big| \geq t\Big)$$

$$= \mathbb{P}\Big(\Big(-\sum_{k=1}^{n} X_k\Big) \geq t\Big) \wedge \Big(\sum_{k=1}^{n} X_k \geq t\Big)\Big)$$

subadditive

$$\leq 2 \exp\Big(-\frac{t^2}{2\sum_{k=1}^{n} B_k^2}\Big).$$

□

$$\mathbb{P}(|X| \geq t) \leq \mathbb{P}(X \geq t) + \mathbb{P}(-X \geq t).$$

3.3 Bernstein inequalities

Thm 3.9: $X_1, \dots, X_n$ indep. mean zero r.v.'s. Assume $\forall\, n \geq 2$

$$\mathbb{E}[|X_k|^n] \leq \frac{n!}{2} R^{n-2} \sigma_k^2$$

we can understand this as the variance: $n=2$: $\mathbb{E}[|X|^2] = \sigma_k^2$, because $\mathbb{E}[X]=0$.

for $R>0$ and $\sigma_k > 0$, $k \in [n]$. Then, $\forall\, t>0$, we get that

$$\mathbb{P}\Big(\Big|\sum_{k=1}^{n} X_k\Big| \geq t\Big) \leq 2 \cdot \exp\Big(-\frac{t^2}{2(\sigma^2 + Rt)}\Big)$$

with $\sigma^2 := \sum_{k=1}^{n} \sigma_k^2$.

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \quad \rightarrow \mathrm{Var}(X) \leq \mathbb{E}[X^2]$$

Corollary 3.10: $X_1, \dots, X_n$ indep. mean zero r.v.'s, such that $|X_k| \leq K$ almost surely (indep. of k). We further assume (additivity w.r.t. itself doing) $\exists\, \sigma_k^2 > 0$ s.t. $\mathbb{E}|X_k|^2 \leq \sigma_k^2$, $k \in [n]$. Then

$$\mathbb{P}\Big(\Big|\sum_{k=1}^{n} X_k\Big| \geq t\Big) \leq 2 \cdot \exp\Big(-\frac{t^2}{2(\sigma^2 + \frac{1}{3}Kt)}\Big) \xrightarrow{t \to \infty} 0$$

for $\sigma^2 = \sum_{k=1}^{n} \sigma_k^2$.

Proof: We check the assumption of Thm. 3.9.

$n=2$ ✓

$1 \cdot 2 \cdot 3 \cdot 4 \cdot \dots \cdot n \geq 1 \cdot 2 \cdot 3 \cdot 3 \cdot \dots \cdot 3$

$n \geq 3$: We have $n! \geq 2 \cdot 3^{n-2}$, and hence

$$\mathbb{E}|X_\ell|^n = \mathbb{E}\left[|X_\ell|^2 \cdot |X_\ell|^{n-2}\right]$$

bounded a.s. $\leq K^{n-2}\, \mathbb{E}\left[|X_\ell|^2\right] \leq K^{n-2}\, \theta_\ell^2$

$$= K^{n-2}\, \frac{n!}{n!}\, \theta_\ell^2$$

$$\leq K^{n-2}\, \frac{n!}{2 \cdot 3^{n-2}}\, \theta_\ell^2$$

$$= \frac{n!}{2}\, R^{n-2} \cdot \theta_\ell^2$$

for $R = \frac{1}{3}K$. ✓

Corollary 3.11: (For r.v. without boundedness assumption)

$X_1, \dots, X_M$ indep. r.v. with zero mean and are subexponential, i.e. $\forall t > 0$, we have

$$\mathbb{P}(|X_\ell| \geq t) \leq \beta e^{-\kappa t}$$

exponential decay in the tails

bounded RVs are fulfilling this as well.

for $\beta, \kappa > 0$.

Proof: Using the Cavalieri formula:

$$\mathbb{E}\left[|X_\ell|^n\right] = n \cdot \int_0^\infty \mathbb{P}(|X_\ell| \geq t)\, t^{n-1}\, dt$$

$$\leq n \cdot \beta \int_0^\infty \exp(-\kappa t)\, t^{n-1}\, dt$$

$s = \kappa t$, $ds = \kappa\, dt$

$$= n \cdot \beta \cdot \kappa^{-n} \underbrace{\int_0^\infty e^{-s}\, s^{n-1}\, ds}_{= (n-1)! =: \Gamma(n)}$$

Gamma function.

$$= \beta\, n(n-1)!\, \kappa^{-n} = \beta\, n!\, \kappa^{-n}$$

$$= \frac{n!}{2}\, \kappa^{-(n-2)} \underbrace{\left(2\beta\kappa^{-2}\right)}_{= \theta_\ell^2}$$

$, R = \kappa^{-1}$

Thm 3.9.

$$\Rightarrow \mathbb{P}\left(\left|\sum_{k=1}^{M} X_k\right| \geq t\right) \leq 2 \cdot \exp\left(-\frac{(Kt)^2}{2(2\beta M + Kt)}\right).$$

Proof: (Thm 3.9.) Based on Cramér's theorem.

$$\mathbb{E}[\exp(\theta X_k)] = 1 + \theta \underbrace{\mathbb{E}[X_k]}_{=0} + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}[X_k^n]}{n!} \quad \text{(series def.)}$$

$$= 1 + \frac{\theta^2 \sigma_k^2}{2} \sum_{n=2}^{\infty} \frac{2\theta^{n-2} \mathbb{E}[X_k^n]}{n! \cdot \sigma_k^2}$$

We define $F_k(\theta) = \sum_{n=2}^{\infty} \frac{2\theta^{n-2} \mathbb{E}[X_k^n]}{n! \cdot \sigma_k^2}$. Then we can write $\mathbb{E}[\exp(\theta X_k)] = 1 + \underbrace{\frac{\theta^2 \sigma_k^2}{2} \cdot F_k(\theta)}_{= \xi}$.

$$1 + \xi \leq e^{\xi} \leq \exp\left(\frac{\theta^2 \sigma_k^2}{2} \cdot F_k(\theta)\right).$$

Defining $F(\theta) = \max_{k \in [M]} F_k(\theta)$ and $\sigma^2 = \sum_{k=1}^{M} \sigma_k^2$ we obtain by Cramér's theorem $\ln(\exp(\frac{\theta^2 \sigma_k^2}{2} F_k(\theta)))$

$$\mathbb{P}\left(\sum_{k=1}^{M} X_k \geq t\right) \leq \inf_{\theta > 0} \exp\left(-\theta t + \frac{\theta^2 \sigma^2 F(\theta)}{2}\right)$$

$$\leq \inf_{0 < R\theta < 1} \exp\left(-\theta t + \frac{\theta^2 \sigma^2 F(\theta)}{2}\right) \quad (*)$$

Since $\mathbb{E}[X_k^n] \leq \mathbb{E}|X_k|^n$, we get

$$F_k(\theta) = \sum_{n=2}^{\infty} \frac{2\theta^{n-2} \mathbb{E}[X_k^n]}{n! \cdot \sigma_k^2} \leq \sum_{n=2}^{\infty} (R\theta)^{n-2}.$$

(ass.)

$$\leq \frac{2\theta^{n-2}}{n! \cdot \sigma_k^2} \cdot \frac{n!}{2} R^{n-2} \sigma_k^2$$

We know the geometric sum:

$$\sum_{n=0}^{\infty} a^n = \frac{1}{1-a} \quad \text{for} \quad 0 < a < 1.$$

Hence:

$$\sum_{u=2}^{\infty} (2\theta)^{u-2} = \frac{1}{1-2\theta} \quad \text{for} \quad 0 < 2\theta < 1.$$

Therefore

$$F(\theta) = \max_{l} F_l(\theta) \leq \frac{1}{1-2\theta} \quad \text{if} \quad 0 < 2\theta < 1$$

and hence

$$P\left(\sum_{l=1}^{u} X_l \geq t\right) \overset{(*)}{\leq} \inf_{0<2\theta<1} \exp\left(-t\theta + \frac{\theta^2\theta^2}{2(1-2\theta)}\right)$$

If we choose $\theta^* = \frac{t}{\theta^2 + 2t}$, (Ex: $2\theta^* < 1$)

then

$$P\left(\sum_{l=1}^{u} X_l \geq t\right) \leq \exp\left(-t\theta^* + \frac{(\theta^*)^2\theta^2}{2(1-2\theta^*)}\right)$$

$$= \exp\left(-\frac{t^2}{\theta^2+2t} + \frac{t^2}{(\theta^2+2t)^2} \cdot \frac{\theta^2}{2\left(1-2\cdot\frac{t}{\theta^2+2t}\right)}\right)$$

basic computations

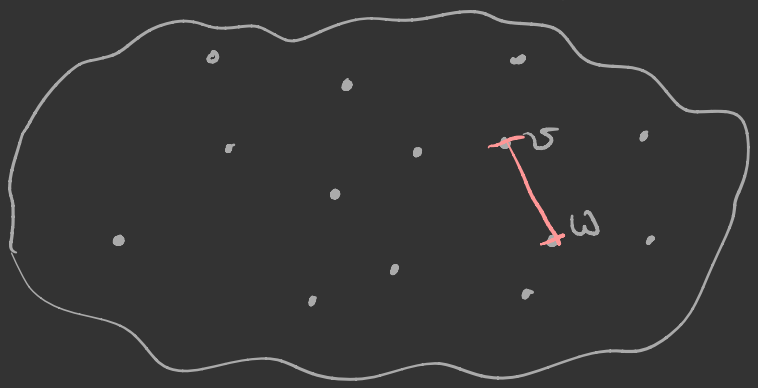
$$= \exp\left(-\frac{t^2}{2(\theta^2+2t)}\right).$$

$\square$

$$2\theta^* = \frac{t2}{\theta^2+2t} < 1 \quad \text{for } \theta^2 > 0$$

4. Johnson-Lindenstrauss embeddings

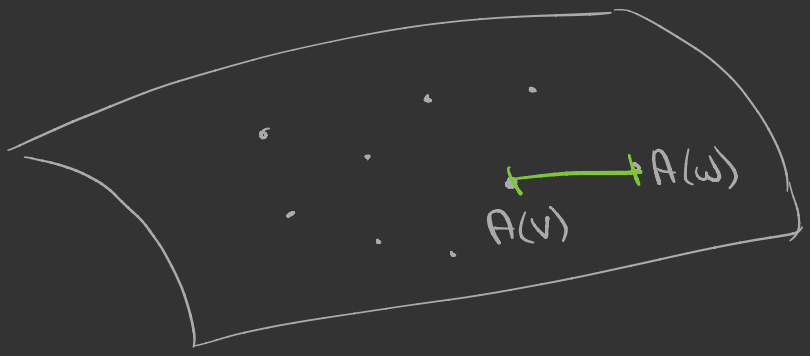
Introduction: Point-cloud in high dimension $\mathcal{C} \subseteq \mathbb{R}^d$.



