Agenda 24.11.23:

- Lösungen EA3
  - insbes. A3
  - ggf. Fragen
- Ideen u. Tipps KE4
  - Kugeln im $\mathbb{R}^d$
  - RIP
  - Orth.

3. Support Vector Machines I (10 Punkte). Wir betrachten die Situation aus der Motivation von KE 3. Wir haben also eine Menge $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ von Datenpunkten $x_i \in \mathbb{R}^d$ und deren Klassifizierungen $y_i \in \{-1, 1\}$. Wir nehmen an, die Daten seien linear separierbar, d.h. es gibt ein $w \in \mathbb{R}^d \setminus \{0\}$ und ein $b \in \mathbb{R}$, sodass für alle $i = 1, \ldots, N$ $y_i(\langle w, x_i \rangle + b) \geq 0$ erfüllt ist. Wir suchen nach der Hypothese $f_{w,b}$ bzw. Hyperebene $H_{w,b}$ für $(w, b) \in \mathbb{R}^{d+1}$, die die Geometric Margin maximiert:

$$\rho = \max_{w,b: y_i(\langle w, x_i\rangle + b) \geq 0\, \forall i} \rho_{f_{w,b}} = \max_{w,b: y_i(\langle w, x_i\rangle + b) \geq 0\, \forall i} \min_{i=1,\ldots,N} \frac{|\langle w, x_i \rangle + b|}{\|w\|}.$$

Ein Beispiel ist in der Abbildung skizziert.

a) (2 Punkte) Zeigen Sie, dass

$$\rho = \max_{w,b} \min_{i=1,\ldots,N} \frac{y_i(\langle w, x_i\rangle + b)}{\|w\|} = \max_{w,b: y_i(\langle w, x_i\rangle + b) \geq 1\, \forall i} \frac{1}{\|w\|}$$

gilt.

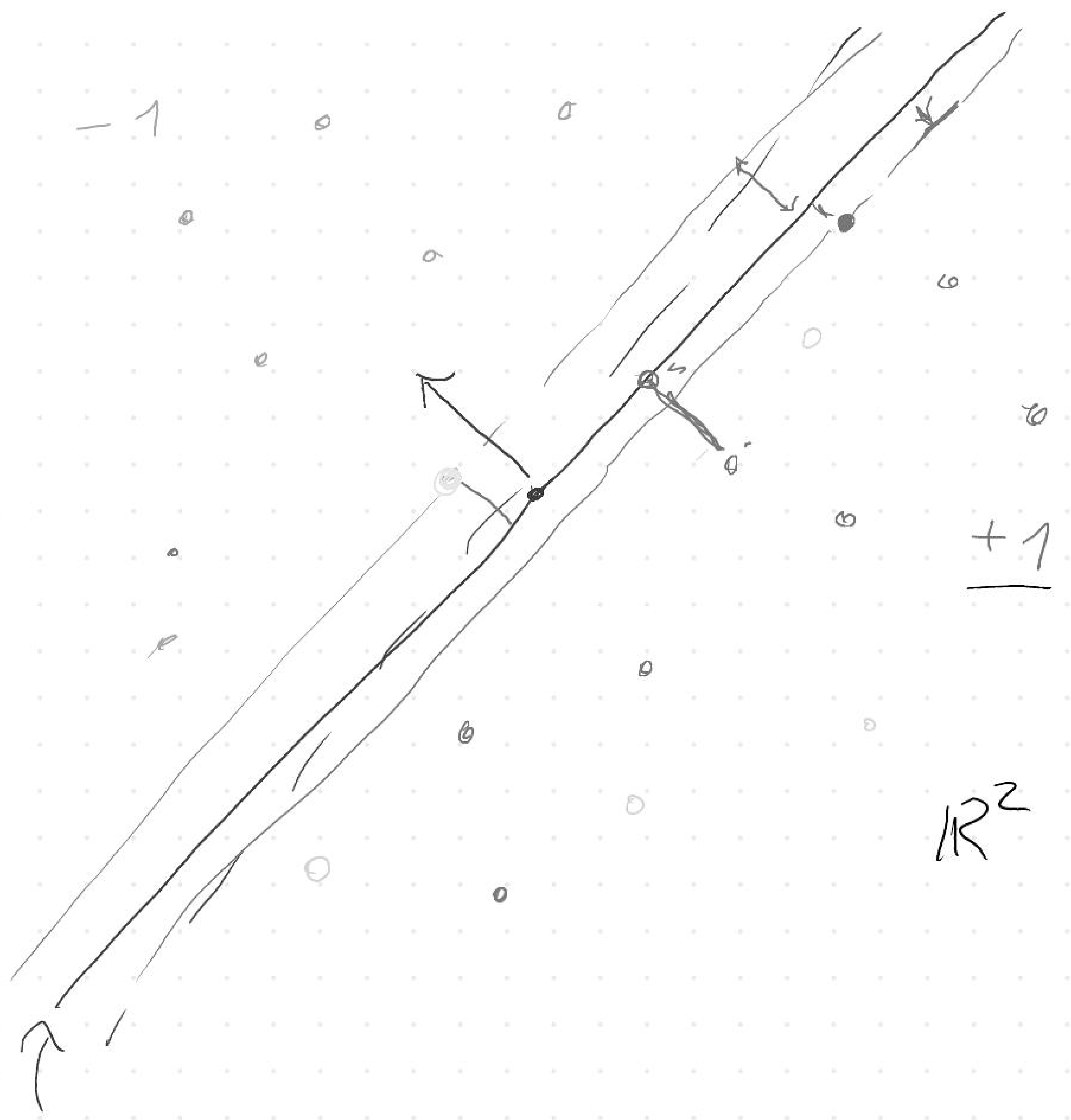
b) (2 Punkte) Begründen Sie, warum das optimale Paar (w, b) für das Problem aus a) genau die Lösung des Optimierungsproblems

