TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta přírodovědně-humanitní
a pedagogická

Tenzorové sítě a hierarchický Tuckerův rozklad

Diplomová práce

Studijní program: N1101 – Matematika
Studijní obory: 7504T089 – Učitelství matematiky pro střední školy
7503T009 – Učitelství anglického jazyka pro 2. stupeň základní školy

Autor práce: **Jana Žáková**
Vedoucí práce: Martin Plešinger

Liberec 2017

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Science, Humanities
and Education

Tensor networks and hierarchical Tucker decomposition

Diploma thesis

Study programme: N1101 – Mathematics
Study branches: 7504T089 – Teacher training for upper-sec. schools: Mathematics
7503T009 – Teacher training for lower-secondary schools: English

Author: **Jana Žáková**
Supervisor: Martin Plešinger

Liberec 2017

Technická univerzita v Liberci
Fakulta přírodovědně-humanitní a pedagogická
Akademický rok: **2015/2016**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Jana Žáková**
Osobní číslo:	**P15000611**
Studijní program:	**N1101 Matematika**
Studijní obory:	**Učitelství anglického jazyka pro 2. stupeň základní školy**
	Učitelství matematiky pro střední školy
Název tématu:	**Tenzorové sítě a hierarchický Tuckerův rozklad**
Zadávající katedra:	**Katedra matematiky a didaktiky matematiky**

Zásady pro vypracování:

Při manipulaci a výpočtech s tenzory velkých řádů snadno narazíme na omezení na straně paměti počítače. Pokusíme-li se takový tenzor naivně uložit jako prosté vícerozměrné pole, rychle může dojít k vyčerpání dostupné paměti, které nelze jednoduše vyřešit použitím lepšího počítače. Tento jev je ve výpočetním světě známý pod anglickým termínem "curse of dimensionality".
Diplomová práce čtenáři ukáže, že na tenzory, resp. sady tenzorů a jejich různých vzájemných součinů lze jednoduše nahlížet jako na grafy, t. zv. tenzorové sítě. V práci bude dále ukázáno jak lze tenzor příliš velkého řádu na to, aby s ním šlo pracovat přímo, resp. jeho Tuckerovo jádro, šikovně rozložit do tenzorové sítě, jejíž uzly jsou tvořeny tenzory malých řádů. Tento přístup, t. zv. hierarchický Tuckerův rozklad povede k úspoře paměťových a výpočetních nákladů při ukládání tenzoru, resp. při manipulaci s tenzorem (násobení tenzoru maticí, lineární kombinace tenzorů) v počítači.
Požadavky: Základní znalosti z lineární algebry, základní znalost anglického jazyka. Práce by měla být psána tak, aby mohla celá, nebo její části, sloužit jako materiál pro úvod studia dané problematiky. Práce by měla být psaná v LaTeXu, bude-li to v možnostech studenta.

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

**T. G. Kolda, B. W. Bader: Tensor Decompositions and Applications.
SIAM Review,
Volume 51 (2009), Number 3, pp. 455–500.
L. Grasedyck: Hierarchical Singular Value Decomposition of Tensors.
SIAM Journal on Matrix Analysis and Applications,
Volume 31 (2010), Number 4, pp. 2029–2054.
I. V. Oseledets: Tensor-Train Decomposition.
SIAM Journal on Scientific Computing,
Volume 33 (2011), Number 5, pp. 2295–2317.
C. Tobler: Low-rank Tensor Methods for Linear Systems and Eigenvalue Problems.
PhD Tesis, Seminar for Applied Mathematics,
Department of Mathematics, ETH Zurich, 2012.**

Vedoucí diplomové práce: **Ing. Martin Plešinger, Ph.D.**
Katedra matematiky a didaktiky matematiky

Datum zadání diplomové práce: **6. května 2016**

Termín odevzdání diplomové práce: **28. dubna 2017**

prof. RNDr. Jan Picek, CSc.
děkan

L.S.

Technická univerzita v Liberci - 10 -

doc. RNDr. Jaroslav Mlýnek, CSc.
vedoucí katedry

V Liberci dne 10. května 2016

Prohlášení

Byla jsem seznámena s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé diplomové práce pro vnitřní potřebu TUL.

Užiji-li diplomovou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědoma povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Diplomovou práci jsem vypracovala samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé diplomové práce a konzultantem.

Současně čestně prohlašuji, že tištěná verze práce se shoduje s elektronickou verzí, vloženou do IS STAG.

Datum:

Podpis:

Anotace

V moderní numerická algebře se stále častěji setkáváme s problémy, kde potřebujeme pracovat s vícerozměrnými daty, uloženými jako tenzory. Při manipulaci a výpočtech s tenzory velkých řádů snadno narazíme na omezení na straně paměti počítače. Pokusíme-li se takový tenzor naivně uložit jako prosté vícerozměrné pole, rychle může dojít k vyčerpání dostupné paměti, které nelze jednoduše vyřešit použitím lepšího počítače. Tento jev je ve výpočetním světě známý pod anglickým termínem „curse of dimensionality“. Jedním z nástrojů, které umožňují snížit paměťové nároky, je Tuckerův rozklad tenzoru. Úspora je ovšem omezena tzv. vektorovou hodností tenzoru a pro tenzory vyšších řádů není dostatečná.

Cílem této práce je ukázat, že na tenzory, resp. sady tenzorů a jejich různých vzájemných součinů lze nahlížet jako na specifické neorientované grafy. Vysvětluje způsob reprezentace tenzorů a další objektů lineární algebry pomocí těchto grafů. Takový způsob reprezentace tenzorů označujeme jako tzv. tenzorové sítě.

V práci je dále ukázáno, jak lze tenzor (příliš velkého řádu na to, aby s ním šlo pracovat přímo) šikovně rozložit do tenzorové sítě se strukturou binárního stromu, jejíž uzly jsou tvořeny tenzory malých řádů; konkrétně řádů tři a dva. Navíc počet těchto tenzorů malých řádů závisí na řádu původního tenzoru lineárně. Tento přístup, tzv. hierarchický Tuckerův rozklad (HTD, z anglického hierarchical Tucker decomposition) může vést k úspoře paměťových a výpočetních nákladů při ukládání tenzoru, resp. při manipulaci s tenzorem (násobení tenzoru maticí, lineární kombinace tenzorů) v počítači.

Práce také vysvětluje, jakým způsobem s tenzory uloženými ve tvaru HTD provádět vybrané základní algebraické operace tak, aby výsledek byl opět tenzor v podobě hierarchického Tuckerova rozkladu.

Klíčová slova:

multilineární algebra; tenzor; tenzrová síť; (hierarchický) Tuckerův rozklad (HTD); tensor train; tensor chain; operace s tenzory; low-rank aproximace

Abstract

In modern numerical algebra, there quite frequently arise problems, where there is a need to work with multidimensional data stored in the form of tensors. While manipulating or calculating with tensors of high order, we often encounter the restricitions by the memory of the computer. The attempt to store such a tensor can lead to the exhaustion of the available memory, which can not be improved by the use of a better computer. This problem is referred as the curse of dimensionality. One of the tools used for the reduction of the storage requirements is the so-called Tucker decomposition. However, the storage savings by this decomposition are restricted by the vector-rank of the given tensor and are not sufficient for tensors of high order.

The aim of this thesis is to explain how tensors, or sets of tensors and tensor products can be interpreted as specific (undirected) graphs. We explain the way of the representation of tensors and other objects from linear algebra. Such representation is called the tensor network.

In the text we show the way to decompose the tensor (of order which is too high) into the tensor network of the binary tree structure. The nodes of such a tree represent tensors of order two or three. Moreover, the number of these tensors of low order depends linearly on the order of the original tensor. This approach, called hierarchical Tucker decomposition (HTD), can lead to storage requirements and computation savings while storing or manipulating with the tensor, respectively.

The thesis also explains how to do some selected basic arithmetic operations so that the result is also a tensor in HTD format.

Key words:

multilinear algebra; tensor; tensor network diagram; (hierarchical) Tucker decomposition (HTD); tensor train; tensor chain; tensor arithmetic; low-rank approximation

Poděkování

Ráda bych na tomto místě poděkovala všem, kteří se zasloužili o to, že jsem mohla vytvořit tuto práci. Děkuji mým rodičům, kteří mě podporovali po celou dobu studia nejen materiálně, ale i psychicky, a také mým přátelům a spolužákům, od nichž se mi vždy dostalo pomoci a povzbuzení. Zejména ale děkuji Martinu Plešingerovi za jeho cenné rady, nadšení, trpělivost a čas věnovaný konzultacím.

Obsah

Seznam obrázků

Seznam tabulek

Použité značení a zkratky

V textu značíme

vektory (tenzory řádu 1)	pomocí malých písmen u_1, u_2, u_r, v_1, v_2, v_r, x, atd.,
matice (tenzory řádu 2)	pomocí velkých písmen (latinských i řeckých) A, B, C, D, E, F, U, V, Σ, atd.,
tenzory řádu k, $k \geq 3$	pomocí velkých písmen psaných kaligraficky $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$, $\mathcal{E}$, $\mathcal{F}$, $\mathcal{T}$, $\mathcal{S}$, atd.,
množiny	pomocí velkých písmen psaných Scriptem $\mathscr{D}$, $\mathscr{S}$, $\mathscr{T}$, $\mathscr{X}_j$,
číselné obory	pomocí velkých zdvojených písmen $\mathbb{N}$, $\mathbb{R}$ atd., speciálně $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

Pomocí malých písmen (latinských i řeckých) také značíme prvky matic a tenzorů a také skaláry (tenzory řádu 0). Speciální význam pak mají písmena i, j, ℓ, jimiž zpravidla indexujeme prvky matic a tenzorů, a k, m, n, r, která používáme k označení řádu tenzoru, dimenze matice nebo tenzoru, resp. hodnosti (ranku) matice nebo tenzoru.

