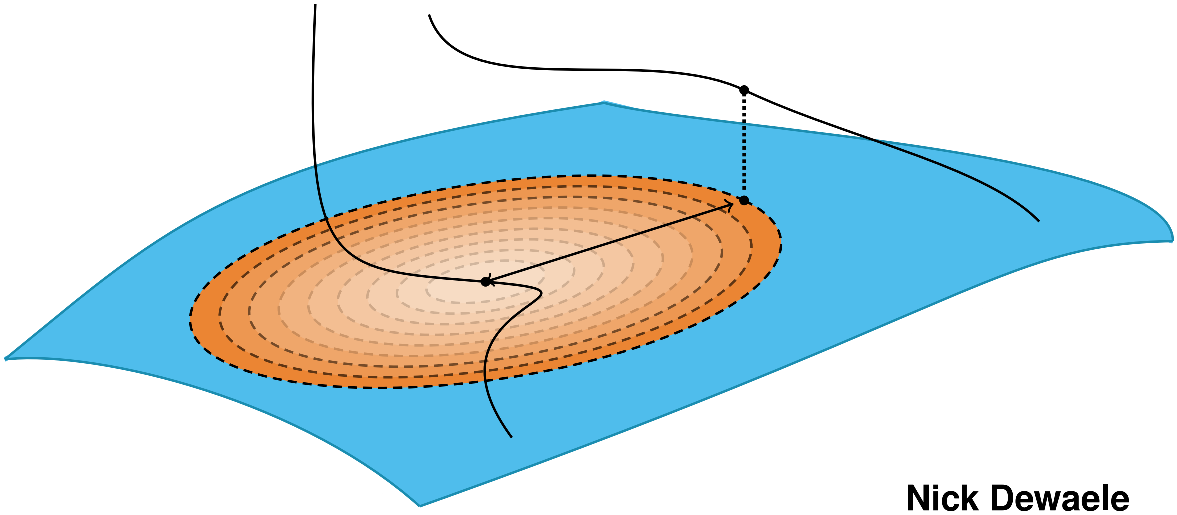
KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Geometry of numerical sensitivity



Nick Dewaele

Supervisor:
Prof. dr. N. Vannieuwenhoven
Co-supervisor:
Prof. dr. rer. nat. P. Breiding
(Universität Osnabrück)

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

5 July 2024

Geometry of numerical sensitivity

Nick DEWAELE

Examination committee:
Em. prof. dr. ir. O. Van der Biest, chair
Prof. dr. N. Vannieuwenhoven, supervisor
Prof. dr. rer. nat. P. Breiding, co-supervisor
(Universität Osnabrück)
Prof. dr. M. Ishteva
Prof. dr. J. Van der Veken
Prof. dr. ir. B. Vandereycken
(Université de Genève)
Prof. dr. C. Beltrán Alvarez
(Universidad de Cantabria)

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

5 July 2024

Uitgegeven in eigen beheer, Nick Dewaele, Celestijnenlaan 200A, B-3001 Leuven (Belgium)

Preface

This dissertation is the product of several years of work. This formative period was shaped by many individuals whose contributions must not be forgotten.

The foremost and most expected mention is, of course, my primary advisor Nick Vannieuwenhoven. It is sometimes said that we all turn into the people we spend the most time with. I am not alone to remark that this is an apt observation in my case. This is true not because we happen to share a name – as has been brought to my attention at least once or twice – but it is allegedly visible in our mannerisms, interests, and overall scientific approach. Despite some of our likenesses, I have grown into an independent academic and I have you to thank for that.

Second on the list is my second advisor Paul Breiding. We worked together closely in the first two years, when I was most in need of guidance. Especially at the early stages, a doctoral research project must have a clear sense of direction, and I feel that you provided this. Alas, our joint scientific output came to a halt around the time when I could see with my own eyes that you had legs and existed in three spatial dimensions.

By extension, I thank the whole jury for reviewing my dissertation, suggesting improvements, and engaging in interesting discussions about the content.

I was fortunate enough to travel to different places in Europe in the context of my research. Two of those stand out in particular. The first one is my month-long stay at AGATES, the semester organised by Weronika Buczyńska, Jarosław Buczyński, Francesco Galuppi, and Joachim Jelisiejew. This gave me a tremendous opportunity to meet people and explore Warsaw. The second one is the unforgettable Oberwolfach seminar on Metric Algebraic Geometry, organised by Paul Breiding, Kathlén Kohn, and Bernd Sturmfels. More than anything, this experience taught me that there is no fundamental divide between so-called “pure” and “applied” mathematics or mathematicians. The aura in Oberwolfach is truly one of a kind, which must be experienced to be understood.

Throughout the years, I took many breaks on workdays with all colleagues at NUMA. You have collectively engrained a Pavlovian response cycle into my head that is triggered whenever I hear footsteps and chatter in a hallway. This motivated me to show up to the office every day more than I would like to admit. Special thanks go to the four office mates I had: Julian, Nikhil, Yiqing, and Evert, who all contributed to an enjoyable experience at the office. I owe a subset of you an apology for any distractions I may or may not have caused.

Finally, I have had the pleasure of being supported by the most caring people in my personal life. My last words of gratitude go to my parents and to my best friend Katerina for having my back whenever I needed support.

Abstract

Numerical analysis is the study of computations with real numbers and the concepts one needs to understand when performing these computations on a modern computer. One such concept is the fact that real numbers are a mathematical abstraction: the field of real numbers is infinite, whereas computers are inherently finite. For this reason, computers cannot represent all real numbers exactly and thereby need to round off almost any number to one of only few which they *can* represent. The extent to which this affects the overal accuracy of the computation is called *numerical sensitivity.*

One of the basic measures of numerical sensitivity is the *condition number*, which depends on the computational problem one wishes to solve. It reflects, roughly, how accurately a problem can be solved (regardless of *how* the solution is obtained) if the input data are subject to small perturbations such as the roundoff that computers introduce. This thesis studies the theory and computation of condition numbers. One of its main tenets is that, by formulating numerical problems geometrically, one uncovers key insights about their sensitivity.

This dissertation follows two main paths, the first of which concerns the condition number of *tensor decomposition problems.* A *tensor* is the mathematical structure of an array of numbers. *Decomposing* a tensor can be viewed as breaking up an array of data into elementary components to reveal hidden structure in the tensor. This computation is the crux of a variety of algorithms used for data analysis. Since the decomposition is used to interpret the structure of the data, it is essential to quantify how sensitive the decomposition is to perturbations. This can be captured with the condition number.

The first major contribution of the thesis is a proof that, for a broad class of tensor decompositions, the condition number is invariant under *Tucker compression.* This property can be exploited to speed up the computation of the condition number by several orders of magnitude, so that it is now practically feasible to compute the condition number of some decompositions of large tensors.

The second path in the thesis concerns the theory of condition numbers of more general problems, specifically the solution of systems of equations. It presents a new framework that can be used to explain *why* a system is ill-conditioned. With this framework, it is possible to compute condition numbers of partially specified systems of equations and partial solutions. Using this information, one can quantify which equations and variables contribute the most to the condition number of a system of equations. The utility of this new theory is illustrated for Tucker decompositions of tensors.

Beknopte samenvatting

De numerieke analyse is de studie van berekeningen met reële getallen en de begrippen die men moet kennen bij het uitvoeren van deze berekeningen op een moderne computer. Één van deze begrippen is het feit dat de reële getallen een wiskundige abstractie zijn: het veld van reële getallen is oneindig, terwijl computers inherent eindig zijn. Om deze reden kunnen computers niet alle reële getallen exact voorstellen en moeten ze bijna elk getal afronden naar één van de weinige die ze wel exact kunnen voorstellen. De mate waarin dit de nauwkeurigheid van de berekening in het geheel beïnvloedt noemt men *numerieke sensitiviteit*.

Één van de voornaamste maten van numerieke sensitiviteit is het *conditiegetal*, dat afhangt van het op te lossen computationele probleem. Dit geeft ongeveer weer hoe nauwkeurig een probleem opgelost kan worden (ongeacht *hoe* de oplossing gevonden wordt) als de invoergegevens onderhevig zijn aan kleine verstoringen, zoals de afrondingen aangebracht door computers. Deze thesis bestudeert de theorie en berekening van conditiegetallen. Één van de voornaamste uitgangspunten erin is dat men door numerieke problemen meetkundig te formuleren essentiële inzichten opdoet omtrent sensitiviteit.

Dit proefschrift volgt in grote lijnen twee paden, waarvan het eerste handelt over het conditiegetal van *tensorontbindingsproblemen*. Een *tensor* is de wiskundige structuur van een meerdimensionale rij getallen. Het *ontbinden* van een tensor kan beschouwd worden als het opdelen van getabelleerde gegevens in bestanddelen die de structuur in de tensor onthullen. Deze berekening staat centraal in verscheidene algoritmen gebruikt in de gegevensanalyse. Aangezien de ontbinding gebruikt wordt om de structuur van de gegevens te interpreteren is het essentieel om te meten hoe gevoelig de ontbinding is aan onnauwkeurigheid op de gegevens. Deze informatie wordt bevat door het conditiegetal.

De eerste voorname bijdrage van de thesis is een bewijs dat het conditiegetal van een brede klasse van tensorontbindingen invariant is onder *Tucker-compressie*.

Deze eigenschap kan gebruikt worden om de berekening van het conditiegetal met enkele grootteordes te verstellen. Hierdoor is het nu praktisch haalbaar om het conditiegetal te berekenen van enkele ontbindingen van grote tensoren.

Het tweede pad in de thesis behandelt de theorie van conditiegetallen van meer algemene problemen, meer bepaald het oplossen van stelsels vergelijkingen. Hierin wordt een nieuw kader voorgesteld dat gebruikt kan worden om te verklaren *waarom* een stelsel slecht geconditioneerd is. Met dit kader is het mogelijk om conditiegetallen te berekenen van gedeeltelijk gespecificeerde stelsels vergelijkingen en gedeeltelijke oplossingen. Met deze informatie kan men kwantitatief uitdrukken welke vergelijkingen en veranderlijken het meest aan het conditiegetal van het stelsel bijdragen. Het nut van deze nieuwe theorie wordt geïllustreerd voor Tuckerontbindingen van tensoren.

List of Abbreviations

(C)PD	(canonical) polyadic decomposition
WD	Waring decomposition
PSTD	partially symmetric tensor decomposition
HOSVD	higher order singular value decomposition
(S)BTD	(structured) block term decomposition
FCRE	feasible constant-rank equation
CREP	constant-rank elimination problem

List of symbols

General mathematics

$\mathbb{R}$	real numbers
$\mathbb{C}$	complex numbers
$\mathbb{K}$	denotes either $\mathbb{R}$ or $\mathbb{C}$
$F\|_S$	restriction of F to S
$[x]$	equivalence class of x
$\mathcal{O}$	Landau's big $\mathcal{O}$
o	Landau's little o

Linear and multilinear algebra

$\\|A\\|$	norm of A (specified by the context)
$\\|A\\|_F$	Frobenius norm of A
$\sigma_i(A)$	ith largest singular value of A
$\sigma_{\min}(A)$	$\sigma_{\min(m,n)}(A)$ where $A \in \mathbb{R}^{m\times n}$
$\mathbb{I}_n$ or $\mathbb{I}$	$n \times n$ identity matrix
e_i	ith canonical basis vector
$\mathbb{V}^\perp$	orthogonal complement of the space $\mathbb{V}$
$A^\dagger$	Moore–Penrose inverse of A
$\otimes$	tensor product
$\mathcal{A}_{(d)}$	dth unfolding of $\mathcal{A}$
$a \otimes_k (a_1 \otimes \cdots \otimes a_D)$	shorthand for $a_1 \otimes \cdots \otimes a_{k-1} \otimes a \otimes a_{k+1} \otimes \cdots \otimes a_D$

Numerical analysis

$\kappa[F](x)$	condition number of F at x
$\kappa[\mathcal{P}](x,y)$	condition number of solving for y at $(x,y) \in \mathcal{P}$
$\kappa_{x\mapsto y}[\mathcal{P}](x,y,z)$	condition number of solving for y at $(x,y,z) \in \mathcal{P}$

Geometry

$\partial\mathcal{M}$	boundary of $\mathcal{M}$
$\mathcal{T}_x\mathcal{M}$	tangent space to $\mathcal{M}$ at x
$\dot{x}$	generic tangent vector at x
$\pi_{\mathcal{X}}$	projection onto $\mathcal{X}$
$\exp_x \dot{x}$	exponential map at x
$\log_x y$	logarithmic map at x
$DF(x)$	differential of F at x
$DF(x)[\dot{x}]$	differential of F in the direction $\dot{x}$
$\frac{\partial}{\partial x}F(x,y)$	partial derivative of F with respect to x
$\langle\cdot,\cdot\rangle$	inner product or Riemannian metric

Named spaces

$\mathbb{R}_k^{m\times n}$	real $m \times n$ matrices of rank k
$\mathbb{R}_\star^{n_1\times\cdots\times n_D}$	real tensors of size $n_1 \times \cdots \times n_D$ of full multilinear rank
$\mathrm{St}(n,m)$	Stiefel manifold of $n \times m$ matrices
$O(n)$	orthogonal group over $\mathbb{R}^n$ (equivalent to $\mathrm{St}(n,n)$)
$\mathrm{GL}(n)$	general linear group over $\mathbb{K}^n$
$\mathcal{S}$	Segre manifold
$\mathcal{V}$	Veronese manifold
$\mathcal{SV}$	Segre–Veronese manifold

Contents

Chapter 1

Introduction

Computation is one of many aspects in life that are remarkably different in theory and in practice. This is because the way we think about numbers in mathematics is far removed from how they are treated in applications.

In mathematics, we paint an ideal picture of just about everything. We teach students in secondary school that all real numbers together form a straight line whose centre is the number zero and which extends to negative infinity on the left and to positive infinity on the right. The number line is infinitely dense: no matter how closely we look, there are always infinitely many real numbers between any two points on the line.

Computer scientists, however, see a different picture. The usual system for translating real numbers into bits in computer memory is the use of so-called *floating-point numbers*, most commonly with *double precision* [OS06, Chapter 2]. This system attempts to fit the whole number line into a language where every word is a number and the length of every word is eight bytes. A priori, this language can only have 256^8 or about 18 quintillion (18×10^{19}) words, and thus, it can only talk about 18 quintillion numbers.

Though this may seem like an immensely expressive language, 18 quintillion is nothing compared to the infinite length and density of the number line. In mathematics, we are used to having unlimited precision (which is why we can talk about numbers like π having infinite digits). When we do calculations with computers, though, we need to accept that the words we have available cannot express anything beyond the sixteenth significant digit. That is, if two numbers have their first sixteen digits in common and only differ at the seventeenth digit, the language of floating point numbers cannot express this difference.

This contrast between the theory and the practice is enhanced when we consider where numbers come from. Most numbers we work with do not come from mathematics itself, but rather come to us as physical measurements or statistical estimates. When we work with numbers in real life, we tend to know at most a few digits after the decimal point. The sixteen digits that the *floating point* system of computers permits look plentiful in comparison.

Because our mathematical view of numbers is vastly different to how applications treat them, we must supplement our idealised mathematical theory with another theory that reasons about numbers in an approximate (but no less rigorous) sense. One of the key concepts for achieving this is that of a *condition number*. The subject of this dissertation is the theory and computation of these numbers.

1.1 Sensitivity in action

Readers who have worked their way through a handful of introductory numerical analysis textbooks might expect me to continue here by explaining what a matrix is and what Gaussian elimination does. After all, the term “condition number” is found so often in the same sentence as words like “matrix” and “Gaussian elimination” that one might be led to believe that all these words are inextricably linked. For a change, let us begin with a less orthodox example that is not usually discussed in this context.

One of the most memorable events involving the clash of two aforementioned perspectives on numbers is the discovery of chaos. I will summarise how Gleick [Gle08] outlines the story here. In 1961, meteorologist Edward Lorenz programmed a simulation tasked with forecasting the weather. This program encoded the atmosphere as a handful of variables that evolve over time, called the *state*. The state at any point could be calculated in terms of the state at some time interval in the past. Therefore, he could program a simulation that took physical measurements of the atmosphere (i.e., the *initial conditions*) and use them to predict the state at the next moment, the moment after that, and so on.

Wanting to study a sequence of predictions in detail, Lorenz took a predicted state from an earlier simulation and used it as the initial conditions for a new simulation. Naturally, one would expect the new simulation to predict the exact same states that the earlier simulation had. In reality, though, the new simulation briefly matched the original one, but diverged from it entirely after a short time. The cause was that Lorenz had rounded off the initial conditions to three digits while the computer calculated with six significant digits. Since small disturbances in the atmosphere can drastically impact the weather over time – a phenomenon later dubbed the *butterfly effect* in popular science – this initial roundoff made all the difference.

1.2 Condition numbers

Let us now shift gears when it comes to abstraction. In the foregoing example and in all other calculations, we have three objects of interest: (i) an *input* x (in the example: the initial weather conditions), (ii) an *output* y (e.g, the state of the atmosphere a month in the future), and (iii) a *model* that expresses how the output depends on the input (e.g., Lorenz' weather equations). This model may or may not be connected to predictions over time. In a very different context, x can be a tabulation of all words in a document and y can be the probability that the document is about the Roman Empire.

If we now remember that the input is not as precise and pure as we prefer it in mathematics, the input that will *actually* be used is not x but some slightly different quantity that we can write as $x+\Delta x$. For example, x could be changed to $x+\Delta x$ by rounding off the last digits, as Lorenz did. Another possible reason is that we may not even know the exact input (with all its digits after the decimal point). Since the input is different from its theoretically exact value, we expect the output to be different to its theoretical value y as well and instead be some value $y+\Delta y$.

If we can estimate how much uncertainty there is on the value of the input, it would be helpful to be able to estimate the uncertainty on the output. Such an estimate is provided by the *condition number*. This number depends on the model, and it is approximately equal to

$$\max_{\Delta x} \frac{||\Delta y||}{||\Delta x||} = \frac{\text{output uncertainty}}{\text{input uncertainty}}$$

in which $||\Delta x||$ and $||\Delta y||$ are the size of Δx and Δy, respectively. The maximum is taken over all sufficiently small[1] perturbations Δx. Concretely, this means that the condition number gives the highest possible factor by which the model could amplify small inaccuracies. In other words, it measures how *sensitive* the output is to small changes in the input.

For any calculation on a computer, the condition number puts a limit on how precise the result of the computation can generally be. Since computers usually process numbers with at most sixteen digits of precision, a condition number of, say, 10^{12} would indicate that the result of the calculation can only be trusted

[1] In the more rigorous Chapter 2, the maximum is replaced by a limit supremum as $\Delta x \to 0$.

up to at most $16 - 12 = 4$ significant figures. Therefore, we want the algorithm that does the calculation to actually deliver a solution this accurate (which not all algorithms do). In jargon: an algorithm should be *forward stable* [Hig02].

1.3 Tensors

This thesis has two main themes: general theory of condition numbers and applications of this theory to *tensor decomposition problems*. To introduce these problems, let us look at a text classification algorithm by Anandkumar et al. [Ana+14].

One of the simplest models for how texts carry meaning is the *bag-of-words model*. This model states that the topic of a text can be inferred by looking only at the words in the text, regardless of how they occur in a sentence. That is, if we cut out every word in the text, place them all in a bag, and shuffle the bag, the topic of the text is still identifiable from the words in the bag. This assumption may not seem accurate, but it has proven to be effective as a baseline model [JM09, §20.2].

The algorithm by Anandkumar et al. takes a corpus of many texts and identifies the document topics covered in the corpus as well as the typical word choices per topic. To do this, we proceed as follows: we make an empty three-dimensional table with as many rows and columns as there are words in the dictionary. In this example, I will use a three-word dictionary, but it will be larger in practice. Then for every set of three words in a text, we add one to the table at position (i, j, k) if the first word is the ith word in the dictionary, the second is the jth word, and the third is the kth word. For instance, if $i = 3, j = 1, k = 2$, then the table looks like this after seeing the first three words:

$$\left[\begin{array}{ccc|ccc|ccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

After going through all texts, every cell in the table will count how often a particular set of three words occurred in the whole corpus.

If we divide every count by the total number of words in all texts, we obtain the *empirical third moment tensor*, denoted as $\mathcal{A}$. The word *tensor* refers to the mathematical structure of arrays of numbers, and *empirical* means that it was constructed based on data rather than theory.

The goal is to find the relative frequency with which each topic is discussed as well as the probability that each word is used, given the topic. The key to identifying these parameters is that, if we *did* know what they were, there would have been another method for constructing the moment tensor.

The other method goes as follows: for each topic, we can predict what the moment tensor would be if all texts were only on that topic. For example, we could have a computer generate arbitrarily large bags of words using the word probabilities and take the moment tensors of these generated bags of words. An alternative (more clever) technique uses a closed formula [Ana+14, Theorem 3.1]. If the number of topics is R, this would generate R moment tensors, called $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_R$. If the relative frequencies with which the topics occur are $p_1, p_2, \ldots, p_R$, then we expect

$$\mathcal{A} = p_1\mathcal{A}_1 + p_2\mathcal{A}_2 + \cdots + p_R\mathcal{A}_R. \tag{1.1}$$

The left-hand side of (1.1) is the moment tensor of the whole corpus, which we have calculated. The right-hand side is unknown, since we do not know the topic and word probabilities. However, the equation (1.1) often has a unique solution in terms of $\mathcal{A}$ [Com+08]. Therefore, we can find the hidden parameters by solving this equation.

Equations such as (1.1) are known as *tensor decomposition problems* because they break up a tensor into simpler parts. Besides text classification, they have found numerous applications in areas such as separation of signals, analysis of psychological or chemical data, and compressed sensing, to name a few [KB09; PFS16; Sid+17]. The condition number can help us understand how sensitive the hidden variables are with respect to small inaccuracies in the tensor that we want to decompose. If the condition number is large, then small errors in the data can cause the parameters of the model to be very different and thereby impossible to interpret.

1.4 Approach and intended audience

This introduction has been an appetiser to appreciate what condition numbers are for. The mathematics I have used to study them is Riemannian geometry. Therefore, this dissertation may appeal to audiences with a background in either of the two historically separate fields of numerical analysis and geometry. Yet, it does not fall squarely in either category. For being a text on numerical analysis, the discussion on algorithms is thin on the ground. Likewise, this cannot be considered a text on differential geometry since it does not even attempt to advance the theory of manifolds. Instead, this dissertation is most appropriately understood as lying somewhere in between. It should be read as a treatise advocating that the geometry of a numerical problem is fundamental to one of the most basic questions one can ask about it: *how accurately can we expect to compute the solutions?*

1.5 Outline

Figure 1.1 gives a graphical overview of the dissertation. It shows the two main themes appearing throughout the thesis: condition numbers and tensor decompositions. The left side of the diagram is the *condition track*, which contributes to the general theory of condition numbers. This contribution is a *modular theory of condition*, which means that it makes it possible to break up a problem into smaller components. The right side is the *tensor track*, which applies the existing theory of condition in the literature to tensor decompositions. After the preliminary chapters, both tracks can be read independently of one another. Although the condition track consists of the later chapters, I consider it to be the most notable innovative contribution of the dissertation.

The remaining chapters can be summarised as follows.

> **Chapter 2** starts with an overview of the geometric prerequisites that have been proven useful for a comprehensive understanding of condition numbers. Then, the theory of condition numbers is built up from evaluation of functions to solving systems of equations.
>
> **Chapter 3** summarises the algebra of tensors, with an emphasis on decompositions such as tensor rank, Waring, block term, and Tucker decompositions. I also illustrate the relevance of the geometry of tensors to the study of their condition number. Finally, some applications of tensor decompositions are presented.
>
> **Chapter 4** is the first original contribution of the tensor track. We derive an invariance property of the condition number of most additive tensor decompositions that allows for a drastic speedup of the computation time. Another contribution is a joint analysis of some basic properties of tensor manifolds that I call *structured Tucker manifolds*.
>
> **Chapter 5** extends the results of the preceding chapter to symmetric tensor decompositions. It also establishes a connection between the condition number of symmetric and non-symmetric additive tensor decompositions.
>
> **Chapter 6** is the heart of the condition track. It introduces a new, modular theory of condition that can quantify the impact that each constraint in a

system of equations has on the condition number. The workhorse for this is a new theory of condition for underdetermined systems. The theory is illustrated by a computation of the condition number of two-factor matrix decompositions and orthogonal Tucker decompositions.

Chapter 7 complements the theory of the preceding chapter. Whilst Chapter 6 is about the contribution of each *equation* to the condition number, this chapter gives a theory that quantifies the impact of each *solution variable*. For example, this allows for the study of the sensitivity of individual factors in tensor decompositions, which I illustrate for the Tucker decomposition.

Finally, **Chapter 8** presents the main conclusions and outlook of the thesis.

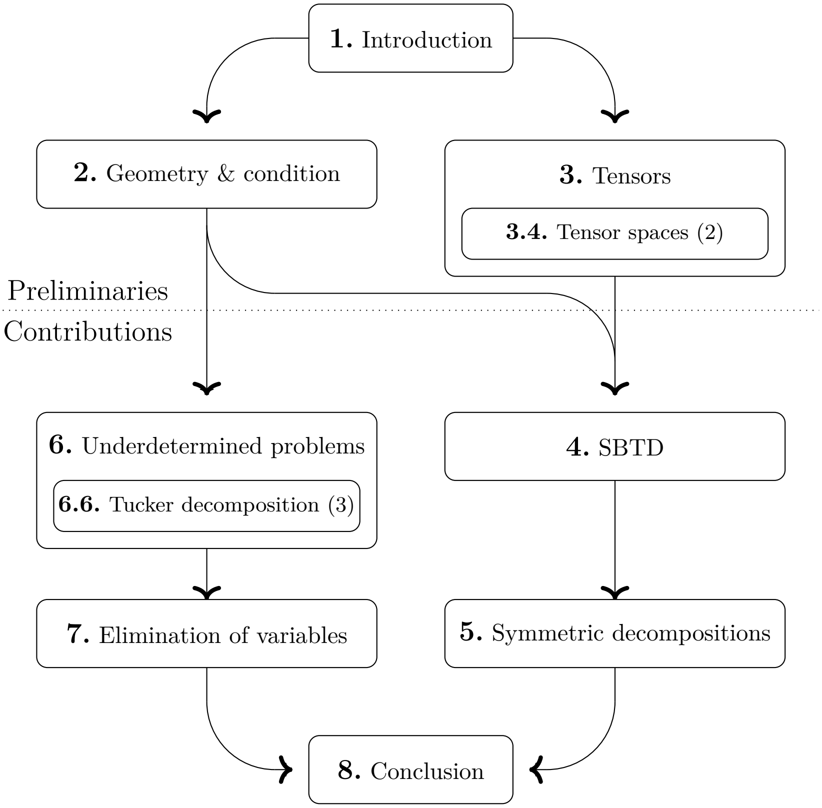


Figure 1.1: Content graph of the dissertation. The arrows represent which chapters serve as background knowledge for another chapter. Parentheses after a section indicate that the section builds on more preliminary information than the remainder of its chapter. The number in parentheses is the additional prerequisite chapter.

Chapter 2

Geometric foundations of condition

The aim of this chapter is to connect the chiefly numerical notion of condition numbers to basic concepts in differential geometry. Condition numbers are covered in every introductory course in numerical analysis, but they are rarely formalised in the numerical literature. An improved understanding of the condition number can come from an appreciation of the geometry of numerical problems. This chapter scratches that itch.

The standard references on geometric condition numbers are [Blu+98; BC13], from which I have borrowed major elements to write this chapter. The literature clustered around these two books emphasises condition as a tool for complexity analysis and homotopy continuation. By contrast, this text stays true to one main aspect of condition: an estimate of rounding errors. This is achieved by tracing all occurrences of the condition number back to Rice's definition, introduced in Section 2.2. Another difference is that I shine the spotlight on the major class of numerical problems in this thesis: inverse problems.

The geometric preliminaries are established in Section 2.1. The key objects introduced in this section are Riemannian manifolds and the differential of a map. Readers unfamiliar with differential geometry are encouraged to gloss over this section and come back to it when geometric concepts are invoked. The basic definition of condition of functions is presented in Section 2.2. In Section 2.3, the concept of condition is lifted to more general numerical problems. This section also gives an overview of how the condition number of several important classes of numerical problems can be computed. Finally, Section 2.4 discusses how the condition number is related to other concepts in numerical analysis besides roundoff error. This last section can be regarded as supplementary content, since it is not used in the remainder of the thesis.

2.1 Fundamentals of differential geometry

This section provides an overview of the geometric concepts used throughout this thesis. All geometric fundamentals below come from the standard references [Lee11; Lee13; Lee18]. For an introduction to differential geometry focused on numerical aspects and optimisation, see e.g. the book by Absil, Mahony, and Sepulchre [AMS08] or Boumal's book [Bou23].

2.1.1 Smooth manifolds

The fundamental objects in differential geometry are *smooth manifolds*. Intuitively, these are spaces that look Euclidean when zoomed in close enough. They are exactly the spaces to which we can generalise standard concepts from multivariable calculus, such as directional derivatives, integrals, vector fields, tangent lines, etc. Many spaces that are familiar to numerical analysts can be described as smooth manifolds, including

- any real vector space (in these spaces, differential geometry is exactly multivariable calculus),
- smooth curves and surfaces in $\mathbb{R}^n$,
- the n-dimensional sphere $\mathbb{S}^n = \{(x_0, \dots, x_n) \in \mathbb{R}^{n+1} \mid \sum_{i=0}^{n} x_i^2 = 1\}$,
- $m \times n$ matrices of some fixed rank $k \leqslant \min\{m, n\}$, and
- $m \times n$ matrices with orthonormal columns, i.e., the *Stiefel manifold*,

to name a few.

A precise definition of smooth manifolds is the following: a topological space $\mathcal{M}$ is a *smooth manifold of dimension n* if all of the following holds:

- $\mathcal{M}$ has an open cover $\mathcal{M} = \bigcup_{\alpha \in A} \mathcal{U}_\alpha$,
- for every α, there exists a homeomorphism (i.e., a map that is continuous in both directions) $\phi_\alpha : \mathcal{U}_\alpha \to \mathcal{B}_\alpha$ where $\mathcal{B}_\alpha$ is an open subset of $\mathbb{R}^n$,

- for all α, β where $\mathcal{U}_{\alpha\beta} := \mathcal{U}_\alpha \cap \mathcal{U}_\beta \neq \emptyset$, the *transition map* $\phi_\alpha \circ (\phi_\beta|_{\mathcal{U}_{\alpha\beta}})^{-1}$ is a smooth map from $\phi_\beta(\mathcal{U}_{\alpha\beta}) \subseteq \mathcal{B}_\beta$ to $\mathcal{B}_\alpha$, and
- the topology of $\mathcal{M}$ is Hausdorff and second countable.

The tuple $(\mathcal{U}_\alpha, \phi_\alpha)$ is called a *chart* and ϕ_α is a *chart map*, but we will call it *chart* for short. If $p \in \mathcal{M}$ and $\phi_\alpha(p) = 0$, we call ϕ_α a *chart centred at p*. The chart is usually thought of as a map that assigns an n-tuple of coordinates to points in an open subset $\mathcal{U}_\alpha$ of $\mathcal{M}$. The properties imposed on the charts imply two main properties: first, that every sufficiently small neighbourhood in $\mathcal{M}$ is topologically the same an open subset of $\mathbb{R}^n$ and (second) that, for regions in $\mathcal{M}$ with multiple coordinate maps, any change of local coordinates is a map from $\mathbb{R}^n$ to $\mathbb{R}^n$ that is smooth in the sense of basic calculus.

The collection of charts $\{(\mathcal{U}_\alpha, \phi_\alpha)\}_{\alpha \in A}$ is referred to as the *smooth structure of* $\mathcal{M}$. This name stems from how smooth maps are defined over $\mathcal{M}$. Let $F : \mathcal{M} \to \mathcal{N}$ be a map between smooth manifolds of dimension m and n, respectively, and let $p \in \mathcal{M}$ be any point. Let ϕ and ψ be charts defined at p and $F(p)$, respectively. Then F is k times *differentiable* at p if $\widehat{F} := \psi \circ F \circ \phi^{-1}$ is k times differentiable at $\phi(p)$. If this is the case for all $k \in \mathbb{N}$, then F is *smooth*. We can interpret $\widehat{F}$ as the local coordinate representation of F, i.e., it maps the coordinates of p to the coordinates of $F(p)$. Since $\widehat{F}$ maps an open set of $\mathbb{R}^n$ into $\mathbb{R}^n$, this definition of differentiability boils down to that of multivariable calculus.

To linearise manifolds around a point, we use the *tangent space*. Intuitively, the tangent space at a point $p \in \mathcal{M}$, denoted $\mathcal{T}_p\mathcal{M}$ is the set of all vectors tangent to $\mathcal{M}$ at p. For instance, for a smooth surface $\mathcal{M} \subseteq \mathbb{R}^N$, the tangent space at any point is the plane tangent to $\mathcal{M}$. With this informal definition, $\mathcal{T}_p\mathcal{M}$ is a subset of $\mathbb{R}^N$. It is sometimes desirable to define the tangent space only in terms of $\mathcal{M}$, rather than any ambient space in which $\mathcal{M}$ is embedded. This can be done using *directional derivatives*.

For manifolds embedded in $\mathbb{R}^N$, the identification of a tangent vector $\xi \in \mathcal{T}_p\mathcal{M}$ with a directional derivative operator is the following. The extension lemma states that every smooth map $F : \mathcal{M} \to \mathbb{R}$ can be extended smoothly just outside $\mathcal{M}$, i.e., there is a smooth map $F^{\mathcal{U}} : \mathcal{U} \to \mathbb{R}$ defined on an open set $\mathcal{U} \subseteq \mathbb{R}^N$ containing $\mathcal{M}$ such that $F = F^{\mathcal{U}}|_{\mathcal{M}}$. Then the directional derivative of F along a tangent vector ξ can be defined just like in calculus:

$$\nabla_\xi F(p) := \lim_{t \to 0} \frac{F(p + t\xi) - F(p)}{t}.$$

The identification between ξ and the operator ∇_ξ can be used to define tangent spaces of general manifolds (not just those that are embedded in $\mathbb{R}^N$) as follows. Let $p \in \mathcal{M}$ be any point and let $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{M}$ be a smooth curve such that $\gamma(0) = p$. The *tangent vector* corresponding to γ is the operator $\xi : C^\infty(\mathcal{M}, \mathbb{R}) \to \mathbb{R}, F \to (F \circ \gamma)'(0)$. The point p is the *origin* of ξ. The set of all tangent vectors whose origin is p forms a vector space, called the *tangent space of* $\mathcal{M}$ *at* p, or $\mathcal{T}_p\mathcal{M}$. Its dimension is $\dim \mathcal{M}$.

Every differentiable map $F : \mathcal{M} \to \mathcal{N}$ between smooth manifolds induces a linear map on the tangent spaces of $\mathcal{M}$ and $\mathcal{N}$, called the *differential*. Let p in $\mathcal{M}$ and let $\xi \in \mathcal{T}_p\mathcal{M}$ be the tangent vector corresponding to some curve γ. Then the *differential of* F *at* p takes ξ to the tangent vector $\eta \in \mathcal{T}_{F(p)}\mathcal{N}$ corresponding to the curve $F \circ \gamma$. This map is written as $\eta = DF(p)[\xi]$. This vector η can be interpreted as the derivative of F in the direction of ξ.

The differential obeys standard properties of differentiation, such as the chain rule, i.e., $D(G \circ F)(p) = DG(F(p)) \circ DF(p)$ for any differentiable map G defined on $\mathcal{N}$. If $DF(p)$ is injective, F is an *immersion*. If $DF(p)$ is surjective, F is a *submersion*. A (topological) *embedding* is a map that is a homeomorphism onto its image. A *smooth embedding* is a smooth immersion which is also an embedding.

The differential plays the same role in differential geometry as the matrix of partial derivatives (or Jacobian matrix) in multivariable calculus. Given charts ϕ and ψ centred at p and $F(p)$, respectively, the coordinate representation of F is $\widehat{F} = \psi \circ F \circ \phi^{-1} : \mathbb{R}^m \to \mathbb{R}^n$, where $m = \dim \mathcal{M}$ and $n = \dim \mathcal{N}$. The matrix representation of the linear map $D\widehat{F}(\phi(p)) : \mathbb{R}^m \to \mathbb{R}^n$ is the Jacobian matrix of $\widehat{F}$.

Given a manifold $\mathcal{M}$, some related spaces of points can be shown to admit a manifold structure. For instance, the product of two smooth manifolds $\mathcal{M}$ and $\mathcal{N}$ is a smooth manifold. A topological subspace $\mathcal{S}$ of $\mathcal{M}$ is an *embedded submanifold* of $\mathcal{M}$ if it has a smooth structure such that the inclusion $\mathcal{S} \hookrightarrow \mathcal{M}$ is a smooth embedding.

2.1.2 Riemannian manifolds

To study numerical analysis, we need a notion of distance to express errors. In differential geometry, the standard definition of distance is induced by a *Riemannian metric*. A *metric* on a smooth manifold $\mathcal{M}$ is an inner product defined on the tangent space, i.e., at any point $p \in \mathcal{M}$, the metric is a positive definite bilinear form $g_p : \mathcal{T}_p\mathcal{M} \times \mathcal{T}_p\mathcal{M} \to \mathbb{R}$. In coordinates induced by a chart ϕ at p, we can write g_p as a positive definite matrix G_p, such that $g_p(\xi, \eta) = \vec{\xi}^T G_p \vec{\eta}$ where $\vec{\xi}$ and $\vec{\eta}$ are the coordinate tuples of ξ and η, respectively.

A *Riemannian metric* is a metric whose coordinate matrix G_p is a smooth function of p. The metric is also written as $g_p(\xi, \eta) = \langle \xi, \eta \rangle$. The norm induced by $\langle \cdot, \cdot \rangle$ is written as $\|\cdot\|$. A smooth manifold with a specified Riemannian metric is called a *Riemannian manifold.* A *Riemannian submanifold* of $\mathcal{M}$ is an embedded submanifold of $\mathcal{M}$ whose metric is a restriction of the metric on $\mathcal{M}$.

The Riemannian metric induces a definition of length of a curve over $\mathcal{M}$. A curve $\gamma : [a, b] \to \mathcal{M}$ is *admissible* if it is piecewise smooth and its differential is never zero. The *length* of an admissible curve γ is the integral $L[\gamma] := \int_a^b \|\gamma'(t)\| \mathrm{d}t$ where $\gamma'(t)$ is the differential of γ.

Between any two points $p, q \in \mathcal{M}$, the *Riemannian* or *geodesic distance* is defined as

$$d(p, q) = \inf \{ L[\gamma] \mid \gamma \text{ is an admissible curve from } p \text{ to } q \} .$$

If there are no admissible curves connecting p and q, then $d(p, q) := \infty$. If $\mathcal{M}$ is a Euclidean space (i.e., a finite-dimensional inner product space), then the shortest curve from p to q is a straight line. Its length is the Euclidean distance between p and q.

A special type of curves are *geodesics*, which are defined by having zero curvature relative to $\mathcal{M}$. They can be thought of as curves of constant speed on $\mathcal{M}$ that are as close to straight lines as possible. That is, geodesics have zero curvature in every direction tangent to $\mathcal{M}$. A precise definition can be found in the textbook [Lee18]. For example, over the unit sphere $\mathbb{S}^n$, the geodesics are precisely the circle segments of unit radius. For any $p, q \in \mathcal{M}$ such that $d(p, q) \neq \infty$, any curve γ such that $L[\gamma] = d(p, q)$ can be parametrised a geodesic.

Given a point $p \in \mathcal{M}$ and a vector $\xi \in \mathcal{T}_p\mathcal{M}$ sufficiently close to 0, there is a unique geodesic γ_ξ such that $\gamma(0) = p$ and $\gamma'(0) = \xi$. The *exponential map* is defined as $\exp_p : \mathcal{T}_p\mathcal{M} \to \mathcal{M}, \xi \mapsto \gamma_\xi(1)$. It is a local diffeomorphism between $\mathcal{M}$ and $\mathcal{T}_p\mathcal{M}$. The local inverse of the exponential map is the *logarithmic map* $\log_p$.

The logarithmic map provides a convenient choice of coordinates, called *normal coordinates*. Let $\mathscr{B}$ be any orthonormal basis for $\mathcal{T}_p\mathcal{M}$ and let $F : \mathcal{T}_p\mathcal{M} \to \mathbb{R}^n$ be the coordinate map relative to $\mathscr{B}$. Then the associated *normal coordinate map* at p is $\phi := F \circ \log_p$. This chart preserves the metric information of $\mathcal{M}$ at p. In particular, $D\phi(p)$ is unitary and geodesics through p are mapped to lines in $\mathbb{R}^n$.

2.2 The condition number of a map

When browsing the literature on computational mathematics, one encounters various objects named *condition numbers*. The earliest use of the term is in Turing's paper on matrix computations [Tur48], where the condition number of a matrix $A \in \mathbb{R}^{n\times n}$ is defined as $\frac{1}{n}\|A\|\|A^{-1}\|$. An entirely different notion of condition was introduced by Renegar [Ren95], where the condition number measures the distance to the nearest ill-posed problem. Additionally, there is a whole zoo of ad hoc definitions of condition that serve to estimate the numerical error of specific algorithms.

This ambiguity is not unlike that of the term *singularity*. Depending on the context, singularity may refer to a property of matrices over arbitrary fields, functions of real or complex variables, or algebraic varieties. While these notions of singularity are all related, the precise definition depends on the context. Such is the case for condition as well.

To avoid any "condition number zoo," we take a general definition of condition due to Rice [Ric66] and show how various notions of condition may be derived from it. This definition can be stated as follows.

Definition 2.1. Let $F : \mathcal{X} \to \mathcal{Y}$ be any map between metric spaces with distances $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$, respectively. Then the *condition number* (or simply *condition*) of F at any point $x_0 \in \mathcal{X}$ is

$$\kappa[F](x_0) := \limsup_{x\to x_0} \frac{d_{\mathcal{Y}}(F(x_0), F(x))}{d_{\mathcal{X}}(x_0, x)},$$

where the limit is to be understood in the metric topology.

Evaluating F at x_0 is said to be *ill-conditioned* if $\kappa[F](x_0)$ is large (by some subjective standard, typically several orders of magnitude) and *well-conditioned otherwise.*

We will omit the subscripts in $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ occurring in Definition 2.1 when it does not compromise on clarity. An equivalent definition reflects the way in which condition numbers are usually applied: $\kappa[F](x_0)$ is the smallest number κ such that

$$d(F(x_0), F(x)) \leqslant \kappa \cdot d(x_0, x) + o(d(x_0, x)) \quad \text{as} \quad x \to x_0. \tag{2.1}$$

Recall that the expression "$o(f(x))$ as $x \to x_0$" is shorthand for an unspecified function $g(x)$ that converges to zero faster than $f(x)$ does. That is, for all $L > 0$, there exists a neighbourhood $\mathcal{U}$ of x_0 such that $|g(x)| \leqslant L|f(x)|$ for all $x \in \mathcal{U}$.

In (2.1), $d(x_0, x)$ is the error on the data and $d(F(x_0), F(x))$ is the induced error on the result of the computation. The asymptotic term $o(d(x_0, x))$ is often neglected when x is sufficiently close to x_0. Thus, (2.1) can be used as a back-of-the-envelope estimate of errors in numerical computations. It is from this estimate that the concept of the condition number as a measure of sensitivity emerges. The right-hand side of (2.1) is more or less the expected error when F is evaluated numerically [Arm10].

Sometimes a distinction is made between *absolute* and *relative* condition numbers. For a metric space $\mathcal{M}$ contained in a normed vector space, the *relative error* between two points $p, q \in \mathcal{M} \setminus \{0\}$ is defined as $d_r(p, q) := d(p, q)/\|p\|$. Unless $\|\cdot\|$ is constant over $\mathcal{M}$, this is not a distance function. A common definition of the *relative condition number* is

$$\kappa_r[F](x_0) := \limsup_{x \to x_0} \frac{d_r(F(x_0), F(x))}{d_r(x_0, x)} = \frac{\|x_0\|}{\|F(x_0)\|} \limsup_{x \to x_0} \frac{d(F(x_0), F(x))}{d(x_0, x)}.$$

Note that $\kappa_r[F](x_0)$ and $\kappa[F](x_0)$ are the same up to a factor $\|x_0\|/\|F(x_0)\|$. For this reason, most of our focus will be on the absolute condition number.

Example 2.2. (Turing's condition number [Tur48; BC13, §1.2]) Let $\Sigma \subseteq \mathbb{R}^{n \times n}$ be the set of singular $n \times n$ matrices. Consider $\mathbb{R}^{n \times n}$ as a metric space defined by the spectral norm $\|\cdot\|$. At any $X_0 \in \mathbb{R}^{n \times n} \setminus \Sigma$, the absolute and relative condition number of $F : \mathbb{R}^{n \times n} \setminus \Sigma \to \mathbb{R}^{n \times n} : X \to X^{-1}$ are, respectively,

$$\kappa[F](X_0) = \left\|X_0^{-1}\right\|^2 \quad \text{and} \quad \kappa_r[F](X_0) = \|X_0\|\left\|X_0^{-1}\right\|.$$

Remark 2.3. Because Turing's condition number $\|X_0\|\left\|X_0^{-1}\right\|$ appears in sensitivity estimates of many problems associated with the matrix X_0 (e.g., those in [SS90], especially Part III), it is referred to as *the condition number of* X_0 by most numerical analysts. I caution against this usage of the term, though, since it obfuscates for what problem the quantity $\|X_0\|\left\|X_0^{-1}\right\|$ is actually the condition number. If it is not emphasised that this usage may not be equivalent to Definition 2.1, the inattentive reader might read "the condition number of X_0" and conclude (erroneously) that just about every problem involving X_0 has $\|X_0\|\left\|X_0^{-1}\right\|$ as its condition number.

If F is a smooth map between Riemannian manifolds, the condition number can be computed in terms of the differential of F. This is expressed by the following theorem, which is sometimes used as the definition of the condition number [BC13, §14.3].

Theorem 2.4 (Rice's theorem). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Riemannian manifolds and let $x_0 \in \mathcal{X}$ be any point. For any $F : \mathcal{X} \to \mathcal{Y}$ that is differentiable at x_0, the* *condition number with respect to the geodesic distance satisfies*

$$\kappa[F](x_0) = \|DF(x_0)\| := \sup_{0 \neq \xi \in \mathcal{T}_{x_0}\mathcal{X}} \frac{\|DF(x_0)[\xi]\|}{\|\xi\|}.$$

This theorem is usually attributed to Rice [Ric66]. However, the original proof requires that the radial distance function $d_{x_0} : x \mapsto d(x_0, x)$ be differentiable at x_0. Riemannian distances never have this property, since the radial distance is locally the composition of a normal coordinate map (i.e., a diffeomorphism) and the Euclidean norm (which has a singularity at the origin) [Lee18, Corollary 6.12]. For this reason, an updated proof is presented below.

Proof. Let $\phi : x \mapsto \hat{x}$ and $\psi : y \mapsto \hat{y}$ be normal coordinate charts of $\mathcal{X}$ and $\mathcal{Y}$, centred at x_0 and $F(x_0)$, respectively. Define $\widehat{F} := \psi \circ F \circ \phi^{-1}$. For all x in some neighbourhood of x_0, it holds that

$$d(x_0, x) = \|\hat{x}\| \quad \text{and} \quad d(F(x_0), F(x)) = \left\|\widehat{F}(\hat{x})\right\|$$

where $\|\cdot\|$ is the Euclidean norm [Lee18, Corollary 6.12]. This equivalence allows for the following exchange of limits:

$$\limsup_{x \to x_0} \frac{d(F(x_0), F(x))}{d(x_0, x)} = \limsup_{\hat{x} \to 0} \frac{d(F(x_0), F(x))}{d(x_0, x)} = \limsup_{\hat{x} \to 0} \frac{\left\|\widehat{F}(\hat{x})\right\|}{\|\hat{x}\|}. \tag{2.2}$$

Applying Taylor's theorem gives $\widehat{F}(\hat{x}) = D\widehat{F}(0)[\hat{x}] + o(\|\hat{x}\|)$ as $\hat{x} \to 0$. Therefore, by the triangle inequality,

$$\left\|D\widehat{F}(0)[\hat{x}]\right\| - o(\|\hat{x}\|) \leqslant \left\|\widehat{F}(\hat{x})\right\| \leqslant \left\|D\widehat{F}(0)[\hat{x}]\right\| + o(\|\hat{x}\|) \quad \text{as} \quad \hat{x} \to 0.$$

Hence,

$$\limsup_{\hat{x} \to 0} \frac{\left\|\widehat{F}(\hat{x})\right\|}{\|\hat{x}\|} = \limsup_{\hat{x} \to 0} \left(\frac{\left\|D\widehat{F}(0)[\hat{x}]\right\|}{\|\hat{x}\|} + o(1) \right). \tag{2.3}$$

Since $\left\|D\widehat{F}(0)[\hat{x}]\right\| / \|\hat{x}\|$ is constant along lines punctured at the origin, the limit supremum is an ordinary supremum over all $\hat{x} \neq 0$. Therefore, the above equals $\left\|D\widehat{F}(0)\right\|$. Since $D\widehat{F}(0) \cong DF(x_0)$ up to multiplication by unitary linear maps, $\left\|D\widehat{F}(0)\right\| = \|DF(x_0)\|$. Combining this with (2.2) and (2.3) gives the desired result.

□

Remark 2.5. Though Rice's theorem applies foremost to the geodesic distance, it is true for any asymptotically equivalent distance. We say that two distances d and $\tilde{d}$ in a metric space $\mathcal{M}$ are asymptotically equivalent if, for any $p \in \mathcal{M}$, it holds that $\tilde{d}(p, q) = d(p, q)(1 + o(1))$ as $q \to p$, for $q \in \mathcal{M}$. The asymptotic term $o(1)$ vanishes in Definition 2.1.

In particular, if $\mathcal{M}$ is a Riemannian submanifold of $\mathbb{R}^n$, then the geodesic distance d and the restriction of the Euclidean distance $\tilde{d}$ from $\mathbb{R}^n$ to $\mathcal{M}$ are asymptotically equivalent. To see this, write $\exp_p \colon \mathcal{T}_p\mathcal{M} \to \mathbb{R}^n$ as a map between normed linear spaces with the Taylor expansion $\exp_p \xi = p + \xi + \mathcal{O}(\|\xi\|^2)$. Since $\exp_p$ is a local embedding, we have

$$\lim_{q \to p} \frac{\|q - p\|}{d(q, p)} = \lim_{\xi \to 0} \frac{\|\exp_p \xi - p\|}{d(\exp_p \xi, p)} = \lim_{\xi \to 0} \frac{\|\xi\| + \mathcal{O}(\|\xi\|^2)}{\|\xi\|} = 1.$$

Hence, d is asymptotically equivalent to the Euclidean distance.

2.3 Geometry of numerical problems

Not all numerical problems are presented to us as explicit functions to be evaluated. For example, consider the problem of finding real roots of $x^3 + ax^2 + bx + c = 0$ given the tuple $(a, b, c) \in \mathbb{R}^3$. Depending on (a, b, c), there may be one, two, or three solutions. How should the problem be modelled as the evaluation of a function $F(a, b, c)$? If the problem is to find *all* real roots, it is not clear what the space is in which the output $F(a, b, c)$ lives. If, instead, we are asked to find *at least one* root, the problem cannot be modelled as a function $F : \mathbb{R}^3 \to \mathbb{R}$, since multiple outputs may correspond to the same input. Given that we cannot easily model the problem as a function, how do we interpret its condition number? This section provides an answer.

2.3.1 The geometric condition number

The aforementioned problem motivates the geometric study of numerical problems and their condition numbers, pioneered by Blum, Cucker, Shub, and Smale [Blu+98] and by Bürgisser and Cucker [BC13]. It allows us to think abstractly about numerical problems without abandoning Rice's theory of condition. Most of this section is my interpretation of the framework in [Blu+98; BC13].

A general numerical problem has an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. The set of all admissible input/output pairs is a subset $\mathcal{P}$ of $\mathcal{X} \times \mathcal{Y}$. I will simply refer to $\mathcal{P}$ as “the problem”, since $\mathcal{P}$ encapsulates all information about the relationship between inputs and outputs. In the literature, $\mathcal{P}$ is also called the *solution variety* or *solution manifold* depending on the geometric properties of $\mathcal{P}$. When we solve a numerical problem, we are given $x \in \mathcal{X}$ such that $(x, y) \in \mathcal{P}$ for some unknown y. The goal, then, is to find (at least one) y such that $(x, y) \in \mathcal{P}$.

For example, solving monic cubic equations over the real numbers can be modelled in terms of the solution variety

$$\mathcal{P} = \big\{(x, y) \,\big|\, x = (a, b, c) \in \mathbb{R}^3 \quad \text{and} \quad y^3 + ay^2 + by + c = 0\big\} \subseteq \mathbb{R}^3 \times \mathbb{R}.$$

Likewise, for problems defined as the evaluation of a map $F : \mathcal{X} \to \mathcal{Y}$ (as in the previous section), $\mathcal{P}$ is the *graph* of F, defined as $\mathcal{P} := \{(x, F(x)) \,|\, x \in \mathcal{X}\}$. When we talk about “the geometry of a numerical problem”, we are interested in properties of the projection maps $\pi_{\mathcal{X}} : \mathcal{P} \to \mathcal{X}, (x, y) \mapsto x$ and $\pi_{\mathcal{Y}} : \mathcal{P} \to \mathcal{Y}, (x, y) \mapsto y$.

The geometry of a problem determines whether Rice’s definition of the condition number can be naturally applied to it. For general numerical problems, we need the concept of *condition of $\mathcal{P}$ at a point* $(x_0, y_0) \in \mathcal{P}$, which we define next.

To generalise condition, we want to formalise the notion that, on $\mathcal{P}$, the variable y is a continuous function of x. The *domain* of $\mathcal{P}$ is the set of all $x \in \mathcal{X}$ that have a corresponding solution, i.e., $\pi_{\mathcal{X}}(\mathcal{P})$. Formally, the notion of continuity that we want to express is that the inverse projection

$$\begin{aligned} \pi_{\mathcal{X}}^{-1} \colon \ \pi_{\mathcal{X}}(\mathcal{P}) &\to \mathcal{P} \\ x &\mapsto (x, y) \end{aligned}$$

is continuous. In other words, $\pi_{\mathcal{X}}$ is a *topological embedding.* If this is the case, we may define the *solution map* as the continuous map $H := \pi_{\mathcal{Y}} \circ \pi_{\mathcal{X}}^{-1}$, which maps x to y such that $(x, y) \in \mathcal{P}$. The condition number of this map may be used as the definition for the condition number of $\mathcal{P}$.

While this assumption on $\pi_{\mathcal{X}}$ can be used to generalise condition slightly, the above can only apply if $\mathcal{P}$ is the graph of a continuous function H. It seems as though we have not made any progress towards our initial goal of studying problems that are *not* merely the graph of a function! The key to remedying this is that condition numbers are a *local* property. That is, instead of imposing $\pi_{\mathcal{X}}$ to be an embedding, it suffices that, at any point (x_0, y_0) of interest, the restriction of $\pi_{\mathcal{X}}$ to some *neighbourhood* $\mathcal{U} \subseteq \mathcal{P}$ of (x_0, y_0) is an embedding. This is visualised in Figure 2.1. We can formalise these insights as follows.

Definition 2.6. Let $\mathcal{X}$ and $\mathcal{Y}$ be metric spaces and let $\mathcal{P}$ be a subspace of $\mathcal{X} \times \mathcal{Y}$ containing some point (x_0, y_0). Let $\pi_{\mathcal{X}}$ and $\pi_{\mathcal{Y}}$ denote the projections

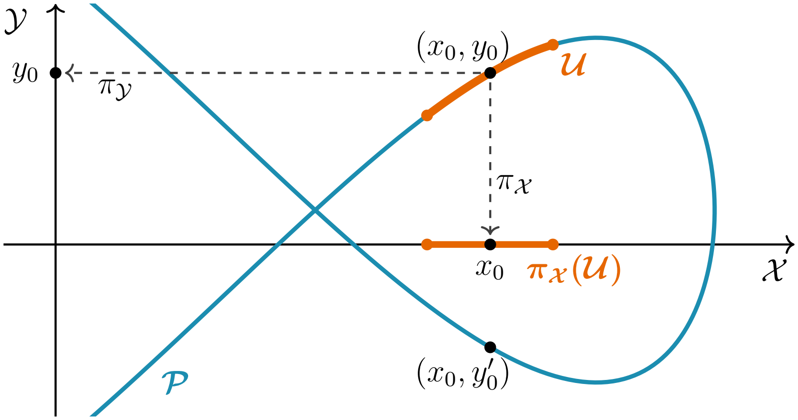


Figure 2.1: Projections of a problem $\mathcal{P} \subseteq \mathcal{X} \times \mathcal{Y}$ with a one-dimensional input and output. At the point (x_0, y_0), the projection $\pi_{\mathcal{X}}$ is a local embedding. Thus, $\mathcal{U}$ is the graph of the locally defined solution map $\pi_{\mathcal{Y}} \circ \pi_{\mathcal{X}}^{-1} \colon x \mapsto y$. This means that every x that is sufficiently close to x_0 corresponds to a unique point $(x, y) \in \mathcal{U}$, although it does not rule out the existence of another point $(x, y') \in \mathcal{P} \setminus \mathcal{U}$ further away from (x_0, y_0). This example problem fails to be the graph of a solution map at the self-intersection and at the rightmost point of $\mathcal{P}$.

from $\mathcal{P}$ onto $\mathcal{X}$ and $\mathcal{Y}$, respectively. Suppose that (x_0, y_0) has a neighbourhood $\mathcal{U}$ such that $\pi_{\mathcal{X}}|_{\mathcal{U}}$ is a topological embedding. Then the *condition number of $\mathcal{P}$* at (x_0, y_0) is

$$\begin{aligned}\kappa[\mathcal{P}](x_0, y_0) &:= \limsup_{\substack{(x,y)\to(x_0,y_0)\\(x,y)\in\mathcal{U}}} \frac{d(y_0, y)}{d(x_0, x)} = \limsup_{\substack{x\to x_0\\x\in\pi_{\mathcal{X}}(\mathcal{U})}} \frac{d\left(y_0, \pi_{\mathcal{Y}}\left(\pi_{\mathcal{X}}^{-1}(x)\right)\right)}{d(x_0, x)} \\ &= \kappa[\pi_{\mathcal{Y}} \circ \pi_{\mathcal{X}}^{-1}](x_0)\end{aligned}$$

where d is the distance and $\pi_{\mathcal{X}}^{-1}$ is shorthand for $(\pi_{\mathcal{X}}|_{\mathcal{U}})^{-1}$. The map $\pi_{\mathcal{Y}} \circ \pi_{\mathcal{X}}^{-1}$ is the *solution map*.

Remark 2.7. Note that Definition 2.6 takes the limit over the domain $\pi_{\mathcal{X}}(\mathcal{U})$, i.e., all x corresponding to a point $(x, y) \in \mathcal{P}$ that is close to (x_0, y_0). This domain may have a lower dimension than $\mathcal{X}$, as in Figure 2.2. While Definition 2.6 is quite natural given the geometric approach above, it is a surprisingly uncommon intuition in numerical analysis that limits can be taken over a subset of $\mathcal{X}$. This can be a point of confusion when literature on geometric condition numbers is put side-by-side with literature in numerical linear algebra (where most problems are functions over an open subset of a linear space). If the domain is not all of

$\mathcal{X}$, then $\kappa[\mathcal{P}](x_0, y_0)$ is sometimes called a *structured condition number* [ANT19; GK93].

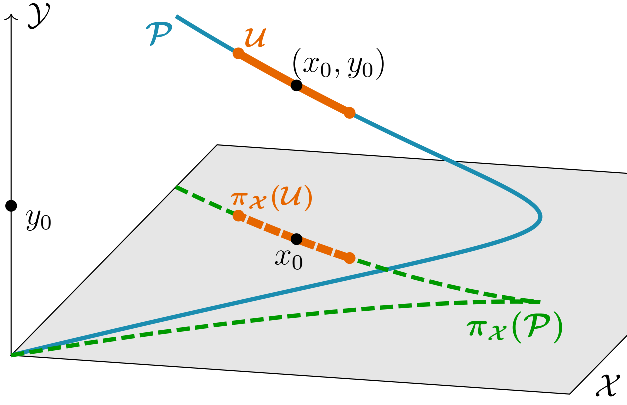


Figure 2.2: Example problem $\mathcal{P}$ whose domain $\pi_{\mathcal{X}}(\mathcal{P}) \subseteq \mathcal{X}$ has a strictly lower dimension than the input space $\mathcal{X}$. Even though $\pi_{\mathcal{X}}$ is not a homeomorphism between $\mathcal{U} \subseteq \mathcal{P}$ and an open subset of $\mathcal{X}$, it is a local embedding. Hence, the condition number is well-defined.