$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$\text{unter } y_i(\langle w, x_i \rangle + b) \geq 1 \forall i$$

ist. Zeigen Sie, dass es sich um ein konvexes Optimierungsproblem handelt, welches Slater's Bedingung erfüllt.

c) (2 Punkte) Formulieren Sie die Lagrange-Funktion und die KKT-Bedingungen.

d) (2 Punkt) Zeigen Sie, dass ein optimaler Vektor w eine Linearkombination $w = \sum_{i=1}^N \tilde{\lambda}_i x_i$ der Vektoren $x_1, \ldots, x_N$ aus dem Trainingsset seien muss.



a) Es gibt mindestens ein Paar (w,b) gibt, sodass $y_i(\langle w, x_i\rangle + b) \geq 0$ für alle i. Das bedeutet, dass die Ungl. auch für das maximierende Paar von

$$\max_{w,b} \min_{i=1,..N} \frac{y_i(\langle w, x_i\rangle + b)}{\|w\|}$$

Für alle w,b für die es ein i gibt, sodass der Zähler negativ ist, ist der ganze Ausdruck negativ, also kann für diese nicht das Max. erreicht werden. Dann folgt

$$\rho = \max_{w,b} \min_{i=1,..N} \frac{y_i(\langle w, x_i\rangle + b)}{\|w\|} \text{ (erste Gleichung, Def. } \rho).$$

Sei w^*, b^* optimal. Dann gilt für $\tilde{w} = \frac{w^*}{(\min_i y_i(\langle w^*, x_i\rangle + b^*))}$, $\tilde{b} = \frac{b^*}{(\min_i y_i(\langle w^*, x_i\rangle + b^*))}$

$$\min_{i=1\dots N} \frac{y_i(\langle \tilde{w}, x_i\rangle + \tilde{b})}{\|\tilde{w}\|} = \min_{i=1,\dots N} \frac{y_i(\langle w^*, x_i\rangle + b^*)}{\|w^*\|} \geq \min_{i=1,\dots N} \frac{y_i(\langle w, x_i\rangle + b)}{\|w\|}.$$

Also ist auch $\tilde{w}, \tilde{b}$ ein optimales Paar. Das bedeutet, wir können uns auf die Paare einschränken, die $\underbrace{\min_{i=1,\dots N} y_i(\langle w, x_i\rangle + b) = 1}$ erfüllen:

$$\rho = \max_{w,b} \min_{i=1,\dots N} \frac{y_i(\langle w, x_i\rangle + b)}{\|w\|} = \max_{w,b:\, \min_i y_i(\langle w, x_i\rangle + b) = 1} \ \min_{i=1,\dots N} \frac{1}{\|w\|}$$

$$= \max_{w,b:\, \min_i y_i(\langle w, x_i\rangle + b) = 1} \frac{1}{\|w\|} = \max_{w,b:\, \min_i y_i(\langle w, x_i\rangle + b) \geq 1 \ \forall i} \frac{1}{\|w\|}. \rightsquigarrow \text{2. Gleichung}$$

---

b) Maximieren von $\frac{1}{\|w\|}$ bedeutet Minimieren von $\|w\|$. Das ist äquiv. zum Minimieren von $\frac{1}{2}\|w\|^2$. D.h.

$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$y_i(\langle w, x_i\rangle + b) \geq 1 \ \forall i.$$

Beispiel 3.4.18 "quadratisches Optimierungsproblem"!

$$\min_{x \in \mathbb{R}^n} \underbrace{\tfrac{1}{2} x^T Q x}_{\frac{1}{2}\|w\|^2} + c^T x$$
$$\text{unter } Ax \leq m.$$

Hier ist $x = (w, b)$, Q ist

$$\begin{pmatrix} 1 & 0 & 0 & & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ & & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (d+d)$$

$$w^T \begin{pmatrix} 1 & 0 & 0 & \dots \\ 0 & 1 & 0 & \dots \\ 0 & 0 & 1 & \dots \end{pmatrix} w = w_1^2 + w_2^2 + \dots$$

Q symmetrisch + pos. semidefinit. $c = 0$. Wir wählen als i-te Zeile

---

von A $(-y_i x_i)$ $(\forall i)$ und m als den Vektor bei dem alle Einträge -1 sind. D.h. konvexes Opt. Prob. + Slater's Bedingung erfüllt.