Matice a vektory

Značení	**Význam**
$A \in \mathbb{R}^{n \times m}$	reálná matice s rozměry n krát m, s prvky $a_{i,j}$
$\operatorname{vec}(A) \in \mathbb{R}^{nm}$	vektorizace matice $A \in \mathbb{R}^{n \times m}$
$A \otimes B$	Kroneckerův součin dvou matic
A^T	transpozice matice A
$\operatorname{rank}(A)$	hodnost matice definovaná jako počet lineárně nazávislých řádků, resp. sloupců matice A
$\|x\| = (\sum_i x_i^2)^{1/2}$	eukleidovská norma vektoru

Tenzory

Značení	Význam
$\mathcal{A} = (a_{i_1,i_2,i_3}) \in \mathbb{R}^{n_1 \times n_2 \times n_3}$	tenzor třetího řádu o rozměrech n_1, n_2, n_3
$\mathcal{A} = (a_{i_1,\ldots,i_k}) \in \mathbb{R}^{n_1 \times \cdots \times n_k}$	tenzor k-tého řádu o rozměrech $n_1, \ldots, n_k$
$a_{:,i_2,i_3} \in \mathbb{R}^{n_1}$	vlákno tenzoru třetího řádu v módu 1
$A_{:,:,i_3} \in \mathbb{R}^{n_1 \times n_2}$	řez tenzoru třetího řádu v módu (1, 2)
$\mathrm{vec}(\mathcal{A}) \in \mathbb{R}^{n_1 n_2 \cdots n_k}$	vektorizace tenzoru $\mathcal{A} \in \mathbb{R}^{n_1 \times \cdots \times n_k}$
$A^{\{\ell\}} \in \mathbb{R}^{n_\ell \times ((\prod_{j=1}^{k} n_j)/n_\ell)}$	rozvoj tenzoru do matice v ℓ-tém módu
$A^{\{t_1,\ldots,t_d\}}$	rozvoj tenzoru do matice dle multiindexu $\{t_1, \ldots, t_d\}$
$A^{\mathscr{C}}$	rozvoj tenzoru do matice dle multiindexu $\mathscr{C}$
$\mathcal{A} \times_\ell M$	násobení tenzoru maticí v ℓ-tém módu; platí $(\mathcal{A} \times_\ell M)^{\{\ell\}} = MA^{\{\ell\}}$
$\mathcal{A} \times_{\ell,s} \mathcal{B}$	tenzorový součin v módech ℓ a s
$\mathcal{A} \times_{(\ell_1,\ell_2),(s_1,s_2)} \mathcal{B}$	tenzorový součin ve dvojici módů

Použité zkratky a akronymy

Zkratka	Význam
QR	QR rozklad matice, $A = QR$
SVD	singulární rozklad matice (singular value decomposition), $A = U\Sigma V^T$
HOSVD	Tuckerův rozklad tenzoru (high-order SVD)
HTD	hierarchický Tuckerův rozklad (hiearachical Tucker decomposition)
TT	tensor train
TC	tensor chain
MMp	součin matice s maticí (matrix-matrix product)
TMp	součin tenzoru s maticí (tensor-matrix product)

Úvod

V numerických výpočtech se často setkáváme s potřebou uložit data uspořádaná podle určitých parametrů. *Tenzory* jsou algebraickými objekty, které nám toto umožňují. Počet parametrů udává řád tenzoru, speciálně tenzor se dvěma rozměry tzv. tenzor druhého řádu je matice, tenzor prvního řádu je vektor. Tenzorem k-tého řádu $\mathcal{T} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k}$ tedy budeme v celém tomto textu rozumět k-rozměrné pole čísel. Obor zabývající se prací s tenzory řádu $k \geq 3$ se nazývá multilineární algebra; ta jistým způsobem rozšiřuje klasickou lineární algebru pracující se skaláry, vektory a maticemi.

Tenzorové výpočty ovšem skýtají problémy praktického rázu, začínající už od potřeby tenzor nějak uložit v počítači. Počet prvků tenzoru roste exponenciálně s řádem tenzoru k, což vede k faktu, že tenzory vysokých řádů prakticky nelze uložit, v anglické literatuře se setkáváme s termínem „curse of dimensionality“. Proto jsou stále hledány způsoby, jak tento problém eliminovat a efektivně tenzor uložit, například v podobě součinu tenzorů s nižší paměťovou náročností – tenzorových rozkladů, a následně možnosti jak data komprimovat, tj. efektivně aproximovat bez velkých ztrát informací.

Úkolem této práce je představit možnosti reprezentace tenzorů a tenzorových rozkladů, s využitím teorie grafů, v podobě tzv. tenzorových sítí. Zavádíme přitom tzv. multigraf s volně visícími hranami a smyčkami, kde vrcholy reprezentují tenzory v síti a hrany mezi nimi znázorňují součiny v odpovídajících módech. Speciálně, grafem tenzorové sítě v podobě binárního stromu lze reprezentovat tzv. hierarchický Tuckerův rozklad tenzoru, jehož analýze se v práci věnujeme podrobně.

Dalším cílem tohoto textu je tedy výklad hierarchického Tuckerova rozkladu (HTD, z anglického hierarchical Tucker decomposition), zaměřující se zejména na principy odvození tohoto nástroje, naopak praktickou implementaci výpočtu pouze naznačíme. Původní idea je ve zobecnění singulárního rozkladu matic, která vede na tzv. Tuckerův rozklad, viz viz [20], [21], [22], [11] nebo [19], který umožňuje uložit tenzor řádu k jako součin menšího tenzoru stejného řádu, tzv. Tuckerova jádra, a k matic. Rozměry jádra jsou dány tzv. vektorovou hodností, tj. k-tící čísel vyjadřující hodnosti rozvojů tenzoru v jednotlivých módech. Jak je ale ukázáno např. v [26], tento rozklad velmi často (pro tenzory vyšších řádů) nepřinese dostatečnou úsporu paměťových nároků.

Hierarchický Tuckerův rozklad fakticky pracuje s jádrem tenzoru, které rozkládá do součinu tenzorů třetího (resp. druhého) řádu, reprezentovaného sítí, konkrétně binárním stromem, námi předepsaného tvaru. Jako HTD však typicky uvažujeme síť ve tvaru maximálně vyváženého binárního stromu. Obdobou HTD, která je od-

vozena stejným způsobem až na volbu struktury stromu je tzv. tensor train, kdy je naopak volen maximálně nevyvážený binární strom.

Text je strukturován následujícím způsobem. Po stručném úvodu v kapitole 1 zavádíme tenzor jako vícerozměrné pole čísel a vysvětujeme základní operace s tenzory, které budeme v textu dále používat, konkrétně součin tenzoru s maticí, rozvoj tenzoru do matice a Tuckerův rozklad tenzoru. Kapitola 2 je věnována pojmům z teorie grafů, které budeme potřebovat proto, abychom v následující kapitole 3 vysvětlili, jakým způsobem grafy využíváme k reprezentaci tenzorů a také jejich součinů – tzv. tenzorových sítí. Kapitola 4 je věnována hierarchickému Tuckerovu rozkladu (HTD), kde ukazujeme princip, na kterém je tento rozklad postaven a dokazujeme jeho existenci. Porovnáváme také paměťové nároky při ukládání tenzoru do počítače různými způsoby a ukazujeme efektivitu uložení dat v HTD. V kapitole 5 vysvětlujeme, jak lze provádět vybrané základní operace s tenzory, jsou-li tyto uloženy v HTD tak, abychom i výsledný tenzor získali v tomto tvaru. V kapitole 6 ukazujeme, jakým způsobem lze HTD tenzoru spočítat v dané konkrétní situaci.

1 Tenzory a základní manipulace s nimi, Tuckerův rozklad

V první kapitole vysvětlíme pojem tenzor a ukážeme základní nástroje potřebné pro práci s tenzory. Připomeneme také Tuckerův rozklad tenzoru jako zobecnění singulárního rozkladu matice.

Definice 1 (Tenzor)**.** *Nechť $\mathcal{T}$ je k-rozměrné pole reálných čísel $t_{i_1,i_2,\ldots,i_k}$ o rozměrech $n_1, n_2, \ldots, n_k$. Potom*

$$\mathcal{T} = (t_{i_1,i_2,\ldots,i_k}) \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k} \tag{1.1}$$

nazýváme tenzor k-tého řádu, viz [10], [11].

Aritmetické vektory a matice považujeme za tenzory prvního, resp. druhého řádu. Tenzory stejného řádu, které mají stejné rozměry, tvoří společně s operacemi sčítání (po prvcích) a násobení skalárem vektorový prostor.

Důležitou operací pro práci s tenzory je násobení. Následující definice ukazuje, jak násobit tenzor s maticí.

Definice 2 (Součin v ℓ-tém módu)**.** *Nechť $\mathcal{T}$ je tenzor (1.1) a $M \in \mathbb{R}^{m \times n_\ell}$ je matice s prvky $m_{i,j}$, kde $i = 1, \ldots, m$, $j = 1, \ldots, n_\ell$. Potom*

$$\mathcal{D} = \mathcal{T} \times_\ell M \equiv \left(\sum_{\alpha=1}^{n_\ell} t_{i_1,\ldots,i_{\ell-1},\alpha,i_{\ell+1},\ldots,i_k} \cdot m_{i,\alpha} \right) \in \mathbb{R}^{n_1 \times \cdots \times n_{\ell-1} \times m \times n_{\ell+1} \times \cdots \times n_k} \tag{1.2}$$

se nazývá součin v ℓ-tém módu viz [11].

Vzhledem k faktu, že tenzor můžeme považovat za zobecnění matice, je samozřejmě možné definovat součin dvou tenzorů obecněji, analogicky k (1.2), což popíšeme následně.

Mějme tenzory $\mathcal{T} = (t_{i_1,i_2,\ldots,i_k}) \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k}$ a $\mathcal{S} = (s_{j_1,j_2,\ldots,j_\ell}) \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_\ell}$. Pokud $n_p = m_q \equiv \mu$, potom

$$\mathcal{F} = \mathcal{T} \times_{(p,q)} \mathcal{S} \equiv \left(\sum_{\alpha=1}^{\mu} t_{i_1,\ldots,i_{p-1},\alpha,i_{p+1}\ldots,i_k} \cdot s_{j_1,\ldots,j_{q-1},\alpha,j_{q+1}\ldots,j_\ell} \right) \in \mathbb{R}^{n_1 \times \cdots \times n_{p-1} \times n_{p+1} \times \cdots \times n_k \times m_1 \times \cdots \times m_{q-1} \times m_{q+1} \times \cdots \times m_\ell} \tag{1.3}$$

považujeme za součin těchto tenzorů. Tento druh součinu můžeme v některých zdrojích nalézt pod pojmy úžení tenzorů, viz [15, str. 70], contracted product, viz [2, str. 643] nebo tensor-tensor contraction, viz [19, str. 31].

S tenzory můžeme někdy pracovat i ve tvaru matice. Pro to bude užitené definovat tzv. rozvoj tenzoru v matici v daném módu, příp. módech.

Definice 3 (Rozvoj tenzoru v matici)**.** *Uvažujme tenzor $\mathcal{T}$ (1.1) a jeho sadu indexů rozdělenou do dvou disjunktních podmnožin $\mathscr{R}$ a $\mathscr{C}$, přičemž $\mathscr{R} \equiv \{r_1, r_2, \ldots, r_R\}$ a $\mathscr{C} \equiv \{c_1, c_2, \ldots, c_C\}$ a zároveň $\mathscr{R} \cup \mathscr{C} = \{1, 2, \ldots, k\}$ a navíc platí $r_1 < r_2 < \cdots < r_R$ a $c_1 < c_2 < \cdots < c_C$. Matice*

$$\mathcal{T}^{\mathscr{R}} = \mathcal{T}^{\{r_1, r_2, \ldots, r_R\}} \in \mathbb{R}^{N_R \times N_C}, \quad kde \quad N_R = \prod_{\ell=1}^{R} n_{r_\ell}, \quad N_C = \prod_{\ell=1}^{C} n_{c_\ell}, \tag{1.4}$$

obsahující prvky $t_{i_1, i_2, \ldots, i_k}$ v řádcích s multiindexy $(i_{r_R}, \ldots, i_{r_2}, i_{r_1})$ a ve sloupcích s multiindexy $(i_{c_C}, \ldots, i_{c_2}, i_{c_1})$ v lexikografickém pořadí se nazývá rozvoj tenzoru v matici (v angličtině matricization), viz [19, kap. 3.1.2].