In general, the condition number in Definition 2.6 depends on both the input x_0 and the output y_0. For problems such as the one visualised in Figure 2.1, the input corresponds to multiple points (x_0, y_0) and (x_0, y_0') on $\mathcal{P}$, which do not necessarily have the same condition number. In the special case of function evaluation (Definition 2.1), the output depends uniquely on the input, so that the condition number is a function of only the input variable x_0.

The explicit dependence of the condition number on both x_0 and y_0 reminds us that computing the condition number of a problem typically requires finding a solution first. That is, to know how difficult a problem is numerically, one must first solve it. After a solution pair (x_0, y_0) is known, the condition number can be used to analyse how the problem behaves in a neighbourhood of (x_0, y_0).

2.3.2 Rice's theorem generalised

Theorem 2.4 gives a convenient expression for the computation of the condition number of a map. We wish to find a similar expression for general numerical problems, under some smoothness assumptions.

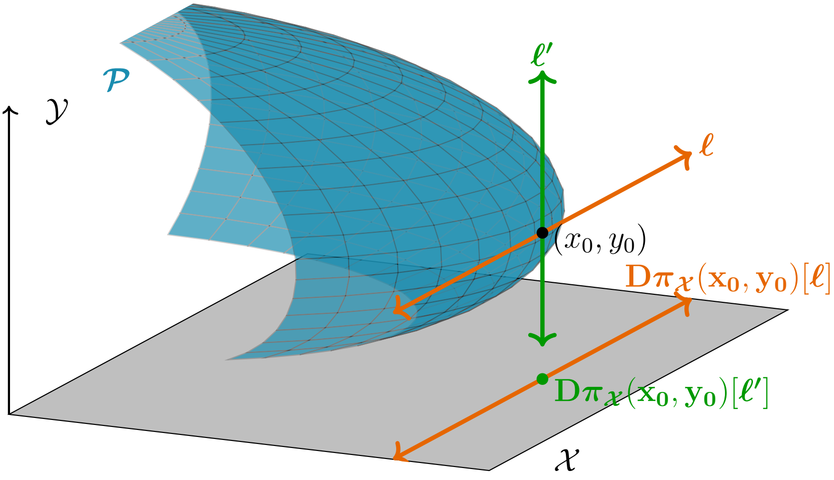

$\mathcal{P}$
$\mathcal{Y}$
ℓ'
ℓ
(x_0, y_0)
$\mathbf{D\pi_{\mathcal{X}}(x_0, y_0)[\ell]}$
$\mathbf{D\pi_{\mathcal{X}}(x_0, y_0)[\ell']}$
$\mathcal{X}$

Figure 2.3: Projections of lines tangent to a problem $\mathcal{P} \subseteq \mathcal{X} \times \mathcal{Y}$. For this problem, $\mathcal{X}$ and $\mathcal{Y}$ are linear spaces, so that they can be identified with $\mathcal{T}_{x_0}\mathcal{X}$ and $\mathcal{T}_{y_0}\mathcal{Y}$, respectively. At (x_0, y_0), the tangent space to $\mathcal{P}$ is spanned by ℓ and ℓ'. Its projection onto $\mathcal{T}_{x_0}\mathcal{X}$ has dimension one (rather than two) because ℓ' is parallel to $\mathcal{Y}$.

If $\mathcal{P}$ is smooth, the condition number turns out to have a geometric characterisation in terms of the slopes of lines tangent to $\mathcal{P}$. Any line ℓ tangent to $\mathcal{P}$ at (x_0, y_0) lives in the vector space $\mathcal{T}_{(x_0,y_0)}\mathcal{P} \subseteq \mathcal{T}_{x_0}\mathcal{X} \times \mathcal{T}_{y_0}\mathcal{Y}$. The projections from $\mathcal{T}_{(x_0,y_0)}\mathcal{P}$ onto $\mathcal{T}_{x_0}\mathcal{X}$ and $\mathcal{T}_{y_0}\mathcal{Y}$ are written formally as linear maps, i.e., $D\pi_{\mathcal{X}}(x_0, y_0)$ and $D\pi_{\mathcal{Y}}(x_0, y_0)$, respectively. The projection of ℓ onto $\mathcal{T}_{x_0}\mathcal{X}$ is either a line or a point. If it is a point (i.e., $\ell \in \ker D\pi_{\mathcal{X}}(x_0, y_0)$), we may visualise ℓ as being parallel to $\mathcal{T}_{y_0}\mathcal{Y}$, and linearly independent of $\mathcal{T}_{y_0}\mathcal{Y}$ otherwise, as Figure 2.3 illustrates.

Recall that, in the definition of the condition number, we invoked the *solution map*, which is defined if $\pi_{\mathcal{X}}$ is a local embedding. If there are no lines tangent to $\mathcal{P}$ at (x_0, y_0) which are parallel to $\mathcal{T}_{y_0}\mathcal{Y}$, then $\pi_{\mathcal{X}}$ is a local embedding (in fact, a diffeomorphsim). The following statement expresses this fact.

Lemma 2.8. *Let $\mathcal{X}$ and $\mathcal{Y}$ be smooth manifolds and let $\mathcal{P}$ be an embedded submanifold of $\mathcal{X} \times \mathcal{Y}$ containing a point (x_0, y_0). Write the projection onto the first component as $\pi_{\mathcal{X}} : \mathcal{P} \to \mathcal{X}, (x, y) \mapsto x$. If $\ker D\pi_{\mathcal{X}}(x_0, y_0) = \{0\}$, then the restriction of $\pi_{\mathcal{X}}$ to some neighbourhood $\mathcal{U}$ of (x_0, y_0) is a diffeomorphism onto its image.*

Proof. Combine [Lee13, Proposition 4.1] (to show that $\pi_{\mathcal{X}}$ is a local immersion) and the local embedding theorem [Lee13, Theorem 4.25]. □

Remark 2.9. Lemma 2.8 is a geometric analogue of the *implicit function theorem.* In real analysis, this theorem states the following: let $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n, (x, y) \mapsto F(x, y)$ be a smooth function such that $\frac{\partial}{\partial y} F(x, y)$ is invertible at some point $(x_0, y_0) \in F^{-1}(0)$. Then $F^{-1}(0)$ is locally the graph of some smooth function $H : x \mapsto y$ [Lee13, Theorem C.40]. In our case, Lemma 2.8 implies that $\mathcal{P}$ is locally the graph of $\pi_{\mathcal{Y}} \circ \pi_{\mathcal{X}}^{-1}$.

Proposition 2.10. *In the context of Lemma 2.8, define a Riemannian metric on $\mathcal{X}$ and $\mathcal{Y}$ and write $H_{x_0,y_0} := \pi_{\mathcal{Y}} \circ (\pi_{\mathcal{X}}|_{\mathcal{U}})^{-1}$. Consider $D\pi_{\mathcal{X}}(x_0, y_0)$ as a surjection onto its image. Then*

$$\kappa[\mathcal{P}](x_0, y_0) = \|DH_{x_0,y_0}(x_0, y_0)\| = \left\|D\pi_{\mathcal{Y}}(x_0, y_0) \circ D\pi_{\mathcal{X}}(x_0, y_0)^{-1}\right\| \tag{2.4}$$

where $\|\cdot\|$ is the operator norm with respect to the Riemannian metrics. The linear map $DH_{x_0,y_0}(x_0, y_0)$ is known as the condition map.

The condition map can be interpreted as follows. The projection of $\mathcal{T}_{(x_0,y_0)}\mathcal{P}$ onto $\mathcal{T}_{x_0}\mathcal{X}$ is a linear subspace $\mathbb{V}_{x_0} \subseteq \mathcal{T}_{x_0}\mathcal{X}$. If $D\pi_X(x_0, y_0)$ is injective, then every vector $\xi_x \in \mathbb{V}_{x_0}$ is the first component of a unique vector (ξ_x, ξ_y) tangent to $\mathcal{P}$ at (x_0, y_0). The condition map is given by $\xi_x \mapsto \xi_y$. In fact, this linear map generalises the concept of slope, as the following example shows.

Example 2.11 (Condition number of a plane curve)**.** Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and let $\mathcal{P} \subseteq \mathbb{R}^2$ be a smooth curve. For any (x_0, y_0), let ℓ be the line tangent to $\mathcal{P}$ at (x_0, y_0) and write its slope as m. Suppose that $m \neq \pm\infty$. Then, for any $\xi_x \in \mathbb{R}$, the unique vector in ℓ with ξ_x as its first component is $(\xi_x, m\xi_x)$. That is, the condition map is given by $\mathbb{R} \to \mathbb{R}, \xi \mapsto m\xi$. The spectral norm of this map is $|m|$.

2.3.3 Problems defined by equations

We have discussed how to characterise condition of a solution manifold $\mathcal{P}$ as a purely geometric property of $\mathcal{P}$ relative to the input and output space. Yet, numerical analysts would prefer a more computational description than Proposition 2.10. To find it, we want to express $\mathcal{P}$ (locally) as the zero set of equations over $\mathcal{X}$ and $\mathcal{Y}$. This is always possible in theory if $\mathcal{P}$ is an embedded submanifold of $\mathcal{X} \times \mathcal{Y}$ [Lee13, Theorem 5.8]. In practice, most numerical problems are *defined* in terms of equations.

Implicit problems

An *implicit problem* is defined as the problem of solving a system of equations $F(x,y)=c$, where c is some constant (typically 0). Under some assumptions, the theory above can be formulated in terms of the defining equations.

Proposition 2.12. *Let $\mathcal{X},\mathcal{Y}$, and $\mathcal{Z}$ be smooth manifolds and let $F:\mathcal{X}\times\mathcal{Y}\to\mathcal{Z},(x,y)\mapsto F(x,y)$ be a smooth map such that $\frac{\partial}{\partial y}F(x,y)$ is invertible for all (x,y). Consider any point (x_0,y_0,c) on the graph of F. Then $\mathcal{P}:=F^{-1}(c)$ is an embedded submanifold of $\mathcal{X}\times\mathcal{Y}$ of the same dimension as $\mathcal{X}$. Its condition map at (x_0,y_0) is*

$$D\pi_{\mathcal{Y}}(x_0,y_0)D\pi_{\mathcal{X}}(x_0,y_0)^{-1}=-\left(\frac{\partial}{\partial y}F(x_0,y_0)\right)^{-1}\frac{\partial}{\partial x}F(x_0,y_0). \tag{2.5}$$

The spectral norm of this map is $\kappa[\mathcal{P}](x_0,y_0)$.

Proof. The first statement follows from the submersion level set theorem [Lee13, Corollary 5.13]. Since F is constant over $\mathcal{P}$, every $\xi=(\xi_x,\xi_y)\in\mathcal{T}_{(x_0,y_0)}\mathcal{P}$ satisfies the relation

$$DF(x_0,y_0)[\xi]=\frac{\partial}{\partial x}F(x_0,y_0)[\xi_x]+\frac{\partial}{\partial y}F(x_0,y_0)[\xi_y]=0. \tag{2.6}$$

If $\xi_x=0$, the above implies that $\frac{\partial}{\partial y}F(x_0,y_0)[\xi_y]=0$ and therefore $\xi_y=0$. Thus, $\ker D\pi_{\mathcal{X}}(x_0,y_0)$ consists of only $(\xi_x,\xi_y)=(0,0)$. This means that $D\pi_{\mathcal{X}}(x_0,y_0)$ is injective, and since $\dim\mathcal{P}=\dim\mathcal{X}$, it is invertible.

Now, let ξ_x be an arbitrary vector in $\mathcal{T}_{x_0}\mathcal{X}$ and let (ξ_x,ξ_y) be the unique vector in $\mathcal{T}_{(x_0,y_0)}\mathcal{P}$ whose first component is ξ_x. Rearranging (2.6) gives

$$\xi_y=-\left(\frac{\partial}{\partial y}F(x_0,y_0)\right)^{-1}\frac{\partial}{\partial x}F(x_0,y_0)\xi_x$$

as required. The last statement follows from Proposition 2.10. □

Inverse problems

One important class of equations are *inverse problems*, i.e, equations of the form $G(y)=x$ for some map G. These equations rarely (if ever) receive special attention in texts on geometric condition numbers. Yet, their usefulness in data-driven applications cannot be overstated. Such equations arise when a model G with hidden parameters y produces measurable data x. The goal is to solve for the parameters, given the data. In the remainder of this section, we work out the condition number of a general inverse problem.

To find an expression of the condition number, we need the following definition.

Definition 2.13. Let $A\colon \mathbb{V} \to \mathbb{W}$ be a linear map between Euclidean spaces and write the projection onto $\operatorname{Im} A$ as $P_{\operatorname{Im} A}$. The *Moore–Penrose inverse* of A, written as $A^\dagger$, is the map $\hat{A}^{-1} P_{\operatorname{Im} A}$, where the bijection $\hat{A}\colon (\ker A)^\perp \to \operatorname{Im} A$ is defined by $x \mapsto Ax$.

We will transform our problem into one to which Proposition 2.12 applies and ultimately obtain an expression for the condition number. As usual, we assume that $\mathcal{X}$ and $\mathcal{Y}$ are Riemannian manifolds and G is smooth. The solution manifold is $\mathcal{P} := \{(G(y), y) \,|\, y \in \mathcal{Y}\}$, i.e., the transposition of the graph of G. Recall that, for the condition number to be defined at all, we want $\pi_{\mathcal{X}}$ to be a local embedding. By Lemma 2.8, this is guaranteed (and $\pi_{\mathcal{X}}$ is locally a *smooth* embedding) if the zero vector is the only vector tangent to $\mathcal{P}$ whose first component is zero. For inverse problems, this means that $\ker DG(y) = \{0\}$, i.e., G is a smooth immersion.

Next, we turn the equation $G(y) = x$ into something of the form $F(x, y) = 0$. Fix any point $(x_0, y_0) \in \mathcal{P}$. Since G is a local smooth embedding, there exists a neighbourhood $\mathcal{U}_{y_0}$ of y_0 such that $\widehat{\mathcal{X}} := G(\mathcal{U}_{y_0})$ is an embedded submanifold of $\mathcal{X}$ of the same dimension as $\mathcal{Y}$ [Lee13, Theorem 5.2]. Pick any chart ϕ of $\widehat{\mathcal{X}}$ centred at x_0 and define $F : \widehat{\mathcal{X}} \times \mathcal{Y} \to \mathbb{R}^r, (x, y) \to \phi(x) - \phi(G(y))$. In a neighbourhood of (x_0, y_0), we have $(x, y) \in \mathcal{P} \Leftrightarrow x = G(y) \Leftrightarrow F(x, y) = 0$.

Finally, we compute the partial derivatives of F. In the following, we define $D\hat{G}(y)$ as $DG(y)$ with its codomain restricted to $\operatorname{Im} DG(y)$. That is, at $y \in \mathcal{U}_{y_0}$, we can write $D\hat{G}(y) : \mathcal{T}_y\mathcal{Y} \to \mathcal{T}_{x_0}\widehat{\mathcal{X}}$. By the chain rule,

$$\frac{\partial}{\partial x} F(x, y) = D\phi(x) \quad \text{and} \quad \frac{\partial}{\partial y} F(x, y) = D\phi(G(y)) \circ D\hat{G}(y).$$

Note that the linear map $\frac{\partial}{\partial y} F(x, y) : \mathcal{T}_y\mathcal{Y} \to \mathbb{R}^{\dim \mathcal{Y}}$ has a domain and codomain of the same dimension and that it is injective. Hence, it is invertible. Note as well that $D\phi(G(y)) = D\phi(x)$ for all $(x, y) \in \mathcal{P}$, since $x = G(y)$. Hence, the right-hand side of (2.5) becomes

$$-\left(\frac{\partial}{\partial y} F(x_0, y_0)\right)^{-1} \frac{\partial}{\partial x} F(x_0, y_0) = -D\hat{G}(y_0)^{-1}.$$

The operator norm of this map is $\left\| D\hat{G}(y_0)^{-1} \right\| = \left\| DG(y_0)^\dagger \right\|$, which is equal to the reciprocal of the smallest singular value of $DG(y_0)$ [GVL13, Chapter 5]. Thus, we can apply Proposition 2.12 to obtain the following.

Proposition 2.14. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Riemannian manifolds and let $G : \mathcal{Y} \to \mathcal{X}$ be a smooth immersion. Consider the solution manifold $\mathcal{P} := \{(G(y), y) \mid y \in \mathcal{Y}\}$. At any point $(x_0, y_0) \in \mathcal{P}$, we have*

$$\kappa[\mathcal{P}](x_0, y_0) = \left\| DG(y_0)^\dagger \right\| = \frac{1}{\sigma_{\min}(DG(y_0))}$$

in which $\sigma_{\min}(DG(y_0))$ is the nth largest singular value of $DG(y_0)$, where $n = \dim \mathcal{Y}$.

Some of the major numerical problems considered in this thesis are *join decomposition problems*. For instance, additive tensor decompositions can be modelled as instances of these problems. Their condition number can be computed as follows.

Example 2.15 (Condition of join decompositions [BV18b])**.** For all $r = 1, \ldots, R$, let $\mathcal{Y}_r$ be an embedded Riemannian submanifold of $\mathbb{R}^N$. Let $n := \sum_{r=1}^{R} \dim \mathcal{Y}_r$. Define the *addition map*

$$\Sigma : \overbrace{\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_R}^{:=\mathcal{Y}} \to \overbrace{\mathbb{R}^N}^{:=\mathcal{X}}$$

$$(a_1, \ldots, a_R) \mapsto a_1 + \cdots + a_R$$

and the solution manifold $\mathcal{P} := \{(\Sigma(y), y) \mid y \in \mathcal{Y}\}$. At any point $y_0 = (a_1, \ldots, a_R)$, the differential $D\Sigma$ takes $(\dot{a}_1, \ldots, \dot{a}_R) \mapsto \dot{a}_1 + \cdots + \dot{a}_R$. We can represent it in coordinates as follows. For all $r = 1, \ldots, R$, let T_r be a matrix whose columns are an orthonormal basis of $\mathcal{T}_{a_r}\mathcal{Y}_r$. In coordinates, $D\Sigma(y_0) \cong T := [T_1 \quad T_2 \quad \cdots \quad T_R] \in \mathbb{R}^{N \times n}$. Let σ_n be the nth largest singular value of T. If $\sigma_n = 0$, then $\kappa[\mathcal{P}](\Sigma(y_0), y_0)$ is defined to be infinite. Otherwise, $\kappa[\mathcal{P}](\Sigma(y_0), y_0) = 1/\sigma_n$.

Remark 2.16. For any problem $\mathcal{P}$, the condition number measures the change in y with respect to small changes in x, with the important constraint that $x \in \pi_{\mathcal{X}}(\mathcal{P}) \subseteq \mathcal{X}$, i.e., $(x, y) \in \mathcal{P}$ has an exact solution. In practice, we may have noisy input x that does not lie in $\pi_{\mathcal{X}}(\mathcal{P})$, i.e., there may not be an exact solution to $x = G(y)$. In such cases, one often relaxes the equation to a minimisation problem $\min_y \|x - G(y)\|$. This is outside the scope of this thesis. Whilst the condition numbers studied above can all be described in terms of the orientation of $\mathcal{T}_{(x,y)}\mathcal{P}$ (i.e., a *linear* approximation of $\mathcal{P}$) relative to the input and output space, the condition number of finding $\min_y \|x - G(y)\|$ depends on the *curvature* of $\mathcal{P}$ as well. See [BV21] for a study of the condition number of these problems.

2.3.4 Summary

The geometric framework has allowed us to lift Rice's definition of the condition number and Rice's theorem to general classes of problems. Table 2.1 gives an overview of these different classes and the corresponding expressions for the condition number.

Problem	Condition number	Formal statement
$y = F(x)$	$\|DF(x_0)\|$	Theorem 2.4
$x = G(y)$	$\|DG(y_0)^\dagger\|$	Proposition 2.14
$F(x, y) = c$	$\left\|\left(\frac{\partial}{\partial y}F(x_0, y_0)\right)^{-1}\frac{\partial}{\partial x}F(x_0, y_0)\right\|$	Proposition 2.12
$(x, y) \in \mathcal{P}$	$\|D\pi_{\mathcal{Y}}(x_0, y_0)D\pi_{\mathcal{X}}(x_0, y_0)^{-1}\|$	Proposition 2.10

Table 2.1: Overview of condition numbers of different classes of numerical problems, under some smoothness assumptions.

2.4 Other aspects of condition

This section gives a few pointers to the literature on how condition numbers occur in other areas of computational mathematics besides sensitivity to numerical perturbations. Unlike the foregoing discussion on the sensitivity aspect of condition, I only give a rough sketch of the other aspects in this section. For more details and rigour, readers are directed to the literature referred to below. Readers may choose to skip this section, as it does not contain prerequisite knowledge for the remaining chapters.

2.4.1 Distance to ill-posedness

A major aspect of condition numbers is the study of so-called *condition number theorems*. For a problem $\mathcal{P} \subseteq \mathcal{X} \times \mathcal{Y}$, the *ill-posed locus* is the set of inputs $x \in \mathcal{X}$ such that $\mathcal{P}$ is singular at x by some broad definition. A condition number theorem is a result that expresses a relation (typically inverse proportionality) between the condition number of $\mathcal{P}$ at some point (x_0, y_0) and the distance from x_0 to the ill-posed locus. Some of the earliest such results were derived with a technique by Demmel [Dem87]. Arguably the most well-known condition number theorem is the following.

Theorem 2.17 (Eckart–Young theorem [Sch07; EY36; Mir60]). *Let $X_0 \in \mathbb{R}^{n\times n}$ be an invertible matrix. Then*

$$\min_{X:\ \det X=0} \frac{\|X - X_0\|}{\|X_0\|} = \frac{1}{\|X_0\|\|X_0^{-1}\|} \tag{2.7}$$

where $\|\cdot\|$ is any unitarily invariant norm.

Note that, if $\|\cdot\|$ is the spectral norm, the right-hand side is the reciprocal of Turing's condition number (Example 2.2).

A similar result exists for the solution of homogeneous polynomial systems. Every array x of coefficients with respect to the monomial basis defines a homogeneous polynomial map $p_x\colon \mathbb{R}^m \to \mathbb{R}^n$. We write the application of p_x as some function $F\colon (x, y) \mapsto p_x(y)$. Then the problem of solving homogeneous polynomial systems is characterised by the equation $F(x, y) = 0$.

The singular locus Σ_y of a point $y \in \mathbb{R}^m$ is defined as the set of all inputs x such that y is a root of p_x whose multiplicity is at least two. Then at any point (x, y) such that $x \notin \Sigma_y$, the distance from p_x to Σ_y in the so-called *Bombieri-Weyl metric* is inversely proportional to (a slight modification of) the condition number $\kappa[F^{-1}(0)](x, y)$ [BC13, Theorem 16.19].

Because of condition number theorems such as the ones above, it may be reasonable to define other quantities as the condition number if Definition 2.1 cannot be applied. For instance, consider the *(primal) feasibility problem* in linear optimisation: given $A \in \mathbb{R}^{m\times n}$, determine whether $\ker A$ contains a vector whose coordinates are all non-negative [NW06, Chapter 13]. The output of this problem is a boolean value (true or false), so that Definition 2.1 cannot meaningfully be applied.

Write the set of all matrices $A \in \mathbb{R}^{m\times n}$ that are feasible by the aforementioned definition as $\mathcal{F}$. For matrices A on the boundary $\partial\mathcal{F}$, arbitrarily small perturbations to A could make the input either feasible or infeasible depending on the direction of perturbation. In this context, Renegar [Ren95] defined the condition number at $A \notin \partial\mathcal{F}$ as $1/d(A, \partial\mathcal{F})$ where $\partial\mathcal{F}$ is the boundary of $\mathcal{F}$. This can be interpreted as an alternative measure of sensitivity. Every condition number studied in this dissertation, though, is an application of Rice's Definition 2.1 rather than Renegar's definition.

2.4.2 Computational complexity

The computational complexity analysis of iterative algorithms often involves complexity bounds based on the condition number of the associated problem.

Perhaps the most well-known such result is the convergence rate estimate of the conjugate gradient method. This method solves linear systems of the form $Ay = b$ with a positive definite matrix A by generating a sequence of iterates $y^{(1)}, y^{(2)}, \ldots$ that converge to the solution $A^{-1}b$. It is well-known that the error, as measured in the norm $\|y\|_A := \sqrt{y^T Ay}$, decreases as

$$\left\| y^{(k)} - A^{-1}b \right\|_A \leqslant 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \left\| y^{(1)} - A^{-1}b \right\|_A, \tag{2.8}$$

where $\kappa := \|A\| \|A^{-1}\|$ is the condition number of matrix inversion as in Example 2.2 [NW06, Section 5.1]. The convergence factor can be approximated as

$$\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} = 1 - \frac{2}{\sqrt{\kappa}} + \mathcal{O}\left(\frac{1}{\kappa}\right) \quad \text{as} \quad \kappa \to \infty.$$

Thus, the larger the condition number, the slower the expected convergence.

Another example of the connection between condition and complexity is in the context of *homotopy continuation*. This is a technique for solving systems of polynomial equations $p_x(y) = 0$, such as in Section 2.4.1. The basic approach is to trace a parametrised curve $\gamma(t) = (p_x^{(t)}, y^{(t)})$ from $t = 0$ to $t = 1$ where, for all t, $p_x^{(t)}$ is a polynomial and $p_x^{(t)}(y^{(t)}) = 0$. The curve γ is chosen such that the equation $p_x^{(0)}(y) = 0$ is easy to solve for y and such that $p_x^{(1)}$ is equal to p_x, i.e., the polynomial whose root is to be computed. The algorithm discretises the interval $[0, 1]$ as $\{t_0 = 0, t_1, \ldots, t_n = 1\}$ and proceeds as follows: for each $i = 1, \ldots, n$, the value $y^{(t_i)}$ is computed by solving $p_x^{(t_i)}(y) = 0$ using Newton's method, where $y^{(t_{i-1})}$ is used as an initial guess. The solution of $p_x(y) = 0$ is obtained when $i = n$. For certain choices of γ, the computational complexity of this algorithm is proportional to the integral of the (squared) condition number over γ [BC13, Theorem 17.3].

The results presented in this subsection suggest that iterative algorithms are not suitable for ill-conditioned problems. A workaround is to use *preconditioning*. Assuming that the computational problem is the evaluation of a map $F : \mathcal{X} \to \mathcal{Y}$, preconditioning is the act of decomposing $F = F_3 \circ F_2 \circ F_1$ in which F_1 and F_3 can be evaluated with a finite algorithm and F_2 is well-conditioned and can be evaluated with an iterative algorithm. This gives a three-step method for evaluating F such that (hopefully) none of the three steps have a computational complexity that is affected significantly by the condition number of the problem[1]. For example, a system of linear equations $Ay = b$ can be preconditioned as

[1] The term *preconditioning* can be somewhat misleading, as it does not actually reduce the inherent sensitivity of F. In fact, the sensitivity is a property of the problem itself that cannot be changed by an algorithm. Preconditioning is used purely for the sake of computational complexity.

$(MA)y = Mb$ where M is some matrix such that the inversion of MA is well-conditioned. Such preconditioning techniques are a major topic in numerical linear algebra [Saa03].

Chapter 3

Tensor decompositions

3.1 Prelude

I want to take a few paragraphs to digress on an abstract debate into which I was involuntarily thrown several years ago and which continued to haunt me ever since.

This dissertation is the product of work between different research communities: engineers and mathematicians. While both work with tensors, they cannot seem to agree on what a tensor even is. In his notorious *pickle paragraph* [Lan12], J.M. Landsberg recounts a comparable experience: *"In the course of preparing this book I have been fortunate to have had many discussions with computer scientists, applied mathematicians, engineers, physicists, and chemists. Often the beginnings of these conversations were very stressful to all involved. [...] While [geometers and scientists] are interested in communicating, there are language and even philosophical barriers to be overcome."*

These differing views are partially explained by different interests: engineers prefer concrete, tangible representations of mathematical objects. In this context, *tangible* means amenable to numerical computations. By contrast, algebraists endeavour to detach an object's representation from its algebraic relations. To the algebraist, the rules an object obeys convey its essence, whereas its concrete representations are merely a costume in which it presents itself. In our case, the quarrel is over whether to represent a tensor as a numerical array or as a point in an abstract space.

While the mathematical perspective is arguably more complete and succinct, its insistence on concealing any and all concrete representations is at times wholly unhelpful for lay audiences. Indeed, when the goal is clear communication, we sometimes want to call an abstract thing by a concrete name and get on with it.

It is for this reason that I am not committed to any definition of tensors in particular. To balance both perspectives, this chapter is written mostly in abstract terms, with indications as to how tensors can be represented numerically. The numerical perspective is predominant throughout the remaining chapters.

3.2 Tensors

We start with an abstract definition of tensors in terms of multilinear maps and derive some basic properties, following Greub's book [Gre78]. Afterwards, we present some common concrete representations of tensors.

3.2.1 The abstract tensor product

Definition 3.1. Let $\mathbb{V}_1, \dots, \mathbb{V}_D, \mathbb{W}$ be vector spaces. A map $f : \mathbb{V}_1 \times \cdots \times \mathbb{V}_D \to \mathbb{W}$ is *multilinear* if fixing any $D-1$ arguments of f gives a linear map in the remaining argument.

Definition 3.2. Let $\mathbb{V}_1, \dots, \mathbb{V}_D, \mathbb{T}$ be vector spaces. A multilinear map $\otimes : \mathbb{V}_1 \times \cdots \times \mathbb{V}_D \to \mathbb{T}$ is a *tensor product* if, for every multilinear map $f : \mathbb{V}_1 \times \cdots \times \mathbb{V}_D \to \mathbb{W}$, there is a unique linear map $L : \mathbb{T} \to \mathbb{W}$ such that $f = L \circ \otimes$. This relation can be expressed as the following commutative diagram:

$$\begin{tikzcd}
\mathbb{V}_1 \times \cdots \times \mathbb{V}_D \arrow[r, "f"] \arrow[d, "\otimes"] & \mathbb{W} \\
\mathbb{T} \arrow[ur, "L"] &
\end{tikzcd} \tag{3.1}$$

In this case, $\mathbb{T}$ is a *tensor product space* of $\mathbb{V}_1 \times \cdots \times \mathbb{V}_D$, written as $\mathbb{T} = \mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$, and its elements are *tensors*.

Tensor products are usually written with infix notation, i.e, $v_1 \otimes \cdots \otimes v_D := \otimes(v_1, \dots, v_D)$. The property defining the tensor product, illustrated by (3.1), is known as a *universal property*, the general form of which comes from category theory. Universal properties can also be used to define free groups [Gal20, Chapter 25] and quotient and product topologies [Mun14], to name a few examples.

Definition 3.2 is not vacuous, since the following abstract tensor product satisfies the universal property [Gre78].

Definition 3.3. The *formal tensor product* of D vector spaces $\mathbb{V}_1, \ldots, \mathbb{V}_D$ is the set of all formal expressions $\sum_{r=1}^{R} v_1^r \otimes \cdots \otimes v_D^r$ in which $v_d^r \in \mathbb{V}_d$ for all d and $v_1, \ldots, v_D \mapsto v_1 \otimes \cdots \otimes v_D$ is multilinear.

The statement that $\otimes$ is multilinear means, for example, that the formal expression $(\lambda u + \mu v) \otimes w$ is considered equivalent to the expression $\lambda(u \otimes w) + \mu(v \otimes w)$. Since the definition of the formal tensor product does not impose any properties other than multilinearity, it can be considered the *least specific* or *most rudimentary* of all multilinear maps.

It is not surprising that the formal tensor product satisfies the universal property. Indeed, (3.1) reads that any multilinear expression in $v_1, \ldots, v_D$ can be formed by applying linear operations to the most rudimentary multilinear expression in $v_1, \ldots, v_D$, i.e., their formal tensor product.

The following theorem states that all tensor products are equivalent to the formal tensor product. That is, all tensor products are “minimally specific” multilinear maps.

Theorem 3.4 (Uniqueness of the tensor product)**.** *Let $\mathbb{V}_1, \ldots, \mathbb{V}_D$ be vector spaces and let $\otimes$ and $\widetilde{\otimes}$ be tensor products over $\mathbb{V}_1 \times \cdots \times \mathbb{V}_D$. Then there exists a unique linear isomorphism A between $\mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$ and $\mathbb{V}_1 \widetilde{\otimes} \ldots \widetilde{\otimes} \mathbb{V}_D$ such that $\widetilde{\otimes} = A \circ \otimes$.*

Proof. Since $\otimes$ is a tensor product and $\widetilde{\otimes}$ is multilinear, there exists a unique linear map A such that $\widetilde{\otimes} = A \circ \otimes$. Likewise, $\otimes = B \circ \widetilde{\otimes}$ for a unique linear map B. By combining these observations, we find that $\otimes = B \circ A \circ \otimes$. By setting $f := \otimes$ in Definition 3.2, we find that the identity map is the unique linear map L such that $\otimes = L \circ \otimes$. Thus, $B \circ A = \mathrm{Id}$. By the same argument, $B \circ A = \mathrm{Id}$. Hence, A is invertible. □

We may summarise the discussion up until this point as follows:

1. the tensor product is multilinear,
2. the tensor product satisfies the universal property (3.1),
3. the formal tensor product satisfies points 1 and 2, and
4. all maps satisfying points 1 and 2 are equivalent.

The fourth point is especially potent, as it allows us to switch between different representations of a tensor without losing the essential information. In the next subsection, we present some more concrete tensor products and illustrate their equivalence.

3.2.2 Equivalence of concrete tensor products

Several multilinear operations other than the formal tensor product satisfy Definition 3.2. The following is a list of alternative definitions, which can all be said to define *the* tensor product in their respective contexts. The fact that they are all tensor products in the sense of Definition 3.2 is shown in standard references such as [Gre78; Lan12].

- (Contraction) For $v_1, \ldots, v_D \in \mathbb{V}_1 \times \cdots \times \mathbb{V}_D$, the tensor product is a multilinear map

  $$\begin{aligned} v_1 \otimes \cdots \otimes v_D \colon \quad & \mathbb{V}_1^* \times \cdots \times \mathbb{V}_D^* \to \mathbb{K} \\ & (\alpha_1, \ldots, \alpha_D) \mapsto \prod_{d=1}^{D} \alpha_d(v_d) \end{aligned} \tag{3.2}$$

  where $\mathbb{V}_d^* = \mathrm{Hom}(\mathbb{V}_d, \mathbb{K})$ is the dual space of $\mathbb{V}_d$. This definition is extended linearly for general tensors, i.e., a sum of tensor products is a sum of multilinear maps over $\mathbb{V}_1^* \times \cdots \times \mathbb{V}_D^*$.

- (Euclidean contraction) If an inner product $\langle \cdot, \cdot \rangle$ is defined on $\mathbb{V}_1, \ldots, \mathbb{V}_D$, we may identify

  $$\begin{aligned} v_1 \otimes \cdots \otimes v_D \colon \quad & \mathbb{V}_1 \times \cdots \times \mathbb{V}_D \to \mathbb{K} \\ & (u_1, \ldots, u_d) \mapsto \prod_{d=1}^{D} \langle u_d, v_d \rangle . \end{aligned}$$

  In finite dimensional inner product spaces $\mathbb{V}_d$, there is an identification $\mathbb{V}_d \cong \mathbb{V}_d^*$. That is, the covector $\alpha_d \in \mathbb{V}_d^*$ associated with $u_d \in \mathbb{V}_d$ is the linear map $v_d \mapsto \langle u_d, v_d \rangle$. Thus, the Euclidean contraction is a special case of the above.

- (Outer product) Let $v_1 \in \mathbb{K}^{n_1}, \ldots, v_D \in \mathbb{K}^{n_D}$. Then $v_1 \otimes \cdots \otimes v_D$ is an array with D indices whose element at the index $(i_1, \ldots, i_D)$ is $v_1^{i_1} \cdots v_D^{i_D}$, where $v_d^{i_d}$ is the i_dth coordinate of v_d, for all $d = 1, \ldots, D$. With this interpretation, we can identify tensors with multi-indexed arrays, i.e.,

  $$\mathbb{K}^{n_1} \otimes \cdots \otimes \mathbb{K}^{n_D} \cong \mathbb{K}^{n_1 \times \cdots \times n_D}.$$

- (Multilinear multiplication) If $A_1, \dots, A_D$ are linear maps defined on $\mathbb{V}_1, \dots, \mathbb{V}_D$, respectively, their tensor product is the linear map over $\mathbb{V}_1 \otimes \dots \otimes \mathbb{V}_D$ defined by

  $$(A_1 \otimes \dots \otimes A_D)(v_1 \otimes \dots \otimes v_D) = (A_1 v_1) \otimes \dots \otimes (A_D v_D).$$

  For a tensor $\mathcal{B} \in \mathbb{V}_1 \otimes \dots \otimes \mathbb{V}_D$, this operation is also written as $(A_1, \dots, A_D) \cdot \mathcal{B}$. Multilinear multiplication is colloquially referred to as *multiplying $\mathcal{B}$ by a matrix on each side* [KB09], since, if $B \in \mathbb{K}^{m \times n}$ is a matrix, then $(A_1, A_2) \cdot B = A_1 B A_2^T$.

- (Kronecker product) The *traditional Kronecker product* of $A \in \mathbb{K}^{m_1 \times n_1}$ and $B \in \mathbb{K}^{m_2 \times n_2}$ is

  $$A \otimes_K B = \begin{bmatrix} a_{11}B & \dots & a_{1n_1}B \\ \vdots & \ddots & \vdots \\ a_{m_11}B & \dots & a_{m_1n_1}B \end{bmatrix}.$$

  The *reversed Kronecker product* is $A \otimes B := B \otimes_K A$. We lift this definition to n-tuples (i.e., numerical vectors) by identifying $\mathbb{K}^n \cong \mathbb{K}^{n \times 1}$.

In the spirit of Theorem 3.4, all tensor products in this list will be written with the symbol $\otimes$ and all objects in the span of these tensor products can be called tensors. If it is essential to use a particular interpretation of the tensor product, this interpretation will be specified wherever necessary.

Multilinear multiplication and the Kronecker product

The equivalence between some of the aforementioned tensor products is illustrated by the so-called *vec trick*, a name coined in K. Borgwardt's doctoral thesis on kernel machines [Bor07]. This is not really a trick, so much as a recognition of the fact that the Kronecker product and the multilinear multiplication operator are both tensors whose representations are isomorphic to each other. This uses the *vectorisation operator*, which we define next.

If we temporarily write the outer product and reversed Kronecker product as $\otimes_{\text{out}}$ and $\oslash$, respectively, then the vectorisation operator $\text{vec} : \mathbb{K}^{n_1 \times \dots \times n_D} \to \mathbb{K}^{n_1 \times \dots \times n_D}$ is defined as the unique linear isomorphism vec such that $\oslash = \text{vec} \circ \otimes_{\text{out}}$. Equivalently, we can define vec as reversing the order of the indices of an array $\mathcal{B}$ and listing the components in lexicographic order. For matrices, this corresponds to vertical stacking of the columns. This is the way most numerical programming languages, such as Julia, NumPy, and GNU Octave implement vectorisation. The *vec trick*, then, is the identity

$$\operatorname{vec}\left((A_1, \ldots, A_D) \cdot \mathcal{X}\right) = (A_1 \otimes_R \cdots \otimes_R A_D) \operatorname{vec} \mathcal{X}$$

for any tensor $\mathcal{X} \in \mathbb{K}^{n_1 \times \cdots \times n_D}$ and matrices $A_d \in \mathbb{K}^{m_d \times n_d}$ for $d = 1, \ldots, D$.

Contraction and multi-indexed arrays

Similarly to the previous example, it is easy to translate between multilinear maps (i.e., contractions) and multi-indexed arrays. For each $d = 1, \ldots, D$, we pick a basis $\{e^d_j\}_{j=1}^{\dim \mathbb{V}^*_d}$ of $\mathbb{V}^*_d$. For a given multilinear map ϕ we construct the D-indexed array $\mathcal{A}_\phi$ whose coordinates are $a_{i_1,\ldots,i_D} := \phi(e^1_{i_1}, \ldots, e^D_{i_D})$.

To see that this construction is invertible, write an arbitrary $\beta = (\beta_1, \ldots, \beta_D) \in \mathbb{V}^*_1 \times \cdots \times \mathbb{V}^*_D$ in coordinates as $\beta_d = \sum_{i_d=1}^{\dim \mathbb{V}^*_d} b_{i_d} e^d_{i_d}$. By multilinearity of ϕ, we have

$$\begin{aligned} \phi(\beta) &= \phi\left(\sum_{i_1=1}^{\dim \mathbb{V}^*_1} b_{i_1} e^1_{i_1}, \ldots, \sum_{i_D=1}^{\dim \mathbb{V}^*_D} b_{i_D} e^D_{i_D}\right) \\ &= \sum_{i_1=1}^{\dim \mathbb{V}^*_1} \cdots \sum_{i_D=1}^{\dim \mathbb{V}^*_D} b_{i_1} \cdots b_{i_D} \underbrace{\phi(e^1_{i_1}, \ldots, e^D_{i_D})}_{a_{i_1,\ldots,i_D}}, \end{aligned} \tag{3.3}$$

which expresses $\phi(\beta)$ only in terms of the coordinates of β and $a_{i_1,\ldots,i_D}$. Thus, we can invert the construction $\phi \mapsto \mathcal{A}_\phi$ as follows. Given a D-indexed array $\mathcal{A}$ with components $a_{i_1,\ldots,i_D}$ we can associate it with a unique multilinear map $\phi_{\mathcal{A}}$, defined by (3.3). It can be verified that the coordinate array corresponding to the multilinear map (3.2) is the outer product of $v_1, \ldots, v_D$. This shows that the contraction tensor product and outer product are equivalent up to a choice of basis.

Flattenings

One more family of representations of tensors are the *standard flattenings*. These turn a tensor $\mathcal{A} \in \mathbb{K}^{n_1 \times \cdots \times n_D}$ in coordinates into a matrix of size $n_d \times \prod_{d' \neq d} n_{d'}$ for some $d \in \{1, \ldots, D\}$. The *dth standard flattening* can be defined through the universal property of the tensor product as the unique linear map that satisfies

$$(\cdot)_{(d)} \colon\ a_1 \otimes \cdots \otimes a_D \mapsto a_d \operatorname{vec}\left(a_1 \otimes \cdots \otimes a_{d-1} \otimes a_{d+1} \otimes \cdots \otimes a_D\right)^T.$$

More concretely, if $\mathcal{A}$ is a coordinate array with indices $(i_1, \ldots, i_D)$, then for all $i_d = 1, \ldots, n_d$, the i_dth row of $\mathcal{A}_{(d)}$ can be formed by taking all coordinates of $\mathcal{A}$ where dth index is i_d and listing them in reverse lexicographic order.

3.3 Decompositions

A frequently encountered problem in applications is to find *tensor decompositions*. These come in two varieties: additive and multiplicative. A good overview of tensor decompositions from a numerical perspective can be found in the review by Kolda and Bader [KB09]. First, we look at additive decompositions.

3.3.1 Polyadic decomposition

Definition 3.5. A tensor $\mathcal{A} \in \mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$ is a *rank-1 tensor* or a *polyad* if it can be expressed as a tensor product of some vectors $v_1, \ldots, v_D \in \mathbb{V}_1 \times \cdots \times \mathbb{V}_D$. The set of all rank-1 tensors in $\mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$ is the *Segre manifold over* $\mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$.

Definition 3.6 ([Hit27])**.** A *polyadic decomposition* of a tensor $\mathcal{A} \in \mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$ is a set of rank-1 tensors $\{\mathcal{A}_1, \ldots, \mathcal{A}_R\}$ whose sum is $\mathcal{A}$. The cardinality of this set is the *length* of the decomposition. The minimal R such that $\mathcal{A}$ has a polyadic decomposition of length R is the *rank* of $\mathcal{A}$. By convention the rank of the zero tensor is zero. A polyadic decomposition of $\mathcal{A}$ of minimal length is a *canonical polyadic decomposition (CPD).*

By Definition 3.3, every tensor has a polyadic decomposition. Computing a polyadic decomposition of $\mathcal{A}$ of length R is equivalent to finding R vectors $v_1^d, \ldots, v_R^d \in \mathbb{V}_d$ for each $d = 1, \ldots, D$ such that

$$\mathcal{A} = \sum_{r=1}^{R} v_r^1 \otimes \cdots \otimes v_r^D. \tag{3.4}$$

In this equivalent problem, each rank-1 tensor $\mathcal{A}_r$ in the expression $\mathcal{A} = \sum_{r=1}^{R} \mathcal{A}_r$ is factorised as a tensor product. Some texts define the polyadic decomposition as the factors v_r^d, but for our definition, the rank-1 terms $\mathcal{A}_r$ need not be presented in factorised form.

The characterisation (3.4) is more ambiguous than Definition 3.6 because, if (3.4) holds and some set of scalars $\{\lambda_{r,d}\}_{r=d=1}^{R,D}$ satisfies $\prod_{d=1}^{D} \lambda_{r,d} = 1$ for all r, then the following is an alternative decomposition of $\mathcal{A}$:

$$\mathcal{A} = \sum_{r=1}^{R} \left(\lambda_{r,1} v_r^1\right) \otimes \cdots \otimes \left(\lambda_{r,D} v_1^D\right).$$

That is, the parametrisation of a rank-1 tensor as a tensor product is unique up to a choice of $D-1$ scalars.

Applying Definition 3.6 to matrices, a polyadic decomposition of $A \in \mathbb{K}^{n_1 \times n_2}$ is a set of rank-1 matrices $u_1 v_1^T, \ldots, u_R v_R^T$ such that $A = \sum_{r=1}^{R} u_r v_r^T = [u_1 \cdots u_R]\,[v_1 \cdots v_R]^T$. The minimal R such that this equality can be satisfied is the rank of A (as defined in linear algebra). Thus, tensor rank generalises matrix rank.

The consistency between matrix and tensor rank prompts the question: *does every tensor in* $\mathbb{K}^{n\times\cdots\times n}$ *of order D have rank at most n?* For matrices, this is true. However, by the following counting argument, this cannot be true for sufficiently large tensors of higher order. By definition, the set of tensors of rank at most R is contained in the image of the polynomial map

$$F\colon\ (\mathbb{K}^n)^{RD} \to \mathbb{K}^{n\times\cdots\times n}$$

$$(v_1^1, v_1^2, \ldots, v_R^D) \mapsto \sum_{r=1}^{R} v_r^1 \otimes \cdots \otimes v_r^D.$$

The image of F cannot contain a set of higher dimension than its domain[1], which has dimension nRD. For example, let $R = n$ and $D = 3$. Then the set of rank-n tensors cannot contain a space of dimension more than $3n^2$. Thus, for $n > 3$, it cannot contain all of $\mathbb{K}^{n\times n\times n}$.

An early result on the rank of general tensors is the following.

Theorem 3.7 (Strassen–Lickteig, Theorem 3.1.4.3 in [Lan12])**.** *All tensors in* $\mathbb{C}^{n\times n\times n}$ *outside of some set of measure zero have rank* $\left\lceil \frac{n^3}{3n-2} \right\rceil$*, except if* $n = 3$*. In that case, the count should be increased by 1.*

3.3.2 Symmetric tensor decomposition

Some tensors found in applications obey certain symmetries. In this case, one is usually interested in decompositions that preserve these symmetries. Some

[1] For sharper estimates of the dimension of the image of F, see Section 3.4, in particular the definition of *expected dimension*, or [Lan12, chapter 3].

basic definitions relating to symmetric tensors are presented below. A more extensive introduction was provided by Comon et al. [Com+08].

Definition 3.8. For a finite set $S \subseteq \mathbb{N}$, the *Dth symmetric group* $\mathfrak{S}(S)$ is the group of all permutations of S. The permutation swapping i and j is written as $\sigma_{i\leftrightarrow j}$. If $\mathbb{V}_1, \dots, \mathbb{V}_D$ are vector spaces, every $\sigma \in \mathfrak{S}(\{1, \dots, D\})$ corresponds to a linear map defined by

$$\begin{aligned} \sigma : \quad & \mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D \to \mathbb{V}_{\sigma(1)} \otimes \cdots \otimes \mathbb{V}_{\sigma(D)} \\ & v_1 \otimes \cdots \otimes v_D \to v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(D)}. \end{aligned}$$

If we interpret $\mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$ as $\mathbb{K}^{n_1 \times \cdots \times n_D}$, then for any array $\mathcal{A}$ with coordinates $a_{i_1,\dots,i_D}$, its permutation $\sigma(\mathcal{A})$ has coordinates $a_{\sigma(i_1),\dots,\sigma(i_D)}$. For example, if $D = 2$, we get $\mathfrak{S}(\{1,2\}) = \{\mathrm{Id}, \sigma_{1\leftrightarrow 2}\}$. The associated permutations on $\mathbb{K}^{n_1 \times n_2}$ are $\mathrm{Id} : A \mapsto A$ and $\sigma_{1\leftrightarrow 2} : A \mapsto A^T$.

Definition 3.9. Let $H \neq \{\mathrm{Id}\}$ be a subgroup of $\mathfrak{S}(\{1, \dots, D\})$ that is generated by pairwise swaps $\sigma_{i\leftrightarrow j}$. A tensor $\mathcal{A} \in \mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$ is *partially symmetric* or *symmetric relative to* H if $\sigma(\mathcal{A}) = \mathcal{A}$ for all $\sigma \in H$. If $\sigma(\mathcal{A}) = \mathcal{A}$ for all $\sigma \in \mathfrak{S}(\{1, \dots, D\})$, then $\mathcal{A}$ is *symmetric*.

Note that a tensor in $\mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$ can only be invariant under $\sigma_{i\leftrightarrow j}$ if $\mathbb{V}_i = \mathbb{V}_j$. Therefore, the only tensor spaces in which symmetric tensors exist are of the form $\mathbb{V} \otimes \cdots \otimes \mathbb{V}$, abbreviated as $\mathbb{V}^{\otimes D}$. Likewise, the tensor product of D copies of $v \in \mathbb{V}$ will be written as $v^{\otimes D}$.

A basic property is the following:

Proposition 3.10 (see e.g. [Com+08]). *Given a vector space* $\mathbb{V}$*, the set of symmetric tensors* $\mathrm{Sym}(\mathbb{V}, D) \subseteq \mathbb{V}^D$ *is a linear space generated by symmetric rank-1 tensors, i.e.,* $\mathrm{Sym}(\mathbb{V}, D) = \mathrm{span}\left\{v^{\otimes D} \,\middle|\, v \in \mathbb{V}\right\}$.

The above property ensures that, if we want to decompose a symmetric tensor $\mathcal{A}$ as a sum of rank-1 tensors, i.e., $\mathcal{A} = \sum_{r=1}^{R} \mathcal{A}_r$, we can impose that all $\mathcal{A}_r$ are symmetric. This kind of polyadic decomposition is called a *Waring decomposition*. If we factor the $\mathcal{A}_r$, this gives

$$\mathcal{A} = \sum_{r=1}^{R} \lambda_r v_r^{\otimes D}$$

where $\lambda_r \in \mathbb{K}$ and $v_r \in \mathbb{V}$.

Partially symmetric tensors have an analogous characterisation, but it is more complicated notationally. Let H be the symmetry group acting on the tensor space $\mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$, as in Definition 3.9. Consider the graph G with nodes $\{1, \dots, D\}$ where i and j are connected if $\sigma_{i \leftrightarrow j} \in H$. Since H is generated by pairwise swaps, G is a union of cliques[2] $C_1, \dots, C_K$. We can relabel the nodes $\{1, \dots, D\}$ such that C_1 consists of nodes 1 to d_1 for some d_1, the second clique consists of d_2 nodes starting at node $d_1 + 1$, and so on. This relabelling corresponds to a permutation π of the tensor space $\mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$. If a tensor $\mathcal{A} = w_1 \otimes \cdots \otimes w_D$ is invariant under H, then (up to relabelling), $\mathcal{A}$ must be of the form $\lambda v_1^{\otimes 1} \otimes \cdots \otimes v_K^{\otimes d_K}$ for some scalar λ and vectors $w_1, \dots, v_K$.

By the preceding argument, the space of partially symmetric tensors is the tensor product of spaces of symmetric tensors. Since symmetric tensors are generated by symmetric rank-1 tensors, partially symmetric tensors are generated by their tensor products. In other words, if $\mathcal{A}$ is partially symmetric, we may write it as

$$\mathcal{A} = \sum_{r=1}^{R} \lambda_r (v_1^r)^{\otimes 1} \otimes \cdots \otimes (v_K^r)^{\otimes d_K}$$

for some choice of R, scalars λ_r, and vectors v_k^r with $k = 1, \dots, K$ and $r = 1, \dots, R$. Such a decomposition is called a *partially symmetric* decomposition.

3.3.3 Tucker decomposition

The Tucker decomposition [Tuc66] is a commonly used model that can be computed by a combination of flattenings and matrix decompositions. Put simply, a Tucker decomposition of a tensor $\mathcal{A}$ is a choice of basis together with the corresponding coordinates of $\mathcal{A}$.

Definition 3.11. Let $\mathcal{A} \in \mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$. For $d = 1, \dots, D$, let $U_d = \{u_d^1, \dots, u_d^{r_d}\}$ be a set of vectors spanning $\mathbb{V}_d$, and let $\mathcal{C} \in \mathbb{K}^{r_1 \times \cdots \times r_D}$ be an array of coordinates of $\mathcal{A}$ with respect to $U_1, \dots, U_D$, that is,

$$\mathcal{A} = \sum_{i_1, \dots, i_D = 1}^{r_1, \dots, r_D} \mathcal{C}_{i_1, \dots, i_D} (u_D^{i_1} \otimes \cdots \otimes u_D^{i_D}).$$

Then the tuple $(\mathcal{C}, U_1, \dots, U_D)$ is a *Tucker decomposition* of $\mathcal{A}$.

If we identify each $U_d = \{u_d^1, \dots, u_d^{r_d}\}$ with the application of coordinates, i.e., the map $(\alpha_1, \dots, \alpha_{r_d}) \mapsto \sum_{i=1}^{r_d} \alpha_i u_d^i$, then we may write the Tucker decomposition as $\mathcal{A} = (U_1 \otimes \cdots \otimes U_D)\mathcal{C}$ where $(U_1 \otimes \cdots \otimes U_D)$ is the multilinear multiplication operator (Section 3.2.2). In the numerical literature, this is also written as $\mathcal{C} \times_1 U_1 \cdots \times_D U_D$ [KB09, §4].

[2] Recall that a *clique* is a graph in which all nodes are connected to each other.

It is clear from the definition that the Tucker decomposition depends on an explicit choice of basis. Hence, the Tucker decomposition of a given tensor can never be unique. For each $d = 1, \ldots, D$, let $G_d \in \mathrm{GL}(\mathbb{V}_d)$ be a change of basis. If $\mathcal{A} = (U_1 \otimes \cdots \otimes U_D)\mathcal{C}$, then

$$\mathcal{A} = (\underbrace{U_1 G_1}_{=:\tilde{U}_1} \otimes \cdots \otimes \underbrace{U_D G_D}_{\tilde{U}_D}) \underbrace{\big((G_1^{-1} \otimes \cdots \otimes G_D^{-1})\mathcal{C}\big)}_{=:\widetilde{\mathcal{C}}}$$

which yields the alternative Tucker decomposition $(\widetilde{\mathcal{C}}, \tilde{U}_1, \ldots, \tilde{U}_D)$ of $\mathcal{A}$.

If $U_1, \ldots, U_D$ are all linearly independent sets, the coordinates $\mathcal{C}$ are uniquely determined given $\mathcal{A}$, and the only indeterminacy of the Tucker decomposition (with fixed cardinalities of $U_1, \ldots, U_D$) is a choice of basis.

To define the notion of rank associated with the Tucker decomposition, we use the following lemma.

Lemma 3.12. *For $d = 1, \ldots, D$, let $\mathbb{V}_d$ be a vector space. Let $\mathbb{U}_1$ and $\mathbb{U}_2$ be linear subspaces of $\mathbb{V}_1$ and $\mathbb{V}_2$, respectively. Then*

$$\mathbb{T} := (\mathbb{U}_1 \otimes \mathbb{V}_2 \otimes \cdots \otimes \mathbb{V}_D) \cap (\mathbb{V}_1 \otimes \mathbb{U}_2 \otimes \cdots \otimes \mathbb{V}_D) = \mathbb{U}_1 \otimes \mathbb{U}_2 \otimes \cdots \otimes \mathbb{V}_D$$

and likewise for subspaces $\mathbb{U}_i$ of other factors $\mathbb{V}_i$ with $1 \leqslant i \leqslant D$.

Proof. It is obvious that $\mathbb{U}_1 \otimes \mathbb{U}_2 \otimes \cdots \otimes \mathbb{V}_D \subseteq \mathbb{T}$. For the reverse inclusion, let $\mathcal{A} \in \mathbb{T}$ be a tensor decomposed as $\mathcal{A} = \sum_{r=1}^{R} v_1^r \otimes u_2^r \otimes v_3^r \otimes \cdots \otimes v_D^r$, where, for each r, $v_d^r \in \mathbb{V}_d$ for all d and $u_2^r \in \mathbb{U}_2$. If $P_{\mathbb{U}_1}$ is the projection from $\mathbb{V}_1$ onto $\mathbb{U}_1$, then $\mathcal{A} = (P_{\mathbb{U}_1} \otimes \mathrm{Id} \otimes \cdots \otimes \mathrm{Id})\mathcal{A}$ by assumption. By substituting the decomposition of $\mathcal{A}$ into the right-hand side of this identity, we see that $\mathcal{A} = \sum_{r=1}^{R} u_1^r \otimes u_2^r \otimes v_3^r \otimes \cdots \otimes v_D^r$ for some vectors $u_1^r \in \mathbb{U}_1$. This proves the first statement. The other statement can be shown analogously by permuting the factors of the tensor product. □

This lemma ensures that the following is well-defined.

Definition 3.13. Let $\mathcal{A} \in \mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$ be a tensor. If the smallest subspaces $\mathbb{U}_d \subseteq \mathbb{V}_d$ such that $\mathcal{A} \in \mathbb{U}_1 \otimes \cdots \otimes \mathbb{U}_D$ are $\mathbb{U}_d = \mathbb{V}_d$ for all d, then the *multilinear rank* of $\mathcal{A}$ is $\operatorname{mlrank} \mathcal{A} := (\dim \mathbb{V}_1, \ldots, \dim \mathbb{V}_D)$. Equivalently, the multilinear rank of a tensor $\mathcal{A}$ is the tuple of minimal integers $(r_1, \ldots, r_d)$ such that $\mathcal{A}$ has a Tucker decomposition $(\mathcal{C}, U_1, \ldots, U_D)$ with $\# U_d = r_d$ for all d.

This definition can be interpreted as follows. Let $\mathcal{A} = \sum_{r=1}^{R} v_r^1 \otimes \cdots \otimes v_r^D$ be any tensor. For each $d \in \{1, \ldots, D\}$, define $\mathbb{V}_d := \operatorname{span}\{v_1^d, \ldots, v_R^d\}$. By Definition 3.3, $\mathcal{A} \in \mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$, so that the multilinear rank of $\mathcal{A}$ is (componentwise) at most $(\dim \mathbb{V}_1, \dots, \dim \mathbb{V}_D)$. Conversely, if a given tensor $\mathcal{A}$ has multilinear rank $(m_1, \dots, m_D)$, then there exist linear spaces $\mathbb{V}_1, \dots, \mathbb{V}_D$ of dimension $m_1, \dots, m_D$, respectively, such that $\mathcal{A} \in \mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$. Thus, $\mathcal{A}$ can be expressed as $\mathcal{A} = \sum_{r=1}^{R} v_r^1 \otimes \cdots \otimes v_r^D$ with $\dim \operatorname{span}\{v_1^d, \dots, v_R^d\} \leqslant m_d$ for all d.

3.3.4 Block term decomposition

In some applications, one is interested in a family of decompositions that generalise the polyadic decomposition. These are defined as follows.

Definition 3.14 ([DL08])**.** Let $\mathcal{A} \in \mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$ be a tensor and choose positive integers R and l_d^r for all $d = 1, \dots, D$ and $r = 1, \dots, R$. A *block term decomposition (BTD) of ranks* $(l_d^r)_{d=r=1}^{D,R}$ is a set of tensors $\{\mathcal{A}_1, \dots, \mathcal{A}_R\}$ such that $\mathcal{A} = \sum_{r=1}^{R} \mathcal{A}_r$ and $\mathcal{A}_r$ has multilinear rank $(l_1^r, \dots, l_D^r)$ for each r.

Similarly to the polyadic decomposition, it may be useful to find a Tucker decomposition of the terms $\mathcal{A}_r$ that reveals their multilinear rank. This gives the following analogue of (3.4) for the BTD:

$$\mathcal{A} = \sum_{r=1}^{R} (U_1^r \otimes \cdots \otimes U_D^r) \mathcal{C}_r$$

where $U_d^r \in \mathbb{R}^{\dim \mathbb{V}_d \times l_d^r}$ and $\mathcal{C}_r \in \mathbb{K}^{l_1^r \times \cdots \times l_D^r}$. In fact, this is the original definition of the BTD. Note that factorising each term as a Tucker decomposition introduces redundancy, since the Tucker decomposition is never unique.

A common special case of the BTD of third-order tensors is the so-called $(l_r, l_r, 1)$-BTD, in which $l_1^r = l_2^r$ and $l_3^r = 1$ for all r.

