BIG Data –

eine Einführung in „Compressive Sensing“

Ehrhard Behrends, FU Berlin, WS 2015/16

Einleitung

Die Theorie des „Big Data“ ist vergleichsweise jung, es gibt verschiedene Teilgebiete:

- Wie kann man in der Statistik in einem großen Datensatz Korrelationen erkennen?
- Mit welchen Methoden kann man das Rauschen bei Bild- oder Tonsignalen wegrechnen?
- Wie komprimiert man Bilder oder Videos?
- ...

„Compressive Sensing“ wird seit etwa 2005 intensiv betrieben. Um die Grundidee zu verstehen, denken wir uns ein Signal $x = (x_i)_{i=1,\dots,N}$ mit einem „sehr großen N“. Das steht uns nicht direkt zur Verfügung, man kennt nur den Vektor $y = Bx$, wobei B eine Matrix (oft eine $N \times N$-Matrix)ist.

- Klassische Verfahren gehen so vor: Man berechnet alle Komponenten von x aus Bx, danach werden die unwesentlichen Komponenten (also die mit $|x_i| \leq \varepsilon$) weggelassen.
- Beim Compressive Sensing versucht man, statt Bx den Vektor $y = Ax$ zu messen, wobei A eine $m \times N$-Matrix mit „kleinem“ m ist. Das bedeutet, dass an x nur vergleichsweise wenige Messungen vorgenommen werden. Dann kann man aus dem Gleichungssystem $y = Ax$ das x natürlich nicht eindeutig rekonstruieren. Wenn man allerdings die Zusatzinformation hat, dass x *schwach besetzt* ist (also nur wenige von Null verschiedene Komponenten hat), könnte es doch klappen.

  Kurz: Man spart sich den Umweg, zunächst das gesamte x zu identifizieren und danach doch nur wenige Komponenten weiter zu verwenden.

Eine oft zitierte und wirklich spektakuläre Anwendung dieser Idee gibt es in der Tomographie. Die Verweildauer im Tomographen kann für manche Anwendungen drastisch reduziert werden, und das hat eventuell lebensrettende Konsequenzen. (Etwa wenn ein Kind für eine Gehirntomographie durch Medikamente absolut ruhiggestellt werden muss.)

Die erste Erkenntnis: Naive Verfahren (alle schwach besetzten Vektoren ausprobieren) führen nicht zum Ziel, denn es würde viel zu lange dauern. Deswegen muss man sich etwas Neues einfallen lassen, und einige der Möglichkeiten, das Problem anzugehen, sollen in der Vorlesung vorgestellt werden. Die verwendeten mathematischen Methoden sind sehr vielfältig: Lineare Algebra (natürlich!), Optimierungstheorie, Hilbertraummethoden und – überraschender Weise – Stochastik.

Wir werden uns dem Thema eher theoretisch nähern. Konkrete Anwendungen und Computerimplementierungen sind nicht geplant.

Ehrhard Behrends, im Sommer 2015

Literatur

Die Theorie nahm ihren Anfang durch die Arbeiten von Candès et al. und Donoho. Diese Vorlesung orientiert sich stark an dem sehr empfehlenswerten Buch von Foucart und Rauhut, dort ist auch eine sehr ausführliche Literaturliste zu finden.

E.J. Candès, J. Romberg, T. Tao: Robust Uncertainty principles. IEEE 52, 2006, 489-509.

D.L. Donoho: Compressed Sensing. IEEE 52, 2006, 1289-1306.

M. Elad: "Sparse and Redundant Representations". Springer 2010.

S. Foucart und H. Rauhut: "A Mathematical Introduction to Compressive Sensing". Birkhäuser 2013.

Für die grundlegenden Ergebnisse aus der elementaren Stochstik beziehe ich mich auf mein Buch

[Be] E. Behrends: "Elementare Stochastik", Springer Spektrum 2012.

Inhaltsverzeichnis

Kapitel 1

Das Problem

1.1 Beispiele zur Einstimmung

Beispiel 1: Single-Pixel-Kamera. Die besteht aus $N_1 \times N_2$ Minispiegeln, die einzeln angesteuert werden können: offen oder geschlossen. Es spiegelt sich in dem Array ein Bild, das ist durch $N_1 N_2 = N$ Graustufen, also durch einen Vektor $b = (b_j)_{j=1,\dots,N}$ charakterisiert.

Wenn man gewisse Spiegel des Microarrays schließt (etwa die mit $j \notin \Delta$) und das gespiegelte Bild auf einen Sensor lenkt, so misst der $\sum_{j \in \Delta} b_i$. Das mache man für verschiedene Δ, die Intensitäten bilden dann einen Vektor y, der als Bb mit einer geeigneten Matrix B aufgefasst werden kann (in den Zeilen stehen die charakteristischen Funktionen der Δ).

Wenn man dann aus Erfahrung weiß, dass das Bild durch wenige Größen charakterisiert ist, dass also $b = Wx$ für eine geeignete Matrix W und einen schwach besetzten Vektor x, so heißt das, dass man x aus $y = (BW)x$ rekonstruieren möchte.

Beispiel 2: Bildkomprimierung. Da geht man ganz ähnlich vor.

Beispiel 3: Radar. Wie ermittelt man aus dem reflektierten Radarsignal die Position der Flugzeuge im Erfassungsbereich?

Beispiel 4: Sampling-Theorie. Ein aus wenigen Grundfrequenzen zusammengesetztes Signal sei gegeben. Wie kann man die identifizieren?

Beispiel 5: Approximation mit wenigen Vektoren. Gegeben seien Vektoren y und $a_1, \dots, a_N \in \mathbb{R}^m$. y soll durch eine Linearkombination der a_j so approximiert werden, dass wenige a_j dabei auftreten und die Approximation möglichst gut ist. Das heißt gerade, dass y „sehr gut“ durch Ax mit einem schwach besetzten Vektor angenähert werden soll, wobei A die Matrix mit Spalten a_j ist.

1.2 Unsere Bezeichnungsweisen

Wir wollen $y = Ax$ mit einem schwach besetzten x lösen. Stets wird dabei A eine $m \times N$-Matrix sein: Meist über $\mathbb{R}$, aber so gut wie alle Ergebnisse lassen sich auf komplexe Zahlen übertragen. Mit $a_1, \ldots, a_N \in \mathbb{R}^m$ werden die Spalten von A bezeichnet.

Für Vektoren $x \in \mathbb{R}^n$ wird die übliche l^p-Norm $||\cdot||_p$ auftreten ($1 \leq p \leq \infty$):

$$||x||_p := \big(\sum_j |x_i|^p\big)^{1/p};\ ||x||_\infty := \max_i |x_i|.$$

Wir nutzen aber die gleiche Bezeichnungsweise auch für beliebige $p > 0$ (obwohl dann keine Norm mehr vorliegt).

Mit $||x||_0$ bezeichnen wir die Anzahl der von Null verschiedenen Komponenten von x. Ein Vektor x ist also schwach besetzt, wenn $||x||_0$ „klein“ ist; was „klein“ ist, kann in verschiedenen Situationen etwas anderes bedeuten. Auch $||\cdot||_0$ ist keine Norm, die Bezeichnung ist durch $\lim_{p\to 0} ||x||_p^p = ||x||_0$ motiviert.

Ist $x \in \mathbb{R}^N$, so ist der *Träger* von x die Menge aller i mit $x_i \neq 0$.

Für eine Teilmenge S von $\{1, \ldots, N\}$ ist $|S|$ die Anzahl der Elemente in S. Meist wird diese Zahl mit s bezeichnet. Unter $\overline{S}$ wollen wir die Menge $\{1, \ldots, N\} \setminus S$ verstehen.

A_S ist die $m \times s$-Matrix, die aus den Spalten a_i mit $i \in S$ gebildet wird.

1.3 Ein naiver Versuch

Die Versuchung ist groß, sich das Leben einfach zu machen. Mal angenommen, A ist eine 100×1000-Matrix, und man weiß, dass $y = Ax$ durch ein x mit $||x||_0 \leq 30$ gelöst wird. Dann muss man doch „nur“ systematisch alle Familien von je 30 Spalten zu einer Matrix 100×30-Matrix A' zusammenfassen und nachprüfen, ob $y = A'x'$ durch ein $x' \in \mathbb{R}^{30}$ lösbar ist.

Leider ist das völlig illusorisch. Es gibt $\binom{1000}{30}$ solche Familien, so viele wären also zu lösen. Die genaue Anzahl ist

$$2429608192173745103270389838576750719302222606198631438800 \approx 2.43{\cdot}10^{57}.$$

Auch wenn 10.000 Einzelprobleme pro Sekunde behandelt werden könnten, müsste man immer noch eine Zeit von 7 mal 10^{45} Jahren einkalkulieren. *Das ist unrealistisch!*

Aber geht es nicht geschickter? Nein! Das zeigen wir im nächsten Abschnitt.

1.4 Das Problem ist NP-schwierig

In den Siebzigern des vorigen Jahrhunderts wurde erstmals quantifiziert, was man unter einem „schwierigen“ Problem verstehen möchte. Die nachstehenden Definitionen haben sich bewährt.

„Einfache“ Probleme: Probleme vom Typ P

Jedes Grundschulkind weiß, dass Addieren und Multiplizieren noch recht einfach sind, dass aber Addieren leichter ist als Multiplizieren. Beim Addieren n-stelliger Zahlen muss man nämlich im Wesentlichen n Rechnungen durchführen, bei der Multiplikaton ist die Größenordnung n^2. Präzisiert wird diese Beobachtung in der

Definition 1.4.1. *Sei V ein Verfahren, dass aus gewissen Eingabegrößen gewisse Ergebnisse produziert. V heißt vom* polynomiellen Typ *oder auch* vom Typ P*, wenn es ein Polynom P so gibt, dass die Anzahl der Rechenschritte bei einer Eingabe von n Ziffern durch $P(n)$ nach oben abgeschätzt werden kann*[1)]*.*

Addition und Multiplikation sind also P-Probleme, der Grad der hier relevanten Polynome ist 1 bzw. 2. P-Probleme gelten als „einfach“, fast alle Verfahren der Schulmathematik können als Beispiele herangezogen werden: Wurzelziehen, Determinanten berechnen, Gleichungssysteme lösen usw.

Ob ein P-Problem immer wirklich „einfach“ ist, kann allerdings bezweifelt werden. Wenn zum Beispiel bei einem Algorithmus für das Auffinden von Primzahlen für die Abschätzung ein Polynom 10-ten Grades benötigt wird, so heißt das, dass man bei einer Anwendung für 1000-stellige Zahlen $1000^{10} = 10^{30}$ Rechenschritte einkalkulieren muss

Probleme vom Typ NP

Die Umkehrung ist aber sicher richtig: Probleme, die nach heutigem Kenntnisstand nicht zur Klasse P gehören, können als schwierig betrachtet werden. (Allerdings wird sich sicher in der Zukunft in manchen Fällen zeigen, dass es doch Algorithmen mit polynomieller Laufzeit gibt.) Es kann bei „schwierigen“ Problemen allerdings vorkommen, dass man schnell feststellen kann, ob man eine Lösung vor sich hat, das sind die *NP-Probleme*. Genauer: Ein Problem gehört zur Klasse NP (= nichtdeterministisch polynomiell), wenn man für ein x in polynomieller Laufzeit (polynomiell in der Eingabegröße) feststellen kann, ob x eine Lösung des Problems ist.

Als Beispiel betrachten wir die Aufgabe, aus dem Produkt $n = pq$ von zwei großen Primzzahlen die Faktoren zu ermitteln. (Das ist ein Problem, auf dem die Sicherheit mancher Kryptographieverfahren beruht.) Es ist zurzeit kein Verfahren bekannt, p und q in polynomieller Laufzeit zu ermitteln. Doch wenn mir jemand ein p vorschlägt, kann ich schnell entscheiden, ob p wirklich ein Faktor von n ist.

Es kann durchaus als Skandal der Mathematikgeschichte bezeichnet werden, dass man noch nicht nachweisen konnte, dass die Klassen P und NP wirklich verschieden sind. Niemand rechnet damit, ein Beweis steht aber noch aus. (Seit dem Jahr 2000 sind von der Clay-Foundation 1.000.000 Dollar für die Lösung der Frage ausgesetzt, ob nun $P = NP$ gilt oder nicht.)

[1)] Wenn man es ganz genau machen möchte, müsste man noch präzisieren, in welchem Zahlensystem die Eingabegrößen dargestellt werden: Dualzahlen, Dezimalzahlen, ...

NP-schwere Probleme

Ein Problem P_s wird *NP-schwer* genannt, wenn gilt: Ist P ein NP-Problem, so kann man eine Lösung von P in polynomieller Zeit finden, falls man eine Lösung von P_s zur Verfügung hat. (P_s muss dabei kein NP-Problem sein.)

NP-vollständige Probleme

Ist P gleichzeitig in der Klasse NP und NP-schwer, so heißt P *NP-vollständig*. Diese Problem sind also irgendwie „gleich schwierig“: Wenn man das eine lösen kann, so auch alle anderen. Das erste Beispiel stammt von Stephan A. Cook (1971). Es ist das so genannte 3SAT-Problem: Kann ein gewisser logischer Ausdruck durch geeignete Belegung der Variablen mit den Wahrheitswerten W und F den Wahrheitswert W annehmen? Inzwischen gibt es einen ganzen Zoo von NP-vollständigen Problemn, insbesondere aus der Graphentheorie und der Kombinatorik. (Ein Beispiel: Hat ein vorgelegter Graph einen Hamiltonschen Kreis[2)] ?)

Wir bemerken: Findet man für ein Problem P' ein NP-vollständiges Problem P, so dass mit einer Lösung von P' in polynomieller Zeit eine Lösung von P gefunden werden kann, so muss P' NP-schwer sein. Diese Beobachtung wird gleich wichtig werden:

Satz 1.4.2. *Das Problem P', ein x mit $Ax = y$ und minimalem $||x||_0$ zu finden, ist NP-schwer.*

Beweis: Wir werden zeigen, dass das 3-Mengen-Problem auf P' zurückgeführt werden kann:

> Das 3-*Mengen-Problem*: Gegeben seien 3-elementige Teilmengen C_i einer m-elementigen Menge, $i = 1, \ldots, N$. Kann man eine Teilfamilie $C_{i_1}, \ldots, C_{i_r}$ so auswählen, dass $\{1, \ldots, m\}$ die disjunkte Vereinigung der $C_{i_1}, \ldots, C_{i_r}$ ist?

Sicher kann das nur dann gehen, wenn m durch 3 teilbar ist, aber wie kann man es in diesem Fall entscheiden? Ohne Beweis verwenden wir, dass das 3-Mengen-Problem NP-vollständig ist. wenn wir es also auf die $||\cdot||_0$-Minimierung zurückführen können, ist der Satz bewiesen.

Die $C_i \subset \{1, \ldots, m\}$ seien für $i = 1, \ldots, N$ vorgelegt, und wir nehmen an, dass m durch 3 teilbar ist. Wir konstruieren damit eine $m \times N$-Matrix A: Die i-te Spalte enthält die charakteristische Funktion von C_i. wir setzen noch $y := (1, \ldots, 1)^\top \in \mathbb{R}^n$ und betrachten die Gleichung $y = Ax$.

Angenommen, wir sind in der Lage, ein x mit $Ax = y$ zu finden, so dass $||x||_0$ minimal ist. Sicher muss dann $||x||_0 \geq m/3$ sein, denn jede nichttriviale Komponente von x erzeugt höchstens 3 Einsen.

[2)] Das ist ein geschlossener Weg, der jeden Knoten genau einmal trifft

Fall 1: $||x||_0 = m/3$. Durch die nichttrivialen Komponenten von x sind dann diejenigen Indizes ausgewählt, die zu einer disjunkten Überdeckung von $\{1, \ldots, m\}$ durch die C_i führen. Die Antwort für das 3-Mengen-Problem ist also „JA“.

Fall 2: $||x||_0 > m/3$. Die Antwort für das 3-Mengen-Problem muss jetzt „NEIN“ sein, denn im Fall „JA“ könnten wir ja leicht ein x mit $||x||_0 = m/3$ finden. □

Das Fazit: Es sind keine Verfahren zu erwarten, mit denen unser compressive-sensing-Problem in allen Fällen in angemessener Rechenzeit gelöst werden könnte.

Kapitel 2

Zwei Lösungsstrategien

In diesem Kapitel beschreiben wir zwei Strategien, mit denen man hoffen kann, die richtige schwach besetzte Lösung zu finden. Bei der ersten wird das Problem darauf zurückgeführt, eine konvexe Funktion über einem konvexen Definitionsbereich zu minimieren. Dafür gibt es gut bewährte Lösungsverfahren. Die zweite Strategie ist „gierig“ („greedy“). Nach und nach konstruiert man Vektoren mit höchstens ein, zwei, usw. von Null verschiedenen Komponenten, für die das Bild unter A möglichst nahe bei y ist. Mit etwas Glück wird nach „wenigen“ Schritten das y exakt getroffen.

In beiden Fällen gibt es eine ausführliche Motiviation, man kann das Verfahren „sehen“, und dabei werden auch jeweils gleich die Grenzen deutlich.

2.1 l^1-Minimierung

Als Vorbereitung sollte man sich die elementare Tatsache klarmachen, dass

$$||x||_0 = \lim_{p \to 0} ||x||_p^p$$

gilt. Ein x mit möglichst kleinem $||x||_0$ ist als fast das gleiche wie ein x mit möglichst kleinem $||x||_p$ für ein $p \approx 0$.

Das motiviert, warum man sich ein Bild der Einheitskugel der p-Norm für „kleine“ p machen sollte. Gleichzeitig stellen wir uns die Einheitskugel der l^1-Norm vor.

Wir kombinieren die folgenden Beobachtungen:

- Wir suchen doch ein x, das erstens die Gleichung $Ax = y$ löst (also in einem speziellen affinen Unterraum U liegt) und dass zweitens möglichst wenige von Null verschiedene Komponenten hat: Im Idealfall liegt x auf einer Koordinatenachse ($||x||_0 = 1$), evtl. auch auf einer von zwei Achsen aufgespannten Ebene usw.

- Mal angenommen, wir suchen ein $x \in U$ mit minimaler l^1-Norm. Dazu müssen wir die l^1-Einheitskugel so weit schrumpfen oder aufblasen, dass sie gerade noch U berührt. Alle x im Schnitt von U mit der geschrumpften/gedehnten Kugel lösen dann das Minimierungsproblem.
- Und man stellt fest: In vielen Fällen ergeben sich die gleichen Vektoren x bei beiden Problemen.

Leider können wir uns nur dreidimensionale Bilder vorstellen, dafür wollen wir gleich den Zusammenhang genauer analysieren.

Zuerst einigen wir uns aber auf eine *Fairnessbedingung*. Zur Motivation untersuchen wir, was passiert, wenn wir die Spalten von A mit Zahlen $\lambda_i \neq 0$ multiplizieren: Wird aus Spalte $a^{(i)}$ der Vektor $\lambda_i a^{(i)}$, so sind statt $x = (x_i)$ nun $(x_i/\lambda_i$ die Lösungen. die Norm $||\cdot||_0$ ändert sich dabei nicht, die l^1-Norm wird aber völlig unterschiedlich sein, so dass sich ganz andere Lösungen des Minimierungsproblems ergeben können.