Speciálním případem rozvoje tenzoru je tzv. *rozvoj tenzoru v ℓ-tém módu*, kde jedna z množin multiindexů obsahuje pouze jeden prvek, tj. $\mathscr{R} = \{\ell\}$, $\mathscr{C} = \{1, \ldots, k\} \setminus \{\ell\}$. V tomto případě dostáváme

$$\mathcal{T}^{\{\ell\}} \in \mathbb{R}^{n_\ell \times ((\prod_{j=1}^{k} n_j)/n_\ell)}, \tag{1.5}$$

viz [11]. Další speciální případ nastává, pokud je množina $\mathscr{C}$ prázdná. Dostáváme potom sloupcový vektor obsahující všechny prvky tenzoru $\mathcal{T}$, tzv. *vektorizaci tenzoru*, kterou označujeme $\mathrm{vec}(\mathcal{T})$.

1.1 Tuckerův rozklad

Podobně jako v lineární algebře je velice užitečným nástrojem singulární rozklad matice (SVD z anglického singular value decomposition), existuje ve vícerozměrném případě (tj. pro řád tři a více) jeho zobecnění, tzv. Tuckerův rozklad. V některých zdrojích se můžeme setkat také s názvy high-order SVD (HOSVD), tedy singulární rozklad vyšších řádů. Zřejmě pro matice (tenzory řádu dva) odpovídá Tuckerův rozklad přímo singulárnímu rozkladu matice.

Definice 4 (Tuckerův rozklad)**.** *Nechť $\mathcal{T}$ je tenzor řádu k. Potom*

$$\mathcal{T} = \mathcal{S} \times_1 U_1 \times_2 U_2 \times \ldots \times_k U_k, \quad U_\ell = U_\ell^{-1}, \tag{1.6}$$

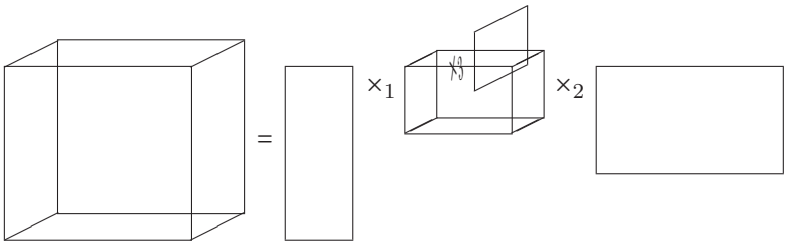
kde U_ℓ jsou matice s levými singulárními vektory matic rozvojů tenzoru $\mathcal{T}$ v daných módech, tj. $\mathcal{T}^{\{\ell\}}$, $\ell = 1, 2, \ldots, k$, nazýváme Tuckerův rozklad tenzoru $\mathcal{T}$. Tenzor $\mathcal{S}$ se nazývá Tuckerovo jádro.

Tuckerův rozklad tenzoru je zřejmě zobecněním singulárního rozkladu matice. Pokud čísla $r_1, r_2, \ldots, r_k$ jsou hodnosti rozvojů v módech $1, 2, \ldots, k$, posledních $(n_\ell - r_\ell)$

řádků matic $\mathcal{T}^{\{\ell\}}$ obsahuje pouze nuly. Tuckerův rozklad potom můžeme stejně jako singulární rozklad tenzoru vyjádřit v tzv. ekonomickém tvaru, tj.

$$\mathcal{T} = \mathcal{S}_{\mathcal{T}} \times_1 U_1' \times_2 U_2' \times \cdots \times_k U_k', \qquad S_{\mathcal{T}} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_k}, \qquad U_\ell \in \mathbb{R}^{n_\ell \times r_\ell}, \tag{1.7}$$

viz [20], [21], [22], případně [11, kap. 4.1] a [19, kap. 4.1], pro ilustraci Tuckerova rozkladu viz obrázek 1.1.



Obrázek 1.1: Tuckerův rozklad (HOSVD) tenzoru řádu 3. Převzato z [11].

Poznamenejme, že minimální rozměry Tuckerova jádra, tj. právě rozměry tenzoru $\mathcal{S}_{\mathcal{T}}$ z (1.7) se nazývají (vektorová) hodnost tenzoru, případně vektorový rank, viz [11, kap. 3]. Tuckerovo jádro v případě obecného tenzoru však na rozdíl od maticového případu nemá diagonální strukturu, ale je obecně hustý tenzor.

Pro podrobnější přehled o manipulaci s tenzory ve tvaru Tuckerova rozkladu nebo jeho využití k aproximaci tenzoru tenzorem nižší hodnosti viz [26].

2 Grafy

V této kapitole se seznámíme se základními pojmy týkající se teorie grafů. Tyto poznatky budeme dále potřebovat v dalších kapitolách pro vysvětlení tenzorových sítí. Při zavádění pojmů vycházíme zejména z [16].

2.1 Základní pojmy teorie grafů

Grafy jsou prostředkem pro vyjádření nějaké množiny bodů a vztahů mezi nimi. Tyto body nazýváme vrcholy nebo uzly grafu a příslušné vztahy mezi nimi jsou vyjádřené spojnicemi, které nazýváme hrany grafu. Grafy lze definovat různě, nejčastěji se setkáváme s následujícími definicemi:

Definice 5 (Orientovaný graf)**.** *Orientovaný graf G je uspořádaná dvojice (V, H), kde $V = \{v_1, v_2, \ldots, v_n\}$ je nějaká neprázdná množina a*

$$H \subseteq V \times V \equiv \big\{(v_i, v_j) : \ i, j \in \{1, \ldots, n\}\big\} \tag{2.1}$$

je množina uspořádaných dvojic množiny V.

Definice 6 (Neorientovaný graf)**.** *Neorientovaný graf G je opět uspořádaná dvojice (V, H), kde $V = \{v_1, v_2, \ldots, v_n\}$ je nějaká neprázdná množina, ale*

$$H \subseteq \binom{V}{2} \equiv \big\{\{v_i, v_j\} : \ i, j \in \{1, \ldots, n\}\big\} \tag{2.2}$$

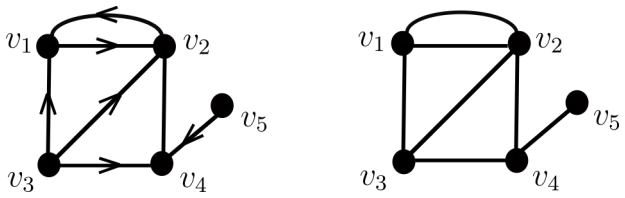
je množina dvouprkvkových podmnožin množiny V.

Prvky množiny V se nazývají *vrcholy grafu* G (někdy také *uzly*) a prvky množiny H *hrany* grafu G. Příklad neorientovaného i orientovaného grafu je na obrázku 2.1.

2.1.1 Volně visící hrany, multi-hrany a smyčky

My budeme dále pracovat pouze s neorientovanými grafy. Budeme ale *potřebovat* navíc zavést:

- tzv. *volně visící hrany* (anglicky *dangling edges*, viz [19, str. 29]), tj. hrany, které mají pouze jeden vrchol;
- více hran mezi jednou dvojicí vrcholů, tzv. *násobnost hran.*



Obrázek 2.1: Orientovaný graf (vlevo), šipkami je znázorněna orientace hran; zatímco např. (v_5, v_4) je hranou, (v_4, v_5) hranou není. Neorientovaný graf (vpravo); zde je hranou $\{v_4, v_5\}$.

Pro zavedení obecné tenzorové sítě navíc bude vhodné uvažovat grafy, které mohou obsahovat:

- hrany, které začínají a končí ve stejném vrcholu, tzv. *smyčky.*

Poznamenejme, že definice 5, na rozdíl od definice 6, existenci smyček umožňuje.

Těchto rozšíření pojmu neorientovaného grafu můžeme docílit např. následujícími konstrukcemi: Množinu V nahradíme množinou $V \cup \{f\} = \{f, v_1, v_2, \ldots, v_n\}$, tedy přidáme speciální vrchol f, přičemž hrany typu $\{f, v_i\}$ budeme nazývat *volně visící hrany* (později budou odpovídat tzv. fyzickým indexům tenzoru). Násobnost vyřešíme zavedením tzv. *multigrafu*, viz [16, str. 139]. Smyčky jsou neorientované hrany s oběma konci ve stejném vrcholu; formálně je můžeme považovat za prvky množiny jednoprvkových podmnožin množiny V, tj. množiny

$$\binom{V}{1} \equiv \big\{\{v_i\},\ i \in \{1, \ldots, n\}\big\}.$$

Všechna tato rozšíření shrneme v následující definici.

Definice 7 (Neorientovaný multigraf s volně visícími hranami a smyčkami)**.** *Uspořádanou dvojici* $G = (V \cup \{f\}, \mu)$*, kde* $V = \{v_1, v_2, \ldots v_n\}$ *a*

$$\mu : \binom{V \cup \{f\}}{2} \cup \binom{V}{1} \longrightarrow \mathbb{N}_0,$$

budeme nazývat neorientovaný multigraf s volně visícími hranami a smyčkami.

Prvky množiny V *nazýváme vrcholy a prvek* f *nazýváme volný vrchol. Prvky množiny* $\binom{V\cup\{f\}}{2} \cup \binom{V}{1}$ *nazýváme hranami, přičemž prvky typu* $\{v_i, v_j\}$ *jsou hrany v klasickém slova smyslu, prvky typu* $\{f, v_i\}$ *jsou volně visící hrany a prvky typu* $\{v_i\}$ *jsou smyčky.*

Zobrazení m *přiřadí každé hraně* $h_\ell \in \binom{V\cup\{f\}}{2} \cup \binom{V}{1}$ *násobnost* $\mu(h_\ell)$*. Pokud je násobnost hrany* $\mu(h_\ell) = 0$*, hrana není v grafu přítomna;* $\mu(h_\ell) = 1$ *znamená, že hrana je jednoduchá;* $\mu(h_\ell) > 1$ *znamená násobnou hran, tzv. multi-hranu.*

V následujícím textu budeme slovem *graf* téměř výhradně rozumět právě neorientovaný multigraf s volně visícími hranami a smyčkami.