3.4 Geometry of low-rank tensors

3.4.1 Basic tensor manifolds

The spaces of general, symmetric, and partially symmetric tensors of rank at most 1 are classical objects in algebraic geometry, known as the Segre, Veronese, and Segre–Veronese varieties, respectively [Har95, Chapter 2][Lan12, Chapter 4]. These are known to be smooth in the sense of projective algebraic geometry, but they are rarely studied from a differential geometric point of view. Recently, however, Lars Swijsen studied the Riemannian geometry of real rank-1 tensors in his doctoral dissertation [Swi22]. From here on, we will only work in real tensor spaces as well. A rudimentary result in said dissertation is the following.

Proposition 3.15. *The set*

$$\mathcal{S}(\mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D) := \{\mathcal{A} \in \mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D \mid \operatorname{rank} \mathcal{A} = 1\}$$

is an embedded submanifold of $\mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$ *of dimension* $1+\sum_{d=1}^{D} \dim(\mathbb{V}_d - 1)$*, called the* Segre manifold*. We will simply refer to it as* $\mathcal{S}$ *if the* $\mathbb{V}_1, \ldots, \mathbb{V}_D$ *are clear from the context.*

This was shown using the fact that

$$\begin{aligned} \phi\colon\ \mathbb{R}_0^+ \times \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_D-1} &\to \mathbb{R}^{n_1 \times \cdots \times n_D} \\ (\lambda, v_1, \ldots, v_D) &\mapsto \lambda v_1 \otimes \cdots \otimes v_D \end{aligned}$$

is a smooth immersion whose image is $\mathcal{S}$ (in coordinates) and that at any point in its domain, ϕ is locally a homeomorphism onto an open subset of $\mathcal{S}$. This argument generalises to (partially) symmetric tensors and gives the following result.

Proposition 3.16. *The partially symmetric tensors of rank 1, i.e.,*

$$\mathcal{SV}(\mathbb{V}_1^{\otimes d_1} \otimes \cdots \otimes \mathbb{V}_K^{d_K}) := \left\{ \lambda v_1^{\otimes d_1} \otimes \cdots \otimes v_k^{\otimes d_K} \;\middle|\; \lambda \in \mathbb{R}, \quad \text{and} \quad \forall k\colon v_k \in \mathbb{V}_k \right\},$$

is an embedded submanifold of $\mathbb{V}_1^{\otimes d_1} \otimes \cdots \otimes \mathbb{V}_K^{d_K}$*, called the* Segre–Veronese manifold*. For* $K = 1$*, the* Veronese manifold *is* $\mathcal{V}(\mathbb{V}^{\otimes D}) := \mathcal{SV}(\mathbb{V}^{\otimes D})$*.*

All three of these manifolds are standard objects in the algebraic geometry of tensor decompositions [Lan12]. A less classical tensor manifold is the set

$$\mathcal{T}_{r_1,\ldots,r_D}(\mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D) := \{\mathcal{A} \in \mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D \mid \operatorname{mlrank} \mathcal{A} = (r_1, \ldots, r_D)\},$$

which we call the *Tucker manifold.* Basic properties of this space, such as a parametrisation of the tangent space, were derived by Koch and Lubich [KL10]. Note that if we set $(r_1, \ldots, r_D) := (1, \ldots, 1)$, we obtain the Segre manifold. Note also that $\mathcal{T}_{n_1,\ldots,n_D}(\mathbb{R}^{n_1 \times \cdots \times n_D})$ is the set of tensors of full multilinear rank, also denoted as $\mathbb{R}_\star^{n_1 \times \cdots \times n_D}$, which is an open subset of $\mathbb{R}^{n_1 \times \cdots \times n_D}$. The set of multilinear rank *at most* $(r_1, \ldots, r_D)$ is called the *subspace variety* in algebraic geometry [Lan12, Definition 3.1.3.4]. In Chapter 4, we generalise Tucker manifolds to *structured Tucker manifolds.*

3.4.2 Join sets

The Segre, Veronese, Segre–Veronese, and Tucker manifolds are the basic objects from which additive tensor decompositions are derived. The domains of such decompositions are known as *join sets* [Lan12, Chapter 5], which are defined as follows.

Definition 3.17. Suppose that $\mathcal{M}_1, \ldots, \mathcal{M}_R$ are subsets of a linear space $\mathcal{E}$. The *addition map* is defined as

$$\Sigma\colon\ \mathcal{M}_1 \times \cdots \times \mathcal{M}_R \to \mathcal{E}$$

$$(p_1, \ldots, p_R) \mapsto p_1 + \cdots + p_R.$$

The image of Σ is the *join set* of $\mathcal{M}_1, \ldots, \mathcal{M}_R$. If $\mathcal{M}_1, \ldots, \mathcal{M}_R$ are clear from the context, we will write the join set as $\mathcal{J}$.

The join set is classically studied for *algebraic varieties* rather than manifolds. In this text, all algebraic varieties are *affine varieties*, i.e., they can be characterised as the zero set of a system of polynomial equations. Given a subset $\mathcal{M}$ of a linear space, the *Zariski closure* $\overline{\mathcal{M}}$ is the smallest algebraic variety containing $\mathcal{M}$. Every algebraic variety has a dense subset which is a smooth manifold [Sha13]. Conversely, the basic manifolds of tensors with a fixed (additive or multilinear) rank, i.e., the Segre–Veronese and Tucker manifolds, are dense in their Zariski closures. More precisely, their Zariski closures are the spaces of tensors of *bounded* rank [Lan12]. Thus, in our context, the difference between a variety and a manifold is a small set of (possibly nonsmooth) points.

If $\mathcal{M}_1 = \cdots = \mathcal{M}_R = \mathcal{M}$, the Zariski closure of the join set is called the *rth secant variety of $\mathcal{M}$*. Such varieties, especially the secants of the Segre and Veronese manifold, are the main objects of interest in the algebraic geometry of additive tensor decompositions. An important example is the join set of R copies of the Segre manifold, i.e.,

$$\Sigma(\mathcal{S}, \ldots, \mathcal{S}) = \{\mathscr{A}_1 + \cdots + \mathscr{A}_R \,|\, \operatorname{rank} \mathscr{A}_r = 1 \text{ for all } r = 1, \ldots, R\}.$$

The decomposition of a tensor into R rank-1 tensors is the problem of inverting Σ on this set. For an overview of the theory of secant varieties and join sets, see Landsberg's book [Lan12], from which we have borrowed some terminology.

One of the key relevant properties of the join set is its dimension. The standard tool for computing this dimension is the following lemma.

Lemma 3.18 (Terracini's lemma - Lemma 5.4.1.1 in [Lan12])**.** *Let $\mathcal{J}$ be the join set of algebraic varieties $\mathcal{M}_1, \ldots, \mathcal{M}_R$. For all $(p_1, \ldots, p_R)$ in an open and* *dense subset of $\mathcal{M}_1 \times \cdots \times \mathcal{M}_R$, we have*

$$\mathcal{T}_{p_1+\cdots+p_R}\overline{\mathcal{J}} = \operatorname{span} D\Sigma(p_1, \ldots, p_R) = \mathcal{T}_{p_1}\mathcal{M}_1 + \cdots + \mathcal{T}_{p_R}\mathcal{M}_R$$

where $\mathcal{T}_{p_r}\mathcal{M}_r$ is treated as a linear subspace of the ambient linear space $\mathcal{E}$, for each r.

It follows from Terracini's lemma that

$$\dim \mathcal{J} = \operatorname{rank} D\Sigma(p_1, \ldots, p_R) \leqslant \dim \mathcal{M}_1 + \cdots + \dim \mathcal{M}_R \tag{3.5}$$

where $p_1 \in \mathcal{M}_1, \ldots, p_R \in \mathcal{M}_R$ are general points. This upper bound is the dimension of the domain of Σ, and I call it the *desired dimension* of the join set. That is, $\mathcal{J}$ has the desired dimension if and only if $\ker D\Sigma(p_1, \ldots, p_R) = \{0\}$ for general $p_1, \ldots, p_R$, i.e., if the linear spaces $\mathcal{T}_{p_1}\mathcal{M}_1, \ldots, \mathcal{T}_{p_R}\mathcal{M}_R$ do not intersect.

The name *desired dimension* is inspired by its relevance for the condition number. We saw in Example 2.15 that the condition number of inverting Σ is defined precisely under the assumption that $D\Sigma$ is injective. More precisely, if $D\Sigma(p_1, \ldots, p_R)$ is injective at generic points $p_1, \ldots, p_R$ (i.e., $\mathcal{J}$ is not defective), there exists a unique smooth map

$$\Phi \colon \mathcal{J} \to \mathcal{M}_1 \times \cdots \times \mathcal{M}_R$$

$$p_1 + \cdots + p_R \mapsto (p_1, \ldots, p_R)$$

that inverts Σ on a neighbourhood of $p_1 + \cdots + p_R$. We call it the *decomposition map*. When we ask the central question in this thesis, i.e., *"How sensitively does the decomposition $(p_1, \ldots, p_R)$ depend on the point $p_1 + \cdots + p_R$?"*, we are referring to the sensitivity of Φ. Thus, for this question to make sense at all, we want to assume that $\mathcal{J}$ has the desired dimension.

If $\mathcal{J}$ does not have the desired dimension, I will call it *defective*. In algebraic geometry, this means something slightly different. The *expected dimension*, denoted as $\operatorname{expdim} \mathcal{J}$, is defined as the minimum of the desired dimension and the dimension of the ambient linear space $\mathcal{E}$. This is a sharper upper bound on $\dim \mathcal{J}$ than (3.5). Most literature on the algebraic geometry of secant varieties [Lan12] defines a defective join set as one that does not have the *expected* dimension, but we require the *desired* dimension instead.

It is important not to overstate the implications of nondefectivity: the local existence of the decomposition map Φ implies that $p_1 + \cdots + p_R$ has a *locally unique* decomposition $(p_1, \ldots, p_R)$. That is, any alternative decomposition must be sufficiently far away from $(p_1, \ldots, p_R)$. For example, the join set of 11 copies of $\mathcal{S}(\mathbb{R}^3 \otimes \mathbb{R}^3 \otimes \mathbb{R}^{11})$ is nondefective. Points in this space generically have 352 716 complex polyadic decompositions, all of which are isolated from each other [Hau+19].

Example 3.19 (secants of the two-factor Segre manifold)**.** Let $\mathcal{S}_{mn} := \mathcal{S}(\mathbb{R}^m \otimes \mathbb{R}^n)$ denote the set of rank-1 matrices. The join set of R copies of $\mathcal{S}_{mn}$ is $\mathbb{R}^{m\times n}_{\leqslant r}$, i.e., the set of matrices of rank at most r. This algebraic variety has dimension $r(m+n)-r^2$ [Lan12], whereas its expected dimension is $r(m+n-1)$. Thus, the secant varieties of $\mathcal{S}_{mn}$ are defective for all $r > 1$. See [Lan12, section 4.6.2] for a proof that the tangent spaces at any pair $p_1, p_2 \in \mathcal{S}_{mn}$ must intersect.

In more algebraic terms, rank-R matrices are defective because of symmetry. That is, for rank-1 matrices $a_1b_1^T, \ldots, a_Rb_R^T$, the addition map is given by

$$\Sigma(a_1b_1^T, \ldots, a_Rb_R^T) = \underbrace{[a_1a_2\cdots a_R]}_{A}\underbrace{[b_1b_2\cdots b_R]^T}_{B^T}.$$

For all $G \in \mathrm{GL}(R)$, we have

$$\begin{aligned}\Sigma(a_1b_1^T, \ldots, a_Rb_R^T) &= (AG)(G^{-1}B^T)\\ &= \Sigma((AGe_1)(e_1^TG^{-1}B^T), \ldots, (AGe_R)(e_R^TG^{-1}B^T)) \quad (3.6)\end{aligned}$$

where e_r is the rth canonical basis vector of $\mathbb{R}^R$. Thus, Σ is constant over orbits of $\mathrm{GL}(R)$, which means it is not invertible, not even locally. Hence, the condition number of inverting Σ in the sense of Chapter 2 is undefined. A generalised notion of condition that overcomes this barrier is introduced in Chapter 6.

Since the most familiar examples of secant varieties (i.e., low-rank matrices) are defective, the term *expected dimension* may seem like a misnomer. Readers coming from the world of matrices might even *expect* symmetries akin to those of (3.6). It turns out that such symmetries are rare for tensors of higher order, though, which leads to the adage *"tensors are normal and matrices are strange [because of unusual symmetries]"*. The following well-known conjecture captures this more precisely.

Conjecture 3.20 (Abo–Ottaviani–Peterson [AOP08])**.** *The join set of R copies of the Segre manifold $\mathcal{S} \subseteq \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_D}$ has the expected dimension if $(n_1, \ldots, n_D)$ is none of the following special cases:*

- $(n_1, \ldots, n_D) = (2, 2n+1, 2n+1)$ *for some* $n \in \mathbb{N}$,
- $(n_1, \ldots, n_D) = (3, 4, 4)$,
- $(n_1, \ldots, n_D) = (2, 2, n, n)$ *for some* $n \in \mathbb{N}$,
- $n_M - 2 \geqslant \prod_{d\neq M} n_d - \sum_{d\neq M}(n_d - 1)$ *where* $1 \leqslant M \leqslant D$ *and* $n_M = \max\{n_1, \ldots, n_D\}$.

Through an exhaustive search using computer algebra, it was found that the only combinations of $R, n_1, \ldots, n_D$ such that $\dim \mathcal{J} \neq \operatorname{expdim} \mathcal{J} < n_1 \cdots n_D \leqslant 15000$ are the exceptions listed above [COV14].

For sufficiently small values of R (i.e., all R such that $R \dim \mathcal{S} \leqslant n_1 \cdots n_D$), the desired dimension is the expected dimension. Thus, the conjecture implies that spaces of tensors of sufficiently low rank, usually called *subgeneric rank*, always have the desired dimension, with a few exceptions.

Even for manifolds $\mathcal{M}_1, \ldots, \mathcal{M}_R$ whose join set has the desired dimension, there may still be a subset of exceptional points $p_1 \in \mathcal{M}_1, \ldots, p_R \in \mathcal{M}_R$ where $D\Sigma$ fails to be injective. These points mark the *Terracini locus*, i.e.,

$$TL := \{p_1 + \cdots + p_R \mid p_r \in \mathcal{M}_r \text{ for all } r \text{ and } \ker D\Sigma(p_1, \ldots, p_R) \neq \{0\}\}\,.$$

These are the points where the join set “looks defective”. Furthermore, it is the locus of points where the condition number of decomposition problem diverges [BV18b]. For recent results on the Terracini locus of specific tensor spaces, see [BBS20; BC21].

3.5 Applications of tensor decompositions

Tensor decompositions are central to many data analysis models. For an overview, see the review articles [KB09; PFS16; Sid+17]. I give two examples below.

3.5.1 Mixture models

Symmetric tensor decompositions are sometimes used for the estimation of several classes of *mixture models* [Ana+14]. These models are probability distributions involving an unobserved discrete variable Z and an observed variable X. The mixture model states that any two samples X_1, X_2 of X are conditionally independent given Z, i.e.,

$$P(X_1 = x_1, X_2 = x_2 | Z = z) = P(X_1 = x_1 | Z = z) P(X_2 = x_2 | Z = z). \quad (3.7)$$

One kind of mixture model is the bag-of-words model for *topic modelling* [JM09, §20.2]. This is an unsupervised machine learning problem where one is given a corpus of documents, each about one topic, and the goal is to detect the most common topics discussed in the corpus, their relative frequencies, and the word frequencies given the topic. I will revisit the algorithm from [Ana+14], first mentioned here in Section 1.3, for the estimation of a bag-of-words model.

In the bag-of-words model, the latent variable Z is the topic, which is always one of R possible values. The observed variables are the words in a text. The vocabulary used by a text is a set of cardinality n. Given any topic z, the assumption (3.7) implies that the probability of any sequence of words $X_1, X_2, \ldots$ in a text about z is invariant under permutation of the sequence. The distribution can be characterised by specifying the probabilities $p_1, \ldots, p_R$ of the R topics and the conditional probabilities $\mu_{ij} := P(X = x_i | Z = j)$.

The algorithm of interest for the estimation of these probabilities from data is based on *one-hot encoding.* That is, the ith word in the vocabulary (say, in alphabetical order) is represented as the n-tuple $e_i \in \mathbb{R}^n$ consisting of zeros in all components, except for the ith component, which is 1.

For each $r = 1, \ldots, R$, the vector $\mu_r := [\mu_{ir}]_{i=1}^n$ encodes the conditional probabilities of all words in the vocabulary given that the document is on the rth topic. If X_1, X_2, X_3 are stochastic variables representing any three words in the same text about an unknown topic, then

$$\mathbb{E}\left[X_1 \otimes X_2 \otimes X_3\right] = \sum_{r=1}^{R} p_r \mu_r^{\otimes 3} \tag{3.8}$$

under the mixture model assumption [Ana+14].

The left-hand of side of (3.8) can be approximated by computing the empirical mean of $X_1 \otimes X_2 \otimes X_3$ over a large corpus of texts. By computing a Waring decomposition of this mean, one can obtain estimates of the parameters p_r and μ_r that define the distribution. The tensor in (3.8) usually lives in a space of very high dimension (i.e, n^3) and its rank is usually much less than n. Such tensors can be decomposed relatively efficiently with a compress-decompose-expand algorithm [BA98], which we will explore in detail in Chapters 4 and 5.

It is important to note, however, that the empirical mean will only be an approximation of the true mean, and therefore, the estimated parameters will be off as well. The accuracy of the decomposition can be estimated in terms of the condition number, which will be discussed in Chapter 5. If the decomposition is sensitive to small inaccuracies in the data, the recovered parameters are essentially uninterpretable.

3.5.2 Blind source separation

Signal separation, or blind source separation, is the process of recovering independently sourced signals (such as time series) of which only a combination was measured. More concretely, suppose that R sources emit signal vectors

$s_1, \ldots, s_R$ and that we have K observers that detect linear combinations of the signals:

$$y_k = \sum_{r=1}^{R} m_{kr} s_r + n_k, \quad \text{where} \quad k = 1, \ldots, K, \tag{3.9}$$

and n_k is a stochastic vector representing measurement noise. The matrix $[m_{kr}]_{k=r=1}^{K,R}$ is called the *mixing matrix* and its columns are called the *mixing vectors* m_r. While the task at hand is to recover the signals $s_1, \ldots, s_R$ as a function of $y_1, \ldots, y_K$, practical algorithms often estimate the mixing matrix first [Com94; DLDMV00b; CJ10]. That is, one estimates *how* the signals are mixed.

Different assumptions about the signals, noise have led to different reconstruction algorithms. The quintessential model is *independent component analysis*, which assumes that the signals are statistically independent stochastic vectors and that the noise is Gaussian and independent of the signals. In this case, the mixing vectors can often be computed through a Waring decomposition of the cumulant tensor of the mixed signals [Com94].

An algorithm based on a different assumption was introduced by De Lathauwer [DL11] and has been successful for the detection of epileptic seizure signals [Hun+14]. This model assumes that the signals $s_1, \ldots, s_R$ consist of samples of an *exponential polynomial* at consecutive integer values of the argument. Recall that an exponential polynomial in a variable t is an expression that is polynomial in t and all b^t where $b \in \mathbb{C}$.

Under this assumption and in the absence of noise, the unobserved variables $m_1, \ldots, m_R$ and $s_1, \ldots, s_R$ can be obtained by constructing an array $\mathcal{Y}$ that is defined componentwise by $\mathcal{Y}_{ijk} := (y_k)_{i+j-1}$. It can be shown that $\mathcal{Y}$ admits an $(l_r, l_r, 1)$-BTD of the form

$$\mathcal{Y} = \sum_{r=1}^{R} H_r \otimes m_r,$$

where m_r is as in (3.9) and the rank of H_r depends on the degree of the exponential polynomial sampled in the rth signal. Moreover, the signals s_r can be read from the first row and last column of H_r.

Chapter 4

Structured block-term decompositions and Tucker compression

This chapter consists of the journal article [DBV23a].

N. Dewaele, P. Breiding and N. Vannieuwenhoven. “The condition number of many tensor decompositions is invariant under Tucker compression”. In: *Numerical Algorithms* (June 2023).

The doctoral candidate derived the theoretical results and performed the experiments. The text was written in collaboration with the coauthors.

Abstract

We characterise the sensitivity of several additive tensor decompositions with respect to perturbations of the original tensor. These decompositions include canonical polyadic decompositions, block term decompositions, and sums of tree tensor networks. Our main result shows that the condition number of all these decompositions is invariant under Tucker compression. This result can dramatically speed up the computation of the condition number in practical applications. We give the example of a $265 \times 371 \times 7$ tensor of rank 3 from a food science application whose condition number was computed in 6.9 milliseconds by

exploiting our new theorem, representing a speedup of four orders of magnitude over the previous state of the art.

4.1 Introduction

In numerous applications, one seeks a decomposition that expresses a tensor $\mathcal{A}$, living in the tensor product of vector spaces $\mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$, as a sum of R elementary terms:

$$\mathcal{A} = \mathcal{A}_1 + \cdots + \mathcal{A}_R, \tag{4.1}$$

where $\mathcal{A}_r \in \mathcal{M}_r$ and $\mathcal{M}_r$ is a low-dimensional *manifold* in the space of tensors. Such a decomposition was called a *join decomposition* in [BV18b]. We consider the case where $\mathbb{V}_1, \ldots, \mathbb{V}_D$ are Euclidean spaces, so that their tensor product has a natural inner product [Gre78]. Tensors in this space are known as Cartesian tensors [Tem60], but they are simply referred to as "tensors" in this chapter. To simplify notation, we identify $\mathbb{V}_1 \otimes \cdots \otimes \mathbb{V}_D$ with D-arrays in $\mathbb{R}^{\dim \mathbb{V}_1 \times \cdots \times \dim \mathbb{V}_D}$ after choosing orthonormal bases of $\mathbb{V}_1, \ldots, \mathbb{V}_D$.

In this chapter, we study the sensitivity properties of a certain subclass of join decompositions related to tensors. We call them *structured block term decompositions* (SBTD). The formal definition of this class is given in Section 4.2 below. Informally, an SBTD involves manifolds $\mathcal{M}_r$ that are defined by imposing certain (manifold) structures on the core tensor of a Tucker decomposition with fixed multilinear rank $(l_1, \ldots, l_D)$. Many commonly used decompositions are SBTDs; for instance,

- sums of rank-1 tensors,[1] i.e., canonical polyadic decomposition (CPD) [Hit27],
- sums of Tucker decompositions, i.e., block term decomposition (BTD) [DL08],
- sums of tensor train decompositions [Ose11; ERL22], and
- sums of hierarchical Tucker decompositions [HK09; Gra10].

These decompositions are sometimes used for data compression (see [ERL22] for sums of tensor trains) or, in the case of the CPD and BTD, for the identification of certain model parameters [AB03; Har70; DLDMV00a]. Especially in the latter case, it is essential to measure the sensitivity of the decomposition relative

[1] Recall that a rank-1 tensor $a_1 \otimes \cdots \otimes a_D$ is naturally represented in coordinates by the d-array $[(a_1)_{i_1} \cdots (a_D)_{i_D}]_{i_1,\ldots,i_D=1}^{\dim \mathbb{V}_1,\ldots,\dim \mathbb{V}_D}$, where $(a_j)_i$ is the ith coordinate of a_j.

to perturbations of the tensor. One way of doing this involves the *condition number*, which was analysed for general join decompositions in [BV18b]. One main result we establish in this chapter is that the condition number of an SBTD is invariant under *Tucker compression.*

Recall that a *Tucker decomposition* [Tuc66] represents $\mathcal{A}$ in a tensor product subspace by expressing it as a multilinear product $\mathcal{A} = (Q_1, \ldots, Q_D) \cdot \mathcal{G}$ where the matrices $Q_d \in \mathbb{R}^{n_d \times m_d}$ with $n_d \geqslant m_d$ have linearly independent columns for each $d = 1, \ldots, D$. That is, if $\mathcal{G} = \sum_{r=1}^{R} g_1^r \otimes \cdots \otimes g_D^r$, for some vectors $\{g_d^r\}_{d=1,r=1}^{D,R}$ then $\mathcal{A} = \sum_{r=1}^{R} (Q_1 g_1^r) \otimes \cdots \otimes (Q_D g_D^r)$.

The *core tensor* $\mathcal{G}$ is often much smaller than $\mathcal{A}$, and it gives the coordinates of $\mathcal{A}$ with respect to the tensor product basis $Q_1 \otimes \cdots \otimes Q_D$. Note that we will switch freely between two equivalent notations for Tucker decomposition: the first, $(Q_1, \ldots, Q_D) \cdot \mathcal{G}$, is a common notation [SL08] for *multilinear multiplication*, while the second, $(Q_1 \otimes \cdots \otimes Q_D)\mathcal{G}$ emphasises that a Tucker decomposition consists of taking a particular linear combination of the tensors in a tensor product basis $Q_1 \otimes \cdots \otimes Q_D$. Herein, $Q_1 \otimes \cdots \otimes Q_D$ denotes the tensor product of matrices, which acts linearly on rank-1 tensors by $(Q_1, \ldots, Q_D) \cdot (v_1 \otimes \cdots \otimes v_D) := (Q_1 v_1 \otimes \cdots \otimes Q_D v_D)$. In coordinates, this matrix is given by the Kronecker product of $Q_1, \ldots, Q_D$; see [Gre78].

Originally proposed for CPD, Tucker compression [BA98] consists of expressing a tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times \cdots \times n_D}$ in coordinates in the smallest tensor product subspace in which it lives, in order to speed up the computation of decompositions of the form (4.1). That is, before computing the CPD, one first computes a Tucker decomposition, expressing $\mathcal{A} = (Q_1, \ldots, Q_D) \cdot \mathcal{G}$. Then, one computes the CPD of the core tensor $\mathcal{G}$. The obtained decomposition can then be extended to a decomposition of the original tensor $\mathcal{A}$ by multilinear multiplication with the basis $(Q_1, \ldots, Q_D)$. Since there are efficient algorithms [DLDMV00a; VVM12] for computing an approximate Tucker decomposition of $\mathcal{A}$, contrary to the mostly optimization-based algorithms for computing CPDs, this compress–decompose–expand approach can often reduce the overall computation time [BA98]. Another main contribution of this chapter is that the SBTD provides a general framework for smoothly varying decompositions of the form (4.1) so that the SBTD of a tensor $\mathcal{G}$ corresponds to an SBTD of $(Q_1, \ldots, Q_D) \cdot \mathcal{G}$.

The topic of this chapter is to characterise how a decomposition of the form (4.1) changes if $\mathcal{A}$ is corrupted by noise. In order to obtain a robust interpretation of the elementary terms, it is essential to quantify how sensitive they are to the perturbations. As explained in [BV18b], under certain mild conditions, $\mathcal{A}$ has an isolated decomposition $a = (\mathcal{A}_1, \ldots, \mathcal{A}_R)$ and we can find a local inverse function Σ_a^{-1} of the *addition map* $\Sigma : \mathcal{M}_1 \times \cdots \times \mathcal{M}_R \to \mathbb{R}^{n_1 \times \cdots \times n_D}$, $(\mathcal{A}_1, \ldots, \mathcal{A}_R) \mapsto \mathcal{A}_1 + \cdots + \mathcal{A}_R$. The local sensitivity of the elementary terms $\mathcal{A}_r$ can be measured by the condition number [Ric66]

$$\kappa^{\mathrm{SBTD}}(\mathcal{A}_1,\ldots,\mathcal{A}_R) := \lim_{\delta\to 0} \sup_{\widetilde{\mathcal{A}}\in\mathcal{I}:\left\|\mathcal{A}-\widetilde{\mathcal{A}}\right\|\leqslant\delta} \frac{\left\|\Sigma_a^{-1}(\mathcal{A})-\Sigma_a^{-1}(\widetilde{\mathcal{A}})\right\|}{\left\|\mathcal{A}-\widetilde{\mathcal{A}}\right\|}, \tag{4.2}$$

where $\mathcal{I}$ is the set of valid perturbations (more on this below), and $\|\cdot\|$ denotes both the Euclidean norm on the ambient space $\mathbb{R}^{n_1\times\cdots\times n_D}$ and the product Euclidean norm on $\mathbb{R}^{n_1\times\cdots\times n_D}\times\cdots\times\mathbb{R}^{n_1\times\cdots\times n_D}$. The condition number measures perturbations to the summands $\mathcal{A}_r\in\mathcal{M}_r, r=1,\ldots,R$. It does not measure how these summands are *parametrised*, which would introduce a number of complications.[2] Furthermore, a priori, the condition number depends on both input $\mathcal{A}$ and output $(\mathcal{A}_1,\ldots,\mathcal{A}_R)$ because it is defined in terms of a *local* inverse [BC13]. However, since $\mathcal{A}$ depends uniquely on $(\mathcal{A}_1,\ldots,\mathcal{A}_R)$ we can write the condition number as a function of the output only. We have

$$\left\|\Sigma_a^{-1}(\mathcal{A})-\Sigma_a^{-1}(\widetilde{\mathcal{A}})\right\| \leqslant \kappa^{\mathrm{SBTD}}(\mathcal{A}_1,\ldots,\mathcal{A}_R)\left\|\mathcal{A}-\widetilde{\mathcal{A}}\right\| + o\left(\left\|\mathcal{A}-\widetilde{\mathcal{A}}\right\|\right) \tag{4.3}$$

as an asymptotically sharp first-order error bound. Eq. (4.2) requires specifying the domain $\mathcal{I}$, which means fixing the space in which the perturbations $\widetilde{\mathcal{A}}$ are allowed to live. There are four increasingly restrictive ways of looking at the problem:

1. $\widetilde{\mathcal{A}}\in\mathbb{R}^{n_1\times\cdots\times n_D}$ is arbitrary and the SBTD of $\widetilde{\mathcal{A}}$ is interpreted as the least-square minimiser $\operatorname{argmin}_{(\mathcal{A}_1,\ldots,\mathcal{A}_R)\in\mathcal{M}_1\times\cdots\times\mathcal{M}_R}\frac{1}{2}\left\|\widetilde{\mathcal{A}}-(\mathcal{A}_1+\cdots+\mathcal{A}_R)\right\|^2$.

2. $\widetilde{\mathcal{A}}$ has an SBTD.

3. $\widetilde{\mathcal{A}}$ can be Tucker compressed to a core $\widetilde{\mathcal{G}}\in\mathbb{R}^{m_1\times\cdots\times m_D}$ and $\widetilde{\mathcal{G}}$ has an SBTD.

4. $\widetilde{\mathcal{A}}$ lives in the same tensor subspace as $\mathcal{A}$, i.e., we have $\mathcal{A}=(Q_1,\ldots,Q_D)\cdot\mathcal{G}$ and $\widetilde{\mathcal{A}}=(Q_1,\ldots,Q_D)\cdot\widetilde{\mathcal{G}}$, and the cores $\mathcal{G}$ and $\widetilde{\mathcal{G}}$ both have an SBTD.

A priori, one should expect the problem to become easier in the more restrictive cases in the sense that the condition number decreases. Indeed, the set of allowed perturbations gets strictly smaller. However, we prove the following surprising result.

Theorem 4.1. *Let $\mathcal{A}=\mathcal{A}_1+\cdots+\mathcal{A}_R$ be an SBTD. The condition number $\kappa^{\mathrm{SBTD}}(\mathcal{A}_1,\ldots,\mathcal{A}_R)$ is the same for all four domains outlined above.*

[2] See [Van17] for how to deal with such complications in the context of the CPD.

This theorem is implied by Theorem 4.14 and Corollary 4.15 below.

Theorem 4.1 is in stark contrast to some other problems in which the condition number depends on the domain. For instance, the condition number of the matrix logarithm for perturbations constrained in the symplectic group was studied in [ANT19]. It was shown that the ratio between the unconstrained and constrained condition number can become arbitrarily large.

Our result indicates that computing the SBTDs of $\mathcal{A}$ and $\mathcal{G}$ are equally difficult from a numerical point of view. Indeed, condition numbers are connected to convergence rates of iterative methods [NW06; AMS08]. For instance, the local convergence rate of the Riemannian Gauss-Newton method applied the minimisation of $\|\Sigma(\mathcal{A}_1, \ldots \mathcal{A}_R) - \mathcal{A}\|^2$ is bounded in terms of the condition number (4.2) [BV18a]. This suggests that compression, surprisingly, will not improve the local rate of convergence, even though the search space can be much smaller. Compression can, nevertheless, reduce the overall computation time when $\mathcal{A}$ is highly compressible [BA98].

A major practical advantage of Theorem 4.1 is that the condition number can be computed more efficiently by considering $\mathcal{A}$ as a point in a tensor product subspace: It suffices to compute the condition number of the core $\mathcal{G}$. An example illustrates the above significant computational advantage. Consider a rank-3 tensor of dimensions $265 \times 371 \times 7$, as in the sugar data set of [BA98]. Its CPD can be compressed to that of a $3 \times 3 \times 3$-tensor. We implemented two algorithms to compute the condition number in Julia v1.6 [Bez+17]; the one from [BV18b] and one based on Theorem 4.1. On a system with an Intel Xeon CPU E5-2697 v3 running on 8 cores and 126GB memory, this took 110 seconds and 6.9 milliseconds, respectively, which gives a speedup ratio of over 15 000. If the CPD is already in compressed form, the time can be reduced further to only 0.089 milliseconds, representing a speedup of more than a million over the state of the art.

4.1.1 Outline

We introduce the SBTD in Section 4.2. In Section 4.3, we derive the geometric foundations of structured Tucker decompositions. Section 4.4 gives an algorithm to compute the condition number and presents special cases and estimates. Section 4.5 introduces subspace-constrained SBTDs and proves the main result, Theorem 4.1, which states that the condition number of SBTDs is invariant under Tucker compression. Section 4.6 contains some numerical experiments confirming the theoretical results.

4.1.2 Notation

The only norms used in this chapter are the Euclidean (or Frobenius) norms for tensors and vectors, all consistently denoted by $\|\cdot\|$. The manifold of real $n \times m$ matrices of rank m is denoted as $\mathbb{R}_\star^{n\times m}$, where $n \geqslant m$. The n-dimensional sphere is $\mathbb{S}^n$. The jth column of the identity $\mathbb{1}_n$ is $e_j^{(n)}$. If the ambient dimension is clear from the context, we also abbreviate $e_j := e_j^{(n)}$. The dth unfolding of a tensor $\mathcal{A}$ is $\mathcal{A}_{(d)}$. For any matrix X and any set of matrices A_n, $n = 1, \ldots, N$, and any $k = 1, \ldots, N+1$, we denote $X \otimes_k (A_1 \otimes \cdots \otimes A_N) := A_1 \otimes \cdots \otimes A_{k-1} \otimes X \otimes A_k \otimes \cdots \otimes A_N$. For a group G acting on a set $\mathcal{M}$, the G-orbit of $x \in \mathcal{M}$ is $[x]_G$.

4.2 The structured block term decomposition

In this section, we give a formal definition of the SBTD, the main tensor decomposition that we study in this chapter. Just as a BTD is a linear combination of Tucker decompositions, an SBTD is a linear combination of *structured Tucker decompositions.* The structure we consider is imposed on the core tensor of the Tucker decomposition.

Definition 4.2 (Tucker core structure)**.** A smooth submanifold $\mathcal{M} \subseteq \mathbb{R}^{l_1\times\cdots\times l_D}$ is a *Tucker core structure* if every $\mathcal{C} \in \mathcal{M}$ satisfies the following:

1. $\mathcal{C}$ has multilinear rank equal to $(l_1, \ldots, l_D)$, and
2. $(A_1, \ldots, A_D) \cdot \mathcal{C} \in \mathcal{M}$ for all $A_d \in \mathrm{GL}(l_d)$ with $d = 1, \ldots, D$.

Next, we can define the $\mathcal{M}$-structured Tucker decomposition.

Definition 4.3 (Structured Tucker decomposition)**.** Let $\mathcal{M} \subseteq \mathbb{R}^{l_1\times\cdots\times l_D}$ be a Tucker core structure. An *$\mathcal{M}$-structured Tucker decomposition* of $\mathcal{A} \in \mathbb{R}^{n_1\times\cdots\times n_D}$ is an expression of the form

$$\mathcal{A} = (U_1, \ldots, U_D) \cdot \mathcal{C} = (U_1 \otimes \cdots \otimes U_D)\mathcal{C}$$

with $\mathcal{C} \in \mathcal{M}$ and all $U_d \in \mathbb{R}_\star^{n_d\times l_d}$ for $d = 1, \ldots, D$.

The first basic result we establish in Section 4.3 below ensures that the results from [BV18b] can be applied to study the condition number.

Proposition 4.4. *The set of all tensors $\mathcal{A}$ admitting an $\mathcal{M}$-structured Tucker decomposition forms a smooth embedded submanifold $\mathcal{M}^{n_1,\ldots,n_D} \subseteq \mathbb{R}^{n_1\times\cdots\times n_D}$, called the $\mathcal{M}$-structured Tucker manifold.*

An important subclass of structured Tucker manifolds in applications are defined by *tensor networks* in which the graph is a tree [Orús14]. This includes tensors with a fixed rank Tucker decomposition [Tuc66], fixed-rank tensor train decomposition [Ose11], and fixed rank hierarchical Tucker decomposition [HK09; Gra10].

The set of tree tensor networks (i.e., hierarchical Tucker formats) with fixed ranks is known to form a manifold [UV13]. This manifold is invariant under the natural action of $\mathrm{GL}(l_1) \times \cdots \times \mathrm{GL}(l_D)$. Since multilinear rank is also invariant under this action [Lan12], all concise (i.e., multilinear rank equals the dimension of the ambient space) tree-based tensor networks are valid Tucker core structures. This includes all aforementioned Tucker, tensor trains, and hierarchical Tucker decompositions in $\mathbb{R}^{l_1 \times \cdots \times l_D}$ of multilinear rank $(l_1, \ldots, l_D)$. In particular, the case where $(l_1, \ldots, l_D) = (1, \ldots, 1)$ is the set of rank-1 tensors.

We will be interested in additive decompositions whose elementary terms lie in structured Tucker manifolds, called *structured block term decompositions (SBTDs).*

Definition 4.5 (Structured block term decomposition)**.** An SBTD of the tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times \cdots \times n_D}$ associated with the $\mathcal{M}_r$-structured Tucker manifolds $\mathcal{M}_r^{n_1,\ldots,n_D}$ is a decomposition of the form $\mathcal{A} = \mathcal{A}_1 + \cdots + \mathcal{A}_R$ with $\mathcal{A}_r \in \mathcal{M}_r^{n_1,\ldots,n_D}$ for $r = 1, \ldots, R$.

Any sum mixing rank-1 tensors, Tucker decompositions, tensor trains decompositions, and hierarchical Tucker decompositions is thus an SBTD.

4.3 Geometry of the structured Tucker manifold

The condition number of join decompositions from [BV18b] requires that the summands in (4.1) live on manifolds. Therefore, we first derive the geometric properties of the manifolds involved in the decomposition. We prove Proposition 4.4, which shows that $\mathcal{M}^{n_1,\ldots,n_D}$ in Definition 4.3 is indeed a manifold. We also derive an expression for its tangent space. The proofs of these statements are standard computations in differential geometry, similar to those of [UV13].

The following result establishes the differential structure of our manifolds.

Proposition 4.6. *Let $\mathcal{M}$ be a Tucker core structure as in Definition 4.2. Define the manifolds*

$$\widetilde{\mathcal{M}} := \mathcal{M} \times \mathbb{R}_\star^{n_1 \times l_1} \times \cdots \times \mathbb{R}_\star^{n_D \times l_D} \quad \textit{and} \quad \mathcal{G} := \mathrm{GL}(l_1) \times \cdots \times \mathrm{GL}(l_D)$$

and the group action

$$\theta : \mathcal{G} \times \widetilde{\mathcal{M}} \to \widetilde{\mathcal{M}}$$

$$((A_1, \dots, A_D), (\mathcal{C}, U_1, \dots, U_D)) \mapsto \left((A_1^{-1}, \dots, A_D^{-1}) \cdot \mathcal{C},\, U_1 A_1, \dots, U_D A_D\right).$$

Then $\widetilde{\mathcal{M}}/\mathcal{G}$ is a quotient manifold with a unique smooth structure so that the quotient map $[\cdot]_{\mathcal{G}} : \widetilde{\mathcal{M}} \to \widetilde{\mathcal{M}}/\mathcal{G}$ is a smooth submersion.

Proof. By [Lee13, Theorem 21.10], we only need to verify that the action is smooth, free (i.e., it fixes the identity), and proper. The first two properties are straightforward to check. To show that it is proper, consider the sequences $\{x_n\}_{n\in\mathbb{N}} \to x$ in $\widetilde{\mathcal{M}}$ and $\{\mathfrak{a}_n\}_{n\in\mathbb{N}}$ in $\mathcal{G}$ where $\{\theta(\mathfrak{a}_n, x_n)\}_{n\in\mathbb{N}}$ converges in $\widetilde{\mathcal{M}}$. By [Lee13, Proposition 21.5], θ is proper if $\{\mathfrak{a}_n\}_{n\in\mathbb{N}}$ converges in $\mathcal{G}$. Consider the map $f : \widetilde{\mathcal{M}} \times \widetilde{\mathcal{M}} \to \mathcal{G}$ taking

$$(\mathcal{C}, U_1, \dots, U_D), (\widehat{\mathcal{C}}, \hat{U}_1, \dots, \hat{U}_D) \mapsto (U_1^\dagger \hat{U}_1, \dots, U_D^\dagger \hat{U}_D),$$

where $U_d^\dagger = (U_d^T U_d)^{-1} U_d^T$ is the Moore-Penrose inverse. Note $f(x_n, \theta(\mathfrak{a}_n, x_n)) = \mathfrak{a}_n$. Furthermore, f is continuous by the continuity of the Moore-Penrose inverse. Since $\{(x_n, \theta(\mathfrak{a}_n, x_n))\}_{n\in\mathbb{N}}$ converges, so does $\{f(x_n, \theta(\mathfrak{a}_n, x_n))\}_{n\in\mathbb{N}} = \{\mathfrak{a}_n\}_{n\in\mathbb{N}}$. □

The tangent space to this quotient manifold is derived next.

Proposition 4.7. *Take the manifold $\widetilde{\mathcal{M}}/\mathcal{G}$ as in Proposition 4.6 and consider a point $x = (\mathcal{C}, U_1, \dots, U_D) \in \widetilde{\mathcal{M}}$. Complete each U_d to a basis $[U_d \quad U_d^\perp]$ of $\mathbb{R}^{n_d}$. Then*

$$\mathcal{T}_{[x]_{\mathcal{G}}}(\widetilde{\mathcal{M}}/\mathcal{G}) \cong \left\{ \left(\dot{\mathcal{C}}, U_1^\perp B_1, \dots, U_D^\perp B_D\right) \mid \dot{\mathcal{C}} \in \mathcal{T}_{\mathcal{C}}\mathcal{M},\ B_d \in \mathbb{R}^{(n_d - l_d)\times l_d} \right\}.$$

Proof. Define the fibre $\mathcal{F}$ of all x' equivalent to x:

$$\mathcal{F}_x = \left\{ \left((A_1, \dots, A_D) \cdot \mathcal{C}, U_1 A_1^{-1}, \dots, U_D A_D^{-1}\right) \mid A_d \in \mathrm{GL}(l_d),\ d = 1, \dots, D \right\}.$$

This allows us to define the vertical space as the tangent space to $\mathcal{F}$ at x:

$$\mathbb{V}_x = \left\{ \left(\sum_{d=1}^{D} \left(\dot{A}_d \otimes_d \bigotimes_{d' \neq d} \mathbb{1}_{l_{d'}} \right) \mathcal{C},\ -U_1 \dot{A}_1, \dots, -U_D \dot{A}_D \right) \middle| \dot{A}_d \in \mathbb{R}^{l_d \times l_d} \right\}. \tag{4.4}$$

In this expression, we used $\mathcal{T}_{A_d}\mathrm{GL}(l_d) \cong \mathbb{R}^{l_d \times l_d}$[Lee13] for each $d = 1, \dots, D$. Now define the horizontal space as

$$\mathbb{H}_x := \left\{ \left(\dot{\mathcal{C}}, U_1^\perp B_1, \dots, U_d^\perp B_d\right) \middle| \dot{\mathcal{C}} \in \mathcal{T}_{\mathcal{C}}\mathcal{M},\ B_d \in \mathbb{R}^{(n_d - l_d)\times l_d} \right\}.$$

We will show that $\mathbb{V}_x \oplus \mathbb{H}_x = \mathcal{T}_x\widetilde{\mathcal{M}}$. First, we verify that the intersection is trivial. Take $\xi \in \mathbb{V}_x$, parametrised as in (4.4). If also $\xi \in \mathbb{H}_x$, by construction of $U_d^\perp$, it must hold that all $\dot{A}_d$ in the parametrisation of ξ are zero and hence $\xi = 0$.

Next, we show that the sum is $\mathcal{T}_x\widetilde{\mathcal{M}}$. We know that $\mathcal{M}$ is invariant under the action of $\mathrm{GL}(l_1) \times \cdots \times \mathrm{GL}(l_D)$. Therefore, for any $\dot{A}_d \in \mathbb{R}^{l_d \times l_d}$ for $d = 1, \ldots, D$, there exist curves over $\mathcal{M}$ of the form $\gamma(t) = (A_1(t), \ldots, A_D(t)) \cdot \mathcal{C}$ with $A_d(0) = \mathbb{1}_{l_d}$ and $\frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}A_d(t) = \dot{A}_d$. Hence, all tensors of the form

$$\gamma'(0) = \sum_{d=1}^{D} \left(\dot{A}_d \otimes_d \bigotimes_{d' \neq d} \mathbb{1}_{l_{d'}} \right) \mathcal{C} \quad \text{with} \quad \dot{A}_d \in \mathbb{R}^{l_d \times l_d}, \quad d = 1, \ldots, D$$

are tangent to $\mathcal{M}$ at $\mathcal{C}$. Because of this, it is easy to check that

$$\mathbb{V}_x \oplus \mathbb{H}_x = \mathcal{T}_\mathcal{C}\mathcal{M} \times \mathbb{R}^{n_1 \times l_1} \times \cdots \times \mathbb{R}^{n_D \times l_D} = \mathcal{T}_x\widetilde{\mathcal{M}}.$$

By the general theory of quotient manifolds, this establishes $\mathbb{H}_x \cong \mathcal{T}_{[x]_\mathcal{G}}\left(\widetilde{\mathcal{M}}/\mathcal{G}\right)$, where the isomorphism is the unique horizontal lift [AMS08, Section 3.5.8]. □

We have established that Tucker decompositions with a structured core form a smooth manifold. By definition, a point on an $\mathcal{M}$-structured Tucker manifold corresponds to a Tucker decomposition that is unique up to basis transform. We now have all the tools we need to show that $\mathcal{M}^{n_1,\ldots,n_D}$ is a manifold. We do this next.

Proposition 4.8. *Let $\widetilde{\mathcal{M}}/\mathcal{G}$ be as in Proposition 4.6 and let $\mathcal{M}^{n_1,\ldots,n_D}$ be the $\mathcal{M}$-structured Tucker manifold. Then $\mathcal{M}^{n_1,\ldots,n_D}$ is a smooth embedded submanifold of $\mathbb{R}^{n_1 \times \cdots \times n_D}$ and the following is a diffeomorphism:*

$$\Phi : \widetilde{\mathcal{M}}/\mathcal{G} \to \mathcal{M}^{n_1,\ldots,n_D}$$

$$[(\mathcal{C}, U_1, \ldots, U_D)]_\mathcal{G} \mapsto (U_1, \ldots, U_D) \cdot \mathcal{C}.$$

Moreover, the tangent space to $\mathcal{M}^{n_1,\ldots,n_D}$ at $(U_1, \ldots, U_D) \cdot \mathcal{C}$ is generated by all tensors

$$(U_1, \ldots, U_D) \cdot \dot{\mathcal{C}} + \sum_{d=1}^{D} \left(\dot{U}_d \otimes_d \bigotimes_{d' \neq d} U_{d'} \right) \mathcal{C} \tag{4.5}$$

with $\dot{\mathcal{C}} \in \mathcal{T}_\mathcal{C}\mathcal{M}$ and $U_d^T \dot{U}_d = 0_{l_d \times l_d}$ for all $d = 1, \ldots, D$.

Proof. By [Lee13, Proposition 5.2], the first claim holds if Φ is both a homeomorphism and a smooth immersion. First, we show that it is a homeomorphism. Note that Φ is a bijection because $\mathcal{M}^{n_1,\ldots,n_D}$ is precisely the set of all tensors with a Tucker decomposition where the core is in $\mathcal{M}$. Since Φ is induced by a polynomial map, it is also continuous. To show that Φ^{-1} is continuous, consider the maps

$$V_d : \mathcal{M}^{n_1,\ldots,n_D} \to \operatorname{Gr}(n_d, l_d)$$
$$(U_1, \ldots, U_D) \cdot \mathcal{C} \mapsto [U_d]_{\operatorname{GL}(l_d)}$$

where $\operatorname{Gr}(n_d, l_d) \cong \mathbb{R}_\star^{n_d \times l_d}/\operatorname{GL}(l_d)$ is the Grassmannian of l_d-dimensional linear spaces in $\mathbb{R}^{n_d}$ [AMS08]. That is, $V_d(\mathcal{X})$ is the column span of its dth flattening $\mathcal{X}_{(d)}$.

We will demonstrate continuity of V_d at any $\mathcal{X} \in \mathcal{M}^{n_1,\ldots,n_D}$ by showing that every open neighbourhood $\mathcal{V}$ of $V_d(\mathcal{X})$ contains the image of a neighbourhood of $\mathcal{X}$ [Mun14, Theorem 18.1]. By the definition of the quotient topology, we can write $\mathcal{V} = [\mathcal{U}]_{\operatorname{GL}(l_d)}$ for some open neighbourhood $\mathcal{U} \subseteq \mathbb{R}_\star^{n_d \times l_d}$ of U_d, where U_d is any representative of $V_d(\mathcal{X})$. Furthermore, for some ball $B_\varepsilon(U_d)$ of radius ε centred at U_d, we have $[B_\varepsilon(U_d)]_{\operatorname{GL}(l_d)} \subseteq [\mathcal{U}]_{\operatorname{GL}(l_d)} = \mathcal{V}$.

Now we exploit the liberty of choosing the representative U_d. Observe that $V_d(\mathcal{X})$ is the span of l_d columns of $\mathcal{X}_{(d)}$. In other words, there exists a column selection operator $P_d \in \mathbb{R}^{(\prod_{d' \neq d} n_{d'}) \times l_d}$, so that $V_d(\mathcal{X}) = [\mathcal{X}_{(d)} P_d]_{\operatorname{GL}(l_d)}$. By the semicontinuity of matrix rank, there exists $0 < \delta < \varepsilon$ so that any perturbation to $\mathcal{X}$ of norm less than δ does not change the rank of $\mathcal{X}_{(d)}$ or $\mathcal{X}_{(d)} P_d$. Hence, $V_d(\tilde{\mathcal{X}}) = [\tilde{\mathcal{X}}_{(d)} P_d]_{\operatorname{GL}(l_d)}$ for any $\tilde{\mathcal{X}}$ in a ball $B_\delta(\mathcal{X})$ of radius δ. Because $\left\|\tilde{\mathcal{X}}_{(d)} P_d - \mathcal{X}_{(d)} P_d\right\| < \delta < \varepsilon$, we have $V_d(B_\delta(\mathcal{X})) \subseteq [B_\varepsilon(\mathcal{X}_{(d)} P_d)]_{\operatorname{GL}(l_d)}$, which proves continuity of V_d.

For any $\mathcal{X} \in \mathcal{M}^{n_1,\ldots,n_D}$, let $V_d(\mathcal{X}) = [U_d]_{\operatorname{GL}(l_d)}$ for each d. It can be verified that the following is independent of the representatives U_d:

$$\Psi(\mathcal{X}) := [((U_1^\dagger, \ldots, U_D^\dagger) \cdot \mathcal{X}, U_1, \ldots, U_D)]_{\mathcal{G}}.$$

The right-hand side is the Tucker decomposition of $\mathcal{X}$, which is unique up to the action of $\mathcal{G}$. Hence, $\Psi = \Phi^{-1}$. This shows that Φ^{-1} is the composition of continuous maps: V_d for each d, the Moore-Penrose inverse, multilinear multiplication, and the canonical projection map. Hence, Φ^{-1} is continuous.

Next, we show that Φ is an immersion, i.e., that its derivative maps a basis to a basis, in which case $\mathcal{T}_{\mathcal{X}} \mathcal{M}^{n_1,\ldots,n_D}$ is the image of $\mathrm{d}\Phi$. Fix a basis $\mathscr{B}_0$ of $\mathcal{T}_{\mathcal{C}} \mathcal{M}$ and, for each $d = 1, \ldots, D$, a basis $\mathscr{B}_d$ of all $\dot{U}_d \in \mathbb{R}^{n_d \times l_d}$ so that $U_d^T \dot{U}_d = 0$. By Proposition 4.7, the tangent space of $\widetilde{\mathcal{M}/\mathcal{G}}$ can be considered as a product space generated by the canonical product basis derived from $\mathscr{B}_0, \ldots, \mathscr{B}_D$.

Applying $\mathrm{d}\Phi$ to this basis gives $\mathrm{d}\Phi(\mathcal{T}_{[(U_1,\ldots,U_D)\cdot\mathcal{C}]_\mathcal{G}}(\widetilde{\mathcal{M}}/\mathcal{G})) = T_0 \cup \cdots \cup T_D$ where

$$T_0 := \left\{(U_1,\ldots,U_D)\cdot\dot{\mathcal{C}} \mid \dot{\mathcal{C}} \in \mathscr{B}_0\right\} \quad \text{and}$$

$$T_d := \left\{\big(U_1,\ldots,U_{d-1},\dot{U}_d,U_{d+1},\ldots,U_D\big)\cdot\mathcal{C} \mid \dot{U}_d \in \mathscr{B}_d\right\} \quad \text{for } d = 1,\ldots,D.$$

Note that the sets T_i and T_j with $i \neq j$ are pairwise orthogonal due to the constraint on $\dot{U}_d$. Since $(U_1 \otimes \cdots \otimes U_D)$ has full rank and $\mathscr{B}_0$ is a basis, T_0 is linearly independent. The tangents in the set T_d with $d \geqslant 1$ are tensors whose dth unfolding is

$$\dot{U}_d\mathcal{C}_{(d)}(U_1 \otimes \cdots \otimes U_{d-1} \otimes U_{d+1} \otimes \cdots \otimes U_D)^T. \tag{4.6}$$

Recall from Definition 4.3 that $\mathcal{C}_{(d)}$ and all U_i^T have full row rank. For a set of linearly independent matrices $\dot{U}_d$, all matrices (4.6) are linearly independent. This shows that Φ is an immersion. By [Lee13, Proposition 5.2], $\mathcal{M}^{n_1,\ldots,n_D}$ is a manifold and Φ is a diffeomorphism. By [Lee13, Theorem 4.14] $\mathcal{T}_\mathcal{X}\mathcal{M}^{n_1,\ldots,n_D}$ is the image of $\mathrm{d}\Phi$. □

Note that Proposition 4.4 is a corollary of the previous statement.

4.4 Computing the condition number

Having shown that the structured Tucker decompositions form a manifold, we can investigate the condition number of the associated SBTD using the tools from [BV18b]. For this, we first derive an orthonormal basis of the structured Tucker manifold, so that the condition number can be computed with efficient algorithms from linear algebra using (4.8) below. We present some examples, as well as useful estimates of the condition number of SBTDs.

4.4.1 A direct algorithm

Let $\mathcal{M}_1^{n_1,\ldots,n_D},\ldots,\mathcal{M}_R^{n_1,\ldots,n_D}$ be structured Tucker manifolds, and recall the addition map

$$\Sigma : \mathcal{M}_1^{n_1,\ldots,n_D} \times \cdots \times \mathcal{M}_R^{n_1,\ldots,n_D} \to \mathbb{R}^{n_1\times\cdots\times n_D},\ (\mathcal{A}_1,\ldots,\mathcal{A}_R) \mapsto \mathcal{A}_1 + \cdots + \mathcal{A}_R$$

from the introduction. Computing an SBTD translates to finding a decomposition $(\mathcal{A}_1,\ldots,\mathcal{A}_R)$ so that $\Sigma(\mathcal{A}_1,\ldots,\mathcal{A}_R) = \mathcal{A}$. The condition number $\kappa^{\mathrm{SBTD}}(\mathcal{A}_1,\ldots,\mathcal{A}_R)$ from (4.2) is computed as follows [BV18b]. For $r = 1,\ldots,R$, compute orthonormal bases of $\mathcal{T}_{\mathcal{A}_r}\mathcal{M}_r^{n_1,\ldots,n_D}$, the tangent space to $\mathcal{M}_r^{n_1,\ldots,n_D}$

at $\mathcal{A}_r$. The basis vectors are the columns of matrices T_r. Then, the so-called *Terracini matrix* is constructed as

$$T_{\mathcal{A}_1,\ldots,\mathcal{A}_R} := \begin{bmatrix} T_1 & \ldots & T_R \end{bmatrix}. \tag{4.7}$$

The condition number satisfies

$$\kappa^{\mathrm{SBTD}}(\mathcal{A}_1, \ldots, \mathcal{A}_R) = \frac{1}{\sigma_{\min}(T_{\mathcal{A}_1,\ldots,\mathcal{A}_R})}, \tag{4.8}$$

where $\sigma_{\min}(A) = \sigma_{\min\{m,n\}}(A)$ denotes the smallest singular value of $A \in \mathbb{R}^{m\times n}$. Thus, the computation of κ^{SBTD} requires orthonormal bases of the tangent spaces to the structured Tucker manifolds. We explain this in the next proposition.

Recall that the compact higher-order singular value decomposition (HOSVD) [Tuc66; DLDMV00a] is an orthogonal Tucker decomposition $\mathcal{X} = (U_1, \ldots, U_D)\cdot\mathcal{C}$ with $U_d \in \mathbb{R}^{n_d\times l_d}$ a basis of left singular vectors of $\mathcal{X}_{(d)}$ corresponding to the nonzero singular values. In particular, $U_d^T U_d = \mathbb{1}_{l_d}$ and the columns of U_d span the column span of $\mathcal{X}_{(d)}$. The core tensor $\mathcal{C}$ is the orthogonal projection of $\mathcal{X}$ onto the orthonormal basis $U_1 \otimes \cdots \otimes U_D$: $\mathcal{C} = (U_1^T, \ldots, U_D^T) \cdot \mathcal{X}$. With this terminology in place, we can state the result.

Proposition 4.9. *Let $\mathcal{M}^{n_1,\ldots,n_D} \subseteq \mathbb{R}^{n_1\times\cdots\times n_D}$ be the $\mathcal{M}$-structured Tucker manifold with Tucker core structure $\mathcal{M} \subseteq \mathbb{R}^{l_1\times\cdots\times l_D}$. Assume that we are given a tensor $\mathcal{X} \in \mathcal{M}^{n_1,\ldots,n_D}$ expressed in HOSVD format $\mathcal{X} = (U_1, \ldots, U_D)\cdot\mathcal{C}$. Complete each U_d to an orthonormal basis $\begin{bmatrix} U_d & U_d^{\perp}\end{bmatrix}$ of $\mathbb{R}^{n_d}$. Let $\sigma_j^d := \left\| e_j^T \mathcal{C}_{(d)} \right\|$ and $\hat{u}_j^d := (\sigma_j^d)^{-1} e_j^{(l_d)}$. Let $\mathscr{B}_C$ be an orthonormal basis of $\mathcal{T}_{\mathcal{C}}\mathcal{M}$. Then an orthonormal basis of $\mathcal{T}_{\mathcal{X}}\mathcal{M}^{n_1,\ldots,n_D}$ is given by*

$$\mathscr{B}_{\mathcal{X}} := \mathscr{B}_{\mathcal{X}}^0 \cup \cdots \cup \mathscr{B}_{\mathcal{X}}^D \quad \textit{where} \quad \mathscr{B}_{\mathcal{X}}^0 := \left\{ (U_1, \ldots, U_D)\cdot\dot{\mathcal{C}} \mid \dot{\mathcal{C}} \in \mathscr{B}_{\mathcal{C}} \right\}, \tag{4.9}$$

$$\mathscr{B}_{\mathcal{X}}^d := \left\{ (U_1, \ldots, U_{d-1}, U_d^{\perp} e_i (\hat{u}_j^d)^T, U_{d+1}, \ldots, U_D)\cdot\mathcal{C} \right\}_{i,j=1}^{n_d-l_d,l_d}$$

for $d = 1, \ldots, D$.