Um diese Mehrdeutigkeit zu vermeiden, vereinbaren wir ab sofort, dass $||a^{(i)}||_2 = 1$ für alle i gelten soll. Es folgt die Analyse:

1. Wir sind im $\mathbb{R}^3$, und $U = \{x \mid Ax = y\}$ ist zweidimensional. Dann ist A eine 1×3-Matrix $(a_{11}a_{12}a_{13})$, wobei wegen unserer Vereinbarung alle a_{1i} Betrag 1 haben. U ist dann die Hyperebene $\{x \mid x_1a_{11} + x_2a_{12} + x_3a_{13} = \alpha\}$ für irgendein $\alpha \neq 0$.
Man sieht: Es gibt drei Lösungen x mit $||x||_0 = 1$ und unendlich viele mit $||x||_0 = 2$ und $||x||_0 = 3$, und alle haben minimale l^1-Norm. (Etwas ganz anderes hätte sich ergeben, wenn wir die Spalten von A anders skaliert hätten[1)].)

1'. Wir sind im $\mathbb{R}^N$, und $U = \{x \mid Ax = y\}$ ist (N-1)-dimensional. Das geht ganz analog: Es gibt nun N Lösungen x von $x_1a_{11} + \cdots + x_Na_{1N} = \alpha$ mit $||x||_0 = 1$.

2. Wir sind im $\mathbb{R}^3$, und $U = \{x \mid Ax = y\}$ ist eindimensional. Wir können das Problem uminterpretieren. Gegeben sind drei auf Eins normierte Vektoren $B, C, D \in \mathbb{R}^2$ (die Spalten von A) und ein $Y \in \mathbb{R}^2 \setminus \{0\}$, und wir suchen alle $(x_1, x_2, x_3) \in \mathbb{R}^3$ so dass die Linearkombination $x_1B + x_2C + x_3D$ gleich Y ist. Es soll – um Trivialfälle zu vermeiden – angenommen werden, dass B, C, D paarweise linear unabhängig sind.
Man sieht: Lösungen mit $||x||_0 = 1$ muss es nicht geben, wohl aber drei x mit $||x||_0 = 2$. Wir behaupten:

Lemma 2.1.1. *(i) In der vorstehenden Situation gebe es eine Lösung x mit $||x||_0 = 1$. Dieses x kann durch l^1-Minimierung gefunden werden.*
(ii) Gibt es kein x mit $||x||_0 = 1$, so gibt es eins mit $||x||_0 = 2$. Auch das kann durch l^1-Minimierung ermittelt werden.

Beweis: (i) O.E. sei $||Y||_2 = 1$. Dann bedeutet die Voraussetzung, dass Y in $\{\pm B, \pm C, \pm D\}$ liegt. Sei etwa $Y = B$, es ist also $x = (1, 0, 0)^\top$ mit $||x||_1 = 1$.

[1)] Evtl. gäbe es dann nur ein x mit $||x||_0 = 1$, das $||\cdot||_1$ minimiert.

Wir behaupten, dass $||x'||_1 > 1$ sein muss, wenn $Y = x'_1B + x'_2C + x'_3D$ und $x' \neq x$. Aufgrund er vorausgesetzten paarweisen linearen Unabhängigkeit hat x' mindestens zwei nichttriviale Komponenten, und da $||\cdot||_2$ strikt konvex ist, folgt

$$1 = ||Y||_2 = ||x'_1B + x'_2C + x'_3D||_2 < |x'_1| + |x'_2| + |x'_3| = ||x'||_1.$$

(ii) Sei etwa das Minimum der l^1-Normen der x mit $Y = x_1B+x_2C+x_3D$ gleich 1. Das bedeutet, dass Y auf dem Rand der konvexen Hülle von $\{\pm B, \pm C, \pm D\}$ liegt, aber – aufgrund der Voraussetzung – kein Extremalpunkt ist. Es gibt also zwei eindeutig bestimmte Elemente in $\{\pm B, \pm C, \pm D\}$, so dass Y auf der Verbindungsgeraden liegt. Das aber zeigt, dass $||x||_0 = 2$ gilt. □

2'. Wir sind im $\mathbb{R}^N$, und $U = \{x \mid Ax = y\}$ ist $(N-2)$-dimensional. Das zeigt man ganz analog.

Wir fassen zusammen

> Ist A eine $(1 \times N)$- oder eine $(2 \times N)$-Matrix mit l^2-normierten Spalten, so können – falls existent – schwach besetzte Löungen von $Ax = y$ mit l^1-Minimierung gefunden werden.

3. Wir sind im $\mathbb{R}^4$, und $U = \{x \mid Ax = y\}$ ist 3-dimensional. (Das ist der erste Fall, der durch die bisherigen Überlegungen nicht abgedeckt ist.) Leider gilt dann

Lemma 2.1.2. *Es gibt eine (3×4)-Matrix mit l^2-normierten Spalten und ein $y \in \mathbb{R}^3$, so dass gilt:*

(i) Man findet ein $x \in \mathbb{R}^4$ mit $Ax = y$ und $||x||_0 = 2$.

(ii) Dieses x hat unter den z mit $Az = y$ nicht die minimale l^1-Norm.

Beweis. Die Spalten von A – wir bezeichnen sie mit $a^{(i)}$ – sind die Einheitsvektoren des $\mathbb{R}^3$ sowie der Vektor $(-\sqrt{2}/\sqrt{3}, 1/\sqrt{6}, 1/\sqrt{6})^\top$, und y ist durch $(1/3, 1/3, 1/3)^\top$ definiert.

a) Es ist y der Mittelwert der ersten drei Spalten, d.h. für $z = (1/3, 1/3, 1/3, 0)^\top$ ist $Az = y$ und $||z||_1 = 1$. Das l^1-Minimum der Lösungen ist also ≤ 1. Es ist $||z||_0 = 3$.

b) Es gibt ein x mit $Ax = y$ und $||x||_0 = 2$, es ist nämlich $y = a^{(1)} + \sqrt{2/3}a^{(4)}$. Dabei ist $||x||_1 = ||(1, 0, 0, \sqrt{2/3})||_1 = 1 + \sqrt{2/3} > 1$.

c) Außer diesem x gibt es keine Lösung z mit $||z||_0 = 2$. Dazu muss man nur nachprüfen, dass y nicht in den linearen Hüllen der Vektoren $a^{(2)}, a^{(4)}$ bzw. $a^{(3)}, a^{(4)}$ bzw. $a^{(1)}, a^{(2)}$ bzw. $a^{(1)}, a^{(3)}$ bzw. $a^{(2)}, a^{(3)}$. liegt. □

(Hätte man nicht bzgl. $||\cdot||_1$ sondern bzgl. $||\cdot||_p$ für ein „genügend kleines“ p minimiert, wäre das richtige x sehr wohl erkannt worden. Das liegt daran, dass $\lim_{p\to 0} ||x||_p^p = 2 < 3 = \lim_{p\to 0} ||z||_p^p$ für alle z mit $||z||_0 = 3$.

Wir fassen zusammen:

Aufgrund der Bilder im $\mathbb{R}^3$ kann man hoffen, Vektoren x mit $Ax = y$ und minimalem $||x||_0$ in manchen Fällen dadurch zu finden, dass man ein $x \in U$ mit minimaler l^1-Norm findet. Wir wissen allerdings schon, dass das ohne Zusatzvoraussetzungen nicht klappen muss.

Oder etwas formaler:

Definition 2.1.3. *Bei der Strategie* l^1-Minimierung *sucht man ein* x *mit* $y = Ax$*, so dass* $||x||_1$ *so klein wie möglich ist. In der Theorie der Optimierung sind Methoden entwickelt worden, solche* x *mit vertretbarem Aufwand zu finden.*

Der Vollständigkeit halber ist noch anzumerken, dass Minimierung der l^p-Norm für $p > 1$ bei dem hier zu behandelnden Problem zu keinen vernünftigen Ergebnissen führt[2)].

2.2 Ein „gieriges“ Verfahren

Zur Motivation des Verfahrens, das wir nun beschreiben wollen, betrachten wir die folgende Situation:

Im $\mathbb{R}^m$ seien ein Unterraum U und ein Vektor y gegeben. y werde durch einen Vektor $y_0 \in U$ optimal aus U heraus approximiert:

$$||y - y_0||_2 = \min_{z \in U} ||y - z||_2.$$

Die Approximation an y soll verbessert werden. Ein Einheitsvektor a wird vorgelegt, und wir wollen nun von U_a, der linearen Hülle von U und a, aus approximieren.

Welche Eigenschaften von a sind günstig für dieses Ziel?

Man beachte dabei: Wir messen den Fehler in der l^2-Norm. Der Hauptgrund: Für diese Norm stehen Hilbertraum-Methoden zur Verfügung.

Lemma 2.2.1. *Es gibt einen Vektor* $z_0 \in U_a$ *mit*

$$||y - z_0||_2^2 \leq ||y - y_0||_2^2 - (\langle y - y_0, a\rangle)^2.$$

Beweis. Alle $y_0 + ta$ mit $t \in \mathbb{R}$ gehören zu U_a. Wir werden dasjenige t wählen, für das die Approximation am besten ist. Das ist das übliche Approximationsproblem, diesmal soll $\tilde{y} := y - y_0$ optimal durch einen Vektor aus dem eindimensionalen Raum $\mathbb{R}\, a$ approximiert werden. Man muss t so wählen, dass $\tilde{y} - ta$ senkrecht auf a steht, d.h.

$$0 = \langle \tilde{y} - ta, a\rangle = \langle \tilde{y}, a\rangle - t\langle a, a\rangle = \langle \tilde{y}, a\rangle - t.$$

[2)] Das ist schade, denn im Fall $p = 2$ gibt es besonders viele Lösungsverfahren.

Für dieses t ist dann der (quadrierte) Abstand von y zu $z_0 := y_0 + ta$ gleich

$$\begin{aligned} ||y - z_0||_2^2 &= ||\tilde{y} - ta||_2^2 \\ &= \langle \tilde{y} - ta, \tilde{y} - ta \rangle \\ &= ||\tilde{y}||_2^2 - 2t\langle \tilde{y}, a \rangle + (\langle \tilde{y}, a \rangle)^2 \\ &= ||\tilde{y}||_2^2 - (\langle \tilde{y}, a \rangle)^2 \\ &= ||y - y_0||_2^2 - (\langle y - y_0, a \rangle)^2. \end{aligned}$$

Es folgt die Behauptung. □

Wenn mehrere a zur Auswahl stehen, könnte eine naheliegende Strategie zur Approximationsverbesserung also darin bestehen, ein a auszuwählen, für das $|\langle y - y_0, a \rangle|$ möglichst groß ist.

Man sollte sich klar machen, wo die Schwäche dieses Arguments liegt: Im Grunde haben wir das a dann nicht so gewählt, dass die Approximation an U_a bestmöglich ist, sondern nur so, dass $y_0 + \mathbb{R}\, a$ dem y möglichst nahe kommt. Deswegen nennt man das Verfahren auch „gierig“ („greedy“).

Die gierige Wahl muss nicht die beste sein:

Lemma 2.2.2. *Es gibt einen Unterraum $U \subset \mathbb{R}^3$ und normierte Vektoren $y, a, b \in \mathbb{R}^3$, so dass gilt:*

(i) $|\langle y, a \rangle| > |\langle y, b \rangle|$.

(ii) y wird durch U_b besser approximiert als durch U_a.

Die gierige Strategie hätte hier also nicht zum optimalen Ergebnis geführt.

Beweis: Wir setzen $U := \mathbb{R}(1,0,0)^\top$, $y := (0,0,1)^\top$, $a := (0, 1/\sqrt{2}, 1/\sqrt{2})^\top$, $b := (0.9, 0, \sqrt{0.19})^\top$.

a) Es ist wirklich

$$0.707... = |\langle y, a \rangle| > |\langle y, b \rangle| = 0.435...$$

Es liegt also nahe, a zur Approximation vorzuziehen. Da y nicht in U_a liegt, ist die beste Approximation größer als Null.

b) y liegt in U_b, man hätte also exakt approximieren können! □

Nach diesen Vorbereitungen beschreiben wir die „gierige“ Strategie in der für das Compressive Sensing relevanten Situation:

Definition 2.2.3. *Es soll das Problem $y = Ax$ mit einem möglichst schwach besetzten x gelöst werden. Dazu approximiert man y durch eine Folge von Vektoren der Form Ax, wobei x nach und nach immer mehr von Null verschiedene Komponenten hat. Die Hoffnung: Man kann nach wenigen Schritten exakt approximieren.*

Genauer sieht das Verfahren so aus. Wir bezeichen mit $a_j \in \mathbb{R}^m, j = 1, \ldots, N$, *die Spalten von* A *und setzen voraus, dass die alle* l^2*-normiert sind. Gegeben ist ein* $y \in \mathbb{R}^m$, *und es soll* y *durch* Ax *mit einem schwach besetzten* x *dargestellt oder wenigstens gut approximiert werden. Bei der gierigen Strategie geht man so vor. Man konstruiert für* $k = 0, 1, 2, \ldots$ *Mengen* $S_k \subset \{1, \ldots, N\}$ *und einen Vektor* $x^{(k)}$ *so, dass der Träger in* S_k *liegt und* $Ax^{(k)}$ *dem* y *möglichst nahe kommt); der Fehler, also der Vektor* $y - Ax^{(k)}$, *heiße* $y^{(k)}$. *Das Verfahren soll abbrechen, sobald* $y^{(k)} = 0$ *gilt oder wenigstens* $||y^{(k)}||_2$ *genügend klein ist. Genauer:*

- *Start:* $S_0 = \emptyset$, $x^{(0)} = 0$ *und* $y^{(0)} := y$.
- *Angenommen, bis zum* k*-ten Schritt sei alles schon konstruiert.*
- *Suche unter den* a_j *mit* $j \notin S_k$ *ein* j' *so, dass* $|\langle a_{j'}, y^{(k)} \rangle|$ *maximal ist. Dann sei* $S_{k+1} := S_k \cup \{j'\}$.

  $x^{(k+1)}$ *soll derjenige Vektor sein, der seinen Träger in* S_{k+1} *hat und für den* $Ax^{(k+1)}$ *den Vektor* y *optimal approximiert. (Anders ausgedrückt:* $Ax^{(k+1)}$ *ist die orthogonale Projektion von* y *auf die lineare Hülle der Menge* $\{a_j \mid j \in S^{k+1}\}$.*)*

  Setze noch $y^{(k+1)} := y - Ax^{(k+1)}$.
- *Setze das Verfahren so lange fort, bis* $y^{(k)} = 0$ *(bis also* $Ax^{(k)} = y$*) oder wenigstens* $||y^{(k)}|| \leq \varepsilon$ *für ein vorgegebenes* ε.

Diese Verfahren heißt Orthogonal Matching Pursuit (kurz OMP).

Aufgrund von Lemma 2.2.1 wissen wir dann, dass die Approximation von y durch $Ax^{(k)}$ in jedem Schritt verbessert wird, und zwar wird das Quadrat des Fehlers mindestens um $(\langle y^{(k)}, a_{j'} \rangle)^2$ kleiner.

Kapitel 3

Theoretisches zur l^1-Minimierung und zu OMP

In diesem Kapitel sollen diejenigen Matrizen charakterisiert werden, bei denen s-schwachbesetzte Vektoren durch l^1-Minimierung bzw. durch OMP gefunden werden können. Das wird die Grundlage dafür sein, geeignete Matrizen A zu finden.

3.1 Eindeutig bestimmte Lösungen

Wir wollen doch $y = Ax$ durch ein x mit $||x||_0 \leq s$ lösen. Angenommen, wir haben eins gefunden: Ist das eindeutig bestimmt? Dazu das

Lemma 3.1.1. *Die folgenden Ausssagen sind äquivalent:*

(i) Aus $Ax = Az = y$ und $||x||_0, ||z||_0 \leq s$ folgt $x = z$.

(ii) Sei x im Kern von A, und es gelte $||x||_0 \leq 2s$. Dann ist $x = 0$.

(iii) Je $2s$ Spalten von A sind linear unabhängig.

Beweis: Der Beweis ist sehr elementar. □

Aufgrund dieses Ergebnisses wäre es natürlich interessant zu wissen, wie man sich konkrete Beispiele solcher Matrizen verschaffen kann. Das ist recht leicht:

Ein Beispiel: Es seien $t_1, \ldots, t_N$ verschiedene reelle Zahlen, und eine $m \times N$-Matrix $A = (a_{ij})$ sei durch $a_{ij} := t_j^{i-1}$ definiert. Dann sind je m Spalten von A linear unabhängig, so dass man bei vorlegtem s nur mit $m = 2s$ arbeiten muss, um das Lemma anwenden zu können.

(*Die Begründung:* Wähle m Spalten aus, die zugehörigen t_i bezeichnen wir mit $s_1, \ldots, s_m$. Dann ist die Determinante der aus den entsprechenden Spalten gebildeten Matrix gerade die *Vandermonde-Matrix*, also gleich $\prod_{i<j}(s_i - s_j)$. Da die t_i nach Voraussetzung paarweise verschieden waren, ist die Determinante ungleich Null.)

3.2 Wann funktioniert l^1-Minimierung?

Wann kann l^1-Minimierung erfolgreich sein? Doch sicher genau dann, wenn gilt:

> Ein x mit $||x||_0 \leq s$ sei die eindeutig bestimmte Lösung von $Ax = y$.
> Ist dann x' eindeutig bestimmte Lösung von $Ax' = y, ||x'||_1 = \min$,
> so ist $x = x'$.

Wir haben schon gesehen, dass das i.A. nicht erfüllt sein muss. Hier ist die für dieses Problem passende

Definition 3.2.1. *(i) Man sagt, dass eine $m \times N$-Matrix A die* Kern-Separierungs-Eigenschaft für ein $S \subset \{1, \ldots, N\}$ *hat, wenn gilt: Ist $x \neq 0$ mit $Ax = 0$, so gilt $||x_S||_1 < ||x_{\overline{S}}||_1$.*
(ii) A hat die Kern-Separierungs-Eigenschaft für s*, wenn A die Kern-Separierungs-Eigenschaft für alle S mit höchstens s Elementen hat.*

Und dann gilt:

Satz 3.2.2. *Wir setzen voraus, dass s-schwachbesetzte Lösungen eindeutig bestimmt sind. Äquivalent sind:*
(i) Ein x mit $||x||_0 \leq s$ sei die eindeutig bestimmte Lösung von $Ax = y$. Ist dann x' eindeutig bestimmte Lösung von $Ax' = y, ||x'||_1 = \min$, so ist $x = x'$.
(ii) A hat die Kern-Separierungs-Eigenschaft für s.