Beispiel 3.4.18. Ein *quadratisches Optimierungsproblem* ist von der Form

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}x^T Q x + c^T x$$
$$\text{unter} \quad Ax \leq b,$$

wobei Q eine symmetrische $d \times d$-Matrix ist und $c \in \mathbb{R}^d$. Weiter ist $A \in \mathbb{R}^{m\times d}$, $b \in \mathbb{R}^m$ und die Bedingung $Ax \leq b$ sei komponentenweise zu verstehen, d.h. jeder Eintrag des Vektors auf der linken Seite sei kleiner oder gleich dem entsprechenden Eintrag des Vektors auf der rechten Seite. Ist Q positiv semi-definit, so ist f konvex, d.h. es handelt sich um ein konvexes Optimierungsproblem, bei dem alle Nebenbedingungen affin sind. Damit ist Slater's Bedingung erfüllt und aus Satz 3.4.17 folgt, dass starke Dualität gilt. Die Lagrangefunktion ist gegeben durch

$$\mathcal{L}(x,\lambda) = \frac{1}{2}x^T Q x + c^T x + \lambda^T(Ax - b).$$

Um $F(\lambda) = \inf_{x\in\mathbb{R}^d} \mathcal{L}(x,\lambda)$ zu bestimmen, betrachten wir den Gradienten und setzen diesen gleich 0. Wir erhalten dann, dass an einer Stelle x^* ein Minimum angenommen wird genau dann wenn

$$Qx^* = -(c + A^T\lambda).$$

Ist Q sogar positiv definit, so ist Q invertierbar und $x^* = -Q^{-1}(c + A^T\lambda)$. Damit ist

$$\begin{aligned} F(\lambda) &= \mathcal{L}(x^*,\lambda) \\ &= -\frac{1}{2}\lambda^T A Q^{-1} A^T \lambda - (c^T Q^{-1} A^T + b^T)\lambda - \frac{1}{2}c^T Q^{-1} c. \end{aligned}$$

Daraus folgt, dass auch das duale Problem ein quadratisches konvexes Optimierungsproblem ist.

(Lagrange-Fkt: Def. 3.4.4)

c) Die Lagrange-Funktion ist

$$\mathcal{L}(x,\lambda) = \tfrac{1}{2} x^T Q x + \lambda^T (Ax - m)$$
$$= \frac{1}{2}\|w\|^2 + \sum_{k=1}^{N} \lambda_k (Ax - m)_k$$
$$= \tfrac{1}{2}\|w\|^2 - \sum_{k=1}^{N} \lambda_k \left(y_k(\langle x_k, w\rangle + b) - 1\right)$$

Wir schreiben das Problem wie in (3.23) aus dem Skript:

Einschub:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \qquad y_i(\langle w, x_i\rangle + b) \geq 1 \ \forall i.$$

$$\min_{x \in G} f(x) \quad \text{unter} \quad g_i(x) \leq 0 \quad i = 1 \dots N \qquad (3.22)$$

$$x = (w, b) = (w_1, \dots, w_d, b) \qquad f(x) = f(w_1, \dots, w_d, b) = \frac{1}{2}(w_1^2 + \dots + w_d^2)$$

$$-y_i\left(\sum_{k=1}^{d} w_k x_k^i + b\right) \leq -1$$

$$-y_i\left(\sum_{k=1}^{d} w_k x_k^i + b\right) - 1 =: g_i(w_1, \dots, w_d, b) \overset{!}{\leq} 0$$

+ Def. 3.4.4

KKT - Bed:

Theorem 3.4.20 (Karush-Kuhn-Tucker). *Wir betrachten ein allgemeines Optimierungsproblem (3.23) mit differenzierbaren Funktionen $f, g_1, \dots, g_m$ sowie $h_1, \dots, h_p$. Es gelte starke Dualität. Sei $x^* \in \mathbb{R}^d$ eine Lösung des primalen Problems und $(\lambda^*, \nu^*) \in \mathbb{R}^{m+p}$ eine Lösung des dualen Problems (3.24). Dann gelten die folgenden Aussagen:*

$$g_i(x^*) \leq 0 \quad \text{für } i = 1, \dots, m, \qquad (3.26)$$
$$h_j(x^*) = 0 \quad \text{für } j = 1, \dots, p, \qquad (3.27)$$
$$\lambda_i^* \geq 0 \quad \text{für } i = 1, \dots, m, \qquad (3.28)$$
$$\lambda_i^* g_i(x^*) = 0 \quad \text{für } i = 1, \dots, m, \qquad (3.29)$$
$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{p} \nu_j^* \nabla h_j(x^*) = 0. \qquad (3.30)$$

$$\begin{pmatrix} \partial x_1 f \\ \partial x_2 f \\ \vdots \end{pmatrix} + \sum \begin{pmatrix} \partial x_1 g \\ \partial x_2 g \\ \vdots \end{pmatrix} \cdot \lambda = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad ! \ \nabla \mathcal{L}$$

Die Bedingungen (3.26) - (3.30) nennt man auch Karush-Kuhn-Tucker *oder kurz* KKT-Bedingungen.