Want to find a lower dimensional space $\mathbb{R}^k$, $k \ll d$, before processing the data.

$k \approx \ln(d)$

transformation map A



Which properties do we want for this embedding / map A?

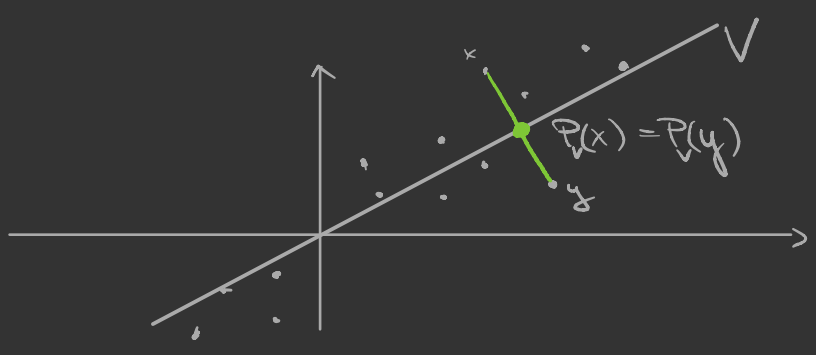
→ Want to preserve the geometry of the data, meaning keeping the distance between our data points, i.e.

$$\exists \text{ small } \varepsilon > 0 \cdot \forall v, w \in \mathcal{C} \quad (1-\varepsilon)\,\|v-w\|_2^2 \leq \|A(v)-A(w)\|_2^2 \leq (1+\varepsilon)\,\|v-w\|_2^2$$

(ε: distortion parameter, $\varepsilon < 1$)

Such an embedding is called Johnson-Lindenstrauss embedding and will exist. This is a strong example for probabilistic arguments to be beneficial for Data Analysis.

Note: SVD does not have the property to preserve distance:



$$\Rightarrow (1-\varepsilon) \underbrace{\|x-y\|_2^2}_{>0} \leq \|A(x)-A(y)\|_2^2 = 0$$

SVD has a completely different goal, the J-L-embedding does not extract common information.

Vorlesung 17

23.06.2020

Theorem 4.1: Johnson-Lindenstrauss Lemma

For any $0 < \varepsilon < 1$ and any integer n, let $k \in \mathbb{N}$ such that

$$k \geq 2\beta \left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3} \right)^{-1}$$

$\rightsquigarrow$ $\frac{1}{\varepsilon^3}$ scaling at $k \to \infty$ for $\varepsilon \to 0$

for some $\beta \geq 2$. Then for any set $\mathcal{P}$ of n points in $\mathbb{R}^d$, there is a map $f: \mathbb{R}^d \to \mathbb{R}^k$ such that for all $v, w \in \mathcal{P}$:

$$(1-\varepsilon) \|v-w\|_2^2 \leq \|f(v) - f(w)\|_2^2 \leq (1+\varepsilon) \|v-w\|_2^2 .$$

Furthermore, the map can be generated at random with probability $1 - (n^{2-\beta} - n^{1-\beta})$.

i.e. not built using knowledge of the data

probability of success for the generation

$\rightarrow$ the larger we choose β, the higher the success probability but the higher the lower bound for k

$\rightarrow$ the smaller we choose ε, the stronger the inequality, but the higher the lower bound for k

★ Does d really not play a significant role here?

Preliminary:

① If f would be linear map, then the inequality would be equivalent to

$$(1-\varepsilon) \leq \|Az\|^2 \leq (1+\varepsilon)$$

or

$$\left| \|Az\|^2 - 1 \right| \leq \varepsilon .$$

for $z = \frac{v-w}{\|v-w\|_2}$.

(2) Aim for this map is to estimate the length of a unit vector projected in $\mathbb{R}^d$ on a randomly picked k-dim. subspace.

From a probabilistic view, this is the same as picking the unit vector at random and projecting on a fixed subspace.

$\rightsquigarrow$ we pick an easy subspace $V = \text{span}\{e_1, \dots, e_k\}$.

(3) We generate uniformly distributed vectors y on the sphere, using $X_i \sim N(0,1)$ and $y = (X_1, \dots, X_d) / \|X\|_2 \in \mathbb{S}^{d-1}$

(4) Projection of y onto V is given by $z = (y_1, \dots, y_k) \in \mathbb{R}^k$ i.e. simply taking the first k coordinates.

$$\Rightarrow \quad \mathbb{E}\left[\|z\|_2^2\right] = \underset{\uparrow}{k/d} =: \mu$$

Exercise 10.1

$$L = \|z\|_2^2$$

(5) Since we want the length to be close to 1, we can simply rescale z accordingly (i.e. later rescale our linear map by this factor).

Summarizing these ideas:

Define $A = \sqrt{\frac{d}{k}}\, P_{\text{span}\{e_1, \dots, e_k\}}$,

$$X \sim N(0,1)$$
$$y = \frac{X}{\|X\|_2} \in \mathbb{S}^{d-1}$$

and

$$P_{\text{span}\{e_1, \dots, e_k\}}\, y = z = (y_1, \dots, y_k).$$

Exercise: $\mathbb{E}\left[\|z\|_2^2\right] = \frac{k}{d}$

and hence

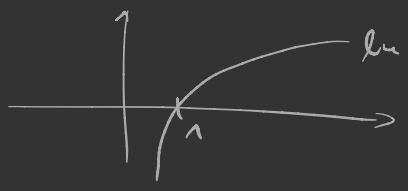
$$\mathbb{E}\left[\|Ay\|_2^2\right] = \mathbb{E}\left[\left\|\sqrt{\tfrac{d}{k}}\, z\right\|_2^2\right] = 1.$$

Problem: We don't want this for the expected value, but for each specific pair of points $v, w \in \mathcal{P}$, so $\forall v, w \in \mathcal{P}$:

$$(1-\varepsilon)\|v-w\|_2^2 \leq \|Av - Aw\|_2^2 \leq (1+\varepsilon)\|v-w\|_2^2.$$

$$\Leftrightarrow \quad (1-\varepsilon) \leq \left\| A\left(\frac{v-w}{\|v-w\|_2}\right)\right\|_2^2 \leq (1+\varepsilon)$$

$$\Leftrightarrow \quad \left| \underbrace{\left\| A\left(\frac{v-w}{\|v-w\|_2}\right)\right\|}_{=:y} - 1 \right| \leq \varepsilon$$



Lemma 4.2: Let $k < d$. Then

(a) If $\alpha < 1$, then

$$P\left(\|z\|_2^2 \leq \alpha \cdot \frac{k}{d}\right) \leq \exp\left(\frac{k}{2}\underbrace{(1-\alpha+\ln(\alpha))}_{\alpha \to 0 \to -\infty}\right)$$

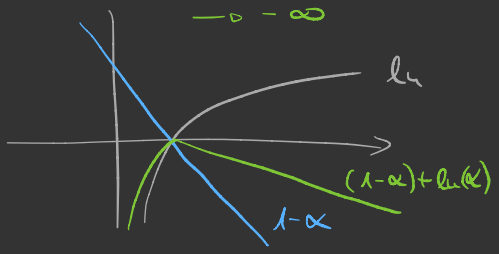
"$\|z\|_2^2$ can't be much smaller than $\frac{k}{d}$."

$\alpha \to 0 \to 0$

(b) If $\alpha > 1$, then

$$P\left(\|z\|_2^2 \geq \alpha \cdot \frac{k}{d}\right) \leq \exp\left(\frac{k}{2}\underbrace{(1-\alpha+\ln(\alpha))}_{\alpha \to \infty \to -\infty}\right)$$

"$\|z\|_2^2$ can't be much larger than $\frac{k}{d}$."



$\alpha \to \infty \to 0$

Proof: for Thm 4.1. using 4.2.

We prove existence, don't give a construction of such a map.

Trivial for $k \geq d$. Assume $k < d$.

Define $P_V, L = \|P_V(v_i) - P_V(v_j)\|_2^2$ for $v_i, v_j \in \mathcal{P}$ and $\mu = \frac{k}{d}\|v_i - v_j\|_2^2$.

By applying Lem. 4.2. a) we get that

$$\mathbb{P}\left(L \le \underbrace{(1-\varepsilon)\cdot\mu}_{\alpha}\right) \le \exp\left(\frac{\kappa}{2}\left(1-(1-\varepsilon)+\ln(1-\varepsilon)\right)\right)$$

Taylor exp. of $\ln(\cdot)$, $0 \le \varepsilon < 1$:
$\ln(1-\varepsilon) \le -\varepsilon - \frac{\varepsilon^2}{2}$

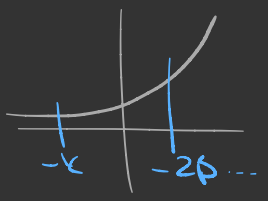
$$\le \exp\left(\frac{\kappa}{2}\left(\varepsilon-\left(\varepsilon+\frac{\varepsilon^2}{2}\right)\right)\right)$$

$$= \exp\left(-\frac{\kappa\varepsilon^2}{4}\right).$$

Taylor: $\ln(1-\varepsilon) = \ln(1) + \frac{d}{dx}\ln(x)\Big|_{x=1}\cdot(1-\varepsilon-1) + \frac{1}{2}\frac{d}{dx^2}\ln(x)\Big|_{x=2}(1-\varepsilon-1)^2 + \Theta(\varepsilon^3)$
$= 0 + \frac{1}{x}\Big|_{x=1}(-\varepsilon) + \frac{1}{2}\cdot(-1)\frac{1}{x^2}\Big|_{x=1}\cdot(-\varepsilon)^2 + \Theta(\varepsilon^3)$
$= -\varepsilon - \frac{\varepsilon^2}{2} + \Theta(\varepsilon^3)$.