2.1.2 Stupeň vrcholu

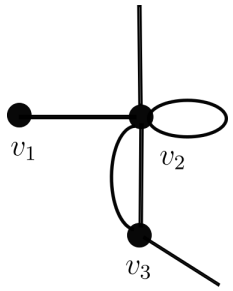
Mějme graf G, který obsahuje vrchol v. Počet hran, ve kterých je přítomen vrcholu v, nazýváme *stupeň vrcholu* v v grafu G. Toto číslo označujeme $\deg(v)$, viz [16]. Vrcholy grafu, které mají stupeň 0 se nazývají *izolované*. Stupěň vrcholu závisí na typech vrcholů a násobnostech hran. Stupeň vrcholu v_i můžeme vyjádřit jako následující součet:

$$\deg(v_i) = \sum_{\substack{j=1 \\ j\neq i}}^{n} \mu(\{v_i, v_j\}) + \mu(\{v_i, f\}) + 2\mu(\{v_i\}), \tag{2.3}$$

kde sčítáme počty klasických, volně visících hran a smyček, které vedou z daného vrcholu, zatímco stupeň vrcholu f je dán následující rovností

$$\deg(f) = \sum_{j}^{n} \mu(\{v_j, f\}), \tag{2.4}$$

kde sčítáme pouze volně visící hrany. Poznamenejme, že smyčka z vrcholu f není přípustná. Pojem stupeň vrcholu ilustruje obrázek 2.2.



Obrázek 2.2: Graf se čtyřmi vrcholy v_1, v_2, v_3 a f; zde $\deg(v_j) = 1, 6, 3$, postupně pro $j = 1, 2, 3$, a $\deg(f) = 2$.

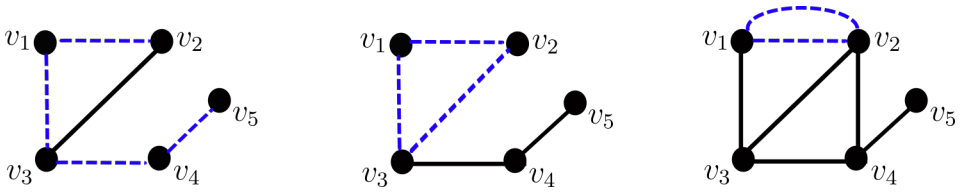
2.1.3 Cesta a kružnice, souvislý graf a strom

Nyní vysvětlíme několik pojmů, které se často vyskytují při práci s grafy a i my je v této práci budemem používat.

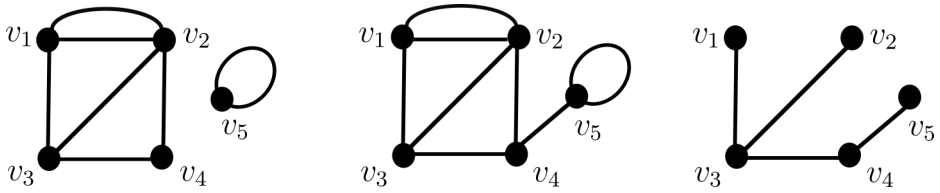
Cesta (délky ℓ) z vrcholu v_i do vrcholu v_j je libovolná posloupnost hran násobnosti alespoň jedna

$$P(v_i, v_j) = \big\{\{v_i, v_{i_1}\}, \{v_{i_1}, v_{i_2}\}, \ldots, \{v_{i_{\ell-1}}, v_j\}\big\}.$$

Poznamenejme, že námi zavedená cesta nemůže obsahovat smyčky, tj. hrany typu $\{v_{i_t}\}$. Příklad cesty je znázorněn na obrázku 2.3 vlevo. *Kružnice* je cesta (vždy délky alespoň 2) z vrcholu v_i do vcholu v_i, viz obrázek 2.3 uprostřed a vpravo. Graf nazveme *souvislý* právě tehdy, když existuje cesta mezi každými dvěma vrcholy v_i, v_j; graf, který není souvisý nazveme *nesouvislý*, viz obázek 2.4 vlevo a uprostřed. Poznamenejme, že každý souvislý podgraf (přesněji řečeno indukovaný podgraf, viz [16, str. 122]), ke kterému nelze přidat žádný další vrchol daného grafu tak, aby



Obrázek 2.3: Příklady grafů, kde modrou přerušovanou čarou je znázorněna cesta (vlevo) a kružnice (uprostřed a vpravo).



Obrázek 2.4: Příklad nesouvislého grafu (vlevo), souvislého grafu (uprostřed) a stromu (vpravo).

zůstal souvislý se nazývá (maximální souvislá) *komponenta*. Jakkoliv obecnou definici grafu (resp. multigrafu s volně visícími hranami a smyčkami) jsme zavedli a pro práci s obecnými tenzorovými sítěmi je budeme potřebovat (viz kap. 3.4, v mnoha praktických případech vystačíme s grafy mnohem jednoduššími, tzv. *stromy*. Stromem nazýváme souvislý graf, který neobsahuje kružnice, smyčky, ani násobné hrany. Příklad stromu je na obrázku 2.4 vpravo.

2.1.4 Binární strom

Dále bude užitečné definovat pojem *binární strom*. Binární strom je speciálním typem stromu, který se skládá z jednoho význačného vrcholu (zvaného kořen) a z uspořádané dvojice binárních stromů – levého a pravého podstromu, viz [16, str. 360]. Pro nás bude binární strom znamenat takový graf, pro který platí:

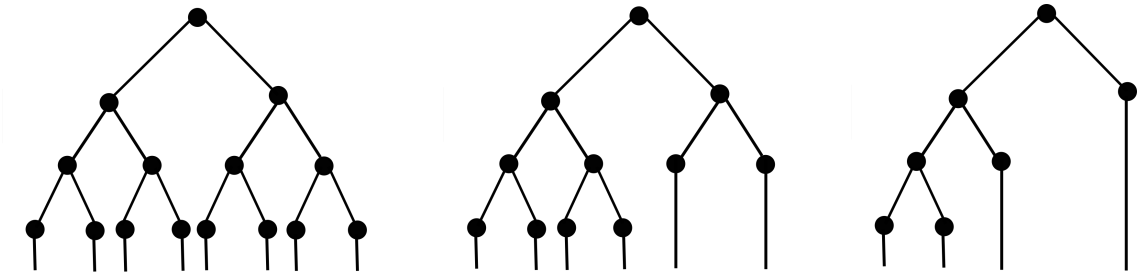
- právě jeden *vnitřní* vrchol (přesněji řečeno vrchol nemající volně visící hrany) má stupeň 2, tento vrchol je tzv. *kořen*;
- ostatní vnitřní vrcholy mají stupeň 3;
- a další vrcholy, které mají volně visící hrany, mají stupeň 2.

Příklady binárních stromů jsou uvedeny na obrázku 2.5.

2.1.5 Násobné hrany a jejich jednotlivé větve

V případě, že budeme chtít jednotlivé části, tzv. *větve*, multihrany

$$h_\ell \in \binom{V \cup \{f\}}{2} \cup \binom{V}{1}$$



Obrázek 2.5: Grafy různých binárních stromů. Tyto stromy udeme nazývat (zleva): ideálně vyvážený binární strom, částečně (ne)vyvážený binární strom, maximálně nevyvážený binární strom.

takové, že $\mu(h_\ell) = m_{h_\ell} > 1$, rozlišit, budeme předpokládat, že máme jejich *jednoznačně dané pořadí*, tj. *ohodnocení* čísly $1, 2, \ldots, m_{h_\ell}$; můžeme je značit

$$h_\ell^{(1)}, h_\ell^{(2)}, \ldots, h_\ell^{(m_{h_\ell})}.$$

Ohodnocení nebudeme formálněji zavádět, pro naši potřebu je postačující vědět, že je jednoznačné.

2.2 Faktorový graf

Protože pomocí grafů budeme později znázorňovat tenzory a speciálně také součiny tenzorů, tedy operace, při kterých např. ze dvou tenzorů vzniká tenzor nový, budeme potřebovat tyto operace něakým způsobem převést do jazyka grafů. K tomu poslouží konstrukce, kterou nazýváme *faktorový graf*.

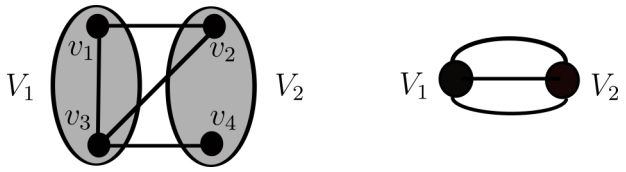
Uvažujme graf G s množinou vrcholů $V = \{v_1, \ldots, v_n\}$. Rozdělíme množinu V na k disjunktních podmnožin, tj.

$$V = V_1 \cup \cdots \cup V_k \qquad \text{a zároveň} \qquad V_i \cap V_j = \varnothing, \qquad i = 1, \ldots, k,\ j = 1 \ldots, k,\ i \neq j.$$

Budeme-li množiny $V_1, \ldots, V_k$ nyní považovat za vrcholy grafu $\widetilde{G}$, platí pro tyto vrcholy

$$\deg(V_i) = \left(\sum_{v_\ell \in V_i} \deg(v_\ell) \right) - 2 \cdot \left(\sum_{h_\ell \in \binom{V_i}{2} \cup \binom{V_i}{1}} \mu(h_\ell) \right),$$

kde v_ℓ jsou vrcholy uvnitř množiny V_i a h_ℓ jsou hrany, které incidují pouze s vrcholy uvnitř množiny V_i. Takový graf budeme nazývat faktorovým grafem. Na obrázku 2.6 je zobrazen příklad takto vzniklého grafu.



Obrázek 2.6: Příklad faktorového grafu. Původní graf se čtyřmi vrcholy $v_1, \ldots, v_4$ (vlevo), kde je naznačeno, jak vznikne faktorový graf s vrcholy V_1 a V_2 (vpravo).

3 Tenzor jako graf

V této kapitole využijeme pojmů zavedených v předchozí kapitole a vysvětlíme, jak je možné znázornit tenzory v podobě grafů. Uvidíme, že je to užitečné zejména pro znázornění součinů tenzorů nebo tenzorových rozkladů.

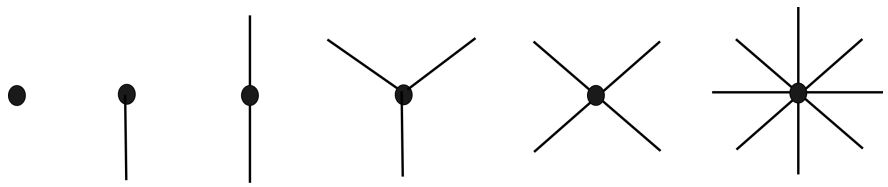
3.1 Tenzor jako graf

Mějme tenzor (1.1), tj.

$$\mathcal{T} = (t_{i_1,i_2,\ldots,i_k}) \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k}.$$

Budeme chtít tento tenzor reprezentovat jako multigraf s volně visícími hranami a smyčkami (dále jen graf), který má jediný vrchol $\mathcal{T}$ (záměrně budeme tenzory a jim odpovídající vrcholy grafu značit stejně) a k volně visících hran – přesněji řečeno jedinou volně visící multi-hranu $h = \{\mathcal{T}, f\}$ s násobností $\mu(h) = k$.