Proof. Eq. (4.5) for $\mathcal{T}_{\mathcal{X}}\mathcal{M}^{n_1,\ldots,n_D}$ suggests a decomposition of the tangent space of the form

$$T_{\mathcal{X}}\mathcal{M}^{n_1,\ldots,n_D} = \mathbb{T}_0 \oplus \mathbb{T}_1 \oplus \cdots \oplus \mathbb{T}_D,$$

where $\mathbb{T}_0$ contains all tangents of the form $(U_1, \ldots, U_D)\cdot\dot{\mathcal{C}}$ and $\mathbb{T}_d$ with $d = 1, \ldots, D$ contains the tangents of the form $(U_1, \ldots, U_{d-1}, \dot{U}_d, U_{d+1}, \ldots, U_D)\cdot\mathcal{C}$. Since $\dot{U}_d^T U_d = 0_{l_d\times l_d}$ as in Proposition 4.8, this is a decomposition of $\mathcal{T}_{\mathcal{X}}\mathcal{M}^{n_1,\ldots,n_D}$ into pairwise orthogonal spaces.

First we verify that (4.9) spans $\mathcal{T}_{\mathcal{X}}\mathcal{M}^{n_1,\dots,n_D}$. Since we have an orthonormal basis of $\mathcal{T}_{\mathcal{C}}\mathcal{M}$ available, we have

$$\mathbb{T}_0 = \operatorname{span}\left\{(U_1,\dots,U_D)\cdot\dot{\mathcal{C}} \mid \dot{\mathcal{C}} \in \mathscr{B}_{\mathcal{C}}\right\} = (U_1\otimes\cdots\otimes U_D)(T_{\mathcal{C}}\mathcal{M}).$$

For the other D subspaces $\mathbb{T}_d$, we require all $\dot{U}_d$ such that $U_d^T\dot{U}_d = 0$, or equivalently $\dot{U}_d = U_d^{\perp}B$ for some $B \in \mathbb{R}^{(n_d-l_d)\times l_d}$. The $e_i^{(n_d-l_d)}(\hat{u}_j^d)^T$ with $i = 1,\dots,n_d-l_d$ and $j = 1,\dots,l_d$ are a basis of $\mathbb{R}^{(n_d-l_d)\times l_d}$, because they are just a rescaling of the canonical basis $e_i^{(n_d-l_d)}(e_j^{(l_d)})^T$. Substituting each of these for B, we get a basis of all allowed $\dot{U}_d$. This parametrises all of $\mathbb{T}_d$ as

$$\mathbb{T}_d = \operatorname{span}\left\{(U_1,\dots,U_{d-1},U_d^{\perp}e_i(\hat{u}_j^d)^T,U_{d+1},\dots,U_D)\cdot\mathcal{C}\right\}_{i=j=1}^{i=n_d-l_d,j=l_d}.$$

Hence, the proposed basis $\mathscr{B}_{\mathcal{X}}$ generates $\mathbb{T}_0\oplus\mathbb{T}_1\oplus\cdots\oplus\mathbb{T}_D$.

We have yet to verify that the proposed basis is orthonormal. We already know that $\mathbb{T}_0,\dots,\mathbb{T}_D$ are pairwise orthogonal. It thus suffices to show that the bases we constructed for each of these spaces separately are orthonormal. The basis for $\mathbb{T}_0$ is orthonormal because $\mathscr{B}_{\mathcal{C}}$ is orthonormal and $U_1\otimes\cdots\otimes U_D$ is an orthonormal tensor product basis.

For the basis of $\mathbb{T}_d$ with $d \geqslant 1$, we use the fact that $\mathcal{X}$ is in HOSVD format. This ensures that an HOSVD of $\mathcal{C}$ is $(\mathbb{1},\dots,\mathbb{1})\cdot\mathcal{C}$. In other words, $\mathcal{C}_{(d)}$ has singular values σ_j^d as defined above and its corresponding left singular vectors are e_j [DLDMV00a]. Hence, the transpose of its jth right singular vector is $(v_j^d)^T := (\sigma_j^d)^{-1}e_j^T\mathcal{C}_{(d)} = (\hat{u}_j^d)^T\mathcal{C}_{(d)}$. With this in mind, we calculate the inner products between the basis vectors of $\mathbb{T}_d$:

$$\begin{aligned}&\big\langle(U_1,\dots,U_{d-1},U_d^{\perp}e_i(\hat{u}_j^d)^T,U_{d+1},\dots,U_D)\cdot\mathcal{C},\\ &\qquad(U_1,\dots,U_{d-1},U_d^{\perp}e_{i'}(\hat{u}_{j'}^d)^T,U_{d+1},\dots,U_D)\cdot\mathcal{C}\big\rangle\\ &= \operatorname{Trace}(U_d^{\perp}e_{i'}(\hat{u}_{j'}^d)^T\mathcal{C}_{(d)}\mathcal{C}_{(d)}^T\hat{u}_{j'}^d e_{i'}^T(U_d^{\perp})^T)\\ &= \langle U_d^{\perp}e_i,U_d^{\perp}e_{i'}\rangle\,\langle v_j^d,v_{j'}^d\rangle.\end{aligned} \tag{4.10}$$

If $i = i'$, the right-hand side is the inner product between two right singular vectors of $\mathcal{C}_{(d)}$, which is $\delta_{jj'}$, the Kronecker delta. Otherwise, it is zero due to the orthogonality of the columns of $U_d^{\perp}$. This ensures that our basis of $\mathbb{T}_d$ is orthogonal, which completes the proof. □

Now we can compute the condition number of several decompositions using the formula in (4.8). Consider the following examples.

Example 1 (BTD) For a BTD with block terms of multilinear rank $(l_1^r, \dots, l_D^r)$, where $r = 1, \dots, R$, we can apply Definition 4.3 in which the Tucker core structure $\mathcal{M}$ is the submanifold of tensors in $\mathbb{R}^{l_1^r \times \cdots \times l_D^r}$ with multilinear rank $(l_1^r, \dots, l_D^r)$. Since $\mathcal{M}$ is an open subset of $\mathbb{R}^{l_1^r \times \cdots \times l_D^r}$, the canonical basis of $\mathbb{R}^{l_1^r \times \cdots \times l_D^r}$ is an orthonormal basis of the tangent space to $\mathcal{M}$ at any point. The algorithm to compute the condition number κ^{BTD} is as follows. For each term $\mathcal{A}_r$ in the BTD, compute its compact HOSVD $(U_1^r, \dots, U_D^r) \cdot \mathcal{C}_r$. An orthonormal basis of the tangent space to the Tucker manifold is given by the columns of

$$T_{\mathcal{A}_r} := \left[\bigotimes_{d=1}^{D} U_d^r \quad \left[\left(U_d^{r\perp} e_i (\hat{u}_d^{rj})^T \otimes_d \bigotimes_{d' \neq d} U_{d'}^r \right) \mathcal{C}_r \right]_{d,i,j=1}^{D, m_d - l_d^r, l_d^r} \right],$$

where $\hat{u}_d^{rj}$ and $U_d^{r\perp}$ are as in Proposition 4.9. The condition number of the BTD with terms $\mathcal{A}_1, \dots, \mathcal{A}_R$ can then be computed by applying (4.8):

$$\kappa^{\mathrm{BTD}}(\mathcal{A}_1, \dots, \mathcal{A}_R) = \sigma_{\min}(T_{\mathcal{A}_1, \dots, \mathcal{A}_R})^{-1}, \quad T_{\mathcal{A}_1, \dots, \mathcal{A}_R} = \begin{bmatrix} T_{\mathcal{A}_1} & \dots & T_{\mathcal{A}_R} \end{bmatrix}. \quad (4.11)$$

Example 2 (CPD) This case was studied in [BV18b]. By applying the Tucker core structure $\mathcal{M} := \mathbb{R} \backslash \{0\}$ to Definition 4.3, we get the Segre manifold of rank-1 tensors. If $\mathcal{A}_r = \lambda u_1^r \otimes \cdots \otimes u_D^r$ is a rank-1 tensor with $\|u_1\| = \cdots = \|u_D\| = 1$, Proposition 4.9 gives the following familiar basis:

$$T_{\mathcal{A}_r} := \left[u_1^r \otimes \cdots \otimes u_D^r \quad \left[u_1^r \otimes \cdots \otimes u_{d-1}^r \otimes U_d^{r\perp} \otimes u_{d+1}^r \otimes \cdots \otimes u_D^r \right]_{d=1}^{D} \right],$$

where $U_d^{r\perp}$ is an orthonormal basis for the complement of u_d^r for each d. Note that this is equivalent to the previous example with all $l_d^r = 1$. Hence, the condition number κ^{CPD} can be computed in a similar fashion as for the BTD.

4.4.2 Examples of well and ill-conditioned SBTDs

In this subsection, we present some qualitative properties that determine the condition number of the SBTD. As a general rule, the tensor subspace in which the summands live already gives some information about the condition number. For instance, one instance where the condition number is perfect is when the subspaces U_d^r in the Tucker decompositions are pairwise orthogonal. This can be considered as the SBTD equivalent of an orthogonally decomposable (odeco) tensor [ZG01]. In such cases, tangent spaces are pairwise orthogonal, so that the following result holds.

Proposition 4.10. *Suppose $\mathcal{A}_1 + \cdots + \mathcal{A}_R$ is an SBTD with $\mathcal{A}_r = (U_1^r, \dots, U_D^r) \cdot \mathcal{C}_r$ in HOSVD form for $r = 1, \dots, R$. Assume that $(U_d^{r_1})^T U_d^{r_2} = 0$ for each d and each $r_1 \neq r_2$. Then $\kappa^{\mathrm{SBTD}}(\mathcal{A}_1, \dots, \mathcal{A}_r) = 1$.*

Proof. The columns of the Terracini matrix can be partitioned into orthonormal bases of subspaces of $\operatorname{span}\{\bigotimes_{d=1}^{D} U_d^r\}$ and $\operatorname{span}\{U_d^{r\perp} \otimes_d \bigotimes_{d' \neq d} U_{d'}^r\}$ for all $r = 1, \ldots, R$ and for all $d = 1, \ldots, D$. By assumption, these spaces are all pairwise orthogonal. Since the bases are orthonormal, all columns of the Terracini matrix are orthonormal. □

The fact that this result does not depend on the cores $\mathcal{C}_r$ may be surprising if the problem is not considered geometrically. $\mathcal{C}_r$ may be arbitrarily close to having a multilinear rank lower than the specified $(l_1^r, \ldots, l_D^r)$ without it affecting the condition number. Despite this, summands which are close to being low multilinear rank are a notorious obstacle in practical algorithms to compute the BTD, the other being correlations between the terms [NDLK08]. Note that the latter is essentially what the condition number measures. For tensors of lower multilinear rank than $(l_1^r, \ldots, l_D^r)$, there are more ways to parametrise it than is accounted for by the usual symmetries. For instance, for a BTD with multilinear ranks $(l_r, l_r, 1)$, the Jacobian of the residual $\mathscr{A} - \sum_{r=1}^{R} \mathscr{A}_r$ with respect to the parameters becomes singular at such points [SVBDL13]. The ALS algorithm for a general block term decomposition requires solving a system which also becomes singular at the boundary [DLN08].

However, if one studies changes to the points $\mathscr{A}_r \in \mathcal{M}_r$ and abstracts away how they are parametrised, summands close to tensors of lower multilinear rank are not an issue. Hence, it is still reasonable to expect the condition number to be 1 even near the boundary. This suggests that Riemannian optimisation algorithms to compute the BTD could have a significant advantage in these cases, as the convergence rate tends to be related to the condition number. This is analogous to the case of the CPD, where [BV18b] showed experimentally and theoretically that classic flat optimisation methods perform worse if the CPD contains summands of small norm—the analogous situation to a lower multilinear rank in $\mathcal{M}$-structured Tucker decompositions—while Riemannian optimisation methods that treat the summands as one geometric object did not suffer as much.

A geometric understanding of the structured Tucker manifold gives insight into the condition number. For instance, we have the following result from [BV18b].

Proposition 4.11. *Let $\mathscr{A} = \sum_{r=1}^{R} \mathscr{A}_r$ be an SBTD with $\mathscr{A}_r \in \mathcal{M}_r$. If there exists an injective continuous curve $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{M}_1 \times \cdots \times \mathcal{M}_R$ with $\varepsilon > 0$ so that $\gamma(0) = (\mathscr{A}_1, \ldots, \mathscr{A}_R)$ and $\Sigma(\gamma(t)) = \mathscr{A}$ for all t, then $\kappa^{SBTD}(\mathscr{A}_1, \ldots, \mathscr{A}_R) = \infty$.*

If such a curve exists, a zero-norm perturbation to $\mathscr{A}$ is sufficient to get different decompositions, so that the assertion follows immediately from (4.2). A trivial example is the following: let $\mathscr{A} = \sum_{r=1}^{R-1} \mathscr{A}_r$. We can generate R-term SBTDs of the form $(\mathcal{A}_1, \ldots, t\mathcal{A}_{R-1}, (1-t)\mathcal{A}_{R-1})$ with $t \in \mathbb{R}$. Thus, if a tensor admits an SBTD with $R-1$ summands, some of its SBTDs with more summands have an infinite condition number.

However, not all SBTDs with more than the minimum number of summands have an infinite condition number, as the following example illustrates. For a generic decomposition with symmetric rank-1 summands $\mathcal{A}_1, \ldots, \mathcal{A}_{18} \in \left(\mathbb{C}^3\right)^{\otimes 9}$, there exist complex symmetric rank-1 tensors $\mathcal{B}_1, \ldots, \mathcal{B}_{17}$ such that no two of $\{\mathcal{A}_1, \ldots, \mathcal{A}_{18}, \mathcal{B}_1, \ldots, \mathcal{B}_{17}\}$ are linearly dependent and $\sum_{r=1}^{18} \alpha_r \mathcal{A}_r = \sum_{r=1}^{17} \beta_r \mathcal{B}_r$ for some choice of the coefficients $\alpha_r, \beta_r \neq 0$ [AC20]. The condition number of the decomposition $\sum_{r=1}^{18} \alpha_r \mathcal{A}_r$ can be calculated from the Terracini matrix of $(\mathcal{A}_1, \ldots, \mathcal{A}_{18})$. This condition number is independent of the choice of $\alpha_r \neq 0$. For generic points $\mathcal{A}_r$, the Terracini matrix is not singular, so that the condition number is generically finite. To illustrate this numerically, we generated 2000 such decompositions with $\mathcal{A}_r = a_r \otimes \cdots \otimes a_r$ where the a_r are sampled uniformly on the (real) sphere. The geometric mean of $\kappa^{CPD}(\mathcal{A}_1, \ldots, \mathcal{A}_{18})$ was about $5 \cdot 10^4$. Thus, the decompositions $\sum_{r=1}^{18} \alpha_r \mathcal{A}_r$ have a finite condition number on average, even though such decompositions are generically not minimal for some choice of α_r.

In general, the condition number of any SBTD can be upper bounded by the condition number of the corresponding BTD, and for the latter we can get a useful lower bound for the condition number. We show this in the next proposition.

Proposition 4.12. *Given any SBTD $\mathcal{A} = \mathcal{A}_1 + \cdots + \mathcal{A}_R$, we can also regard it as a BTD of $\mathcal{A}$. The condition numbers satisfy*

$$\kappa^{\mathrm{BTD}}(\mathcal{A}_1, \ldots, \mathcal{A}_R) \geqslant \kappa^{\mathrm{SBTD}}(\mathcal{A}_1, \ldots, \mathcal{A}_R).$$

If the terms $\mathcal{A}_r = (U_1^r, \ldots, U_D^r) \cdot \mathcal{C}_r$ are in HOSVD form for $r = 1, \ldots, R$, then

$$\kappa^{\mathrm{BTD}}(\mathcal{A}_1, \ldots, \mathcal{A}_R) \geqslant \sigma_{\min}\left([U_1^r \otimes \cdots \otimes U_D^r]_{r=1}^R \right)^{-1}.$$

Proof. Assume $\kappa^{\mathrm{BTD}}(\mathcal{A}_1, \ldots, \mathcal{A}_R) < \infty$, since otherwise the statement is trivially true. For each $r = 1, \ldots, R$, the rth structured Tucker manifold $\mathcal{M}_r^{n_1,\ldots,n_D}$ of the SBTD is a subset of the manifold $\mathcal{N}_r^{n_1,\ldots,n_D}$ of tensors of fixed multilinear rank. By assumption, the addition map $\Sigma(\mathcal{A}_1, \ldots, \mathcal{A}_R) = \mathcal{A}_1 + \cdots + \mathcal{A}_R$ of the BTD has a local inverse function $\Sigma^{-1}_{\mathcal{A}_1,\ldots,\mathcal{A}_R}$, defined on a neighbourhood $\mathcal{I} \subseteq \Sigma(\mathcal{N}_1^{n_1,\ldots,n_D}, \ldots, \mathcal{N}_R^{n_1,\ldots,n_D})$ of $\mathcal{A}$. On the other hand, for any $\widetilde{\mathcal{A}} \in \mathcal{I}' := \mathcal{I} \cap \Sigma(\mathcal{M}_1^{n_1,\ldots,n_D}, \ldots, \mathcal{M}_R^{n_1,\ldots,n_D})$, the (locally unique) SBTD of $\widetilde{\mathcal{A}}$ is $\Sigma^{-1}_{\mathcal{A}_1,\ldots,\mathcal{A}_R}(\widetilde{\mathcal{A}})$. The condition numbers of the BTD and SBTD are (4.2) applied to $\Sigma^{-1}_{\mathcal{A}_1,\ldots,\mathcal{A}_R}$

and the restriction of $\Sigma^{-1}_{\mathcal{A}_1,\ldots,\mathcal{A}_R}$ onto $\mathcal{I}'$, respectively. Since $\mathcal{I}' \subseteq \mathcal{I}$, the first statement follows.

For the second assertion, observe that the columns of $[U_1^r \otimes \cdots \otimes U_D^r]_{r=1}^R$ are a subset of the columns of the Terracini matrix T of the BTD. By [GVL13, Theorem 8.1.7], $\sigma_{\min}([U_1^r \otimes \cdots \otimes U_D^r]_{r=1}^R) \geqslant \sigma_{\min}(T)$. The result follows from (4.8). □

The second item in the above proposition shows that

$$\text{if } \ker [U_1^r \otimes \cdots \otimes U_D^r]_{r=1}^R \neq \{0\}, \text{ then } \kappa^{\mathrm{BTD}}(\mathcal{A}_1,\ldots,\mathcal{A}_R) = \infty.$$

Another way to see this is the following: if the spaces intersect, there exist cores $\widetilde{\mathcal{C}}_r$ with $r = 1,\ldots,R$ so that $\sum_{r=1}^R (U_1^r \otimes \cdots \otimes U_D^r)\widetilde{\mathcal{C}}_r = 0$ and not all $\widetilde{\mathcal{C}}_r = 0$. Then we can define the curve $\gamma(t) := (\gamma_1(t),\ldots,\gamma_R(t))$, where $\gamma_r(t) := (U_1^r \otimes \cdots \otimes U_D^r)(\mathcal{C}_r + t\widetilde{\mathcal{C}}_r)$, so that the condition number is ∞ by Proposition 4.11.

This leads to the following observation: if the condition number is finite, the BTD can be determined purely from subspace information. That is, suppose that for a given tensor $\mathcal{A}$, only the subspaces $U_1^r \otimes \cdots \otimes U_D^r$ are computed for each rth block term, with $r = 1,\ldots,R$. Because the subspaces do not intersect, the cores $\mathcal{C}_r$ can be uniquely recovered from the linear system $\mathcal{A} = \sum_{r=1}^R (U_1^r \otimes \cdots \otimes U_D^r)\mathcal{C}_r$ This is exactly the principle behind the variable projection methods in [OADL18].

It is worth pointing out that the second lower bound from Proposition 4.12 is not necessarily sharp. To see this, let $\mathcal{M}^{n_1,\ldots,n_D}$ be a structured Tucker manifold and consider the SBTD $\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2$ with the two summands having Tucker compressions $\mathcal{A}_1 = (U_1,\ldots,U_D) \cdot \mathcal{C} \in \mathcal{M}^{n_1,\ldots,n_D}$ and $\mathcal{A}_2 = (V_1, U_2,\ldots,U_D) \cdot \mathcal{C} \in \mathcal{M}^{n_1,\ldots,n_D}$. We assume that $U_1^T V_1 = 0$ and we define the two curves $\gamma_1(t) := (U_1 + tV_1, U_2,\ldots,U_D) \cdot \mathcal{C}$ and $\gamma_2(t) := ((1-t)V_1, U_2,\ldots,U_D) \cdot \mathcal{C}$. Assuming t is small enough, we have that $\gamma_1(t), \gamma_2(t) \in \mathcal{M}^{n_1,\ldots,n_D}$. As in the previous example $\gamma_1(t) + \gamma_2(t) = \mathcal{A}$, for all t. Hence, the condition number is also infinite in this case. Despite this, the estimated lower bound in Proposition 4.12 is $\sigma_{\min}([U_1 \otimes U_2 \otimes \cdots \otimes U_D, V_1 \otimes U_2 \otimes \cdots \otimes U_D]) = 1$.

4.5 Invariance of the condition number under Tucker compression

Next, we discuss the main contribution of this work. Our main result was informally stated as Theorem 4.1 in the introduction. Here, we present its formal version, Theorem 4.14. These two theorems show that the condition number of computing SBTDs is invariant under Tucker compression. As we explain in the Section 4.5.2, this can yield a computationally attractive approach for computing the condition number.

First, we introduce *subspace-constrained SBTDs* as the formal model of decompositions resulting from Bro and Andersson's [BA98] compress-decompose-expand approach. Subspace-constrained CPDs were also considered in the recent paper [Pha+21]. The compress-decompose-expand approach assumes that a tensor that lives in a subspace $\mathbb{W}_1 \otimes \cdots \otimes \mathbb{W}_D$ and has an R-term decomposition also has an R-term decomposition where each summand lives in $\mathbb{W}_1 \otimes \cdots \otimes \mathbb{W}_D$. This is the case for the CPD [SL08, Proposition 3.1]. The following is a slightly weaker statement in the case of the SBTD.

Proposition 4.13. *Suppose $\mathcal{A} = (Q_1, \ldots, Q_D) \cdot \mathcal{G} \in \mathbb{R}^{n_1 \times \cdots \times n_D}$ where $\mathcal{G} \in \mathbb{R}^{m_1 \times \cdots \times m_D}$ and $Q_d \in \mathbb{R}_\star^{n_d \times m_d}$ for all d. If $\mathcal{A}$ has an SBTD $\mathcal{A} = \sum_{r=1}^R \mathcal{A}_r$ with $\mathcal{A}_r \in \mathcal{M}_r^{n_1,\ldots,n_D}$ for some Tucker core structures $\mathcal{M}_r \subseteq \mathbb{R}^{l_1^r \times \cdots \times l_D^r}$ and $l_d^r \leqslant m_d$ for all r and d, then at least one of the following statements holds:*

1. *$\mathcal{G}$ has an SBTD $\mathcal{G} = \sum_{r=1}^R \mathcal{G}_r$ with $\mathcal{G}_r \in \mathcal{M}_r^{m_1,\ldots,m_D}$.*
2. *There exist Tucker core structures $\mathcal{N}_r \subseteq \mathbb{R}^{\ell_1^r \times \cdots \times \ell_D^r}$ where $\ell_d^r \leqslant l_d^r$ for all r and d where at least one inequality is strict and $\mathcal{A} = \sum_{r=1}^R \widetilde{\mathcal{A}}_r$ for $\widetilde{\mathcal{A}}_r \in \mathcal{N}_r^{n_1,\ldots,n_D}$.*[3]

The first statement of Proposition 4.13 implies that $\mathcal{A}$ has an SBTD of the form $\mathcal{A} = \sum_{r=1}^R (Q_1, \ldots, Q_D) \cdot \mathcal{G}_r$, in which each summand lies in the same subspace as $\mathcal{A}$, i.e., $\operatorname{span}(Q_1 \otimes \cdots \otimes Q_D)$. The second statement says that the original SBTD of $\mathcal{A}$ can be simplified in the sense that, for at least one $r \in \{1, \ldots, R\}$, the multilinear rank of $\widetilde{\mathcal{A}}_r$ is less than that of $\mathcal{A}_r$. By repeatedly applying Proposition 4.13 to the SBTD that comes out of the second statement, one eventually reaches an SBTD model where the first statement holds.

Proof of Proposition 4.13. Write the summands in the SBTD as $\mathcal{A}_r = (U_1^r, \ldots, U_D^r) \cdot \mathcal{C}_r$. Then $\mathcal{G} = \sum_{r=1}^R (Q_1^\dagger U_1^r, \ldots, Q_D^\dagger U_D^r) \cdot \mathcal{C}_r$. Let ℓ_d^r be the rank of $Q_d^\dagger U_d^r$. We distinguish between two cases.

If $\ell_d^r = l_d^r$ for all d and r, the first statement holds.

Otherwise, define $\mathcal{N}_r := \left\{ (V_1^r, \ldots, V_D^r) \cdot \mathcal{X} \,\middle|\, \mathcal{X} \in \mathcal{M}_r,\ (V_d^r)^T \in \mathbb{R}_\star^{l_d^r \times \ell_d^r} \right\}$. If any $\ell_d^r = 0$, then $\mathcal{N}_r = \emptyset$ and we drop the rth summand. The two requirements in Definition 4.2 of a core structure can be verified respectively by the fact

[3] For notational convenience, $\ell_d^r = 0$ means that the rth summand is omitted.

that $\operatorname{rank}(V_d^r \mathcal{X}_{(d)}) = \operatorname{rank}(V_d^r)$ whenever $\operatorname{rank}(\mathcal{X}_{(d)}) = l_d^r$ and the fact that $(V_d^r)^T$ is universally quantified over $\mathbb{R}_\star^{l_d^r \times \ell_d^r}$. If we write $Q_d^\dagger U_d^r = W_d^r V_d^r$ with $W_d^r \in \mathbb{R}_\star^{n_d \times \ell_d}$ and $(V_d^r)^T \in \mathbb{R}_\star^{l_d \times \ell_d}$, we can expand out $\mathcal{A} = (Q_1, \dots, Q_D) \cdot \mathcal{G}$ and the SBTD of $\mathcal{G}$ to obtain

$$\mathcal{A} = \sum_{r=1}^{R} (Q_1 W_1^r, \dots, Q_D W_D^r) \cdot ((V_1^r, \dots, V_D^r) \cdot \mathcal{C}_r)$$

as desired. □

Proposition 4.13 justifies the following method by Bro and Andersson [BA98] to compute a subspace-constrained SBTD of $\mathcal{A}$.

Compress

$\mathcal{A}$ lives in a minimal tensor product subspace of $\mathbb{R}^{n_1 \times \cdots \times n_D}$ (possibly trivial). Its minimal Tucker decomposition is $\mathcal{A} = (Q_1, \dots, Q_D) \cdot \mathcal{G}$ with core tensor $\mathcal{G} \in \mathbb{R}^{m_1 \times \cdots \times m_D}$ and matrices $Q_d \in \mathbb{R}_\star^{n_d \times m_d}$ for $d = 1, \dots, D$. The decomposition is minimal if $\mathcal{G}$ has multilinear rank equal to $(m_1, \dots, m_D)$. It can be computed with a (sequentially) truncated higher-order singular value decomposition [VVM12; DLDMV00a].

Decompose

Decompose $\mathcal{G}$ as a sum of $\mathcal{M}_r$-structured Tucker tensors, i.e.,

$$\mathcal{G} = \mathcal{G}_1 + \cdots + \mathcal{G}_R \quad \text{with} \quad \mathcal{G}_r \in \mathcal{M}_r^{m_1, \dots, m_D}. \tag{4.12}$$

This is performed by an algorithm specific to the core structure (e.g. [SVBDL13; ERL22]).

Expand

We expand $\mathcal{A}_r = (Q_1 U_1^r, \dots, Q_D U_D^r) \cdot \mathcal{C}_r$ and find a decomposition $\mathcal{A} = \mathcal{A}_1 + \cdots + \mathcal{A}_R$ of $\mathcal{A}$. In this decomposition the summands are also $\mathcal{M}_r$-structured Tucker tensors: if we have the $\mathcal{M}_r$-structured Tucker decomposition $\mathcal{G}_r = (U_1^r, \dots, U_D^r) \cdot \mathcal{C}_r$ such that $U_d^r \in \mathbb{R}_\star^{m_d \times l_d}$ and $\mathcal{C}_r$ is a point of $\mathcal{M}_r \subset \mathbb{R}^{l_1 \times \cdots \times l_D}$ satisfying the assumptions of

Definition 4.3, then for all r and d the matrices $Q_d U_d^r$ are of full rank. Hence, $\mathcal{A}_r = (Q_1 U_1^r, \ldots, Q_D U_D^r) \cdot \mathcal{C}_r$ is a point of the $\mathcal{M}_r$-structured Tucker manifold $\mathcal{M}_r^{n_1,\ldots,n_D}$. Summarising, the SBTD (4.12) of the compressed tensor $\mathcal{G}$ can be expanded to an SBTD of $\mathcal{A}$:

$$\mathcal{A} = \mathcal{A}_1 + \cdots + \mathcal{A}_R \quad \text{with} \quad \mathcal{A}_r \in \mathcal{M}_r^{n_1,\ldots,n_D}, \quad r = 1, \ldots, R. \tag{4.13}$$

We call the resulting SBTD of $\mathcal{A}$ a *subspace-constrained SBTD*, because it is an SBTD all of whose summands are contained in the same tensor subspace $Q_1 \otimes \cdots \otimes Q_d$ that $\mathcal{A}$ lives in.

This procedure is implemented for several decompositions in software packages such as Tensorlab [VDDL17]. If the "decompose" step is performed by an algorithm that iterates over the manifold, the convergence of such an algorithm would depend on geometric properties of $\mathcal{M}_r^{m_1,\ldots,m_D}$ and $\mathcal{M}_r^{n_1,\ldots,n_D}$ such as path-connectedness. For instance, the set of tensors of multilinear rank $(l_1, \ldots, l_D)$ is path-connected if $l_d < \prod_{d' \neq d} l_{d'}$ for all $d = 1, \ldots, D$ [Com+20]. However, issues related to the convergence of these algorithms are beyond the scope of this thesis.

Given that a subspace-constrained SBTD can be computed by the foregoing compress-decompose-expand approach, it is natural to wonder about the relationship between the condition numbers of $\mathcal{A}$ and $\mathcal{G}$. Since $\mathcal{G}$ lives in a much more constrained space, it seems natural to assume that its condition number could be much lower, similar to the ideas in [ANT19]. In Section 4.5.1 below, we prove the following main result about the condition numbers of computing the SBTD (4.13) of the original tensor and computing the SBTD of the compressed Tucker core (4.12). A priori, the condition number of the decomposition problem (4.12) is bounded above by the condition number of problem (4.13). The next result shows that they are, in fact, always equal.

Theorem 4.14. *Let $\mathcal{M}_r \subset \mathbb{R}^{l_1^r \times \cdots \times l_D^r}$ be Tucker core structures. Assume that the tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times \cdots \times n_D}$ has an orthogonal Tucker decomposition $\mathcal{A} = (Q_1, \ldots, Q_D) \cdot \mathcal{G}$ with $\mathcal{G} \in \mathbb{R}^{m_1 \times \cdots \times m_D}$ and all $Q_d \in \mathbb{R}^{n_d \times m_d}$ having orthonormal columns. Let the subspace-constrained SBTD be $\mathcal{A} = \mathcal{A}_1 + \cdots + \mathcal{A}_R$ with $\mathcal{A}_r \in \mathcal{M}_r^{n_1,\ldots,n_D}$ and the SBTD of the Tucker core be $\mathcal{G} = \mathcal{G}_1 + \cdots + \mathcal{G}_R$ with $\mathcal{G}_r \in \mathcal{M}_r^{m_1,\ldots,m_D}$, and assume that they are related by $\mathcal{A}_r = (Q_1, \ldots, Q_D) \cdot \mathcal{G}_r$ for each $r = 1, \ldots, R$. Then,*

$$\kappa^{\mathrm{SBTD}}(\mathcal{A}_1, \ldots, \mathcal{A}_R) = \kappa^{\mathrm{SBTD}}(\mathcal{G}_1, \ldots, \mathcal{G}_R).$$

Before presenting the proof in Section 4.5.1, let us investigate some consequences of Theorem 4.14. In the subspace-constrained SBTD there are two levels of multilinear multiplication. The first level is in Definition 4.3 of the structured Tucker decompositions. This level is always written with matrices $U_1^r, \dots, U_D^r$ that depend on the index of the summand. The second level is the multilinear multiplication defining the subspace constraint on the tensor $\mathcal{A}$. This level is denoted with matrices $Q_1, \dots, Q_D$ and it is the same for all summands. This is summarised in the following diagram:

$$\begin{tikzcd} \mathcal{A}_r \in \mathcal{M}_r^{n_1,\dots,n_D} & \mathcal{G}_r \in \mathcal{M}_r^{m_1,\dots,m_D} \arrow[l, "(Q_1\otimes\cdots\otimes Q_D)"] & \mathcal{C}_r \in \mathcal{M}_r. \arrow[l, "(U_1^r\otimes\cdots\otimes U_D^r)"] \end{tikzcd}$$

It is imperative to note, however, that $(Q_1 \otimes \cdots \otimes Q_D)\mathcal{M}_r^{m_1,\dots,m_D} \subsetneq \mathcal{M}_r^{n_1,\dots,n_D}$.

When evaluating the sensitivity of a subspace-constrained SBTD (4.13) via the condition number (4.2), there are at least four natural sets of perturbations $\mathcal{I}$ to consider. Let $\widetilde{\mathcal{A}}$ denote the perturbed tensor. It could have resulted from one of the following increasingly restrictive perturbations of the subspace-constrained SBTD $\mathcal{A}$:

1. $\mathcal{A}$ was perturbed with no constraints and $\widetilde{\mathcal{A}}$ was approximated by the closest SBTD $\widetilde{\mathcal{A}} \approx \widetilde{\mathcal{A}}_1 + \cdots + \widetilde{\mathcal{A}}_R$ with $\widetilde{\mathcal{A}}_r \in \mathcal{M}_r^{n_1,\dots,n_D}$;
2. $\mathcal{A}$ was perturbed so $\widetilde{\mathcal{A}}$ has an SBTD $\widetilde{\mathcal{A}} = \widetilde{\mathcal{A}}_1 + \cdots + \widetilde{\mathcal{A}}_R$ with $\widetilde{\mathcal{A}}_r \in \mathcal{M}_r^{n_1,\dots,n_D}$;
3. $\mathcal{A}$ was perturbed so $\widetilde{\mathcal{A}}$ has a subspace-constrained SBTD $\widetilde{\mathcal{A}} = (\widetilde{Q}_1, \dots, \widetilde{Q}_D) \cdot \widetilde{\mathcal{G}}$ with core $\widetilde{\mathcal{G}} = \widetilde{\mathcal{G}}_1 + \cdots + \widetilde{\mathcal{G}}_R$ and $\widetilde{\mathcal{G}}_r \in \mathcal{M}_r^{m_1,\dots,m_D}$; or
4. $\mathcal{A}$ was perturbed inside the fixed subspace $Q_1 \otimes \cdots \otimes Q_D$ so $\widetilde{\mathcal{A}}$ has a subspace-constrained SBTD $\widetilde{\mathcal{A}} = (Q_1, \dots, Q_D) \cdot \widetilde{\mathcal{G}}$ with core $\widetilde{\mathcal{G}} = \widetilde{\mathcal{G}}_1 + \cdots + \widetilde{\mathcal{G}}_R$ and terms in the decomposition $\widetilde{\mathcal{G}}_r \in \mathcal{M}_r^{m_1,\dots,m_D}$.

Since there are 4 domains of perturbations we can consider in (4.2), there are also 4 associated, a priori distinct, condition numbers. Let us denote the condition number corresponding to the ith type of perturbation by κ_i. Then we have

$$\kappa_1 \geqslant \kappa_2 = \kappa^{\mathrm{SBTD}}(\mathcal{A}_1, \dots, \mathcal{A}_R) \geqslant \kappa_3 \geqslant \kappa_4 = \kappa^{\mathrm{SBTD}}(\mathcal{G}_1, \dots, \mathcal{G}_R). \tag{4.14}$$

However, it was already proven in [BV21, Corollary 5.5] that $\kappa_1 = \kappa_2$, i.e., arbitrary perturbations in combination with a least-squares approximation are no worse than structured perturbations. Combining this with Theorem 4.14 immediately implies the following more formal restatement of Theorem 4.1.

Corollary 4.15. *Suppose that we have SBTDs $\mathcal{G} = \mathcal{G}_1 + \cdots + \mathcal{G}_R \in \mathbb{R}^{m_1\times\cdots\times m_D}$ and $\mathcal{A} = \mathcal{A}_1 + \cdots + \mathcal{A}_R \in \mathbb{R}^{n_1\times\cdots\times n_D}$ related by $\mathcal{A}_r = (Q_1 \otimes \cdots \otimes Q_D)\,\mathcal{G}_r$ for each $r = 1, \dots, R$. If all Q_d have orthonormal columns, then* (4.14) *is an equality.*

4.5.1 Proof of the main result

Proposition 4.9 allows us to prove our main result, Theorem 4.14. Before we do this, we need the following lemma.

Lemma 4.16. *For any set of matrices $A_k \in \mathbb{R}^{m \times n_k}$ and any set of orthogonal matrices $Q_k \in \mathbb{R}^{p \times p}$ where $k = 1, \ldots, K$, the matrices*

$$X := \begin{bmatrix} A_1 & \cdots & A_K \end{bmatrix} \quad \textit{and} \quad Y := \begin{bmatrix} A_1 \otimes Q_1 & \cdots & A_K \otimes Q_K \end{bmatrix}$$

have the same singular values up to multiplicities.

Proof. Define the block diagonal matrix $D := \operatorname{diag}(\mathbb{1}_{n_1} \otimes Q_1, \ldots, \mathbb{1}_{n_K} \otimes Q_K)$. Then $Y = [A_1 \otimes \mathbb{1}_p \ \ldots \ A_K \otimes \mathbb{1}_p]\, D$. Since D is orthogonal, Y has the same singular values as $[A_1 \otimes \mathbb{1}_p \ldots A_K \otimes \mathbb{1}_p]$. Up to a permutation of rows and columns, this is $X \otimes \mathbb{1}_p$. The proof is completed by applying the singular value property of Kronecker products [GVL13, section 12.3.1]. □

Remark 4.17. If each A_k in the above lemma is itself a Kronecker product of at least d matrices and we replace Y by $\begin{bmatrix} Q_1 \otimes_d A_1 & \ldots & Q_K \otimes_d A_K \end{bmatrix}$, the statement still holds, because changing the order of the factors only changes their Kronecker product by a permutation of rows and columns [GVL13, Equation 12.3.1].

Now we can prove that the condition number of the SBTD is invariant under Tucker compression.

Proof of Theorem 4.14. For the SBTDs $\mathcal{A} = \mathcal{A}_1 + \cdots + \mathcal{A}_R$ and $\mathcal{G} = \mathcal{G}_1 + \cdots + \mathcal{G}_R$ we denote their associated Terracini matrices by $T_{\mathcal{A}_1,\ldots,\mathcal{A}_R}$ and $T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}$, respectively (see (4.7)). Our strategy is assembling the Terracini matrices in an appropriate way so that we can compare their singular values.

For each $r = 1, \ldots, R$, we apply Proposition 4.9 to obtain a basis $\mathscr{B}_{\mathcal{G}_r}$ for $\mathcal{T}_{\mathcal{G}_r}\mathcal{M}_r^{m_1,\ldots,m_D}$. That is, we write $\mathcal{G}_r$ in HOSVD form $(U_1^r, \ldots, U_D^r) \cdot \mathcal{C}_r$ and compute matrices $U_d^{r\perp}$ so that $[U_d^r \quad U_d^{r\perp}]$ is orthogonal for each d. The vectors $\hat{u}_{rj}^d$ are scaled versions of $e_j^{(l_d^r)}$ as in Proposition 4.9. That is, they are defined such that the transpose of jth right singular vector of the dth flattening of the rth core $(\mathcal{C}_r)_{(d)}$ is $(\hat{u}_j^d)^T (\mathcal{C}_r)_{(d)}$. This gives the basis

$$\mathscr{B}_{\mathcal{G}_r} := \left\{(U_1^r, \ldots, U_D^r) \cdot \dot{\mathcal{C}}_r\right\} \cup \left\{(U_1^r, \ldots, U_{d-1}^r, U_d^{r\perp} e_i (\hat{u}_{rj}^d)^T, U_{d+1}^r, \ldots, U_D^r) \cdot \mathcal{C}_r\right\}$$

with $\dot{\mathcal{C}} \in \mathscr{B}_{\mathcal{C}}$, $d = 1, \ldots, D$, $i = 1, \ldots, m_d - l_d$ and $j = 1, \ldots, l_d^r$.

For $\mathcal{T}_{\mathcal{A}_r}\mathcal{M}_r^{n_1,\dots,n_D}$, we can use a basis of the same form, constructed as follows: We form $Q_d^\perp$ so that the columns of $[Q_d \quad Q_d^\perp]$ are an orthonormal basis of $\mathbb{R}^{n_d}$. Then define

$$(Q_d U_d^r)^\perp := \begin{bmatrix} Q_d U_d^{r\perp} & Q_d^\perp \end{bmatrix} \in \mathbb{R}^{n_d \times (n_d - l_d)}.$$

The columns of this matrix are a basis for the orthogonal complement of the column space of $Q_d U_d^r$. A basis $\mathscr{B}_{\mathcal{A}_r}$ of $\mathcal{T}_{\mathcal{A}_r}\mathcal{M}_r^{n_1,\dots,n_D}$ is obtained by applying (4.9) where $Q_d U_d^r$ fulfils the role of U_d^r and $(Q_d U_d^r)^\perp$ fulfils that of $U_d^\perp$.

By rearranging the order of the basis vectors and factoring out all Q_d and $Q_d^\perp$, we get a partition of this basis:

$$\mathscr{B}_{\mathcal{A}_r} = (Q_1 \otimes \cdots \otimes Q_D)(\mathscr{B}_{\mathcal{G}_r}) \cup \mathscr{B}_1^{r,\perp} \cup \cdots \cup \mathscr{B}_D^{r,\perp}, \tag{4.15}$$

where

$$\mathscr{B}_d^{r,\perp} = \left\{ \left(Q_d^\perp \otimes_d \bigotimes_{d' \neq d} Q_{d'} \right) \left((e_i^{(n_d - m_d)} (\hat{u}_{rj}^d)^T) \otimes_d \bigotimes_{d' \neq d} U_d^r \right) \mathcal{C}_r \right\}_{i,j=1}^{n_d - m_d, l_d^r}.$$

By construction of $Q_d^\perp$, the subspaces that for a fixed r are spanned by the $D+1$ bases in (4.15) are pairwise orthogonal. Therefore, collecting $\mathscr{B}_{\mathcal{A}_r}$ for all r gives a Terracini matrix of $\mathcal{A}$, which splits into $D+1$ pairwise orthogonal blocks. Up to a permutation of the columns,

$$T_{\mathcal{A}_1,\dots,\mathcal{A}_R} = \begin{bmatrix} (Q_1 \otimes \cdots \otimes Q_D) T_{\mathcal{G}_1,\dots,\mathcal{G}_R} & T_1^\perp & \dots & T_D^\perp \end{bmatrix}, \tag{4.16}$$

where the columns of each $T_d^\perp$ are $\mathscr{B}_d^{1,\perp} \cup \cdots \cup \mathscr{B}_d^{R,\perp}$. Explicitly,

$$T_d^\perp = \left(Q_d^\perp \otimes_d \bigotimes_{d' \neq d} Q_{d'} \right) \left[e_i^{(n_d - m_d)} \otimes_d \left((\hat{u}_{rj}^d)^T \otimes_d \bigotimes_{d' \neq d} U_d^r \right) \mathcal{C}_r \right]_{r,i,j=1}^{R, n_d - m_d, l_d^r}.$$

Because the blocks are pairwise orthogonal, the singular values of $T_{\mathcal{A}_1,\dots,\mathcal{A}_R}$ are the union of those of $T_{\mathcal{G}_1,\dots,\mathcal{G}_R}$ and those of each $T_d^\perp$ separately.

The factor $\left(Q_d^\perp \otimes_d \bigotimes_{d' \neq d} Q_{d'} \right)$ is orthogonal and thus can be omitted for the purpose of computing singular values. By the definition of the Kronecker product, the columns of the remaining factor can be permuted to

$$\tilde{T}_d^\perp := \left[\mathbb{1}_{n_d - m_d} \otimes_d \left((\hat{u}_{rj}^d)^T \otimes_d \bigotimes_{d' \neq d} U_d^r \right) \mathcal{C}_r \right]_{r,j=1}^{R, l_d^r}$$

hence, its singular values are just those of $T_d^{\perp}$. We will show that this is effectively a submatrix of $T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}$ so that the desired result follows from the interlacing property of singular values [GVL13, Theorem 8.1.7].

For any r and d, take all tangent vectors at $\mathcal{G}_r$ in the set $\mathscr{V}_r^d \cup \mathscr{W}_r^d$ with

$$\mathscr{V}_r^d := \left\{ \left(\left(U_d^r e_i^{(l_d^r)} (\hat{u}_{rj}^d)^T \right) \otimes_d \bigotimes_{d' \neq d} U_d^r \right) \mathcal{C}_r \right\}_{i,j=1}^{l_d^r, l_d^r} \quad \text{and}$$

$$\mathscr{W}_r^d := \left\{ \left(\left(U_d^{r\perp} e_i^{(m_d - l_d^r)} (\hat{u}_{rj}^d)^T \right) \otimes_d \bigotimes_{d' \neq d} U_d^r \right) \mathcal{C}_r \right\}_{i,j=1}^{m_d - l_d^r, l_d^r} .$$

In the proofs of Propositions 4.6 and 4.8, we showed that all vectors in the same form as $\mathscr{V}_r^d$ are tangent to $\mathcal{M}_r^{m_1,\ldots,m_D}$. $\mathscr{W}_r^d$ is just a subset of $\mathscr{B}_{\mathcal{G}_r}$, the basis we used for $\mathcal{G}_r$. By construction of $U_d^{r\perp}$, the spaces spanned by $\mathscr{V}_r^d$ and $\mathscr{W}_r^d$ are orthogonal. The inner products between the elements of $\mathscr{V}_r^d$ (respectively, $\mathscr{W}_r^d$) are of the same form as (4.10). Hence, they are also zero. By collecting $\mathscr{V}_r^d \cup \mathscr{W}_r^d$ for all r, we get a subset of the columns of $T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}$:

$$\tilde{T}_d^{\text{part}} := \left[[U_d^r \quad U_d^{r\perp}] \otimes_d \left((\hat{u}_{rj}^d)^T \otimes_d \bigotimes_{d' \neq d} U_d^r \right) \mathcal{C}_r \right]_{r,j=1}^{R, l_r^d} ,$$

which has the same singular values as $\tilde{T}_d^{\perp}$ by Lemma 4.16. Hence, the singular values of $T_d^{\perp}$ are interlaced between those of $T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}$. By reminding ourselves that (4.16) is a decomposition into pairwise orthogonal blocks, we can see that $T_{\mathcal{A}_1,\ldots,\mathcal{A}_R}$ and $T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}$ must have the same extreme singular values. □

4.5.2 Improved algorithm for computing the condition number

One computational advantage of Theorem 4.14 is that for any SBTD that is computed using the compress-decompose-expand strategy, the condition number can be computed at a low extra cost right after the decompose phase. That is, it is not necessary to compute the expanded decomposition in order to know its condition number. Furthermore, if a subspace-constrained SBTD $\mathcal{A}_1 + \cdots + \mathcal{A}_R$ is given, it can be compressed prior to computing its condition number. This gives Algorithm 1.

This algorithm was applied to the numerical example mentioned in the introduction. Its computational complexity compares to that of the naive approach as follows.

Algorithm 1 Computation of $\kappa^{\mathrm{SBTD}}(\mathcal{A}_1,\ldots,\mathcal{A}_R)$ with $\mathcal{A}_r = (U_1^r,\ldots,U_D^r)\cdot\mathcal{C}_r$.

```
**for** $d = 1,\ldots,D$ **do**
    Compute a QR decomposition $Q_d R_d = [U_d^1,\ldots,U_d^r]$.
**end for**
**for** $r = 1,\ldots,R$ **do**
    $\mathcal{G}_r \leftarrow (Q_1^T,\ldots,Q_D^T)\cdot\mathcal{A}_r$
**end for**
Compute $\kappa^{\mathrm{SBTD}}(\mathcal{G}_1,\ldots,\mathcal{G}_R)$ using the algorithm from Section 4.4.1.
```

Proposition 4.18. *Let $\mathcal{A} = \mathcal{A}_1 + \cdots + \mathcal{A}_R \in \mathbb{R}^{n\times\cdots\times n}$ be a subspace-constrained SBTD with core structures $\mathcal{M}_r$, where each $\mathcal{M}_r$ is an open submanifold of $\mathbb{R}^{l\times\cdots\times l}$. Assume that the summands $\mathcal{A}_r$ are given in HOSVD form. Assume that computing the QR and singular value decomposition of an $m\times n$-matrix with $m \geqslant n$ both take $O(mn^2)$ arithmetic operations. The number of arithmetic operations involved in applying* (4.8) *directly and applying Algorithm 1 is*

$$O(n^D R^2 l^{2D} + n^D R^2 D^2 l^2 (n-l)^2) \ \ \textit{and}\ \ O(DnR^2l^2 + R^{D+2}l^{3D} + R^{D+4}l^{D+4}D^2),$$

respectively.

Proof. First, we apply (4.9) directly to $\mathcal{A}_r = (U_1^r,\ldots,U_D^r)\cdot\mathcal{C}_r$. Computing the complement $U_d^{r\perp}$ of U_d^r is negligible. The basis vectors in (4.9) with indices i,j,d can be computed as tensors whose dth unfolding is $U_d^{r\perp}e_i(U_d e_j)^T(\mathcal{A}_r)_{(d)}$, which takes $O(ln^D)$ operations per basis vector. Computing $U_1^r\otimes\cdots\otimes U_D^r$ takes $O(n^Dl^D)$ time. This gives a time of $O(Rn^Dl^D + Rl^2(n-l)n^D)$ to construct the full Terracini matrix, whose dimensions are $n^D\times p$ where $p = R(l^D + Dl(n-l))$. Computing its singular values requires $O(n^Dp^2) = O(n^DR^2l^{2D} + n^DR^2D^2l^2(n-l)^2)$ operations [GVL13].

Next, we consider Algorithm 1. The matrices $[U_d^1,\ldots,U_d^r]$ have $m := Rl$ columns and n rows, which gives a complexity of $O(nR^2l^2)$ for each QR decomposition [GVL13]. Converting each $\mathcal{G}_r$ to HOSVD form takes $O(Dm^{D+1})$ time [VVM12]. Constructing the Terracini matrix is negligible compared to computing its singular values, as before. In this case, the Terracini matrix has dimensions $m^D\times q$ where $q = R(l^D + Dl(m-l)) = O(Rl^D + R^2Dl^2)$. The computation of the singular values requires $O(m^Dq^2) = O(R^Dl^Dq^2) = O(R^{D+2}l^{3D} + R^{D+4}l^{D+4}D^2)$ operations. □

If n is significantly larger than Rl in this proposition, the complexity is approximated by $O(n^{D+2}R^2D^2l^2)$ and $O(Rln^D)$, respectively, which shows the superiority of Algorithm 1. On the other hand, if $n \leqslant Rl$, the algorithm does not compress the decomposition and merely adds overhead.

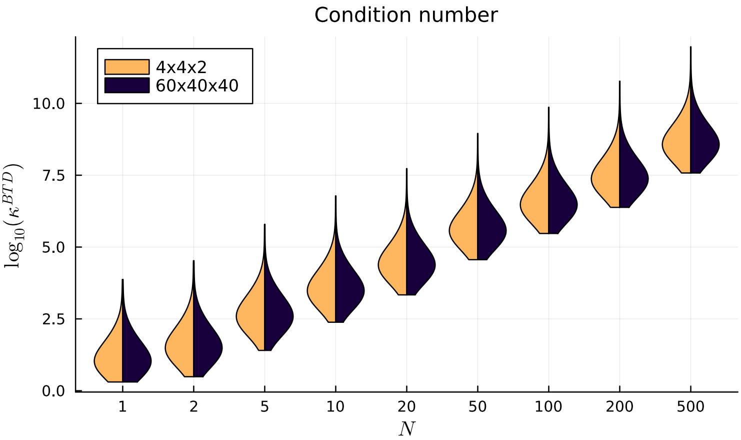


Figure 4.1: Condition number of the BTD of $\mathcal{G}_N \in \mathbb{R}^{4\times4\times2}$ and that of $\mathcal{A}_N \in \mathbb{R}^{60\times40\times40}$ from the experiments in Section 4.6

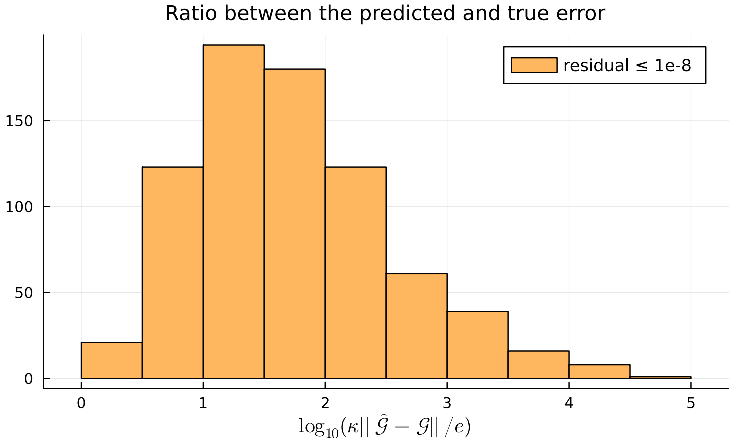


Figure 4.2: Ratio between the estimated forward error based on (4.3) and the true forward error for $\mathcal{G}_N$ in the experiments in Section 4.6. Only cases with a residual $\left\|\hat{\mathcal{G}} - \mathcal{G}\right\| \leqslant 10^{-8}$ were considered.

4.6 Numerical experiments

We present a few numerical experiments illustrating the main result, Theorem 4.14, with a sequence of ill-conditioned block term decompositions. All numerical computations were performed on an Intel Xeon CPU E5-2697

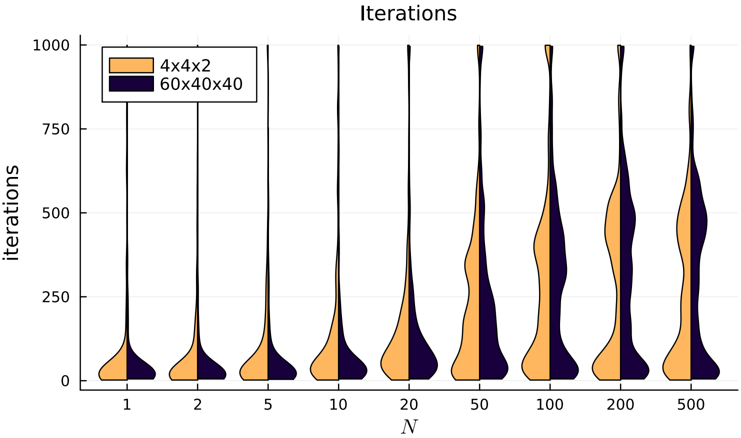


Figure 4.3: Number of iterations of `btd_nls` applied to $\mathcal{G}_N \in \mathbb{R}^{4\times4\times2}$ and $\mathcal{A}_N \in \mathbb{R}^{60\times40\times40}$ from the experiments in Section 4.6.

v3 running on 16 out of 28 physical cores and 126GB memory. The tensor decompositions were computed in MATLAB R2018b with Tensorlab 3.0 [VDDL17] and the other computations were performed in Julia v1.6[4] [Bez+17].

De Silva and Lim [SL08] give an explicit parametrisation of a general curve of rank-2 tensors $\mathcal{X}_N$ that converges to a rank-3 tensor as $N \to \infty$. In such cases, the condition number diverges to infinity [BV18b]. Given the vectors x_d and y_d for $d = 1, 2, 3$, the sequence $\{\mathcal{X}_N\}_{N=1}^{\infty}$ is given by

$$N \bigotimes_{d=1}^{3} \left(x_d + \frac{y_d}{N} \right) - N \bigotimes_{d=1}^{3} x_d = y_1 \otimes x_2 \otimes x_3 + x_1 \otimes y_2 \otimes x_3 + x_1 \otimes x_2 \otimes y_3 + \mathcal{O}\left(\frac{1}{N}\right).$$

This example can easily be generalised to block term decompositions. Take any third-order core tensor $\mathcal{C}$ of full multilinear rank and any two sets of full-rank matrices $\{A_d\}_{d=1}^{3}$ and $\{B_d\}_{d=1}^{3}$. Then set

$$\mathcal{G}_N := \left(N \bigotimes_{d=1}^{3} \left(B_d + \frac{1}{N} A_d \right) - N \bigotimes_{d=1}^{3} B_d \right) \mathcal{C}. \tag{4.17}$$

Both blocks have the same multilinear rank assuming B_d and $B_d + \frac{1}{N} A_d$ have full rank. Similarly to $\mathcal{X}_N$, we can see that $\mathcal{G}_N$ equals a three-term BTD independent of N, plus $o(N^{-1})$ terms. Its condition number diverges as $N \to \infty$ by a special case of [BV18b, Theorem 1.4].

[4] The code for all experiments is available in a public repository at `https://gitlab.kuleuven.be/u0072863/paper.experiments_dbv2022_sbtd_compression`

We generated tensors of this model where the core tensor $\mathcal{C} \in \mathbb{R}^{2\times 2\times 1}$ and the matrices $A_1, A_2 \in \mathbb{R}^{4\times 2}, A_3 \in \mathbb{R}^{2\times 1}$ all have standard normally distributed entries and B_d is the Q-factor of the QR decomposition of a matrix with standard normal entries. For several values of N, we generated 2000 tensors of model (4.17). For each of these we generated an expanded representation $\mathcal{A}_N = (Q_1, Q_2, Q_3) \cdot \mathcal{G}_N$ for some Q_1, Q_2, Q_3 with orthonormal columns. The dimensions of the tensors are $\mathcal{G}_N \in \mathbb{R}^{4\times 4\times 2}$ and $\mathcal{A}_N \in \mathbb{R}^{60\times 40\times 40}$.

We used Tensorlab's Gauss–Newton method `ll1_nls` [VDDL17] to compute a two-term $(2,2,1)$-BTD of both the (sequences of) tensors $\mathcal{A}_N$ and $\mathcal{G}_N$ independently. Since $\mathcal{A}_N$ has a subspace-constrained BTD with core tensor $\mathcal{G}_N$, by Theorem 4.14 their condition numbers are the same. Some built-in optimisations were disabled, namely automatic Tucker compression and the use of the iterative solver to solve the linear system to compute the quasi–Newton update direction. This is to ensure the same algorithm is used for both tensors. Since `ll1_nls` stops when the backward error reaches a certain threshold, this generates exact decompositions of nearby tensors, which allows us to compare the forward and backward error.

A violin plot of the condition number of both BTDs is shown in Figure 4.1. The condition number does indeed increase with the parameter N. Moreover, the distribution of the condition number of the BTD of $\mathcal{G}_N$ is the same as that of $\mathcal{A}_N$. We did find that the ratio between the *computed* condition numbers $\hat{\kappa}$ deviated slightly from one in the more ill-conditioned cases. The most extreme case was $\hat{\kappa}(\mathcal{A}_1, \ldots, \mathcal{A}_R) \approx (1 - 2\cdot 10^{-5})\hat{\kappa}(\mathcal{G}_1, \ldots, \mathcal{G}_R)$ where $\kappa > 10^{12}$. We attribute this to numerical roundoff. These results thus provide a numerical verification of Theorem 4.14.

A major application of the condition number is to estimate the forward error. For a true decomposition $\mathcal{G} = \sum_{r=1}^{R} \mathcal{G}_r$ and a computed decomposition $\hat{\mathcal{G}} = \sum_{r=1}^{R} \hat{\mathcal{G}}_r$, the forward error is measured as

$$e = \min_{\pi \in \mathscr{S}_R} \sqrt{\sum_{r=1}^{r} \left\| \mathcal{G}_r - \hat{\mathcal{G}}_{\pi(r)} \right\|^2},$$

where $\mathscr{S}_R$ is the symmetric group of R elements. By (4.3) we can estimate that $e \lesssim \kappa^{\mathrm{BTD}} \left\| \mathcal{G} - \hat{\mathcal{G}} \right\|$ as long as the residual $\left\| \mathcal{G} - \hat{\mathcal{G}} \right\|$ is not too large. Figure 4.2 shows that this bound tends to hold when the residual is at most 10^{-8}.

Finally, the condition number is related to convergence rates of iterative algorithms to compute the decomposition. The estimates of the convergence rate of the Riemannian Gauss-Newton method from [BV18a] would be the same for $\mathcal{G}_N$ and $\mathcal{A}_N$. The influence of the condition number (4.2) on flat optimisation methods like `ll1_nls` has not yet been studied. Nonetheless, Figure 4.3 seems to indicate that the required number of iterations is not affected by compression. It also shows that convergence gets slower as the condition number increases.