We have that $\kappa \ge 2\beta\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}\right)^{-1}\ln(n)$ by assumption

and hence $-\kappa \le -2\beta\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}\right)^{-1}\ln(n)$



$$\Rightarrow \mathbb{P}\left(L \le (1-\varepsilon)\cdot\mu\right) \le \exp\left(\frac{-2\beta\ln(n)\varepsilon^2}{4\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}\right)}\right)$$

$$= \exp\left(-\frac{\beta\ln(n)\varepsilon^2}{2\varepsilon^2\left(\frac{1}{2}-\frac{\varepsilon}{3}\right)}\right)$$

$$= \exp\left(\underbrace{-\frac{3}{3-2\varepsilon}\cdot\beta\ln(n)}_{\le -\beta\ln(n)}\right)$$

($\frac{3}{3-2\varepsilon} \ge 1$)

$$\le \exp\left(-\beta\ln(n)\right)$$

$$= n^{-\beta}.$$

Similarly, we can prove that

$$\mathbb{P}\left(L \ge (1+\varepsilon)\mu\right) \overset{4.2.(6)}{\le} \exp\left(\frac{\kappa}{2}\left(1-(1+\varepsilon)+\ln(1+\varepsilon)\right)\right)$$

$\ln(1+\varepsilon) \le \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3}$ for $\varepsilon \ge 0$

$$\le \exp\left(\frac{\kappa}{2}\left(-\varepsilon+\left(\varepsilon-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)\right)\right)$$

$$= \exp\left(-\frac{\kappa\left(\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}\right)}{2}\right)$$

$-\kappa \le -2\beta\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}\right)^{-1}\ln(n)$

$$\le \exp\left(-\beta\ln(n)\right) = n^{-\beta}.$$

Summarizing these results, we get that for fixed $v_i, v_j \in \mathcal{B}$:

$$\mathbb{P}\left(\frac{k}{d}\|v_i - v_j\|_2^2 (1-\varepsilon) \le \|P_V(v_i) - P_V(w_i)\|_2^2 \le \frac{k}{d}\|v_i - v_j\|_2^2 (1+\varepsilon)\right)$$

$$\geq 1 - 2n^{-\beta}$$

i.e. we are able to concentrate their distances quite well.

Problem: We want this to be true for all pairs at the same time / simultaneously. $\Rightarrow$ need a union bound.

Recall:

$f(v) = Av = \sqrt{\frac{d}{k}} P_V(v)$ where V is a random k-dim. subspace. So far we know that

$$\left\| \frac{f(v_i) - f(v_j)}{\|v_i - v_j\|_2^2} \right\|_2^2 \in [1-\varepsilon, 1+\varepsilon]$$

with probability $\geq 1 - 2n^{-\beta}$, or

$$\left\| \frac{f(v_i) - f(v_j)}{\|v_i - v_j\|_2^2} \right\|_2^2 \notin [1-\varepsilon, 1+\varepsilon]$$

with probability $\leq 2n^{-\beta}$.

Now, we know that we have $\binom{n}{2}$ pairs to choose in $\mathcal{B}$. Hence, if we want the statement to hold for all pairs simultaneously, i.e.

$$\left\| \frac{f(v_i) - f(v_j)}{\|v_i - v_j\|_2^2} \right\|_2^2 \notin [1-\varepsilon, 1+\varepsilon] \quad \forall v_i, v_j \in \mathcal{B}$$

with probability $\leq 2n^{-\beta}\binom{n}{2} = n^{2-\beta} - n^{1-\beta}$. ∎

$\mathbb{P}(\forall ij \ \ldots) = \mathbb{P}\left(\bigcap_{ij} \ldots\right) \leq \sum_{ij} \mathbb{P}(\ldots)$ union bound

$$\Rightarrow (1-\varepsilon)\|v-w\|_2^2 \le \|Av - Aw\|_2^2 \le (1-\varepsilon)\|v-w\|_2^2 \quad \forall v, w \in \mathcal{B}$$

with probability $\geq 1 - (n^{2-\beta} - n^{1-\beta})$.

$\beta > 2$ ⟶ $\geq 1 - \frac{1}{n^{\beta-2}} \xrightarrow{n \to \infty} 1$

Proof: for Lemma 42.(a).

"(a) If $\alpha < 1$, then

$$\mathbb{P}\left(\|Z\|_2^2 \leq \alpha \cdot \frac{k}{d} \right) \leq \exp\left(\frac{k}{2} (1 - \alpha + \ln(\alpha)) \right)^{"}$$

We use the following fact: $X \sim N(0,1)$, then

$$\mathbb{E}\left[e^{sX^2} \right] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{sx^2} e^{-x^2/2} \, dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-x^2/\frac{1}{2} - s} \, dx$$

$$= \frac{1}{\sqrt{1-2s}} \qquad \forall \ -\infty < s < \frac{1}{2}. \qquad (*)$$

Assume $0 < \alpha < 1$.

① Claim: $\mathbb{P}\left(X_1^2 + \dots + X_k^2 \leq \frac{k}{d} \alpha \left(X_1^2 + \dots + X_d^2 \right) \right)$

$$\leq \alpha^{k/2} \left(1 + \frac{(1-\alpha)k}{d-k} \right)^{\frac{d-k}{2}}$$

$$\mathbb{P}\left(d(X_1^2 + \dots + X_k^2) \leq k\alpha (X_1^2 + \dots + X_d^2) \right)$$

$$= \mathbb{P}\left(0 \leq k\alpha(X_1^2 + \dots + X_d^2) - d(X_1^2 + \dots + X_k^2) \right)$$

$\exp(t \cdot)$, $t > 0$

$$= \mathbb{P}\left(1 \leq \exp\left(t \left(k\alpha(X_1^2 + \dots + X_d^2) - d(X_1^2 + \dots + X_k^2) \right) \right) \right)$$

Markov inequality

$$\leq \frac{1}{1} \mathbb{E}\left[\exp\left(t \left(k\alpha(X_1^2 + \dots + X_d^2) - d(X_1^2 + \dots + X_k^2) \right) \right) \right]$$

$$= \mathbb{E}\left[\exp\left(t(k\alpha-d)(X_1^2+\dots+X_k^2) + tk\alpha(X_{k+1}^2+\dots+X_d^2)\right)\right]$$

independence ↓

$$= \mathbb{E}\left[\exp\left(t(k\alpha-d)(X_1^2+\dots+X_k^2)\right)\right]\mathbb{E}\left[\exp\left(tk\alpha(X_{k+1}^2+\dots+X_d^2)\right)\right]$$

$X \sim$ identically ↓

$$= \mathbb{E}\left[\exp\left(t(k\alpha-d)X^2\right)\right]^k \mathbb{E}\left[tk\alpha X^2\right]^{d-k}$$

(*)

$$= (1-2t(k\alpha-d))^{-\frac{k}{2}} (1-2tk\alpha)^{-\frac{d-k}{2}} =: g(t).$$

(*) $\rightarrow$ $tk\alpha < \frac{1}{2}$ and $t(k\alpha-d) < \frac{1}{2}$

$\Rightarrow$ $0 < t < \frac{1}{2k\alpha}$

We want to minimize $g(t)$ on $(0, \frac{1}{2k\alpha})$

$\Leftrightarrow$ $\max_{t \in (0, \frac{1}{2k\alpha})} f(t) = \frac{1}{g(t)^2} = (1-2tk\alpha)^{d-k}(1-2t(k\alpha-d))^k$

For this, we try to solve $f'(t_0) = 0$. With this we find

$$t_0 = \frac{1-\alpha}{2\alpha(d-k\alpha)} \in (0, \frac{1}{2k\alpha})$$

$$g(t_0) = \sqrt{\frac{1}{f(t_0)}} \quad \text{and} \quad f(t_0) = \left(\frac{d-k}{d-k\alpha}\right)^{d-k}\left(\frac{1}{\alpha}\right)^k.$$

$$= \alpha^{k/2}\left(1+\frac{(1-\alpha)k}{d-k}\right)^{\frac{d-k}{2}}$$

(1)

(2) Exercise:

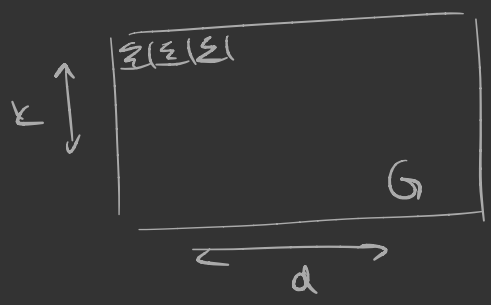
$$\alpha^{k/2}\left(1+\frac{(1-\alpha)k}{d-k}\right)^{\frac{d-k}{2}} \leq \exp\left(\frac{k}{2}(1-\alpha+\ln(\alpha))\right)$$

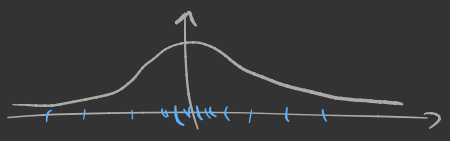
(a)

(b) can be proven similarly.

Remark: We have proven the existence now, but still don't know how to derive / construct such a map.

Improvement 1: Generate a matrix $G \in \mathbb{R}^{k \times d}$ elementwise $G_{ij} \sim N(0,1)$





$\Rightarrow$ $A = \frac{1}{\sqrt{k}} G$ and define $f(v) = Av$ the linear map.

Ito: $f(v)$ is a Johnson-Lindenstrauss embedding:

Theorem 4.3: Let $0 < \varepsilon < \frac{1}{2}$, any $n \in \mathbb{N}$, $k \in \mathbb{N}$ st.

$$k \geq \beta \varepsilon^{-2} \ln(n).$$

Then, for any $\mathcal{P} \subseteq \mathbb{R}^d$ with $|\mathcal{P}| = n$, the map $f(v)$ with

$$f(v) = \underbrace{\left(\frac{1}{\sqrt{k}} G\right)}_{= A} v \quad \text{for} \quad G = (G_{ij})_{ij},\ G_{ij} \sim N(0,1)$$

fulfills $\forall v, w \in \mathcal{P}$:

$$(1-\varepsilon)\|v-w\|_2^2 \leq \|f(v)-f(w)\|_2^2 \leq (1+\varepsilon)\|v-w\|_2^2$$

with probability $\geq \underbrace{1 - \left(n^{2-\beta(1-\varepsilon)} - n^{1-\beta(1-\varepsilon)}\right)}_{\text{for } \beta \gg 2 \xrightarrow{n \to \infty} 1}$.

Remark: Compared to 4.1 we now have a recipe to generate our mapping.

Proof: next time.