Beweis: Zunächst setzen wir (i) voraus. Wir beginnen mit einem $x \neq 0$ mit $Ax = 0$ und einer Menge S mit höchstens s Elementen. Ziel: $||x_S||_1 < ||x_{\overline{S}}||_1$. Setze $y' := Ax_S$. Dann ist x_S die eindeutig bestimmte s-schwachbesetzte Lösung von $Ax = y'$, hat folglich minimale l^1-Norm unter aller Lösungen. Auch $x_S - x = -x_{\overline{S}}$ ist eine derartige Lösung, sie ist von x_S verschieden. So folgt (ii).
Nun sei (ii) erfüllt. Sei $Ax = y$, und x habe den Träger in einer Menge S mit höchstens s Elementen. Weiter sei auch $A\tilde{x} = y$, und es gelte $||\tilde{x}||_1 \leq ||x||_1$. Es ist $x = \tilde{x}$ zu zeigen. Setze $x^* := \tilde{x} - x$. Das ist ein Element im Kern von A, wir nehmen an, dass es von Null verschieden ist. Nach Voraussetzung ist dann $||x^*_S||_1 < ||x^*_{\overline{S}}||_1$. So folgt der Widerspruch

$$\begin{aligned} ||x||_1 &= ||x + x^*_S - x^*_S||_1 \\ &\leq ||x + x^*_S||_1 + ||x^*_S||_1 \\ &< ||x + x^*_S||_1 + ||x^*_{\overline{S}}||_1 \\ &= ||\tilde{x}_S||_1 + ||\tilde{x}_{\overline{S}}||_1 \\ &= ||\tilde{x}||_1. \end{aligned}$$

Hier haben wir $x_{\overline{S}} = 0$ und $x^*_{\overline{S}} = \tilde{x}^*_{\overline{S}}$ ausgenutzt. □

3.3 Wann funktioniert OMP?

Sei $S \subset \{1, \ldots, N\}$. Ist dann $x \in \mathbb{R}^N$ mit Träger in S, so ist $Ax = \sum_{j \in S} x_i a_i$ eine Linearkombination der a_j, $j \in S$; wie vorher sind dabei die a_j die Spalten von A, die o.B.d.A. normalisiert sind.

Das hat eine wichtige Konsequenz. Ist x' mit Träger in S so, dass Ax' einen Vektor y optimal approximiert, so steht $y - Ax'$ auf allen a_j mit $j \in S$ senkrecht. (Das gilt übrigens auch umgekehrt: Ist $y - Ax' \perp a_j$ für alle $j \in S$, so ist Ax' die beste Approximation an y aus der linearen Hülle der a_j, $j \in S$.)

Wann genau klappt das OMP-Verfahren? Zur Vorbereitung betrachen wir einen Vektor $v = \sum_{j \in S} x_j a_j$. Dann ist

$$||v||^2 = \langle v, v \rangle = \sum_{j \in S} \langle v, a_j \rangle x_j,$$

insbesondere müssen im Fall $v \neq 0$ gewisse $\langle v, a_j \rangle$ mit $j \in S$ ungleich Null sein. Das Kriterium besagt, dass die $\langle v, a_j \rangle$ mit $j \in \overline{S}$ keine betragsmäßig größeren Werte liefern dürfen:

Satz 3.3.1. *Mit den vorstehenden Bezeichnungsweisen sind äquivalent:*
(i) OMP ist für alle x mit Träger in S nach höchstens $s = |S|$ Schritten erfolgreich: Ist der Träger von x in S und ist $y = Ax$, so ist $x^{(k)} = x$ (und damit insbesondere $y^{(k)} = 0$) für ein $k \leq s$.
(ii) Aus $\sum_{j \in S} x_j a_j = 0$ folgt $x_j = 0$ für alle j, und es gilt

$$\max_{j \in S} |\langle v, a_j \rangle| > \max_{j \in \overline{S}} |\langle v, a_j \rangle|$$

für alle $v = \sum_{j \in S} x_j a_j \neq 0$.

Beweis: Wir setzen zunächst (i) voraus. Es sei $\sum_{j \in S} x_j a_j = 0$. Wir füllen die x_j nurch Nullen zu einem Vektor $x \in \mathbb{R}^N$ auf, dann ist $Ax = 0$. Schon im ersten Schritt meldet OMP „fertig!“, wenn auf das Problem $Ax = 0$ angewendet, und zwar bei $x^{(0)} = 0$. Und da Vektoren nach Voraussetzung rekonstruiert werden können, heißt das $x = 0$. Das beweist die erste Bedingung in (ii).

Nun zur zweiten, wir geben ein $v = \sum_{j \in S} x_j a_j \neq 0$ vor. Wir wollen aus $y := v = Ax$ (x entsteht aus den x_j durch Nullen-Ergänzung) das x rekonstruieren. Das werden wir nicht schaffen, wenn schon im ersten Schritt ein j' mit $j' \in \overline{S}$ gewählt werden würde. Damit das nicht passieren kann, muss $\max_{j \in S} |\langle v, a_j \rangle| > \max_{j \in \overline{S}} |\langle v, a_j \rangle|$ gelten.
Nun gelte (ii). Wir starten mit einem x mit Träger in S und setzen $y := Ax$. Aufgrund der Bedingung (ii) wird bei allen OMP-Schritten der fragliche Index j' in S liegen. j' kann aufgrund der Vorbemerkung auch nicht in S_k liegen, denn $y^{(k)} \perp a_j = 0$ für $j \in S_k$. Es kommt also in jedem Schritt ein weiteres Element

aus S dazu, nach höchstens s Schritten sind alle $a_j, j \in S$ im Spiel. Dann ist y natürlich exakt zu approximieren, nach spätestens s Schritten ist also $y^{(k)} = 0$. Wegen der ersten Bedingung in (ii) wurde auch das richtige x gefunden. □

3.4 Pseudoinverse und Matrixnormen

Um den Zusammenhang zwischen den Bedingungen aus Abschnitt 3.1 und 3.2 analysieren zu können, brauchen wir einige Vorbereitungen.

Pseudoinverse

C sei eine $k \times l$-Matrix, wir identifizieren C mit einer linearen Abbildung von $\mathbb{R}^l$ nach $\mathbb{R}^k$. Es soll $l \leq k$ sein, und es soll vorausgesetzt werden, dass C vollen Rang hat: Die zugehörige Abbildung ist also injektiv. Natürlich kann sie im Fall $l < k$ nicht surjektiv sein, es gibt also i.A.keine Inverse. Um trotzdem „bestmöglich zu invertieren“, führt man die *Pseudoinverse* C^+ ein:

> Sei $x \in \mathbb{R}^k$. Bestimme ein x_0 aus dem Bild von C so, dass der Abstand zu x (im l^2-Sinn) minimal ist und schreibe $x_0 = Cy$ für ein (nach Voraussetzung eindeutig bestimmtes) $y \in \mathbb{R}^l$. Dann ist $C^+x := y$.

Im Fall $k = l$ ist wirklich $C^+ = C^{-1}$, dadurch ist der Name gerechtfertigt. Da C injektiv ist, ist auch $C^\top C : \mathbb{R}^l \to \mathbb{R}^l$ injektiv. (Angenommen nämlich, es ist $C^\top Cx = 0$. Dann ist

$$0 = \langle C^\top Cx, x\rangle = \langle Cx, Cx\rangle = ||Cx||_2^2,$$

und daraus folgt zunächst $Cx = 0$ und dann $x = 0$.) Wir können also $(C^\top C)^{-1}$ bilden, und damit kann C^+ explizit angegeben werden:

Satz 3.4.1. *Es ist $C^+ = (C^\top C)^{-1}C^\top$. Insbesondere folgt, dass C^+ eine lineare Abbildung ist.*

Beweis: Der $\mathbb{R}^k$ zerfällt in das Bild von C und sein orthogonales Komplement V. Sei $x \in \mathbb{R}^l$, wir schreiben $x = x_0 + z$ mit $x_0 \in C(\mathbb{R}^l)$ und $z \in V$. Dann ist x_0 die beste Approximation an x aus $C(\mathbb{R}^l)$, es ist also $C^+x = y$, wo $Cy = x_0$. Wir bemerken nun, dass $C^\top z = 0$. Das folgt aus

$$||C^\top z||_2^2 = \langle C^\top z, C^\top z\rangle = \langle z, CC^\top z\rangle = 0$$

(denn $z \perp C(\mathbb{R}^l)$). Zusammen:

$$\begin{aligned}(C^\top C)^{-1}C^\top x &= (C^\top C)^{-1}C^\top(x_0 + z)\\ &= (C^\top C)^{-1}C^\top x_0\\ &= (C^\top C)^{-1}C^\top(Cy)\\ &= y.\end{aligned}$$

Das beweist die Behauptung. □

Matrixnormen

C sei wie vorstehend. Versieht man $\mathbb{R}^l$ und $\mathbb{R}^k$ mit irgendwelchen Normen $||\cdot||$ und $||\cdot||'$, so versteht man unter der zugehörigen *Matrixnorm* $||C||$ die kleinste Zahl K, so dass

$$||Cx||' \leq K||x||$$

für alle x gilt. Es gilt dann stets $||Cx||' \leq ||C||\,||x||$.

Für uns wird nur der Fall interessant sein, dass beide Räume mit $||\cdot||_1$ oder $||\cdot||_\infty$ versehen sind. Wir schreiben dann $||C||_{1\to 1}$ bzw. $||C||_{\infty\to\infty}$.

Lemma 3.4.2. *Ist $||Cx||' < ||x||$ für alle $x \neq 0$, so ist $||C|| < 1$. Die Umkehrung gilt offensichtlich auch.*

Beweis: Man kombiniere die Beobachtung $||C|| = \sup_{||x||=1} ||Cx||'$ mit einem Kompaktheitsschluss. □

Lemma 3.4.3.

$$||C||_{1\to 1} = ||C^\top||_{\infty\to\infty}$$

Beweis: (Die Aussage ist ein Spezialfall der Gleichung $||T|| = ||T^*||$ für Operatoren.) Sei $x \in \mathbb{R}^l$ mit $||x||_1 = 1$. Wähle $y \in \mathbb{R}^k$ so, dass $||y||_\infty = 1$ und

$$||Cx||_1 = \langle Cx, y\rangle = (Cx)^\top y = x^\top C^\top y.$$

Dann ist $||C^\top y||_\infty \leq ||C^\top||_{\infty\to\infty}$, und folglich ist $x^\top C^\top y \leq ||C^\top||_{\infty\to\infty}$. Das beweist $||C||_{1\to 1} \leq ||C^\top||_{\infty\to\infty}$.
Sei umgekehrt $y \in \mathbb{R}^k$ mit $||y||_\infty = 1$. Wir wählen $x \in \mathbb{R}^l$ mit $||x||_1 = 1$ und

$$||C^\top y||_\infty = \langle C^\top y, x\rangle = y^\top Cx.$$

Es ist $||Cx||_1 \leq ||C||_{1\to 1}$, deswegen ist $y^\top Cx \leq ||C||_{1\to 1}$. Damit ist auch $||C||_{1\to 1} \geq ||C^\top||_{\infty\to\infty}$ bewiesen. □

3.5 Der Zusammenhang

In den Abschnitten 3.1 und 3.2 haben wir Bedingungen kennengelernt, unter denen l^1-Minimierung bzw. OMP garantiert zum Erfolg führen. Die zweite Bedingung impliziert die erste:

Satz 3.5.1. *Die $m \times N$-Matrix A habe normalisierte Spalten, es sei $S \subset \{1, \ldots, N\}$ eine s-elementige Teilmenge, und die Spalten $a_j, j \in S$ seien linear unabhängig. Gilt dann*

$$\max_{j\in S} |\langle v, a_j\rangle| > \max_{j\in\overline{S}} |\langle v, a_j\rangle|$$

für alle $v = \sum_{j\in S} x_j a_j \neq 0$, so hat A die Kern-Separierungseigenschaft für S.

Beweis: Die Matrix A_S bestehe aus den zu S gehörigen Spalten. Es ist eine $m \times s$-Matrix, wir fassen sie als lineare Abbildung von $\mathbb{R}^s$ nach $\mathbb{R}^m$ auf. Nach Voraussetzung ist diese Abbildung injektiv.

Für $u \in \mathbb{R}^s \setminus \{0\}$ ist damit $A_S u \neq 0$, und die Voraussetzung kann als

$$||A_S^\top A_S u||_\infty > ||A_{\overline{S}}^\top A_S u||_\infty$$

umgeschrieben werden.

Setze $v := A_S^\top A_S u$, auch das ist wegen der Injektivität von $A_S^\top A_S$ ein typisches Element von $\mathbb{R}^s \setminus \{0\}$. Für v gilt nach Satz 3.4.1[1)]:

$$||v||_\infty > ||A_{\overline{S}}^\top A_S (A_S^\top A_S)^{-1} v||_\infty = ||A_{\overline{S}}^\top (A_S^+)^\top v||_\infty.$$

Also ist $||A_{\overline{S}}^\top (A_S^+)^\top||_{\infty\to\infty} < 1$ (Lemma 3.4.2), und wegen Lemma 3.4.3 heißt das

$$||A_S^+ A_{\overline{S}}||_{1\to 1} < 1.$$

Wir zeigen nun, dass A die Kern-Separierungseigenschaft für S hat. Sei dazu $Av = 0$ mit $v \neq 0$. Dann ist $A_S v_S = -A_{\overline{S}} v_{\overline{S}}$, und der Vektor v_S ist von Null verschieden. (Das ergibt sich aus der Injektivität von A_S.) So folgt

$$\begin{aligned} ||v_S||_1 &= ||A_S^+ A_S v_S||_1 \\ &= ||A_S^+ A_{\overline{S}} v_{\overline{S}}||_1 \\ &\leq ||A_S^+ A_{\overline{S}}||_{1\to 1} ||v_{\overline{S}}||_1 \\ &< ||v_{\overline{S}}||_1. \end{aligned}$$

□

3.6 „Beinahe“ schwach besetzte Vektoren: Stabilität

Um unsere Untersuchungen besser für praktische Anwendungen nutzbar zu machen, sind noch einige Zusatzüberlegungen notwendig. Bisher waren wir doch davon ausgegangen, dass der gesuchte Vektor x s-schwach besetzt ist, also nur wenige von Null verschiedene Komponenten hat. In der Praxis jedoch gibt es so gut wie nie exakte Nullen: x wird s wesentliche Komponenten haben, die anderen werden so klein sein, dass man sie vernachlässigen kann. Um das präzisieren zu können, definieren wir:

Definition 3.6.1. *Für $x \in \mathbb{R}^N$ und $s < N$ sei $\sigma_s(x)_1$ das Infimum der Zahlen $||x - z||_1$, wobei z über alle s-schwach besetzten Vektoren läuft.*

Es ist dann x s-schwach besetzt genau dann, wenn $\sigma_s(x)_1 = 0$ gilt, und x hat genau dann höchstens s nicht zu vernachlässigende Einträge, wenn $\sigma_s(x)_1$

[1)] Beachte, dass $A_S^\top A_S$ und damit die Inverse selbstadjungiert sind.

„sehr klein“ ist. $\sigma_s(x)_1$ ist übrigens leicht zu berechnen: Suche eine s-elementige Menge S, so dass $|x_i| \geq |x_j|$ für alle $i \in S, j \in \overline{S}$; S enthält also die Indizes mit den s betragsmäßig größten Einträgen. Dann ist

$$\sigma_s(x)_1 = \sum_{j \in \overline{S}} |x_j|.$$

Wie verhalten sich denn unsere Strategien (l^1-Minimierung und OMP) in dieser allgemeineren Situation? Wir wollen das nur für die l^1-Minimierung genauer analysieren.

Definition 3.6.2. *Wie üblich habe A normalisierte Spalten, je $2s$ Spalten seien linear unabhängig und es sei $\rho \in]\,0,1\,[$. Wir sagen, dass A die* stabile Kern-Separierungseigenschaft zum Parameter ρ *für s hat, wenn*

$$||v_S||_1 \leq \rho ||v_{\overline{S}}||_1$$

für alle v mit $Av = 0$ und alle s-elementigen Mengen S gilt.

Offensichtlich folgt die Kern-Separierungseigenschaft aus der stabilen Kern-Separierungseigenschaft.

Hier ist das Hauptergebnis:

Satz 3.6.3. *Die $m \times N$-Matrix A habe die* stabile Kern-Separierungseigenschaft zum Parameter ρ *für s. Ein $x \in \mathbb{R}^n$ sei gegeben, und $x_0 \in \mathbb{R}^N$ sei so, dass $Ax = Ax_0$ gilt und $||x_0||_1$ minimal ist. Dann gilt*

$$||x - x_0|_1 \leq 2\frac{1+\rho}{1-\rho}\sigma_s(x)_1.$$

Wenn also x beinahe s-schwach besetzt ist, so findet l^1-Minimierung eine Approximation von x. Die wird umso besser sein, je kleiner ρ ist. (Man mache sich durch eine Skizze klar, dass die Aussage qualitativ plausibel ist.)

Der Beweis wird aufgeschoben, als Vorbereitung beweisen wir ein Lemma und eine Charakterisierung:

Lemma 3.6.4. *Für $S \subset \{1, \ldots, N\}$ und $x, z \in \mathbb{R}^N$ gilt*

$$||(x-z)_{\overline{S}}||_1 \leq ||z||_1 - ||x||_1 + ||(x-z)_S||_1 + 2||x_{\overline{S}}||_1.$$

Beweis: Zum Beweis addieren wir die Ungleichungen

$$\begin{aligned} ||x||_1 &= ||x_{\overline{S}}||_1 + ||x_S||_1 \\ &\leq ||x_{\overline{S}}||_1 + ||(x-z)_S||_1 + ||z_S||_1 \end{aligned}$$

und

$$||(x-z)_{\overline{S}}||_1 \leq ||x_{\overline{S}}||_1 + ||z_{\overline{S}}||_1.$$

Damit folgt

$$||x||_1 + ||(x-z)_{\overline{S}}||_1 \leq 2||x_{\overline{S}}||_1 + ||(x-z)_S||_1 + ||z||_1,$$

und das ist offensichtlich gleichwertig zur Behauptung. □

Und hier die Charakterisierung:

Satz 3.6.5. *Die folgenden Aussagen sind (für ein ρ mit $0 < \rho < 1$ und $S \subset \{1, \ldots, N\}$) äquivalent:*

(i) Für v mit $Av = 0$ gilt $||v_S||_1 \leq \rho||v_{\overline{S}}||_1$. (D.h., A hat die stabile Kern-Separierungseigenschaft zum Parameter ρ für S.)

(ii) Für $x, z \in \mathbb{R}^N$ mit $Ax = Az$ gilt

$$||z-x||_1 \leq \frac{1+\rho}{1-\rho}\big(||z||_1 - ||x||_1 + 2||x_{\overline{S}}||_1\big).$$

Beweis:

(ii) ⇒ (i): Ein v mit $Av = 0$ sei vorgelegt. Dann ist $A(-v_S) = Av_{\overline{S}}$, wir können also (ii) mit $x = -v_S$ und $z = v_{\overline{S}}$ anwenden. Unter Beachtung von $z - x = v$ und $x_{\overline{S}} = 0$ folgt

$$\begin{aligned} ||v||_1 &\leq \frac{1+\rho}{1-\rho}\big(||z||_1 - ||x||_1 + 2||x_{\overline{S}}||_1\big) \\ &= \frac{1+\rho}{1-\rho}\big(||v_{\overline{S}}||_1 - ||v_S||_1\big). \end{aligned}$$

Nun ist $||v||_1 = ||v_S||_1 + ||v_{\overline{S}}||_1$, und da – wie leicht zu sehen – aus

$$\alpha + \beta \leq \frac{1+\rho}{1-\rho}(\alpha - \beta)$$

stets $\beta \leq \rho\alpha$ folgt, ist (i) gezeigt.