Jednotlivé větve multi-hrany $h^{(1)}, h^{(2)}, \ldots, h^{(k)}$ odpovídají indexům $i_1, i_2, \ldots, i_k$ tenzoru $\mathcal{T}$. V dalším textu o nich budeme mluvit jako o *fyzických* indexech, resp. hranách (resp. větvích multi-hrany), viz obrázek 3.1.

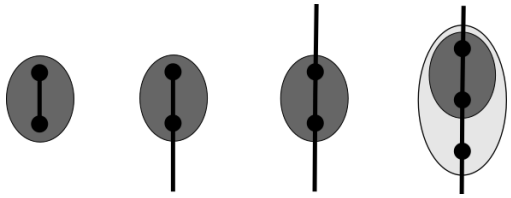


Obrázek 3.1: Grafy odpovídající tenzorům různých řádů (zleva): skalár (tenzor nultého řádu), vektor (tenzor prvního řádu), matice (tenzor druhého řádu), tenzor třetího, čtvrtého a osmého řádu.

3.2 Tenzorový součin

Nyní se zaměříme na znázornění různých tenzorových interakcí v podobě grafu. Klasické hrany spojující dva vrcholy, tj. dva tenzory, budou představovat součin těchto tenzorů v příslušných módech, viz [19, str. 29].

Zaměříme se nejprve na součiny, které dobře známe z lineární algebry. Příklady operací s vektory a maticemi, tj. tenzory prvního a druhého řádu, jsou zobrazeny

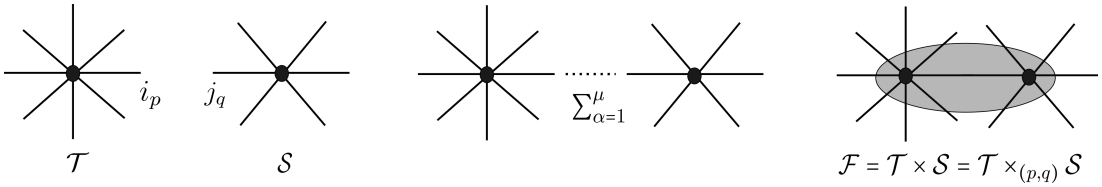


Obrázek 3.2: Znázornění součinů vektorů a matic (zleva): skalární součin dvou vektorů, vektor ve tvaru součinu matice s vektorem, matice ve tvaru součinu dvou matic, matice ve tvaru součinu tří matic. Jednotlivé oválné slupky jsou de-facto jednotlivé faktorové grafy.

na obrázku 3.2. Máme-li dva tenzory $\mathcal{T}$ a $\mathcal{S}$, potom, abychom mohli provést jejich součin (1.3), tj.

$$\mathcal{F} = \mathcal{T} \times_{(p,q)} \mathcal{S} \equiv \left(\sum_{\alpha=1}^{\mu} t_{i_1,\ldots,i_{p-1},\alpha,i_{p+1}\ldots,i_k} \cdot s_{j_1,\ldots,j_{q-1},\alpha,j_{q+1}\ldots,j_\ell} \right) \in \mathbb{R}^{n_1\times\cdots\times n_{p-1}\times n_{p+1}\times\cdots\times n_k\times m_1\times\cdots\times m_{q-1}\times m_{q+1}\times\cdots\times m_\ell},$$

musí existovat indexy i_p v tenzoru $\mathcal{T}$ a j_q v tenzoru $\mathcal{S}$ nabývající stejného rozsahu hodnot $1,\ldots,\mu$. Znázornění součinu tenzorů v podobě grafu je ilustrováno na obrázku 3.3.



Obrázek 3.3: Princip zápisu tenzorového součinu (úžení) dvou tenzorů řádů osm a šest. Volně visící hrany dvou tenzorů, které odpovídají módům i_p a i_q stejných rozměrů a ve kterých probíhá násobení, jsou nahrazeny hranou spojující oba tenzory (v terminologii grafů jde o tzv. kontrakci hrany). Šedý ovál představuje výsledný součin – tenzor řádu $8 + 6 - 2 = 12$.

Uvědomme si, že součin tenzoru a matice v ℓ-tém módu a součin dvou tenzorů v daných módech, viz kap. 1, jsou definovány téměř stejně až na permutaci indexů (analogii transpozice matice). Srovnej např.

$$\mathcal{T} \times_\ell M \in \mathbb{R}^{n_1\times\cdots\times n_{\ell-1}\times m\times n_{\ell+1}\times\cdots\times n_k} \qquad \text{a} \qquad \mathcal{T} \times_{(\ell,2)} M \in \mathbb{R}^{n_1\times\cdots\times n_{\ell-1}\times n_{\ell+1}\times\cdots\times n_k\times m},$$

viz (1.2), (1.3) a také [26, kap. 2.2 a 2.6]. Při zápisu v podobě grafu toto odpovídá pouze přečíslování větví volně visící multihrany součinu.

3.3 Další objekty lineární algebry interpretovatelné jako tenzorové součiny

Poznamenejme, že kromě standardních maticových součinů lze tímto způsobem vyjádřit i řadu dalších objektů běžně užívaných v lineární algebře, které ovšem většinou jako součiny nevykládáme.

3.3.1 Stopa matice

Stopa čtvercové matice $A \in \mathbb{R}^{n\times n}$ je v lineární algebře definována jako součet diagonálních prvků, tj.

$$\operatorname{trace}(A) = \sum_{i=1}^{n} a_{i,i}.$$

Využijeme-li graf, lze stopu matice interpretovat jako součin čtvercové matice sama se sebou, viz obrázek 3.4.

3.3.2 Skalární součin na prostoru matic

Podobně i skalární součin dvou matic $A, B \in \mathbb{R}^{n\times m}$ definovaný jako

$$\langle A, B\rangle = (\operatorname{vec}(A))^T \cdot (\operatorname{vec}(B)) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j} \cdot b_{i,j}$$

lze znázornit pomocí grafu, viz opět obrázek 3.4.

Podobným způsobem lze zavést také např. následující „nestandardní součin“ tří (či více) matic, jejichž výsledkem je skalár. Pro

$$A \in \mathbb{R}^{n\times m},\ B \in \mathbb{R}^{m\times o},\ C \in \mathbb{R}^{o\times n} \qquad \textit{definujme} \text{ součin} \qquad \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{o} a_{i,j} \cdot b_{j,k} \cdot c_{k,i}.$$

Tento součin je ilustrován na obrázku 3.4 jako třetí zleva.

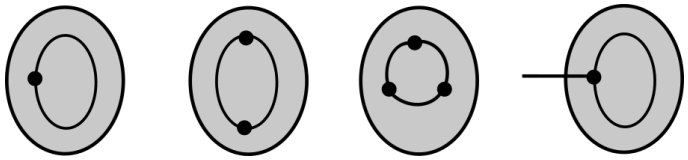
3.3.3 Méně obvyklé objekty

Posledním příkladem, který uvádíme na obrázku 3.4, je vektor vzniklý z tenzoru třetího řádu $\mathcal{A} \in \mathbb{R}^{n\times m\times m}$. Jednotlivé prvky tenzoru $\mathcal{A}$ označíme $a_{i,j,k}$ a řezy v prvním módu $a_{1,:,:}, a_{2,:,:}, \ldots, a_{n,:,:} \in \mathbb{R}^{1\times m\times m}$, tzv. horizontántální řezy, viz [11] a [26, kap. 2.1.2, obr. 2.2]. Potom *definujme* vektor v po složkách tak, že

$$v_i = \sum_{j=1}^{m} a_{i,j,j}.$$

Tedy i-tá složka vektoru v je stopou matice, která je triviálně izomorfní s i-tým řezem tenzoru $\mathcal{A}$ v prvním módu (horizontálním řezem) $a_{i,:,:}$.

Je zřejmé, že pokud rozumíme grafům, lze názorným způsobem zapsat nejrůznější objekty. Možnosti však nejsou neomezené, pokud bychom např. chtěli vyjádřit „trojrozměrnou stopu kubického tenzoru“ třetího řádu $\mathcal{A} \in \mathbb{R}^{n\times n\times n}$, tj. součet prvků na tělesové úhlopříčce $\sum_{i=1}^{n} a_{i,i,i}$, potřebovali bychom k tomu „hranu se třemi konci“.



Obrázek 3.4: Méně obvyklé typy součinů (zleva): stopa matice (viz kap. 3.3.1; skalární součin dvou matic viz kap. 3.3.2; zvláštní součin tří matic a vektor, jehož složky jsou stopy matic – řezů tenzoru třetího řádu (viz kap 3.3.3).

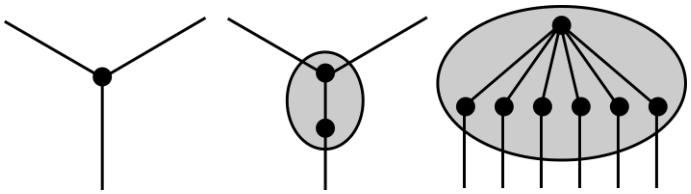
3.4 Obecné tenzorové sítě

V předchozím textu jsme vysvětlili, jak lze interpretvat graf jako tenzor. Nyní budeme chtít postupovat opačně. Budeme mít daný tenzor a danou strukturu sítě (graf), případně i některé další vlastnosti, a naším úkolem bude najít faktory tohoto tenzoru, tj. vrcholy grafu (sítě). Grafu, který představuje nějaký tenzor jako výsledek součinů jiných tenzorů, říkáme *tenzorová síť*.

Tenzorovou síť lze proto použít k zápisu různých tenzorových rozkladů, které mají právě podobu součinů. V dalších kapitolách tohoto textu se s některými z nich seznámíme podrobněji.

Pro znázornění tenzorové sítě se používá i zvláštní terminologie k rozlišení hran různých typů. Hrany klasického typu, tj. typu $\{v_i, v_j\}$, v tenzorové síti nazýváme *sčítací indexy*, případně vnitřní nebo virtuální indexy; volně visící hrany, tj. hrany typu $\{v_i, f\}$, se nazývají *fyzické* (příp. vnější) indexy a jejich počet udává řád celého tenzoru.

Na obrázku 3.5 můžeme porovnat znázornění jednoduchého tenzoru třetího řádu a tenzorové sítě – tenzoru třetího řádu ve tvaru součinu tenzoru třetího řádu s maticí. Ve stejném obrázku dále ilustrujeme Tuckerův rozklad tenzoru šestého řádu. Připoměňme, že abychom mohli takovouto síť nazvat Tuckerovým rozkladem, předpokládáme kromě dané struktury také vlastnost, že matice v této síti mají *ortonormální sloupce*, viz kap. 1.1.