Die Bedingungen (3.26), (3.27) und (3.27) müssen gelten, da x^* und (λ^*, ν^*) notwendigerweise zulässige Punkte des primalen bzw. dualen Problems sein müssen. Die Bedingung (3.29) haben wir schon in Satz 3.4.19 gefolgert.

Wir werden nun sehen, dass im Fall eines konvexen Problems die KKT-Bedingungen auch hinreichend sind für die Existenz von Lösungen des dualen sowie des primalen Problems.

Satz 3.4.21. *Wir betrachten ein konvexes Optimierungsproblem (3.23). Sei (x^*, λ^*, ν^*) ein Punkt, der die KKT-Bedingungen erfüllt. Dann gilt starke Dualität und x^* löst das primale, (λ^*, ν^*) löst das duale Problem.*

für $f(x) = f(w, b) = \frac{1}{2}\|w\|^2$

$$g_i(w, b) = -(y_i(\langle x_i, w\rangle + b) - 1)$$

(3.26) $g_i(w,b) = -(y_i(\langle x_i, w\rangle + b) - 1) \leq 0 \quad \forall i$

(3.28) $\lambda \geq 0$

(3.29) $\lambda_i g_i(w, b) = 0 \ \forall i$

$\Leftrightarrow \forall i: \lambda_i = 0$ oder $y_i(\langle x_i, w\rangle + b) = 1$ $\quad \in \{-1, 1\}$

(3.30) $w_i + \sum_{k=1}^{N} \lambda_k y_k x_{i,k} = 0$ * Abl. nach $w_1, \dots, w_d$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \qquad \text{Abl. nach } b$$

---

d) Um das Optimum zu finden, setzen wir den Gradienten von $\mathcal{L}$ gleich Null:

$$D_w \mathcal{L}(x, \lambda) = w_i - \sum_{k=1}^{N} \lambda_k y_k x_{k,i} \overset{!}{=} 0 \qquad \text{I)}$$

$$D_b \mathcal{L}(x, \lambda) = -\sum_{i=1}^{N} \lambda_i y_i \overset{!}{=} 0$$

I) $w^* = \sum_{k=1}^{N} \lambda_k y_k x_k$ und $\sum_{i=1}^{N} \lambda_i y_i = 0$.

$$\begin{pmatrix} w_1 \\ \vdots \\ w_i \\ \vdots \\ w_d \end{pmatrix} = \sum_{k=1}^{N} \lambda_k y_k \underbrace{\begin{pmatrix} x_{k,1} \\ \vdots \\ x_{k,i} \\ \vdots \\ x_{k,d} \end{pmatrix}}_{x_k}$$

e) duales Problem: Wir setzen das optimale $x=(w,b)$ in die Lagrange-Fkt. ein: $\left(F(\lambda) = \inf_{x\in G} \mathcal{L}(x,\lambda) = \mathcal{L}(x^*(\lambda),\lambda) \right)$

$$\mathcal{L}(x^*,\lambda) = \frac{1}{2}\Big\| \underbrace{\sum_{k=1}^{N} \lambda_k y_k x_k}_{w^*} \Big\|^2 - \sum_{k=1}^{N} \lambda_k \Big(y_k \cdot \Big\langle x_k, \underbrace{\sum_{j=1}^{N} \lambda_j y_j x_j}_{w^*} \Big\rangle - 1 \Big) - b \cdot \underbrace{\sum_{k=1}^{N} \lambda_k y_k}_{=0}$$

$$\overset{\text{umformen}}{=} \dots = \sum_{k=1}^{N} \lambda_k - \frac{1}{2} \Big\| \sum_{k=1}^{N} \lambda_k y_k x_k \Big\|^2$$

Also wird das duale Problem:

$$\max_{\lambda\in\mathbb{R}^N} F(\lambda) = \max_{\lambda\in\mathbb{R}^N} \sum_{k=1}^{N} \lambda_k - \frac{1}{2} \sum_{k=1}^{N}\sum_{j=1}^{N} \lambda_k y_k \lambda_j y_j \langle x_k, x_j \rangle$$

unter $\sum_{k=1}^{N} \lambda_k y_k = 0$ und $\lambda \geq 0$.

Tipps KE 4:

A1) a) $K^d(\delta) = \delta^{1/4} K^d(1)$ $*$

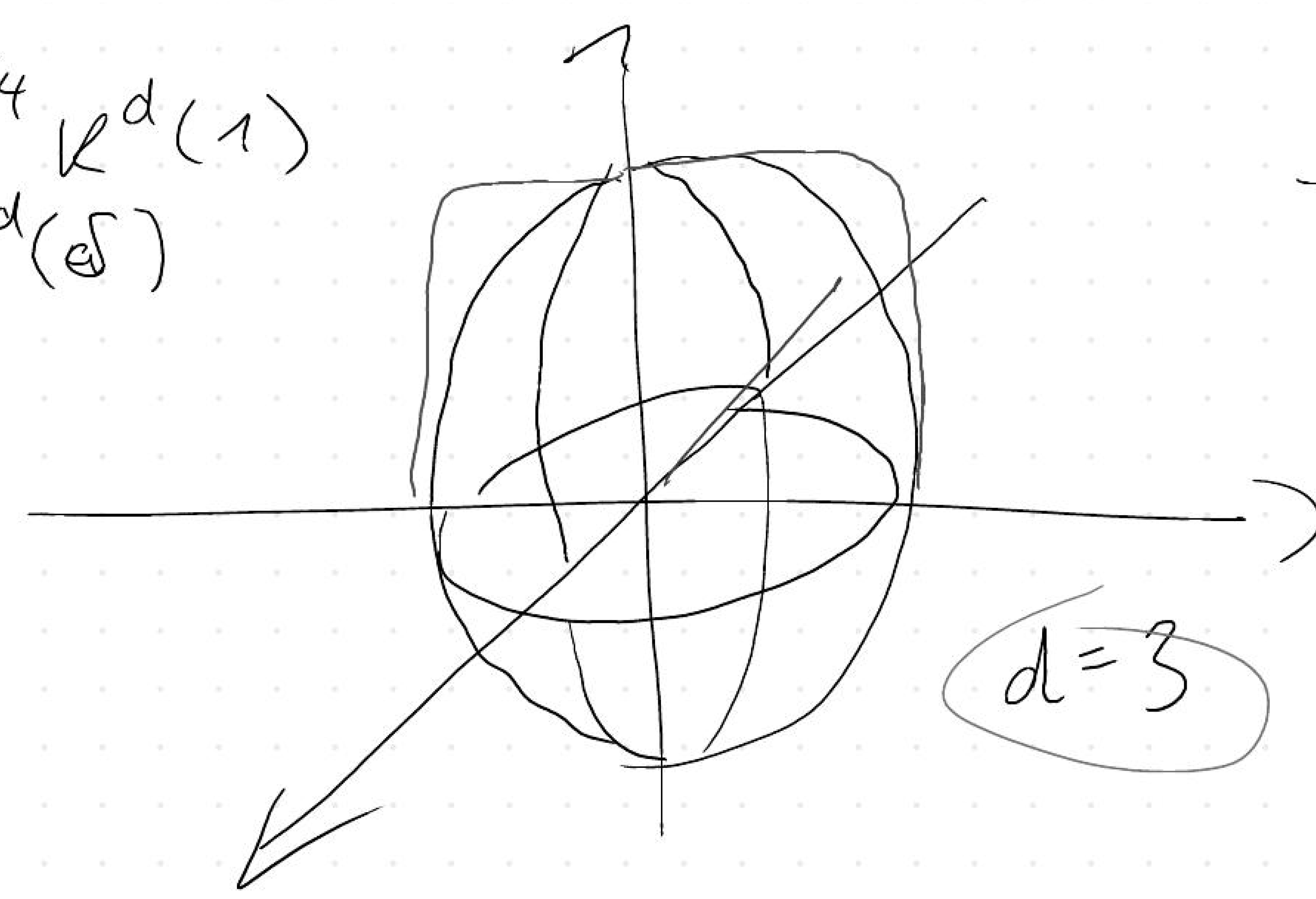
$K^d(\delta) = \{ x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i^4 \leq \delta \}$