(i) ⇒ (ii): Es sei $Ax = Az$, wir setzen $v = z - x$. Wegen $Av = 0$ folgt $||v_S||_1 \leq \rho||v_{\overline{S}}||_1$. Das Lemma liefert

$$||v_{\overline{S}}||_1 \leq ||z||_1 - ||x||_1 + ||v_S||_1 + 2||x_{\overline{S}}||_1,$$

zusammen ergibt das

$$||v_{\overline{S}}||_1 \leq ||z||_1 - ||x||_1 + \rho||v_{\overline{S}}||_1 + 2||x_{\overline{S}}||_1.$$

Es folgt

$$(1-\rho)||v_{\overline{S}}||_1 \leq ||z||_1 - ||x||_1 + 2||x_{\overline{S}}||_1.$$

Nun das Finale:

$$\begin{aligned} ||x - z||_1 &= ||v||_1 \\ &= ||v_{\overline{S}}||_1 + ||v_S||_1 \\ &\leq (1+\rho)||v_{\overline{S}}||_1 \\ &\leq \frac{1+\rho}{1-\rho}\big(||z||_1 - ||x||_1 + 2||x_{\overline{S}}||_1\big). \end{aligned}$$

Damit ist (ii) gezeigt. □

Es ist nun leicht, Satz 3.6.3 zu beweisen. x sei vorgelegt, und eine s-elementige Menge S wird so gewählt, dass die Indizes mit den s größten Komponenten von x enthalten sind. Dann ist $\sigma_s(x)_1 = ||x_{\overline{S}}||_1$. Wähle ein x_0, so dass $Ax = Ax_0$ gilt und $||x_0||_1$ minimal ist. Insbesondere ist also $||x_0||_1 \leq ||x||_1$, und damit folgt die Ungleichung in Satz 3.6.3 sofort aus der in Satz 3.6.4.

Kapitel 4

Kohärenz

4.1 Definitionen

Wie kann man erreichen, dass die vorgeschlagenen Verfahren auch wirklich das gewünschte Ergebnis liefern? Offensichtlich muss man dazu für die Matrix A die „richtigen“ Eigenschaften fordern. Dabei gibt es mehrere Antworten auf die Frage, was „richtig“ bedeuten könnte. Wir beginnen mit

Definition 4.1.1. *Sei A eine $m \times N$-Matrix; die Spalten $a_j, j = 1, \dots, N$ seien normiert. Unter der* Kohärenz von A *verstehen wir die Zahl*

$$\mu := \max_{i \neq j} |\langle a_i, a_j \rangle|.$$

Allgemeiner: Für $s < N$ bezeichnet $\mu_1(s)$ das Maximum der Zahlen $\sum_{i \in S} |\langle a_i, a_j \rangle|$, wobei S alle s-elementigen Teilmengen von $\{1, \dots, N\}$ und i alle Elemente aus $\{1, \dots, N\} \setminus S$ durchläuft.

Bemerkungen: 1. μ misst also, wie weit die Winkel zwischen den Spaltenvektoren von 90 Grad abweichen: In einem Orthogonalsystem wäre $\mu = 0$.

2. Für $s + t < N$ ist

$$\max\{\mu_1(s), \mu_2(t)\} \leq \mu_1(s + t) \leq \mu_1(s) + \mu_1(t).$$

(Übungsaufgabe.)

3. $\mu = \mu_1(1) \leq \mu_1(2) \leq \cdots \leq \mu_1(s) \leq s\mu$.

Es wird sich zeigen, dass „kleine“ Kohärenz ausreicht, um mit den Ansätzem l^1-Minimierung und OMP Erfolg zu haben. Doch wie klein kann die Kohärenz sein?[1)] Diese Frage untersuchen wir im nächsten Abschnitt.

[1)] Im Extremfall $\mu = 0$ bilden die $a_1, \dots, a_N$ ein Orthonormalsystem (ONS) im $\mathbb{R}^m$, und so etwas gibt es nur im Fall $N \leq m$.

4.2 Matrizen mit kleiner Kohärenz

Vektoren b_j, $j = 1, \ldots, N$ im $\mathbb{R}^m$ seien gegeben. Sie heißen ein *reproduzierendes System* („tight frame “), wenn für alle $x \in \mathbb{R}^m$ gilt:

$$x = \sum_j \langle x, b_j \rangle b_j.$$

Ein einfaches Beispiel: Man kann k beliebige Orthonormalbasen im $\mathbb{R}^m$ zusammenfassen und alle Vektoren dieser Vereinigung durch k teilen.

Satz 4.2.1. *Äquivalent sind:*
(i) Die b_j sind ein reproduzierendes System.
(ii) Für alle x ist

$$||x||_2^2 = \sum_j |\langle x, b_j \rangle|^2.$$

(iii) Sei $B = (b_{ij})$ die Matrix, die als Spalten die b_j hat. Dann sind die Zeilen $c_i, i = 1, \ldots, m$ ein ONS im $\mathbb{R}^N$.

Beweis: (iii)⇒(i): Aus Linearitätsgründen ist das nur für die Einheitsvektoren $x = e_{i_0}$ zu zeigen, $i_0 = 1, \ldots, N$. Dann ist

$$\begin{aligned} \sum_j \langle x, b_j \rangle b_j &= \sum_j b_{i_0,j} b_j \\ &= \left(\langle c_{i_0}, c_i \rangle \right)_i \\ &= (\delta_{i_0 i})_i \\ &= e_{i_0}. \end{aligned}$$

(iii)⇒(ii): x sei gegeben, wir betrachten $X := \sum_i x_i c_i \in \mathbb{R}^N$. Da die c_i ein ONS bilden, ist

$$\begin{aligned} ||x||_2^2 &= \sum_i |x_i|^2 \\ &= ||X||_2^2 \\ &= \sum_j |X_j|^2 \\ &= \sum_j |\langle x, b_j \rangle|^2. \end{aligned}$$

(i)⇒(iii): Es ist $\left(\sum_j \langle x, b_j \rangle b_j \right)_{i_0} = \sum_i x_i \langle c_i, c_{i_0} \rangle$ für alle x, i, i_0. Setzt man das speziell für $x = e_{i_0}$ ein, so folgt aus (i), dass die c_i ein ONS bilden.
(ii)⇒(iii): Hier muss man sich vorbereitend klar machen, dass X, Y ein ONS genau dann bilden, wenn $||\alpha X + \beta Y||^2$ stets mit $\alpha^2 + \beta^2$ übereinstimmen. (Entsprechendes gilt für m Summanden.) Damit folgt (iii) sofort aus (ii), denn

$$||\sum_i x_i c_i||^2 = \sum_j |\langle x, b_j \rangle|^2 = ||x||^2.$$

□

Es ist also leicht, sich ein reproduzierendes System zu verschaffen. Bestimme ein ONS aus m Elementen im $\mathbb{R}^N$ und schreibe die entsprechenden Vektoren zeilenweise in eine Matrix B. Die Spalten bilden dann ein reproduzierendes System im $\mathbb{R}^m$.

Etwas allgemeiner gilt der

Satz 4.2.2. *Für eine $m \times N$-Matrix A mit Spalten a_j und ein $\lambda > 0$ sind äquivalent:*

(i) Für alle x ist $x = \lambda \sum_j \langle x, a_j\rangle a_j$.

(ii) Für alle x ist $||x||_2^2 = \lambda \sum_j |\langle x, a_j\rangle|^2$.

(iii) Die Zeilen von A sind orthogonal und haben l^2-Norm $1/\sqrt{\lambda}$. (In Kurzfassung: $AA^\top = \lambda^{-1} E_m$, dabei ist E_m die $m \times m$- Einheitsmatrix.)

Wenn diese Bedingungen erfüllt sind, soll $a_1, \ldots, a_N$ ein λ-reproduzierendes System *heißen.*

Beweis: Mit $b_j := a_j\sqrt{\lambda}$ kann diese Behauptung auf den vorigen Satz zurückgeführt werden. □

Definition 4.2.3. *l^2-normalisierte Vektoren $a_1, \ldots, a_N$ im $\mathbb{R}^m$ seien gegeben. Wir sagen, dass sie ein* gleichwinkliges System *bilden, wenn alle Zahlen $|\langle a_j, a_{j'}\rangle|$ für $j \neq j'$ gleich sind.*

Man denke im $\mathbb{R}^2$ an einen Mercedesstern oder im Fall $N \leq m$ an orthogonale Vektoren. Mit den vorstehenden Definitionen können wir die Frage beantworten, wie klein die Kohärenz bestenfalls sein kann:

Satz 4.2.4. *Für die Kohärenz μ einer $m \times N$-Matrix A mit normalisierten Spalten gilt stets*

$$\mu \geq \sqrt{\frac{N-m}{m(N-1)}}.$$

(Welch-Ungleichung.)

Gleichheit gilt genau dann wenn die Spalten ein gleichwinkliges System bilden, das gleichzeitig λ-reproduzierend für ein geeignetes $\lambda > 0$ ist.

Der Beweis wird verschoben, da wir einige Vorbereitungen benötigen.

Spuren von Matrizen

Ist $C = (c_{ij})_{i,j=1,\ldots,n}$ eine quadratische Matrix, so ist die *Spur* die Summe der Elemente über die Hauptdiagonale:

$$\mathrm{Spur}(C) := \sum_{i=1,\ldots,n} c_{ii}.$$

Wichtig ist zu wissen, dass stets $\mathrm{Spur}(CD) = \mathrm{Spur}(DC)$ gilt, wenn beide Matrixprodukte definiert sind: Ist C $k \times l$-Matrix und D $l \times k$-Matrix, so gilt nämlich

$$\mathrm{Spur}(CD) = \mathrm{Spur}(DC) = \sum_{i=1,\ldots,l,j=1,\ldots,k} c_{ji} d_{ij}.$$

Die Gram-Matrix

Das ist eine Erinnerung an die lineare Algebra. Sind $a_1, \ldots, a_N$ Vektoren im $\mathbb{R}^m$, so ist die zugehörige *Gram-Matrix* die $N \times N$-Matrix mit den Einträgen $\langle a_i, a_j \rangle$. Man weiß, dass G genau dann invertierbar ist, wenn die a_j linear unabhängig sind.

Fasst man die a_i als Spalten einer Matrix A auf, so ist G gerade die Matrix $A^\top A$.

Die Frobenius-Norm einer Matrix

Sei X der Vektorraum der $n \times n$-Matrizen. Dann definiert

$$\langle U, V \rangle_F := \mathrm{Spur}(UV^\top)$$

ein inneres Produkt auf X. Die zugehörige Norm ($||U||_F := \sqrt{\langle U, U \rangle_F}$) heißt die *Frobeniusnorm* auf X.

Klar ist, dass $\langle \cdot, \cdot \rangle_F$ bilinear ist. Die positive Definitheit folgt sofort aus der Formel

$$\langle U, U \rangle_F = \mathrm{Spur}(UU^\top) = \sum_{i,j} |u_{ij}|^2.$$

Beweis von Satz 4.2.4

Setze $G := A^\top A$ (die Gram-Matrix) und $H := AA^\top$. Da die a_i normalisiert sind, ist $\mathrm{Spur}(G) = N$. Andererseits folgt aus der Cauchy-Schwarzschen Ungleichung für $\langle \cdot, \cdot \rangle_F$, dass

$$\mathrm{Spur}(H) = \langle H, E_m \rangle_F \leq ||H||_F ||E_m||_F = \sqrt{m} \sqrt{\mathrm{Spur}(HH^\top)}.$$

(E_m ist die m-dimensionale Einheitsmatrix.) So folgt

$$\begin{aligned} \mathrm{Spur}(HH^\top) &= \mathrm{Spur}(AA^\top AA^\top) \\ &= \mathrm{Spur}(A^\top AA^\top A) \\ &= \mathrm{Spur}(GG^\top) \\ &= \sum_{i,j} |\langle a_i, a_j \rangle|^2 \\ &= N + \sum_{i,j,i \neq j} |\langle a_i, a_j \rangle|^2, \end{aligned}$$

und daraus schließen wir

$$\begin{aligned} N^2 &= \big(\mathrm{Spur}(G)\big)^2 \\ &= \big(\mathrm{Spur}(H)\big)^2 \\ &\le m\,\mathrm{Spur}(HH^\top) \\ &\le m\big(N + \sum_{i,j,1\neq j} |\langle a_i, a_j\rangle|^2\big) \\ &\le m\big(N + (N^2-N)\mu^2\big). \end{aligned}$$

Dabei haben wir ausgenutzt, dass $|\langle a_i, a_j\rangle| \le \mu$ für $i \neq j$ und dass es $N^2 - N$ Summanden $i, j, i \neq j$ gibt. Durch Umstellen folgt die Ungleichung des Satzes.

Angenommen, es gilt die Gleichheit. Dann muss die auch bei den beiden „$\le$“ des Beweises gelten: Erstens muss in der Cauchy-Schwarzschen Ungleichung ein „$=$“ stehen, und zweitens müssen alle $|\langle a_i, a_j\rangle|$ gleich μ sein.
Die erste Bedingung liefert $H = \lambda E_m$ für ein geeignetes $\lambda > 0$, und damit ist alles gezeigt. □

Bemerkungen: 1. Meist ist N groß gegen m, und dann besagt der Satz, dass μ bestenfalls von der Größenordnung $1/\sqrt{m}$ ist.
2. Ganz analog kann man auch $\mu_1(s)$ abschätzen: Es gilt stets

$$\mu_1(s) \ge s\sqrt{\frac{N-m}{m(N-1)}}.$$

(Vgl. Foucart-Rauhut, Theorem 5.8.)
3. Der Satz besagt natürlich nicht, dass bei beliebigen m und N eine $m \times N$-Matrix existieren muss, für die die Kohärenz nahe an der Welch-Schranke ist. (Man mache sich das für $m = 3$ und „große“ N klar.)

Es ist naheliegend zu versuchen, möglichst große gleichwinklige Systeme zu bilden, optimale μ werden wegen von Satz 4.2.4 auch nur so gefunden. Der folgende Satz beantwortet die Frage, wie groß die höchstens sein können. Leider ist damit noch kein konkretes Verfahren gefunden, solche S¡steme auch wirklich zu finden, dazu sind noch viele Fragen offen.

Satz 4.2.5. *Die l^2-normierten Vektoren $a_1, \ldots, a_N \in \mathbb{R}^m$ sollen ein gleichwinkliges System bilden. Dann ist $N \le m(m+1)/2$.*

Beweis: Als Vorbereitung betrachten wir eine $n \times n$-Matrix der Form

$$M_z = \begin{pmatrix} 1 & z & z & \cdots & z \\ z & 1 & z & \cdots & z \\ \cdots & \cdots\cdots & \cdots & \cdots & \cdots \\ z & z & z & \ldots & 1 \end{pmatrix}.$$

Wir behaupten, dass die Matrix genau dann invertierbar ist, wenn $z \neq 1$ und $z \neq -1/(n-1)$ gilt.

Beweis dazu: Im Fall $z = 1$ ist M_z sicher nicht invertierbar, und im Fall $z = -1/(n-1)$ liegt $(1, 1, \ldots, 1)^\top$ im Kern: M_z ist dann auch nicht injektiv.

Wir nehmen nun an, dass $1 \neq z \neq -1/(n-1)$ gilt. Für ein $x = (x_1, \ldots, x_n)^\top$ gelte $M_z x = 0$. (Ziel: $x = 0$.) Schreibt man das in n Gleichungen aus und addiert, so folgt $(x_1 + \cdots + x_n)\big(1 + (n-1)z\big) = 0$. Also gilt $x_1 + \cdots + x_n = 0$. Setzt man das in die Gleichungen ein, so ergibt sich für jedes i die Beziehung $x_i(1 - z) = 0$. Es folgt $x = 0$, und damit ist M_z invertierbar.

Es sei X der Raum der symmetrischen $m \times m$-Matrizen; X ist $m(m+1)/2$-dimensional. Wir versehen diesen Raum mit dem Frobenius-Skalarprodukt (s.o.) und betrachten für jedes i die Abbildung $P_i : v \mapsto \langle v, a_i \rangle a_i$. Schreibt man ein a_i als $(\alpha_1, \ldots, \alpha_m)^\top$, so hat P_i die Matrixdarstellung $(\alpha_s \alpha_t)_{s,t=1,\ldots,m}$. Das zeigt, dass P_i zu X gehört und dass $\mathrm{Spur}(P_i) = 1$ gilt. Auch ist klar, dass $(P_i)^2 = P_i$ gilt. Wir folgern:

$$\langle P_i, P_i \rangle_F = \mathrm{Spur}(P_i P_i^\top) = \mathrm{Spur}(P)_i = 1$$

sowie, für $i \neq j$ und $a_j = (\beta_1, \ldots, \beta_m)^\top$,

$$\begin{aligned} \langle P_i, P_j \rangle_F &= \mathrm{Spur}(P_i P_j)^\top \\ &= (\alpha_1 \beta_1 + \cdots \alpha_m \beta_m)^2 \\ &= (\langle a_i, a_j \rangle)^2. \end{aligned}$$

Nach Voraussetzung stimmen alle $(\langle a_i, a_j \rangle)^2$ mit einer Zahl c^2 überein, und deswegen ist die Gram-Matrix der $P_1, \ldots, P_N$ gleich M_{c^2}. Es ist $0 \leq c^2 < 1$, und deswegen ist die Gram-Matrix aufgrund unserer Vorbereitung invertierbear. Die P_j sind also linear unabhängig in X, und da dieser Raum $m(m+1)/2$-dimensional ist, bedeutet das $N \leq m(m+1)/2$ □

4.3 Kleine Kohärenz impliziert den Erfolg für OMP

Wie üblich sei A eine $m \times N$-Matrix mit normalisierten Spalten $a_j, j = 1, \ldots, N$. Wenn die Kohärenz klein genug ist, kann man garantieren, dass OMP zum Erfolg führt:

Satz 4.3.1. *Es gelte $\mu_1(s-1) + \mu_1(s) < 1$. Dann können s-schwachbesetzte Vektoren x als Lösung von $Ax = y$ nach höchstens s BMP-Schritten identifizeirt werden.*

Wegen $\mu_1(s) \leq s\mu$ ist die Bedingung insbesondere dann erfüllt, wenn $\mu < 1/(2s-1)$ gilt.

Beweis: Sei $S \subset \{1, \ldots, N\}$ eine Teilmenge mit s Elementen. Wir müssen wegen Satz 3.3.1 zeigen:

- A_S ist injektiv.
- Hat $x \in \mathbb{R}^N$ den Träger in S und ist $v := \sum_{i \in S} x_i a_i \neq 0$, so gilt:

$$\max_{j \in S} |\langle v, a_j \rangle| > \max_{l \in \overline{S}} |\langle v, a_l \rangle|.$$

1. Sei $\sum_{i \in S} x_i a_i = 0$. Wir wollen zeigen, dass alle x_i gleich Null sind.

Angenommen, das ist nicht der Fall: x_{i_0} sei die betragsmäßig größte Komponente. Dann ist $-x_{i_0} a_{i_0} = \sum_{i \in S, i \neq i_0} x_i a_i$, und es folgt

$$\begin{aligned} |x_{i_0}| &= |\sum_{i \in S, i \neq i_0} \langle a_{i_0}, x_i a_i \rangle| \\ &\leq \sum_{i \in S} |x_{i_0}| |\langle a_{i_0}, a_i \rangle| \\ &\leq |x_{i_0}| \mu_1(s-1). \end{aligned}$$

Das würde zu dem Widerspruch $1 \leq \mu_1(s-1)$ führen.