Vorlesung 18

24.06.2020

For the proof of Thm. 4.3, we first state the following Lemma:

Lemma 4.4: Let $x \in \mathbb{R}^d$ be fixed. Assume that $G \in \mathbb{R}^{k \times d}$ has iid entries $G_{ij} \sim N(0,1)$. Then for $A = \frac{1}{\sqrt{k}} G$,

(but J-L-lemma is about differences of pairs)

(union bound → result we need)

$$\mathbb{P}\left(\left| \|Ax\|_2^2 - \|x\|_2^2 \right| > \varepsilon \|x\|_2^2 \right) \leq 2 \exp\left(- \frac{(\varepsilon^2 - \varepsilon^3) k}{4} \right)$$

or equivalently

$$\mathbb{P}\left((1-\varepsilon)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\varepsilon)\|x\|_2^2 \right) \geq 1 - 2\exp\left(- \frac{(\varepsilon^2 - \varepsilon^3) k}{4} \right).$$

Proof: (Thm 4.3)

Note again that there are $\binom{n}{2}$ pairs of points $v, w \in \mathcal{P}$. Then

$$\mathbb{P}\left(\text{for some } v,w \in \mathcal{P}: \left| \|f(v) - f(w)\|_2^2 - \|v-w\|_2^2 \right| > \varepsilon \|v-w\|_2^2 \right)$$

(at least one pair but can we write in "intersections")

$$\leq \mathbb{P}\left(\bigcup_{v,w} \dots \right)$$

(can have intersections)

$$\leq \sum_{v,w \in \mathcal{P}} \mathbb{P}\left(\left| \|A(\underbrace{v-w}_{=:x})\|_2^2 - \|v-w\|_2^2 \right| > \varepsilon \|v-w\|_2^2 \right)$$

$$\leq \sum_{v,w \in \mathcal{P}} 2 \exp\left(- \frac{(\varepsilon^2 - \varepsilon^3) k}{4} \right)$$

($k \geq \beta \varepsilon^{-2} \ln(n)$)

$$\leq 2\binom{n}{2} \cdot \exp\left(- \frac{(\varepsilon^2 - \varepsilon^3) \beta \varepsilon^{-2} \ln(n)}{4} \right)$$

$$= 2\binom{n}{2} n^{-\beta(1-\varepsilon)}$$

$$= n^{2-\beta(1-\varepsilon)} - n^{1-\beta(1-\varepsilon)}.$$

$\Rightarrow$ Complementary probability: $\forall\, v,w \in \mathcal{P}$

$$(1-\varepsilon)\|v-w\|_2^2 \leq \|Av - Aw\|_2^2 \leq (1+\varepsilon)\|v-w\|_2^2$$

with probability $\left(1 - n^{2-\beta(1-\varepsilon)} - n^{1-\beta(1-\varepsilon)}\right)$.

Proof: (Lemma 4.4.)

$G_{ij} \overset{iid}{\sim} N(0,1)$, $A = \frac{1}{\sqrt{k}} G$.

1. Step: $\mathbb{E}[\|Ax\|_2^2] = \|x\|_2^2$ for any $x \in \mathbb{R}^d$ fixed.

Let's first check about $\mathbb{E}[(Gx)_j^2]$ j-th coordinate of Ax.

We recall that $\mathbb{E}[G_{ij}] = 0$, moreover $\mathbb{E}[G_{ij}^2] = 1$.

$$\mathbb{E}[(Gx)_j^2] = \mathbb{E}\left[\left(\sum_{i=1}^d G_{ij} x_i\right)^2\right]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^d G_{ij} x_i\right)\left(\sum_{i'=1}^d G_{i'j} x_{i'}\right)\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^d \sum_{i'=1}^d G_{i'j} G_{ij} \underbrace{x_{i'} x_i}_{\text{deterministic}}\right]$$

linear

$$= \sum_{i=1}^d \sum_{i'=1}^d \mathbb{E}[G_{i'j} G_{ij}] x_{i'} x_i$$

indep.

$$= \sum_{i=1}^d \sum_{i'=1}^d \underbrace{\mathbb{E}[G_{i'j}]\mathbb{E}[G_{ij}]}_{= \delta_{i'i}} x_{i'} x_i$$

$$= \sum_{i=1}^d x_i^2 = \|x\|_2^2.$$

$$\Rightarrow \mathbb{E}\left[\|Ax\|_2^2\right] = \frac{1}{k} \sum_{j=1}^k \mathbb{E}\left[(Gx)_j^2\right] \overset{\text{1. Step}}{=} \frac{1}{k} \sum_{j=1}^k \|x\|_2^2 = \|x\|_2^2.$$

$$\Rightarrow Z_j = \frac{(Gx)_j}{\|x\|_2^2} = \frac{\sum_{i=1}^d G_{ij} x_i}{\|x\|_2^2} \sim N(0,1).$$

1. Step

↑ Sum of iid Gaussians is again Gaussian

Indeed:

$$\mathbb{E}[Z_j] = \frac{\sum_{i=1}^d \mathbb{E}[G_{ij}] x_i}{\|x\|_2^2} = 0 \quad (\mathbb{E}[G_{ij}] = 0)$$

$$\mathbb{E}[Z_j^2] = [\text{1. Step}] = 1.$$

2. Step: $\mathbb{P}\left(\frac{1}{k} \sum_{j=1}^{k} z_j^2 > (1+\varepsilon) \right) \leq ?$

$$\Leftrightarrow \frac{\|Ax\|_2^2}{\|x\|_2^2} > (1+\varepsilon)$$

$\theta > 0$

$$= \mathbb{P}\left(\exp\left(\theta \sum_{j=1}^{k} z_j^2 \right) > \exp\left(\theta \cdot k(1+\varepsilon) \right) \right)$$

Markov

$$\leq \frac{\mathbb{E}\left[\exp\left(\theta \sum_{j=1}^{k} z_j^2 \right) \right]}{\exp(\theta \cdot k(1+\varepsilon))}$$

independence

$$= e^{-\theta k(1+\varepsilon)} \left(\mathbb{E}\left[\exp(\theta z^2) \right] \right)^k$$

$\mathbb{E}\left[e^{sX^2} \right] = \frac{1}{\sqrt{1-2s}}$ for $-\infty < s < \frac{1}{2}$

$$= e^{-\theta k(1+\varepsilon)} \cdot \left(\frac{1}{1-2\theta} \right)^{k/2} \quad \text{as soon as } -\infty < \theta < \frac{1}{2}.$$

Choose for instance $\theta = \frac{\varepsilon}{2(1+\varepsilon)} < \frac{1}{2}$, then

$$\mathbb{P}\left(\frac{1}{k} \sum_{j=1}^{k} z_j^2 > (1+\varepsilon) \right) \leq e^{-\frac{\varepsilon}{2(1+\varepsilon)} k (1+\varepsilon)} \left(\frac{1}{1-2\left(\frac{\varepsilon}{2(1+\varepsilon)}\right)} \right)^{k/2}$$

$$= \left(e^{-\varepsilon} \right)^{k/2} \left(\frac{1}{1-\frac{\varepsilon}{(1+\varepsilon)}} \right)^{k/2}$$

$$= \left((1+\varepsilon) e^{-\varepsilon} \right)^{\frac{k}{2}}.$$

For $\varepsilon > 0$: $1+\varepsilon \leq \exp\left(\varepsilon - \frac{(\varepsilon^2 - \varepsilon^3)}{2} \right)$

$$\leq \exp\left(-\frac{k}{4} (\varepsilon^2 - \varepsilon^3) \right).$$

2. Step

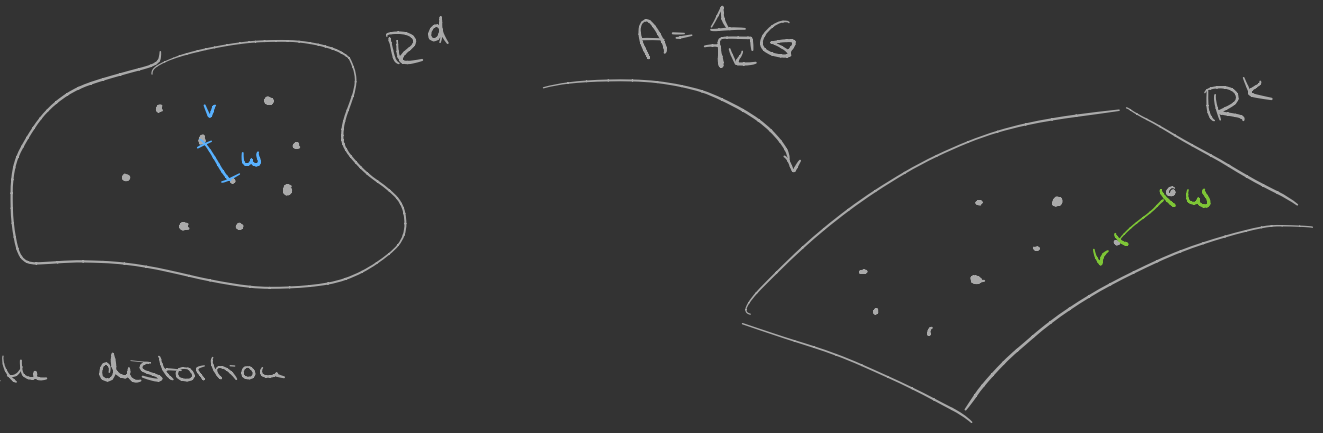
Analogously we can show that

$$\mathbb{P}\left(\frac{\|Ax\|_2^2}{\|x\|_2^2} < 1-\varepsilon \right) \leq \exp\left(-\frac{k}{4} (\varepsilon^2 - \varepsilon^3) \right).$$

$$\Rightarrow \mathbb{P}\left(\frac{\|Ax\|_2^2}{\|x\|_2^2} \notin [1-\varepsilon, 1+\varepsilon] \right) \leq 2\exp\left(-\frac{k}{4} (\varepsilon^2 - \varepsilon^3) \right).$$

∎

What we just proved:



with distortion

$$\frac{\text{—}}{\text{—}} \approx (1 \pm \varepsilon).$$

Problem: Gaussian R.V. can be unbounded, but very large entries in G could be problematic for certain applications.

Improvement 2:

$$G_{ij} = \begin{cases} 1, & \text{w.p. } \frac{1}{2} \\ -1, & \text{w.p. } \frac{1}{2} \end{cases} \rightsquigarrow G = \begin{pmatrix} -1 & 1 & 1 \\ 1 & -1 & -1 \\ \cdots & \cdots & \cdots \end{pmatrix}$$

and $A = \frac{1}{\sqrt{k}} G$ will actually give the exact same theorem as Thm 4.3.

Problem: G might be huge and therefore difficult to store.