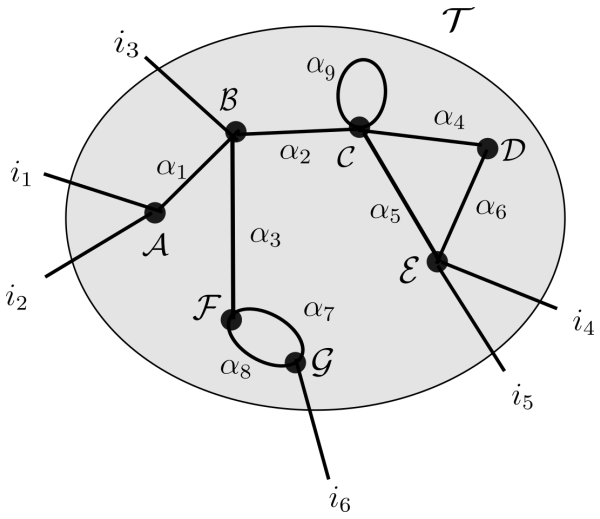
Obrázek 3.5: Znázornění tenzorů vyšších řádů ve formě grafu (zleva): tenzor třetího řádu, tenzor třetího řádu ve tvaru součinu tenzoru třetího řádu s maticí, tenzor šestého řádu ve tvaru Tuckerova rozkladu.

V principu, máme-li daný tenzor a předepsaný graf (tenzorovou síť), můžeme se pokusit vyjádřit tento tenzor v podobě rozkladu, který v grafickém znázornění má právě podobu předepsaného grafu. Pro přesnější představu poslouží příklad 1. U skutečných úloh samozřejmě síť nepředepisujeme zcela svévolně. Zpravidla se

snažíme tenzor poskládat z objektů, které mají nějaký, např. fyzikální, význam, viz [13] nebo [18], který navíc umožňuje předepsat strukturu (např. symetrii, nebo hodnosti) těchto objektů (např. symetrická matice, nebo symetrické řezy v daném módu, toeplitzovská matice, matice hodnosti nejvýše r, atp.).

Příklad 1. *Pro daný tenzor $\mathcal{T}$ chceme najít tenzory $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$, $\mathcal{E}$, $\mathcal{F}$, $\mathcal{G}$ tak, aby tvořily síť tenzoru $\mathcal{T}$, takovou jako je na obrázku 3.6. Tato síť odpovídá součinu definovanému vztahem:*

$$\begin{aligned}\mathcal{T} &\approx (t_{i_1,i_2,i_3,i_4,i_5,i_6,i_7,i_8}) \\ &= \left(\sum_{\alpha_1,\ldots,\alpha_9} a_{i_1,i_2,\alpha_1} \cdot b_{i_3,\alpha_1,\alpha_2\alpha_3} \cdot c_{\alpha_2,\alpha_9,\alpha_9,\alpha_4,\alpha_5} \cdot d_{\alpha_4,\alpha_6} \cdot e_{\alpha_5,\alpha_6,i_4,i_5} \cdot f_{\alpha_3,\alpha_7,\alpha_8} \cdot g_{\alpha_7,\alpha_8,i_6} \right).\end{aligned} \tag{3.1}$$



Obrázek 3.6: Rozklad tenzoru $\mathcal{T}$ šestého řádu do (resp. aproximace pomocí) tenzorové sítě tvořené tenzory $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$, $\mathcal{E}$, $\mathcal{F}$ a $\mathcal{G}$ nižších řádů; viz příklad 1.

3.5 Speciální tenzorové sítě

Obecně je motivací pro konstrukci tenzorových sítí zejména umožnit práci s rozsáhlými vícerozměrnými daty. Například tenzor $\mathcal{T} \in \mathbb{R}^{2\times 2\times\cdots\times 2}$ řádu 100 obsahuje $2^{100} \approx 1.2676506 \times 10^{30}$ prvků, což zřejmě nelze uložit do paměti počítače.[1] Navíc zde ani případná komprese, např. pomocí klasického Tuckerova rozkladu (viz [26, kap. 4], [19, str. 20], a [3, str. 1267]), nepomůže. Další motivací může být snaha pomocí tenzorových sítí zpřehlednit mnohorozměrná data, viz zejména [18]. Naším cílem tedy bude najít takové sítě, které umožní

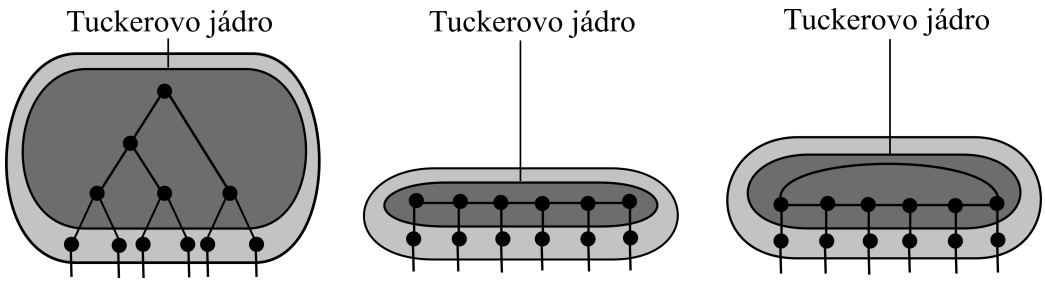
- snížit paměťové nároky (tj. budeme chtít najít rozklad do sítě s co nejméně tenzory co nejnižších řádů);

[1] Uvažujeme-li, že na uložení jednoho čísla např. v přesnosti *double* potřebujeme 8 bytů, pak $2^{100} = 2^{60} \cdot 2^{40}$ a na uložení $2^{40} = 1024^4$ čísel potřebujeme 8 terabytů.

- snadno dále manipulovat s tenzory ve tvaru sítě (tj. hledáme topologicky jednoduché sítě).

Tyto požadavky vedou k tenzorové síti ve tvaru binárního stromu. Nejčastěji používané jsou: *hierarchický Tuckerův rozklad* (HTD; z anglického *hierarchical Tucker decomposition*), viz např. [19, kap. 3] nebo [7], *tensor train* (TT), viz [17], a další, viz také obrázek 3.7. V ideálním případě hierarchický Tuckerův rozklad dostáváme ve tvaru *vyváženého binárního stromu*. Obecně se snažíme dostat strom, který není příliš nevyvážený. Případná nevyváženost může být způsobena:

- řádem tenzoru (je-li různý od $k = 2^s$);
- praktickými důvody (významem komponent – tj. když fyzické indexy odpovídají určitému jevu, např. tepelné vodivosti jako v [13, kap. 4.1]).



Obrázek 3.7: Příklady tenzorových sítí odpovídající tenzoru šestého řádu (zleva): hierarchický Tuckerův rozklad tenzoru (HTD) – síť ve tvaru ne zcela vyváženého binárního stromu, tensor train (TT) – maximálně nevyvážený binární strom a tzv. *tensor chain* (TC), viz [9, str. 5]. Ten z předchozího rozkladu vzniká přidáním jediné hrany; na rozdíl od obou předchozích obsahuje kružnici a je tedy výpočetně náročnější na konstrukci. Pokud se budeme na vnitřní tmavší blok dívat jako na jediný tenzor, všechny tři obrázky mohou představovat obyčejný Tuckerův rozklad.

4 Hierarchický Tuckerův rozklad (HTD)

Jedním z rozkladů, jehož struktura je znázorňována ve tvaru tenzorové sítě, je hierarchický Tuckerův rozklad. Klasický Tuckerův rozklad, který jsme již připomněli v kapitole 1, umožňuje vyjádření tenzoru řádu k ve tvaru součinu tenzoru (jehož rozměry jsou omezené vektorovým rankem původního tenzoru) řádu k, tzv. jádra tenzoru, s k maticemi. Hierarchický Tuckerův rozklad (HTD z anglického hierarchical Tucker decomposition) spočívá navíc v rozložení Tuckerova jádra daného tenzoru do tvaru součinu jednodušších tenzorů. Takový rozklad můžeme reprezentovat tenzorovou sítí (předem dané struktury). V této kapitole vysvětlíme základní princip vytvoření hierarchického Tuckerova rozkladu a tím zároveň ověříme jeho existenci.

4.1 Struktura HTD

V této části si ukážeme, jakým způsobem lze tenzor transformovat do potřebné struktury dané tenzorvou sítí, která bude mít podobu binárního stromu, jako např. na obrázku 3.7 vlevo. Uvažujme tedy tenzor (1.1) řádu k

$$\mathcal{T} = (t_{i_1,i_2,\dots,i_k}) \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k}. \tag{4.1}$$

Nultým krokem může být klasický Tuckerův rozklad, přičemž HTD se provede pro Tuckerovo jádro. My zde HTD vyložíme pro obecný tenzor (ne nezbytně Tuckerovo jádro).

4.1.1 Nalezení tenzoru druhého řádu – kořene binárního stromu

Nyní budeme hledat rozklad tenzoru do sítě předepsaného tvaru. Konkrétně budeme chtít dosáhnout *co nejvíce vyváženého* stromu. Z kapitoly 2.1.4 víme, že kromě vrcholů s volně visícími hranami obsahuje binární strom vrcholy stupňů 3 a jeden vrchol (kořen) stupně 2. Důležitými nástroji pro nalezení takového binárního stromu dále budou:

- rozvoj tenzoru v matici (viz definici 3),
- singulární rozklad matice (viz např. [4, kap. 5]).

Tenzorovou síť konkrétního tvaru získáme tak, že každá hrana tenzorové sítě bude představovat rozvoj tenzoru v módech odpovídajích rozdělení množiny indexů tenzoru. V našem případě, kdy chceme získat vyvážený binární strom pro jádro tenzoru

(4.1), rozdělíme množinu jeho indexů

$$\{1, 2, \ldots, k\}$$

do dvou disjunktních podmnožin

$$\{1, \ldots, s\} \quad \text{a} \quad \{s+1, \ldots, k\},$$

pro nějaké s, $1 \leq s < k$. Potom lze vytvořit rovoj tenzoru v matici podle první podmnožiny, tj.

$$\mathcal{T}^{\{1,\ldots,s\}} \in \mathbb{R}^{N_{\mathfrak{L}} \times N_{\mathfrak{R}}}, \qquad \text{kde} \qquad N_{\mathfrak{L}} = n_1 \cdot n_2 \cdot \cdots \cdot n_s, \quad N_{\mathfrak{R}} = n_{s+1} \cdot n_{s+2} \cdot \cdots \cdot n_k.$$

Uvažujme *ekonomický* tvar singulárního rozkladu této matice

$$\mathcal{T}^{\{1,\ldots,s\}} = U_{(1-s)} \Sigma_{(1-s)} \left(V_{(1-s)}\right)^T, \tag{4.2}$$

kde

$$U_{(1-s)} \in \mathbb{R}^{N_{\mathfrak{L}} \times r_{(1-s)}}, \quad \Sigma_{(1-s)} \in \mathbb{R}^{r_{(1-s)} \times r_{(1-s)}}, \quad V_{(1-s)} \in \mathbb{R}^{N_{\mathfrak{R}} \times r_{(1-s)}}, \tag{4.3}$$

přičemž

$$r_{(1-s)} = \operatorname{rank}\left(\mathcal{T}^{\{1,\ldots,s\}}\right). \tag{4.4}$$

Tímto jsme dosáhli prvního kroku při hledání HTD, jelikož máme tenzor řádu 2, matici $\Sigma_{(1-s)}$, která je kořenem hledaného binárního stromu, viz obrázek 3.7 vlevo.