The cost per iteration is expected to be a function of only the dimensions of the tensor and the block terms, as only direct linear algebra routines are used to compute the iteration steps [SVBDL13]. By using the compressed tensor $\mathcal{G}_N$ instead of $\mathcal{A}_N$, the geometric mean of the speedup per iteration was 9.5. In [BA98], speedup factors of up to 40 were observed for the ALS algorithm applied to tensors used in chemometrics. Note that this is the speedup of computing the decomposition. For the sugar data set of [BA98], the computation of the *condition number*, as mentioned in the introduction, was sped up by a factor of 15 000 by first Tucker compressing the tensor from size $265 \times 371 \times 7$ to $3 \times 3 \times 3$.

4.7 Conclusion

In this chapter, we introduced the structured block term decomposition, a generalisation of the tensor rank and block term decomposition. We studied the geometry of the associated manifold in Section 4.3 and provided an orthonormal basis for the tangent space in Section 4.4. This gives an algorithm to compute the condition number, as well as some estimates of it.

In Section 4.5, we generalised Tucker compression to the SBTD. If a tensor $\mathcal{A}$ can be compressed as $\mathcal{A} = (Q_1, \ldots, Q_D) \cdot \mathcal{G}$ and the core $\mathcal{G}$ has an SBTD, this SBTD can be converted to an SBTD of $\mathcal{A}$. Our main result states that the condition numbers of these two SBTDs are identical. This is unlike some other problems, where the condition number of the structured problem can be significantly lower [ANT19].

In Section 4.5.2, we exploited our theorem to provide an algorithm to compute the condition number of the SBTD that can speed up the state-of-the-art method by over four orders of magnitude in certain practical cases. The invariance of the condition number under Tucker compression suggests that the local convergence rate of certain optimisation methods is unaffected by compression, even though the search space is reduced. In particular, we observed that Tensorlab's [VDDL17] `ll1_nls` required about the same number of iterations for computing the compressed and uncompressed decomposition.

Chapter 5

Condition of symmetric tensor decompositions

Sections 5.1-5.3 consist of the journal article [DBV23b].

N. Dewaele, P. Breiding and N. Vannieuwenhoven. “Three decompositions of symmetric tensors have similar condition numbers”. In: *Linear Algebra and its Applications* 664 (May 2023), pp. 253–263.

The preprint version of this article contained an appendix, which was turned into Section 5.4. This section is a straightforward generalisation of Section 5.2, but it is kept separate because it introduces complicated notation. The results presented in this chapter were obtained in collaboration with all authors of the article.

Abstract

We relate the condition numbers of computing three decompositions of symmetric tensors: the polyadic decomposition, the Waring decomposition, and a Tucker-compressed Waring decomposition. Based on this relation we can speed up the computation of these condition numbers by orders of magnitude through Tucker compression.

5.1 Introduction

Many problems in machine learning and signal processing involve computing a decomposition of a *symmetric tensor* [Ana+14]; an order-D tensor $\mathcal{A} = [a_{i_1,\ldots,i_D}]_{i_1,\ldots,i_D=1}^{n} \in \mathbb{R}^{n\times\cdots\times n}$ is symmetric if its entries $a_{i_1,\ldots,i_D}$ are invariant under all permutations of the indices $i_1,\ldots,i_D$. We establish a close connection between the numerical sensitivity of the following three increasingly structured decomposition problems associated with a symmetric tensor $\mathcal{A}$:

1. A *polyadic decomposition (PD)* of $\mathcal{A}$ expresses $\mathcal{A}$ as a sum of (not necessarily symmetric) tensors of rank 1. In other words, $\mathcal{A} = \sum_{r=1}^{R} \mathcal{A}_r$ where $R \in \mathbb{N}$, $\mathcal{A}_r = \alpha_r \, a_r^{(1)} \otimes \cdots \otimes a_r^{(D)}$, $\alpha_r \in \mathbb{R} \setminus \{0\}$ and each $a_r^{(i)}$ is a point on the sphere $\mathbb{S}^{n-1} = \{a \in \mathbb{R}^n \mid \|a\|_2 = 1\}$.

2. A *Waring decomposition (WD)* is a special case of the PD where all summands are symmetric. That is, for $r = 1,\ldots,R$, we have that $\mathcal{A}_r = \alpha_r \, a_r^{\otimes D}$ where $\alpha_r \in \mathbb{R} \setminus \{0\}$, $a_r \in \mathbb{S}^{n-1}$, and $a_r^{\otimes D}$ is the tensor product of D copies of a_r.

3. A *Q-compressed Waring decomposition (Q-WD)* is defined as follows. A symmetric tensor $\mathcal{A}$ can be represented in a minimal subspace by a symmetric Tucker decomposition [Tuc66; DLDMV00a], i.e., $\mathcal{A} = (Q,\ldots,Q) \cdot \mathcal{G}$ where $Q \in \mathbb{R}^{n\times m}$ has orthonormal columns and $\mathcal{G} \in \mathbb{R}^{m\times\cdots\times m}$ is symmetric with $m < n$. We write this as $\mathcal{A} = Q^{\otimes D}\mathcal{G}$. If $\mathcal{G}$ has a WD $\mathcal{G} = \sum_{r=1}^{R} \mathcal{G}_r$, then it can be converted to a WD $\mathcal{A} = \sum_{r=1}^{R} Q^{\otimes D}\mathcal{G}_r$. We call a WD of this form a *Q-WD*.

One is often interested in a *minimal* PD and WD of a given tensor $\mathcal{A}$, i.e., a decomposition of $\mathcal{A}$ consisting of the smallest possible number of summands R. This number is known as the *rank* of $\mathcal{A}$ in the case of the PD and *symmetric rank* for the WD. A well-known conjecture attributed to Comon states that the rank and symmetric rank are equal for most symmetric tensors [Com+08]. This conjecture holds generically for small ranks [Fri16; COV17]. Unless stated otherwise, we do not assume that any of the decompositions studied here attains the rank.

For the above three types of decompositions, the summands are points on a smooth manifold $\mathcal{M} \subset \mathbb{R}^{n\times\cdots\times n}$, so they are *join decompositions* [BV18b]. For the PD, the summands lie on the *Segre manifold* $\mathcal{S}_{n,D}$, for the WD they lie on the *Veronese manifold* $\mathcal{V}_{n,D}$ [Lan12], and for the Q-WD they lie on the manifold $\mathcal{W}_{Q,D} = Q^{\otimes D}(\mathcal{V}_{m,D})$. In the remainder, we drop the subscripts on the manifolds if they are clear from the context.

We study the sensitivity of the summands in these three decompositions with respect to perturbations of $\mathcal{A}$. Consider a decomposition of $\mathcal{A}$ with summands $\mathfrak{a} = (\mathcal{A}_1, \dots, \mathcal{A}_R) \in \mathcal{M}^{\times R}$, where $\mathcal{M}$ is one of the three manifolds described above and $\mathcal{M}^{\times R}$ is the product of R copies of $\mathcal{M}$. Under mild conditions [BV18b], $\mathfrak{a}$ is an *isolated* decomposition of $\mathcal{A}$ and the addition map $\Sigma : \mathcal{M}^{\times R} \mapsto \mathbb{R}^{n\times\cdots\times n}$, $(\mathcal{A}_1, \dots, \mathcal{A}_R) \mapsto \mathcal{A}_1 + \cdots + \mathcal{A}_R$ admits a local inverse $\Sigma_{\mathfrak{a}}^{-1}$. In this case, the sensitivity of the decomposition with respect to $\mathcal{A}$ can be measured by the condition number [Ric66]:

$$\kappa_{\mathcal{M}}(\mathcal{A}_1, \dots, \mathcal{A}_R) := \lim_{\delta\to 0} \sup_{\substack{\widetilde{\mathcal{A}}\in\Sigma(\mathcal{M}^{\times R}),\\ \|\mathcal{A}-\widetilde{\mathcal{A}}\|\leqslant\delta}} \frac{\|\Sigma_{\mathfrak{a}}^{-1}(\mathcal{A}) - \Sigma_{\mathfrak{a}}^{-1}(\widetilde{\mathcal{A}})\|}{\|\mathcal{A} - \widetilde{\mathcal{A}}\|}, \qquad (5.1)$$

where $\|\cdot\|$ is the Euclidean or Frobenius norm. The condition number $\kappa_{\mathcal{M}}(\mathfrak{a}) = \infty$ if and only if the kernel of the derivative of Σ is nontrivial. The latter can be caused by two situations. Either $\mathfrak{a}$ is not isolated or the polynomial system defined by $\Sigma(\mathfrak{a}) - \mathcal{A} = 0$ is singular. In accordance with this, we say that an isolated decomposition $\mathfrak{a}$ is singular if its condition number is infinite and nonsingular otherwise.

There are many known spaces of symmetric tensors in which, for a generic tensor $\mathcal{A}$, the minimal PD and WD of $\mathcal{A}$ are unique (up to a permutation of the summands) and nonsingular [Lan12]. Thus, these decompositions have a finite condition number. Moreover, there exist non-minimal WDs with a finite condition number where the number of summands exceeds the rank. The earliest example that we know of was given by Angelini and Chiantini [AC21]. They generated a WD $\mathcal{A} = \sum_{r=1}^{18} \mathcal{A}_r$ where $\mathcal{A} \in \left(\mathbb{R}^3\right)^{\otimes 9}$ has rank 17 over $\mathbb{C}$. For this example, we calculated $\kappa_{\mathcal{S}}(\mathcal{A}_1, \dots, \mathcal{A}_{18})$ and $\kappa_{\mathcal{V}}(\mathcal{A}_1, \dots, \mathcal{A}_{18})$ with the algorithm from [BV18b]. Both condition numbers are equal to about 50610.

Suppose $\mathcal{A}$ has a Q-WD $\mathfrak{a} = (\mathcal{A}_1, \dots, \mathcal{A}_R)$. It can also be regarded as a WD or PD by ignoring symmetry or subspace constraints. We investigate the relationship between the condition numbers of these three problems at $\mathfrak{a}$. The difference between $\kappa_{\mathcal{W}}, \kappa_{\mathcal{V}}$, and $\kappa_{\mathcal{S}}$ is the domain of the supremum in (5.1), i.e., the set of considered perturbations. Since $\mathcal{W} \subseteq \mathcal{V} \subseteq \mathcal{S}$, it follows from (5.1) that $\kappa_{\mathcal{W}}(\mathfrak{a}) \leqslant \kappa_{\mathcal{V}}(\mathfrak{a}) \leqslant \kappa_{\mathcal{S}}(\mathfrak{a})$. In general, restricting the domain of a map can drastically reduce the condition number. For instance, consider the problem of evaluating the matrix logarithm. In this case, the condition number for perturbations restricted to the symplectic group can be much smaller than the one for unconstrained perturbations [ANT19]. This motivates us to study if a similar relation exists for $\kappa_{\mathcal{W}}, \kappa_{\mathcal{V}}$, and $\kappa_{\mathcal{S}}$. Similarly to Chapter 4, we show the following:

Theorem 5.1. *Let $\mathcal{G} = \mathcal{G}_1 + \cdots + \mathcal{G}_R \in \mathbb{R}^{m\times\cdots\times m}$ be a WD of an order-D tensor.*

1. *Take $Q \in \mathbb{R}^{n\times m}$ with orthonormal columns and set $\mathcal{A}_r = Q^{\otimes D}\mathcal{G}_r$, for $r = 1,\ldots,R$. Then*

$$\kappa_{\mathcal{W}_{Q,D}}(\mathcal{A}_1,\ldots,\mathcal{A}_R) \leqslant \kappa_{\mathcal{V}_{n,D}}(\mathcal{A}_1,\ldots,\mathcal{A}_R) \leqslant \sqrt{D}\kappa_{\mathcal{V}_{m,D}}(\mathcal{G}_1,\ldots,\mathcal{G}_R),$$

   where the right-hand side is equal to $\sqrt{D}\kappa_{\mathcal{W}_{Q,D}}(\mathcal{A}_1,\ldots,\mathcal{A}_R)$.

2. *Let $U \in \mathbb{R}^{\ell\times m}$ have orthonormal columns and $\mathcal{B}_r := U^{\otimes D}\mathcal{G}_r$ for all r. If $\min(\ell,n) > m$, then $\kappa_{\mathcal{V}_{n,D}}(\mathcal{A}_1,\ldots,\mathcal{A}_R) = \kappa_{\mathcal{V}_{\ell,D}}(\mathcal{B}_1,\ldots,\mathcal{B}_R)$; i.e., the condition number is invariant under non-minimal symmetric Tucker compressions.*

Numerical evidence indicates a stronger connection, which can be proved in the rank-2 case:

Conjecture 5.2. *If $\mathcal{A} = \sum_{r=1}^{R}\mathcal{A}_r$ is a WD of an order-D symmetric tensor $\mathcal{A} \in \mathbb{R}^{n\times\cdots\times n}$ with $D \geqslant 3$, then $\kappa_{\mathcal{V}}(\mathcal{A}_1,\ldots,\mathcal{A}_R) = \kappa_{\mathcal{S}}(\mathcal{A}_1,\ldots,\mathcal{A}_R)$.*

Proposition 5.3. *Conjecture 5.2 holds for $R \leqslant 2$.*

In conjunction with Theorem 4.14, Conjecture 5.2 would imply that $\kappa_{\mathcal{W}}(\mathfrak{a}) = \kappa_{\mathcal{V}}(\mathfrak{a}) = \kappa_{\mathcal{S}}(\mathfrak{a})$ for any Q-WD $\mathfrak{a}$, which is sharper than Theorem 5.1. This entails that the supremum in (5.1) applied to the Segre manifold (i.e., $\mathcal{M} = \mathcal{S}$) can be attained locally with a perturbation $\widetilde{\mathcal{A}} \in \Sigma(\mathcal{W}^{\times R})$. Another implication of Conjecture 5.2 would be that a WD of a tensor $\mathcal{A}$ is isolated and nonsingular if and only if it is also isolated and nonsingular as a PD without symmetry conditions.

A practical consequence of Theorem 5.1 relates to the following procedure from [BV18b] to compute the condition number. Let $\mathcal{M}$ be either $\mathcal{S}$ or $\mathcal{V}$. Let the matrix $T_{\mathcal{A}_r}^{\mathcal{M}}$ contain as columns an orthonormal basis of $\mathcal{T}_{\mathcal{A}_r}\mathcal{M}$ for $r = 1,\ldots,R$, where $\mathcal{T}_{\mathcal{A}_r}\mathcal{M}$ is the tangent space to $\mathcal{M}$ at $\mathcal{A}_r$. Then the condition number is characterized by the *Terracini matrix* $T_{\mathcal{A}_1,\ldots,\mathcal{A}_R}^{\mathcal{M}}$:

$$T_{\mathcal{A}_1,\ldots,\mathcal{A}_R}^{\mathcal{M}} := \begin{bmatrix} T_{\mathcal{A}_1}^{\mathcal{M}} & \cdots & T_{\mathcal{A}_R}^{\mathcal{M}} \end{bmatrix} \quad \text{and} \quad \kappa_{\mathcal{M}}(\mathcal{A}_1,\ldots,\mathcal{A}_R) = \sigma_{\min}(T_{\mathcal{A}_1,\ldots,\mathcal{A}_R}^{\mathcal{M}})^{-1}, \tag{5.2}$$

where $\sigma_{\min}(A)$ is the smallest singular value of A. Consider a WD $\mathcal{A} = \sum_{r=1}^{R}\mathcal{A}_r$ with $\mathcal{A}_r = \alpha_r a_r^{\otimes D}$ for some $\alpha_r \in \mathbb{R}\setminus\{0\}$ and $a_r \in \mathbb{S}^{n-1}$. For this decomposition the Terracini matrices for the PD and the WD, respectively, are given as follows: for any two matrices X and A, let $X \otimes_d A^{\otimes D-1} := A^{\otimes d-1} \otimes X \otimes A^{\otimes D-d}$. The tangent space to the sphere at a_r is the orthogonal complement $a_r^{\perp}$ of a_r. Let

$U(a_r) \in \mathbb{R}^{n\times(n-1)}$ be a matrix whose columns form an orthonormal basis of $a_r^{\perp}$. Then the Terracini matrices are

$$T^{\mathcal{S}}_{\mathscr{A}_1,\ldots,\mathscr{A}_R} = \left[a_r^{\otimes D} \quad \left[U(a_r) \otimes_d a_r^{\otimes D-1}\right]_{d=1}^{D}\right]_{r=1}^{R}, \quad \text{and} \tag{5.3}$$

$$T^{\mathcal{V}}_{\mathscr{A}_1,\ldots,\mathscr{A}_R} = \left[a_r^{\otimes D} \quad \frac{1}{\sqrt{D}}\sum_{d=1}^{D} U(a_r) \otimes_d a_r^{\otimes D-1}\right]_{r=1}^{R}. \tag{5.4}$$

A major implication of Theorem 5.1 is that we can speed up the computation of $\kappa_{\mathcal{V}_{n,D}}(\mathscr{A}_1,\ldots,\mathscr{A}_R)$. Assuming $n > R$ and $\mathscr{A}_r = \alpha_r a_r^{\otimes D}$ with $\alpha_r \neq 0$ and $a_r \in \mathbb{S}^{n-1}$, the following computes $\kappa_{\mathcal{V}_{n,D}}(\mathscr{A}_1,\ldots,\mathscr{A}_R)$:

1. Compute a thin singular value decomposition $[a_r]_{r=1}^R = Q\Sigma V^T$ and set $[g_r]_{r=1}^R := \Sigma V^T \in \mathbb{R}^{m\times R}$.
2. Construct $b_r = [g_r^T \quad 0]^T \in \mathbb{R}^{\ell}$ where $\ell = m+1$ and set $\mathscr{B}_r = \alpha_r b_r^{\otimes D}$ for each r.
3. Construct $T^{\mathcal{V}}_{\mathscr{B}_1,\ldots,\mathscr{B}_R}$ as in (5.3) and compute $\kappa_{\mathcal{V}}(\mathscr{B}_1,\ldots,\mathscr{B}_R)$ by applying (5.2).

Steps 1-2 give one possible choice of Q and $U = [I \quad 0]^T$ and $\mathscr{G}_r = \alpha_r g_r^{\otimes D}$ that satisfy Theorem 5.1. A Julia [Bez+17] implementation of this method is provided along with the arXiv version of this manuscript. Since $T^{\mathcal{V}}_{\mathscr{B}_1,\ldots,\mathscr{B}_R} \in \mathbb{R}^{\ell^D \times R\ell}$ and $\ell \leqslant R+1$, step 3 can be performed in $\mathcal{O}(R^{D+4})$ operations, adding to the $\mathcal{O}(nR^2)$ cost of step 1. Applying (5.2) to $T^{\mathcal{V}}_{\mathscr{A}_1,\ldots,\mathscr{A}_R} \in \mathbb{R}^{n^D\times Rn}$ would involve $\mathcal{O}(n^{D+2}R^2)$ operations. The algorithm can reach significant speedups if $n \gg R$. For instance, we applied the Julia code to a WD with $(n, D, R, \ell) = (100, 3, 10, 11)$ on an Intel Xeon CPU E5-2697 v3 running on 8 cores and 126GB memory. The computation times were 115.6 and 0.0092 seconds for the original and improved algorithm, respectively.

5.2 Condition number of a Q-WD

In this section, we prove Theorem 5.1 based on the following insight: $\Sigma(\mathcal{V}^{\times R})$ is locally a manifold whose tangent space is decomposed as $\mathbb{T} \oplus \mathbb{T}^{\perp}$ where $\mathbb{T}$ is the tangent space to $\Sigma(\mathcal{W}^{\times R})$ and $\mathbb{T}^{\perp}$ is its orthogonal complement. As long as $n > m$, the effect of the worst perturbation to $\mathscr{A}$ inside $\mathbb{T}^{\perp}$ is independent of n and can be bounded as in the first statement. From this, the second statement follows as well.

Proof of Theorem 5.1. The first inequality follows from the inclusion $\mathcal{W}_{Q,D} \subseteq \mathcal{V}_{n,D}$. The last follows from the fact that $Q^{\otimes D}$ is an isometry between $\mathcal{V}_{m,D}$ and $\mathcal{W}_{Q,D}$. It remains to show the middle inequality. If $m = n$, Q is an orthogonal change of basis, which preserves the condition number. Thus, we assume $n > m$.

For each r, let $\mathcal{G}_r = \alpha_r g_r^{\otimes D}$ with $\alpha_r \in \mathbb{R} \setminus \{0\}$ and $g_r \in \mathbb{S}^{m-1}$, let $a_r = Qg_r$ and define U_r so that the matrix $[g_r \quad U_r] \in \mathbb{R}^{m \times m}$ is orthogonal. Construct $T_{\mathcal{G}_r}^{\mathcal{V}}$ by applying (5.3) to $\mathcal{G}_r$. Complete Q to an orthonormal basis $[Q \quad Q_\perp]$ of $\mathbb{R}^n$. The columns of $U(a_r) := [QU_r \quad Q_\perp]$ form an orthonormal basis of $\mathcal{T}_{a_r}\mathbb{S}^{n-1}$. Substituting this into (5.3) gives $T_{\mathcal{A}_1,\ldots,\mathcal{A}_R}^{\mathcal{V}_{n,D}} = \left[T_r \quad T_r^\perp\right]_{r=1}^R$ with

$$T_r = \left[a_r^{\otimes D} \quad \tfrac{1}{\sqrt{D}}\left(\textstyle\sum_{d=1}^D QU_r \otimes_d a_r^{\otimes D-1}\right)\right] \text{ and } T_r^\perp = \frac{1}{\sqrt{D}}\sum_{d=1}^D Q_\perp \otimes_d a_r^{\otimes D-1}.$$

Since $a_r = Qg_r$, we have $T_r = Q^{\otimes D} T_{\mathcal{G}_r}^{\mathcal{V}_{m,D}}$. Thus, up to a column permutation, $T_{\mathcal{A}_1,\ldots,\mathcal{A}_R}^{\mathcal{V}_{n,D}}$ is the horizontal concatenation of $Q^{\otimes D} T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}^{\mathcal{V}_{m,D}}$ and $T^\perp := [\, T_1^\perp \; \ldots \; T_R^\perp \,]$. The column spaces of $Q^{\otimes D}$ and $T^\perp$ are orthogonal, so that the singular values of $T_{\mathcal{A}_1,\ldots,\mathcal{A}_R}^{\mathcal{V}_{n,D}}$ are the union of those of $Q^{\otimes D} T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}^{\mathcal{V}_{m,D}}$ and $T^\perp$ separately. Since Q has orthonormal columns, $Q^{\otimes D} T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}^{\mathcal{V}_{m,D}}$ has the same singular values as $T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}^{\mathcal{V}_{m,D}}$, so it suffices to show $\sigma_{\min}(T^\perp) \geqslant \sigma_{\min}(T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}^{\mathcal{V}_{m,D}})/\sqrt{D}$.

To do this, we compute $(T^\perp)^T T^\perp = [(T_{r_1}^\perp)^T (T_{r_2}^\perp)]_{r_1,r_2=1}^R$, where the block at (r_1, r_2) is

$$\begin{aligned} \frac{1}{D}\left(\sum_{d=1}^D Q_\perp \otimes_d a_{r_1}^{\otimes D-1}\right)^T \left(\sum_{d=1}^D Q_\perp \otimes_d a_{r_2}^{\otimes D-1}\right) &= \langle a_{r_1}, a_{r_2}\rangle^{D-1} I_{n-m} \\ &= \langle g_{r_1}, g_{r_2}\rangle^{D-1} I_{n-m}. \quad (5.5) \end{aligned}$$

Consider two modifications of $T^\perp$ that preserve the singular values. First, let $\hat{T}^\perp := [I_{n-m} \otimes g_r^{\otimes D-1}]_{r=1}^R$, then by (5.5), we have $(\hat{T}^\perp)^T \hat{T}^\perp = (T^\perp)^T (T^\perp)$, so they have the same singular values. Second, if we define $\widetilde{T}^\perp := [[g_r \quad U_r] \otimes g_r^{\otimes D-1}]_{r=1}^R$, then $\widetilde{T}^\perp$ and $\hat{T}^\perp$ also have the same singular values, since $[g_r \quad U_r]$ and I_{n-m} are orthogonal up to multiplicities by Lemma 4.16. Hence, we can proceed with $\widetilde{T}^\perp$ instead of $T^\perp$. Similarly, we modify $T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}^{\mathcal{V}}$. Scaling up all its columns of the form $g_r^{\otimes D}$ by $\sqrt{D}$ gives

$$\begin{aligned}\widetilde{T}^{\mathcal{V}}_{\mathcal{G}_1,\ldots,\mathcal{G}_R} &:= \begin{bmatrix} \sqrt{D} g_r^{\otimes D} & \frac{1}{\sqrt{D}} \sum_{d=1}^{D} U_r \otimes_d g_r^{\otimes D-1} \end{bmatrix}_{r=1}^{R} \\ &= \begin{bmatrix} \frac{1}{\sqrt{D}} \sum_{d=1}^{D} [g_r \quad U_r] \otimes_d g_r^{\otimes D-1} \end{bmatrix}_{r=1}^{R},\end{aligned}$$

i.e., $\widetilde{T}^{\mathcal{V}}_{\mathcal{G}_1,\ldots,\mathcal{G}_R} = T^{\mathcal{V}}_{\mathcal{G}_1,\ldots,\mathcal{G}_R}\Delta$ where Δ is diagonal and $\sigma_{\min}(\Delta) = 1$. Hence, $\sigma_{\min}(\widetilde{T}^{\mathcal{V}}_{\mathcal{G}_1,\ldots,\mathcal{G}_R}) \geqslant \sigma_{\min}(T^{\mathcal{V}}_{\mathcal{G}_1,\ldots,\mathcal{G}_R})$.

To compare the singular values of $\widetilde{T}^{\perp}$ and $\widetilde{T}^{\mathcal{V}}_{\mathcal{G}_1,\ldots,\mathcal{G}_R}$, take the singular vector $v = [v_r \in \mathbb{R}^m]_{r=1}^R$ of $\widetilde{T}^{\perp}$ corresponding to the smallest singular value and compute

$$\begin{aligned}\widetilde{T}^{\mathcal{V}_{m,D}}_{\mathcal{G}_1,\ldots,\mathcal{G}_R} v &= \sum_{r=1}^{R} \left(\frac{1}{\sqrt{D}} \sum_{d=1}^{D} [g_r \quad U_r] \otimes_d g_r^{\otimes D-1} \right) v_r \\ &= \frac{1}{\sqrt{D}} \sum_{d=1}^{D} \sum_{r=1}^{R} ([g_r \quad U_r] v_r) \otimes_d g_r^{\otimes D-1}.\end{aligned}$$

Since all the summands in the outer sum have the same norm, the triangle inequality gives

$$\begin{aligned}\left\| \widetilde{T}^{\mathcal{V}_{m,D}}_{\mathcal{G}_1,\ldots,\mathcal{G}_R} v \right\| &\leqslant \sqrt{D} \left\| \sum_{r=1}^{R} ([g_r \quad U_r] v_r) \otimes g_r^{\otimes D-1} \right\| \\ &= \sqrt{D} \left\| \left[[g_r \quad U_r] \otimes g_r^{\otimes D-1} \right]_{r=1}^{R} v \right\| = \sqrt{D} \cdot \sigma_{\min}(\widetilde{T}^{\perp}).\end{aligned}$$

As $\sigma_{\min}(T^{\mathcal{V}}_{\mathcal{G}_1,\ldots,\mathcal{G}_R}) \leqslant \sigma_{\min}(\widetilde{T}^{\mathcal{V}}_{\mathcal{G}_1,\ldots,\mathcal{G}_R}) \leqslant \left\| \widetilde{T}^{\mathcal{V}}_{\mathcal{G}_1,\ldots,\mathcal{G}_R} v \right\|$, and $\sigma_{\min}(\widetilde{T}^{\perp}) = \sigma_{\min}(T^{\perp})$ this gives the desired bound.

For the second statement, recall that the singular values of $T^{\mathcal{V}_{n,D}}_{\mathcal{A}_1,\ldots,\mathcal{A}_R}$ are the union of those of $T^{\mathcal{V}_{m,D}}_{\mathcal{G}_1,\ldots,\mathcal{G}_R}$ and those of $\widetilde{T}^{\perp}$ whenever $n > m$. Observe that both of these matrices are independent of n and Q. Hence, applying the above calculation to $\mathcal{B}_1, \ldots, \mathcal{B}_R \in \mathcal{V}_{\ell,D} \subset \mathbb{R}^{\ell \times \cdots \times \ell}$ and orthogonal $U \in \mathbb{R}^{\ell \times m}$ under the assumption $\ell > m$ would reveal the same singular values. □

5.3 Equivalence between the PD and WD

Conjecture 5.2 is a stronger statement than Theorem 5.1, but it seems too challenging to show in general. We present a proof for the case where $R = 2$ and present numerical evidence for the general case.

Proof of Proposition 5.3. For $R = 1$, both condition numbers are equal to 1 by (5.2) and (5.3). For $R = 2$, the proof comprises computing the singular values of (5.2) for the PD. Let $\mathcal{A}_1 = \lambda_1 u^{\otimes D}$ and $\mathcal{A}_2 = \lambda_2 v^{\otimes D}$ with $u, v \in \mathbb{S}^{n-1}$ and $\lambda_1, \lambda_2 \neq 0$. Note that both matrices in (5.3) contain $[u^{\otimes D},\ v^{\otimes D}]$ as a subset of their columns. If u and v are collinear, this submatrix is singular, in which case $\kappa_{\mathcal{S}}(\mathcal{A}_1, \mathcal{A}_2) = \kappa_{\mathcal{V}}(\mathcal{A}_1, \mathcal{A}_2) = \infty$. In the remainder, we will assume that u and v are linearly independent. Let $U, V \in \mathbb{R}^{n\times(n-1)}$ be orthonormal bases of $\mathcal{T}_u\mathbb{S}^{n-1} = u^\perp$ and $\mathcal{T}_v\mathbb{S}^{n-1} = v^\perp$, respectively. Applying (5.3) and using as before the notation $X \otimes_d A^{\otimes D-1} := A^{\otimes d-1} \otimes X \otimes A^{\otimes D-d}$ gives

$$T^{\mathcal{S}}_{\mathcal{A}_1,\mathcal{A}_2} = \begin{bmatrix} u^{\otimes D} & \left[U \otimes_d u^{\otimes D-1}\right]_{d=1}^{D} & v^{\otimes D} & \left[V \otimes_d v^{\otimes D-1}\right]_{d=1}^{D} \end{bmatrix}.$$

Next, define the vectors $q_D^j := \begin{bmatrix} \frac{1}{\sqrt{j(j+1)}} 1_j^T & \frac{-j}{\sqrt{j(j+1)}} & 0_{D-j-1}^T \end{bmatrix}^T \in \mathbb{R}^D$ where $1_N, 0_N \in \mathbb{R}^N$ are the vectors consisting of ones and zeros, respectively. We set $Q_D := \begin{bmatrix} \frac{1}{\sqrt{D}} 1_D & q_D^1 & \cdots & q_D^{D-1} \end{bmatrix} \in \mathbb{R}^{D\times D}$. This matrix is called *Helmert's orthogonal matrix* [Hel76]; the rows of its right $D \times (D-1)$ submatrix are the vertices of a regular simplex in $\mathbb{R}^{D-1}$. We transform $T^{\mathcal{S}}_{\mathcal{A}_1,\mathcal{A}_2}$ into a matrix $\widetilde{T}^{\mathcal{S}}$ with the same singular values using an orthogonal change of basis: $\widetilde{T}^{\mathcal{S}} := T^{\mathcal{S}}_{\mathcal{A}_1,\mathcal{A}_2} \operatorname{diag}(1, I \otimes Q_D, 1, I \otimes Q_D)$. This gives

$$\widetilde{T}^{\mathcal{S}} = \begin{bmatrix} u^{\otimes D} & S_u & S^1_{u,\perp} & \cdots & S^{D-1}_{u,\perp} & v^{\otimes D} & S_v & S^1_{v,\perp} & \cdots & S^{D-1}_{v,\perp} \end{bmatrix},$$

$$\text{where } S_u = \frac{1}{\sqrt{D}} \sum_{d=1}^{D} (U \otimes_d u^{\otimes D-1}) \text{ and } \quad S^j_{u,\perp} = \sum_{i=1}^{j+1} (q_D^j)_i (U \otimes_i u^{\otimes D-1}),$$

and analogously for v. After rearranging the blocks, we get the following partition of $\widetilde{T}^{\mathcal{S}}$:

$$\widetilde{T}^{\mathcal{S}} \cong \begin{bmatrix} T^{\mathcal{V}} & T^1_\perp & \cdots & T^{D-1}_\perp \end{bmatrix} \tag{5.6}$$

$$\text{where } T^{\mathcal{V}} = \begin{bmatrix} u^{\otimes D} & S_u & v^{\otimes D} & S_v \end{bmatrix} \text{ and } T^d_\perp := \begin{bmatrix} S^d_{u,\perp} & S^d_{v,\perp} \end{bmatrix};$$

herein, we recognise (5.3).

Now, we will show that these D blocks are pairwise orthogonal, so that the singular values of $\widetilde{T}^{\mathcal{S}}$ are the union of the singular values of the blocks. To see this, we compute $(S^j_{u,\perp})^T(S^{j'}_{v,\perp})$. Let $\alpha := \langle u,\, v\rangle$. Without loss of generality, we can assume that $j < j'$. Let

$$X_i = U \otimes_i u^{\otimes D-1} \quad \text{and} \quad Y_{i'} = V \otimes_{i'} v^{\otimes D-1};$$

$$\beta_= := \alpha^{D-1} U^T V \quad \text{and} \quad \beta_{\neq} := \alpha^{D-2} U^T v u^T V.$$

Note that $X_i^T Y_{i'} = \beta_=$, if $i = i'$, and that $X_i^T Y_{i'} = \beta_{\neq}$, if $i \neq i'$. Up to the constant $(j(j+1)j'(j'+1))^{-\frac{1}{2}}$, the inner products $(S^j_{u,\perp})^T(S^{j'}_{v,\perp})$ are

$$\left(X_1 + \cdots + X_j - jX_{j+1}\right)^T \left(Y_1 + \cdots + Y_j + \cdots + Y_{j'} - j'Y_{j'+1}\right), \qquad (5.7)$$

which is a linear form in $\beta_=$ and $\beta_{\neq}$. First, we calculate the terms in (5.7) involving the case $i = i'$. There are j terms of the form $X_i^T Y_i$ with $i \leqslant j < j'$ and one of the form $-jX_{j+1}Y_{j+1}$. Adding these terms together shows that the coefficient of $\beta_=$ is zero. Second, we identify all terms in (5.7) where the coefficient of $\beta_{\neq}$ is positive. For each X_i with $i \leqslant j$, there are $j' - 1$ terms $X_i^T Y_{i'}$ with $i' \neq i$. One more term has a positive coefficient of $\beta_{\neq}$, which is $jj'X_{j+1}Y_{j'+1}$. Together, these terms add up to $j(j'-1)\beta_{\neq} + jj'\beta_{\neq}$. Third, we accumulate the negative coefficients of $\beta_{\neq}$, which involve either X_{j+1} or $Y_{j'+1}$. For X_{j+1}, there are $j' - 1$ terms $Y_{i'}$ with $j + 1 \neq i' \leqslant j'$. For $Y_{j'+1}$, there are j terms X_i with $i \leqslant j$. Hence, the terms with a negative coefficient of $\beta_{\neq}$ add up to $-j(j'-1)\beta_{\neq} - jj'\beta_{\neq}$. This means the terms involving $\beta_{\neq}$ also vanish. Therefore, all inner products $(S^j_{u,\perp})^T(S^{j'}_{v,\perp})$ vanish for $j \neq j'$.

Furthermore, the columns of $T^{\mathcal{V}}$ are symmetric tensors. The space of symmetric tensors is the linear span of the Veronese manifold $\mathcal{V} := \{\alpha z^{\otimes D} \mid \alpha \in \mathbb{R} \setminus \{0\}, z \in \mathbb{S}^{n-1}\}$. Since $\sum_{i=1}^{j+1}(q_D^j)_i = 0$, we have $(z^{\otimes D})^T S^j_{u,\perp} = \sum_{i=1}^{j+1}(z^T u)^{D-1} z^T U (q_D^j)_i = 0$, so that the columns of $S^j_{u,\perp}$ and $T^{\mathcal{V}}$ are pairwise orthogonal. We can therefore conclude that (5.6) partitions $\widetilde{T}^{\mathcal{S}}$ into pairwise orthogonal blocks.

Next, we compute all singular values of $\widetilde{T}^{\mathcal{S}}$ by computing the singular values of the blocks in (5.6) separately. Using the same notation as before, we compute the blocks of $(T^j_\perp)^T T^j_\perp$:

$$\begin{aligned}(S^j_{u,\perp})^T(S^j_{v,\perp}) &= \frac{1}{j(j+1)} \left(X_1 + \cdots + X_j - jX_{j+1}\right)^T \left(Y_1 + \cdots + Y_j - jY_{j+1}\right) \\ &= \frac{1}{j(j+1)} \left(a + b + c + d\right)\end{aligned}$$

where

$$\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} (X_1 + \cdots + X_j)^T (Y_1 + \cdots + Y_j) \\ -(X_1 + \cdots + X_j)^T j Y_{j+1} \\ -j X_{j+1}^T (Y_1 + \cdots + Y_j) \\ j^2 X_{j+1}^T Y_{j+1} \end{bmatrix} = \begin{bmatrix} j\beta_= + (j^2 - j)\beta_{\neq} \\ -j^2 \beta_{\neq} \\ -j^2 \beta_{\neq} \\ j^2 \beta_= \end{bmatrix}.$$

This gives $(S_{u,\perp}^j)^T (S_{v,\perp}^j) = \beta_= - \beta_{\neq} = \alpha^{D-1} U^T V - \alpha^{D-2} U^T v u^T V$. Hence, the Gramian of $T_\perp^j$ is $G_\perp := (T_\perp^j)^T T_\perp^j$, where

$$G_\perp = \begin{bmatrix} I_{n-1} & \alpha^{D-1} U^T V - \alpha^{D-2} U^T v u^T V \\ \alpha^{D-1} V^T U - \alpha^{D-2} V^T u v^T U & I_{n-1} \end{bmatrix},$$

which is independent of j. The Gramian of $T^{\mathcal{V}}$ is $G_S := (T^{\mathcal{V}})^T T^{\mathcal{V}}$, i.e.,

$$G_S = \begin{bmatrix} 1 & 0 & \alpha^D & \sqrt{D}\alpha^{D-1} u^T V \\ \times & I_{n-1} & \sqrt{D}\alpha^{D-1} U^T v & \alpha^{D-1} U^T V + (D-1)\alpha^{D-2} U^T v u^T V \\ \times & \times & 1 & 0 \\ \times & \times & \times & I_{n-1} \end{bmatrix},$$

where each × should be replaced by the transpose of corresponding element in the upper diagonal part.

To continue, we exploit the liberty of choosing the bases U and V for the orthogonal complements $u^\perp$ and $v^\perp$ respectively. By planar geometry, we can choose these bases such that $Ue_1 = \frac{v - \alpha u}{\|v - \alpha u\|}$, $Ve_1 = \frac{u - \alpha v}{\|u - \alpha v\|}$ and $Ue_j = Ve_j$ for all $j = 2, \ldots, n-1$. Consequently, $U^T v = \sqrt{1 - \alpha^2} e_1$, $V^T u = \sqrt{1 - \alpha^2} e_1$, and $U^T V = \operatorname{diag}(-\alpha, 1, \ldots, 1)$. Plugging these into $G_\perp$, we get

$$\begin{aligned} G_\perp &= \begin{bmatrix} I_{n-1} & \operatorname{diag}(-\alpha^D - \alpha^{D-2}(1-\alpha^2), \alpha^{D-1}, \ldots, \alpha^{D-1}) \\ \times & I_{n-1} \end{bmatrix} \\ &= I_{2(n-1)} + \begin{bmatrix} 0 & A_\perp \\ A_\perp & 0 \end{bmatrix}, \end{aligned} \tag{5.8}$$

where $A_\perp := \operatorname{diag}(-\alpha^{D-2}, \alpha^{D-1}, \ldots, \alpha^{D-1})$. Recall that the eigenvalues of $\left[\begin{smallmatrix} 0 & A^T \\ A & 0 \end{smallmatrix}\right]$ are $\pm\sigma(A)$, where σ are the singular values of A. Therefore, the eigenvalues of $G_\perp$ are $\lambda(G_\perp) = \{1 \pm \alpha^{D-1}, 1 \pm \alpha^{D-2}\}$. We only need the extreme eigenvalues, which are $1 \pm \alpha^{D-2}$ since $|\alpha|<1$. For G_S, we obtain

$$G_S = \begin{bmatrix} 1 & 0 & \alpha^D & \sqrt{D}\alpha^{D-1}\sqrt{1-\alpha^2} e_1^T \\ \times & I & \sqrt{D}\alpha^{D-1}\sqrt{1-\alpha^2} e_1 & A_S \\ \times & \times & 1 & 0 \\ \times & \times & \times & I_{n-1} \end{bmatrix}, \tag{5.9}$$

where $A_S = \operatorname{diag}(-\alpha^D + (D-1)\alpha^{D-2}(1-\alpha^2), \alpha^{D-1}, \dots, \alpha^{D-1})$. Define the two matrices

$$Z = \begin{bmatrix} \alpha^D & \sqrt{D}\alpha^{D-1}\sqrt{1-\alpha^2}e_1^T \\ \sqrt{D}\alpha^{D-1}\sqrt{1-\alpha^2}e_1 & A_S \end{bmatrix},$$

$$Z' := \begin{bmatrix} \alpha^D & \sqrt{D}\alpha^{D-1}\sqrt{1-\alpha^2} \\ \sqrt{D}\alpha^{D-1}\sqrt{1-\alpha^2} & -D\alpha^D + (D-1)\alpha^{D-2} \end{bmatrix}.$$

The eigenvalues of G_S are $1 \pm \sigma(Z)$. Due to the sparse structure of Z, its singular values are α^{D-1} and the singular values of Z'. Since Z' is symmetric, its eigenvalues and singular values coincide. We factor out α^{D-2} and compute the eigenvalues in terms of the trace τ and determinant Δ. This gives $\tau = (D-1)(1-\alpha^2)$, $\Delta = -\alpha^2$. The eigenvalues of Z' are $\lambda_1(Z') = \frac{\alpha^{D-2}}{2}\left(\tau + \sqrt{\tau^2 - 4\Delta}\right)$ and $\lambda_2(Z') = \frac{\alpha^{D-2}}{2}\left(\tau - \sqrt{\tau^2 - 4\Delta}\right)$. Finally, we compare the eigenvalues of G_S to the extreme eigenvalues of $G_\perp$. Since $\alpha^2 < 1$ and $D \geqslant 3$, we can derive a bound on the eigenvalues of Z' as follows. Straightforward calculations show that

$$4\tau \geqslant 4(1+\Delta) \quad \Rightarrow \quad \tau^2 - 4\Delta \geqslant \tau^2 - 4\tau + 4$$

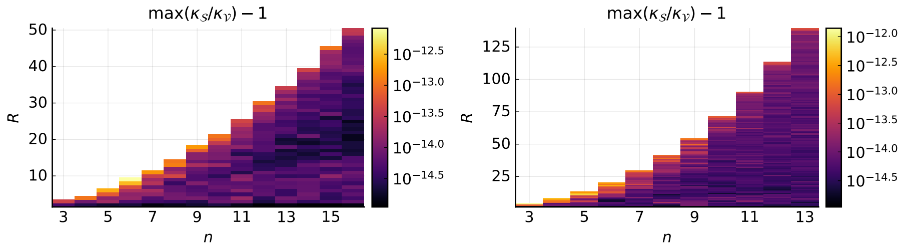
$$\Rightarrow \quad \sqrt{\tau^2 - 4\Delta} \geqslant 2 - \tau \quad \Rightarrow \quad \frac{1}{2}(\tau + \sqrt{\tau^2 - \Delta}) \geqslant 1.$$

Hence, G_S has at least one eigenvalue less than or equal to the smallest eigenvalue of $G_\perp$, namely $1 + \lambda_2(Z') \leqslant 1 + \alpha^{D-2}$ if α^{D-2} is negative, and $1 - \lambda_1(Z') \leqslant 1 - \alpha^{D-2}$ otherwise. This shows that the smallest singular value of $\widetilde{T}^{\mathcal{S}}$ in (5.6) is a singular value of $T^{\mathcal{V}}$, as required. □

We conclude this section with a note on the possibility of extending the proof to demonstrate Conjecture 5.2. The approach relies on the following two principles.

First, the set of singular values of $T^{\mathcal{S}}_{\mathcal{A}_1,\dots,\mathcal{A}_R}$ can be partitioned into the singular values of $T^{\mathcal{V}}_{\mathcal{A}_1,\dots,\mathcal{A}_R}$ and a set of additional singular values corresponding to singular vectors that are orthogonal to all symmetric tensors. These two sets of singular values are the square roots of the eigenvalues of G_S and $G_\perp$, respectively. This would still hold in the case where $R > 2$, since the first part of the proof generalises straightforwardly.

Second, the Gramian matrix can be decomposed as the sum of the identity and a sparse matrix, as in (5.8) and (5.9). For the case where $R > 2$, the Gramian matrix would be less sparse. Therefore, we believe it to be significantly more difficult to calculate the eigenvalues of (the analogues of) $G_S - I$ and $G_\perp - I$ exactly. To prove Conjecture 5.2, it seems natural to use estimates of these



(a) Ratio for third-order tensors

(b) Ratio for fourth-order tensors

Figure 5.1: Ratio between the condition numbers of the PD and WD of an $n \times \cdots \times n$ symmetric tensor of rank R. The displayed value is the maximum over 100 test cases.

eigenvalues instead. However, numerical evidence indicates that the smallest eigenvalues of $G_S - I$ and $G_\perp - I$, respectively, can get arbitrarily close to each other for ill-conditioned decompositions. Therefore, any estimates of the eigenvalues would have to be sharp.

5.3.1 Numerical experiments

We tested Conjecture 5.2 for third- and fourth-order tensors. For several small values of n, we generated 100 random symmetric rank-R decompositions $\sum_{r=1}^{R} a_r^{\otimes D}$ where $a_r \sim \mathcal{N}(0, I_n)$ using Julia v1.6 [Bez+17]. For each decomposition, we computed the two condition numbers. By dimensionality arguments, the condition number can only be finite if $Rn < \binom{n+D-1}{D}$, where the right-hand side is the dimension of the space of symmetric $n \times \cdots \times n$ tensors of order D [AH95]. We tested all values of R below this upper bound.

Figure 5.1 shows the ratio between the condition number of the PD and the WD. A priori, it can never be less than 1. In practice, numerical computations would sometimes find a ratio of $1 - 10^{-11}$ or less. This suggests that ratios exceeding 1 by less than 10^{-11} can be explained by numerical roundoff. All measurements displayed on the figure lie below this threshold.

5.4 The partially symmetric decomposition

In this section, we present a generalisation of Theorem 5.1 to the partially symmetric case. We say that a tensor $\mathcal{A}$ of order D is partially symmetric if it is invariant under the permutation of some of its indices. That is, $\mathcal{A}$ is invariant under some subgroup G of the symmetric group that is generated by pairwise swaps. When this symmetry constraint is imposed on the summands in its PD, the PD is called a *partially symmetric tensor decomposition (PSTD).* Write the size and degree of the tensors as $n = (n_1, \ldots, n_K)$ and $d = (d_1, \ldots, d_K)$, respectively. Then partially symmetric tensors of rank 1 form the image of the map

$$\Phi : \mathbb{R} \setminus \{0\} \times \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_K-1} \to \mathbb{R}^{n_1 \times \cdots \times n_1 \times \cdots \times n_K}$$

$$(\alpha, a_1, \ldots, a_K) \mapsto \alpha a_1^{\otimes d_1} \otimes \cdots \otimes a_K^{\otimes d_K}.$$

The image of Φ is known as the Segre–Veronese manifold $\mathcal{SV}_{n,d}$ [Lan12]. Analogous to the Q-WD, a $(Q_1, \ldots, Q_K)$-PSTD is a PSTD of the form $\mathcal{A} = \sum_{r=1}^{R} (Q_1^{\otimes d_1} \otimes \cdots \otimes Q_K^{\otimes d_K}) \mathcal{G}_r$ where each Q_k has orthonormal columns and $\mathcal{G}_r \in \mathcal{SV}_{m,d}$ where $m < n$ elementwise. We write $\mathcal{W} = (Q_1^{\otimes d_1} \otimes \cdots \otimes Q_K^{\otimes d_K})(\mathcal{SV}_{m,d})$.

To determine the condition number, we apply (5.2) to the PSTD. Let $x = (\alpha, a_1, \ldots, a_K)$ and $\mathcal{A} := \Phi(x)$. Then the differential of Φ is the linear map defined by

$$D\Phi(x)[\dot{\alpha}, 0, \ldots, 0] = \dot{\alpha} \bigotimes_{k=1}^{K} a_k^{\otimes d_k},$$

$$D\Phi(x)[0, \ldots, 0, \dot{a}_k, 0, \ldots, 0] = \left(\sum_{d=1}^{d_k} \dot{a}_k \otimes_d a_k^{\otimes d_k - 1} \right) \otimes_k \left(\bigotimes_{k' \neq k} a_{k'}^{\otimes d_{k'}} \right).$$

The tangent space to $\mathcal{SV}_{n,d}$ at $\mathcal{A}$ is the image of $D\Phi(x)$. To express it as the column span of a matrix, we proceed as follows. Let $U(a_k)$ be a matrix whose columns are an orthonormal basis of $\mathcal{T}_{a_k}\mathbb{S}^{n_k-1}$. Then $\mathcal{T}_{\mathcal{A}}\mathcal{SV}_{\mathbf{n},\mathbf{d}}$ is the span of

$$T_{\mathcal{A}}^{\mathcal{SV}_{n,d}} := \begin{bmatrix} \bigotimes_{k=1}^{K} a_k^{\otimes d_k} & T_{\mathcal{A}}^1 & \ldots & T_{\mathcal{A}}^K \end{bmatrix} \tag{5.10}$$

$$\text{where} \quad T_{\mathcal{A}}^k := \frac{1}{\sqrt{d_k}} \left(\sum_{d=1}^{d_k} U(a_k) \otimes_d a_k^{\otimes d_k - 1} \right) \otimes_k \left(\bigotimes_{k' \neq k} a_{k'}^{\otimes d_{k'}} \right)$$

for all $k = 1, \ldots, K$. Observe that all $K + 1$ blocks of $T_{\mathcal{A}}^{\mathcal{SV}_{n,d}}$ have orthonormal columns and are pairwise orthogonal by construction of U_k. Therefore, the condition number of any PSTD can be computed using (5.2) where the blocks in the Terracini matrix are as in (5.10). Now we can present a generalisation of Theorem 5.1.

Proposition 5.4. *Let $\mathcal{G} = \mathcal{G}_1 + \cdots + \mathcal{G}_R$ be a PSTD with summands in $\mathcal{SV}_{m,d}$. For $k = 1, \ldots, K$, take $Q_k \in \mathbb{R}^{n_k \times m_k}$ with orthonormal columns and set $\mathcal{A}_r := (Q_1^{\otimes d_1} \otimes \cdots \otimes Q_K^{\otimes d_K})\mathcal{G}_r$. Then*

$$\kappa_{\mathcal{SV}_{n,d}}(\mathcal{A}_1, \ldots, \mathcal{A}_R) \leqslant \sqrt{\max d}\, \kappa_{\mathcal{SV}_{m,d}}(\mathcal{G}_1, \ldots, \mathcal{G}_R).$$

Similarly, for $k = 1, \ldots, K$, let $\tilde{Q}_k \in \mathbb{R}^{\tilde{n}_k \times m_k}$ have orthonormal columns and $\mathcal{B}_r := (\tilde{Q}_1^{\otimes d_k} \otimes \cdots \otimes \tilde{Q}_k^{\otimes d_k})\mathcal{A}_r$ for $r = 1, \ldots, R$. If $\min(\tilde{n}_k, n_k) > m_k$ for all k, then

$$\kappa_{\mathcal{SV}_{n,d}}(\mathcal{A}_1, \ldots, \mathcal{A}_R) = \kappa_{\mathcal{SV}_{\tilde{n},d}}(\mathcal{B}_1, \ldots, \mathcal{B}_R).$$

Remark 5.5. The case $K = 1$ is exactly Theorem 5.1. The case $d_1 = \cdots = d_K = 1$ is a statement about the PD. In this case, the proposition reads $\kappa_{\mathcal{S}_{n,K}}(\mathcal{A}_1, \ldots, \mathcal{A}_R) = \kappa_{\mathcal{S}_{n,K}}(\mathcal{G}_1, \ldots, \mathcal{G}_R)$, which is a special case of Theorem 4.14.

Proof. For each r, let $\mathcal{G}_r = \alpha_r (g_1^r)^{\otimes d_1} \otimes \cdots \otimes (g_K^r)^{\otimes d_K}$ with $\alpha_r \neq 0$ and $g_k^r \in \mathbb{S}^{m_k - 1}$ for all k. Let $a_k^r = Q_k g_k^r$ and define U_k^r so that $[g_k^r \quad U_k^r] \in \mathbb{R}^{m_k \times m_k}$ is orthogonal. Construct $T_{\mathcal{G}_r}^{\mathcal{SV}_{n,d}}$ by applying (5.10) to $\mathcal{G}_r$. Complete each Q_k to an orthonormal basis $[Q_k \quad Q_k^\perp]$ of $\mathbb{R}^{n_k}$. If $n_k = m_k$, $Q_k^\perp$ is an $n_k \times 0$ matrix. The columns of $U(a_k^r) := [Q_k U_k^r \quad Q_k^\perp]$ form an orthonormal basis of $T_{a_k^r}\mathbb{S}^{n_k - 1}$. For each r, these can be substituted into (5.10) applied to $\mathcal{A}_1, \ldots, \mathcal{A}_R$, respectively. Similarly to the symmetric case, this gives

$$T_{\mathcal{A}_r}^{\mathcal{SV}_{n,d}} = \begin{bmatrix} T_r & T_r^{1\perp} & \cdots & T_r^{K\perp} \end{bmatrix} \quad \text{where} \quad T_r = \left(\bigotimes_{k=1}^{K} Q_k^{\otimes d_k} \right) T_{\mathcal{G}_r}^{\mathcal{SV}_{m,d}}$$

$$\text{and} \quad T_r^{k\perp} = \frac{1}{\sqrt{d_k}} \left(\sum_{d=1}^{d_k} Q_k^\perp \otimes_d (a_k^r)^{\otimes d_k - 1} \right) \otimes_k \left(\bigotimes_{k' \neq k} (a_{k'}^r)^{\otimes d_{k'}} \right)$$

for each r and k. Define $T = [T_r]_{r=1}^R$ and $T^{k\perp} = [T_r^{k\perp}]_{r=1}^R$. Observe that these $K + 1$ matrices are pairwise orthogonal since $Q_k^T Q_k = 0$ and $a_k^r \in \operatorname{span} Q_k$. Furthermore, note that $T_{\mathcal{A}_r}^{\mathcal{SV}_{n,d}} = [T \quad T^{1\perp} \quad \cdots \quad T^{K\perp}]$ up to a column permutation. Finally, $T = \left(\bigotimes_{k=1}^K Q_k \right) T_{\mathcal{G}_1, \ldots, \mathcal{G}_R}^{\mathcal{SV}_{m,d}}$ has the same singular values as $T_{\mathcal{G}_1, \ldots, \mathcal{G}_R}^{\mathcal{SV}_{m,d}}$ by the orthogonality of all Q_k. The combination of these three observations implies that the singular values of $T_{\mathcal{A}_r}^{\mathcal{SV}_{n,d}}$ are the union of the singular values of T, $T^{1\perp}, \ldots, T^{K\perp}$ separately. Consequently, it suffices to show that $\sigma_{\min}(T^{k\perp}) \geqslant \sigma_{\min}(T_{\mathcal{G}_1, \ldots, \mathcal{G}_R}^{\mathcal{SV}_{m,d}})/\sqrt{d_k}$ for each k.

To do this, we compute the Gramian $(T^{k\perp})^T T^{k\perp}$. Define the following auxiliary matrices:

$$A_r^k := \bigotimes_{k'\neq k} (a_{k'}^r)^{\otimes d_{k'}}, \quad G_r^k := \bigotimes_{k'\neq k} (g_{k'}^r)^{\otimes d_{k'}}, \quad \text{and}$$

$$S_r^k := \frac{1}{\sqrt{d_k}} \left(\sum_{d=1}^{d_k} Q_k^\perp \otimes_d (a_k^r)^{\otimes d_k - 1} \right).$$

This allows us to write $T^{k\perp} = S_r^k \otimes_k A_r^k$. For all r_1, r_2, the inner products between the columns of $S_{r_1}^k$ and $S_{r_2}^k$ are

$$(S_{r_1}^k)^T S_{r_2}^k = \langle a_k^{r_1}, a_k^{r_2} \rangle^{d_k - 1} I_{n_k - m_k} = \langle g_k^{r_1}, g_k^{r_2} \rangle^{d_k - 1} I_{n_k - m_k}.$$

Hence, if we replace the factors S_r^k in $T_r^{k\perp}$ by $I_{n_k - m_k} \otimes (g_k^r)^{\otimes d_k - 1}$, the Gramian remains unchanged. Similarly, $(A_{r_1}^k)^T A_{r_2}^k = (G_{r_1}^k)^T G_{r_2}^k$ for all r_1 and r_2, so that we can replace each A_r^k in $T^{k\perp}$ by G_r^k. Define

$$\hat{T}^{k\perp} := \left[I_{n_k - m_k} \otimes (g_k^r)^{\otimes d_k - 1} \otimes_k G_k^r \right]_{r=1}^R \quad \text{and}$$

$$\tilde{T}^{k\perp} := \left[[g_k^r \quad U_k^r] \otimes (g_k^r)^{\otimes d_k - 1} \otimes_k G_k^r \right]_{r=1}^R.$$

$\hat{T}^{k\perp}$ is $T^{k\perp}$ with the aforementioned replacements applied. Since $[g_k^r \quad U_k^r]$ is orthogonal, the singular values of $\hat{T}^{k\perp}$ and $\tilde{T}^{k\perp}$ are the same up to multiplicities by Lemma 4.16. Hence, for the purpose of comparing singular values, we can proceed with $\tilde{T}^{k\perp}$ instead of $T^{k\perp}$.

Next, we also modify $T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}^{\mathcal{SV}_{m,d}}$. First, take the following subset of its columns:

$$T^k := \left[\bigotimes_{k=1}^K (g_k^r)^{\otimes d_k} \quad \frac{1}{\sqrt{d_k}} \left(\sum_{d=1}^{d_k} U_k^{r\perp} \otimes_d (g_k^r)^{\otimes d_k - 1} \right) \otimes_k G_r^k \right]_{r=1}^R$$

The first column of the rth block is $\bigotimes_{k=1}^K (g_k^r)^{\otimes d_k} = (g_k^r)^{\otimes d_k} \otimes_k G_r^k$. Define $\tilde{T}^k$ as a modification of T^k where these R columns are scaled up by $\sqrt{d_k}$. Rearranging the columns gives

$$\tilde{T}^k = \left[\frac{1}{\sqrt{d_k}} \left(\sum_{d=1}^{d_k} [g_k^r \quad U_k^{r\perp}] \otimes_d (g_k^r)^{\otimes d_k - 1} \right) \otimes_k G_r^k \right]_{r=1}^R.$$

Since T^k is a submatrix of $T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}^{\mathcal{SV}_{m,d}}$, we have $\sigma_{\min}(T_{\mathcal{G}_1,\ldots,\mathcal{G}_R}^{\mathcal{SV}_{m,d}}) \leqslant \sigma_{\min}(T^k)$. Because of how we defined $\tilde{T}^k$, we also have $\sigma_{\min}(T^k) \leqslant \sigma_{\min}(\tilde{T}^k)$. From here on, we can compare the singular values of $\tilde{T}^k$ and $\tilde{T}^{k\perp}$ the same way as their counterparts in the proof of Theorem 5.1. This completes the proof. □

5.5 Conclusion

In this chapter, we showed a connection between the condition numbers of related decompositions of (partially) symmetric tensors.

The first main conclusion is that the condition number of a (partially) symmetric tensor decomposition is invariant under symmetric orthogonal Tucker compression if the compressed tensor does not have full multilinear rank. Similarly to Chapter 4, this property yields an efficient method for computing the condition number of a symmetric tensor decomposition of low rank in high dimension: the tensor is compressed to one dimension more than its minimal size (i.e., the multilinear rank) and the condition number is evaluated for the compressed tensor. Compression to the minimal size conjecturally yields the same result and certainly does not change the result by more than a constant factor $\sqrt{D}$.

The second conclusion is that a rank-2 Waring decomposition has the same condition number regardless of whether it is interpreted as a solution to the symmetric or polyadic decomposition problem. Numerical evidence suggests that this holds for an arbitrary number of summands. This would imply that theoretical results about the condition number of the polyadic decomposition such as those of Chapter 4 apply equally to the Waring decomposition. It would also imply that every Waring decomposition is either a singular solution to both the symmetric and polyadic decomposition problem or neither.