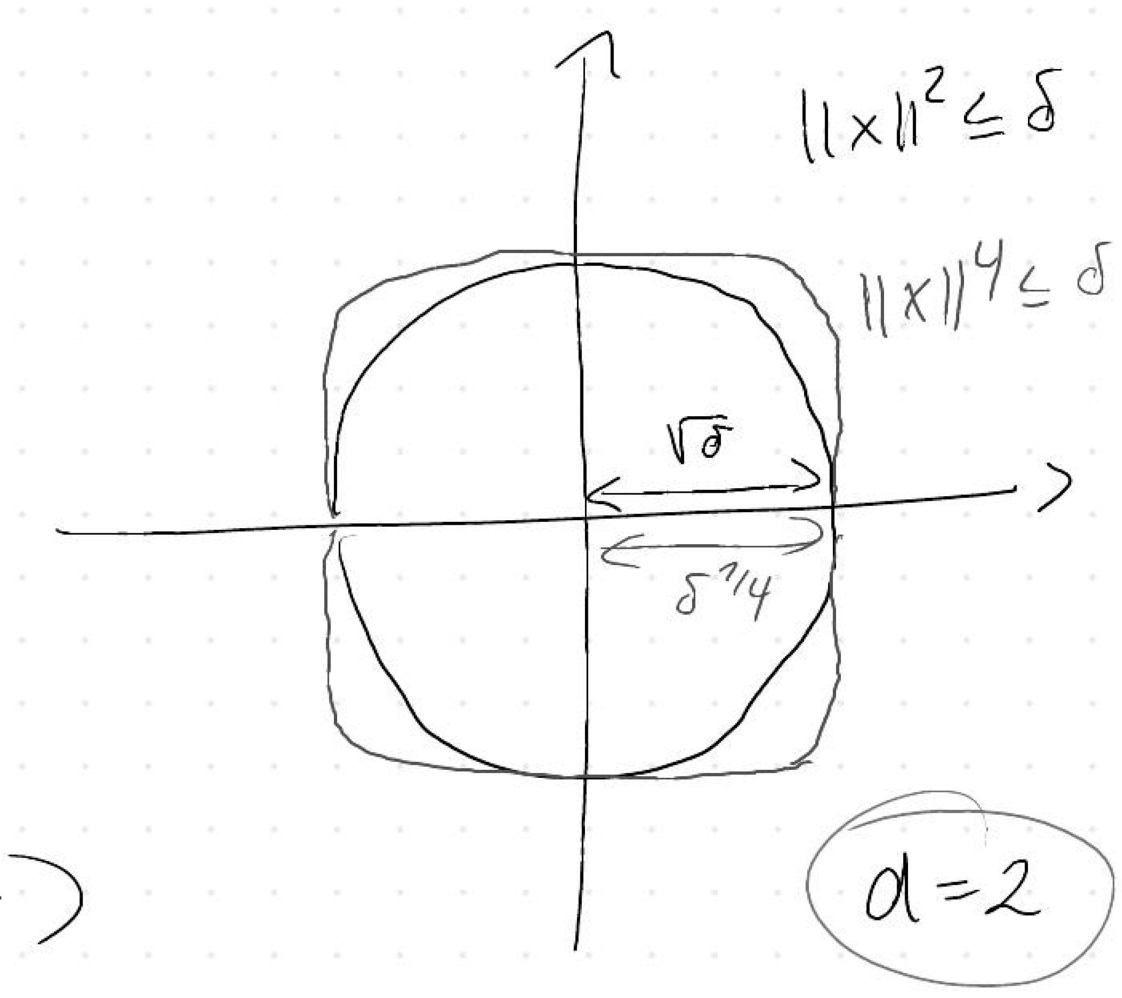
Man zeigt $*$, indem man

$x \in K^d(\delta) \Rightarrow x \in \delta^{1/4} K^d(1)$

$x \in \delta^{1/4} K^d(1) \Rightarrow x \in K^d(\delta)$

beweist

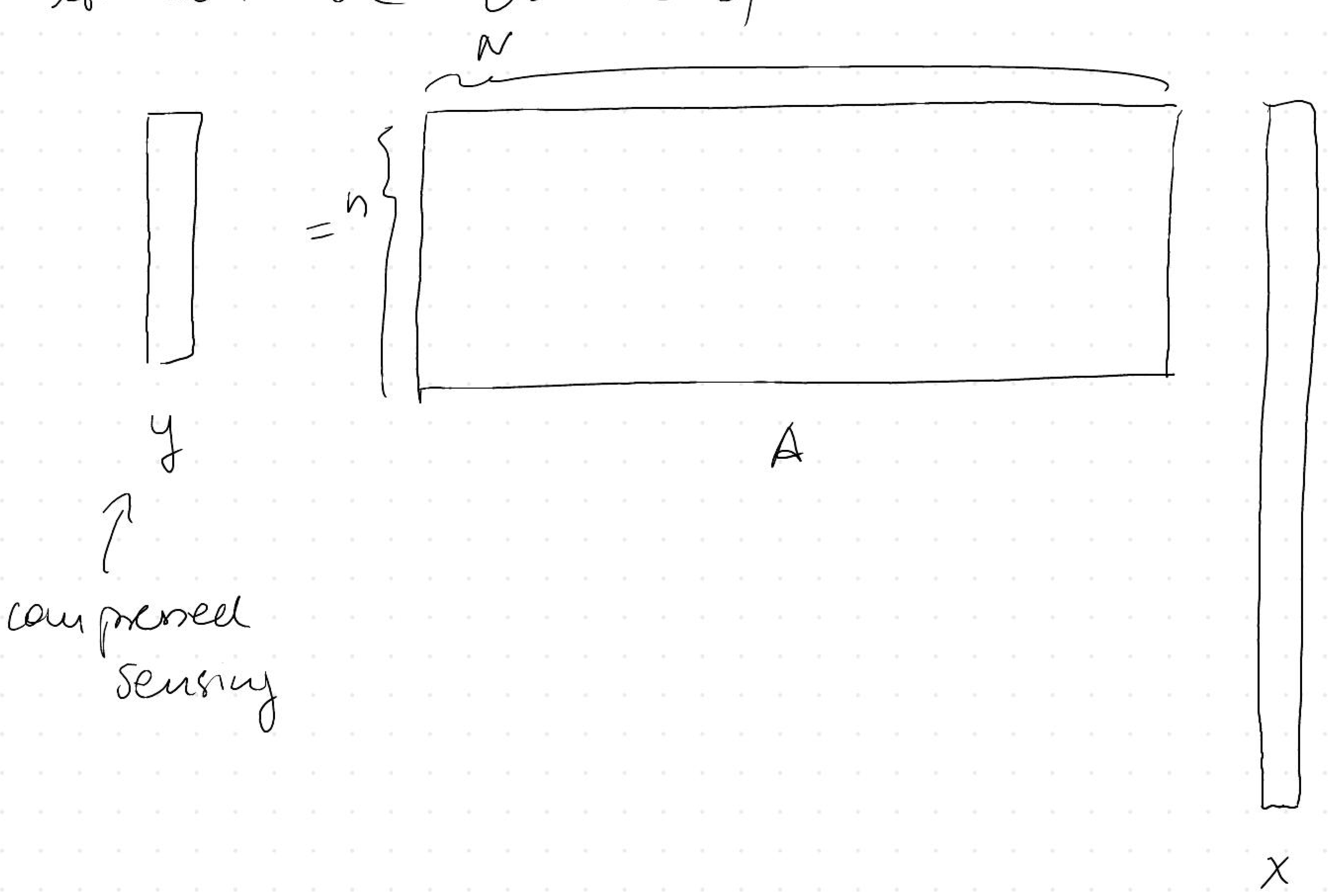




b) Satz 4.2.6

c) Beispiel 4.25

A2) Ähnlich wie JL-Transform!



Idee: Wenn A die RIP hat, d.h. "y"

$(1-\delta)\|x\|^2 \leq \|Ax\|^2 \leq (1+\delta)\|x\|^2$,

dann kann A für compr. sensing genutzt werden

A3) $\Rightarrow$ A mit normalvt. Einträgen kann dafür genutzt werden

1. $a \sim \mathcal{N}(\mu, \sigma \quad \Rightarrow \quad \frac{a-\mu}{\sigma} \sim \mathcal{N}(0,1)$