Improvement 3:

$$A_{ij} = \begin{cases} +\sqrt{\frac{3}{k}}, & \text{w.p. } \frac{1}{6} \\ 0, & \text{w.p. } \frac{2}{3} \\ -\sqrt{\frac{3}{k}}, & \text{w.p. } \frac{1}{6} \end{cases}$$

gives a very sparse matrix, but we can find that

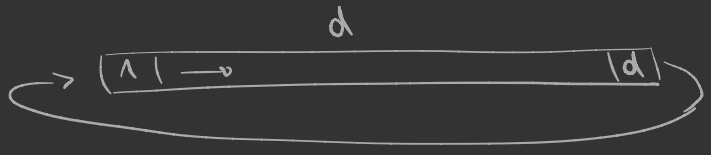
$$\rightsquigarrow A = \sqrt{\frac{3}{k}} \cdot \begin{pmatrix} -1 & 0 & 0 & 1 & 0 & 0 & \cdots \\ 1 & 0 & -1 & 0 & 0 & 0 & \\ 0 & 0 & 0 & -1 & \cdots & & \\ & & \ddots & & & & \end{pmatrix},$$

is much easier to store.

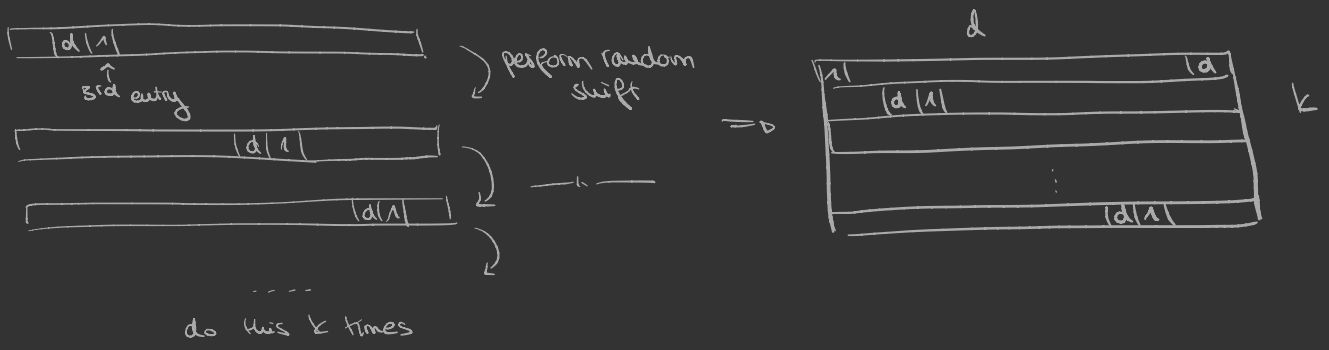
Problem: High randomization, i.e. need to generate and store the full matrix G or A respectively.

Improvement 4: Use a circulate-random matrix

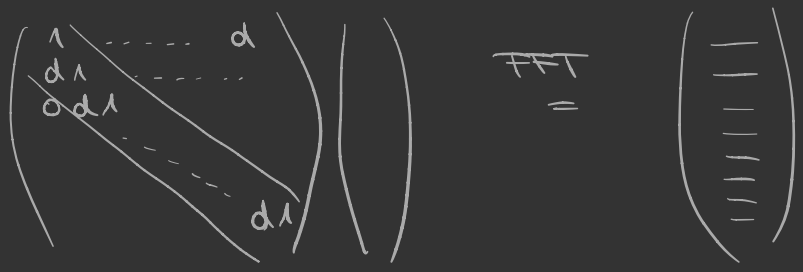
It starts with vector $(1, 0, \dots, 0, d) \in \mathbb{R}^d$, and is constructed using random shifts:



Do this k times to construct the matrix:



If we use a full circular matrix

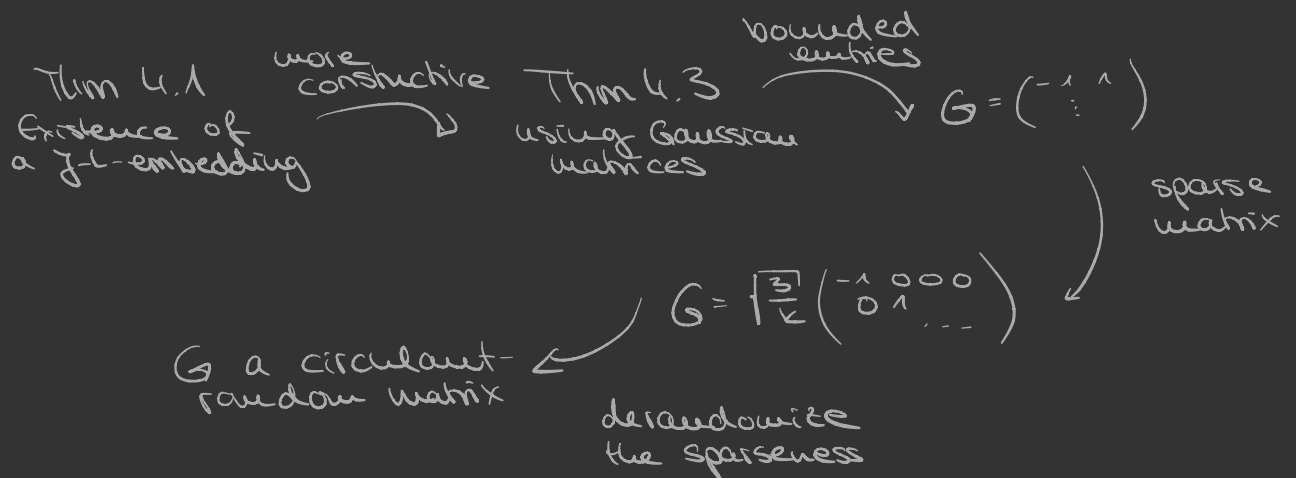


the matrix vector multiplication can be executed very fast using the Fast-Fourier-Transformation with complexity $(d \cdot \ln(d))$

In a second step we then randomly select k entries from the result vector (this corresponds with performing k random shifts).

Such circular-random matrices are actually also a Johnson-Lindenstrauss-embedding.

Summary of Chapter 4:



But there is no deterministic construction for Johnson-Lindenstrauss-embeddings.

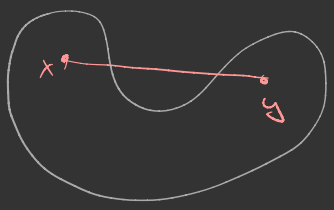
Lecture 19

30.06.2020

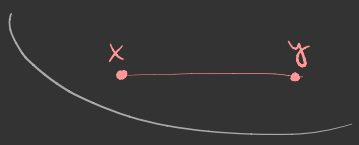
5. Convex Analysis

5.1 Convex Sets

not convex:



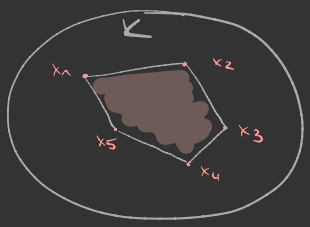
convex:



Def. 5.1: A subset $K \subset \mathbb{R}^N$ is called convex if $\forall x, y \in K$, the segment between x and y is completely contained in K, i.e.

$$tx + (1-t)y \in K \qquad \forall\, t \in [0,1]$$

Property: $K \subseteq \mathbb{R}^N$ is convex



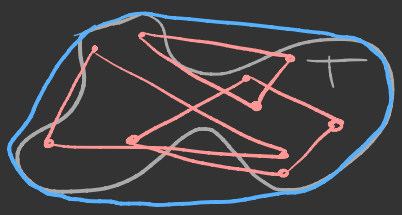
$\Leftrightarrow$ $\forall x_1, \dots, x_u \in K$ and $t_1, \dots, t_u \geq 0$ s.t. $\sum_{j=1}^{u} t_j = 1$, the convex combination $\sum_{j=1}^{u} t_j x_j \in K$.

Visualization: $x_1, \dots, x_u$ define vertices and all the space between them is contained in K.

Def. 5.2: Let $T \subseteq \mathbb{R}^u$ be a set. We define the convex hull of T as

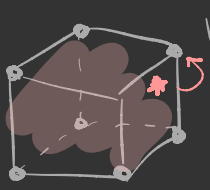
$$\mathrm{Conv}(T) = \left\{ \sum_j t_j x_j \;:\; t_j \geq 0,\ \sum_j t_j = 1,\ x_j \in T \right\}.$$

union of all these polytopes we can form using points in T



By definition $\mathrm{Conv}(T)$ is the smallest convex set, which contains T.

Property: Often, we need to optimize over low-convex sets, i.e. in integer optimizations over $[-1,1]^u \cap \mathbb{Z}^u$.

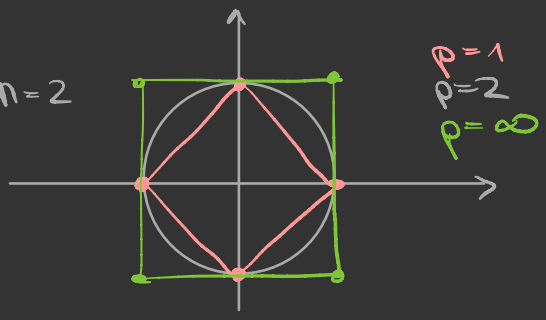


In these situations we can consider a convex relaxation of the problem, in this case optimizing over the domain $[-1,1]^u$ and then find the closest point to the optimal solution of the relaxed problem.

Example: $B_p(0,1)$, the ball around 0 with radius 1, w.r.t. the p-norm.

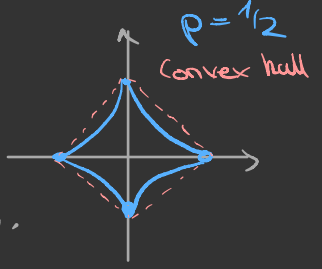
$$\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p}$$

$$B_p(0,1) = \{ \|x\|_p \leq 1 \}.$$



$B_p(0,1)$ is convex for $1 \leq p \leq \infty$.

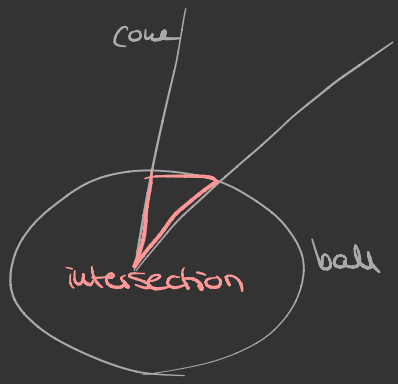
$B_p(0,1)$ is not convex for $0 < p < 1$.



Exercise: For $0 < p < 1$: $\text{Conv}(B_p(0,1)) = B_1(0,1)$.