2. Wähle $i_0 \in S$ mit $|x_{i_0}| = \max_{i \in S} |x_i|$. Für $l \in \overline{S}$ ist dann

$$\begin{aligned} |\langle v, a_l \rangle| &= |\sum_{i \in S} x_i \langle a_i, a_l \rangle| \\ &\leq \sum_{i \in S} |x_i| |\langle a_i, a_l \rangle| \\ &\leq |x_{i_0}| \mu_1(s). \end{aligned}$$

Andererseits gilt

$$\begin{aligned} |\langle v, a_{i_0} \rangle| &= |\sum_{i \in S} x_i \langle a_i, a_{i_0} \rangle| \\ &\geq |x_{i_0}| \langle a_{i_0}, a_{i_0} \rangle - \sum_{i \in S, i \neq i_0} |x_i| |\langle a_i, a_{i_0} \rangle| \\ &\geq |x_{i_0}| (1 - \mu_1(s-1)). \end{aligned}$$

Wegen $\mu_1(s) < (1 - \mu_1(s-1))$ heißt das

$$\max_{i \in \overline{S}} |\langle v, a_i \rangle| \leq |x_{i_0}| \mu_1(s)) < |x_{i_0}| (1 - \mu_1(s-1)) \leq |\langle v, a_{i_0} \rangle|.$$

Das beweist die Behauptung. □

4.4 Kleine Kohärenz impliziert den Erfolg bei l^1-Minimierung

Bei kleiner Kohärenz findet man schwach besetzte Lösungen durch l_1-Minimierung:

Satz 4.4.1. *Es sei $\mu_1(s)+\mu_1(s-1) < 1$. Ist dann x ein s-schwachbesetzter Vektor, so ist x die eindeutig bestimmte Lösung des Minimierungsproblems $Ax = y$ mit $||x||_1 = \min$.*

Beweis: Eigentlich ist ein Beweis nicht erforderlich, denn die Aussage folgt, wenn man Satz 3.5.1 und Satz 4.3.1 kombiniert. Ein einfacher direkter Beweis kann aber auch gegeben werden.

$S \subset \{1, \ldots, N\}$ sei eine Menge mit s Elementen, und $v \in \mathbb{R}^N \backslash \{0\}$ mit $Av = 0$ ist gegeben. Wegen Satz 3.2.2 ist nur zu zeigen, dass dann $||v_S||_1 < ||v_{\overline{S}}||_1$ folgt. Es ist also $\sum_j v_j a_j = 0$. Sei zunächst $i \in S$. Dann gilt

$$\begin{aligned} v_i &= v_i \langle a_i, a_i \rangle \\ &= -\sum_{j \neq i} v_j \langle a_j, a_i \rangle \\ &= -\sum_{j \in S, j \neq i} v_j \langle a_j, a_i \rangle - \sum_{j \in \overline{S}} v_j \langle a_j, a_i \rangle. \end{aligned}$$

Folglich ist

$$|v_i| \leq \sum_{j \in S, j \neq i} |v_j| |\langle a_j, a_i \rangle| + \sum_{j \in \overline{S}} |v_j| |\langle a_j, a_i \rangle|.$$

Summiert man über die $i \in S$, ergibt sich

$$||v_S||_1 \leq \mu_1(s-1)||v_S||_1 + \mu_1(s)||v_{\overline{S}}||_1,$$

und das impliziert

$$||v_S||_1 \leq \frac{\mu_1(s)}{1-\mu_1(s-1)} ||v_{\overline{S}}||_1 < ||v_{\overline{S}}||_1.$$

□

Kapitel 5

Fast-Isometrie-Konstanten

In diesem Kapitel lernen wir eine weitere Maßzahl kennen, mit der vorausgesagt werden kann, ob unsere Verfahren zum Auffinden schwach besetzter Lösungen der Gleichungen $Ax = y$ zum Erfolg führen. Das Problem, wie man Matrizen mit „günstiger“ Maßzahl findet, wird erst später mit stochastischen Methoden gelöst werden können.

5.1 Definitionen

Gegeben seien s Vektoren a_i, $i = 1, \ldots, s$. Im Allgemeinen ist die Menge

$$\{\sum_i \alpha_i a_i \mid \alpha_i \in \mathbb{R}, \sum_i |\alpha_i|^2 \leq 1\}$$

ein Ellipsoid. Die folgende Definition misst, wie nahe diese Ellipsoide einer perfekten Kugel sind, wenn alle s-elementigen Teilmengen der Spalten von einer Matrix A durchlaufen werden.

Definition 5.1.1. *Sei A eine $m \times N$-Matrix und $s \leq N$. Die* s-Isometriekonstante *δ_s ist die kleinste Zahl $\delta \geq 0$, so dass gilt:*

$$(1 - \delta) \leq ||Ax||_2^2 \leq (1 + \delta)$$

für alle x mit $||x||_0 \leq s$ und $||x||_2 = 1$.

In vielen Fällen wird zusätzlich vorausgesetzt werden, dass A l^2-normalisierte Spalten hat.

Bemerkungen: 1. Insbesondere kann man für x die Einheitsvektoren einsetzen. Das bedeutet, dass die Norm der Spalten von A im Intervall $[\,1 - \delta_1, 1 + \delta_1\,]$ liegt.

Sei V der s-dimensionale Raum der $x \in \mathbb{R}^N$, die ihren Träger in einer festen s-elementigen Menge $S \subset \{1, \ldots, N\}$ haben. Unter A wird die l^2-Einheitskugel

in einen s-dimensionalen Teilraum von $\mathbb{R}^m$ abgebildet[1)]. Ist δ_2 „klein“, so ist das Bild der Einheitskugel „beinahe“ eine Kugel.

3. Offensichtlich ist $\delta_1 \leq \delta_2 \leq \delta_3 \leq \cdots$.

Die δ_s haben eine operatortheoretische Bedeutung:

Lemma 5.1.2. *Sei $S \subset \{1, \ldots, N\}$ s-elementig. Dann sind für ein $\delta > 0$ äquivalent:*

(i) Für alle x mit Träger in S und $||x||_2 = 1$ ist $||Ax||_2^2 \in [\,1-\delta, 1+\delta\,]$.

(ii) Für die Norm des Operators $A_S^\top A_S - Id$ (von $\mathbb{R}^s$ nach $\mathbb{R}^s$, jeweils mit der l^2-Norm) gilt

$$||A_S^\top A_S - Id||_{2\to 2} \leq \delta.$$

Es folgt:

$$\delta_s = max_{S\subset\{1,\ldots,N\},|S|=s}||A_S^\top A_S - Id||_{2\to 2}.$$

Beweis: Zunächst bemerken wir: Ist $B : \mathbb{R}^s \to \mathbb{R}^s$ ein symmetrischer Operator, so ist $||B||_{2\to 2} = \max_{||x||_2=1} |\langle Bx, x\rangle|$. Für Diagonaloperatoren ist das klar, und B kann durch eine orthogonale Transformtion in diese Form gebracht werden.

(i) $\Rightarrow$ (ii): Die Voraussetzung besagt, dass $|\langle A_S x, A_S x\rangle - \langle x, x\rangle| \leq \delta$ für alle normalisierten Vektoren gilt. Es ist also $|\langle (A_S^\top A_S - Id)x, x\rangle| \leq \delta$. Da $A_S^\top A_S - Id$ symmetrisch ist, impliziert das aufgrund der Vorbemerkung die Aussage (ii).

(ii) $\Rightarrow$ (i): Dazu muss man nur den vorigen Beweis rückwärts lesen. □

5.2 Allgemeine Eigenschaften der s-Isometriekonstanten

Zunächst untersuchen wir den Zusammenhang zur Kohärenz. Als Vorbereitung beweisen wir einen Spezialfall des Gershgorin-Theorems:

Lemma 5.2.1. *Sei B eine symmetrische $s \times s$-Matrix, für die alle Diagonalelemente gleich 1 sind. Ist dann λ ein Eigenvektor, so gilt*

$$|\lambda - 1| \leq \max_i \sum_{j=1,\ldots,s, j\neq i} |b_{ij|}.$$

Beweis: Wähle ein $x \neq 0$ im $\mathbb{R}^s$ mit $Bx = \lambda x$. Die betragsmäßig größte Komponente stehe beim Index i_0, und o.E. gelte $x_{i_0} = 1$.
$Ax = \lambda x$ besagt insbesondere, dass

$$\lambda x_{i_0} = \sum_j b_{i_0 j} x_j = x_{i_0} + \sum_{j, j\neq i_0} b_{i_0 j} x_j.$$

[1)] Jedenfalls wenn je s Spalten linear unabhängig sind.

Folglich ist

$$\begin{aligned} |1-\lambda| &= |\sum_{j,j\neq i_0} b_{i_0 j}x_j| \\ &\leq \sum_{j,j\neq i_0} |b_{i_0 j}||x_j| \\ &\leq \sum_{j,j\neq i_0} |b_{i_0 j}|. \end{aligned}$$

□

Satz 5.2.2. *A habe l^2-normalisierte Spalten. Dann gilt für jeden s-schwachbesetzten Vektor x die Ungleichung*

$$\big(1-\mu_1(s-1)\big)||x|_2^2 \leq ||Ax||_2^2 \leq \big(1+\mu_1(s-1)\big)||x|_2^2.$$

Insbesondere ist $\delta_s \leq \mu_1(s-1)$.

Beweis: Es ist zu zeigen, dass $||A_S^\top A_S - Id||_{2\to 2} \leq \mu_1(s-1)$ für s-elementige Mengen S gilt. Die symmetrische $s\times s$-Matrix $A_S^\top A_S$ hat die Einträge $\langle a_i, a_j\rangle$ mit $i,j\in S$. Die Hauptdiagonalelemente sind 1, und deswegen ist der Abstand der Eigenwerte zu 1 aufgrund des Lemmas durch $\max_{i_0\in S}\sum_{j\in S\setminus\{i_0\}}|\langle a_{i_0}, a_j\rangle|$, also durch $\mu_1(s-1)$, abschätzbar. Folglich ist $||A_S^\top A_S - Id||_{2\to 2} \leq \mu_1(s-1)$ wie behauptet. □

Als Ergänzung zeigen wir noch

Satz 5.2.3. *A habe l^2-normalisierte Spalten. Dann gilt $\delta_1 = 0$ und $\delta_2 = \mu$.*

Beweis: Der erste Teil ist klar. Für den zweiten Teil betrachten wir zwei beliebige verschiedene $i,j\in\{1,\ldots,N\}$ und setzen $S=\{i,j\}$. $A_S^\top A_S$ ist die Matrix

$$\begin{pmatrix} 1 & \langle a_i, a_j\rangle \\ \langle a_i, a_j\rangle & 1 \end{pmatrix}.$$

Die Eigenwerte von $A_S^\top A_S - Id$ sind $\pm\langle a_i, a_j\rangle$, die Operatornorm ist also $|\langle a_i, a_j\rangle|$. Wenn i,j alle Indizes durchlaufen, ist das Maximum dieser Zahlen gleich μ, und das beweist die Behauptung. □

5.3 Kleine δ_s garantieren den Erfolg von l^1-Minimierung

Wir wollen zeigen, dass bei kleinem δ_{2s} l_1-Minimierung erfolgreich sein wird. Wir wollen Satz 3.2.2 anwenden, also zeigen, dass $||v_S||_1 < ||v_{\overline{S}}||_1$ für alle nichttrivialen v mit $Av=0$ gilt.

Als Vorbereitung müssen wir uns mit dem Verhältnis von $||\cdot||_1$ und $||\cdot||_2$ bschäftigen. Aus der Cauchy-Schwarz-Ungleichung, angewendet auf den Hilbertraum $\mathbb{R}^s$, folgt für $x\in\mathbb{R}^s$ leicht die Ungleichung

$$||x||_1 \leq \sqrt{s}||x||_2;$$

man muss nur $|\langle x, 1\rangle|$ abschätzen, wobei 1 der Vektor ist, der aus s Einsen besteht. Wir benötigen aber eine Ungleichung in der anderen Richtung:

Lemma 5.3.1. *Es seien $x, y \in \mathbb{R}^s$, und es gelte $\max_i |x_i| \leq \min_j |y_j|$. Dann ist*

$$||x||_2 \leq \frac{1}{\sqrt{s}}||y||_1.$$

Beweis: Man kombinere die Voraussetzung mit den offensichtlichen Ungleichungen

$$\frac{1}{\sqrt{s}}||x||_2 \leq \max |x_i|$$

und

$$\frac{1}{s}||y||_1 \geq \min |y_j|.$$

□

(Wenn man das für $x = y$ anwenden möchte, so müssen alle x_i den gleichen Betrag haben. Dann liefern das Lemma und die Vorbemerkung die Aussage $||x||_2 = ||x||_1/\sqrt{s}$. Das ist allerdings wenig überraschend.)
Als weitere Vorbereitung beweisen wir einen Zusammenhang zwischen den δ_s und dem Skalarprodukt:

Lemma 5.3.2. *$x, y \in \mathbb{R}^n$ seien s-schwachbesetzte Vektoren mit disjunktem Träger. Dann ist*

$$|\langle Ax, Ay\rangle| \leq \delta_{2s}||x||_2||y||_2.$$

Beweis: Sei S die Vereinigung der Träger von x und y. S hat höchstens $2s$ Elemente, und $\langle x_S, y_S\rangle = 0$. So folgt unter Verwendung der Cauchy-Schwarz-Ungleichung und Lemma 5.1.2

$$\begin{aligned} |\langle Ax, Ay\rangle| &= |\langle Ax_S, Ay_S\rangle - \langle x_S, y_S\rangle| \\ &= |\langle (A_S^* A_S - Id)x_S, y_S\rangle| \\ &\leq ||(A_S^* A_S - Id)x_S||_2||y_S||_2 \\ &\leq ||A_S^* A_S - Id||_{2\to 2}||x||_2||y||_2 \\ &\leq \delta_{2s}||x||_2||y||_2. \end{aligned}$$

□

Und hier ist unser Hauptergebnis:

Satz 5.3.3. *Es gelte $\delta_{2s} < 1/3$. Dann kann l^1-Minnimierung erfolgreich angewendet werden.*

Beweis: Es seien $v \neq 0$ mit $Av = 0$ und ein s-elementiges S vorgegeben. Wir müssen $||v_S||_1 < ||v_{\overline{S}}||_1$ zeigen. Zunächst bemerken wir, dasss das äquivalent zu $||v_S||_1 < ||v||_1/2$ ist, und deswegen dürfen und werden wir annehmen, dass S die s größten Komponenten von v enthält.

Setze $S_0 = S$; und weiter: S_1 enthält die nächstgrößeren s Komponenten, S_2 die s größten mit Indizes in $\{1, \ldots, N\} \setminus (S_0 \cup S_1)$ usw. Auf diese Weise sind die Voraussetzungen von Lemma 5.3.1 jeweils für $x = v_{S_k}$ und $y = v_{S_{k-1}}$ erfüllt. Es gilt also $||v_{S_k}||_2 \leq ||v_{S_{k-1}}||_1/\sqrt{s}$.

Da $Av = 0$ ist, gilt $Av_{S_0} = \sum_{k\geq 1} A(-v_{S_k})$, und wegen

$$||v_{S_0}||_2^2 \leq ||A(v_{S_0})||_2^2/(1-\delta_{2s})$$

folgt unter Verwendung von Lemma 5.3.2

$$\begin{aligned} ||v_{S_0}||_2^2 &\leq \frac{1}{1-\delta_{2s}} |\langle A(v_{S_0}), A(v_{S_1} + Av_{S_2} + \cdots)\rangle| \\ &= \frac{1}{1-\delta_{2s}} \sum_{k\geq 1} |\langle A(v_{S_0}), A(v_{S_k})\rangle| \\ &\leq \frac{\delta_{2s}}{1-\delta_{2s}} ||v_{S_0}||_2 \sum_{k\geq 1} ||v_{S_k}||_2. \end{aligned}$$

Nach Kürzen wird daraus

$$||v_{S_0}||_2 \leq \frac{\delta_{2s}}{1-\delta_{2s}} \sum_{k\geq 1} ||v_{S_k}||_2,$$

also auch (Lemma 5.3.1)

$$\begin{aligned} ||v_{S_0}||_2 &\leq \frac{\delta_{2s}}{(1-\delta_{2s})\sqrt{s}} \sum_{k\geq 1} ||v_{S_{k-1}}||_1 \\ &\leq \frac{\delta_{2s}}{(1-\delta_{2s})\sqrt{s}} ||v||_1. \end{aligned}$$

Und nun das Finale:

$$\begin{aligned} ||v_S||_1 &\leq \sqrt{s}||v_S||_2 \\ &\leq \frac{\delta_{2s}}{(1-\delta_{2s})} ||v||_1 \\ &< \frac{1}{2}||v||_1, \end{aligned}$$

denn $\delta_{2s} < 1/3$ impliziert $\delta_{2s}/(1-\delta_{2s}) < 1/2$, und es ist $v \neq 0$. □

Kapitel 6

Vorbereitungen aus der Stochastik

Hier werden die Untersuchungen des nächsten Kapitels vorbereitet. Eine besondere Rolle spielen dabei Abschätzungen für die Wahrscheinlichkeit des Auftretens „großer“ Abweichungen einer Zufallsvariablen von ihrem Erwartungswert.

6.1 Warum stochastische Methoden?

Wir haben gesehen, dass man im Fall von Matrizen mit speziellen Eigenschaften die Chance hat, schwach besetzte Vektoren x aus Ax zu rekonstruieren. Insbesondere ist es günstig, wenn die Zeilen von A „möglichst orthonormal“ sind.

Stochastik kommt durch die folgende Überlegung ins Spiel. Wir denken uns eine Zufallsvariable X mit Erwartungswert 0 und Varianz 1. Ein m-Vektor x werde durch m unabhängige Zufallsabfragen von X erzeugt. Dann gilt doch:

- Der Erwartungswert von $\sum_i x_i^2$ ist der Erwartungswert der Varianz von X, also gleich 1. Anders ausgedrückt: „Fast immer“ wird die l^2-Norm von x sehr nahe bei 1 sein.

- Ist auch y (unabhängig von x) so entstanden, so ist der Erwartungswert von $\langle x, y\rangle$ gleich dem Produkt der Erwartungswerte von x und y, also gleich 0. Anders ausgedrückt: „Fast immer“ werden x, y „beinahe orthogonal“ sein.

Wenn man also die Zeilen von A durch unabhängige Abfragen von X erzeugt, so hat man gute Chancen, zu kleinen Fastisometriekonstanten δ_s zu kommen.

So wird es auch gehen, doch ist der technische Aufwand, um dieses Ziel zu erreichen, nicht unerheblich.

6.2 Stochastik: Erinnerungen

Es wird im Folgenden vorausgesetzt, dass die folgenden Sachverhalte bekannt sind:

Wahrscheinlichkeitsräume

- Eine *σ-Algebra* $\mathcal{E}$ auf einer Menge Ω ist eine Teilmenge der Potenzmenge, die unter allen Mengenoperationen stabil ist, bei denen höchstens abzählbar viele Elemente von $\mathcal{E}$ beteilgt sind.