2. Satz 4.4.2 (+ Beweis)

A3) Beweis JL-T. + Fragen im Forum

c) iii) no "Mindestdimension" $A \in \mathbb{R}^{N \times n}$

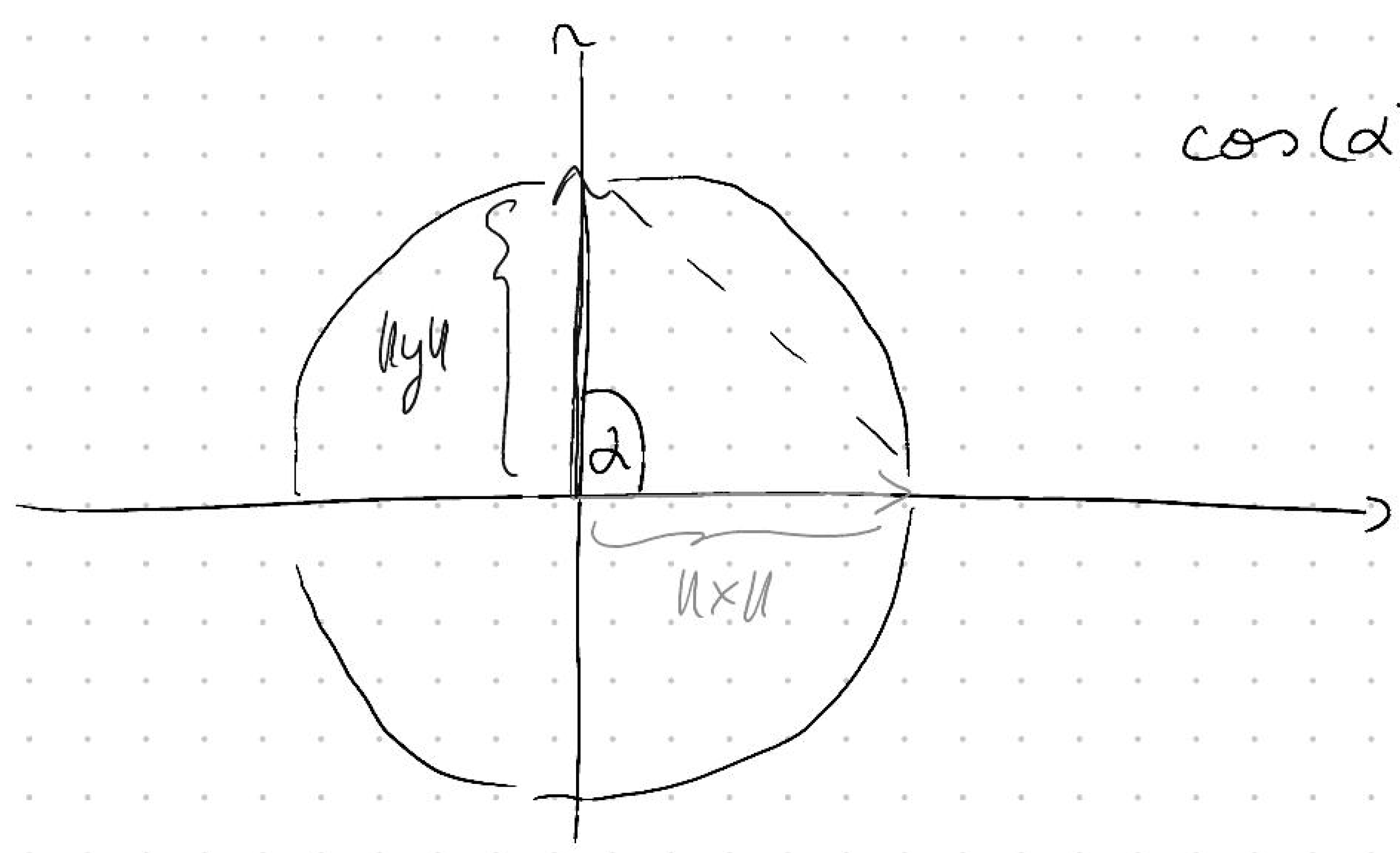
δ klein $\rightarrow$ genauer

ε klein $\rightarrow$ wahrscheinlicher

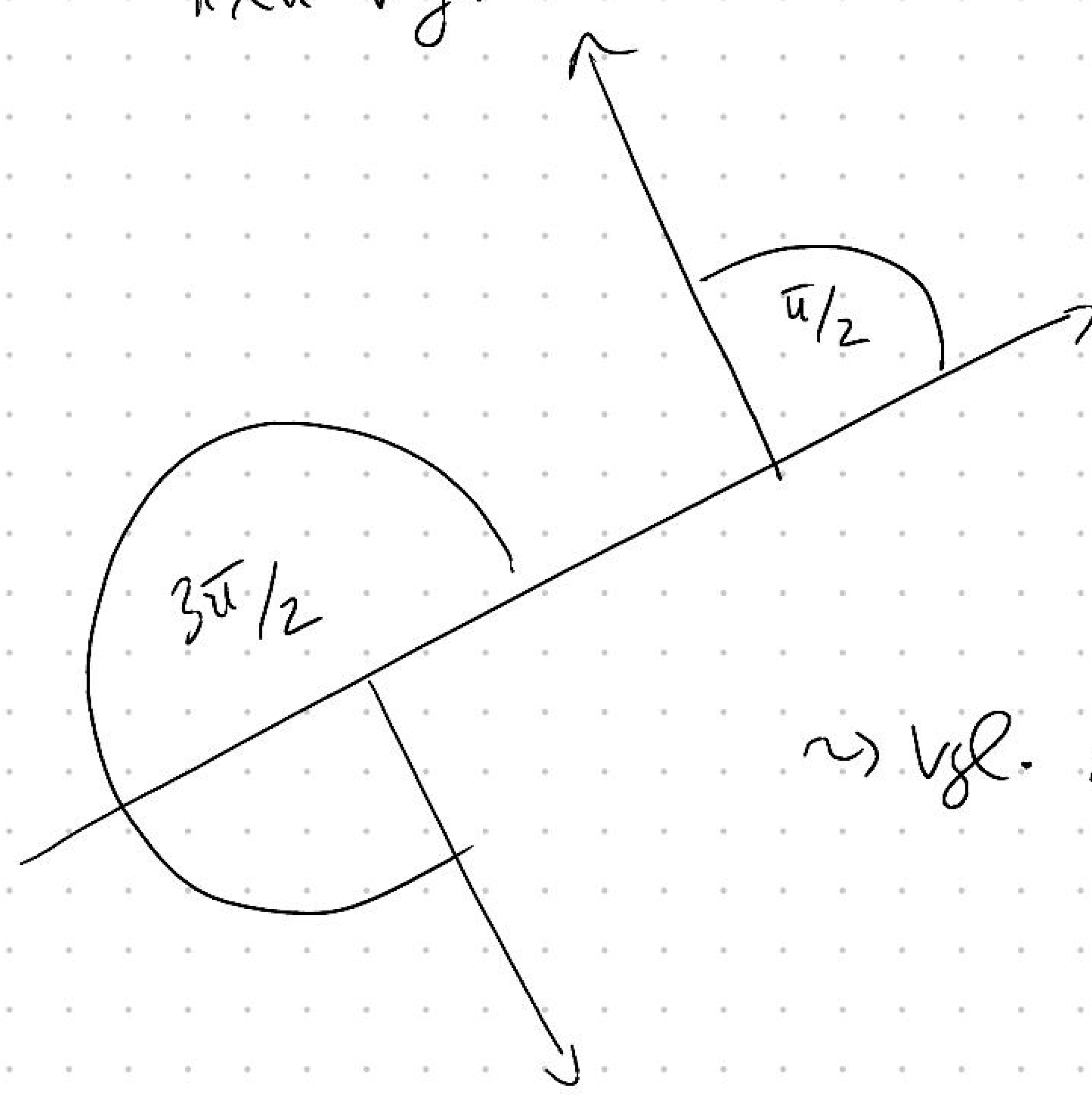
A4) Ziel:

$$\mathbb{P}\left[\, |\langle x, y \rangle| < \frac{c}{\sqrt{d-1}} \right) \geq 1 - \frac{2}{c} e^{-\frac{c^2}{2}} \xrightarrow{c \to \infty} 1$$

$\mathbb{R}^2$



$$\cos(\alpha) = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$$



⇝ vgl. Anhang B