More examples:
- linear subspaces
- convex cones
- balls of norms
- intersection of convex sets
- systems of linear (in)equalities

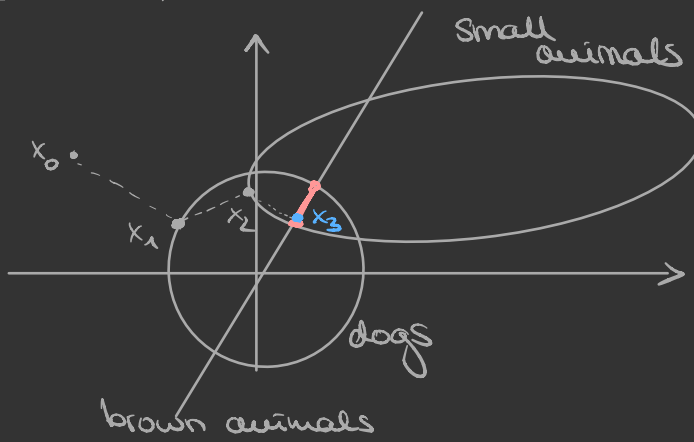
$$a_1 x_1 + \dots + a_n x_n \leq b.$$



Example: POCS: "Projection onto convex sets"

Want to find x, s.t.

$x \in C_1$ dogs
$x \in C_2$ small animals
$x \in C_3$ brown animals



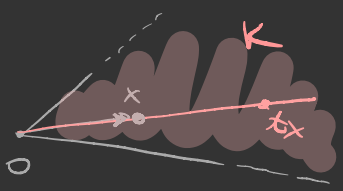
POCS defines an iteration to find such a point:

$$x_{n+1} = P_{C_3} P_{C_2} P_{C_1} x_n.$$

see Exercise 3.2: Alternating Projection Theorem

Def. 5.3: A set $K \subset \mathbb{R}^n$ is called <u>cone</u>, if

$$\forall x \in K \ \forall t \geq 0 : \quad t \cdot x \in K.$$



If in addition, K is convex, then K is called <u>convex cone</u>.

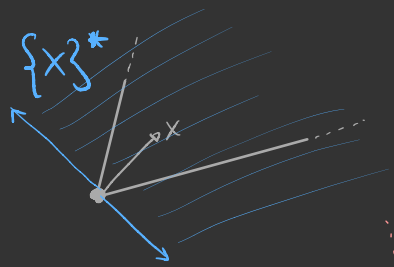
Remark: By definition, since $t=0$ is allowed, $0 \in K$.

Exercise: If K is a convex cone, then

$$\forall x, y \in K \ \forall s, t \geq 0 : \quad sx + ty \in K.$$

Def.: Let $K \subseteq \mathbb{R}^n$ be a cone. Then, its <u>dual cone K^*</u> is defined by

$$K^* = \left\{ z \in \mathbb{R}^n \,\middle|\, \langle x, z \rangle \geq 0 \ , \ x \in K \right\}.$$



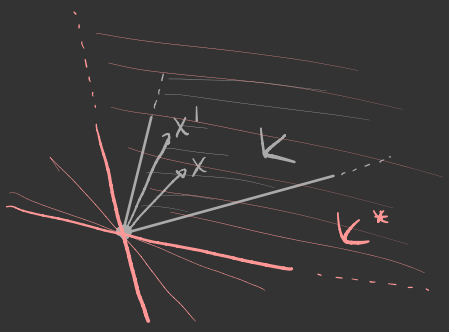
$\langle x, z \rangle \geq 0$ describes a half space (convex) for a fixed $x \in K$.

$\| \quad$

$x_1 z_1 + x_2 z_2 + \dots + x_n z_n \geq 0$

$\Downarrow \ \forall x \in K$

intersection of convex sets.



Exercise: If K is a closed cone, then $(K^*)^* = K$.

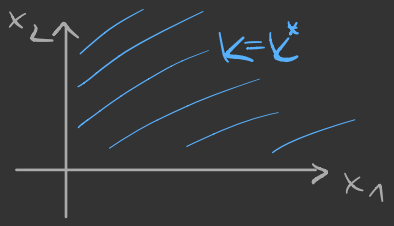
Exercise: $H, K \subset \mathbb{R}^n$ cones. Then

$$H \subset K \quad \Rightarrow \quad K^* \subset H^*.$$

↝ "switches" due to more constraints by larger set.

Property: There exist cones which are <u>self-dual</u>, i.e. $K^* = K$.
One example is

$$K = \mathbb{R}^n_+ = \left\{ x \in \mathbb{R}^n \,\middle|\, x_i \geq 0 \ , \ i \in [n] \right\}.$$



Theorem 5.4: Hahn-Banach

useful to understand SVMs

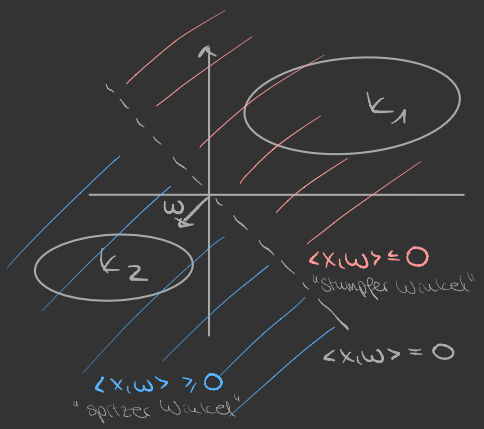
Let $K_1, K_2 \subseteq \mathbb{R}^n$ convex sets, s.t. their interior have an empty intersection, i.e. $\mathring{K}_1 \cap \mathring{K}_2 = \emptyset$. Then $\exists\, w \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, s.t.

$$K_1 \subset \{x \in \mathbb{R}^n \mid \langle x, w \rangle \leq \lambda\}$$

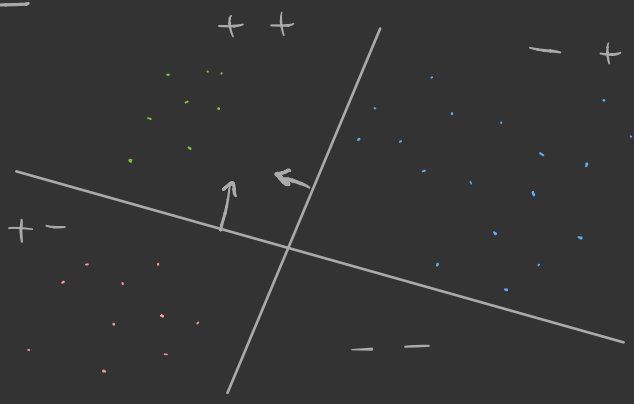
and

$$K_2 \subset \{x \in \mathbb{R}^n \mid \langle x, w \rangle \geq \lambda\}.$$

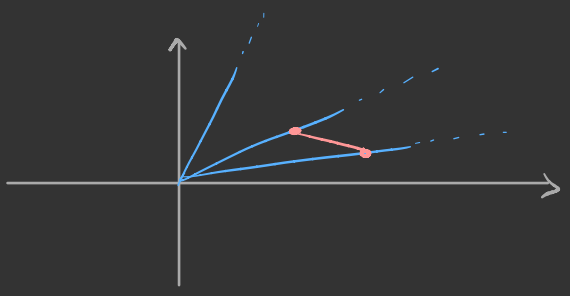
i.e. $\langle x, w \rangle = \lambda$ separates K_1, K_2.



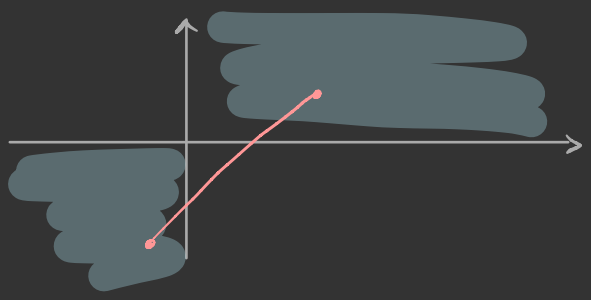
Approach in SVM:

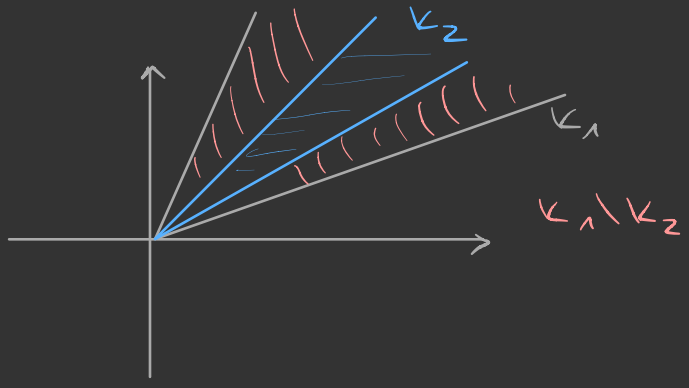


Example for non-convex cone:



- union of lines is a cone
- not convex





5.2. Convex Functions

Definition: Let $F: \mathbb{R}^n \to (-\infty, \infty]$. We define the **domain of F** by

$$\operatorname{dom}(F) = \{ x \in \mathbb{R}^n \mid F(x) \neq +\infty \}.$$

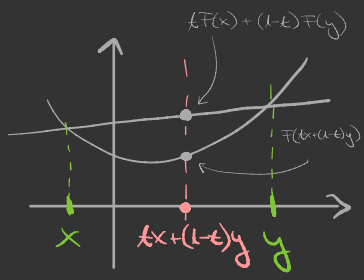
A function whose domain is not empty, i.e. $\operatorname{dom}(F) \neq \emptyset$, is called **proper**.

↝ function might be equal to infinity in some points, but not everywhere / there are some function values $< \infty$.

Remark: Let $F: K \to \mathbb{R}$, $\emptyset \neq K \subseteq \mathbb{R}^N$. We can always extend F to be an **extended valued function** $\tilde{F}: \mathbb{R}^n \to (-\infty, \infty]$ by setting $\tilde{F}(x) = \infty$ for all $x \in \mathbb{R}^n \setminus K$. Then $\operatorname{dom}(\tilde{F}) = K$.

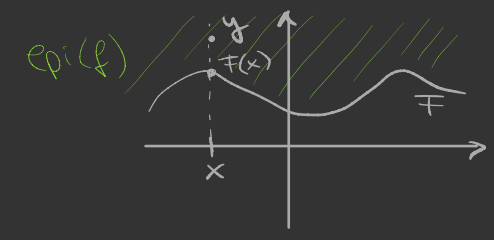
Def. 5.7: Let $F: \mathbb{R}^n \to (-\infty, \infty]$. Then F is called **convex**, if $\forall x, y \in \mathbb{R}^N$, $t \in [0,1]$, we have that

$$F(tx + (1-t)y) \leq t F(x) + (1-t) F(y).$$



i.e. the connecting segment is always "over" the graph.