4.1.2 Větvení binárního stromu pomocí tenzorů třetího řádu

Z obrázku je dále patrné, že síť HTD obsahuje také velké množství tenzorů třetího řádu, které budeme hledat ve směru k listům, tak jak naznačuje obrázek 4.1.

Pro jejich nalezení využijeme důležitou vlastnost, kterou později zobecníme ve větě 1. Uvažujme m takové, že $1 \leq m < s$. Uvažujme dále matice $U_{(1-m)}$ a $U_{((m+1)-s)}$ levých singulárních vektorů získané z *ekonomických* singulárních rozkladů rozvojů $\mathcal{T}^{\{1,\ldots,m\}}$ a $\mathcal{T}^{\{m+1,\ldots,s\}}$ podobně jako ve (4.2)–(4.4). Pro obor hodnot matice $U_{(1-s)}$ platí

$$\mathcal{R}(U_{(1-s)}) \subseteq \mathcal{R}\left(U_{((m+1)-s)} \otimes U_{(1-m)}\right), \tag{4.5}$$

neboli každý sloupec matice $U_{(1-s)}$ lze zapsat jako lineární kombinaci sloupců Kroneckerova součinu matic $U_{(1-m)}$ a $U_{((m+1)-s)}$. Tedy existuje matice B taková, že

$$U_{(1-s)} = \left(U_{((m+1)-s)} \otimes U_{(1-m)}\right) \cdot B_{(1-s)}, \qquad B_{(1-s)} \in \mathbb{R}^{(r_{(1-m)} \cdot r_{((m+1)-s)}) \times r_{(1-s)}}; \tag{4.6}$$

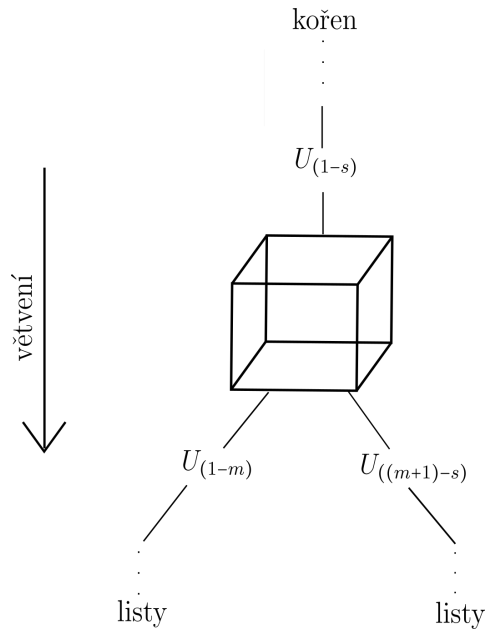
sloupce matice $B_{(1-s)}$ obsahují koeficienty výše zmíněných lineárních kombinací. Tuto matici je možné chápat jako rovoj tenzoru třetího řádu $\mathcal{B}$ tak, že

$$B_{(1-s)} = \mathcal{B}^{\{1,2\}}_{(1-s)}, \qquad \text{kde} \qquad \mathcal{B}_{(1-s)} \in \mathbb{R}^{r_{(1-m)} \times r_{((m+1)-s)} \times r_{(1-s)}}. \tag{4.7}$$

Analogickým postupem budeme dál pracovat s maticemi $U_{(1-m)}$ a $U_{((m+1)-s)}$, čímž se postupně rozvětvuje binární strom a získáváme tak další faktory – tenzory třetího řádu – tenzorové sítě, viz obrázek 4.2. S maticí $V_{(1-s)}$ naložíme podobně, stačí si uvědomit, že platí

$$V_{(1-s)} \equiv U_{((s+1)-k)}, \qquad \text{neboť} \qquad \left(\mathcal{T}^{\{1,\ldots,s\}}\right)^T = \mathcal{T}^{\{s+1,\ldots,k\}}. \tag{4.8}$$

Uvědomme si, že podobu tenzorové sítě (větvení binárního stromu) určuje vždy rozdělení množiny indexů a tomu odpovídající rozvoje v jednotlivých krocích.



Obrázek 4.1: Znázornění větvení binárního stromu – tenzorové sítě při hierarchickém Tuckerově rozkladu.

4.1.3 Listy stromu – tenzory druhého řádu

Způsobem popsaným v předchozí části se postupně dostaneme až k množině matic

$$U_{(1)}, U_{(2)}, \ldots, U_{(k)}, \qquad U_{(\ell)} \in \mathbb{R}^{n_\ell \times r_{(\ell)}}, \qquad \ell = 1, \ldots, k.$$

Připomeňme, že pro hodnoty r_ℓ z klasického Tuckerova rozkladu a hodnosti $r_{(\ell)}$, které získáváme při hierarchickém Tuckerově rozkladu platí

$$r_\ell = r_{(\ell)} = \operatorname{rank}(\mathcal{T}^{\{\ell\}}),$$

jelikož hodnosti Tuckerova jádra jsou dány jednoznačně. Dále také zřejmě platí

$$\mathcal{R}(U'_\ell) = \mathcal{R}(U_{(\ell)}).$$

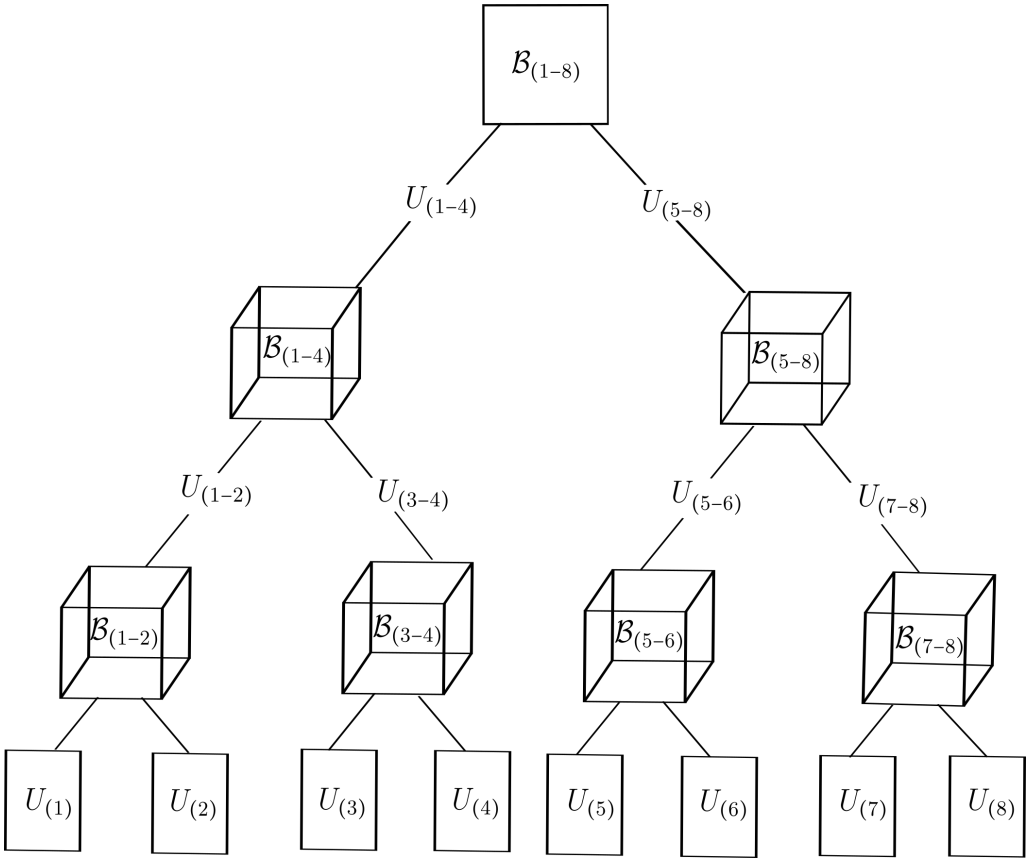
4.1.4 Příklad rozkladu tenzoru osmého řádu

Příklad 2. *V tomto příkladu postupně použijeme vztah (4.6) k ilustraci postupu při HTD tenzoru $\mathcal{T}$ řádu $k = 8 = 2^3$, viz také obrázek 4.2. Zřejmě platí*

$$\operatorname{vec}(\mathcal{T}) = \mathcal{T}^{\{1,\ldots,8\}} \quad = \quad (U_{(5-8)} \otimes U_{(1-4)}) \cdot B_{(1-8)},$$

kde $B_{(1-8)} = \operatorname{vec}(\Sigma_{(1-4)})$ a $U_{(5-8)} = V_{(1-4)}$, viz (4.2) a (4.8), a kde

$$\begin{aligned} U_{(1-4)} &= (U_{(3-4)} \otimes U_{(1-2)}) \cdot B_{(1-4)}, \\ U_{(5-8)} &= (U_{(7-8)} \otimes U_{(5-6)}) \cdot B_{(5-8)}, \end{aligned}$$



Obrázek 4.2: Ilustrace hierarchického Tuckerova rozkladu tenzoru osmého řádu, kde jsme formálně označili $\mathcal{B}_{(1-8)} = \Sigma_{(1-4)} \in \mathbb{R}^{r_{(1-4)} \times r_{(1-4)}}$, viz (4.4).

kde dále

$$\begin{aligned} U_{(1-2)} &= (U_{(2)} \otimes U_{(1)}) \cdot B_{(1-2)}, \\ U_{(3-4)} &= (U_{(4)} \otimes U_{(3)}) \cdot B_{(3-4)}, \\ U_{(5-6)} &= (U_{(6)} \otimes U_{(5)}) \cdot B_{(5-6)}, \\ U_{(7-8)} &= (U_{(8)} \otimes U_{(7)}) \cdot B_{(7-8)}. \end{aligned}$$

Potom, s využitím asociativity Kroneckerova součinu, po dosazení platí

$$\begin{aligned} \operatorname{vec}(\mathcal{T}) = (U_{(8)} \otimes U_{(7)} \otimes U_{(6)} \otimes U_{(5)} \otimes U_{(4)} \otimes U_{(3)} \otimes U_{(2)} \otimes U_{(1)}) \\ \cdot (B_{(7-8)} \otimes B_{(5-6)} \otimes B_{(3-4)} \otimes B_{(1-2)}) \cdot (B_{(5-8)} \otimes B_{(1-4)}) \cdot B_{(1-8)}. \end{aligned} \tag{4.9}$$

Matice B, nazývané také matice přenosu, lze (s výjimkou kořene stromu) převést do tvaru tenzorů třetích řádů, tj.