- Sei $\mathcal{E}$ eine σ-Algebra auf Ω. Eine Abbildung $\mathbb{P} : \mathcal{E} \to [0,1]$ heißt ein *Wahrscheinlichkeitsmaß*, wenn $\mathbb{P}(\Omega) = 1$ ist und

$$\mathbb{P}\big(\bigcup_n E_n\big) = \sum_n \mathbb{P}(E_n)$$

  für jede Folge (E_n) von paarweise disjunkten Mengen in $\mathcal{E}$ gilt.

- Ein *Wahrscheinlichkeitsraum* ist ein Tripel $(\Omega, \mathcal{E}, \mathbb{P})$; dabei ist Ω eine Menge, $\mathcal{E}$ eine σ-Algebra auf Ω und $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{E})$.

- Die σ-Algebra der *Borelmengen* auf dem $\mathbb{R}^n$ ist die kleinste σ-Algebra, die alle offenen Teilmengen enthält. Faustregel: *Jede* Teilmenge, die in den Anwendungen jemals vorkommen kann, ist eine Borelmenge.

Wichtige Beispiele für Wahrscheinlichkeitsräume

- Ist Ω endlich oder höchstens abzählbar, so ist $\mathcal{E}$ in der Regel die Potenzmenge. Ein Wahrscheinlichkeitsmaß ist dann durch die Angabe der Zahlen $\mathbb{P}(\{\omega\})$ definiert. (Diese Zahlen müssen nichtnegativ sein und sich zu Eins summieren.)

- Die wichtigsten Beispiele dazu sind

  - *Laplaceräume*: Da ist Ω endlich, und alle Elementarereignisse haben die gleiche Wahrscheinlichkeit.
  - *Bernoulliräume*. Hier ist $\Omega = \{0,1\}$, und es reicht die Angabe der Zahl $p = \mathbb{P}(\{1\})$ („Wahrscheinlichkeit für Erfolg“), um das Wahrscheinlichkeitsmaß festzulegen.
  - Abgeleitet von Bernoulliräumen sind die *geometrische Verteilung* (warten auf den ersten Erfolg), die *Binomialverteilung* (k Erfolge in n Versuchen), die *hypergeometrische Verteilung* (Ziehen ohne Zurücklegen) und die *Poissonverteilung* (Grenzwert von Binomialverteilungen).

- Sei zunächst Ω eine „einfache“ Teilmenge von $\mathbb{R}$ (etwa ein Intervall) und $f : \Omega \to \mathbb{R}$ eine „gutartige“ (etwa eine stetige) nichtnegative Funktion mit

Integral Eins. Dann wird dadurch ein Wahrscheinlichkeitsraum durch die Festsetzung

$$\mathbb{P}(E) := \int_E f(x)\,dx$$

definiert. Dabei kann E eine beliebige Borelmenge sein. Für die Anwendungen reicht es aber so gut wie immer, sich für E ein Teilintervall von Ω vorzustellen. f heißt die *Dichtefunktion* zu dem so definierten Wahrscheinlichkeitsmaß.

- Die wichtigsten Beispiele sind
  - Die *Gleichverteilung* auf $[a, b]$; da ist $f(x) := 1/(b-a)$.
  - Die *Exponentialverteilung* zum Parameter $\lambda > 0$; sie ist durch die Dichtefunktion
    $$f(x) := \lambda \cdot e^{-\lambda x}$$
    auf $\mathbb{R}^+$ definiert. Durch die Exponentialverteilung kann gedächtnisloses Warten beschrieben werden.
  - Die *Normalverteilungen* $N(\mu, \sigma^2)$ auf $\mathbb{R}$. Sie haben – für $\mu \in \mathbb{R}$ und $\sigma > 0$ – die Dichtefunktion
    $$f(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$
    Sie spielen in der Statistik eine ganz besonders wichtige Rolle.
- Die gleiche Idee kann in allen Situationen ausgenutzt werden, in denen ein Integral zur Verfügung steht. Wer also auf $\mathbb{R}$ das Lebesgue-Integral kennen gelernt hat, kann integrierbare Dichten zulassen, wer die Integration im $\mathbb{R}^n$ beherrscht, kann leicht Wahrscheinlichkeitsmaße auf den Borelmengen dieses Raumes angeben usw.

Wahrscheinlichkeitstheorie: Grundbegriffe

- Bedingte Wahrscheinlichkeit.
- Was bedeutet „Unabhängigkeit“ für zwei, endlich viele bzw. beliebig viele Ereignisse?
- Zufallsvariable.
- Erwartungswert und Streuung.
- Unabhängigkeit für Zufallsvariable.

Grenzwertsätze

Die Grenzwertsätze besagen, „dass der Zufallseinfluss verschwindet“, wenn sich „viele“ Zufallseinflüsse unabhängig überlagern. Man sollte kennen:

- Die Definitionen „Konvergenz in Wahrscheinlichkeit“, „Konvergenz in Verteilung“, „Fast sichere Konvergenz“.
- Das Wurzel-n-Gesetz.
- Die Lemmata von Borel-Cantelli.
- Die Tschebyscheff-Ungleichung und die Markov-Ungleichung.
- Das schwache Gesetz der großen Zahlen.
- Das starke Gesetz der großen Zahlen.
- Den zentralen Grenzwertsatz.

6.3 Die Normalverteilung: weitere Ergebnisse

Die Normalverteilung wird im Folgenden eine wichtige Rolle spielen. In diesem Abschnitt sammeln wir einige Vorbereitungen.

Es ist offensichtlich, dass die Dichtefunktion der Normalverteilung „schnell abfällt“, und deswegen sind große Abweichungen vom Erwartungswert „unwahrscheinlich“. Das soll nun quantifiziert werden:

Lemma 6.3.1. *Die Zufallsvariable X sei $N(0,1)$-verteilt. Dann gilt für $t > 0$:*
(i) $\mathbb{P}(|X| \geq t) \leq e^{-t^2/2}$.
(ii) $\mathbb{P}(|X| \geq t) \leq \big(\sqrt{2/\pi}\big)e^{-t^2/2}/t$.
(iii) $\mathbb{P}(|X| \geq t) \geq \big(\sqrt{2/\pi}\big)\big(1/t - 1/t^3\big)e^{-t^2/2}$.
(iv) $\mathbb{P}(|X| \geq t) \geq \big(1 - t\sqrt{2/\pi}\big)e^{-t^2/2}$.

Beweis: Aus Symmetriegründen hat $|X|$ die Dichtefunktion

$$\frac{2}{\sqrt{2\pi}}e^{-x^2/2}\left(= \sqrt{\frac{2}{\pi}}e^{-x^2/2}\right),$$

und deswegen gilt

$$\mathbb{P}(|X| \geq t) = \sqrt{\frac{2}{\pi}}\int_t^\infty e^{-x^2/2}\,dx.$$

(i) Es ist

$$\begin{aligned}
\sqrt{\frac{2}{\pi}}\int_t^\infty e^{-x^2/2}\,dx &= \sqrt{\frac{2}{\pi}}\int_0^\infty e^{-(x+t)^2/2}\,dx \\
&= \sqrt{\frac{2}{\pi}}e^{-t^2/2}\int_0^\infty e^{-x^2/2}e^{-xt}\,dx \\
&\leq \sqrt{\frac{2}{\pi}}e^{-t^2/2}\int_0^\infty e^{-x^2/2}\,dx \\
&= e^{-t^2/2}.
\end{aligned}$$

Hier haben wir ausgenutzt, dass $e^{-xt} \leq 1$ und dass $\frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-x^2/2}\,dx = 1/2$.
(ii) Für $x \geq t$ ist $1 \leq x/t$, und deswegen gilt

$$\begin{aligned}\sqrt{\frac{2}{\pi}} \int_t^\infty e^{-x^2/2}\,dx &\leq \sqrt{\frac{2}{\pi}} \frac{1}{t} \int_t^\infty x e^{-x^2/2}\,dx \\ &= \sqrt{\frac{2}{\pi}} \frac{1}{t} \left[-e^{-x^2/2}\right]_t^\infty \\ &= \sqrt{\frac{2}{\pi}} \frac{1}{t} e^{-t^2/2}.\end{aligned}$$

(iii) Vorbereitend bemerken wir:

- $e^{-x} \geq 1 - x$ für $x \geq 0$.
  (Bei 0 stimmen beide Funktionen überein, und $(e^{-x})' = -e^{-x} \geq -1 = (1-x)'$). Insbesondere ist $e^{-x^2/2} \geq 1 - x^2/2$.
- $\int_0^\infty e^{-xt}\,dx = 1/t$ (klar).
- $\int_0^\infty x^2 e^{-xt}\,dx = 2/t^3$ (partielle Integration).

Wir wissen schon, dass

$$\sqrt{\frac{2}{\pi}} \int_t^\infty e^{-x^2/2}\,dx = \sqrt{\frac{2}{\pi}} e^{-t^2/2} \int_0^\infty e^{-x^2/2} e^{-xt}\,dx,$$

und es geht weiter mit

$$\begin{aligned}\int_0^\infty e^{-x^2/2} e^{-xt}\,dx &\geq \int_0^\infty (1 - x^2/2) e^{-xt}\,dx \\ &= 1/t - 1/t^3.\end{aligned}$$

Das beweist (iii).
(iv) Man arbeitet diesmal mit der Abschätzung $e^{-tx} \geq 1 - tx$ und erhält

$$\begin{aligned}\sqrt{\frac{2}{\pi}} \int_t^\infty e^{-x^2/2}\,dx &= \sqrt{\frac{2}{\pi}} e^{-t^2/2} \int_0^\infty e^{-x^2/2} e^{-xt}\,dx \\ &\geq \sqrt{\frac{2}{\pi}} e^{-t^2/2} \int_0^\infty e^{-x^2/2} (1 - xt)\,dx \\ &= e^{-t^2/2} \left(1 - t\sqrt{\frac{2}{\pi}} \int_0^\infty x e^{-x^2/2}\,dx\right) \\ &= \left(1 - t\sqrt{\frac{2}{\pi}}\right) e^{-t^2/2}.\end{aligned}$$

□

Bemerkung: Je nachdem, ob t in der Nähe der 0 liegt oder nicht, wird man sich für eine der Varianten entscheiden.

Nun betrachten wir Funktionen einer $N(0,1)$-verteilten Zufallsvariablen. Bekanntlich gilt (s. z. B. [Be], Satz 3.3.5)

> Für eine reellwertige Zufallsvariable X habe $\mathbb{P}_X$ eine Dichtefunktion h. Dann gilt für $\phi : \mathbb{R} \to \mathbb{R}$:
>
> $$\mathbb{E}\big(\phi(X)\big) = \int_{\mathbb{R}} \phi(x)h(x)\,dx.$$

Damit zeigen wir:

Satz 6.3.2. *X sei $N(0,1)$-verteilt. Für jedes $\theta \in \mathbb{R}$ ist dann*

$$\mathbb{E}(e^{\theta X}) = e^{\theta^2/2}.$$

Beweis: Wir müssen

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

auswerten. Durch quadratische Ergänzung wird daraus

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\theta x} e^{-x^2/2}\,dx &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\theta x - x^2/2 + \theta^2/2 - \theta^2/2}\,dx \\
&= \frac{1}{\sqrt{2\pi}} e^{\theta^2/2} \int_{\mathbb{R}} e^{-(x+\theta)^2/2}\,dx \\
&= \frac{1}{\sqrt{2\pi}} e^{\theta^2/2} \int_{\mathbb{R}} e^{-x^2/2}\,dx \\
&= e^{\theta^2/2}.
\end{aligned}$$

□

Das hat eine interessante Konsequenz:

Korollar 6.3.3. *X sei $N(0,1)$-verteilt. Dann gilt für $n \in \mathbb{N}$:*

$$\mathbb{E}(X^{2n}) = \frac{(2n)!}{2^n\, n!}.$$

Beweis: Man muss nur $\mathbb{E}(e^{\theta X})$ auf zwei verschiedene Weisen berechnen: einmal durch den vorigen Satz, und dann durch gliedweises Integrieren der Entwicklung von $e^{\theta X}$. Durch Koeffizientenvergleich (Koeffizient bei θ^{2n}) folgt

$$\frac{\mathbb{E}(X^{2n})}{(2n)!} = \frac{1}{2^n n!}.$$

□

6.4 Große Abweichungen

Durch die Gesetze der großen Zahlen wird in der elementaren Stochastik präzisiert, inwiefern größere Abweichungen vom Erwartungswert unwahrscheinlich sind, wenn man Mittelwerte unabhängiger Abfragen betrachtet. Was aber lässt sich für eine einzelne Zufallsvariable aussagen? Für die Normalverteilung gab es schon recht scharfe Ergebnisse in Lemma 6.3.1, wir wollen hier einen Ansatz studieren, mit dem man alle Zufallsvariablen behandeln kann.

Sei X eine reellwertige Zufallsvariable. Wir wollen ihr eine im folgenden wichtige Funktion $C_X : \mathbb{R} \to \mathbb{R}$ zuordnen, die *Kumulantenerzeugende Funktion*:

Definition 6.4.1. *$C_X : \mathbb{R} \to \mathbb{R}$ ist durch*

$$C_X(\theta) := \log \mathbb{E}(e^{\theta X})$$

definiert. (Wir werden stets voraussetzen, dass die hier betrachteten Erwartungswerte existieren).

Bemerkungen:

1. Ist X gleich der Konstanten r, so ist $C_X(\theta) = \theta r$.

2. Für $N(0,1)$-verteilte X haben wir schon gerechnet: $C_X(\theta) = \log(e^{\theta^2/2}) = \theta^2/2$ (Satz 6.3.2).

3. Sind X, Y unabhängig, so ist $C_{X+Y} = C_X + C_Y$. Dabei haben wir ausgenutzt, dass mit X, Y auch $e^{\theta X}, e^{\theta Y}$ unabhängig sind und dass deswegen der Erwartungswert mit Produkten vertauscht.

4. Klar ist, dass stets $C_{\alpha X}(\theta) = C_X(\alpha\theta)$ gilt.

5. Fixiere ein $\alpha > 0$, wir definieren X als diejenige Zufallsvariable, für die $\mathbb{P}_X = (\delta_{-\alpha} + \delta_\alpha)/2$ gilt: Mit jeweils Wahrscheinlichkeit 0.5 wird α oder $-\alpha$ erzeugt. Es ist dann $\mathbb{E}(e^{\theta X}) = \cosh(\alpha\theta)$, für „große“ positive θ ist also $C_X(\theta) \approx \theta\alpha - \log 2$. Vergleicht man nur Zufallsvariable mit Erwartungswert 0, so scheint der Anstieg von C_X ein Maß für die Streuung zu sein.

Hier ist der Satz von Cramér:

Satz 6.4.2. *$X, X_1, \ldots, X_M$ seien Zufallsvariable, und $X_1, \ldots, X_M$ seien unabhängig. Für $t > 0$ gilt dann*

(i) $\mathbb{P}(X \geq t) \leq \exp\big(\inf_{\theta>0} -\theta t + C_X(\theta)\big)$.

(ii) $\mathbb{P}\big(\sum_{i=1,\ldots,M} X_i \geq t\big) \leq \exp\big(\inf_{\theta>0} -\theta t + \sum_i C_{X_i}(\theta)\big)$.

Beweis: (i) Fixiere ein $\theta > 0$. Aufgrund der *Markovungleichung* gilt für jede positive Zufallsvariable Y:

$$\mathbb{P}(Y \geq s) \leq \frac{\mathbb{E}(Y)}{s}.$$

Das nutzen wir wie folgt aus:

$$\begin{aligned}\mathbb{P}(X \geq t) &= \mathbb{P}\big(\exp(\theta X) \geq \exp(\theta t)\big) \\ &\leq e^{-\theta t}\mathbb{E}\big(\exp(\theta X)\big) \\ &= \exp(-\theta t + C_X(\theta)).\end{aligned}$$

Nun muss man nur noch das Infimum über alle θ bilden.
(ii) Hier ist nur (i) mit Bemerkung 3 zu kombinieren. □

Wir testen die Cramér-Ungleichung an einigen Beispielen.
1. X sei konstant gleich r, also $C_X(\theta) = \theta r$. Auf der rechten Seite der Cramérungleichung steht dann

$$\inf_{\theta>0} \theta(r-t),$$

und das ist 0 für $t \leq r$ und $-\infty$ für $t > r$. Die Cramérungleichung sagt also richtig voraus:

- $\mathbb{P}(X \geq t) = 1$ für $t \leq r$.
- $\mathbb{P}(X \geq t) = 0$ für $t > r$.

2. X sei nun $N(0,1)$-verteilt. Dann ist $C_X(\theta) = \theta^2/2$. Für $t > 0$ ist

$$\inf_{\theta>0} \theta t - \theta^2/2 = -t^2/2$$

(mit elementarer Analysis), und damit liefert die Ungleichung

$$\mathbb{P}(X \geq t) \leq e^{-t^2/2}.$$

Das ist etwas schlechter als das Ergebnis in Lemma 6.3.1 (i). Dort wurde die Ungleichung $\mathbb{P}|X| \geq t) \leq e^{-t^2/2}$ gezeigt, hier kommt nur $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/2}$ heraus.

Die Cramérungleichung ist etwas unhandlich. Deswegen zeigen wir noch etwas leichter anwendbare Varianten. Wir beginnen mit einer eher technischen Vorbereitung.

Lemma 6.4.3. *(i)* $\cosh x \leq e^{x^2/2}$ *für alle* x.
(ii) X *sei eine beschränkte Zufallsvariable:* $|X| \leq B$ *fast sicher. Es sei auch* $\mathbb{E}X = 0$. *Dann ist* $C_X(\theta) \leq B^2\theta^2/2$.
(iii) Seien $t, \alpha > 0$. *Dann ist* $\min_{\theta>0} -\theta t + \alpha\theta^2/2 = t^2/(2\alpha)$.

Beweis: (i) Die Potenzreihenentwicklungen von $\cosh x$ und $e^{x^2/2}$ sind

$$\frac{1}{2}\big((1 + x + \frac{x^2}{2!} \pm \cdots) + (1 - x + \frac{x^2}{2!} \pm \cdots)\big) = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots$$

und

$$1 + \frac{x^2}{2} + \frac{x^4}{2^2 2! + \cdots}.$$

Der typische Summand in der der ersten bzw. zweiten Reihe lautet

$$\frac{x^{2n}}{(2n)!} \text{ bzw. } \frac{x^{2n}}{2^n n!}.$$

Die Behauptung folgt dann aus der offensichtlichen Ungleichung $(2n)! \geq 2^n n!$.

(ii) Schreibe X als Konvexkombination von $\pm B$, wobei die Parameter variieren:

$$X = Y(-B) + (1-Y)B, \text{ mit } Y := \frac{B-X}{2B} \in [\,0,1\,].$$

Die Funktion $f : x \mapsto e^{\theta x}$ ist konvex, für jedes $\lambda \in [\,0,1\,]$ ist also

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

In unserem Fall heißt das: punktweise gilt

$$e^{\theta X} \leq Y e^{-\theta B} + (1-Y) e^{\theta B}.$$

Wir integrieren diese Ungleichung, setzen die konkrete Form von Y ein und nutzen aus, dass $\mathbb{E}(X) = 0$ gilt. So folgt

$$\mathbb{E}(e^{\theta X}) \leq \frac{B}{2B} e^{-\theta B} + \frac{B}{2B} e^{\theta B} = \cosh(\theta B).$$

Wegen (i) kann das mit $\exp\big((\theta B)^2/2\big)$ weiter abgeschätzt werden, und das beweist (wegen der Monotonie des Logarithmus) die Behauptung.