F is called **strictly convex**, if the inequality is strict for $t \in (0,1)$.



Remark: Define the **epigraph of F** by

$$\operatorname{epi}(F) = \{(x,y) \in \mathbb{R}^n \times \mathbb{R} \mid F(x) \leq y\}.$$

Then F is convex (as a function)

$\Leftrightarrow$ $\operatorname{epi}(F)$ is convex (as a set).

↝ Relationship between convex sets & convex functions.

Vorlesung 20

01.07.2020

Def: $F: \mathbb{R}^N \to (-\infty, \infty]$ is called **strongly convex with parameter $\gamma > 0$** if for all $x, y \in \mathbb{R}^N$ and $t \in [0,1]$:

$$F(tx+(1-t)y) \leq tF(x)+(1-t)F(y) - \frac{\gamma}{2} t(1-t)\|x-y\|_2^2$$

($t(1-t)$: parabola)

F is called **(strictly, strongly) concave** if $-F$ is (strictly, strongly) convex.

$\rightsquigarrow$ concavity $\sim$ maximization

convexity $\sim$ minimization

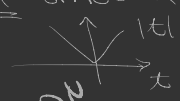
Exercise: strongly convex $\Rightarrow$ strictly convex $\Rightarrow$ convex

Exercise: The domain of a convex function is convex (as a set).

Def: A function $F: K \to \mathbb{R}$ defined on a convex set $K \subseteq \mathbb{R}^N$ is called convex, if its canonical extension $\tilde{F}: \mathbb{R}^N \to (-\infty, \infty]$ is convex.

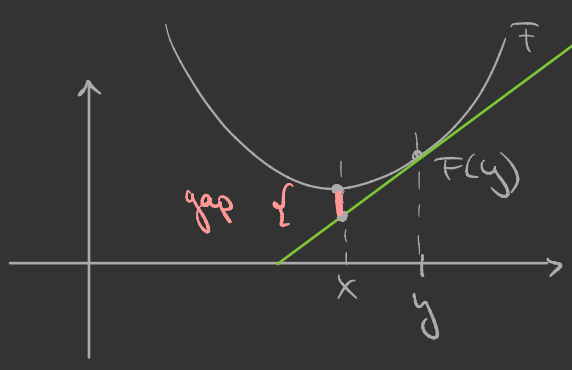
Proposition 5.8: Let $F: \mathbb{R}^N \to \mathbb{R}$ be diff'able.

$\rightsquigarrow$ In general convex function are not smooth



(i) F is convex

$\Leftrightarrow$ $F(x) \geq F(y) + \langle \nabla F(y), x-y \rangle$ $\quad \forall x, y \in \mathbb{R}^N$



(ii) F is strongly convex with parameter $\gamma > 0$

$$\Leftrightarrow F(x) \geq F(y) + \langle \nabla F(y), x-y \rangle + \frac{\gamma}{2} \|x-y\|_2^2 \quad \forall x,y \in \mathbb{R}^N$$

(iii) If F is twice diffable. Then F is convex

$$\Leftrightarrow \nabla^2 F(x) \succeq 0 \qquad \forall x \in \mathbb{R}^N$$

↖ Hessian is pos. semi definite

Remark: Gradient of $F : \mathbb{R}^N \to \mathbb{R}$:

$$\nabla F(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} F(x) \\ \vdots \\ \frac{\partial}{\partial x_n} F(x) \end{pmatrix}$$

Hessian matrix

$$\nabla^2 F(x) = \left(\frac{\partial^2}{\partial x_i \partial x_j} F(x) \right)_{i,j \in [N]}$$

$$= \begin{pmatrix} \frac{\partial^2}{\partial^2 x_1} F(x) & \frac{\partial^2}{\partial x_1 \partial x_2} F(x) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n} F(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} F(x) & \frac{\partial^2}{\partial^2 x_2} F(x) & \cdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \frac{\partial^2}{\partial^2 x_n} F(x) \end{pmatrix}$$

A matrix $M \in \mathbb{R}^{N \times N}$ is called positive semi-definite, short $M \succeq 0$

$$:\Leftrightarrow \forall v \in \mathbb{R} : v^T M v = \langle v, Mv \rangle \geq 0$$

Proof (Idea): Taylor expansion:

$$F(x) = F(y) + \langle \nabla F(y), x-y \rangle + \frac{1}{2} \langle (x-y), \nabla^2 F(y)(x-y) \rangle$$

$$= F(y) + \langle \nabla F(y), x-y \rangle + \int_0^1 \xi_t^T \nabla^2 F(y) \xi_t \, dt$$

$\nabla^2 F \geq 0$ $\Rightarrow$ ≥ 0, for $\nabla^2 F \geq 0$

$$\geq F(y) + \langle \nabla F(y), x-y \rangle$$

and hence exactly the definition of being convex.

Prop. 5.9:

(i) Let F, G be convex function, then for $\alpha, \beta > 0$
$\alpha F + \beta G$ is convex.

(ii) Let $F: \mathbb{R} \to \mathbb{R}$, $G: \mathbb{R}^N \to \mathbb{R}$ with F convex and non-decreasing and G convex. Then $H(x) = F(G(x))$ is convex.

Examples:

- Norms are convex due to Δ-inequality and homogeneity.
- l_p-norms, $1 < p < \infty$, strictly convex.
- l_p-norms, $p = 1, \infty$, not strictly convex. $\rightsquigarrow$ straight / flat sides of unit balls
- For $A \in \mathbb{R}^{N \times N}$, $A \geq 0$, then $f(x) = x^T A x$ is convex.
  f is strictly convex, if $A > 0$, i.e. strictly pos. definite.
- For a convex set K the characteristic function (≠ indicator function)

$$\chi_K(x) = \begin{cases} 0 & , x \in K \\ \infty & , \text{else} \end{cases}$$

is convex.

Proposition 5.11: Let $F: \mathbb{R}^N \to \mathbb{R}$ (in particular $\operatorname{dom}(F) = \mathbb{R}^N$) be convex. Then F is continuous.

Remark: No function which takes on the value ∞ can be continuous. → for these, we need the following concept

Def. 5.12: A function $F: \mathbb{R}^N \to (-\infty, \infty]$ is called **lower semicontinuous** if for all $x \in \mathbb{R}^N$ and every sequence $(x_j)_{j \in \mathbb{N}} \subset \mathbb{R}^N$ with $x_j \to x$, it holds that

$$\liminf_{j \to \infty} F(x_j) \geq F(x).$$

Example: •) χ_K if K is closed, convex.

Remark: A function F is lower-semicontinuous
$\iff$ $\operatorname{epi}(F)$ is closed.

Def.: A **(global) minimizer** of a function $F: \mathbb{R}^N \to (-\infty, \infty]$ is a point $x \in \mathbb{R}^N$ s.t. $\forall y \in \mathbb{R}^N: F(x) \leq F(y)$.

A local minimum of F is a point $x \in \mathbb{R}^N$ s.t.
$\exists \varepsilon > 0$ s.t. $F(x) \leq F(y)$ $\forall y \in B_\varepsilon(x)$.

Prop 5.13: Let $F: \mathbb{R}^N \to (-\infty, \infty]$ be convex.

(i) A local minimum of F is a global minimum.

(ii) The set of minima of F is convex. ↳ equal function value

(iii) If F is strictly convex, then the minimum is unique.

Proof (of iii): Assume that F is strictly convex and $\exists x, \tilde{x} \in \mathbb{R}^N$, $x \neq \tilde{x}$, global minimizers, i.e. $F(x) = F(\tilde{x}) \leq F(y)$ $\forall y \in \mathbb{R}^N$.

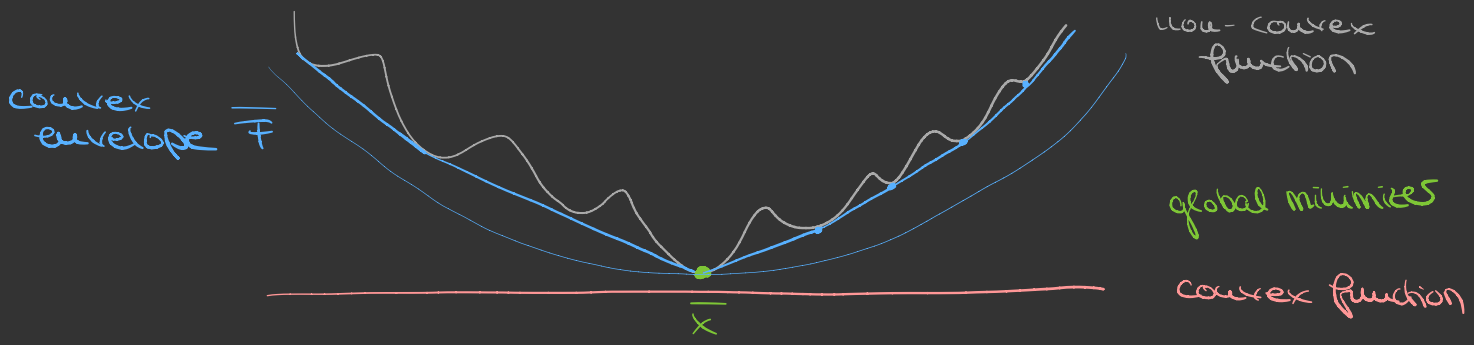
Then

$$F(\underbrace{tx + (1-t)\tilde{x}}_{\neq x, \tilde{x}}) < tF(x) + (1-t)F(\tilde{x}) = F(x)(t+1-t)$$
$$= F(x)$$
$$= \min_z F(z)$$

↯ $x, \tilde{x}$ global minimizers

$\Rightarrow x = \tilde{x}$.

Revisiting the idea of convex relaxation



$$\overline{F}(x) = \max_{\substack{G \leq F \\ G \text{ convex}}} G(x) \quad \rightsquigarrow \quad \overline{x} = \operatorname*{argmin}_x \overline{F}(x) = \operatorname*{argmin}_x F(x)$$

Depending on F, getting to $\overline{F}$ can be easy. Also a "good enough" estimation can be sufficient. Sometimes it's as hard as solving the original problem in the first place.