Chapter 6

Which constraints of a numerical problem are ill-conditioned?

This chapter consists of the revised text of a manuscript submitted to Numerische Mathematik. At the time of writing, the manuscript has not been accepted. Its preprint version is the following:

N. Dewaele and N. Vannieuwenhoven. "What part of a numerical problem is ill-conditioned?" In: *arXiv preprint arXiv:2305.11547* (2023).

The doctoral candidate derived the theoretical results and performed the experiments. The text was written in collaboration with the coauthors.

Abstract

Many numerical problems with input x and output y can be formulated as a system of equations $F(x, y) = 0$ where the goal is to solve for y. The condition number measures the change of y for small perturbations to x. From this numerical problem, one can derive a (typically underdetermined) relaxation by omitting any number of equations from F. We propose a condition number for underdetermined systems that relates the condition number of a numerical problem to those of its relaxations, thereby detecting the ill-

conditioned constraints. We illustrate the use of our technique by computing the condition of two problems that do not have a finite condition number in the classic sense: two-factor matrix decompositions and Tucker decompositions.

6.1 Introduction

Any computational problem with input $x \in \mathcal{X}$ and output $y \in \mathcal{Y}$ can be characterised by defining a set $\mathcal{P} \subseteq \mathcal{X} \times \mathcal{Y}$ containing all admissible input-output pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$. From here on, we refer to $\mathcal{P}$ as the problem. Given two problems $\mathcal{P}$ and $\mathcal{P}'$, we say that $\mathcal{P}'$ is *less constrained* than $\mathcal{P}$ or a *relaxation of* $\mathcal{P}$ if $\mathcal{P} \subseteq \mathcal{P}'$. The following are typical examples of numerical problems and relaxations.

- A problem defined by a system of equations $F(x, y) = 0$ can be relaxed by removing any number of equations.
- In many applications, a matrix $X \in \mathbb{R}^{m \times n}$ of a known low rank k is decomposed as a product $X = LR$ where $L \in \mathbb{R}^{m \times k}$ and $R \in \mathbb{R}^{k \times n}$. In practice, this problem is usually made more constrained by imposing structure on the tuple $Y = (L, R)$, such as nonnegativity, orthogonality of the columns of L, or by imposing that L contains a subset of the columns of X [TB97; MD09]. Such decompositions are especially preferred for large matrices of low rank [HMT11].
- For a matrix $X \in \mathbb{R}^{m \times n}$ of rank k, computing an *(arbitrary)* basis of the column space is less constrained than computing *(specifically)* the first k columns of U in a singular value decomposition $X = U\Sigma V^T$.
- A *Tucker decomposition* [Tuc66] of a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_D}$ of multilinear rank $(k_1, \ldots, k_D)$ is a relaxation of the *higher-order singular value decomposition* [Tuc66; DLDMV00a]. Similarly, computing a *Tensor train* decomposition is a relaxation of the TT-SVD problem [Ose11].

We say that a problem $\mathcal{P}$ is *identifiable* if a unique $(x, y) \in \mathcal{P}$ exists for every input x. If this y is a continuous function of x, then $\mathcal{P}$ is *well-posed.* In numerical analysis, well-posed problems have a *condition number*, which measures the local sensitivity of the output with respect to small changes in the input. The goal of this chapter is to understand the condition number of $\mathcal{P}$ in relation to the condition number of any relaxation $\mathcal{P}'$ of $\mathcal{P}$. The main obstacle is that $\mathcal{P}'$ *may be too unconstrained to be well-posed and have a condition number in the usual sense.*

To elaborate the concept, consider this issue in the context of linear systems. For $m \leqslant n$, equip $\mathbb{R}^m$ and $\mathbb{R}^n$ with an inner product and let $A \in \mathbb{R}^{m \times n}$ be a fixed matrix of rank m. Consider the system $Ay = x$ for some input $x \in \mathbb{R}^m$. The sensitivity of "solving for y" can be interpreted in several ways.

0. Since the problem is not identifiable, its forward error and condition number are left undefined.

1. We may add constraints to the system to obtain a unique solution. A common way to do this is to minimise $\|y\|$ over all y such that $Ay = x$ [GVL13, Section 5.5]. This more constrained problem is well-posed if A is fixed and can be solved by $y = A^\dagger x$ where $A^\dagger$ is the Moore–Penrose inverse of A.

2. We write the solution corresponding to x as a set $S_x := \{y \mid Ay = x\} \subseteq \mathbb{R}^n$. In the language of [DR14], the map $x \mapsto S_x$ is a *set-valued solution map* and the forward error can be quantified in the *Pompeiu–Hausdorff distance* d_{PH}. For two subsets $S, \tilde{S} \subseteq \mathbb{R}^n$, this is defined as

$$d_{PH}(S, \tilde{S}) := \max\left\{ \sup_{y \in S} d(y, \tilde{S}), \sup_{\tilde{y} \in \tilde{S}} d(\tilde{y}, S) \right\},$$

where $d(y, \tilde{S})$ is the distance from y to its least-squares projection onto $\tilde{S}$. One may define a condition number that measures the error in this distance. To the best of our knowledge, this approach would be difficult to generalise to nonlinear problems, since it is computationally infeasible to compute the Pompeiu–Hausdorff distance for most sets.

3. As before, the solutions are sets S_x, which are $(n-m)$-dimensional affine subspaces of $\mathbb{R}^n$. The set of all such subspaces was called the *affine Grassmannian* in [LWY21]. Since the affine Grassmannian is a Riemannian manifold, it has an induced distance (and hence a measure of forward error). The condition number with respect to this Riemannian distance can be studied using the techniques of [BC13, Chapter 14].

We propose a fourth, practical alternative, which does not require the space of solution sets to be a manifold (as in item 3) and results in a condition number that can be computed using numerical linear algebra. In our approach, we fix any solution y_0 corresponding to the noiseless input x_0. For a noisy input x, the condition number we propose satisfies the error bound

$$\min_{y\colon (x,y) \in \mathcal{P}} d_{\mathcal{Y}}(y_0, y) \leqslant \kappa[\mathcal{P}](x_0, y_0) \cdot d_{\mathcal{X}}(x_0, x) + o(d_{\mathcal{X}}(x_0, x)) \text{ as } x \to x_0, \quad (6.1)$$

where $\mathcal{P} = \{(x, y) \in \mathbb{R}^m \times \mathbb{R}^n \mid Ay = x\}$ and $\kappa[\mathcal{P}](x_0, y_0)$ is the proposed condition number at (x_0, y_0). The left-hand side is the optimal forward error

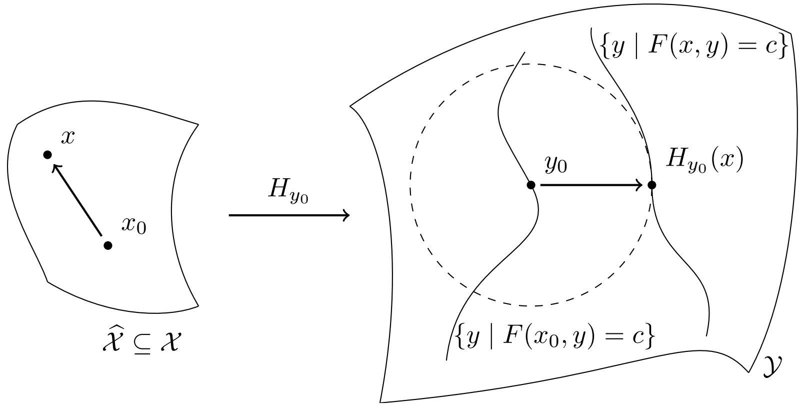

x
x_0
$\widehat{\mathcal{X}} \subseteq \mathcal{X}$
H_{y_0}
$\{y \mid F(x, y) = c\}$
y_0
$H_{y_0}(x)$
$\{y \mid F(x_0, y) = c\}$
$\mathcal{Y}$

Figure 6.1: Simplified view of the solution sets of an FCRE. The point y_0 is a particular solution of $F(x_0, y) = c$ for the noiseless input x_0 and x is a noisy input close to x_0. The point $H_{y_0}(x)$ is the projection of y_0 onto the solution set $\{y \mid F(x, y) = c\}$. If F is linear, then the solution sets are affine spaces which are parallel to each other for all values of x.

over all values of y that solve $\mathcal{P}$ given x. The higher-order term $o(d_{\mathcal{X}}(x_0, x))$ can be neglected if x is close to x_0.

We use this approach to derive an expression of the condition number of a wide class of problems we call *feasible constant-rank equations (FCREs)* and write as $F(x, y) = c$ for some constant c. Deferring a precise definition to Section 6.3, FCREs can be defined informally as systems of (linear or non-linear) equations whose solution sets are smooth[1] manifolds whose points depend smoothly on the input, generalising the linear problem above. Many problems, such as matrix and tensor decompositions, can be modelled as FCREs, even though they are usually not thought of as a system of equations. A fortiori, if $G : \mathbb{R}^m \to \mathbb{R}^n$ is any polynomial map, the inverse problem $F(x, y) := G(y) - x = 0$ is an FCRE. The error measure introduced in (6.1) is visualised for FCREs in Figure 6.1. Our main result can be stated as follows.

Theorem 6.1 (informal version of Theorem 6.6). *Let $\mathcal{X}, \mathcal{Y}$, and $\mathcal{Z}$ be smooth manifolds where $\mathcal{X}$ and $\mathcal{Y}$ have a Riemannian metric. Let $c \in \mathcal{Z}$ be any constant and let $F : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a map so that the equation $F(x, y) = c$ is an FCRE. Finally, let (x_0, y_0) be any pair that solves $F(x_0, y_0) = c$. Writing $\mathcal{P} := F^{-1}(c)$,*

[1] Here and in the rest of the chapter, *smooth* means infinitely differentiable.

the condition number in (6.1) *is*

$$\kappa[\mathcal{P}](x_0, y_0) = \left\| \left(\frac{\partial}{\partial y} F(x_0, y_0) \right)^{\dagger} \frac{\partial}{\partial x} F(x_0, y_0) \right\|,$$

where $\|\cdot\|$ *is the spectral norm.*

The most rudimentary application of our results is the sensitivity of linear systems. Applying Theorem 6.1 to the FCRE $Ay - x = 0$ gives $\|A^{\dagger}\|$ as the (absolute) condition number. Note that this also the condition number of the problem $x \mapsto A^{\dagger}x$ described in item 1 above [TB97, eq. 12.2].

We will show that relaxing a problem defined by an FCRE can never increase the condition number defined in this sense. The precise statement is as follows.

Corollary 6.2. *Given* $R\colon \mathcal{X} \times \widehat{\mathcal{Y}} \to \mathcal{Z}$ *and* $S\colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ *where* $\widehat{\mathcal{Y}} \subseteq \mathcal{Y}$ *is a Riemannian submanifold of* $\mathcal{Y}$*, consider the FCREs* $R(x, y) = \hat{c}$ *and* $S(x, y) = c$ *and assume that the former is more constrained. If* $R(x_0, y_0) = \hat{c}$ *for some* (x_0, y_0)*, then*

$$\kappa[S^{-1}(c)](x_0, y_0) \leqslant \kappa[R^{-1}(\hat{c})](x_0, y_0).$$

Hence, if a problem $\mathcal{P}$ can be relaxed to $\mathcal{P}'$, then the condition number of $\mathcal{P}'$ is a lower bound for the condition number of $\mathcal{P}$. This is essentially because relaxing a problem adds more possible solutions, so that the left-hand side of (6.1) may decrease but not increase. This identity is useful for explaining the condition of a problem $\mathcal{P}$: if $\mathcal{P}$ can be relaxed to $\mathcal{P}'$, and $\mathcal{P}'$ is ill-conditioned, then so is $\mathcal{P}$. Conversely, if $\mathcal{P}$ is ill-conditioned but its relaxation $\mathcal{P}'$ is not, then the additional constraints that $\mathcal{P}$ adds to $\mathcal{P}'$ explain the condition of $\mathcal{P}$.

The simplest instance of this is again the underdetermined linear system $Ay = x$ with $A \in \mathbb{R}^{m \times n}$, whose condition number is $\|A^{\dagger}\|$ at any solution pair (x_0, y_0). A more constrained system could introduce $n - m$ additional equations and be written as $\overline{A}y = \overline{x}$ where $\overline{A} \in \mathbb{R}^{n \times n}$, the first m rows of $\overline{A}$ are A, and the first m components of $\overline{x}$ are x. The absolute condition number of this system is $\left\|\overline{A}^{-1}\right\|$. By the singular value interlacing theorem [HJ12, Theorem 4.3.17], $\left\|\overline{A}^{-1}\right\| \geqslant \|A^{\dagger}\|$, which confirms Corollary 6.2.

Suppose that a problem $\mathcal{P}$ has a solution pair (x_0, y_0) and $\mathcal{P}'$ is a relaxation of $\mathcal{P}$. We say that $\mathcal{P}$ is an *optimal refinement* of $\mathcal{P}'$ at (x_0, y_0) if relaxing $\mathcal{P}$ to $\mathcal{P}'$ does not decrease its condition number, i.e., $\kappa[\mathcal{P}](x_0, y_0) = \kappa[\mathcal{P}'](x_0, y_0)$. By the above example, if we have an underdetermined linear system $Ay = x$, then imposing the constraint that $\|y\|$ be minimal gives an optimal refinement. Thus, our results justify the widespread use of this minimality constraint.

In another example, which we elaborate in Section 6.6, we present a family of matrices $\{X_\varepsilon\}_{\varepsilon\in\mathbb{R}^+}$ for which the computation of a singular value decomposition $X_\varepsilon = U\Sigma V^T$ can be arbitrarily ill-conditioned, but the less constrained *orthogonal Tucker decomposition* (which is the same as a singular value decomposition, except it does not enforce Σ to be diagonal) has a near-optimal condition number. Thus, the singular value decomposition is not an optimal refinement of the Tucker decomposition. That is, for a perturbation of X_ε, it is possible to find a Tucker decomposition close to the decomposition of X_ε, but only if its second factor is not diagonal. We may say that the constraint that Σ be diagonal is responsible for the difference in condition.

6.1.1 Notation

The $n \times n$ identity matrix is denoted by $\mathbb{I}_n$. For a fixed x, we write

$$F_x(y) := F(x,y) \quad \text{and} \quad F_x^{-1}(c) := \{y \mid F(x,y) = c\}.$$

We define the following manifolds: $\mathrm{St}(m,n) := \{U \in \mathbb{R}^{m\times n} \,|\, U^TU = \mathbb{I}_n\}$ is the Stiefel manifold, $O(n) := \mathrm{St}(n,n)$ is the orthogonal group, and $\mathbb{R}_k^{m\times n}$ is the manifold of $m \times n$ matrices of rank k. The kth largest singular value of a matrix A is $\sigma_k(A)$. The canonical basis vectors of $\mathbb{R}^n$ are $e_1, \ldots, e_n$. The tangent space to a manifold $\mathcal{M}$ at p is $\mathcal{T}_p\mathcal{M}$. A generic vector in this space is denoted as $\dot{p}$.

6.1.2 Summary of contributions and outline

The main contribution of this work is an asymptotically sharp estimate for the optimal forward error of a general FCRE. This error estimate is given by Theorem 6.6 and Corollary 6.9, which are proved in Section 6.3.1. The advantage of our approach is that it requires little geometric information about the solution sets. For certain underdetermined systems, though, the solution set can be seen as a unique point on a quotient manifold. We compare this point of view to our approach in Section 6.4.

Another contribution is the computation of the condition number of two specific problems of independent interest: two-factor matrix decomposition and Tucker decomposition of tensors. They are studied in Section 6.5 and Section 6.6, respectively. Numerical experiments for the accuracy of the error bound (6.5) in the case of the Tucker decomposition are presented in Section 6.7.

We start with an overview of the theory of condition numbers of nonlinear equations in the next section.

6.2 Classic theory of condition

Rice [Ric66] defined the condition number of a map $H : \mathcal{X} \to \mathcal{Y}$ between metric spaces at a point $x_0 \in \mathcal{X}$ as

$$\kappa[H](x_0) := \limsup_{x \to x_0} \frac{d_{\mathcal{Y}}(H(x_0), H(x))}{d_{\mathcal{X}}(x, x_0)} \tag{6.2}$$

where $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are the distances in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Equivalently, $\kappa[H](x_0)$ is the smallest number κ such that

$$d_{\mathcal{Y}}(H(x_0), H(x)) \leqslant \kappa \cdot d_{\mathcal{X}}(x_0, x) + o(d_{\mathcal{X}}(x_0, x)) \quad \text{as} \quad x \to x_0. \tag{6.3}$$

The latter property is called *asymptotic sharpness* of the bound (6.3). Since (6.2) depends on H (and consequently its domain) as well as the distances $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$, it is more precise to call $\kappa[H](x_0)$ *the condition number of H with respect to $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$*, but we refer to it simply as *the condition number of H*. Many condition numbers in the literature are instances of (6.2) for some choice of $\mathcal{X}, \mathcal{Y}$, and their distances. The following examples are common.

- If $\mathcal{X}$ and $\mathcal{Y}$ are normed vector spaces, there is a natural distance $d_{\mathcal{X}}(x, x') := \|x - x'\|$ and likewise in $\mathcal{Y}$. In this case, (6.2) is the *absolute normwise condition number* $\kappa_{\text{abs}}[H]$. Alternatively, we may fix x_0 and define $d_{\mathcal{X}}(x, x') := \|x - x'\| / \|x_0\|$ and likewise in $\mathcal{Y}$. This corresponds to the *relative normwise condition number* at x_0 [TB97, Lecture 12].
- If $\mathcal{X}$ is not a linear space, then (6.2) is often referred to as a *structured* condition number [HU17; ANT19]. For instance, if $\mathcal{X} \subseteq \mathbb{R}^{n \times n}$ is a manifold of structured, orthogonal, or low-rank matrices, then (6.2) only considers those matrices x as possible perturbations.
- If $\mathcal{X}$ and $\mathcal{Y}$ are expressed in coordinates and one is interested in perturbations of only one coordinate, a *componentwise* condition number may be used [GK93]. This is less straightforward to define as a special case of (6.2) and will not be considered in this chapter.

We will consider general maps H where $\mathcal{X}$ and $\mathcal{Y}$ are Riemannian manifolds. In this case, Rice's theorem [Ric66] says that $\kappa[H](x_0) = \|DH(x_0)\|$ where DH is the differential of H and $\|\cdot\|$ is the operator norm. In particular, for a map H between Euclidean spaces, $\kappa_{\text{abs}}[H]$ is the spectral norm of the Jacobian matrix of H, as in [TB97, Lecture 12].

Rice's definition (6.2) can be extended to numerical problems $\mathcal{P} \subseteq \mathcal{X} \times \mathcal{Y}$ that cannot be described as a map $H : \mathcal{X} \to \mathcal{Y}$, such as the solution of a polynomial equation $\sum_{i=0}^{d} a_i y^i = 0$, where $x = (a_0, \ldots, a_d)$, $a_d \neq 0$, and $d \geqslant 2$. Such problems may have multiple isolated solutions. Demmel [Dem87] defined a condition number for univariate polynomial rootfinding, which was generalised by Shub and Smale [SS93] to homogeneous polynomial systems and by Bürgisser and Cucker [BC13, Section 14.3] to problems $\mathcal{P} \subseteq \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are Riemannian manifolds.

This extended definition of the condition number goes as follows: let the problem $\mathcal{P}$ be defined by a system of equations $F(x, y) = c$ where $F : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ is a smooth map (meaning infinitely differentiable) and $c \in \mathcal{Z}$ is a constant. If (x_0, y_0) is any solution and $\frac{\partial}{\partial y} F(x_0, y_0)$ is invertible, the implicit function theorem implies the existence of a unique smooth map, often called the *solution map* $H : \widehat{\mathcal{X}} \to \mathcal{Y}$ defined on a neighbourhood $\widehat{\mathcal{X}} \subseteq \mathcal{X}$ of x_0 such that $H(x_0) = y_0$ and $F(x, H(x)) = c$ for all $x \in \widehat{\mathcal{X}}$. The condition number of $\mathcal{P}$ as discussed in [BC13, Section 14.3] is defined as $\kappa[F^{-1}(c)](x_0, y_0) := \kappa[H](x_0)$, where the right-hand side is given by (6.2). Working this out gives

$$\kappa[F^{-1}(c)](x_0, y_0) = \left\| \left(\frac{\partial}{\partial y} F(x_0, y_0) \right)^{-1} \frac{\partial}{\partial x} F(x_0, y_0) \right\|. \tag{6.4}$$

Remark 6.3. If the equation $F(x, y) = c$ has two different solutions y_0 and y_0' for the same input x_0, then their solution maps are different and may generally have a different condition number. Hence, the condition number may depend on the solution as well as the input.

Example 6.4 (Eigenvalue problems)**.** Let $\mathcal{X} := \mathbb{C}^{n \times n}$ and $\mathcal{Y} = \mathbb{C}$ and define $F(X, \lambda) := \det(X - \lambda \mathbb{I})$. The eigenvalues of X are precisely the solutions of $F(X, \lambda) = 0$. Endow $\mathbb{C}^n$ with the Hermitian inner product $\langle \cdot, \cdot \rangle$ and $\mathbb{C}^{n \times n}$ with the spectral norm. Let X_0 be a matrix with a simple eigenvalue λ_0 and left and right eigenvectors v and w, respectively. A basic result in matrix analysis states that there exists a locally unique smooth function $\lambda(X)$ such that $F(X, \lambda(X)) = 0$ and $\lambda(X_0) = \lambda_0$, see e.g. [HJ12, Theorem 6.3.12]. Its differential at X_0 is $D\lambda(X_0)[\dot{X}] = \frac{\langle v, \dot{X} w \rangle}{\langle v, w \rangle}$. Hence, $\kappa[F^{-1}(0)](X_0, \lambda_0) = \frac{\|v\| \|w\|}{|\langle v, w \rangle|}$. For a generic X_0, all its eigenvalues have a different condition number, as per Remark 6.3.

6.2.1 Related work

The condition number of systems with unique solutions is well understood [BC13]. We know of two works studying a condition number in the sense of (6.1). First, Dedieu [Ded96] introduced the *inverse condition number* of a numerical problem $G : \mathcal{Y} \to \mathcal{X}$, where $\mathcal{Y}$ and $\mathcal{X}$ are Euclidean spaces. The expression for this condition number is equivalent to Corollary 6.9, but its interpretation is different. Dedieu interpreted $\mathcal{Y}$ as the input space, $\mathcal{X}$ as the output space, and was interested in measuring backward errors. Conversely, we study the equation $G(y) = x$ as a problem taking $\mathcal{X} \to \mathcal{Y}$ and consider the forward error as in (6.5). Another occurrence of a latent condition number is due to Vannieuwenhoven, who studied the sensitivity of the tensor rank decomposition in its factor matrix representation [Van17]. Theorem 6.1 is a generalisation of both of these results. Another approach for defining a condition number of certain underdetermined systems is based on quotient manifolds. We explain this in detail in Section 6.4.

For several problems in numerical analysis, there is a connection between first-order sensitivity as in (6.3), the distance to the nearest ill-posed problem [Dem87], and the convergence of iterative algorithms [SS93]. Constants appearing in estimates of any of these measures are often called condition numbers, even if the asymptotic sharpness of the estimate is not demonstrated. Specifically, Dégot [Dé00] introduced a condition-like number for underdetermined homogeneous polynomial systems that measures distance to ill-posedness. The same number provides an error estimate of the solution, but little was said about the asymptotic sharpness of this estimate. Dedieu and Kim [DK02] analysed a generalised Newton method for solving the equation $G(x) = 0$, where rank $DG(x)$ is constant. The rate of convergence can be estimated in terms of $\left\| DG(x)^\dagger \right\|$, i.e., the expression appearing in Corollary 6.9. In the context of linear least-squares problems of the form $Ax = b$, the similar expression $\|A\| \left\| A^\dagger \right\|$ is sometimes referred to as the condition number of A, even if this number is not the condition number of the problem as defined by (6.2) (see e.g. [SS90, Corollary III.3.10] and the discussion thereafter).

Another conceptually similar condition number is that of Riemannian approximation [BV21]. In that context, the problem is to project a variable point $x \in \mathbb{R}^n$ onto a fixed manifold $\mathcal{M} \subseteq \mathbb{R}^n$. The latent condition number, by contrast, measures how a fixed point y_0 is projected onto a variable solution set $F_x^{-1}(c)$.

6.3 Proposed theory of condition

The condition number in the sense of Rice is defined for the evaluation of a map $H : \mathcal{X} \to \mathcal{Y}$ and for the solution of systems of equations $F(x, y) = c$ where $\frac{\partial}{\partial y} F(x, y)$ is invertible. In this section, we loosen this constraint on $\frac{\partial}{\partial y} F(x, y)$ and propose a corresponding condition number.

Definition 6.5. Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be smooth manifolds and let $c \in \mathcal{Z}$ be a constant. Let $F : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a smooth map. We call the equation $F(x, y) = c$ a *feasible constant-rank equation (FCRE)* if the following holds:

1. for all $x \in \mathcal{X}$, there exists a point $y \in \mathcal{Y}$ such that $F(x, y) = c$,
2. there exists a number $r \in \mathbb{N}$ so that $\operatorname{rank} \frac{\partial}{\partial y} F(x, y) = r$ and $\operatorname{rank} DF(x, y) = r$ for all $x, y \in \mathcal{X} \times \mathcal{Y}$.

Note that if $r = \dim \mathcal{Y} = \dim \mathcal{Z}$, then $\frac{\partial}{\partial y} F(x, y)$ is invertible, which is the usual condition under which the condition number in the sense of [BC13, Section 14.3] is defined. Note as well that if a map F only satisfies this definition locally, the restriction of F to a subset of its domain defines an FCRE.

The idea behind the second item in Definition 6.5 is as follows. If $\operatorname{rank} DF$ is constant, then F is a *map of constant rank*, which is a fundamental concept in differential geometry [Lee13, Chapter 4]. One such example are polynomial maps: if $F : \mathbb{R}^m \to \mathbb{R}^n$ is a polynomial, then the locus of points $p \in \mathbb{R}^m$ such that $\operatorname{rank} DF(p)$ is *not* maximal is a subvariety of $\mathbb{R}^m$ of dimension less than m. Thus, polynomials have constant rank almost everywhere [BC13, Proposition A.35].

For any map F of constant rank, the problem $\mathcal{P} := F^{-1}(c)$ is a smooth manifold of dimension $\operatorname{null} DF$ where null is the nullity [Lee13, Theorem 4.12]. It then follows that $\operatorname{rank} \frac{\partial F}{\partial y} = \operatorname{rank} DF$ if and only if $\dim \mathcal{P} = \dim \mathcal{X} + \operatorname{null} \frac{\partial F}{\partial y}$. This should be compared to identifiable problems, which have only $\dim \mathcal{X}$ degrees of freedom (since every $x \in \mathcal{X}$ would determine a unique $(x, y) \in \mathcal{P}$). By the foregoing, FCREs are defined exactly by the maps of constant rank that offer $\operatorname{null} \frac{\partial F}{\partial y}$ additional degrees of freedom.

Our main theorem underlies the definition of the condition number of an FCRE. It is proved in Section 6.3.1.

Theorem 6.6. *Let $F(x, y) = c$ be an FCRE as in Definition 6.5 and let (x_0, y_0) be any pair such that $F(x_0, y_0) = c$. If $\mathcal{X}$ and $\mathcal{Y}$ are Riemannian manifolds, then there exist a neighbourhood $\widehat{\mathcal{X}} \subseteq \mathcal{X}$ of x_0 and a smooth map, called the* canonical solution map

$$H_{y_0} : \widehat{\mathcal{X}} \to \mathcal{Y}$$

$$x \mapsto \arg \min_{\substack{y \in \mathcal{Y} \\ F(x,y)=c}} d_{\mathcal{Y}}(y_0, y),$$

where $d_{\mathcal{Y}}$ is the geodesic distance in $\mathcal{Y}$. Its differential at x_0 is

$$DH_{y_0}(x_0) = \left(\frac{\partial}{\partial y}F(x_0, y_0)\right)^{\dagger} \frac{\partial}{\partial x}F(x_0, y_0).$$

That is, out of all possible solution maps, H_{y_0} locally minimises the distance to y_0. Figure 6.1 shows a visualisation of H_{y_0}. The preceding theorem allows us to define our primary object of interest.

Definition 6.7. In the context of Theorem 6.6, the *latent condition number* of F at (x_0, y_0) is

$$\kappa[F^{-1}(c)](x_0, y_0) := \kappa[H_{y_0}](x_0) = \left\| \left(\frac{\partial}{\partial y}F(x_0, y_0)\right)^{\dagger} \frac{\partial}{\partial x}F(x_0, y_0) \right\|$$

where $\|\cdot\|$ is the operator norm.

Note that this is exactly the generalisation of (6.4) that would be obtained if $\frac{\partial F}{\partial y}$ in (6.4) were naively replaced by a pseudoinverse. The value of Theorem 6.6 is that it shows that this generalised formula has a precise interpretation: κ expresses whether the equation $F(x, y) = c$ has a solution close to y_0 if x is a slight perturbation of x_0. If $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are the geodesic distances in $\mathcal{X}$ and $\mathcal{Y}$, respectively, we have the following asymptotically sharp error bound:

$$\min_{\substack{y \in \mathcal{Y},\\ F(x,y)=c}} d_{\mathcal{Y}}(y_0, y) \leqslant \kappa[F^{-1}(c)](x_0, y_0) \cdot d_{\mathcal{X}}(x_0, x) + o(d_{\mathcal{X}}(x_0, x)) \text{ as } x \to x_0. \tag{6.5}$$

Since this is a bound on the asymptotic behaviour as $x \to x_0$, the same bound holds for any distance d such that $d(x_0, x) = d_{\mathcal{X}}(x_0, x)(1 + o(1))$ as $x \to x_0$ and likewise for $d_{\mathcal{Y}}$. For instance, if $\mathcal{X}$ is an embedded Riemannian submanifold of a Euclidean space $\mathcal{E}$, we may take d to be the Euclidean distance in $\mathcal{E}$. Then (6.5) gives a bound in the (more practical) Euclidean distance d.

Remark 6.8. The classic condition number of equations on manifolds, as defined in [BC13, Section 14.3], requires a *unique* solution map at the given solution pair (x_0, y_0). If this map does not exist or is not unique, the condition number is either undefined or infinite by definition [BV18b]. If the classic condition number of an FCRE is finite, the unique solution map is the map from Theorem 6.6 and the condition number is the latent condition number.

The condition number of a relaxation of a FCRE can be used as a lower bound for the condition number of the original FCRE, as stated in Corollary 6.2. In other words, if a (relaxed or underdetermined) problem has a high latent condition number, then all ways to make it well-posed by adding constraints will be ill-conditioned. This is the intuition behind the name *latent condition number*. Corollary 6.2 is a straightforward consequence of the definition, but we prove it here for completeness.

Proof of Corollary 6.2. Let H_R and H_S be the solution maps arising from the application of Theorem 6.6 to R and S, respectively. For all $x \in \mathcal{X}$ sufficiently close to x_0, we have

$$d_{\mathcal{Y}}(H_S(x_0), H_S(x)) = \min_{\substack{y \in \mathcal{Y} \\ S(x,y)=c}} d_{\mathcal{Y}}(y_0, y).$$

An upper bound on this can be obtained by restricting the domain of the minimum and applying the identity that $d_{\mathcal{Y}}(y, y') \leqslant d_{\widehat{\mathcal{Y}}}(y, y')$ for all $y, y' \in \mathcal{Y}$. Thus,

$$d_{\mathcal{Y}}(H_S(x_0), H_S(x)) \leqslant \min_{\substack{y \in \widehat{\mathcal{Y}} \\ R(x,y)=\hat{c}}} d_{\widehat{\mathcal{Y}}}(y_0, y) = d_{\widehat{\mathcal{Y}}}(H_R(x_0), H_R(x))$$

so that the result follows from (6.2). □

An important class of systems of equations are equations of the form $G(y) - x = 0$ for some smooth map $G : \mathcal{Y} \to \mathbb{R}^m$. In this case, Theorem 6.6 specialises to the following Riemannian generalisation of [Ded96, Theorem C].

Corollary 6.9. *Let $\mathcal{Y}$ be a Riemannian manifold and let $G : \mathcal{Y} \to \mathbb{R}^m$ be a smooth map such that* $\operatorname{rank} DG(y)$ *is constant. Pick any point (x_0, y_0) on the graph of G. Then y_0 has a neighbourhood $\widehat{\mathcal{Y}}$ such that $\mathcal{X} := G(\widehat{\mathcal{Y}})$ is an embedded submanifold of $\mathbb{R}^m$. Define*

$$F : \mathcal{X} \times \widehat{\mathcal{Y}} \to \mathbb{R}^m$$

$$(x, y) \mapsto G(y) - x.$$

Then $F(x, y) = 0$ is an FCRE and $\kappa[F^{-1}(0)](x_0, y_0) = \left\| DG(y_0)^\dagger \right\|$.

For a map G satisfying the assumptions in Corollary 6.9, we call the equation $G(y) - x = 0$ a *constant-rank inverse problem* and we write $\kappa^{inv}[G](y_0) := \kappa[F^{-1}(c)](x_0, y_0)$. The dependence on x_0 is not written explicitly since x_0 is determined by y_0. Note that $\left\| DG(y_0)^\dagger \right\|$ is the reciprocal of the smallest nonzero singular value of $DG(y_0)$.

6.3.1 Proof of Theorem 6.6

The proof of Theorem 6.6 is an application of standard concepts from differential geometry and numerical analysis. We will use the following lemma to parametrise the tangent space to the solution sets.

Lemma 6.10. *Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be smooth manifolds of dimensions m, n and k, respectively, and let $F : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a smooth map. Suppose that $\operatorname{rank} \frac{\partial}{\partial y} F(x,y) = r$ is constant. In a neighbourhood of any point $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, there exists a linearly independent tuple of smooth vector fields $((0, E_1(x,y)), \ldots, (0, E_{n-r}(x,y)))$ over $\mathcal{X} \times \mathcal{Y}$ whose span is $\{0\} \times \ker \frac{\partial}{\partial y} F(x,y)$.*

Proof. Consider the map

$$\tilde{F} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Z}$$

$$(x,y) \mapsto (x, F(x,y)).$$

Then, $(\dot{x}, \dot{y}) \in \ker D\tilde{F}(x,y)$ if and only if $\dot{x} = 0$ and $(\dot{x}, \dot{y}) \in \ker DF(x,y)$. Since $DF(x,y)[\dot{x}, \dot{y}] = \frac{\partial}{\partial x} F(x,y)[\dot{x}] + \frac{\partial}{\partial y} F(x,y)[\dot{y}]$, $\dot{x} = 0$, and $\frac{\partial}{\partial x} F(x,y) \colon \mathcal{T}_x\mathcal{X} \to \mathcal{T}_{F(x,y)}\mathcal{Z}$ is a linear map, we have

$$\ker D\tilde{F}(x,y) = \{0\} \times \ker \frac{\partial}{\partial y} F(x,y) \tag{6.6}$$

and $\operatorname{rank} D\tilde{F}(x,y) = m + r$ for all x, y.

By the constant rank theorem [Lee13, Theorem 4.12], there exist charts for the domain and codomain of $\tilde{F}$ in which $\tilde{F}$ is represented as

$$(u^1, \ldots, u^{m+n}) \mapsto (u^1, \ldots, u^{m+r}, 0, \ldots, 0).$$

In these coordinates, the basis $\mathcal{B} := \left\{\frac{\partial}{\partial u^i}\right\}_{i=m+r+1}^{m+n}$ spans $\ker D\tilde{F}(x,y)$. By (6.6), we may write these $\frac{\partial}{\partial u^i}$ as smooth vector fields $(0, E_{i-m-r}(x,y))$, where $E_{i-m-r}(x,y) \in \ker \frac{\partial}{\partial y} F(x,y)$. □

Now we can prove the existence of the canonical solution map.

Proof of Theorem 6.6. Let $n = \dim \mathcal{Y}$ and let $g_{\mathcal{Y}}$ be the Riemannian metric on $\mathcal{Y}$. Let $\{E_i(x,y)\}_{i=1}^{n-r}$ be the vector fields from Lemma 6.10.

By the constant rank theorem [Lee13, Theorem 4.12], there exists a neighbourhood $\mathcal{U} \subseteq \mathcal{X} \times \mathcal{Y}$ of (x_0, y_0) and a chart $\phi_{\mathcal{Z}} : \mathcal{Z} \to \mathbb{R}^{\dim \mathcal{Z}}$

such that $\phi_{\mathcal{Z}}(c) = 0$ and $\phi_{\mathcal{Z}}(F(x,y)) = (\widehat{F}(x,y), 0)$ for some smooth map $\widehat{F} : \mathcal{U} \to \mathbb{R}^r$. Thus, in this neighbourhood, the equation $F(x,y) = c$ is equivalent to $\widehat{F}(x,y) = 0$.

Let $\log_y : \mathcal{Y} \to \mathcal{T}y\mathcal{Y}$ be the inverse of the exponential map in $\mathcal{Y}$. Informally, $\log_y y_0$ is the vector in $\mathcal{T}_y\mathcal{Y}$ that "points towards" y_0. Define $\phi_i(x,y) := g_{\mathcal{Y}}(E_i(x,y), \log_y y_0)$ and consider the system of n equations

$$\Phi(x,y) := (\widehat{F}(x,y),\, \phi_1(x,y), \ldots, \phi_{n-r}(x,y)) = (0, 0, \ldots, 0). \tag{6.7}$$

The last $n - r$ equations specify that $\log_y y_0$ is orthogonal to $\ker \frac{\partial}{\partial y} F(x,y)$ and thus normal to $F_x^{-1}(c)$. We will show, using the implicit function theorem, that (6.7) has a locally unique solution.

Let $g_{\mathcal{X}\times\mathcal{Y}}$ be the product metric in $\mathcal{X} \times \mathcal{Y}$. Then

$$\phi_i(x,y) = g_{\mathcal{X}\times\mathcal{Y}}\left((0, E_i(x,y)), (0, \log_y y_0)\right).$$

Let $(\xi, \eta) \in \mathcal{T}(x_0, y_0)(\mathcal{X}\times\mathcal{Y})$ be any tangent vector and let ∇ be the Levi–Civita connection for $g_{\mathcal{X}\times\mathcal{Y}}$. We calculate $D\phi_i(x_0, y_0)$ using the product rule:

$$\begin{aligned} D\phi_i(x_0, y_0)[\xi, \eta] = g_{\mathcal{X}\times\mathcal{Y}} &\left(\nabla_{(\xi,\eta)}(0, E_i(x,y)), (0, \log_{y_0} y_0)\right) \\ &+ g_{\mathcal{X}\times\mathcal{Y}}\left((0, E_i(x_0, y_0)), \nabla_{(\xi,\eta)}(0, \log_y y_0)\right). \end{aligned} \tag{6.8}$$

The first term vanishes because $\log_{y_0} y_0 = 0$. The second term simplifies to $g_{\mathcal{Y}}(E_i(x_0, y_0), \nabla_\eta \log_y y_0)$. Using normal coordinates centred at y_0, the vector field $\log_y y_0$ can be written as $\log_y y_0 = -\sum_{i=1}^n y^i \frac{\partial}{\partial y^i}$, so that $\nabla_\eta \log_y y_0 = -\eta$ [Lee13, Proposition 5.24]. Hence, (6.8) is equal to $-g_{\mathcal{Y}}(E_i(x_0, y_0), \eta)$.

To apply the implicit function theorem to (6.7), we verify that $\frac{\partial}{\partial y}\Phi(x_0, y_0)$ is invertible. It suffices to show that the kernel of $\frac{\partial}{\partial y}\Phi(x_0, y_0)$ is trivial. If a vector $\dot{y} \in \mathcal{T}y_0\mathcal{Y}$ is such that $\frac{\partial}{\partial y}\Phi(x_0, y_0)[\dot{y}] = 0$, then $\dot{y} \in \ker \frac{\partial}{\partial y}\widehat{F}(x_0, y_0) = \ker \frac{\partial}{\partial y} F(x_0, y_0)$. Furthermore, if $\frac{\partial}{\partial y}\phi_i(x_0, y_0)[\dot{y}] = 0$ for all i, then $\dot{y}$ is orthogonal to $E_i(x_0, y_0)$ for all i. By the definition of E_i, it follows that $\dot{y} \perp \ker \frac{\partial}{\partial y} F(x_0, y_0)$ and thus $\dot{y} = 0$. Therefore, $\ker \frac{\partial}{\partial y}\Phi(x_0, y_0) = \{0\}$. By the implicit function theorem [Lee13, Theorem C.40], there exists a neighbourhood $\widehat{\mathcal{X}} \times \widehat{\mathcal{Y}}$ of (x_0, y_0) and a smooth function H_{y_0} such that $\Phi(x,y) = 0$ for $(x,y) \in \widehat{\mathcal{X}} \times \widehat{\mathcal{Y}}$ if and only if $y = H_{y_0}(x)$.

Next, we show that $H_{y_0}(x)$ is the map from the theorem statement. Consider a variable point $x \in \widehat{\mathcal{X}}$. By continuity of H_{y_0}, if x is sufficiently close to x_0, then

$H_{y_0}(x)$ lies in the interior of some compact geodesic ball $\overline{B} \subseteq \widehat{\mathcal{Y}}$ of radius ρ around y_0. Since the level set $F_x^{-1}(c)$ is properly embedded [Lee13, Theorem 4.12], the minimum of $d_{y_0}(y) := d_{\mathcal{Y}}(y_0, y)$ over all $y \in F_x^{-1}(c) \cap \overline{B}$ is attained. The interior of $F_x^{-1}(c) \cap \overline{B}$ contains at least $H_{y_0}(x)$ and, since $d(y_0, H_{y_0}(x)) < \rho$, it follows that d_{y_0} attains a minimum in the interior of $F_x^{-1}(c) \cap \overline{B}$. Thus, at the minimiser $y_\star$, we must have

$$\operatorname{grad} d_{y_0}^2(y) \perp \mathcal{T}y_\star F_x^{-1}(c) = \ker \frac{\partial}{\partial y} F(x, y_\star),$$

where $\operatorname{grad} d_{y_0}^2(y) = -2 \log_y y_0$. As we established above, $H_{y_0}(x)$ is the unique point that solves (6.7). In other words, it is the only $y \in F_x^{-1}(c) \cap \widehat{\mathcal{Y}}$ such that $\log_y y_0 \perp \ker \frac{\partial}{\partial y} F(x, y)$. Thus, $H_{y_0}(x) = y_\star$, as required.

We obtain an expression for $DH_{y_0}(x_0)$ as follows. Since $\Phi(x, H_{y_0}(x))$ is constant for all x, it follows by implicit differentiation that

$$\frac{\partial}{\partial x} \Phi(x_0, y_0) + \frac{\partial}{\partial y} \Phi(x_0, y_0) DH_{y_0}(x_0) = 0.$$

By substituting the partial derivatives of Φ obtained in the proof, we get

$$\begin{cases} \frac{\partial}{\partial x} F(x_0, y_0) + \frac{\partial}{\partial y} F(x_0, y_0) DH_{y_0}(x_0) & = 0, \\ E_i^*(x_0, y_0) DH_{y_0}(x_0) & = 0 \quad \text{for all} \quad i = 1, \dots, n - r, \end{cases}$$

where $\cdot^*$ is the dual (or adjoint). In other words, $DH_{y_0}(x_0)$ is the unique matrix that solves $\frac{\partial}{\partial y} F(x_0, y_0) DH_{y_0}(x_0) = -\frac{\partial}{\partial x} F(x_0, y_0)$ and has a column space orthogonal to $\ker \frac{\partial}{\partial y} F(x_0, y_0)$. Hence, $DH_{y_0}(x_0) = -\left(\frac{\partial}{\partial y} F(x_0, y_0)\right)^\dagger \frac{\partial}{\partial x} F(x_0, y_0)$. □

6.4 Problems invariant under orthogonal symmetries

A notable advantage of the latent condition number of an FCRE $F(x, y) = c$ is that it only requires information about the local behaviour of F around a particular solution (x_0, y_0). In particular, it does not require an explicit parametrisation of all solutions in terms of (x_0, y_0). By contrast, for some underdetermined systems studied in the literature, the derivation of their condition number relies on the solutions being unique up to a known equivalence relation [BC13; Van17].

When it *is* known that the system is invariant under certain symmetries, however, more can be said about the condition number. For instance, since the condition number generally depends on the solution y and the parameter x, it is natural to ask when it depends on the parameter alone. That is, when do two distinct y_1, y_2 that solve $F(x,y) = c$ for the same x satisfy $\kappa[F^{-1}(c)](x,y_1) = \kappa[F^{-1}(c)](x,y_2)$? An obvious sufficient condition for this is that both F and its solutions are invariant under some family of isometries. This is captured by the following statement.

Proposition 6.11. *Let $F(x,y) = c$ be an FCRE with $F : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$. Let $\psi : \mathcal{Y} \to \mathcal{Y}$ be an isometry such that $F \circ (\mathrm{Id} \times \psi) = F$. For any $x, y \in \mathcal{X} \times \mathcal{Y}$, we have*

$$\kappa[F^{-1}(c)](x,y) = \kappa[F^{-1}(c)](x,\psi(y)).$$

Proof. Compute

$$DF(x,y) = D(F \circ (\mathrm{Id} \times \psi))(x,y) = DF(x,\psi(y))(\mathrm{Id} \times D\psi(y))$$

so that $\frac{\partial}{\partial y}F(x,y) = \frac{\partial}{\partial y}F(x,\psi(y))D\psi(y)$ and $\frac{\partial}{\partial x}F(x,y) = \frac{\partial}{\partial x}F(x,\psi(y))$. Since $D\psi(y)$ is an orthogonal matrix, applying Theorem 6.6 gives the desired result. □

This proposition is useful when the solutions are determined up to certain isometries. That is, suppose that $x \in \mathcal{X}$ is any point and $\{\psi_i\}_{i \in I}$ is a family of isometries such that $F = F \circ (\mathrm{Id} \times \psi_i)$ for all i. If, for every y_1, y_2 where $F(x,y_1) = F(x,y_2) = c$, there exists an $i \in I$ such that $y_1 = \psi_i(y_2)$, then the above implies that all solutions of $F(x,y) = c$ have the same condition number.

For several problems in numerical linear algebra, the solutions are unique up to multiplication by an orthogonal matrix, i.e., a linear isometry. Thus, their condition number depends only on the input by Proposition 6.11. Some examples include:

1. *Positive-semidefinite matrix factorisation*: $\mathcal{X} = (\mathbb{S}_k^{n \times n})^+$ is the set of symmetric positive semidefinite matrices of rank k and $\mathcal{Y} = \mathbb{R}_k^{n \times k}$. A symmetric factorisation of $X \in \mathcal{X}$ is a solution of $F(X,Y) = 0$, where $F(X,Y) := X - YY^T$. The use of this factorisation in optimisation was popularised by Burer and Monteiro [BM03].

2. *Computation of an orthonormal basis of the kernel*: $\mathcal{X} = \mathbb{R}_k^{m \times n}$ and $\mathcal{Y} = \mathrm{St}(n, n-k)$ and $F(X,Y) = 0$, where $F(X,Y) := XY$.

3. *Computation of an orthonormal basis of the column space*: if $\mathcal{X} = \mathbb{R}_k^{m \times n}$ $\mathcal{Y} = \mathrm{St}(m,k)$, then $Y \in \mathcal{Y}$ is a basis of $\operatorname{span} X$ for some $X \in \mathcal{X}$ if and only if $F(X,Y) := (\mathbb{I}_m - XX^\dagger)Y = 0$.

4. *Orthogonal Tucker decomposition*: see Section 6.6.

In the first three examples, the solution Y is unique up to the isometries $\psi : Y \mapsto YQ$, where Q is any orthogonal matrix in $O(k)$.

6.4.1 Comparison to the quotient-based approach

If the solutions to a problem are invariant under a known symmetry group, they can be considered as uniquely defined points in a quotient space as opposed to a set of many solutions. For example, consider the problem of computing the eigenvector corresponding to a given simple eigenvalue of a matrix if $A \in \mathbb{C}^{n\times n}$. Depending on the precise formulation of the problem, the solution can either be considered a set of points in $\mathbb{C}^n$ or as a unique point in projective space.

For some underdetermined problems, a notion of condition has been worked out by quotienting out symmetry group of the solution set [BC13]. The fundamentals of this technique are recapped below. In the remainder of this section, we investigate whether the condition number arising from this method agrees with the latent condition number.

Suppose that $F(x, y) = c$ is an FCRE and that there exists an equivalence relation $\sim$ so that $F(x, y) = F(x, y')$ for all x if and only if $y \sim y'$. If $\pi : \mathcal{Y} \to \mathcal{Y}/{\sim}$, $y \mapsto [y]$ is the projection of a point onto its equivalence class, there exists a unique map $\widetilde{F}$ such that the following diagram commutes.

$$\begin{tikzcd}
\mathcal{X} \times \mathcal{Y} \arrow[r, "F"] \arrow[d, "\mathrm{Id}\times\pi"'] & \mathcal{Z} \\
\mathcal{X} \times (\mathcal{Y}/{\sim}) \arrow[ur, "\widetilde{F}"'] &
\end{tikzcd} \tag{6.9}$$

Under certain conditions, the projection map π and the metric in $\mathcal{Y}$ induce a Riemannian structure on $\mathcal{Y}/{\sim}$. That is, the differential $D\pi$ is a formal linear map such that every smooth function $G \circ \pi$ where G is of the form $\mathcal{Y}/{\sim} \to \mathcal{Z}$ obeys the chain rule (as in [Lee13, Theorem 4.29]) and the restriction of $D\pi$ to the orthogonal complement of its kernel is a linear isometry. In this case, π is called a *Riemannian submersion.* For example, the orbits of certain groups acting isometrically on $\mathcal{X}$ form a Riemannian manifold such that the quotient projection is a Riemannian submersion [Lee18, Theorem 2.28].

Riemannian submersions give an alternative perspective on the system $F(x, y) = c$: it can be formulated equivalently as $\widetilde{F}(x, [y]) = c$ where the goal is to solve for a representative of $[y]$. For this equation, the condition number at a point

$(x_0, [y_0])$ is given by [BC13]:

$$\kappa[\widetilde{F}](x_0, [y_0]) = \left\| \left(\frac{\partial}{\partial [y]} \widetilde{F}(x_0, [y_0]) \right)^{-1} \frac{\partial}{\partial x} \widetilde{F}(x_0, [y_0]) \right\|. \tag{6.10}$$

This can be pulled back to a more concrete expression over the original domain $\mathcal{X} \times \mathcal{Y}$, so that the derivative over the quotient space is not explicitly needed. Because of the way the metric in $\mathcal{Y}/\sim$ is defined, the above turns out to be equal to latent condition number by the following proposition.

Proposition 6.12. *Let $F(x, y) = c$ be an FCRE, where $F : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ is smooth. Let $\pi : \mathcal{Y} \to \mathcal{Y}/\sim$, $y \mapsto [y]$ be a Riemannian submersion such that* (6.9) *commutes. Assume that $\ker \frac{\partial}{\partial y} F(x, y) = \ker D\pi(y)$ and that $\frac{\partial}{\partial [y]} \widetilde{F}(x, [y])$ is invertible. Then, at every $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, we have*

$$\kappa[F^{-1}(c)](x_0, y_0) = \kappa[\widetilde{F}](x_0, [y_0])$$

where the right-hand side is given by (6.10)*.*

Proof. Define $H_0 := (\ker D\pi(y_0))^{\perp} = \left(\ker \frac{\partial}{\partial y} F(x_0, y_0)\right)^{\perp}$. By the definition of a Riemannian submersion, H_0 is isometric to $\mathcal{T}[y_0](\mathcal{Y}/\sim)$. Thus, we may write

$$\frac{\partial}{\partial [y]} \widetilde{F}(x_0, [y_0]) : H_0 \to \mathcal{T}_{F(x_0, y_0)} \mathcal{Z},$$

so that $\frac{\partial}{\partial [y]} \widetilde{F}(x, [y]) = \frac{\partial}{\partial y} F(x_0, y_0)\big|_{H_0}$. If A is a surjective linear map, then $A^{\dagger}$ is the inverse of the restriction of A to $(\ker A)^{\perp}$. Thus,

$$\left(\frac{\partial}{\partial y} F(x_0, y_0) \right)^{\dagger} = \left(\frac{\partial}{\partial [y]} \widetilde{F}(x_0, [y_0]) \right)^{-1}$$

with the identification $H_0 \cong \mathcal{T}[y_0](\mathcal{Y}/\sim)$. In addition, $\frac{\partial}{\partial x} F(x_0, y_0) = \frac{\partial}{\partial x} \widetilde{F}(x_0, [y_0])$. Combining this with Theorem 6.6 gives the desired result. □

This proposition adds a new interpretation to (6.10): *the solution map $\mathcal{X} \to \mathcal{Y}/\sim$ has the same condition number as the canonical solution map $\mathcal{X} \to \mathcal{Y}$.* The main advantage of this is that our approach does not require an explicit equivalence relation up to which the solution is defined. That is, one only needs to know that the problem is an FCRE. Moreover, the latent condition number applies to more general problems, as the quotient $\mathcal{Y}/\sim$ is not required to be a smooth manifold. Such situations can occur when attempting to quotient by a Lie group that does not act freely; this is exactly what happens when viewing tensor rank decomposition as the problem of recovering factor matrices up to permutation and scaling indeterminacies [Van17].

One manifold to which Proposition 6.12 can be applied is the *Grassmannian* of n-dimensional linear subspaces of $\mathbb{R}^m$, i.e., $\mathrm{St}(m,n)/O(n)$ where $O(n)$ is the orthogonal group. An equation $\tilde{F}(x,[y]) = c$ defining a point $[y]$ on the Grassmannian can be thought of as an underdetermined system $F(x,y) = c$ with outputs on $\mathrm{St}(m,n)$. The condition number (6.10) can be obtained by combining Proposition 6.12 and Theorem 6.6. The same conclusion holds for a general problem over the manifold of positive semidefinite $n \times n$ matrices of rank k, which is sometimes identified with $\mathbb{R}_k^{n\times k}/O(k)$ where $Y_1 \sim Y_2 \Leftrightarrow Y_1 Y_1^T = Y_2 Y_2^T$ [Jou+10].

6.5 Condition number of two-factor matrix decompositions

One of the most basic examples of an FCRE is the factorisation of a matrix X of rank k as a product $X = LR$ where L and R^T have k columns. This decomposition is used as a lift for optimisation over the set of low-rank matrices [LKB24]. It is formally defined as follows.

Definition 6.13. The *rank-revealing two-factor matrix decomposition problem* at $X \in \mathbb{R}_k^{m\times n}$ is the inverse problem $G_{\mathcal{M}}(L,R) - X = 0$ where

$$G_{\mathcal{M}} : \overbrace{\mathbb{R}_k^{m\times k} \times \mathbb{R}_k^{k\times n}}^{\mathcal{Y}} \to \mathbb{R}^{m\times n},\ (L,R) \to LR.$$

Proposition 6.14. *Let $G_{\mathcal{M}}(L,R) := LR$, where $L \in \mathbb{R}^{m\times k}$ and $R \in \mathbb{R}^{k\times n}$ have rank k. Let $\sigma_i(\cdot)$ denote the ith largest singular value of its argument if $i \leqslant k$ and $\sigma_i(\cdot) := 0$ if $i > k$. At every point (L,R), we have*

$$\kappa^{inv}[G_{\mathcal{M}}](L,R) = \frac{1}{\sqrt{\min\left\{\sigma_k(L)^2 + \sigma_n(R)^2,\ \sigma_m(L)^2 + \sigma_k(R)^2\right\}}} \tag{6.11}$$

with respect to the Euclidean inner product on $\mathbb{R}^{m\times n}$ and $\mathbb{R}^{m\times k} \times \mathbb{R}^{k\times n}$. If $k < \min\{m,n\}$, then $\kappa^{inv}[G_{\mathcal{M}}](L,R) = \min\{\sigma_k(L), \sigma_k(R)\}^{-1}$.

Proof. We will derive the condition number of this problem using Corollary 6.9. We can isometrically identify $\mathbb{R}^{m\times n} \cong \mathbb{R}^{mn}$ in the Euclidean distances on both spaces and analogously for $\mathbb{R}^{m\times k}$ and $\mathbb{R}^{k\times n}$. Then,

$$DG_{\mathcal{M}}(L,R)[\dot{L},\dot{R}] = \dot{L}R + L\dot{R} \cong \underbrace{\begin{bmatrix}\mathbb{I}_m \otimes R^T & L \otimes \mathbb{I}_n\end{bmatrix}}_{=:J} \begin{bmatrix}\dot{L} \\ \dot{R}\end{bmatrix}.$$

It remains to compute the rth largest singular value of J, where $r = \operatorname{rank}(J)$. The singular values of J are the square roots of the eigenvalues of $JJ^T = \mathbb{I}_m \otimes (R^T R) + (LL^T) \otimes \mathbb{I}_n$. This matrix is a Kronecker sum and its eigenvalues are $\lambda + \mu$ where λ and μ run over all eigenvalues of $R^T R$ and LL^T, respectively [HJ10, Theorem 4.4.5]. Therefore, all singular values of J are

$$\sigma(J) = \left\{ \sqrt{\sigma_i(L)^2 + \sigma_j(R)^2} \;\middle|\; 1 \leqslant i \leqslant m,\, 1 \leqslant j \leqslant n \right\}. \tag{6.12}$$

The number of nonzero singular values of J is thus constant for all L and R of rank k (counted with multiplicity).

By Corollary 6.9, the condition number is the reciprocal of the smallest nonzero singular value of J. An element of (6.12) is zero if and only if both $i > k$ and $j > k$. Thus,

$$\kappa^{inv}[G_{\mathcal{M}}](L, R) = \left(\min_{i \leqslant k \text{ or } j \leqslant k} \sqrt{\sigma_i(L)^2 + \sigma_j(R)^2} \right)^{-1}. \tag{6.13}$$

Since the singular values are sorted in descending order, the minimum is attained when $i = k$ or $j = k$. If it is attained for $i = k$, the right-hand side of (6.13) is $(\sigma_k(L)^2 + \sigma_n(R)^2)^{-1/2}$. Analogously, if the minimum is attained for $j = k$, we get $(\sigma_m(L)^2 + \sigma_k(R)^2)^{-1/2}$. This concludes the general case. The expression for the case where $k < \min\{m, n\}$ is obtained by substituting $\sigma_m(L) = \sigma_n(R) = 0$ in (6.11). □

Not all two-factor decompositions of a given matrix $X \in \mathbb{R}_k^{m \times n}$ have the same condition number. Therefore, one may be interested in a decomposition whose condition number is as small as possible. In the context of tensor decompositions, the *norm-balanced CPD* was introduced for the same purpose [Van17]. Intuitively, one may expect to find an optimal two-factor decomposition by computing a singular value decomposition $X = U\Sigma V^T$ and setting $L := U\Sigma^{1/2}$ and $R := \Sigma^{1/2} V^T$. This turns out to be correct, by the following lemma.

Lemma 6.15. *Suppose that $X = LR$ with $L \in \mathbb{R}_k^{n \times k}$ and $R \in \mathbb{R}_k^{k \times n}$. Then $\min\{\sigma_k(L), \sigma_k(R)\} \leqslant \sqrt{\sigma_k(X)}$.*

Proof. Let U and V be matrices whose columns are orthonormal bases of $\operatorname{span} X$ and $\operatorname{span} X^T$, respectively. If we set $(\widehat{L}, \widehat{R}, \widehat{X}) := (U^T L, RV, U^T X V)$, then the $k \times k$ matrices $\widehat{L}$, $\widehat{R}$, and $\widehat{X}$ have the same k largest singular values as L, R, and X, respectively. Suppose that $\sigma_k(\widehat{X}) = \left\| \widehat{X} v \right\|$ for some unit vector $v \in \mathbb{R}^k$, then $\sigma_k(\widehat{X}) = \left\| \widehat{L}\widehat{R} v \right\| \geqslant \sigma_k(\widehat{L}) \sigma_k(\widehat{R})$ by the Courant–Fisher theorem [HJ10, Theorem 3.1.2]. Hence, $\sigma_k(\widehat{L})$ and $\sigma_k(\widehat{R})$ cannot both be larger than $\sqrt{\sigma_k(X)}$. □

Corollary 6.16. *Let $X_0 \in \mathbb{R}_k^{m\times n}$ be any matrix and let $G_{\mathcal{M}}$ be the map from Proposition 6.14. Then, the best latent condition number of computing a two-factor matrix factorisation is*

$$\min_{L_0 R_0 = X_0} \kappa^{inv}[G_{\mathcal{M}}](L_0, R_0) = \sigma_k(X_0)^{-1/2}.$$

If $X_0 = U\Sigma V^T$ is a compact singular value decomposition[2]*, then the minimum is attained at $(L_0, R_0) = (U\Sigma^{1/2}, \Sigma^{1/2}V^T)$.*

Remark 6.17. Corollary 6.16 should *not* be interpreted as saying that the evaluation of the map $X \mapsto (U\Sigma^{1/2}, \Sigma^{1/2}V^T)$ is well-conditioned or that it refines the two-factor decomposition optimally in the sense defined in the introduction. This is clearly false, since the singular vectors of X may not even be unique. Instead, Corollary 6.16 says that, if one is interested in a two-factor decomposition (L_0, R_0) of X_0 such that rank-preserving perturbations X of X_0 have *any* decomposition close to (L_0, R_0), then $(U\Sigma^{1/2}, \Sigma^{1/2}V^T)$ is optimal.