(iii) Das ist mit elementarer Analysis klar. □

Wir zeigen nun die *Hoeffdingungleichung*:

Satz 6.4.4. *$X_1, \ldots, X_M$ seien beschränkte und unabhängige Zufallsvariable mit Erwartungswert 0, und zwar sei jeweils $|X_l| \leq B_l$. Für $t > 0$ ist dann*
(i) $\mathbb{P}\big(\sum_l X_l \geq t\big) \leq \exp\big(-t^2/(2\sum_l B_l^2)\big)$.
(ii) $\mathbb{P}\big(|\sum_l X_l| \geq t\big) \leq 2\exp\big(-t^2/(2\sum_l B_l^2)\big)$.

Beweis: (i) Es ist

$$C_{X_1+\cdots+X_M} = C_{X_1} + \cdots + C_{X_M} \leq \frac{\theta^2}{2}(B_1^2 + \cdots + B_M^2)$$

aufgrund der obigen Bemerkung 3 und Lemma 6.4.3 (ii). Die Aussage folgt nun aus der Cramérschen Ungleichung und 6.4.3 (iii) (mt $\alpha = \sum_l B_l^2$).

(ii) Man wende (i) für die $-X_l$ an, beachte, dass $\{|Y| \geq t\} = \{Y \geq t\} \cup \{-Y \geq t\}$ und erinnere sich an $\mathbb{P}(E \cup F) \leq \mathbb{P}(E) + \mathbb{P}(F)$. □

Die Hoeffdingungleichung hat ein interessantes Korollar. Man stelle sich vor, dass eine Münze geworfen wird, die mit ± 1 beschriftet ist. Unabhängige Abfragen sollen durch Zufallsvariable $\varepsilon_1, \ldots, \varepsilon_M$ modelliert werden (eine so genannte *Rademacherfolge*). Wir geben uns Zahlen $a_1, \ldots, a_M$ vor und fragen, wie groß

$$a_1\varepsilon_1 + a_2\varepsilon_2 + \cdots + a_M\varepsilon_M$$

wohl sein wird. Hier ist die Antwort:

Korollar 6.4.5. *Für jedes $u > 0$ ist*

$$\mathbb{P}\left(|\sum_l a_l\varepsilon_l| \geq u\sqrt{\sum_l a_l^2} \right) \leq 2e^{-u^2/2}.$$

Beweis: Wir wenden die Hoeffdingungleichung mit $X_l := a_l\varepsilon_l$ an. Es ist dann $B_l^2 = a_l^2$, und man muss nur noch $t = u\sqrt{\sum_l a_l^2}$ setzen. □

Überraschenderweise spielt hier also die l^2-Norm des Vektors (a_l) eine Rolle. Als Spezialfall betrachten wir $a_1 = \cdots = a_M = 1$. Dann erhalten wir das folgende Ergebnis:

$|\varepsilon_1 + \cdots + \varepsilon_M|$ wird nicht wesentlich größer als $\sqrt{M}$. Genauer:

$$\mathbb{P}\big(|\varepsilon_1 + \cdots + \varepsilon_M| \geq u\sqrt{M}\big) \leq 2e^{-u^2/2}.$$

So kommt es zum Beispiel nur mit einer Wahrscheinlichkeit von höchstens $e^{-3^2/2} \approx 0.11\ldots$ vor, dass $\varepsilon_1 + \cdots + \varepsilon_{10000}$ außerhalb von $[-300, 300]$ liegt.

Das nächste Ziel ist der Beweis der *Bernstein-Ungleichung*. Einige Vorbereitungen sammeln wir in

Lemma 6.4.6. *(i) X sei eine Zufallsvariable mit $\mathbb{E}X = 0$. Für geeignete Zahlen R, σ soll*

$$\mathbb{E}\big(|X|^n\big) \leq n!R^{n-2}\sigma^2/2$$

für alle $n \in N$ gelten (wobei meist $\sigma^2 := \mathbb{E}X^2$). Für die $\theta > 0$ mit $R\theta < 1$ gilt dann

$$C_X(\theta) \leq \frac{\theta^2\sigma^2}{2}\frac{1}{1 - R\theta}.$$

(ii) Für $t, \alpha > 0$ gilt

$$\inf_{0<R\theta<1}\left(-\theta t + \frac{\theta^2\alpha^2}{2(1 - R\theta)}\right) \leq \frac{-t^2/2}{\alpha^2 + Rt}$$

Beweis: (i) Wir beginnen mit der Berechnung des Erwartungswerts von $e^{\theta X}$. Wir nutzen aus, dass $\mathbb{E}(X) = 0$ und dass (wegen $X^n \leq |X|^n$) $\mathbb{E}(X^n) \leq \mathbb{E}(|X|^n)$ gilt.

$$\begin{aligned}
\mathbb{E}e^{\theta X} &= 1 + \theta\mathbb{E}X + \frac{\theta^2\sigma^2}{2}\big(\sum_{n\geq 2}\frac{\theta^{n-2}}{n!(\sigma^2/2)}\mathbb{E}(X^n)\big) \\
&= 1 + \frac{\theta^2\sigma^2}{2}\big(\sum_{n\geq 2}\frac{\theta^{n-2}}{n!(\sigma^2/2)}\mathbb{E}(X^n)\big) \\
&\leq 1 + \frac{\theta^2\sigma^2}{2}\big(\sum_{n\geq 2}\frac{\theta^{n-2}}{n!(\sigma^2/2)}\mathbb{E}(|X|^n)\big) \\
&\leq 1 + \frac{\theta^2\sigma^2}{2}\sum_{n\geq 2}(R\theta)^{n-2} \\
&= 1 + \frac{\theta^2\sigma^2}{2}\frac{1}{1-R\theta} \\
&\leq \exp\big(\frac{\theta^2\sigma^2}{2}\frac{1}{1-R\theta}\big).
\end{aligned}$$

Durch Logartihmieren folgt die Behauptung.
(ii) Setze $\theta_0 := t/(\alpha^2 + Rt)$. Dann ist $R\theta_0 \in]\,0, 1\,[$, also

$$\begin{aligned}
\inf_{0<R\theta<1}\frac{\theta^2\alpha^2}{2(1-R\theta)} &\leq \frac{\theta_0^2\alpha^2}{2(1-R\theta_0)} \\
&= \frac{t^2\alpha^2}{2(\alpha^2+Rt)^2}\frac{1}{1-\frac{Rt}{\alpha^2+Rt}} - \frac{t^2}{\alpha^2+Rt} \\
&= \frac{-t^2/2}{\alpha^2+Rt}.
\end{aligned}$$

□

Wir zeigen nun

Satz 6.4.7. *(Bernsteinungleichung) $X_1, \ldots, X_M$ seien unabhängige Zufallsvariable mit $\mathbb{E}(X_l) = 0$ und $\sigma_l := \mathbb{E}(X_l)^2$. Es gebe $R > 0$, so dass für alle n*

$$\mathbb{E}\big(|X_l|^n\big) \leq n!R^{n-2}\sigma_l^2/2$$

gilt. Wir setzen $\sigma^2 := \sigma_1^2 + \cdots + \sigma_M^2$. (Das ist die Varianz von $X_1 + \cdots + X_M$). Für $t > 0$ ist dann
(i)

$$\mathbb{P}\big(\sum_l X_l \geq t\big) \leq \exp\big(\frac{-t^2/2}{\sigma^2+Rt}\big).$$

(ii)

$$\mathbb{P}\big(|\sum_l X_l| \geq t\big) \leq 2\exp\big(\frac{-t^2/2}{\sigma^2+Rt}\big).$$

Beweis: (i) Durch Kombination von Bemerkung 3 zu Beginn dieses Abschnitts und Lemma 6.4.6 folgt

$$C_{\sum X_l}(\theta) \leq \frac{\theta^2\sigma^2}{2}\frac{1}{1-R\theta}.$$

Damit schließen wir aus der Cramérschen Ungleichung:

$$\begin{aligned}
\mathbb{P}\big(\sum_l X_l \geq t\big) &\leq \inf_{\theta>0} \exp\big(-\theta t + C_{\sum_l X_l}(\theta)\big) \\
&\leq \inf_{0<R\theta<1} \exp\big(-\theta t + C_{\sum_l X_l}(\theta)\big) \\
&\leq \inf_{0<R\theta<1} \exp\big(-\theta t + \frac{\theta^2\sigma^2}{2}\frac{1}{1-R\theta}\big) \\
&\leq \exp\big(-\frac{t^2/2}{\sigma^2+Rt}\big)
\end{aligned}$$

(ii) Das folgt sofort aus (i), wenn man (i) auch noch für die $-X_l$ ausnutzt. □

Ein Nachtrag: Es gibt verschiedene Varianten der Bernsteinungleichung. Manchmal wird auch das folgende Ergebnis als Bernsteinungleichung bezeichnet:

> $X_1, \ldots, X_M$ seien unabhängige Zufallsvariable mit $\mathbb{E}(X_l) = 0$, alle X_i seien durch eine Konstante B und alle Varianzen $\sigma_i^2(X_i)$ durch eine Zahl σ^2 beschränkt. Dann gilt für alle $t > 0$:
>
> $$\mathbb{P}\bigg(\frac{1}{M}\sum_i X_i \geq \sqrt{\frac{2\sigma^2 t}{M}} + \frac{2Bt}{3M}\bigg) \leq e^{-t}.$$

6.5 Große Abweichungen bei subexponentiellen Zufallsvariablen

Im nächsten Kapitel werden wir Abschätzungen für Zufallsvariable benötigen, die „sehr schnell“ abfallen. Genauer:

Definition 6.5.1. *X sei eine reellwertige Zufallsvariable.*
(i) X heißt subexponentiell*, wenn es $\beta, \kappa > 0$ so gibt, dass*

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

für alle $t > 0$ gilt.
(ii) X heißt subgaußsch*, wenn es $\beta, \kappa > 0$ so gibt, dass*

$$\mathbb{P}(|X| \geq t) \leq \beta e^{-\kappa t^2}$$

für alle $t > 0$ gilt.

Der Name rührt daher, dass exponentialverteilte X subexponentiell und normalverteilte subgaußsch sind.

Wir benötigen einige allgemeine Eigenschaften solcher Zufallsvariablen. Insbesondere wollen wir zeigen, dass wir für sie die Bernsteinungleichung anwenden können.

Dazu brauchen wir eine Vorbereitung. Zur Motivation betrachten wir eine $\mathbb{N}$-wertige Zufallsvariable X. Dann ist doch

$$\begin{aligned}\mathbb{E}(X) &= \sum_{n\in\mathbb{N}} n\mathbb{P}(X=n)\\ &= \sum_{n\in\mathbb{N}} \mathbb{P}(X\geq n).\end{aligned}$$

Die allgemeine Variante liest sich so:

Lemma 6.5.2. *X sei $\mathbb{R}^+$-wertig. Dann ist*

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X\geq x)\,dx.$$

Beweis: Wir brauchen das Ergebnis nur für solche X, die eine stetige Dichte ϕ haben, und wir wollen auch annehmen, dass X nur Werte in $[\,0,b\,]$ annimmt[1)].

Es gibt also ein ϕ, so dass $\mathbb{P}(X\geq x) = \int_x^b \phi(t)\,dt =: A(x)$ für alle x gilt. Wir setzen $\psi(x) := \int_0^x \phi(t)\,dt = 1 - A(x)$. Man beachte, dass $\psi' = \phi$ gilt.
Zum Beweis des Lemmas wenden wir nun partielle Integration an:

$$\begin{aligned}\mathbb{E}(X) &= \int_0^b t\phi(t)\,dt\\ &= t\psi|_0^b - \int_0^b \psi(t)\,dt\\ &= b - \int_0^b (1-A(t))\,dt\\ &= \int_0^b A(t)\,dt.\end{aligned}$$

□

Durch Substitution $s = t^n$ erhält man daraus

Lemma 6.5.3. $\mathbb{E}(|X|^n) = n\int_0^\infty \mathbb{P}(|X|\geq t)t^{n-1}\,dt$.

Und *damit* lassen sich subexponenielle Zufallsvariable weiter abschätzen. Vorbereitend zeigen wir:

[1)] Der allgemeine Fall kann dann durch ein Approximationsargument bewiesen werden.

Lemma 6.5.4. *(i) Quadrate von subgaußschen Zufallsvariablen sind subexponentiell.*

(ii) Konstante Zufallsvariable sind subgaußsch und subexponentiell.

(iii) Summen und Vielfache subexponentieller (bzw. subgaußscher) Zufallsvariable sind subexponentiell (bzw. subgaußsch).

(iv) Für alle $n \in \mathbb{N}$ ist $\int_0^\infty e^{-x}x^{n-1}\,dx = (n-1)!$.

Beweis: (i) bis (iii): Übungsaufgabe.

(iv) Allgemeiner definiert man $\Gamma(t) := \int_0^\infty e^{-t}x^{t-1}\,dx$, und „man weiß“, dass $\Gamma(n) = (n-1)!$. (S. z.B. Abschnit 6.3 in meinem Buch zur Analysis 2.) Es geht aber auch ganz elementar durch Induktion mit Hilfe partieller Integration (Setze $A_n := \int_0^\infty e^{-x}x^{n-1}\,dx$ und zeige $A_1 = 1$ sowie $A_{n+1} = nA_n$).

□

Wir wollen die Bernsteinungleichung für subexponentielle Zufallsvariable anwenden. Dazu müssen wir eine Ungleichung für $\mathbb{E}|X|^n$ aus einer Ungleichung für $\mathbb{P}(|X| \geq t)$ herleiten. (Die Situation ist also gerade umgekehrt wie bei der Markovungleichung.)

Satz 6.5.5. *$X_1, \ldots, X_M$ seien unabhängige subexponentiele Zufallsvariable mit Erwartungswert 0. Wir wählen $\beta, \kappa > 0$ so, dass*

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

für alle $t > 0$ und alle l gilt. Für jedes $t > 0$ ist dann

$$\mathbb{P}\Big(|\sum_l X_l| \geq t\Big) \leq 2\exp\Big(\frac{(-\kappa t)^2/2}{2\beta M + \kappa t}\Big).$$

Beweis: Mit Hilfe von Lemma 6.5.3 schließen wir so, wobei wir $s = \kappa t$ substituieren:

$$\begin{aligned}
\mathbb{E}|X_l|^n &= n\int_0^\infty \mathbb{P}(|X_l| \geq t)t^{n-1}\,dt \\
&\leq \beta n\int_0^\infty e^{-\kappa t}t^{n-1}\,dt \\
&= \frac{\beta n}{\kappa^n}\int_0^\infty e^{-s}s^{n-1}\,ds \\
&= \frac{n!\beta}{\kappa^n} \\
&= n!\kappa^{-(n-2)}\big(\frac{2\beta}{2\kappa^2}\big).
\end{aligned}$$

Die Voraussetzungen der Bernsteinungleichung sind also erfüllt, wenn man $R = 1/\kappa$ setzt und mit $\sigma_l^2 = 2\beta/\kappa^2$ arbeitet[2)].

[2)] Achtung: σ_l^2 muss nicht die Varianz von X_l sein.

Beachte noch:

$$\begin{aligned}\frac{t^2/2}{\sigma^2+Rt} &= \frac{t^2/2}{2M\beta/\kappa^2+(1/\kappa)t}\\ &= \frac{(\kappa t)^2/2}{2M\beta+\kappa t}.\end{aligned}$$

□

Kapitel 7

Rekonstruktion mit Zufallsmatrizen

In den ersten Kapiteln haben wir gesehen, wie man mit „geeigneten“ Matrizen schwach besetzte Vektoren rekonstruieren kann. Es gab auch gute Gründe zu glauben, dass man mit Zufallsmatrizen „gute“ Matrizen finden könnte. (Vgl. die Einleitung zu Kapitel 6). Dieses Ziel soll in diesem Kapitel unter Verwendung der Ergebnisse aus Kapitel 6 verwirklicht werden.

7.1 Zufallsmatrizen, die Strategie

Eine Zufallsmatrix entsteht dadurch, dass für jeden Eintrag eine Zufallsabfrage durchgeführt wird. In der Regel werden unabhängige Zufallsvariable verwendet, und hier werden wir mit $N(0,1)$-Variablen arbeiten. Wir haben also eine Familie $\xi_{i,j}$ $(i = 1, \dots, m, j = 1, \dots, N)$ von unabhängigen $N(0,1)$-Variablen vor uns, und A ist als $(\xi_{i,j}(\omega))_{i,j}$ entstanden.

Wir zeigen zunächst ein vorbereitendes Ergebnis. Mit $Y = (Y_1, \dots, Y_N)^\top$ bezeichnen wir eine typische Zeile von A, also einen $1 \times N$-Zufallsvektor.

Lemma 7.1.1. *(i) Für jedes $x \in \mathbb{R}^N$ ist $\mathbb{E}(\langle Y, x\rangle) = 0$ und $\mathbb{E}|\langle Y, x\rangle|^2 = ||x||_2^2$.*
(ii) Für jedes x ist $\mathbb{E}||Ax||_2^2 = m||x||_2^2$.

Beweis: (i)

$$\mathbb{E}(\langle Y, x\rangle) = \sum_j x_j \mathbb{E}(Y_j) = 0,$$

da $\mathbb{E}(Y_j) = 0$.

$$\begin{aligned}\mathbb{E}|\langle Y,x\rangle|^2 &= \sum_{l,l'} x_l x_{l'} \mathbb{E}(Y_l Y_{l'}) \\ &= \sum_l x_l^2 \\ &= ||x||_2^2.\end{aligned}$$

(ii) Das folgt sofort aus (i). □

Unser Ziel ist es zu erreichen, dass die Fastisometriekonstante δ_s „mit hoher Wahrscheinlichkeit“ klein wird, um die Ergebnisse aus Abschnitt 5 anwenden zu können. Aufgrund des vorigen Lemmas sollte man es nicht mit A, sondern mit $\tilde{A} := \frac{1}{\sqrt{m}} A$ versuchen. Unsere Strategie wird die folgende sein:

Schritt 1: Man gebe $\delta, \varepsilon > 0$ vor. Wenn m groß genug ist, gilt mit mindestens Wahrscheinlichkeit $1-\varepsilon$ und beliebige $x \in \mathbb{R}^N$:

$$\left| ||\tilde{A}x||_2^2 - ||x||_2^2 \right| \leq \delta ||x||_2^2.$$

Anders ausgedrückt: In jeder festen Richtiung ist $\tilde{A}$ mit hoher Wahrscheinlichkeit fast isometrisch.

Schritt 2: Sei $S \subset \{1,\ldots,N\}$ eine feste s-elementige Menge. Sind dann $\delta, \varepsilon > 0$, so kann man „nicht zu große“ m finden, so dass bei zufälliger Wahl von $\tilde{A}$ die Ungleichung

$$\left| ||\tilde{A}x||_2^2 - ||x||_2^2 \right| \leq \delta ||x||_2^2$$

mit mindestens Wahrscheinlichkeit $1-\varepsilon$ für alle x mit Träger in S gilt.