Another option: Graduating Non-Convexity

We define

$$F(x) = \underbrace{F(y)}_{\text{might want to move in a "non-convex way"}} + \underbrace{\lambda \|x-y\|_2^2}_{\text{adding a parabola which is a strong convex component}}$$

Then

$$y \longmapsto F(y) + \lambda \|x-y\|_2^2$$

is convex. With this we can write

$$x^{u+1} = \operatorname*{argmin}_y \underbrace{F(y) + \lambda_u \|x^u - y\|^2}_{\text{convex}}.$$

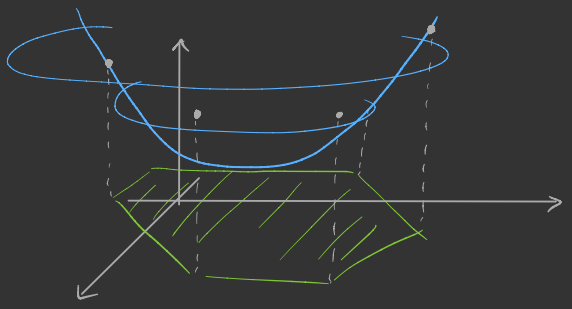
As soon as $\|x^{u+1} - x^u\| \longrightarrow 0$, this means that we are approaching a minimizer, because $\lambda_n \|x^u - x^{u+1}\|$ is getting more and more irrelevant in our optimizing steps.

$$\Leftrightarrow \quad \nabla F(x^{u+1}) + 2\lambda_n (x^u - x^{u+1}) = 0$$

$$\Leftrightarrow \quad 2\lambda_n (x^{u+1} - x^u) = \nabla F(x^{u+1})$$

$$\Leftrightarrow \quad x^{u+1} = x^u - \frac{1}{2\lambda_n} \nabla F(x^{u+1}) \qquad \text{"gradient descent"}$$

Theorem 5.15: Let $K \subset \mathbb{R}^N$ be a compact, convex set, and $F: K \to \mathbb{R}$ be a convex function. Then F attains its maximum at an extreme point of K.



Example: Finding 1-rank-matrices in subspaces which are generated by 1-rank-matrices.

$\{u_1, \dots, u_m\}$ orthonormal vectors. We compute

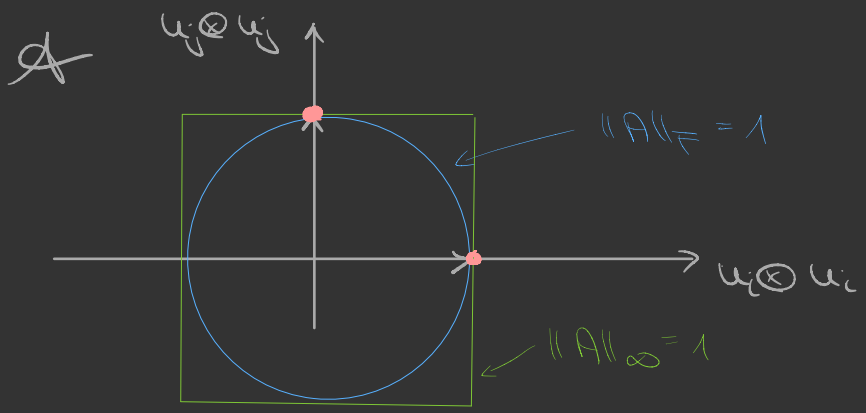
$$u_j \otimes u_j = u_j u_j^T \quad \text{1-rank-matrix.}$$

We define

$$\mathcal{A} = \operatorname{span}\{ u_j \otimes u_j \mid j \in [m] \}.$$

Then $\forall A \in \mathcal{A}$:

$$A = \sum_{j=1}^{m} \alpha_j u_j u_j^T = \sum_{j=1}^{m} |\alpha_j| \operatorname{sign}(\alpha_j) u_j u_j^T$$



Then $\langle u_i \otimes u_i, u_j \otimes u_j \rangle_F = \langle u_i, u_j \rangle^2 = \delta_{ij}$ (mutually $\perp_F$)

and hence

$$\{ u_j \otimes u_j \} = \operatorname*{argmax}_{A \in \mathcal{A},\ \|A\|_F \le 1} \|A\|_F$$

- $u_j \otimes u_j$: 1-rank matrix we want to find
- $\|A\|_F \le 1$: convex set
- $\|A\|_F$: convex function
- any matrix $A \in \mathcal{A}$.

Thm 5.15 $\Rightarrow$ maxima at extremal points / boundary.

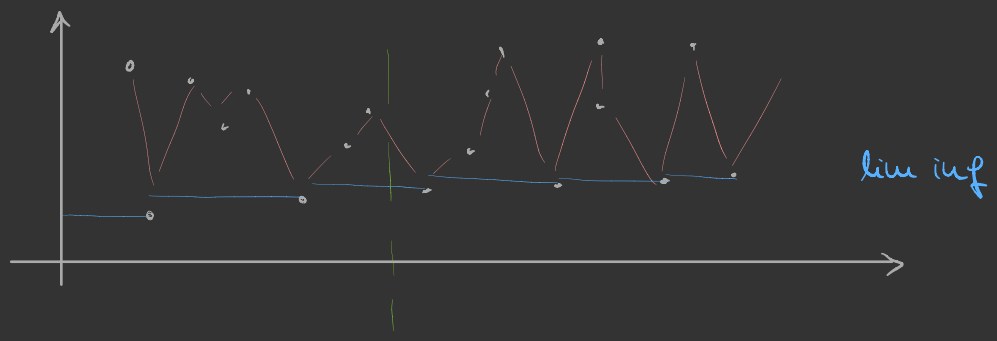
Example: One hidden layer neural network

$$F(x) = \sum_{j=1}^{m} g_j(\langle u_j, x \rangle)$$

$$\nabla F(x) = \sum_{j=1}^{m} g_j'(\langle u_j, x \rangle) \cdot u_j \quad \rightsquigarrow \text{linear combination of the first layer weights}$$

$$\nabla^2 F(x) = \sum_{j=1}^{m} g_j''(\langle u_j, x \rangle)\, u_j \otimes u_j \quad \in \mathcal{S}.$$

Recall: Let $(x_n)_{n \in \mathbb{N}}$ be a sequence.



$$\liminf_{n \to \infty} x_n = \sup_{n} \inf_{m \geq n} x_m.$$

Vorlesung 21:

07.07.2020

5.3. Convex Conjugate

Def 5.16: Let $F: \mathbb{R}^n \to (-\infty, \infty]$. Then we define its convex conjugate $F^*: \mathbb{R}^n \to (-\infty, \infty]$ by

$$F^*(y) := \sup_{x \in \mathbb{R}^n} \{ \underbrace{\langle x, y \rangle - F(x)}_{\text{linear in } y} \}.$$

Remark: F^* is always a convex function.

Property: Fenchel inequality

By definition we have that

$$\langle x, y \rangle \leq F(x) + F^*(y) \qquad \forall x, y \in \mathbb{R}^n.$$

Example: $0 \leq \|x - y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2\langle x, y \rangle$

$$\Rightarrow \langle x, y \rangle \leq \frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|y\|_2^2$$

Defining $F(x) = \frac{1}{2}\|x\|_2^2$, we get that

$$\langle x, y \rangle \leq F(x) + F(y).$$

We can prove that $F^*(x) = \frac{1}{2}\|x\|_2^2 = F(x)$.

Exercise: $1 \leq p < \infty$. If we consider

$$F(x) = \frac{1}{p}\|x\|_p^p, \quad F^*(x) = ?$$

Proposition 5.17: Let $F: \mathbb{R}^n \to (-\infty, \infty]$. Then

(i) F^* is lower semicontinuous.

$\forall (x_n) \xrightarrow{n \to \infty} x : \liminf_n F^*(x_n) \geq F(x)$.

(ii) The biconjugate F^{**} is the largest, lower semicontinuous function satisfying

$$F^{**} \leq F(x).$$

If F is convex and lower semicontinuous, then $F^{**} = F$.

$F^{**} = \sup_y \{ \langle y, z \rangle - F^*(y) \}$ defines the convex envelope.

But: computing the convex conjugate means to produce a minimization over F^*, which is itself a difficult problem itself.

(iii) Define $F_\tau(x) = F(x \cdot \tau)$ for $\tau \neq 0$. Then

$$(F_\tau)^*(x) = F^*\left(\frac{y}{\tau}\right).$$

Proof: $(F_\tau)^*(x) \overset{\text{Def.}}{=} \sup_x \{ \langle x, y \rangle - F_\tau(x) \}$

$$= \sup_x \left\{ \left\langle \tfrac{\tau}{\tau} x, y \right\rangle - F(x \cdot \tau) \right\}$$

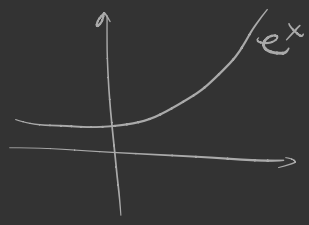
$$= \sup_{\tau x} \left\{ \left\langle \underbrace{\tau x}_{=:\bar{x}}, \frac{y}{\tau} \right\rangle - F(\underbrace{x \tau}_{=:\bar{x}}) \right\} = F^*\left(\frac{y}{\tau}\right).$$

(iv) For $\tau > 0$: $(\tau F)^*(y) = \tau F^*\left(\frac{y}{\tau}\right) = \tau \cdot F_\tau^*(y)$

(v) For $z \in \mathbb{R}^n$ define $F^{(z)} = F(x - z)$. Then

$$\left(F^{(z)}\right)^*(y) = \langle z, y \rangle + F^*(y).$$

Example: Define $F(x) = e^x$, $n = 1$.



Observations: •) F convex.

•) $x \longmapsto x \cdot y - \exp(x)$ takes its maximum (for y fixed) at $x = \ln(y)$, if $y > 0$.

$\frac{\partial}{\partial x}\, x \cdot y - \exp(x) = y - \exp(x) \overset{!}{=} 0$

$\Leftrightarrow y = \exp(x) \Leftrightarrow x = \ln(y) \quad \forall y > 0.$

$\frac{\partial^2}{\partial^2 x}\, x \cdot y - \exp(x) = -\exp(x) < 0 \quad \forall x \in \mathbb{R}$, hence maximum.

Then $F^*(x) = \sup_x (xy - \exp(x))$

$$\downarrow$$

$$= \begin{cases} y \cdot \ln(y) - \exp(\ln(y)) = y \cdot \ln(y) - y & , y > 0. \\ 0 & , y = 0 \\ \infty & , y < 0 \end{cases}$$

Fenchel

$$\Rightarrow \quad xy \leq \exp(x) + y \cdot \ln(y) - y \qquad \forall x \in \mathbb{R},\ y > 0.$$