Corollary 6.16 connects the condition number to the distance from $X \in \mathcal{X} = \mathbb{R}_k^{m\times n}$ to the boundary ∂X of $\mathcal{X}$. Since $\partial\mathcal{X}$ is the set of $m \times n$ matrices of rank strictly less than k, the Eckart–Young theorem implies that $\min_{\widehat{X}\in\partial\mathcal{X}} \left\|X - \widehat{X}\right\|_2 = \sigma_k(X)$, which is the inverse square of the condition number in Corollary 6.16. Consequently, the *ill-posed locus*, defined as the set of (limits of) inputs where the condition number diverges, is precisely the boundary of $\mathcal{X}$. For many numerical problems, there is a connection between the condition number and the reciprocal distance to the ill-posed locus, often called a *condition number theorem* [Dem87; Blu+98]. The above shows that the two-factor decomposition admits such a connection as well.

6.6 Condition number of orthogonal Tucker decompositions

As another application of the proposed theory, in this section, we study the condition number of a different rank-revealing decomposition, this time in the context of tensors. Given a tensor

$$\mathcal{C} = \sum_{i=1}^{R} v_{i,1} \otimes v_{i,2} \otimes \cdots \otimes v_{i,D} \in \mathbb{R}^{k_1\times k_2\times\cdots\times k_D},$$

[2] We call a singular value decomposition *compact* if $\Sigma \in \mathbb{R}^{k\times k}$ and $k = \operatorname{rank} X_0$.

where $v_{i,j} \in \mathbb{R}^{k_j}$ are vectors, the tensor product[3] of the matrices $U_j \in \mathbb{R}^{n_j \times k_i}, j = 1, \dots, D$, acts linearly on $\mathcal{C}$ as

$$\mathcal{X} = (U_1 \otimes \cdots \otimes U_D)\mathcal{C} = \sum_{i=1}^{R} (U_1 v_{i,1}) \otimes (U_2 v_{i,2}) \otimes \cdots \otimes (U_D v_{i,D}) . \tag{6.14}$$

The resulting tensor $\mathcal{X}$ lives in $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_D}$. When $D = 2$, $\mathcal{C}$ is a matrix and the above expression can be simplified to $U_1 \mathcal{C} U_2^T$. The *Tucker decomposition problem* [Tuc66] takes a tensor $\mathcal{X}$ as in (6.14) and asks to recover the factors $U_1, \dots, U_D, \mathcal{C}$. It is common to impose that all columns of U_i are orthonormal for each i, in which case (6.14) is sometimes called an *orthogonal Tucker decomposition.*

To formulate the problem more precisely, we introduce some notation. For the above tensor $\mathcal{C}$, the jth flattening is the matrix

$$\mathcal{C}_{(j)} = \sum_{i=1}^{R} v_{i,j} \cdot \operatorname{vec}(v_{i,1} \otimes \cdots \otimes v_{i,j-1} \otimes v_{i,j+1} \otimes \cdots \otimes v_{i,D})^T,$$

where $\operatorname{vec}(v_1 \otimes \cdots \otimes v_D)$ is the Kronecker product of the vectors $v_1, \dots, v_D$. If $\mathcal{C}$ is a matrix (i.e., $D = 2$), then $\mathcal{C}_{(1)} = \mathcal{C}$ and $\mathcal{C}_{(2)} = \mathcal{C}^T$. The *multilinear rank* of $\mathcal{C}$ is the tuple $\mu(\mathcal{C}) := (\operatorname{rank} \mathcal{C}_{(1)}, \dots, \operatorname{rank} \mathcal{C}_{(D)})$. We say that $\mathcal{C}$ has *full* multilinear rank if $\mu(\mathcal{C}) = (k_1, \dots, k_D)$. The set of such tensors is written as $\mathbb{R}_\star^{k_1 \times \cdots \times k_D}$.

Definition 6.18. The *rank-revealing orthogonal Tucker decomposition problem* at $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_D}$ is the inverse problem $G_\mathcal{T}(U_1, \dots, U_D, \mathcal{C}) - \mathcal{X} = 0$ where

$$\begin{aligned} G_\mathcal{T} : \overbrace{\mathbb{R}_\star^{k_1 \times \cdots \times k_D}, \operatorname{St}(n_1, k_1) \times \cdots \times \operatorname{St}(n_D, k_D)}^{\mathcal{Y}} &\to \mathbb{R}^{n_1 \times \cdots \times n_D} \\ (\mathcal{C}, U_1, \dots, U_D) &\mapsto (U_1 \otimes \cdots \otimes U_D)\mathcal{C}. \end{aligned}$$

The reason for considering $\mathbb{R}_\star^{k_1 \times \cdots \times k_D}$ in the domain rather than its closure $\mathbb{R}^{k_1 \times \cdots \times k_D}$ is to ensure that $\operatorname{rank} DG_\mathcal{T}$ is constant [KL10]. If $y = (\mathcal{C}, U_1, \dots, U_D)$ solves the orthogonal Tucker decomposition problem, then all other solutions can be parametrised as

$$G_\mathcal{T}^{-1}(G_\mathcal{T}(y)) = \left\{ ((Q_1 \otimes \cdots \otimes Q_D)\mathcal{C}, U_1 Q_1^T, \dots, U_D Q_D^T) \mid Q_j \in O(k_j) \right\} . \tag{6.15}$$

[3] In case of unfamiliarity with the tensor product, all occurrences of $\otimes$ may be interpreted as the Kronecker product of matrices and vectors.

In the literature on multi-factor principal component analysis, these invariances are sometimes called "rotational" degrees of freedom [KVM01].

To eliminate some degrees of freedom, it has been proposed to impose constraints on the core tensor $\mathcal{C}$. For instance, one could optimise a measure of sparsity on $\mathcal{C}$ to enhance the interpretability of the decomposition [KVM01; MVL08]. Alternatively, the *higher-order singular value decomposition* (HOSVD) [DLDMV00a] imposes pairwise orthogonality of the slices of $\mathcal{C}$. The HOSVD has the advantages of being definable in terms of singular value decompositions and giving a quasi-optimal solution to Tucker approximation problems [Hac12, Theorem 10.3]. It is unique if and only if the singular values of all $\mathcal{X}_{(i)}$ are simple for all i.

However, for these constrained Tucker decompositions, important geometric properties of the set of feasible values of $\mathcal{C}$ remain elusive. For the HOSVD, it remains unknown precisely what sets of singular values of the flattenings $\mathcal{C}_{(i)}$ are feasible [HU17]. Furthermore, for two tensors $\mathcal{C}, \mathcal{C}'$ with HOSVD constraints, the singular values of $\mathcal{C}_{(i)}$ and $\mathcal{C}'_{(i)}$ may be identical for all i even if $\mathcal{C}$ and $\mathcal{C}'$ are in distinct $O(k_1) \times \cdots \times O(k_D)$-orbits [HU17]. For these reasons, we ignore the constraints to make the orthogonal Tucker decomposition (usually) unique and study the underdetermined problem of Definition 6.18 instead.

To determine the condition number, we need a Riemannian metric for the domain and codomain of $G_{\mathcal{T}}$. A simple metric is the Euclidean or Frobenius inner product, which is defined on (the tangent spaces of) $\mathbb{R}^{n_1 \times \cdots \times n_D}$, $\mathbb{R}_\star^{k_1 \times \cdots \times k_D}$, and $\mathrm{St}(n_i, k_i) \subset \mathbb{R}^{n_i \times k_i}$. Thus, we may use the associated product metric for $\mathcal{Y}$. We call this metric of $\mathcal{Y}$ and the Euclidean inner product in $\mathbb{R}^{n_1 \times \cdots \times n_D}$ *absolute (Riemannian) metrics.* The norm induced by these metrics is the Euclidean or Frobenius norm, which we denote by $\|\cdot\|_F$.

Since the Stiefel manifold is bounded in the Euclidean metric and $\mathbb{R}_\star^{k_1 \times \cdots \times k_D}$ is not, it may be more interesting to work with *relative* metrics. For a punctured Euclidean space $\mathbb{E}\backslash\{0\}$ with inner product $\langle \cdot, \cdot \rangle$, the relative metric for two vectors $\xi, \eta \in \mathcal{T}p\mathbb{E}$ is $\frac{\langle \xi, \eta \rangle}{\langle p, p \rangle}$. Note that this defines a smooth Riemannian metric. We lift this definition so that the relative metric in $\mathcal{Y}$ is the product metric of the relative metric in $\mathbb{R}_\star^{k_1 \times \cdots \times k_D}$ and the Frobenius inner products on all $\mathrm{St}(n_i, k_i)$.

Proposition 6.19. *Let $(U_1 \otimes \cdots \otimes U_D)\mathcal{C}$ be an orthogonal Tucker decomposition of a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_D}$ such that $\mathcal{C} \in \mathbb{R}_\star^{k_1 \times \cdots \times k_D}$ and $k_i < n_i$ for at least one i. Let $\sigma := \min_{i:\, k_i < n_i} \sigma_{k_i}(\mathcal{C}_{(i)})$. Then,*

1. *$\kappa^{inv}[G_{\mathcal{T}}](\mathcal{C}, U_1, \ldots, U_D) = \max\left\{\frac{1}{\sigma}, 1\right\}$ for the absolute metric, and*

2. $\kappa^{inv}[G_{\mathcal{T}}](\mathcal{C}, U_1, \ldots, U_D) = \frac{\|\mathcal{X}\|_F}{\sigma}$ *for the relative metric.*

Proof. Throughout this proof, we abbreviate $DG_{\mathcal{T}}(\mathcal{C}, U_1, \ldots, U_D)$ to $DG_{\mathcal{T}}$. Following Corollary 6.9, we compute the smallest nonzero singular value of $DG_{\mathcal{T}}$ for both metrics. The proof consists of a construction of this matrix with respect to an orthonormal basis and a straightforward calculation of its singular values. To use the same derivation for the two metrics, we let $\alpha = 1$ for the absolute metric and $\alpha = \|\mathcal{C}\|_F = \|\mathcal{X}\|_F$ for the relative metric.

For all i, let $\{\Omega_i^j \mid 1 \leqslant j \leqslant \frac{1}{2}k_i(k_i-1)\}$ be an orthonormal basis of the $k_i \times k_i$ skew-symmetric matrices. Let $U_i^\perp \in \mathrm{St}(n_i, n_i - k_i)$ be any matrix so that $[U_i \quad U_i^\perp]$ is orthogonal. If $n_i = k_i$, we write $U_i^\perp$ formally as an $n_i \times 0$ matrix. Let $\{V_i^p \mid 1 \leqslant p \leqslant (n_i - k_i)k_i\}$ be a basis of $\mathbb{R}^{(n_i-k_i)\times k_i}$. Then, $\mathcal{B}_i := \{U_i\Omega_i^j + U_i^\perp V_i^p\}$ forms an orthonormal basis of $\mathcal{T}U_i\mathrm{St}(n_i, k_i)$. If $\{E_l \mid 1 \leqslant l \leqslant \prod_{i=1}^D k_i\}$ is the canonical basis of $\mathbb{R}^{k_1\times\cdots\times k_D}$, then $\mathcal{B}_0 := \{\alpha E_l\}$ is an orthonormal basis of $\mathbb{R}^{k_1\times\cdots\times k_D}$. The product of these bases gives a canonical orthonormal basis for $\mathcal{Y}$.

Similarly, an orthonormal basis of $\mathbb{R}^{n_1\times\cdots\times n_D}$ is $\{\alpha\hat{E}_j\}$ where the $\hat{E}_j$ are the canonical basis vectors of $\mathbb{R}^{n_1\times\cdots\times n_D}$. In other words, expressing a vector in orthonormal coordinates is equivalent to division by α.

Next, we compute the differential of $G_{\mathcal{T}}$. For general tangent vectors $\dot{\mathcal{C}} \in \mathcal{T}\mathcal{C}\mathbb{R}_\star^{k_1\times\cdots\times k_D}$ and $\dot{U}_i \in \mathcal{T}U_i\mathrm{St}(n_i, k_i)$, we have, in coordinates,

$$DG_{\mathcal{T}}[\dot{\mathcal{C}}, 0, \ldots, 0] = \alpha^{-1}(U_1 \otimes \cdots \otimes U_D)\dot{\mathcal{C}} \quad \text{and}$$

$$DG_{\mathcal{T}}[0, \ldots, \dot{U}_i, \ldots, 0] = \alpha^{-1}(U_1 \otimes \cdots \otimes U_{i-1} \otimes \dot{U}_i \otimes U_{i+1} \otimes \cdots \otimes U_D)\mathcal{C},$$

which is extended linearly for all tangent vectors. The condition that $U_i^T U_i^\perp = 0$ for all i splits the image of DG into pairwise orthogonal subspaces. That is, for any i and k, $DG_{\mathcal{T}}[0, \ldots, U_i^\perp V_i^p, \ldots, 0]$ is orthogonal to both $DG_{\mathcal{T}}[\dot{\mathcal{C}}, 0, \ldots, 0]$ for all $\dot{\mathcal{C}}$ and $DG_{\mathcal{T}}[0, \ldots, \dot{U}_{i'}, \ldots, 0]$ for all $\dot{U}_{i'}$ where $i' \neq i$.

To decompose the domain of $DG_{\mathcal{T}}$ as a direct sum of pairwise orthogonal subspaces, we write $\mathcal{T}U_i\mathrm{St}(n_i, m_i) = W_i \oplus W_i^\perp$ for all i, where $W_i := \mathrm{span}\{U_i\Omega_i^j\}$ and $W_i^\perp = \mathrm{span}\{U_i^\perp V_i^p\}$. The restriction of $DG_{\mathcal{T}}$ to $\mathcal{T}_{\mathcal{C}}\mathbb{R}_\star^{k_1\times\cdots\times k_D} \times W_1 \times \cdots \times W_D$ can be represented in coordinates by a matrix J_0. Likewise, for $i = 1, \ldots, D$, we write the restriction of $DG_{\mathcal{T}}$ to $\{0\} \times \cdots \times W_i^\perp \times \cdots \times \{0\}$ in coordinates as J_i. Then $DG_{\mathcal{T}}$ can be represented as $J = \begin{bmatrix} J_0 & J_1 & \ldots & J_D \end{bmatrix}$.

By the preceding argument, these $D + 1$ blocks that make up J are pairwise orthogonal. Therefore, the singular values of J are the union of the singular values of $J_0, \ldots, J_D$. If $n_i = k_i$ for some i, then J_i is the matrix with zero columns, whose singular values are the empty set.

Let $\Pi := \prod_{i=1}^{D} k_D$. To bound the singular values of J_0, we compute its first Π columns as $DG_{\mathcal{T}}[\alpha E_l, 0, \ldots, 0] = (U_1 \otimes \cdots \otimes U_D)E_l$ for all $l = 1, \ldots, \Pi$. Note that all other columns have a factor $(U_1 \otimes \cdots \otimes U_D)$ as well. Thus, we have

$$J_0 = (U_1 \otimes \cdots \otimes U_D) \begin{bmatrix} \mathbb{I}_{\Pi} & \tilde{J} \end{bmatrix},$$

where $\tilde{J}$ is an unspecified matrix. We can omit the orthonormal factor $U_1 \otimes \cdots \otimes U_D$ when computing the singular values of J_0. Therefore, J_0 has Π nonzero singular values, which are the square roots of the eigenvalues of $\mathbb{I}_{\Pi} + \tilde{J}\tilde{J}^T$. It follows that the Π largest singular values of J_0 are bounded from below by 1 and all other singular values of J_0 are 0.

Next, consider any J_i where $i \geqslant 1$ and $n_i > k_i$. It represents the linear map

$$J_i : V \mapsto \alpha^{-1}(U_1 \otimes \cdots \otimes U_{i-1} \otimes U_i^{\perp} V \otimes U_{i+1} \otimes \cdots \otimes U_D)\mathcal{C}.$$

Up to reshaping, the above is equivalent to

$$J_i : \operatorname{vec} V \mapsto \alpha^{-1} \left(U_i^{\perp} \otimes ((U_1 \otimes \cdots \otimes U_{i-1} \otimes U_{i+1} \otimes \cdots \otimes U_D)\mathcal{C}_{(i)}^T) \right) \operatorname{vec} V.$$

To calculate the singular values of J_i, we factor out $U_i^{\perp}$ and all U_j to obtain $J_i \cong \alpha^{-1}\mathbb{I}_{n_i - k_i} \otimes \mathcal{C}_{(i)}^T$. Its singular values are the singular values of $\alpha^{-1}\mathcal{C}_{(i)}$ with all multiplicities multiplied by $n_i - k_i$.

Finally, we show geometrically that the smallest nonzero singular value ς of $DG_{\mathcal{T}}$ is at most 1. By the Courant–Fisher theorem, $\varsigma \leqslant \|DG_{\mathcal{T}}[\xi]\| / \|\xi\|$ for all $\xi \in \ker(DG_{\mathcal{T}})^{\perp} \backslash \{0\}$. Pick $\xi := (0, \ldots, 0, \mathcal{C})$. Since

$$G_{\mathcal{T}}^{-1}(\mathcal{X}) = \left\{ ((Q_1 \otimes \cdots \otimes Q_D)\mathcal{C}, U_1 Q_1^T, \ldots, U_D Q_D^T) \mid Q_i \in O(k_i) \right\},$$

the projection of $G_{\mathcal{T}}^{-1}(\mathcal{X})$ onto the first component is a submanifold of the sphere over $\mathbb{R}^{k_1 \times \cdots \times k_D}$ of radius $\|\mathcal{C}\|_F$. It follows that

$$\xi \in N_{(\mathcal{C}, U_1, \ldots, U_D)} G_{\mathcal{T}}^{-1}(\mathcal{X}) = (\ker DG_{\mathcal{T}})^{\perp},$$

where N denotes the normal space. Since $DG_{\mathcal{T}}[\xi] = (U_1 \otimes \cdots \otimes U_D)\mathcal{C}$, we obtain $\varsigma \leqslant \|DG_{\mathcal{T}}[\xi]\| / \|\xi\| = 1$.

In conclusion, we have established the following three facts. First, for all i such that $k_i < n_i$, all singular values of $\alpha^{-1}\mathcal{C}_{(i)}$ are singular values of $DG_{\mathcal{T}}$. Second, any other nonzero singular values of $DG_{\mathcal{T}}$ must be bounded from below by 1. Third, the smallest nonzero singular value of $DG_{\mathcal{T}}$ is at most 1. By Corollary 6.9, this proves the statement about the absolute metric. For the relative metric, note that $\sigma \leqslant \|\mathcal{C}\|_F = \|\mathcal{X}\|_F = \alpha$ for all i. Thus, the smallest nonzero singular value of $DG_{\mathcal{T}}$ is $\sigma / \alpha = \sigma / \|\mathcal{X}\|_F \leqslant 1$. □

The expression for the absolute metric can be interpreted as follows: if $k_i < n_i$ and $\sigma_{(i)} := \sigma_{k_i}(\mathcal{C}_{(i)})$ is small, then the factor U_i is sensitive to perturbations of $\mathcal{X}$. Indeed, assume that the last column of U_i is a left singular vector corresponding to the singular value $\sigma_{(i)}$. Then, we can generate the following small perturbation of $\mathcal{X}$ that corresponds to a unit change in the decomposition. Let $\tilde{U}_i$ be a matrix such that $U_i e_j = \tilde{U}_i e_j$ for $1 \leqslant j < k_1$ and $U_i^T \tilde{U}_i e_{k_i} = 0$. The perturbed tensor $\widetilde{\mathcal{X}} = G_{\mathcal{T}}(\mathcal{C}, U_1, \ldots, U_{d-1}, \tilde{U}_i, U_{d+1}, \ldots, U_D)$ is only at a distance $\sigma_{(i)}$ away from $\mathcal{X}$.

On the other hand, if $\sigma \geqslant 1$, then no unit perturbation of $\mathcal{X}$ tangent to $G_{\mathcal{T}}(\mathcal{Y})$ can change the orthogonal Tucker decomposition more than the tangent vector $\Delta\mathcal{X} = \mathcal{X}/\|\mathcal{X}\|_F$ constructed in the proof of Proposition 6.19.

Remark 6.20 (Condition number of a singular value decomposition)**.** For a matrix X, computing an orthogonal Tucker decomposition of the form $X = U_1 S U_2^T$ is a relaxation of the singular value decomposition that does not impose a diagonal structure on S. A condition number for the subspaces spanned by the singular vectors was studied in [Sun96; Van23]. The condition number for the ith singular vector diverges as $|\sigma_i(X) - \sigma_{i+1}(X)| \to 0$. By contrast, the condition number of computing an orthogonal Tucker decomposition depends on $\sigma_{k_1}(X)$ and $\|X\|_F$ only. Thus, computing individual singular vectors may be arbitrarily ill-conditioned even if the condition number of the Tucker decomposition is arbitrarily close to one. For instance, this occurs for the 3×3 diagonal matrix X whose diagonal elements are $(1+\varepsilon, 1, 0)$ and $0 < \varepsilon \ll 1$. Informally, this observation shows that restricting S to be diagonal in an orthogonal Tucker decomposition can make computing the resulting singular value decomposition arbitrarily more ill-conditioned than computing an orthogonal Tucker decomposition.

6.7 Numerical verification of the error estimate

Eq. (6.5) is only an asymptotic estimate of the optimal forward error. A common practice for working with condition numbers is to neglect the asymptotic term $o(d_{\mathcal{X}}(x_0, x))$ and turn (6.5) into the approximate upper bound

$$\min_{\substack{y \in \mathcal{Y}, \\ F(x,y)=c}} d_{\mathcal{Y}}(y_0, y) \lesssim \kappa[F^{-1}(c)](x_0, y_0) \cdot d_{\mathcal{X}}(x_0, x). \tag{6.16}$$

In this section, we determine numerically if this approximation is accurate for random initial solutions (x_0, y_0) and random perturbations x. We restrict ourselves to the orthogonal Tucker decomposition of third-order tensors.

6.7.1 Model

The model is defined as follows. We pick a parameter $\alpha > 0$ to control the condition number. We generate matrices $A, B \in \mathbb{R}^{k\times(k-1)}$ and a tensor $\mathcal{H} \in \mathbb{R}^{k\times k\times k}$ with i.i.d. standard normally distributed entries. Then, we set $\mathcal{C}' \in \mathbb{R}^{k\times k\times k}$ so that $\mathcal{C}'_{(1)} = (AB^T + \alpha\mathbb{I})\mathcal{H}_{(1)}$ and normalise $\mathcal{C}_0 := \mathcal{C}'/\|\mathcal{C}'\|_F$. Finally, we generate the factor U_i^0 by taking the Q-factor in the QR decomposition of a normally distributed $n_i \times k$ matrix, for $i = 1, 2, 3$. To the above matrices and tensor, we associate the Tucker decomposition $\mathcal{X}_0 = G_{\mathcal{T}}(\mathcal{C}_0, U_1^0, U_2^0, U_3^0)$. In the following, we abbreviate $\kappa := \kappa^{inv}[G_{\mathcal{T}}](\mathcal{C}_0, U_1^0, U_2^0, U_3^0)$. Note that choosing a small value of α generally makes the problem ill-conditioned by Proposition 6.19.

The condition number measures the change to the decomposition for *feasible* perturbations of $\mathcal{X}_0 \in \mathbb{R}^{n_1\times n_2\times n_3}$. That is, the perturbed input is a point $\mathcal{X}$ in the image of $G_{\mathcal{T}}$. Such a point can be generated by first perturbing $\mathcal{X}_0$ in the ambient space $\mathbb{R}^{n_1\times n_2\times n_3}$ as $\mathcal{X}' := \mathcal{X}_0 + \varepsilon\Delta\mathcal{X}$ where $\Delta\mathcal{X}$ is uniformly distributed over the unit-norm tensors in $\mathbb{R}^{n_1\times n_2\times n_3}$ and $\varepsilon > 0$ is some parameter. Then, a feasible input can be obtained by applying the ST-HOSVD algorithm [VVM12] to $\mathcal{X}'$ with truncation rank (k, k, k). This gives a quasi-optimal projection $\mathcal{X}$ of $\mathcal{X}'$ onto the image of $G_{\mathcal{T}}$ and a decomposition $\mathcal{X} = G_{\mathcal{T}}(\mathcal{C}, U_1, U_2, U_3)$.

6.7.2 Estimate of the optimal forward error

Since $\|\mathcal{X}_0\|_F = \|\mathcal{C}_0\|_F = 1$, the absolute and relative metric in Proposition 6.19 coincide, and they both correspond to the product of the Euclidean inner products in $\mathbb{R}^{k_1\times k_2\times k_3}$ and all $\mathrm{St}(n_i, k_i)$. By (6.15), the orthogonal Tucker decomposition of $\mathcal{X}$ is determined up to a multiplication by $Q_1, Q_2, Q_3 \in O(k)$. Thus, the square of the left-hand side of (6.16) is[4]

$$E^2 := \min_{Q_1,Q_2,Q_3\in O(k)} \left\{ \left\|\mathcal{C}_0 - (Q_1^T \otimes Q_2^T \otimes Q_3^T)\mathcal{C}\right\|_F^2 + \sum_{i=1}^{3} \left\|U_i^0 - U_iQ_i\right\|_F^2 \right\}. \tag{6.17}$$

Determining the accuracy of (6.16) requires evaluating E numerically. Since we are not aware of any closed-form expression of E, we approximate this by solving the above optimisation problem using a simple Riemannian gradient descent method in `Manopt.jl` [Ber22]. In the following, $\widehat{E}^2$ denotes the numerical solution to (6.17).

[4] Although the definition of the condition number uses the geodesic distance d_γ, the following expression uses the Euclidean distance d_E. However, it can be shown that for a point $x_0 \in \mathcal{X}$ where $\mathcal{X}$ is a Riemannian submanifold of Euclidean space, we have $d_E(x_0, x) = d_\gamma(x_0, x)(1 + o(1))$ as $x \to x_0$.

Assuming that $\widehat{E}$ is an accurate approximation of E, we check (6.16) by verifying that $\widehat{E} \lesssim \kappa \|\mathcal{X} - \mathcal{X}_0\|_F$. A priori, there are at least two scenarios in which this may fail to be a tight upper bound:

1. The tolerance of the gradient descent method that computes $\widehat{E}$ is 5×10^{-8}, so that the numerical error in computing E is about the same order of magnitude. If $E \ll 10^{-8}$, then $\widehat{E}$ is probably a poor approximation of E.

2. E cannot be much larger than 1. Since $\mathrm{St}(n_i, k)$ is a subset of the $n_i \times k$ matrices with Frobenius norm equal to $\sqrt{k}$, we have $\left\|U_i^0 - U_i\right\|_F \leqslant 2\sqrt{k}$. Furthermore, since $\|\mathcal{C}\|_F = \|\mathcal{X}\|_F$ and $\|\mathcal{C}_0\|_F = \|\mathcal{X}_0\|_F = 1$, it follows from the triangle inequality that $\|\mathcal{C} - \mathcal{C}_0\|_F \leqslant 1 + \|\mathcal{X}\|_F$. Thus, if $\kappa\|\mathcal{X} - \mathcal{X}_0\|_F \gg 1$, this would overestimate E.

For these reasons, we are only interested in verifying the estimate $\widehat{E} \lesssim \kappa\|\mathcal{X} - \mathcal{X}_0\|_F$ if $\widehat{E} \geqslant 5 \times 10^{-8}$ and $\kappa\|\mathcal{X} - \mathcal{X}_0\|_F \leqslant 1$.

6.7.3 Experimental results

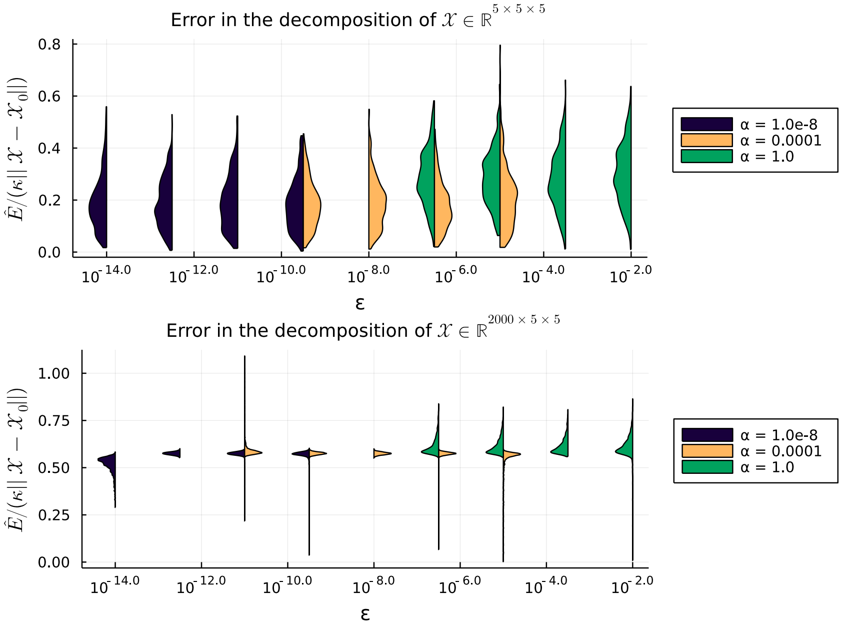
We generated two datasets as specified by the model above. In the first dataset, we used the parameters $k = 3$ and $(n_1, n_2, n_3) = (5, 5, 5)$. For each pair $(\alpha, \varepsilon) \in \{10^{-8}, 10^{-4}, 1\} \times \{10^{-14}, 10^{-12.5}, \ldots, 10^{-2}\}$, we generated 2000 Tucker decompositions and perturbations and measured the error. The second dataset was generated the same way, the only difference being that $n_1 = 2000$.

Since the condition number depends only on $\mathcal{C}_0$, the distribution of the condition number is the same for both datasets. We found that κ is approximately equal to $10/\alpha$, with $1 \leqslant \kappa\alpha \leqslant 100$ in 94.5% of samples. The empirical geometric mean of $\kappa\alpha$ is about 12. This means that we can roughly control the condition number of the sampled tensor by controlling the parameter α.

Figure 6.2 shows the distribution of $\widehat{E}/\kappa\|\mathcal{X} - \mathcal{X}_0\|_F$ for both datasets. The smaller this quantity, the more pessimistic $\kappa\|\mathcal{X} - \mathcal{X}_0\|_F$ is as an estimate of the forward error for random perturbed tensors $\mathcal{X}$. In most cases displayed on Figure 6.2, $\widehat{E}$ is at least a fraction 0.1 of its approximate upper bound $\kappa\|\mathcal{X} - \mathcal{X}_0\|_F$. In the case where $2000 = n_1 \gg k = 3$, we have $\widehat{E} \approx 0.5\kappa\|\mathcal{X} - \mathcal{X}_0\|_F$. These experiments indicate that the approximation $E \lesssim \kappa\|\mathcal{X} - \mathcal{X}_0\|$ is reasonably sharp on average.

The estimate $E \lesssim \kappa\|\mathcal{X} - \mathcal{X}_0\|$ could be sharpened by using a *stochastic condition number*, which estimates the forward error corresponding to uniform random perturbations $\mathcal{X}$ on a sphere around $\mathcal{X}_0$ rather than worst-case perturbations. It was shown in [Arm10] that the stochastic condition number of a map $H : \mathcal{X} \to \mathcal{Y}$

could be as low as $\kappa/\sqrt{\dim \mathcal{X}}$ where κ is the condition number. The idea is that there may only be one bad perturbation direction in $\mathcal{X}$, which is unlikely to manifest in practice if $\mathcal{X}$ is high-dimensional. In our case, though, the error is empirically closer to the worst case in the high-dimensional experiment ($n_1 = 2000$) than in the low-dimensional one ($n_1 = 5$). This is evidence that, for the decomposition of large tensors of low multilinear rank, the ill-conditioned perturbation directions fill up more of the space. A full stochastic analysis is beyond the scope of this thesis.


Error in the decomposition of $\mathcal{X} \in \mathbb{R}^{5\times5\times5}$

Error in the decomposition of $\mathcal{X} \in \mathbb{R}^{2000\times5\times5}$

Figure 6.2: Distribution of $\frac{\widehat{E}}{\kappa\|X - X_0\|_F}$ for a Tucker decomposition of the perturbed tensor X. An estimate of the probability density is plotted for all combinations of (α, ε) such that 90% of samples satisfy $\widehat{E} \geqslant 5 \times 10^{-8}$ and 90% of samples satisfy $\kappa\|X - X_0\|_F \leqslant 1$.

6.8 Conclusion

In this chapter, we proposed a theory of condition for a general class of (potentially underdetermined) systems of equations, which we call FCREs. The *latent condition number* measures the asymptotic behaviour in the error in a least-squares sense. Specifically, the latent condition number estimates the smallest change in the solution for the worst possible perturbation. Our definition is an extension of an earlier definition based on a quotient-based formulation of the problem.

The proposed theory can be used to explain why a system of equations is ill-conditioned: if a problem $\mathcal{P}$ can be relaxed to a less constrained problem $\mathcal{P}'$ that has a high latent condition number, this gives a lower bound for the condition number of $\mathcal{P}$. In other words: if solving $\mathcal{P}$ *requires* solving an ill-conditioned relaxation $\mathcal{P}'$, then $\mathcal{P}$ is ill-conditioned.

The expression for the condition number can be simplified for the problems of two-factor matrix decomposition and Tucker decomposition. In both cases, the condition number can be expressed in terms of the singular values of (matrix unfoldings of) the factors. This confirms the common intuition that the decomposition problem is ill-conditioned insofar as the factors have a small singular value.

Chapter 7

Which variables of a numerical problem are ill-conditioned?

This chapter consists of my own original work. It has not been submitted as a separate article.

Abstract

In the previous chapter, we looked at systems of equations $F(x,y) = c$ and formulated a concept of condition that is applicable to underdetermined systems. This allowed us to quantify which equations in the system cause the problem to be ill-conditioned. In this chapter, we build on the theory from the previous chapter to answer a dual question: if $y = (y_1, \ldots, y_n)$ is a tuple of several variables, how sensitive is each component y_i to perturbations in x? In conjunction with Chapter 6, the results of this chapter make it possible to isolate the constraints of the system $F(x,y) = c$ that make it ill-conditioned, as well as the solution variables $y_1, \ldots, y_n$ that cause those constraints to be ill-conditioned.

7.1 Introduction

All numerical problems we have studied thus far involve an input variable x and a solution variable y whose relationship can be expressed by a known equation, say $F(x, y) = c$. However, not all interesting problems come in this form, especially those involving *hidden variables.* This means that the equation that is to be solved contains unknown variables z which are neither the input x nor the solution y. That is, we have an equation $F(x, y, z) = c$ where $F\colon \mathcal{X}\times\mathcal{Y}\times\mathcal{Z} \to \mathcal{W}$ is any map, $x \in \mathcal{X}$ is given and a value of $y \in \mathcal{Y}$ is needed such that $F(x, y, z) = c$ for some $z \in \mathcal{Z}$.

To define a condition number that expresses the sensitivity of y with respect to x and is consistent with existing theory of condition, one may consider one of following two avenues:

- *Algebraic approach:* We attempt to convert the equation $F(x, y, z) = c$ into an equivalent equation $\widetilde{F}(x, y) = c'$. That is, $F(x, y) = c'$ if and only if there exists a value of z such that $F(x, y, z) = c$. If F is linear, such an equation $\widetilde{F}$ can be found by Gaussian elimination. If F is polynomial, the same can be achieved with symbolic algorithms based on Gröbner bases [CLO07, Chapter 2]. Then the existing theory of condition such as that of Chapter 6 can be applied to the equation $\widetilde{F}(x, y) = c'$. A disadvantage of this approach is that the relationship between F and $\widetilde{F}$ may be complicated in general.
- *Geometric approach:* We can describe the given equation geometrically as its graph $\mathcal{P} := F^{-1}(c) \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. The variable z can be eliminated out of the equation by projecting $\mathcal{P}$ onto $\mathcal{X} \times \mathcal{Y}$. The connections between this viewpoint and the algebraic one are central to elimination theory [CLO07, Chapter 3]. Then the usual theory of condition can be applied to the projection $\pi_{\mathcal{X}\times\mathcal{Y}}(\mathcal{P})$, assuming that defining equations for this set can be found.

The geometric perspective provides a straightforward extension to the prior theory of condition if the projection of $\mathcal{P}$ onto $\mathcal{X}$ is an immersion. In this case, it follows from the results in Section 2.3 (specifically Lemma 2.8) that every point $x \in \mathcal{X}$ that is sufficiently close to x_0 corresponds to a (locally) unique pair (y, z) that solves $F(x, y, z) = c$. Put simply, y and z would be functions of x. In this case, the condition number of solving for y would be unambiguously defined as the condition number of the map $x \mapsto y$. More formally, one would define the following.

Definition 7.1 (special case of Definition 7.4)**.** Let $F : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathcal{W}$ be a smooth map between Riemannian manifolds and assume that a level set $\mathcal{P} := \{(x, y, z) \mid F(x, y, z) = c\}$ is an embedded submanifold of $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. If the projection $\pi_{\mathcal{X}} : \mathcal{P} \to \mathcal{X}$ has a local inverse $\pi_{\mathcal{X}}^{-1}$ at $(x_0, y_0, z_0) \in \mathcal{P}$, the *condition number of* y *at* (x_0, y_0, z_0) is

$$\kappa_{x \mapsto y}[\mathcal{P}](x_0, y_0, z_0) := \kappa[\pi_{\mathcal{Y}} \circ \pi_{\mathcal{X}}^{-1}](x_0, y_0, z_0).$$

Whilst this would be a natural extension of the known theory (Chapter 2) and it is probably known to the experts, I have not found this concept in the literature.

The above definition is not applicable if solving for (y, z) is an underdetermined problem, i.e., if every $x \in \mathcal{X}$ corresponds to a positive-dimensional manifold of points $(y, z) \in \mathcal{Y} \times \mathcal{Z}$ which all solve $F(x, y, z) = c$. Since the previous chapter showed the usefulness of the condition number of underdetermined problems, we would be remiss to limit our study to only those problems where the straightforward definition 7.1 applies. Therefore, the focus of this chapter is to formulate a theory of condition for equations $F(x, y, z) = c$ where (y, z) is not assumed to be locally unique given x. The problems we consider for this are the following variant of the FCRE model.

Definition 7.2. Let $F\colon \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathcal{W}$ be a smooth map and let c be a constant in the image of F. Suppose that

- $\operatorname{rank} DF(x, y, z) = \operatorname{rank} \frac{\partial}{\partial(y,z)} F(x, y, z) = r$ everywhere for some $r \in \mathbb{N}$,
- $\operatorname{rank} \frac{\partial}{\partial z} F(x, y, z) = k$ everywhere for some $k \in \mathbb{N}$.

Then the equation $F(x, y, z) = c$ is a *constant-rank elimination problem (CREP)*. The spaces $\mathcal{X}$ and $\mathcal{Y}$ are the *input* and *output* space, respectively.

If we apply the geometric approach to assigning a condition number to CREPs, one issue is that an understanding of the geometry of $\mathcal{P} := F^{-1}(c)$ in a neighbourhood of a point $(x_0, y_0, z_0) \in \mathcal{P}$ may not suffice in order to describe the local geometry of $\pi_{\mathcal{X}\times\mathcal{Y}}(\mathcal{P})$ at (x_0, y_0). For example, the projection of a smooth manifold may be singular, as illustrated by the examples of [Smi+00, §7.1].

For this reason, we define the condition number in terms of the projection of a *neighbourhood* of a solution tuple (x_0, y_0, z_0). This leads to the following theorem (which is proven in Section 7.3) and corresponding definition.

Theorem 7.3. *Let* $F(x, y, z) = c$ *be a CREP with* $F\colon \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathcal{W}$ *and let* (x_0, y_0, z_0) *be a solution. Then for sufficiently small neighbourhoods* $\widehat{\mathcal{X}}, \widehat{\mathcal{Y}}$, *and*

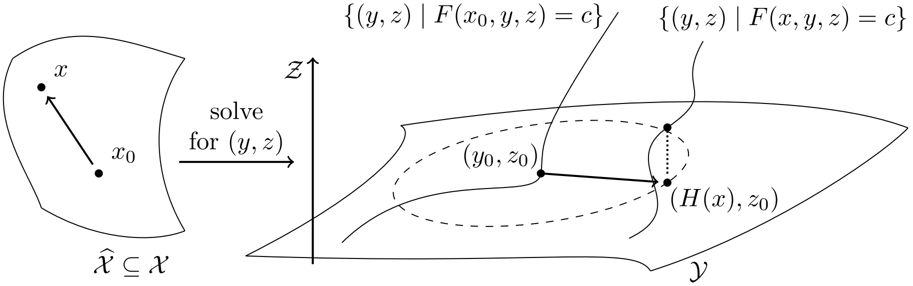

$\{(y,z) \mid F(x_0,y,z)=c\}$
$\{(y,z) \mid F(x,y,z)=c\}$
x
x_0
solve for (y,z)
$\mathcal{Z}$
(y_0,z_0)
$(H(x),z_0)$
$\widehat{\mathcal{X}} \subseteq \mathcal{X}$
$\mathcal{Y}$

Figure 7.1: Solution sets of a CREP and the canonical solution map H. Given an exact and perturbed input x_0 and x, the solution sets for both inputs are subsets of $\mathcal{Y} \times \mathcal{Z}$ (in this case, curves). For sufficiently small neighbourhoods $\widehat{\mathcal{Z}} \subseteq \mathcal{Z}$ of z_0, the solution map can be visualised as follows. For a time parameter $t \geqslant 0$, define the cylinder $C(t) := B(t) \times \widehat{\mathcal{Z}}$ where $B(t)$ is the closed disc around y_0 of radius t. Let t increase from 0 until $C(t)$ touches the solution set of x at some point (y,z). The y-coordinate of this unique point is $H(x)$ by definition. Projecting the right side of the figure onto $\mathcal{Y}$ retrieves the FCRE model.

$\widehat{\mathcal{Z}}$ *of* x_0, y_0*, and* z_0*, respectively, the set*

$$\widetilde{\mathcal{P}} := \left\{(x,y) \in \widehat{\mathcal{X}} \times \widehat{\mathcal{Y}} \,\middle|\, \exists z \in \widehat{\mathcal{Z}} \colon F(x,y,z) = c\right\} \tag{7.1}$$

is the zero set of some FCRE $\widetilde{F}(x,y) = 0$ *where* $\widetilde{F} \colon \widehat{\mathcal{X}} \times \widehat{\mathcal{Y}} \to \mathbb{R}^{\dim(\mathcal{X}\times\mathcal{Y})}$. *Moreover, if* $\mathcal{Y}$ *is Riemannian, then the derivative of the canonical solution map of* $\widetilde{\mathcal{P}}$ *at* x_0*, as defined by Theorem 6.6, is the unique matrix* $DH(x_0)$ *that satisfies the linear system*

$$\begin{cases} (\dot{x}, DH(x_0)[\dot{x}]) \in D\pi_{\mathcal{X}\times\mathcal{Y}}\,[\ker DF] & \textit{for all} \quad \dot{x} \in \mathcal{T}_{x_0}\mathcal{X} \\ \operatorname{span} DH(x_0) \perp D\pi_{\mathcal{Y}} \left[\ker \frac{\partial F}{\partial(y,z)}\right] \end{cases} \tag{7.2}$$

in which all derivatives are evaluated at (x_0, y_0, z_0) *or its projections.*

This theorem is visualised in Figure 7.1, for the specific case where $\frac{\partial F}{\partial z}$ has full rank. If $\dim \mathcal{Z} > 1$ and $\frac{\partial F}{\partial z}$ does not have full rank, the solution sets $\{(y,z) \mid F(x,y,z) = c\}$ have a higher dimension than their projections onto $\mathcal{Y}$.

The preceding theorem ensures that we may define the condition number of CREPs in terms of the condition number from Chapter 6.

Definition 7.4. If $F(x,y,z) = c$ is a CREP whose input and output space are Riemannian manifolds and (x_0, y_0, z_0) is any solution, the *condition number of*

y in $F^{-1}(c)$ is

$$\kappa_{x\mapsto y}[F^{-1}(c)](x_0,y_0,z_0) := \kappa[\widetilde{\mathcal{P}}](x_0,y_0),$$

where the right-hand side is the condition number from Definition 6.7 applied to (7.1). The condition number of z is defined by reversing the roles of y and z. That is, if $\overline{F}(x,z,y)=c$ is a CREP where $\overline{F}\colon (x,z,y)\mapsto F(x,y,z)$, then

$$\kappa_{x\mapsto z}[F^{-1}(c)](x_0,y_0,z_0) := \kappa_{x\mapsto z}[\overline{F}^{-1}(c)](x,z,y).$$

By Theorems 6.6 and 7.3, the condition number in Definition 7.4 is equal to the operator norm of $DH(x_0)$, i.e., the matrix defined by (7.2). Section 7.4 gives two ways of computing $DH(x_0)$ that are more concrete than (7.2).

Similarly to (6.5), the condition number bounds the optimal forward error as

$$\min_{\substack{y\in\mathcal{Y},\, z\in\widehat{\mathcal{Z}}\\ F(x,y,z)=c}} d(y_0,y) \leqslant \kappa_{x\mapsto y}[F^{-1}(c)](x_0,y_0,z_0)\cdot d(x_0,x)+o(d(x_0,x)) \tag{7.3}$$

where $\widehat{\mathcal{Z}}$ is some neighbourhood of z_0. The left-hand side can be interpreted as the optimal forward error $d(y_0,y)$ that can be attained with a solution (y,z) close to (y_0,z_0). It should be noted that closeness to z_0 is defined topologically rather than metrically and that a distance on $\mathcal{Z}$ is not required to define the condition number. Whilst this may seem unintuitive, this is ultimately because the condition number is a local (infinitesimal) property of the CREP and because measuring errors the z-coordinate is irrelevant by assumption. Visually, we can picture the solution curve of $F(x,y,z)=c$ on Figure 7.1 as merging into the solution curve of $F(x_0,y,z)=c$ as x approaches x_0. The distance we keep track of (i.e., the left-hand side of (7.3)) is measured only in the y-coordinate.

In the previous chapter, we established that if one removes constraints from an FCRE, the condition number of this less constrained problem is a lower bound for the condition number of the original problem. In this chapter, we are not interested in solving a subset of the *equations*, but rather in solving for a subset of the *variables*. This raises a natural question: if we solve a system $F(x,y,z)=c$ for y, is this a more well-conditioned problem than solving for y and z combined? The answer is affirmative, as the following proposition shows.

Proposition 7.5. *Let $F(x,y,z)=c$ be a CREP with a Riemannian input and output space and a solution (x_0,y_0,z_0). Consider the equation $\overline{F}(x,(y,z))=c$ with output space $\mathcal{Y}\times\mathcal{Z}$, where $\overline{F}$ is defined by $\overline{F}(x,(y,z)) := F(x,y,z)$. Endow $\mathcal{X},\mathcal{Y}$, and $\mathcal{Z}$ with a Riemannian metric and $\mathcal{Y}\times\mathcal{Z}$ with the product metric. Then, at any solution (x_0,y_0,z_0), we have*

$$\kappa_{x\mapsto y}[F^{-1}(c)](x_0,y_0,z_0) \leqslant \kappa[\overline{F}^{-1}(c)](x_0,y_0,z_0). \tag{7.4}$$

This result, which is proven in Section 7.3, can be used to determine which solution variables of a system of equations are the most sensitive to perturbations. If the ratio between right-hand side and the left-hand side of (7.4) is large, then y is a part of the solution (y, z) that is relatively insensitive to small changes in x (compared to the full solution (y, z) of $\overline{F}(x, (y, z)) = c$).

The remainder of the chapter is organised as follows. In Section 7.2, we define a purely geometric criterion for recognising that a problem can be defined by an FCRE. Then, in Section 7.3, we use this criterion to prove the results stated in the introduction. Section 7.4 shows how to compute the condition number using numerical linear algebra. Finally, Section 7.5 gives a concrete expression for the condition number of computing one of the factors in a Tucker decomposition.

7.2 Geometric characterisation of FCREs

The goal of this section is to find an alternative definition of problems defined by an FCRE that does not reference any defining equations. This is useful because Definition 7.4 involves a projection of the graph of a CREP $F(x, y, z) = c$ onto the x and y variables. Although we do not have access to the equations of this projected problem, it can still be verified geometrically that the projected problem can be defined by an FCRE and has a condition number in the sense of Chapter 6.

Proposition 7.6. *Let $F : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a smooth map of constant rank and let $\mathcal{P}$ be any non-empty level set of F. Then the following three statements are equivalent:*

1. $\operatorname{rank} DF(x, y) = \operatorname{rank} \frac{\partial}{\partial y} F(x, y)$ *for all* (x, y), *i.e.,* F *defines an FCRE,*
2. $\dim \mathcal{P} = \dim \mathcal{X} + \operatorname{null} \frac{\partial}{\partial y} F(x, y)$ *for all* (x, y),
3. *the projection* $\pi_{\mathcal{X}} : \mathcal{P} \to \mathcal{X}$ *is a smooth submersion.*

Proof. We will show the following implications: (1) ⇔ (2), (1) ⇒ (3), and (3) ⇒ (2). For ease of notation, all derivatives in this proof are implicitly evaluated at an arbitrary point (x, y).

For (1) $\Leftrightarrow$ (2), we use the fact that $\dim \mathcal{P} = \operatorname{null} DF$ where null is the nullity [Lee13, Theorem 5.12]. Applying the rank-nullity theorem gives:

$$\begin{aligned}\dim \mathcal{P} = \operatorname{null} DF &= \dim \mathcal{X} + \dim \mathcal{Y} - \operatorname{rank} DF \\ &= \dim \mathcal{X} + \operatorname{null} \frac{\partial F}{\partial y} + \operatorname{rank} \frac{\partial F}{\partial y} - \operatorname{rank} DF,\end{aligned}$$

from which it follows that (1) $\Leftrightarrow$ (2).

To show (1) $\Rightarrow$ (3), we write the tangent space to $\mathcal{P}$ at a general (x, y) as

$$\mathcal{T}_{(x,y)}\mathcal{P} = \ker DF(x,y) = \left\{ (\dot{x}, \dot{y}) \;\middle|\; \frac{\partial F(x,y)}{\partial x}\dot{x} + \frac{\partial F(x,y)}{\partial x}\dot{y} = 0 \right\}.$$

Since $\operatorname{rank} DF = \operatorname{rank} \frac{\partial F}{\partial y}$ and $\operatorname{Im} \frac{\partial F}{\partial y} \subseteq \operatorname{Im} DF$, it follows that $V := \operatorname{Im} \frac{\partial F}{\partial y} = \operatorname{Im} DF$. Pick any right inverse $\left(\frac{\partial F}{\partial y}\right)^{RI} : V \to \mathcal{T}_y\mathcal{Y}$ of $\frac{\partial F}{\partial y}$, i.e., $\frac{\partial F}{\partial y} \circ \left(\frac{\partial F}{\partial y}\right)^{RI} = \operatorname{Id}_V$. For any $\dot{x} \in \mathcal{T}_x\mathcal{X}$, the vector $\dot{y} = -\left(\frac{\partial F}{\partial y}\right)^{RI} \frac{\partial F}{\partial x}\dot{x}$ is well-defined because $\frac{\partial F}{\partial x}\dot{x} \in V$. It can be verified that $(\dot{x}, \dot{y}) \in \mathcal{T}_{(x,y)}\mathcal{P}$. Hence, $D\pi_{\mathcal{X}}$ is surjective.

Finally, we prove (3) $\Rightarrow$ (2). Since $D\pi_{\mathcal{X}}$ is surjective, it has a right inverse $D\pi_{\mathcal{X}}^{RI} : \mathcal{T}_x\mathcal{X} \to \mathcal{T}_{(x,y)}\mathcal{P}$. That is, for any $\dot{x} \in \mathcal{T}_x\mathcal{X}$, the vector $D\pi_{\mathcal{X}}^{RI}(\dot{x})$ is a tuple $(\dot{x}, \dot{y}) \in \ker DF$ for some $\dot{y} \in \mathcal{T}_y\mathcal{Y}$. Hence, $\mathcal{T}_{(x,y)}\mathcal{P} = \ker DF$ contains at least the set

$$W := \left\{ \left(\dot{x}, D\pi_{\mathcal{Y}}\left(D\pi_{\mathcal{X}}^{RI}\dot{x}\right) + v\right) \;\middle|\; \dot{x} \in \mathcal{T}_x\mathcal{X}, v \in \ker \frac{\partial F}{\partial y} \right\}.$$

It is straightforward to check that W is the image of the injective linear map $(\dot{x}, v) \mapsto (\dot{x}, D\pi_{\mathcal{Y}}\left(D\pi_{\mathcal{X}}^{RI}\dot{x}\right) + v)$. Hence, W is a $(\dim \mathcal{X} + \operatorname{null} \frac{\partial F}{\partial y})$-dimensional linear subspace of $\mathcal{T}_{(x,y)}\mathcal{P}$.

We complete the proof by showing that $\mathcal{T}_{(x,y)}\mathcal{P} \subseteq W$. Pick any $(\dot{x}, \dot{y}) \in \mathcal{T}_{(x,y)}\mathcal{P}$. We have both

$$\frac{\partial F}{\partial x}\dot{x} + \frac{\partial F}{\partial y}\dot{y} = 0 \quad \text{and} \quad \frac{\partial F}{\partial x}\dot{x} + \frac{\partial F}{\partial y}\left(D\pi_{\mathcal{Y}}\left(D\pi_{\mathcal{X}}^{RI}\dot{x}\right)\right) = 0.$$

By subtracting the latter equation from the former, we see that $\dot{y} - D\pi_{\mathcal{Y}}\left(D\pi_{\mathcal{X}}^{RI}\dot{x}\right) \in \ker \frac{\partial F}{\partial y}$. Thus, $(\dot{x}, \dot{y}) \in W$, and therefore $\mathcal{T}_{(x,y)}\mathcal{P} = W$. □

Remark 7.7. The assumption in Proposition 7.6 that $\mathcal{P}$ is the level set of a map of constant rank is satisfied by any sufficiently small submanifold of $\mathcal{X} \times \mathcal{Y}$. More precisely, suppose that $\mathcal{P}$ is any embedded submanifold of $\mathcal{X} \times \mathcal{Y}$. By the local slice criterion [Lee13, Theorem 5.8], every point on $\mathcal{P}$ has a neighbourhood $\mathcal{U}$, which is open in $\mathcal{X} \times \mathcal{Y}$, such that $\mathcal{U} \cap \mathcal{P}$ is the zero set of a map $F : \mathcal{U} \to \mathbb{R}^{\dim(\mathcal{X}\times\mathcal{Y})}$ of constant rank. Thus, Proposition 7.6 can be applied to $\mathcal{P} \cap \mathcal{U}$.

The utility of Proposition 7.6 is that FCRE problems can be defined either purely in terms of equations (point 1) or purely geometrically (point 3). Both approaches define the same class of problems.

7.3 Proofs of the main results

An essential part of the proof of Theorem 7.3 is that projecting the problem onto $\mathcal{X} \times \mathcal{Y}$ does not introduce singularities if we look at the problem locally. This is guaranteed by the following lemma.

Lemma 7.8. *Let $F : \mathcal{M} \to \mathcal{N}$ be a smooth map of constant rank r. Then every $p \in \mathcal{M}$ has a neighbourhood $\mathcal{U} \subseteq \mathcal{M}$ such that $F(\mathcal{U})$ is an r-dimensional embedded submanifold of $\mathcal{N}$.*

Proof. By [Lee13, Theorem 4.12], there exist charts at p and $F(p)$ such that F is defined in coordinates by $(x^1, \dots, x^{\dim \mathcal{M}}) \mapsto (x^1, \dots, x^r, 0, \dots, 0)$ on some neighbourhood $\mathcal{U}'$ of p. Let $\pi_r : \mathcal{U}' \to \mathcal{U}'$ be the map that sets all but the first r coordinates to zero, so that, locally, $F = F \circ \pi_r$. By [Lee13, Theorem 5.8], $\pi_r(\mathcal{U}')$ is an embedded submanifold of $\mathcal{M}$. Since $F|_{\pi_r(\mathcal{U}')}$ is an immersion, $F|_{\pi_r(\mathcal{U}')}$ is a local smooth embedding. The result follows from [Lee13, Proposition 5.2]. □

Now we can prove the main result of this chapter. The main idea is that the projected problem $\widetilde{\mathcal{P}}$ satisfies the geometric definition of FCREs (i.e., item 3 of Proposition 7.6).

Proof of Theorem 7.3. By assumption, the equation $F(x, (y, z)) = c$ defines an FCRE over the variables x and (y, z). Define the smooth manifold $\mathcal{P} := F^{-1}(c)$ and define the projection $\pi_{\mathcal{X}\times\mathcal{Y}} : \mathcal{P} \to \mathcal{X} \times \mathcal{Y}$. At any point $(x, y, z) \in \mathcal{P}$, it holds that $\mathcal{T}_{(x,y,z)}\mathcal{P} = \ker DF(x, y, z)$. Thus,

$$\ker D\pi_{\mathcal{X}\times\mathcal{Y}}(x,y,z) = \{(0,0,\dot{z}) \in \ker DF(x,y,z)\} = \{(0,0)\} \times \ker \frac{\partial}{\partial z} F(x,y,z), \tag{7.5}$$

where the last equality follows from the fact that DF is the sum of all partial derivatives of F. By assumption, the dimension of the right-hand side is independent of (x, y, z). Thus, $\pi_{\mathcal{X}\times\mathcal{Y}}$ has constant rank, so that, by Lemma 7.8, the point (x_0, y_0, z_0) has a neighbourhood $\widehat{\mathcal{P}} \subseteq \mathcal{P}$ whose projection onto $\mathcal{X} \times \mathcal{Y}$ is an embedded submanifold of $\mathcal{X} \times \mathcal{Y}$.

Since $F(x,(y,z)) = c$ defines an FCRE, it follows from Proposition 7.6 that $\pi_{\mathcal{X}} : \mathcal{P} \to \mathcal{X}$ is a smooth submersion. In other words, the projection of $\mathcal{T}_{(x,y,z)}\mathcal{P}$ onto $\mathcal{T}_x\mathcal{X}$ is always surjective. Hence, the projection of $\mathcal{T}_{(x,y)}\pi_{\mathcal{X}\times\mathcal{Y}}(\widehat{\mathcal{P}}) = D\pi_{\mathcal{X}\times\mathcal{Y}}[\mathcal{T}_{(x,y,z)}\widehat{\mathcal{P}}]$ is surjective as well. By Proposition 7.6 and Remark 7.7, $\pi_{\mathcal{X}\times\mathcal{Y}}(\widehat{\mathcal{P}})$ is locally defined by the FCRE $\widetilde{F}(x,y) = 0$ for some unspecified map $\widetilde{F} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{\dim(\mathcal{X}\times\mathcal{Y})}$. This proves the first assertion.

By Theorem 6.6, the derivative of the canonical solution map of the equation $\widetilde{F}(x,y) = 0$ is

$$DH(x_0) = -\left(\frac{\partial}{\partial y}\widetilde{F}(x_0,y_0)\right)^{\dagger} \frac{\partial}{\partial x}\widetilde{F}(x_0,y_0), \tag{7.6}$$

which we will reformulate in terms of the derivatives of F. In the following, all derivatives are implicitly evaluated at (x_0, y_0, z_0) or its projections. By the definition of the Moore–Penrose inverse, (i.e., $A^{\dagger} := A|^{-1}_{(\ker A)^{\perp}}$ for any A), (7.6) is equivalent to the system

$$\begin{cases} \frac{\partial \widetilde{F}}{\partial x} + \frac{\partial \widetilde{F}}{\partial y} DH = 0 \\ \operatorname{span} DH \perp \ker \frac{\partial \widetilde{F}}{\partial y} \end{cases} . \tag{7.7}$$

The first line of (7.7) says that $(\dot{x}, DH[\dot{x}]) \in \ker D\widetilde{F}$ for all $\dot{x} \in \mathcal{T}_{x_0}\mathcal{X}$. Since $\pi_{\mathcal{X}\times\mathcal{Y}}(\widehat{\mathcal{P}})$ is a level set of $\widetilde{F}$ and, likewise, $\widehat{\mathcal{P}}$ is locally a level set of F, we have

$$\ker D\widetilde{F} = \mathcal{T}_{(x_0,y_0)}\pi_{\mathcal{X}\times\mathcal{Y}}(\widehat{\mathcal{P}}) = D\pi_{\mathcal{X}\times\mathcal{Y}}\left[\mathcal{T}_{(x_0,y_0,z_0)}\widehat{\mathcal{P}}\right] = D\pi_{\mathcal{X}\times\mathcal{Y}}\left[\ker DF\right].$$

Thus, the first lines of (7.2) and (7.7) are equivalent.