Schritt 3: Dann kommt das Finale. Man wünsche sich $\varepsilon, \delta > 0$, und dann kann man garantieren, dass bei zufälligem $\tilde{A}$ mit „genügend großem“ m die Bedingung $\delta_s \leq \delta$ mit mindestens Wahrscheinlichkeit $1-\varepsilon$ erfüllt ist. Insbesondere darf man sich aufgrund der Ergebnisse von Abschnitt 5.3 wünschen, dass durch das Verfahren der l^1-Minimierung schwach besetzte Lösungsvektoren gefunden werden.

7.2 Fastisometrie: ein einziger Vektor

In diesem Abschnitt soll *Schritt 1* behandelt werden. Das Ziel ist bescheiden: Eine Zufallsmatrix $\tilde{A}$ ist wie in Abschnitt 7.1 gegeben, $x \in \mathbb{R}^N$ ist vorgelegt und t ist positiv. Mit welcher Wahrscheinlichkeit kann man garantieren, dass $\left| ||\tilde{A}x||_2^2 - ||x||_2^2 \right| \leq t||x||^2$ gilt?

Hier eine Vorbereitung:

Lemma 7.2.1. *Sei Y_l eine Zeile von A ($l = 1,\ldots,N$) und $x \in \mathbb{R}^N$ ein Vektor mit $||x||_2^2 = 1$. Dann gilt:*

(i) $\langle Y_l, x\rangle$ ist $N(0,1)$ verteilt.

(ii) Setzt man $Z_l := |\langle Y_l, x\rangle|^2 - 1$*, so ist* Z_l *subexponentiell mit Erwartungswert* 0*. Genauer gilt*

$$\mathbb{P}(|Z_l| \geq u) \leq \beta_0 e^{-u/2};$$

dabei ist $\beta_0 := \sqrt{e} + 1/\sqrt{e} \approx 2.6$*.*

Beweis: (i) Das folgt aus allgemeinen Eigenschaften von Normalverteilungen: Sind X, Y unabhängig und $N(0, \sigma_1^2)$- bzw. $N(0, \sigma_2^2)$-verteilt, so ist $X + Y$ $N(0, \sigma_1^2 + \sigma_2^2)$-verteilt.

(ii) Es ist doch (für $b \geq 0$) $|a| \geq b$ genau dann, wenn $a \geq b$ oder $-a \geq b$. Hier heißt das:

$$\{|Z_l| \geq u\} = \{Z_l \geq u\} \cup \{-Z_l \geq u\} =: M_1 \cup M_2.$$

Und dann ist $\mathbb{P}(|Z_l| \geq u) \leq \mathbb{P}(M_1) + \mathbb{P}(M_2)$.

Nun ist wegen Lemma 6.3.1

$$\begin{aligned} \mathbb{P}(Z_l \geq u) &= \mathbb{P}(\langle Y_l, x\rangle^2 \geq 1 + u) \\ &= \mathbb{P}(|\langle Y_l, x\rangle| \geq \sqrt{1+u}) \\ &\leq e^{-(1+u)/2} \\ &= \frac{1}{\sqrt{e}} e^{-u/2}. \end{aligned}$$

Und für $u \geq 0$ gilt:

$$\begin{aligned} \mathbb{P}(M_2) &= \mathbb{P}(-(\langle Y_l, x\rangle^2 - 1) \geq u) \\ &= \mathbb{P}(\langle Y_l, x\rangle^2 \leq 1 - u) \\ &\leq \chi_{[\,0,1\,]}(u) \\ &\leq \sqrt{e} e^{-u/2}; \end{aligned}$$

Dabei bezeichnet $\chi_{[\,0,1\,]}$ die charakteristische Funktion des Einheitsintervalls. □

Hier ist das erste Hauptergebnis:

Satz 7.2.2. *Es ist*

$$\mathbb{P}\big(\big|||\tilde{A}x||_2^2 - ||x||_2^2\big| \geq t||x||^2\big) \leq 2\exp(-\tilde{c}t^2 m).$$

Dabei kann $\tilde{c}$ *als* $1/(16\beta_0 + 4) \approx 1/20$ *gewählt werden.*

D. h.: Ist m *groß genug, so wird* $\tilde{A}$ *in beliebigen Richtungen mit hoher Wahrscheinlichkeit fastisometrisch sein*[1)]*.*

Beweis: Fixiere x mit (o.B.d.A.) $||x||_2^2 = 1$, erzeuge eine Zufallsmatrix A und definiere die Y_l, Z_l wie vorstehend. Für die Z_l sind wegen des vorstehenden Lemmas die Voraussetzungen von Satz 6.5.5 erfüllt: Sie sind unabhängig, der

[1)] Achtung: Das heißt nicht, dass $\tilde{A}$ mit hoher Wahrscheinlichkeit fastisometrisch ist, denn die „Versagermengen“ können für verschiedenen Richtungen verschieden sein.

Erwartungswert ist Null, und mit $\beta = \beta_0$, $\kappa = 1/2$ gilt $\mathbb{P}(|Z_l| \geq t) \leq \beta e^{-\kappa t}$. Satz 6.5.5 garantiert dann

$$\mathbb{P}\Big(|\sum_l Z_l| \geq u\Big) \leq 2\exp\Big(\frac{-(\kappa u)^2/2}{2\beta m + \kappa u}\Big).$$

Beachte nun, dass

$$\begin{aligned} ||\tilde{A}x||_2^2 - ||x||_2^2 &= \frac{||Ax||_2^2}{m} - ||x||_2^2 \\ &= \frac{\sum_l Z_l}{m}. \end{aligned}$$

Und deswegen ist

$$\begin{aligned} \mathbb{P}\big(\big|||\tilde{A}x||_2^2 - ||x||_2^2\big| \geq t||x||^2\big) &= \mathbb{P}\big(|\sum_l Z_l| \geq tm\big) \\ &\leq 2\exp\big(\frac{-(\kappa tm)^2/2}{2\beta m + \kappa tm}\big) \\ &\leq 2\exp(-\check{c}t^2 m). \end{aligned}$$

Hier wurde ausgenutzt, dass $t \in \,]\,0, 1\,[$ gilt. □

7.3 Fastisometrie: ein s-dimensionaler Unterraum

Wie kann man Fastisometrie von einzelnen Richtungen auf Unterräume „liften“. Genauer: Was folgt daraus. dass man für eine Matrix C weiß, dass

$$\big|||Cx||^2 - ||x||^2\big|$$

für die x in einer Teilmenge des Raumes klein ist?

Diese Frage wird durch die folgenden Lemmata beantwortet werden. Zunächst führen wir einen neuen Begriff ein:

Definition 7.3.1. *Sei D eine Teilmenge eines metrischen Raumes (M, d), und η sei positiv. Ein* η-Netz für *D ist dann eine Teilmenge Δ von M, so dass zu jedem $x \in D$ ein $y \in \Delta$ mit $d(x, y) \leq \eta$ existiert.*

Sei $\eta > 0$. Wie viele Punkte muss man sich in der Einheitskugel eines n-dimensionalen Raumes aussuchen, um ein η-Netz zu erhalten?

Lemma 7.3.2. *Sei X ein s-dimenionaler normierter Raum, die Einheitskugel werde mit B bezeichnet. Zu $\eta > 0$ gibt es dann ein η-Netz in X für B mit höchstens $(1 + 2/\eta)^s$ Elementen.*

Beweis: Für die Kugeln $B(x,\alpha)$ (um x, mit Radius α) ist das euklidische Volumen das α^s-fache des Volumens V von B. Wir wählen eine maximale Teilmenge $x_1,\ldots,x_r$ von B mit der Eigenschaft, dass $||x_i - x_j|| > \eta$ für $i \neq j$ ist. Dann gilt:

- $\{x_1,\ldots,x_r\}$ ist ein η-Netz für B. (Das ist klar.)
- Die r Kugeln $B(x_i,\eta/2)$ sind disjunkt und liegen in $B(0,1+\eta/2)$. Deswegen ist das Volumen von $\bigcup_i B(x_i,\eta/2)$ höchstens gleich dem Volumen von $B(0,1+\eta/2)$.

  Aufgrund der Vorbemerkung heißt das:

$$r\left(\frac{\eta}{2}\right)^s V \leq \left(1+\frac{\eta}{2}\right)^s V.$$

  Und daraus folgt sofort $r \leq (1+2/\eta)^s$. □

Lemma 7.3.3. *$B \in \mathbb{R}^{s\times s}$ sei selbstadjungiert. Dann ist*

$$||B||_{2\to 2} = \max_{||x||=1} |\langle Bx, x\rangle| =: K.$$

Beweis: $K \leq ||B||_{2\to 2}$ ist wegen der Cauchy-Schwarz-Ungleichung klar. Für normierte x, y gilt (wegen $|\langle Bz, z,\rangle| \leq K||z||^2$ und der Parallelogrammidentität $||x+y||^2 + ||x-y||^2 = 2(||x||^2 + ||y||^2)$)

$$\begin{aligned} 4\langle Bx, y\rangle &= \langle B(x+y), x+y\rangle - \langle B(x-y), x-y\rangle \\ &\leq K(||x+y||^2 + ||x-y||^2) \\ &= 2K(||x||^2 + ||y||^2) \\ &= 4K. \end{aligned}$$

Also ist $|\langle Bx, y\rangle| \leq K$, und daraus folgt sofort, dass $||Bx|| \leq K$ für alle normierten x gilt. Das beweist $||B||_{2\to 2} \leq K$.

(Vgl. auch den Beweis von Lemma 5.1.2.) □

Lemma 7.3.4. *Es sei U (für ein $\rho \in \left]0, 0.5\right[$) ein ρ-Netz in der Einheitskugel des $\mathbb{R}^s$ und C eine $m \times s$-Matrix. Gilt dann*

$$\left|||Cu||^2 - ||u||^2\right| \leq t$$

für alle $u \in U$, so ist

$$\left|||Cx||^2 - ||x||^2\right| \leq \frac{t}{1-2\rho}||x||^2$$

für alle $x \in \mathbb{R}^s$.

Beweis: Sei B die selbstadjungierte Matrix $C^\top C - Id$. Die Voraussetzung besagt dann, dass $|\langle Bu, u\rangle| \leq t$ für $u \in U$ gilt. Für beliebige normierte x können wir ein u mit $||x - u|| \leq \rho$ wählen, d.h.

$$\begin{aligned} |\langle Bx, x\rangle| &= |\langle Bu, u\rangle + \langle B(x+u), x-u\rangle| \\ &\leq |\langle Bu, u\rangle| + |\langle B(x+u), x-u\rangle| \\ &\leq t + ||B||_{2\to 2}||x+u||\,||x-u|| \\ &\leq t + 2\rho||B||_{2\to 2}. \end{aligned}$$

Das gilt für alle normierten x, und deswegen ist nach dem vorstehenden Lemma

$$||B||_{2\to 2} \leq t + 2\rho||B||_{2\to 2}.$$

Es folgt $||B||_{2\to 2} \leq t/(1-2\rho)$, was sofort die Behauptung liefert:

$$\begin{aligned} \big|||Cx||^2 - ||x||^2\big| &= |\langle Bx, x\rangle| \\ &\leq ||B||_{2\to 2}||x||^2 \\ &\leq \frac{t}{1-2\rho}||x||^2. \end{aligned}$$

□

7.4 δ_s ist mit hoher Wahrscheinlichkeit klein

Wir beginnen mit der

Definition 7.4.1. *Eine $m \times N$-Matrix A sei durch einen Zufallsprozess entstanden. Wir sagen, dass A die* Konzentrationsungleichung *erfüllt, wenn es eine Konstante $\tilde{c}$ so gibt, dass*

$$\mathbb{P}\big(\big|||Ax||_2^2 - ||x||_2^2\big| \geq t||x||^2\big) \leq 2\exp(-\tilde{c}t^2 m)$$

für alle $x \in \mathbb{R}^N$ gilt.

Wir wissen schon aus Abschnitt 7.2, dass das zum Beispiel (mit einem geeigneten $\tilde{c}$) erfüllt ist, wenn die Einträge von A unabhängig und $N(0,1)$ sind und nachträglich durch $\sqrt{m}$ geteilt wurde. (Das neue A ist also das $\tilde{A}$ des vorigen Abschnitts.)

Sei nun $S \subset \{1, \ldots, N\}$ eine Menge mit s Elementen. Lemma 5.1.2 besagt, dass

$$\big|||Ax||_2^2 - ||x||_2^2\big| \leq \delta||x||_2^2$$

für alle x mit Träger in S genau dann gilt, wenn $||A_S^\top A_S - Id||_{2\to 2} \leq \delta$. Im vorigen Abschnitt haben die die Untersuchung dieser Situation vorbereitet: Interpretiert man die Matrix C aus Lemma 7.3.4 als A_S, so ist die Matrix B gerade die Matrix $A_S^\top A_S - Id$.

Nach dieser Erinnerung gehen wir so vor. Mit einem $\rho > 0$, das erst später fixiert wird, wählen wir ein ρ-Netz U in der Einheitskugel des Unterraums der Vektoren, die ihren Träger in S haben. Das ist mit $(1 + 2/\rho)^s$ Vektoren zu erreichen (Lemma 7.3.2).
Angenommen, der Zufall will es, dass

$$\big|||Au||_2^2 - ||u||_2^2\big| \le t||u||_2^2$$

für alle $u \in U$. Wegen Lemma 7.3.4 ist dann

$$\big|||Ax||_2^2 - ||x||_2^2\big| \le \frac{t}{1-2\rho}||x||_2^2$$

für alle x mit Träger in S, d.h. bzgl. S ist A bis auf den Fehler $t/(1-2\rho)$ isometrisch. Es empfiehlt sich also, mit $t = (1-2\rho)\delta$ weiterzuarbeiten, um am Ende bei δ-Fastisometrien anzukommen.

Doch mit welcher Wahrscheinlichkeit passiert das? Die Wahrscheinlichkeit, dass es mindestens einen „Versager“ unter den $u \in U$ gibt, ist wegen $\mathbb{P}(\bigcup_i B_i) \le \sum_i \mathbb{P}(B_i)$ durch

$$2\left(1+\frac{2}{\rho}\right)^s \exp(-\tilde{c}t^2 m)\left(= 2\big(1+\frac{2}{\rho}\big)^s \exp(-\tilde{c}(1-2\rho)^2\delta^2 m)\right)$$

abschätzbar. Damit folgt: Es ist $||A_S^\top A_S - Id||_{2\to 2} \le \delta$ mit einer Wahrscheinlichkeit von mindestens $1-\varepsilon$, falls

$$\varepsilon \ge 2\left(1+\frac{2}{\rho}\right)^s \exp(-\tilde{c}(1-2\rho)^2\delta^2 m).$$

Aufgelöst nach m heißt das

$$m \ge \frac{1}{\tilde{c}(1-2\rho)^2\delta^2}\big(s\log(1+2/\rho) + \log(2/\varepsilon)\big).$$

Das gilt für beliebige ρ. Um zu einem etwas glatteren Ergebnis zu kommen, setzen wir $\rho = 2/(e^{7/2}-1) \approx 0.062$. Dann ist $1/(1-2\rho)^2) \le 4/3$ und gleichzeitig $(\log(1+2/\rho))/(1-2\rho)^2 \le 14/3$. So erhalten wir als hinreichende Bedingung an m die Abschätzung

$$m \ge \frac{2}{3\tilde{c}}\delta^{-2}\big(7s + 2\log(2/\varepsilon)\big).$$

Wir fassen zusammen:

Satz 7.4.2. *Eine Zufallsmatrix A genüge der Konzentrationsungleichung. $S \subset \{1,\dots,N\}$ habe s Elemente, und ε, δ seien positiv. Falls dann*

$$m \ge C\delta^{-2}\big(7s + 2\log(2/\varepsilon)\big)$$

war, so gilt mit Wahrscheinlichkeit von mindestens $1-\varepsilon$, dass

$$||A_S^\top A_S - Id||_{s\to 2} \le \delta;$$

dabei ist $C := 2/(3\tilde{c})$.

Es fehlt noch der letzte Schritt. Es soll ja nicht ein spezielles S ausgezeichnet sein, wir wollen ja alle s-schwachbesetzten Vektoren gleichberechtigt behandeln. Hier ist unser Hauptergebnis:

Satz 7.4.3. *Eine Zufallsmatrix A genüge der Konzentrationsungleichung zu $\tilde{c}$, und ε, δ seien Zahlen in $]\,0,1\,[$. Falls dann*

$$m \geq \frac{C}{\delta^2}\big(9s + 2s\log(N/s) + 2\log(2/\varepsilon)\big),$$

so ist die Fastisometriekonstante δ_s höchstens gleich δ mit Wahrscheinlichkeit mindestens $1-\varepsilon$. Dabei ist wieder $C := 2/(3\tilde{c})$.

Bei geeigneter Wahl von δ kann also aufgrund der Ergebnisse aus Kapitel 5 garantiert werden, dass s-schwachbesetzte Vektoren durch l^1-Minimierung gefunden werden.

Beweis: Wir verwenden die vorstehenden Bezeichnungen. δ sei vorgegeben.

- Die Wahrscheinlichkeit, dass $||A_S^\top A_S - Id||_{2\to 2} \geq \delta$ bei einer festen Wahl von S ist, kann durch

  $$(1+2/\rho)^s \exp(-\tilde{c}\delta^2(1-2\rho)^2 m)$$

  abgeschätzt werden. (Vgl. die Diskussion vor Satz 7.4.2.)

- Folglich ist die Wahrscheinlichkeit, dass es für *irgendein* s-elementiges S schiefgeht, dass also $\delta_s \geq \delta$ gilt, höchstens gleich

  $$\binom{N}{s}(1+2/\rho)^s \exp(-\tilde{c}\delta^2(1-2\rho)^2 m)$$

Nun kann $\binom{N}{s}$ durch $(eN/s)^s$ abgeschätzt werden:

$$\binom{N}{s} = \frac{n\cdot(N-1)\cdots(N-s+1)}{s!} \leq \frac{N^s s^s}{s^s s!} \leq e^s (N/s)^s.$$

Zusammen:

$$\mathbb{P}(\delta_s \geq \delta) \leq (eN/s)^s (1+2/\rho)^s \exp(-\tilde{c}\delta^2(1-2\rho)^2 m).$$

Nun das Finale: Wählt man wie oben speziell $\rho = 2/(e^{7/2}-1) \approx 0.062$, so ergibt sich wieder $1/(1-2\rho)^2) \leq 4/3$ und $(\log(1+2/\rho))/(1-2\rho)^2 \leq 14/3$, und man erhält die etwas bequemere Abschätzung

$$m \geq \frac{1}{\tilde{c}\delta^2}\left(\frac{4s}{3}\log(eN/s) + \frac{14s}{3} + \frac{4}{3}\log(2/\varepsilon)\right).$$

Das impliziert die Behauptung:

$$\begin{aligned}\frac{1}{\tilde{c}\delta^2}\left(\frac{4s}{3}\log(eN/s)+\frac{14s}{3}+\frac{4}{3}\log(2/\varepsilon)\right) &= \frac{3C}{2\delta^2}\left(\frac{4s}{3}(1+\log(N/s))+\frac{14s}{3}+\frac{4}{3}\log(2/\varepsilon)\right)\\ &= \frac{C}{\delta^2}\left(2s(1+\log(N/s))+7s+2\log(2/\varepsilon)\right)\\ &= \frac{C}{\delta^2}\left(9s+2s\log(N/s)+2\log(2/\varepsilon)\right).\end{aligned}$$

□