For the second condition, we have

$$\begin{aligned} \ker \frac{\partial \widetilde{F}}{\partial y} &= \left\{ \dot{y} \in \mathcal{T}_{y_0}\mathcal{Y} \,\middle|\, (0,\dot{y}) \in \ker D\widetilde{F} \right\} \\ &= \{ \dot{y} \in \mathcal{T}_{y_0}\mathcal{Y} \,|\, \exists \dot{z} \in \mathcal{T}_{z_0}\mathcal{Z} \colon (0,\dot{y},\dot{z}) \in \ker DF \} \\ &= D\pi_{\mathcal{Y}}\left[\ker \frac{\partial F}{\partial(y,z)}\right], \end{aligned}$$

which shows that the second condition in (7.2) is equivalent to that in (7.7). □

Finally, we show that solving for y in the equation $F(x,y,z)=c$ is at least as well-conditioned as solving for (y,z).

Proof of Proposition 7.5. Let $\overline{H}$ be the canonical solution map corresponding to the FCRE $\overline{F}(x,(y,z))=c$ and let H be the canonical solution map of the problem $\widetilde{\mathcal{P}} \subseteq \widehat{\mathcal{X}} \times \widehat{\mathcal{Y}}$ in Theorem 7.3.

For x in the domain of both H and $\overline{H}$, we have

$$\begin{aligned} d_{\mathcal{Y}\times\mathcal{Z}}(\overline{H}(x_0),\overline{H}(x)) &= \min_{\substack{(y,z)\in\mathcal{Y}\times\mathcal{Z}\\ F(x,y,z)=c}} d_{\mathcal{Y}\times\mathcal{Z}}((y_0,z_0),(y,z)) \\ &\geqslant \min_{\substack{(y,z)\in\mathcal{Y}\times\mathcal{Z}\\ F(x,y,z)=c}} d_{\mathcal{Y}}(y_0,y) \\ &\geqslant \min_{\substack{(y,z)\in\widehat{\mathcal{Y}}\times\widehat{\mathcal{Z}}\\ F(x,y,z)=c}} d_{\mathcal{Y}}(y_0,y) \\ &= d_{\mathcal{Y}}(H(x_0),H(x)). \end{aligned}$$

If we divide both sides by $d_{\mathcal{X}}(x_0,x)$ and take the limit supremum as $x\to x_0$, we obtain $\kappa[\overline{H}](x_0)\geqslant\kappa[H](x_0)$, as desired.

□

7.4 Computation of the condition number

This section presents two alternative characterisations of the condition number that may be more useful for computational purposes. The first one is a translation of the system (7.2) into concrete linear equations.

Proposition 7.9. *Let $F(x,y,z)=c$ be a CREP where $F\colon \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}\to\mathbb{R}^N$ and $\mathcal{X}$ and $\mathcal{Y}$ have a Riemannian metric. At any solution (x_0,y_0,z_0), write the partial derivatives $\frac{\partial F}{\partial x}$, $\frac{\partial F}{\partial y}$, and $\frac{\partial F}{\partial z}$ as matrices J_x, J_y, J_z in coordinates with respect to orthonormal bases of $\mathcal{T}_{x_0}\mathcal{X}$ and $\mathcal{T}_{y_0}\mathcal{Y}$ and an arbitrary basis of $\mathcal{T}_{z_0}\mathcal{Z}$. Let Q be a matrix such that* $\operatorname{span} Q = (\operatorname{span} J_z)^\perp$ *and let $U = [U_y^T \quad U_z^T]^T$ be a matrix such that $J_yU_y + J_zU_z = 0$ and* $\operatorname{span} U = \ker\,[J_y \quad J_z]$. *Then the columns of the matrix $A := \begin{bmatrix} Q^TJ_y \\ U_y^T \end{bmatrix}$ are linearly independent. Furthermore,*

$$\kappa_{x\mapsto y}[F^{-1}(c)](x_0,y_0,z_0) = \left\| A^+ \begin{bmatrix} -Q^TJ_x \\ 0 \end{bmatrix} \right\|,$$

where A^+ is any left inverse of A, such as the Moore–Penrose inverse, and $\|\cdot\|$ is the operator norm.

Proof. In the following, we evaluate all derivatives implicitly at (x_0, y_0, z_0) or its projections. We start by finding concrete equations for the first constraint in (7.2). For any $(\dot{x}, \dot{y}) \in \mathcal{T}_{(x_0,y_0)}(\mathcal{X} \times \mathcal{Y})$, we have

$$\begin{aligned}
(\dot{x}, \dot{y}) \in D\pi_{\mathcal{X}\times\mathcal{Y}}\,[\ker DF] &\Leftrightarrow \frac{\partial F}{\partial x}\dot{x} + \frac{\partial F}{\partial y}\dot{y} + \frac{\partial F}{\partial z}\dot{z} = 0 \text{ for some } \dot{z} \in \mathcal{T}_{z_0}\mathcal{Z} \\
&\Leftrightarrow \frac{\partial F}{\partial x}\dot{x} + \frac{\partial F}{\partial y}\dot{y} \in \operatorname{span}\left(\frac{\partial F}{\partial z}\right) \\
&\Leftrightarrow Q^T\left(\frac{\partial F}{\partial x}\dot{x} + \frac{\partial F}{\partial y}\dot{y}\right) = 0 \\
&\Leftrightarrow Q^T J_y \hat{\dot{y}} = -Q^T J_x \hat{\dot{x}}, \qquad (7.8)
\end{aligned}$$

where $\hat{\dot{x}}$ and $\hat{\dot{y}}$ are the coordinates of $\dot{x}$ and $\dot{y}$, respectively.

Similarly, for the second requirement in (7.2), we have

$$\begin{aligned}
\dot{y} \perp D\pi_{\mathcal{Y}}\left[\ker \frac{\partial F}{\partial(y,z)}\right] &\Leftrightarrow \hat{\dot{y}} \perp [\mathbb{I} \quad 0] \ker\left([J_y \quad J_z]\right) \\
&\Leftrightarrow \hat{\dot{y}} \perp \operatorname{span} U_y \\
&\Leftrightarrow U_y^T \hat{\dot{y}} = 0.
\end{aligned}$$

By combining these two observations, we see that (7.2) is equivalent to the system

$$A \cdot DH = \begin{bmatrix} -Q^T J_x \\ 0 \end{bmatrix}, \quad \text{where} \quad A = \begin{bmatrix} Q^T J_y \\ U_y^T \end{bmatrix}. \qquad (7.9)$$

Since this system has a unique solution, A has full rank and is thereby left-invertible. Hence, the Moore–Penrose inverse of A is a left inverse [GVL13, §5.5.2]. Since the desired condition number is $\|DH\|$, this concludes the proof. □

Remark 7.10. Proposition 7.9 suggests computing the derivative of the solution map by solving (7.9). Although this system has precisely one exact solution, it may be overdetermined. It is possible to reduce the number of equations and keep the same solution. For instance, (7.8) expresses that a vector that is known to lie in span $\frac{\partial F}{\partial(x,y)}$ is also an element of span $\frac{\partial F}{\partial z}$. The minimal number of linear equations needed to express this is the codimension of span $\frac{\partial F}{\partial z} \cap$ span $\frac{\partial F}{\partial(x,y)}$

as a subspace of span $\frac{\partial F}{\partial(x,y)}$, but the number of equations used in (7.8) is the codimension of span $\frac{\partial F}{\partial z}$ in $\mathbb{R}^N$. Methods to reduce the number of equations are omitted from this discussion for simplicity.

Another characterisation of the condition number is given by the following lemma. It captures the intuition that $DH(x_0)[\dot{x}]$ gives the smallest possible change to y that solves the CREP when the input is perturbed by $\dot{x}$. Essentially, it uncovers where the defining equations (7.2) come from: they are the critical point equations of a convex optimisation problem.

Lemma 7.11. *Suppose that the equation $F(x,y,z) = c$ satisfies the assumptions of Theorem 7.3 and that (x_0, y_0, z_0) is any solution of the equation. Then the solution of (7.2) is*

$$DH(x_0)\colon \dot{x} \mapsto \arg\min_{\dot{y}} \|\dot{y}\| \text{ s.t. } DF(x_0,y_0,z_0)[\dot{x},\dot{y},\dot{z}] = 0 \text{ for some } \dot{z} \in \mathcal{T}_{z_0}\mathcal{Z}.$$

Consequently, $\kappa_{x\mapsto y}[F^{-1}(c)](x_0,y_0,z_0)$ is the operator norm of $DH(x_0)$.

Proof of Lemma 7.11. Given any $\dot{x} \in \mathcal{T}_{x_0}\mathcal{X}$, write the set of $(\dot{y},\dot{z})$ that solve the linearisation of $F(x,y,z) = c$ as

$$L_{\dot{x}} := \{(\dot{y},\dot{z}) \in \mathcal{T}_{y_0}\mathcal{Y} \times \mathcal{T}_{z_0}\mathcal{Z} \mid DF(x_0,y_0,z_0)[\dot{x},\dot{y},\dot{z}] = 0\}.$$

Note that $L_0 = D\pi_{\mathcal{Y}\times\mathcal{Z}}\left(\ker \frac{\partial F}{\partial(y,z)}\right)$. Thus, (7.2) defines $DH(x_0)[\dot{x}]$ as the unique element in $D\pi_{\mathcal{Y}}[L_{\dot{x}}] \cap D\pi_{\mathcal{Y}}[L_0]^{\perp}$.

Fix any vector $\dot{x} \in \mathcal{T}_{x_0}\mathcal{X}$. The space $L_{\dot{x}}$ is defined by the linear system $\frac{\partial F}{\partial(y,z)}[\dot{y},\dot{z}] = -\frac{\partial F}{\partial x}\dot{x}$. Since L_0 is defined by the same equations, but with a different right-hand side, it follows from elementary linear algebra that L_0 and $L_{\dot{x}}$ are parallel. That is, $L_{\dot{x}} = L_0 + v_{\dot{x}}$ for some $v_{\dot{x}} \in \mathcal{T}_{y_0}\mathcal{Y} \times \mathcal{T}_{z_0}\mathcal{Z}$. Consequently, $D\pi_{\mathcal{Y}}[L_{\dot{x}}]$ and $D\pi_{\mathcal{Y}}[L_0]$ are parallel. Hence, the vector in $D\pi_{\mathcal{Y}}[L_{\dot{x}}]$ with the smallest norm is orthogonal to $D\pi_{\mathcal{Y}}[L_0]$ and thereby satisfies the defining equations of $DH(x_0)[\dot{x}]$.

□

7.5 Example: individual factors of the Tucker decomposition

We studied the condition number of the (orthogonal) Tucker decomposition in Section 6.6. This condition number measures the sensitivity of all factors combined with respect to the input tensor. Using the techniques from this chapter, we can study the sensitivity of individual factors. Again, we restrict ourselves to the orthogonal variant of the Tucker decomposition.

To model the problem, let $\mathcal{X} \subseteq \mathbb{R}^{n_1 \times \cdots \times n_D}$ be the set of tensors of multilinear rank $(m_1, \ldots, m_D)$, which is a manifold by the results in Sections 3.4.1 and 4.3. Define

$$\begin{aligned} G_{\mathcal{T}} \colon \mathbb{R}_\star^{m_1 \times \cdots \times m_D} \times \mathrm{St}(n_1, m_1) \times \cdots \times \mathrm{St}(n_D, m_D) \to \mathcal{X} \\ (\mathcal{C}, U_1, \ldots, U_D) \mapsto (U_1 \otimes \cdots \otimes U_D)\mathcal{C} \end{aligned}$$

and

$$F_{\mathcal{T}}(\mathcal{X}, \mathcal{C}, U_1, \ldots, U_D) := \mathcal{X} - G_{\mathcal{T}}(\mathcal{C}, U_1, \ldots, U_D), \tag{7.10}$$

where $\mathcal{X} \in \mathcal{X}$. Then the Tucker decomposition problem is defined by the zero set of $F_{\mathcal{T}}$.

The sensitivity of the factor U_d in the Tucker decomposition with respect to perturbations in $\mathcal{X}$ is measured by $\kappa_{\mathcal{X} \mapsto U_d}[F_{\mathcal{T}}^{-1}(0)]$. The generic variable names x, y, and z used in the introduction would then refer to $\mathcal{X}$, U_d, and $(\mathcal{C}, U_2, \ldots, U_D)$, respectively. Likewise, the sensitivity of $\mathcal{C}$ is measured by $\kappa_{\mathcal{X} \mapsto \mathcal{C}}[F_{\mathcal{T}}^{-1}(0)]$, in which case $y = \mathcal{C}$ and $z = (U_1, \ldots, U_D)$. For simplicity, we omit the subscripts (e.g., x_0, y_0, z_0) when referring to the particular solution where the condition number is evaluated. The condition number is given by the following proposition. As in Proposition 6.19, the condition number can be formulated in terms of the singular values in the higher-order singular value decomposition [DLDMV00a].

Proposition 7.12. *Let $\mathcal{P}$ be the zero set of* (7.10) *and let $(\mathcal{X}, \mathcal{C}, U_1, \ldots, U_D) \in \mathcal{P}$ be any point. Endow the domain of $F_{\mathcal{T}}$ with the Euclidean (i.e., Frobenius) norm. Then, for all $d = 1, \ldots, D$,*

$$\kappa_{\mathcal{X} \to U_d}[\mathcal{P}](\mathcal{X}, \mathcal{C}, U_1, \ldots, U_D) = \begin{cases} 0 & \text{if } U_d \text{ is square,} \\ \sigma_{\min}(\mathcal{C}_{(d)})^{-1} & \text{otherwise,} \end{cases} \tag{7.11}$$

where $\sigma_{\min}(\mathcal{C}_{(d)})$ is the smallest singular value of the dth flattening of $\mathcal{C}$. Furthermore,

$$\kappa_{\mathcal{X} \to \mathcal{C}}[\mathcal{P}](\mathcal{X}, \mathcal{C}, U_1, \ldots, U_D) = 1. \tag{7.12}$$

Proof of (7.11). Since we can permute the arguments of $F_{\mathcal{T}}$, we can assume without loss of generality that $d = 1$. Let $DH(\mathcal{X})$ be the differential of the canonical solution map and let $\dot{\mathcal{X}} \in \mathcal{T}_{\mathcal{X}}\mathcal{X}$ be a generic tangent vector. The equations (7.2) that define DH can be specialised to the Tucker decomposition problem as follows:

$$\begin{cases} \dot{\mathcal{X}} = DG_{\mathcal{T}}[\dot{\mathcal{C}}, \dot{U}_1, \dots, \dot{U}_D] \quad \text{for some} \quad \dot{\mathcal{C}}, \dot{U}_2, \dots, \dot{U}_D \\ \dot{U}_1 \perp D\pi_{\mathrm{St}(n_1,m_1)} \left[\ker DG_{\mathcal{T}}\right] \end{cases}, \tag{7.13}$$

where $\dot{U}_1 := DH(\mathcal{X})[\dot{\mathcal{X}}]$ and $DG_{\mathcal{T}}$ is evaluated at $(\mathcal{C}, U_1, \dots, U_D)$.

Before solving this, we simplify the second condition in (7.13). By (6.15),

$$D\pi_{\mathrm{St}(n_1,m_1)} \left[\ker DG_{\mathcal{T}}\right] = \{U_1 \dot{Q} \mid \dot{Q} \in \mathcal{T}_{\mathbb{I}} O(m_1)\} \subseteq \{U_1 \dot{Q} \mid \dot{Q} \in \mathbb{R}^{m_1 \times m_1}\},$$

where $O(m_1)$ is the orthogonal group. Thus, the second constraint in (7.13) is satisfied on the sufficient (but not necessary) condition that $\dot{U}_1^T U_1 = 0$

We can solve (7.13) for $\dot{U}_1$ as a function of $\dot{\mathcal{X}}$ as follows. We know from Proposition 4.9 that every $\dot{\mathcal{X}} \in \mathcal{T}_{\mathcal{X}}\mathcal{X}$ admits a unique decomposition of the form

$$\begin{aligned} \dot{\mathcal{X}} &= (\dot{U}_1 \otimes \cdots \otimes U_D)\mathcal{C} + \cdots + (U_1 \otimes \cdots \otimes \dot{U}_D)\mathcal{C} + (U_1 \otimes \cdots \otimes U_D)\dot{\mathcal{C}} \\ &= DG_{\mathcal{T}}[\dot{\mathcal{C}}, \dot{U}_1, \dots, \dot{U}_D] \end{aligned} \tag{7.14}$$

where $\dot{\mathcal{C}} \in \mathbb{R}^{m_1 \times \cdots \times m_D}$ and $\dot{U}_d^T U_d = 0$ for all d. The factor $\dot{U}_1$ in this decomposition solves (7.13) given $\dot{\mathcal{X}}$. Since $DH(\mathcal{X})$ is the unique linear map that takes $\dot{\mathcal{X}} \in \mathcal{T}_{\mathcal{X}}\mathcal{X}$ to the solution $\dot{U}_1$ of (7.13), $DH(\mathcal{X})[\dot{\mathcal{X}}]$ evaluates to the factor $\dot{U}_1$ in the decomposition (7.14), for any $\dot{\mathcal{X}} \in \mathcal{T}_{\mathcal{X}}\mathcal{X}$.

Assume that U_1 is square. Then the only matrix $\dot{U}_1$ that satisfies the constraint $\dot{U}_1^T U_1 = 0$ is the zero matrix. Since $DH(\mathcal{X})[\dot{\mathcal{X}}]$ satisfies this constraint for any $\dot{\mathcal{X}}$, it follows that $DH(\mathcal{X})$ is the zero map. Thus, the condition number is zero. In the remainder, we assume that U_1 is not square, so that the constraint $\dot{U}_1^T U_1 = 0$ is nontrivial.

The operator norm of $DH(\mathcal{X})$ can be calculated as $\left\| DH(\mathcal{X})|_{(\ker DH(\mathcal{X}))^\perp} \right\|$. That is, we can restrict DH to its row space. To determine this space, note that all summands on the right-hand side of (7.14) are pairwise orthogonal in the Euclidean inner product on $\mathbb{R}^{n_1 \times \cdots \times n_D}$, since $\dot{U}_d^T U_d = 0$ for all d. In addition, $\ker DH(\mathcal{X})$ consists of all $\dot{\mathcal{X}}$ such that the first term in (7.14) vanishes. Hence,

$$(\ker DH(\mathcal{X}))^\perp = \left\{ (\dot{U}_1 \otimes U_2 \otimes \cdots \otimes U_D)\mathcal{C} \,\middle|\, \dot{U}_1^T U_1 = 0 \right\},$$

so that

$$DH(\mathcal{X})|_{(\ker DH(\mathcal{X}))^\perp} : (\dot{U}_1 \otimes U_2 \otimes \cdots \otimes U_D)\mathcal{C} \mapsto \dot{U}_1.$$

If we represent tensors as their first standard flattening, the inverse of this map is $L : \dot{U}_1 \mapsto \dot{U}_1 \mathcal{C}_{(1)}(U_2 \oslash \cdots \oslash U_D)^T$ where $\oslash$ is the Kronecker product. The singular values of L are the singular values of $\mathcal{C}_{(1)}$. In conclusion:

$$\|DH(\mathcal{X})\| = \left\| DH(\mathcal{X})|_{(\ker DH(\mathcal{X}))^\perp} \right\| = 1/\sigma_{m_1}(\mathcal{C}_{(1)}).$$

□

Proof of (7.12). We use the characterisation of DH from Lemma 7.11, i.e.,

$$DH(\mathcal{X})[\dot{\mathcal{X}}] = \arg\min_{\dot{\mathcal{C}}} \left\|\dot{\mathcal{C}}\right\| \text{ s.t. } \dot{\mathcal{X}} = DG_{\mathcal{T}}[\dot{\mathcal{C}}, \dot{U}_1, \ldots, \dot{U}_D] \text{ for some } \dot{U}_1, \ldots, \dot{U}_D.$$

For any $\dot{\mathcal{X}} \in \mathcal{T}_{\mathcal{X}}\mathcal{X}$, the minimum can be estimated by decomposing $\dot{\mathcal{X}}$ uniquely as $\dot{\mathcal{X}} = DG_{\mathcal{T}}[\dot{\mathcal{C}}, \dot{U}_1, \ldots, \dot{U}_D]$ with $\dot{U}_d^T U_d = 0$ for all d, as in (7.14). If we multiply (7.14) on the left by $(U_1^T \otimes \cdots \otimes U_D^T)$, all but one term vanish and we obtain $\dot{\mathcal{C}} = (U_1^T \otimes \cdots \otimes U_D^T)\dot{\mathcal{X}}$. Thus, in this specific decomposition of $\dot{\mathcal{X}}$, we have $\left\|\dot{\mathcal{C}}\right\| \leqslant \left\|\dot{\mathcal{X}}\right\|$. It follows that the operator norm of $DH(\mathcal{X})$ is at most 1.

To show that $\|DH(\mathcal{X})\| = 1$, it suffices to find a tangent vector $\dot{\mathcal{X}}$ such that $\left\|DH(\mathcal{X})[\dot{\mathcal{X}}]\right\| = \left\|\dot{\mathcal{X}}\right\|$. Pick the radial direction $\dot{\mathcal{X}} := \mathcal{X} = (U_1 \otimes \cdots \otimes U_D)\mathcal{C}$. We can see that $DH(\mathcal{X})[\dot{\mathcal{X}}] = \mathcal{C}$ by an argument from the proof of Proposition 6.19, which we repeat here.

The kernel of $DG_{\mathcal{T}}$ is the tangent space to the preimage $G_{\mathcal{T}}^{-1}(\mathcal{X})$. Since the projection of $G_{\mathcal{T}}^{-1}(\mathcal{X})$ onto the first component is the orbit of $\mathcal{C}$ under the action of $O(m_1) \times \cdots \times O(m_D)$ (see e.g. (6.15)), it is contained in the sphere of radius $\|\mathcal{C}\|$. Hence, the radial direction $\dot{\mathcal{C}} = \mathcal{C}$ is normal to $D\pi_{\mathbb{R}_\star^{n_1 \times \cdots \times n_D}}[\ker DG_{\mathcal{T}}]$ and thereby solves the critical point equations (7.2). It follows that $\mathcal{C} = DH(\mathcal{X})[\dot{\mathcal{X}}]$ when $\dot{\mathcal{X}} = \mathcal{X}$. Since $\|\mathcal{C}\| = \|\mathcal{X}\|$, this completes the proof.

□

Proposition 7.12 has the following heuristic explanation. The factor U_1 gives a basis for the column space of the flattened tensor $\mathcal{X}_{(1)} \in \mathbb{R}^{n_1 \times n_2 \cdots n_D}$. If U_1 is square, then $\operatorname{span}(\mathcal{X}_{(1)}) = \mathbb{R}^{n_1}$. If a perturbed tensor $\widetilde{\mathcal{X}}$ is sufficiently close to $\mathcal{X}$, its column space is also $\mathbb{R}^{n_1}$ by the Eckart–Young theorem [SS90, Theorem IV.4.18]. Therefore, $\widetilde{\mathcal{X}}$ admits a Tucker decomposition whose first factor is U_1. Alternatively, if U_1 is not square, its column space may rotate as $\mathcal{X}$ is perturbed to $\widetilde{\mathcal{X}}$. Standard results such as Wedin's $\sin\Theta$-theorem [SS90, Theorem V.4.4] bound the largest rotation angle in terms of the size of the perturbation to $\widetilde{\mathcal{X}}$ and $\sigma_{m_1}(\mathcal{X}_{(1)})^{-1}$. Thus, it is natural to expect the latter quantity to be the condition number, and this expectation is confirmed by Proposition 7.12.

Remark 7.13. By [BC13, Section 14.3.2], the condition number of computing the image of a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ of rank m is $1/\sigma_m(X)$. Since computing the U_1 factor of the Tucker decomposition of a tensor $\mathcal{X}$ is equivalent to computing an orthonormal basis of $\operatorname{span}(\mathcal{X}_{(1)})$, one might wonder about the difference between this condition number and the expression in Proposition 7.12. There are two main conceptual differences:

1. The formulation used in [BC13] quotients out the choice of basis for the image in order to obtain a unique solution, as described in Section 6.4. By contrast, Proposition 7.12 is about the condition number in Definition 7.4, which defines the error in terms of a least-squares projection.

2. The input space in Proposition 7.12 consists of all tensors of multilinear rank (m_1, m_2, m_3). If we flatten a tensor $\mathcal{X}$ to $\mathcal{X}_{(1)}$ and apply the approach from [BC13], the corresponding condition number takes all perturbations $\widetilde{\mathcal{X}}$ of $\mathcal{X}$ into account such that $\widetilde{\mathcal{X}}_{(1)}$ has rank m_1. By contrast, our approach considers only perturbations of $\mathcal{X}$ to $\mathcal{X}$ that preserve the multilinear rank. That is, our approach has a more constrained input space.

Remark 7.14. Proposition 7.12 fits in perfectly with two results stated earlier. We know from the general result of Proposition 7.5 that solving for any one solution variable (e.g., one factor of the decomposition) is at most as ill-conditioned as solving for all solution variables combined. It does not follow in general that there is a single variable that is as ill-conditioned as all variables combined. However, by Proposition 6.19 and Proposition 7.12, this is the case for the Tucker decomposition: the condition number of the full decomposition is $\max\left(\{1\} \cup \{\sigma_{\min}(\mathcal{C}_{(d)})^{-1} \mid m_d < n_d\}\right)$, i.e., the maximum of the condition numbers of the individual factors.

7.6 Conclusion

In this chapter, we extended the theory of condition from Chapter 6. The new condition number is defined for (possibly underdetermined) constant-rank elimination problems (CREPs). The condition number estimates the minimal change in the solution variable y for the worst-case infinitesimal perturbation to the input x, keeping the latent variable z close to its reference value. By measuring the error this way, we find a lower bound for the condition number of a CREP based on the eliminated variables: if solving for any subset of the variables is ill-conditioned, so must be solving for all variables combined.

The condition number of a CREP can be characterised in terms of the partial derivatives of the defining equations. By using this result, we derived a condition number of each factor in an orthogonal Tucker decomposition. The results confirm two simple intuitions. First, the dth basis in a Tucker decomposition is ill-conditioned insofar as the dth flattening of the tensor has a small singular value. Second, the condition number of the decomposition as a whole equals that of the most sensitive factor.

Chapter 8

Conclusion

Every numerical problem has a condition number, which is a measure of the difficulty of solving the problem in finite precision. This dissertation presents new results on the theory and computation of condition numbers, particularly with applications to tensor decompositions.

The main conceptual innovation of the thesis is the modular theory of condition developed in chapters 6 and 7. This theory adds a new layer of explainability to the condition number. At the heart of this layer of explainability is the *condition number of underdetermined problems*, which was not a thoroughly explored concept in the literature.

With the proposed theory of condition, it is possible to isolate the constraints that make a problem ill-conditioned. This is achieved by comparing the condition number of the problem to any relaxation, for which a condition number was introduced in Chapter 6. The interpretability is further enhanced by the condition numbers of each individual solution variable. Thus, the techniques we introduced make it possible to explain the condition number at the level of either the constraints, the solution variables, or some combination thereof.

Aside from the modular theory of condition, the results of chapters 4 and 5 are contributions to the theory and computation of condition numbers of tensor decompositions. In particular, we found a practical algorithm to compute the sensitivity of additive decompositions of large tensors of low rank.

Taking all chapters together, the results in this dissertation make it practically feasible to compute the condition number of many tensor decompositions, such as polyadic, Waring, (structured) block term, and Tucker decompositions. The techniques introduced in this dissertation can also be applied to models such as tensor trains or the factor matrices of block-term decompositions.

8.1 Contributions by chapter

The contributions of each chapter can be summarised as follows.

Chapter 2

- An updated proof of Rice's theorem (Theorem 2.4) that rectifies one of the assumptions in the original publication.
- A presentation of the geometric theory of condition numbers that emphasises the sensitivity of solving equations numerically. The presentation is non-standard in that popular texts on condition either disregard geometric aspects [Hig02; TB97] or emphasise the relevance to complexity theory rather than numerical analysis [Blu+98; BC13]. The unique combination of geometry and sensitivity is intended to connect the literature on numerical analysis and differential geometry.

Chapter 4

- A joint analysis of basic manifolds (called *structured Tucker manifolds*) involved in tensor decompositions. The analysis establishes smoothness (Proposition 4.8), group symmetries (Proposition 4.6), and an orthonormal basis (Proposition 4.9).
- A characterisation of the condition number of block term decompositions in terms of the so-called *Terracini matrix* (eq. 4.11). Estimates of the condition number are provided as well.
- A proof of the existence of subspace constrained SBTDs (Proposition 4.13), generalising the analogous result for polyadic decompositions [SL08, Proposition 3.1].
- An invariance property of the condition number of SBTDs (Theorem 4.14) that can be exploited to speed up the computation of the condition number (Section 4.5.2). Numerical evidence (Section 4.6) confirms the expected implications for sensitivity and complexity.

Chapter 5

- An invariance property of symmetric tensor decompositions (Theorem 5.1) that is slightly weaker than the analogous Theorem 4.14.
- A proof of the equality of the condition numbers of polyadic and Waring decompositions of rank 2 (Proposition 5.3), supplemented with numerical evidence for the case of higher ranks.

Chapter 6

- A theory of condition for a general class of underdetermined problems. Notable features of the general theory include:
  - a theorem on the existence and smoothness of the canonical solution map (Theorem 6.6),
  - an expression of the condition number that can be computed numerically (Theorem 6.6 and Definition 6.7),
  - the relation to the constraining or relaxation of numerical problems (Corollary 6.2),
  - consistency with problems with unique solutions ((6.4) and Definition 6.7),
  - equivalence to the known approach based on quotient manifolds whenever that approach is applicable (Proposition 6.12).
- A computation of the condition number of two-factor matrix decompositions (Proposition 6.14) and a derivation of an optimal two-factor decomposition (Corollary 6.16).
- An expression for the condition number of orthogonal Tucker decompositions in terms of the singular values of the canonical flattenings (Proposition 6.19). Numerical evidence confirms that the associated asymptotic error bound is a good approximation in practice (Section 6.7).

Chapter 7

- A theory of condition of elimination of variables. This includes two equivalent definitions of the condition number (Definition 7.4 and Lemma 7.11) and an expression that can be used to compute the condition number numerically (Theorem 7.3). The condition number measures the contribution of each variable to the overall sensitivity, as supported by Proposition 7.5.

- Concrete expressions of the condition numbers of the factors in any orthogonal Tucker decomposition (Proposition 7.12).

8.2 Suggested further research

One unsolved problem encountered in this dissertation is Conjecture 5.2, i.e, the statement that the condition number of any Waring decomposition is equal the condition number of the equivalent solution to the polyadic decomposition problem. Two strictly weaker conjectures may be more feasible to resolve:

1. *At any WD $(\mathcal{A}_1, \ldots, \mathcal{A}_R)$, its condition number as a WD and as a PD are either both finite or both infinite.* This conjecture can also be phrased purely algebraically (i.e., without a metric), because the condition number of either decomposition problem is finite if and only if the associated Terracini matrix has full rank. Geometrically, it says that the tangent spaces $\mathcal{T}_{\mathcal{A}_1}\mathcal{V}, \ldots, \mathcal{T}_{\mathcal{A}_R}\mathcal{V}$ intersect if and only if $\mathcal{T}_{\mathcal{A}_1}\mathcal{S}, \ldots, \mathcal{T}_{\mathcal{A}_R}\mathcal{S}$ intersect, where $\mathcal{V}$ and $\mathcal{S}$ are the Veronese and Segre manifold, respectively. This suggests a proof based on algebraic geometry. Section 5.3 discusses the limitations of the proof technique in Chapter 5.
2. *The condition number of a Waring decomposition is invariant under orthogonal Tucker compression.* Only the case of non-minimal compression (i.e., Theorem 5.1) has been solved for arbitrary rank. I believe that a generalised proof similar to those of Theorem 5.1 and Theorem 4.14 is possible, but I have not found it. The main obstacle trying to prove the general case was that the basis of the tangent space of the Veronese manifold consists of *sums* of tensor products, whereas the tangent space to the structured Tucker manifold admits basis vectors that are simple multilinear products (Proposition 4.9). This complicates the estimation of inner products.

Another continuation of the work in this text would be the development of algorithms to compute the condition number of some concrete problems. The following examples would be natural applications of Chapters 6 and 7.

1. The block-term decomposition has several variants, one of which is the decomposition of a tensor $\mathcal{A}$ as a $\mathcal{A} = \sum_{r=1}^{R} \mathcal{A}_r$ where each $\mathcal{A}_r$ is a point on the Tucker manifold (i.e., a tensor of low multilinear rank). The condition number of this variant was discussed in Chapter 4. A more common variant would factor each $\mathcal{A}_r$ in this decomposition as a multilinear product [DL11],

which gives a decomposition of the form

$$\mathcal{A} = \sum_{r=1}^{R} (U_1^r \otimes \cdots \otimes U_D^r)\mathcal{C}_r.$$

This variant combines a join decomposition and Tucker decompositions. The sensitivity of all factors with respect to $\mathcal{A}$ can be analysed using the techniques of Chapter 6. Similarly, the sensitivity of each *individual* factor is expressed by the condition number defined in Chapter 7. This would be relevant for a perturbation analysis of the blind source separation problem in Section 3.5.2.

2. In algebraic computer vision, the *3D reconstruction problem* is the estimation of m camera configurations and points $p_1, \ldots, p_n$ in three-dimensional space given m pictures of all n points [HZ03; KK22]. Taking the ith picture is modelled as a linear map $C_i : \mathbb{P}^3 \to \mathbb{P}^2$, where $\mathbb{P}^N$ denotes N-dimensional projective space. The 3D reconstruction problem can be defined as the inversion of the map

$$G\colon (C_1, \ldots, C_m, p_1, \ldots, p_n) \mapsto (C_1 p_1, \ldots, C_1 p_n, C_2 p_1, \ldots, C_m p_n).$$

This problem is underdetermined because there is no canonical choice of coordinates of $\mathbb{P}^3$. That is, given any change of basis $M : \mathbb{P}^3 \to \mathbb{P}^3$, we have $G(C_1 M^{-1}, \ldots, C_m M^{-1}, M p_1, \ldots, M p_n) = G(C_1, \ldots, C_m, p_1, \ldots, p_n)$.

Because underdetermined problems do not have a condition number in the usual sense, the sensitivity analysis of 3D reconstruction has been limited to the perturbation theory of so-called *fundamental* or *essential matrices* [FKK22]. Barring a precise definition, these matrices are a somewhat difficult-to-interpret invariant of the solution that is independent of the choice of basis in $\mathbb{P}^3$. In discussions with Joe Kileel, I have determined that the condition number from Chapter 7 would be a viable alternative measure of sensitivity and can either be used for the sensitivity of the cameras, the points, or both.

Finally, some work could be done to connect the condition number of an FCRE to other notions of condition than sensitivity, such as those in Section 2.4. Dégot [Dé00] introduced a condition-like number that measures the sensitivity of the solutions of underdetermined homogeneous polynomial systems. This condition number is inversely proportional to the distance to ill-posedness in the sense of Section 2.4. This makes it plausible that the condition number of more general underdetermined FCREs admits a similar characterisation.

Bibliography

[AB03] C. M. Andersen and R. Bro. “Practical aspects of PARAFAC modeling of fluorescence excitation-emission data”. In: *Journal of Chemometrics* 17.4 (Apr. 2003), pp. 200–215.

[AC20] E. Angelini and L. Chiantini. “On the identifiability of ternary forms”. In: *Linear Algebra and its Applications* 599 (Aug. 2020), pp. 36–65.

[AC21] E. Angelini and L. Chiantini. “Minimality and uniqueness for decompositions of specific ternary forms”. In: *Mathematics of Computation* 91.334 (Nov. 2021), pp. 973–1006.

[AH95] J. Alexander and A. Hirschowitz. “Polynomial interpolation in several variables”. In: *Journal of Algebraic Geometry* 4.2 (1995), pp. 201–222.

[AMS08] P.-A. Absil, R. Mahony and R. Sepulchre. *Optimization algorithms on matrix manifolds.* Princeton: Princeton University Press, 2008.

[Ana+14] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade and M. Telgarsky. “Tensor decompositions for learning latent variable models”. In: *Journal of Machine Learning Research* 15 (2014), pp. 2773–2832.

[ANT19] B. Arslan, V. Noferini and F. Tisseur. “The structured condition number of a differentiable map between matrix manifolds, with applications”. In: *SIAM Journal on Matrix Analysis and Applications* 40.2 (2019), pp. 774–799.

[AOP08] H. Abo, G. Ottaviani and C. Peterson. “Induction for secant varieties of Segre varieties”. In: *Transactions of the American Mathematical Society* 361.2 (Sept. 2008), pp. 767–792.

[Arm10] D. Armentano. “Stochastic perturbations and smooth condition numbers”. In: *Journal of Complexity* 26.2 (Apr. 2010), pp. 161–171.

[BA98] R. Bro and C. A. Andersson. “Improving the speed of multiway algorithms part II: Compression”. In: *Chemometrics and Intelligent Laboratory Systems* 42.1-2 (1998), pp. 105–113.

[BBS20] E. Ballico, A. Bernardi and P. Santarsiero. *Terracini locus for three points on a Segre variety.* Dec. 2020.

[BC13] P. Bürgisser and F. Cucker. “Condition”. In: *Media.* Vol. 349. Grundlehren der mathematischen Wissenschaften. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[BC21] E. Ballico and L. Chiantini. “On the Terracini locus of projective varieties”. In: *Milan Journal of Mathematics* 89.1 (2021).

[Ber22] R. Bergmann. “Manopt.jl: Optimization on manifolds in Julia”. In: *Journal of Open Source Software* 7.70 (Feb. 2022), p. 3866.

[Bez+17] J. Bezanson, A. Edelman, S. Karpinski and V. B. Shah. “Julia: A fresh approach to numerical computing”. In: *SIAM Review* 59.1 (2017), pp. 65–98.

[Blu+98] L. Blum, F. Cucker, M. Shub and S. Smale. *Complexity and real computation.* New York: Springer-Verlag, 1998.

[BM03] S. Burer and R. D. Monteiro. “A Nonlinear Programming Algorithm for Solving Semidefinite Programs via Low-Rank Factorization”. In: *Mathematical Programming* 95.2 (Feb. 2003), pp. 329–357.

[Bor07] K. M. Borgwardt. “Graph kernels”. PhD thesis. Ludwig–Maximilians University Munich, 2007.

[Bou23] N. Boumal. *An Introduction to Optimization on Smooth Manifolds.* Cambridge New York, N.Y: Cambridge university press, 2023.

[BV18a] P. Breiding and N. Vannieuwenhoven. “Convergence analysis of Riemannian Gauss–Newton methods and its connection with the geometric condition number”. In: *Applied Mathematics Letters* 78 (2018). Publisher: Elsevier Ltd, pp. 42–50.

[BV18b] P. Breiding and N. Vannieuwenhoven. “The condition number of join decompositions”. In: *SIAM Journal on Matrix Analysis and Applications* 39.1 (Jan. 2018), pp. 287–309.

[BV21] P. Breiding and N. Vannieuwenhoven. “The condition number of Riemannian approximation problems”. In: *SIAM Journal on Optimization* 31.1 (Jan. 2021), pp. 1049–1077.

[CJ10] P. Comon and C. Jutten, eds. *Handbook of Blind Source Separation: Independent Component Analysis and Applications.* 1st ed. Amsterdam ; Boston: Elsevier, 2010.

[CLO07] D. A. Cox, J. B. Little and D. O'Shea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra.* 3rd ed. Undergraduate texts in mathematics. New York: Springer, 2007.

[Com+08] P. Comon, G. Golub, L. H. Lim and B. Mourrain. "Symmetric tensors and symmetric tensor rank". In: *SIAM Journal on Matrix Analysis and Applications* 30.3 (2008), pp. 1254–1279.

[Com+20] P. Comon, L.-H. Lim, Y. Qi and K. Ye. "Topology of tensor ranks". In: *Advances in Mathematics* 367 (June 2020), p. 107128.

[Com94] P. Comon. "Independent Component Analysis, A New Concept?" In: *Signal Processing* 36.3 (Apr. 1994), pp. 287–314.

[COV14] L. Chiantini, G. Ottaviani and N. Vannieuwenhoven. "An Algorithm for Generic and Low-Rank Specific Identifiability of Complex Tensors". In: *SIAM Journal on Matrix Analysis and Applications* 35.4 (2014), pp. 1265–1287.

[COV17] L. Chiantini, G. Ottaviani and N. Vannieuwenhoven. "Effective criteria for specific identifiability of tensors and forms". In: *SIAM Journal on Matrix Analysis and Applications* 38.2 (2017), pp. 656–681.

[DBV23a] N. Dewaele, P. Breiding and N. Vannieuwenhoven. "The condition number of many tensor decompositions is invariant under Tucker compression". In: *Numerical Algorithms* (June 2023).

[DBV23b] N. Dewaele, P. Breiding and N. Vannieuwenhoven. "Three decompositions of symmetric tensors have similar condition numbers". In: *Linear Algebra and its Applications* 664 (May 2023), pp. 253–263.

[Ded96] J.-P. Dedieu. "Approximate solutions of numerical problems, condition number analysis and condition number theorem". In: *The Mathematics of Numerical Analysis.* Vol. 32. Lectures in Applied Mathematics. Park City, Utah, United States: American Mathematical Society, 1996, pp. 263–283.

[Dem87] J. Demmel. "On condition numbers and the distance to the nearest ill-posed problem". In: *Numerische Mathematik* 51.3 (May 1987), pp. 251–289.

[DK02] J.-P. Dedieu and M.-H. Kim. "Newton's method for analytic systems of equations with constant rank derivatives". In: *Journal of Complexity* 18.1 (Mar. 2002), pp. 187–209.

[DL08] L. De Lathauwer. “Decompositions of a higher-order tensor in block Terms—Part II: Definitions and uniqueness”. In: *SIAM Journal on Matrix Analysis and Applications* 30.3 (Jan. 2008), pp. 1033–1066.

[DL11] L. De Lathauwer. “Blind separation of exponential polynomials and the decomposition of a tensor in rank-(L r,L r,1) terms”. In: *SIAM Journal on Matrix Analysis and Applications* 32.4 (2011), pp. 1451–1474.

[DLDMV00a] L. De Lathauwer, B. De Moor and J. Vandewalle. “A multilinear singular value decomposition”. In: *SIAM Journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1253–1278.

[DLDMV00b] L. De Lathauwer, B. De Moor and J. Vandewalle. “An Introduction to Independent Component Analysis”. In: *Journal of Chemometrics* 14.3 (2000), pp. 123–149.

[DLN08] L. De Lathauwer and D. Nion. “Decompositions of a higher-order tensor in block Terms—Part III: Alternating least squares algorithms”. In: *SIAM Journal on Matrix Analysis and Applications* 30.3 (Jan. 2008), pp. 1067–1083.

[DR14] A. L. Dontchev and R. T. Rockafellar. *Implicit functions and solution mappings*. Springer Series in Operations Research and Financial Engineering. New York, NY: Springer New York, 2014.

[DV23b] N. Dewaele and N. Vannieuwenhoven. “What part of a numerical problem is ill-conditioned?” In: *arXiv preprint arXiv:2305.11547* (2023).

[Dé00] J. Dégot. “A condition number theorem for underdetermined polynomial systems”. In: *Mathematics of Computation* 70.233 (July 2000), pp. 329–335.

[ERL22] V. Ehrlacher, M. F. Ruiz and D. Lombardi. “SOTT: Greedy approximation of a tensor as a sum of tensor trains”. In: *SIAM Journal on Scientific Computing* 44.2 (Apr. 2022), A664–A688.

[EY36] C. Eckart and G. Young. “The Approximation of One Matrix by Another of Lower Rank”. In: *Psychometrika* 1.3 (Sept. 1936), pp. 211–218.

[FKK22] H. Fan, J. Kileel and B. Kimia. “On the instability of relative pose estimation and RANSAC’s role”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, pp. 8925–8933.

[Fri16] S. Friedland. "Remarks on the symmetric rank of symmetric tensors". In: *SIAM Journal on Matrix Analysis and Applications* 37.1 (Jan. 2016). Publisher: Society for Industrial & Applied Mathematics (SIAM), pp. 320–337.

[Gal20] J. A. Gallian. *Contemporary Abstract Algebra.* Tenth edition. Boca Raton: Chapman & Hall/CRC, 2020.

[GK93] I. Gohberg and I. Koltracht. "Mixed, componentwise, and structured condition numbers". In: *SIAM Journal on Matrix Analysis and Applications* 14.3 (July 1993), pp. 688–704.

[Gle08] J. Gleick. *Chaos: Making a New Science.* 28th printing. London, England: Penguin Books, 2008.

[Gra10] L. Grasedyck. "Hierarchical singular value decomposition of tensors". In: *SIAM Journal on Matrix Analysis and Applications* 31.4 (2010), pp. 2029–2054.

[Gre78] W. Greub. *Multilinear algebra.* 2nd ed. New York: Springer-Verlag, 1978.

[GVL13] G. H. Golub and C. F. Van Loan. *Matrix computations.* Vol. 3. Baltimore: JHU press, 2013.

[Hac12] W. Hackbusch. *Tensor spaces and numerical tensor calculus.* Vol. 42. Springer Series in Computational Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[Har70] R. Harshman. "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis". In: *UCLA Working Papers in Phonetics* 16.10 (1970), pp. 1–84.

[Har95] J. Harris. *Algebraic Geometry: A First Course.* Corr. 3rd print. Graduate Texts in Mathematics 133. New York: Springer, 1995.

[Hau+19] J. D. Hauenstein, L. Oeding, G. Ottaviani and A. J. Sommese. "Homotopy techniques for tensor decomposition and perfect identifiability". In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 2019.753 (Aug. 2019), pp. 1–22.

[Hel76] F. R. Helmert. "Die genauigkeit der formel von peters zur berechnung des wahrscheinlichen beobachtungsfehlers directer beobachtungen gleicher genauigkeit". In: *Astronomische Nachrichten* 88 (1876), p. 113.

[Hig02] N. J. Higham. *Accuracy and Stability of Numerical Algorithms.* 2002.

[Hit27] F. L. Hitchcock. “The expression of a tensor or a polyadic as a sum of products”. In: *Journal of Mathematics and Physics* 6.1-4 (Apr. 1927), pp. 164–189.

[HJ10] R. A. Horn and C. R. Johnson. *Topics in matrix analysis.* Transferred to digital printing. Cambridge: Cambridge Univ. Press, 2010.

[HJ12] R. A. Horn and C. R. Johnson. *Matrix analysis.* 2nd ed. Cambridge ; New York: Cambridge University Press, 2012.

[HK09] W. Hackbusch and S. Kühn. “A new scheme for the tensor representation”. In: *Journal of Fourier Analysis and Applications* 15.5 (2009), pp. 706–722.

[HMT11] N. Halko, P. G. Martinsson and J. A. Tropp. “Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions”. In: *SIAM Review* 53.2 (Jan. 2011), pp. 217–288.

[HU17] W. Hackbusch and A. Uschmajew. “On the interconnection between the higher-order singular values of real tensors”. In: *Numerische Mathematik* 135.3 (Mar. 2017), pp. 875–894.

[Hun+14] B. Hunyadi, D. Camps, L. Sorber, W. V. Paesschen, M. D. Vos, S. V. Huffel and L. D. Lathauwer. “Block Term Decomposition for Modelling Epileptic Seizures”. In: *EURASIP Journal on Advances in Signal Processing* 2014.1 (Dec. 2014), p. 139.

[HZ03] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* 2nd ed. Cambridge, UK ; New York: Cambridge University Press, 2003.

[JM09] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* 2. ed. [Nachdr.] Upper Saddle River, NJ: Prentice Hall, 2009.

[Jou+10] M. Journée, F. Bach, P.-A. Absil and R. Sepulchre. “Low-rank optimization on the cone of positive semidefinite matrices”. In: *SIAM Journal on Optimization* 20.5 (Jan. 2010), pp. 2327–2351.

[KB09] T. G. Kolda and B. W. Bader. “Tensor decompositions and applications”. In: *SIAM Review* 51.3 (2009), pp. 455–500.

[KK22] J. Kileel and K. Kohn. *Snapshot of algebraic vision.* Oct. 2022.

[KL10] O. Koch and C. Lubich. “Dynamical tensor approximation”. In: *SIAM Journal on Matrix Analysis and Applications* 31.5 (Jan. 2010), pp. 2360–2375.

[KVM01] H. A. L. Kiers and I. Van Mechelen. “Three-way component analysis: Principles and illustrative application.” In: *Psychological Methods* 6.1 (2001), pp. 84–110.

[Lan12] J. M. Landsberg. *Tensors: Geometry and applications.* AMS, 2012.

[Lee11] J. M. Lee. *Introduction to topological manifolds.* Vol. 202. Graduate Texts in Mathematics. New York, NY: Springer New York, 2011.

[Lee13] J. M. Lee. *Introduction to smooth manifolds.* Springer New York, 2013.

[Lee18] J. M. Lee. *Introduction to Riemannian manifolds.* Vol. 176. Cham: Springer, 2018.

[LKB24] E. Levin, J. Kileel and N. Boumal. “The Effect of Smooth Parametrizations on Nonconvex Optimization Landscapes”. In: *Mathematical Programming* (Mar. 2024).

[LWY21] L.-H. Lim, K. S.-W. Wong and K. Ye. “The Grassmannian of affine subspaces”. In: *Foundations of Computational Mathematics* 21.2 (Apr. 2021), pp. 537–574.

[MD09] M. W. Mahoney and P. Drineas. “CUR matrix decompositions for improved data analysis”. In: *Proceedings of the National Academy of Sciences* 106.3 (Jan. 2009), pp. 697–702.

[Mir60] L. Mirsky. “Symmetric Gauge functions and unitarily invariant norms”. In: *The Quarterly Journal of Mathematics* 11.1 (1960), pp. 50–59.

[Mun14] J. Munkres. *Topology.* 2nd ed. Harlow: Pearson Education, 2014.

[MVL08] C. D. M. Martin and C. F. Van Loan. “A Jacobi-type method for computing orthogonal tensor decompositions”. In: *SIAM Journal on Matrix Analysis and Applications* 30.3 (Jan. 2008), pp. 1219–1232.

[NDLK08] C. Navasca, L. De Lathauwer and S. Kindermann. “Swamp reducing technique for tensor decomposition”. In: *European Signal Processing Conference* Eusipco (2008).

[NW06] J. Nocedal and S. J. Wright. “Numerical optimization”. In: *Numerical optimization.* Springer series in operations research and financial engineering. New York: Springer New York, 2006.

[OADL18] G. Olikier, P. A. Absil and L. De Lathauwer. “Variable projection applied to block term decomposition of higher-order tensors”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10891 LNCS.30468160 (2018). ISBN: 9783319937632, pp. 139–148.

[Orús14] R. Orús. “A practical introduction to tensor networks: Matrix product states and projected entangled pair states”. In: *Annals of Physics* 349 (2014), pp. 117–158.

[OS06] S. Oliveira and D. Stewart. *Writing Scientific Software: A Guide for Good Style.* Cambridge ; New York: Cambridge University Press, 2006.

[Ose11] I. V. Oseledets. “Tensor-train decomposition”. In: *SIAM Journal on Scientific Computing* 33.5 (Jan. 2011), pp. 2295–2317.

[PFS16] E. E. Papalexakis, C. Faloutsos and N. D. Sidiropoulos. “Tensors for Data Mining and Data Fusion: Models, Applications, and Scalable Algorithms”. In: *ACM Transactions on Intelligent Systems and Technology* 8.2 (2016).

[Pha+21] A.-H. Phan, P. Tichavský, K. Sobolev, K. Sozykin, D. Ermilov and A. Cichocki. “Canonical polyadic tensor decomposition with low-rank factor matrices”. In: *ICASSP 2021 - 2021 IEEE international conference on acoustics, speech and signal processing (ICASSP).* 2021, pp. 4690–4694.

[Ren95] J. Renegar. “Incorporating Condition Measures into the Complexity Theory of Linear Programming”. In: *SIAM Journal on Optimization* 5.3 (1995), pp. 506–524.

[Ric66] J. R. Rice. “A theory of condition”. In: *SIAM Journal on Numerical Analysis* 3.2 (June 1966), pp. 287–310.

[Saa03] Y. Saad. *Iterative methods for sparse linear systems.* SIAM, 2003.

[Sch07] E. Schmidt. “Zur Theorie der linearen und nichtlinearen Integralgleichungen: I. Teil: Entwicklung willkuerlicher Funktionen nach Systemen vorgeschriebener”. In: *Mathematische Annalen* 63.4 (Dec. 1907), pp. 433–476.

[Sha13] I. R. Shafarevich. “Basic algebraic geometry 1: Varieties in projective space”. In: *Basic algebraic geometry 1: Varieties in projective space.* Vol. 9783642379. Springer-Verlag Berlin Heidelberg, 2013, pp. 1–310.

[Sid+17] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis and C. Faloutsos. “Tensor decomposition for signal processing and machine learning”. In: *IEEE Transactions on Signal Processing* 65.13 (2017). Publisher: IEEE, pp. 3551–3582.

[SL08] V. de Silva and L.-H. Lim. “Tensor rank and the ill-posedness of the best low-rank approximation problem”. In: *SIAM Journal on Matrix Analysis and Applications* 30.3 (Jan. 2008), pp. 1084–1127.

[Smi+00] K. E. Smith, L. Kahanpää, P. Kekäläinen and W. Traves. *An invitation to algebraic geometry.* Universitext. New York, NY: Springer New York, 2000.

[SS90] G. Stewart and J. Sun. *Matrix perturbation theory.* Academic Press, Inc., 1990.

[SS93] M. Shub and S. Smale. “Complexity of Bezout’s theorem i: geometric aspects”. In: *Journal of the American Mathematical Society* 6.2 (Apr. 1993), p. 459.

[Sun96] J.-g. Sun. “Perturbation analysis of singular subspaces and deflating subspaces”. In: *Numerische Mathematik* 73.2 (Apr. 1996), pp. 235–263.

[SVBDL13] L. Sorber, M. Van Barel and L. De Lathauwer. “Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-(Lr, Lr, 1) terms, and a new generalization”. In: *SIAM Journal on Optimization* 23.2 (2013), pp. 695–720.

[Swi22] L. Swijsen. “Tensor Decompositions and Riemannian Optimization”. PhD thesis. KU Leuven, 2022.

[TB97] L. N. Trefethen and D. Bau. *Numerical linear algebra.* Philadelphia: Society for Industrial and Applied Mathematics, 1997.

[Tem60] G Temple. “Cartesian tensors”. In: *Methuen & Co., London and John Wiley & Sons, New York* (1960).

[Tuc66] L. R. Tucker. “Some mathematical notes on three-mode factor analysis”. In: *Psychometrika* 31.3 (1966). Publisher: Springer New York, pp. 279–311.

[Tur48] A. M. Turing. “Rounding-off Errors in Matrix Processes”. In: *The Quarterly Journal of Mechanics and Applied Mathematics* 1.1 (1948), pp. 287–308.

[UV13] A. Uschmajew and B. Vandereycken. “The geometry of algorithms using hierarchical tensors”. In: *Linear Algebra and Its Applications* 439.1 (2013), pp. 133–166.

[Van17] N. Vannieuwenhoven. “Condition numbers for the tensor rank decomposition”. In: *Linear Algebra and its Applications* 535 (Dec. 2017), pp. 35–86.

[Van23] N. Vannieuwenhoven. *The condition number of singular subspaces, revisited.* arXiv preprint. Aug. 2023.

[VDDL17] N. Vervliet, O. Debals and L. De Lathauwer. “Tensorlab 3.0 - Numerical optimization strategies for large-scale constrained and coupled matrix/tensor factorization”. In: *Conference Record - Asilomar Conference on Signals, Systems and Computers* 32 (2017), pp. 1733–1738.

[VVM12] N. Vannieuwenhoven, R. Vandebril and K. Meerbergen. “A new truncation strategy for the higher-order singular value decomposition”. In: *SIAM Journal on Scientific Computing* 34.2 (2012), pp. 1027–1052.

[ZG01] T. Zhang and G. H. Golub. “Rank-one approximation to high order tensors”. In: *SIAM Journal on Matrix Analysis and Applications* 23.2 (Jan. 2001), pp. 534–550.

Curriculum Vitae

Personal details

Name	Nick Dewaele
Born	10/04/1998, Wilrijk, Antwerpen, Belgium
ORCiD	0000-0002-5558-4782

Employment

2020-2024 Doctoral researcher at KU Leuven

2018-2020 Student teacher at KU Leuven. Courses *Probleemoplossen en ontwerpen, Analyse deel 3, Lineaire Algebra*, including development of course material for courses *Analyse deel 1, 2, 3.*

2016 and 2018 Intern at SCK-CEN, user interface development

Education

2020-2024
Doctor of engineering science: computer science, KU Leuven
Supervisors: Nick Vannieuwenhoven and Paul Breiding

2018-2020
Master of mathematical engineering *(Master in de ingenieurswetenschappen: wiskundige ingenieurstechnieken)*, KU Leuven
Thesis: Computing the tensor geometric mean. Supervisors: Raf Vandebril and Nick Vannieuwenhoven

2015-2018
Bachelor of computer science *(Bachelor in de informatica)*, KU Leuven

Teaching activities

- Teaching assistant for courses at KU Leuven: *Probleemoplossen en ontwerpen* (2018), *Analyse deel 3/Aanvullingen wiskunde* (2018-2019), *Lineaire Algebra/Toegepaste algebra en differentiaalvergelijkingen* (2018-2019), *Toepassingen van meetkunde in de informatica* (2020-2023), *Computergesteund probleemoplossen in de natuurkunde* (2022-2024).

- Development of course material for courses at KU Leuven *Analyse, deel 1, 2, 3* (2020).

- Mentoring for master's theses in mathematical engineering (3 students, 2020-2022) and bachelor project in computer science (one group, 2023-2024).

List of publications

Articles in international peer-reviewed journals

- N. Dewaele, P. Breiding and N. Vannieuwenhoven. “The condition number of many tensor decompositions is invariant under Tucker compression”. In: *Numerical Algorithms* (June 2023)
- N. Dewaele, P. Breiding and N. Vannieuwenhoven. “Three decompositions of symmetric tensors have similar condition numbers”. In: *Linear Algebra and its Applications* 664 (May 2023), pp. 253–263

Preprint articles

- N. Dewaele and N. Vannieuwenhoven. “What part of a numerical problem is ill-conditioned?” In: *arXiv preprint arXiv:2305.11547* (2023)

Talks at international conferences

- N. Dewaele and N. Vannieuwenhoven. “Condition Numbers of Tensor Factorisations”. In: *SIAM AG 2023, University of Eindhoven, Eindhoven, Netherlands.* 2023
- N. Dewaele and N. Vannieuwenhoven. “What part of a numerical problem is ill-conditioned?” In: *FOCM 2023, Sorbonne university, Paris, France.* 2023
- N. Dewaele, P. Breiding and N. Vannieuwenhoven. “Condition numbers of tensor decompositions”. In: Algebraic Geometry with Applications to TEnsors and Secants (AGATES), Warsaw, Poland, 2022

- N. Dewaele. “A condition number for underdetermined systems”. In: Algebraic Geometry with Applications to TEnsors and Secants (AGATES), Warsaw, Poland, 2022
- N. Dewaele, P. Breiding and N. Vannieuwenhoven. “Computing the condition number of tensor decompositions through Tucker compression”. In: DMV Jahrestagung 2022, Berlin, Germany, 2022
- N. Dewaele, P. Breiding and N. Vannieuwenhoven. “Computing the condition number of tensor decompositions through Tucker compression”. In: 7th IMA Conference on Numerical Linear Algebra and Optimization, Birmingham, United Kingdom, 2022
- N. Dewaele, P. Breiding and N. Vannieuwenhoven. “Computing the condition number of tensor decompositions through Tucker compression”. In: Matrix Equations and Tensor Techniques IX (METTIX), Perugia, Italy, 2021

Posters at international conferences

- N. Dewaele. “Sensitivity of roots of underdetermined systems”. In: Centrum Wiskunde en Informatica, Amsterdam, 2021
- N. Dewaele, P. Breiding and N. Vannieuwenhoven. “Sensitivity of block term decompositions”. In: IPAM Workshop on Mathematical Foundations and Algorithms for Tensor Computations (online), 2021

Seminars

- N. Dewaele and N. Vannieuwenhoven. “Which constraints of a numerical problem cause ill-conditioning?” In: NUMA Seminar, Leuven, Belgium, 2024
- N. Dewaele. “A condition number for underdetermined systems”. In: Oberseminar Algebra, University of Osnabrueck, Germany, 2023
- N. Dewaele, P. Breiding and N. Vannieuwenhoven. “Computing the condition number of tensor decompositions through Tucker compression”. In: Max Planck institute for mathematics in the sciences, Leipzig, Germany (hosted online), 2021
- N. Dewaele, P. Breiding and N. Vannieuwenhoven. “Computing the condition number of tensor decompositions through Tucker compression”. In: NUMA Seminar, Leuven, Belgium, 2021

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
NUMERICAL ANALYSIS AND APPLIED MATHEMATICS
Celestijnenlaan 200A
B-3001 Leuven
http://numa.cs.kuleuven.be