$$\begin{aligned} B_{(1-4)} \in \mathbb{R}^{(r_{(1-2)} \cdot r_{(3-4)}) \times r_{(1-4)}} &\iff \mathcal{B}_{(1-4)} \in \mathbb{R}^{r_{(1-2)} \times r_{(3-4)} \times r_{(1-4)}}, \\ B_{(5-8)} \in \mathbb{R}^{(r_{(5-6)} \cdot r_{(7-8)}) \times r_{(5-8)}} &\iff \mathcal{B}_{(5-8)} \in \mathbb{R}^{r_{(5-6)} \times r_{(7-8)} \times r_{(5-8)}}, \end{aligned}$$

a

$$
\begin{aligned}
B_{(1-2)} \in \mathbb{R}^{(r_{(1)}\cdot r_{(2)})\times r_{(1-2)}} &\iff \mathcal{B}_{(1-2)} \in \mathbb{R}^{r_{(1)}\times r_{(2)}\times r_{(1-2)}},\\
B_{(3-4)} \in \mathbb{R}^{(r_{(3)}\cdot r_{(4)})\times r_{(3-4)}} &\iff \mathcal{B}_{(3-4)} \in \mathbb{R}^{r_{(3)}\times r_{(4)}\times r_{(3-4)}},\\
B_{(5-6)} \in \mathbb{R}^{(r_{(5)}\cdot r_{(6)})\times r_{(5-6)}} &\iff \mathcal{B}_{(5-6)} \in \mathbb{R}^{r_{(5)}\times r_{(6)}\times r_{(5-6)}},\\
B_{(7-8)} \in \mathbb{R}^{(r_{(7)}\cdot r_{(8)})\times r_{(7-8)}} &\iff \mathcal{B}_{(7-8)} \in \mathbb{R}^{r_{(7)}\times r_{(8)}\times r_{(7-8)}}.
\end{aligned}
$$

Vidíme, že dostáváme tenzor $\mathcal{T}$ v podobě zcela vyváženého binárního stromu, viz obrázek 4.2.

4.2 Základní věta HTD

Popsali jsme strukturu celého hierarchického Tuckerova rozkladu. Jediné, co zbývá dokázat je, že platí vlastnost (4.5), kterou v nepatrně obecnější podobě formuluje věta 1, viz také [19, kap. 3.1.3]. Předtím ale uvedeme následující lemma, které bude užitečné pro pochopení důkazu.

Lemma 1. *Nechť M je libovolná matice nad $\mathbb{R}$, pak $MM^\dagger$ (kde $M^\dagger$ značí Moorovu–Penroseovu pseudoinverzi) je ortogonální projektor (viz [4, str. 15]) na $\mathcal{R}(M)$.*

Důkaz. Nechť $r = \operatorname{rank}(M)$. Uvažujme ekonomický singulární rozklad matice $M = U_r\Sigma_r V_r^T$, tj. $\mathcal{R}(M) = \mathcal{R}(U_r)$ a Σ_r je regulární. Pak $M^\dagger = V_r\Sigma_r^{-1}U_r^T$. Zřejmě platí

$$
MM^\dagger = U_r\Sigma_r \underbrace{V_r^TV_r}_{I}\Sigma_r^{-1}U_r^T = U_r\underbrace{\Sigma_r\Sigma_r^{-1}}_{I}U_r^T = U_rU_r^T = \sum_{j=1}^{r} u_ju_j^T,
$$

kde $U_r = [u_1, \ldots, u_r]$. □

Pro libovolný vektor x tedy platí $MM^\dagger x \in \mathcal{R}(M)$, speciálně pro $x \in \mathcal{R}(M)$ platí $x = MM^\dagger x$ a tedy $(MM^\dagger)^2 = MM^\dagger$.

4.2.1 Důkaz základní věty hierarchického Tuckerova rozkladu

Nyní se dostáváme k avizované větě.

Věta 1. *Nechť $\mathcal{T} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_k}$ je tenzor řádu k a $\mathscr{C}_\mathfrak{L}, \mathscr{C}_\mathfrak{R} \subset \{1, \ldots, k\}$ jsou libovolné podmnožiny množiny jeho indexů takové, že platí $\mathscr{C}_\mathfrak{L} \cap \mathscr{C}_\mathfrak{R} = \varnothing$. Označme $\mathscr{C} = \mathscr{C}_\mathfrak{L} \cup \mathscr{C}_\mathfrak{R}$. Potom*

$$
\mathcal{R}(\mathcal{T}^{\mathscr{C}}) \subseteq \mathcal{R}\left(\mathcal{T}^{\mathscr{C}_\mathfrak{R}} \otimes \mathcal{T}^{\mathscr{C}_\mathfrak{L}}\right), \tag{4.10}
$$

kde $\mathcal{T}^{\mathscr{C}}, \mathcal{T}^{\mathscr{C}_\mathfrak{L}}, \mathcal{T}^{\mathscr{C}_\mathfrak{R}}$ jsou rozvoje tenzoru $\mathcal{T}$ do matice podle multiindexů $\mathscr{C}, \mathscr{C}_\mathfrak{L}, \mathscr{C}_\mathfrak{R}$.

Důkaz. Nechť pro jednoduchost $\mathscr{C}_\mathfrak{L}$, $\mathscr{C}_\mathfrak{R}$ obsahují po sobě jdoucí indexy, tj.

$$\mathscr{C}_\mathfrak{L} = \{i_{\ell+1}, i_{\ell+2}, \ldots, i_m\}, \quad \mathscr{C}_\mathfrak{R} = \{i_{m+1}, i_{m+2}, \ldots, i_r\},$$

tj.

$$|\mathscr{C}_\mathfrak{L}| = m - \ell, \quad |\mathscr{C}_\mathfrak{R}| = r - m, \quad |\mathscr{C}| = r - \ell.$$

Nechť $\mathcal{T}^{\mathscr{C}}$ je tozvoj tenzoru $\mathcal{T}$ podle multiindexu $\mathscr{C}$. Jakýkoli sloupec matice $\mathcal{T}^{\mathscr{C}}$ lze považovat za vektorizaci $\mathrm{vec}(\mathcal{C})$ nějakého tenzoru $\mathcal{C} \in \mathbb{R}^{n_{i_{\ell+1}} \times \cdots \times n_{i_r}}$ řádu $r - \ell$. Tento tenzor rozvineme do matice tak, abychom (původní) multiindex $\mathscr{C}$ rozdělili na $\mathscr{C}_\mathfrak{L}$ a $\mathscr{C}_\mathfrak{R}$. Dostaneme tak matici $\mathscr{C}^{\{1,\ldots,m-\ell\}} \in \mathbb{R}^{(m-\ell)\times(r-m)}$. Sloupce matice $\mathcal{C}^{\mathscr{C}_\mathfrak{L}}$ přitom musí být také sloupci matice $\mathcal{T}^{\mathscr{C}_\mathfrak{L}}$. Tedy jsou zřejmě obsaženy v $\mathcal{R}(\mathcal{T}^{\mathscr{C}_\mathfrak{L}})$. (Pro snadnější porozumění a ověření pro jednoduchý tenzor odkazujeme na příklad 3 uvedený pod tímto důkazem.) Platí tedy

$$\mathcal{C}^{\mathscr{C}_\mathfrak{L}} = \mathcal{T}^{\mathscr{C}_\mathfrak{L}} (\mathcal{T}^{\mathscr{C}_\mathfrak{L}})^\dagger \mathcal{C}^{\mathscr{C}_\mathfrak{L}},$$

viz lemma 1. Analogicky platí

$$(\mathcal{C}^{\mathscr{C}_\mathfrak{L}})^T \equiv \mathcal{C}^{\mathscr{C}_\mathfrak{R}} = \mathcal{T}^{\mathscr{C}_\mathfrak{R}} (\mathcal{T}^{\mathscr{C}_\mathfrak{R}})^\dagger \mathcal{C}^{\mathscr{C}_\mathfrak{R}}.$$

Transpozicí druhého vztahu a následným dosazením dostaneme

$$\begin{aligned} \mathcal{C}^{\mathscr{C}_\mathfrak{L}} = (\mathcal{C}^{\mathscr{C}_\mathfrak{R}})^T &= (\mathcal{C}^{\mathscr{C}_\mathfrak{R}})^T \big((\mathcal{T}^{\mathscr{C}_\mathfrak{R}})^\dagger\big)^T (\mathcal{T}^{\mathscr{C}_\mathfrak{R}})^T = \mathcal{C}^{\mathscr{C}_\mathfrak{L}} \big((\mathcal{T}^{\mathscr{C}_\mathfrak{R}})^\dagger\big)^T (\mathcal{T}^{\mathscr{C}_\mathfrak{R}})^T \quad (4.11) \\ &= \mathcal{T}^{\mathscr{C}_\ell} \underbrace{\Big((\mathcal{T}^{\mathscr{C}_\mathfrak{L}})^\dagger \mathcal{C}^{\mathscr{C}_\mathfrak{L}} \big((\mathcal{T}^{\mathscr{C}_\mathfrak{R}})^\dagger\big)^T \Big)}_{W} (\mathcal{T}^{\mathscr{C}_\mathfrak{R}})^T \quad (4.12) \end{aligned}$$

a tedy

$$\mathrm{vec}(\mathcal{C}) = \mathrm{vec}(\mathcal{C}^{\mathscr{C}_\mathfrak{L}}) = (\mathcal{T}^{\mathscr{C}_\mathfrak{R}} \otimes \mathcal{T}^{\mathscr{C}_\mathfrak{L}}) \cdot \mathrm{vec}(W),$$

viz [26, str. 39, Poznámka 3], [5] nebo [23]. Tedy libovolný sloupec $\mathrm{vec}(C)$ matice $\mathcal{T}^{\mathscr{C}}$ lze napsat jako lineární kombinaci sloupců matice $\mathcal{T}^{\mathscr{C}_\mathfrak{R}} \otimes \mathcal{T}^{\mathscr{C}_\mathfrak{L}}$. A tedy obor hodnot matice $\mathcal{T}^{\mathscr{C}}$ je podmnožinou oboru hodnot matice $\mathcal{T}^{\mathscr{C}_\mathfrak{R}} \otimes \mathcal{T}^{\mathscr{C}_\mathfrak{L}}$.

□

Příklad 3. *Zde ukážeme vlastnosti sloupců matic rozvojů tenzorů $\mathcal{T}$ a $\mathcal{C}$, které využíváme v důkazu věty 1, na jednoduchém příkladu. Mějme tenzor*

$$\mathcal{T} = \begin{array}{|cc|cc|} \hline 1 & 5 & 2 & 6 \\ 3 & 7 & 4 & 8 \\ \hline \end{array} \in \mathbb{R}^{2\times2\times2}. \quad (4.13)$$

Vybereme množinu jeho indexů $\mathscr{C}_\mathfrak{L} = \{1\}$ a $\mathscr{C}_\mathfrak{R} = \{2\}$. Potom rozvoj rozvoje tenzoru $\mathcal{T}$ jsou

$$\mathcal{T}^{\{1,2\}} = \begin{bmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \\ 4 & 8 \end{bmatrix}, \qquad \mathcal{T}^{\{1\}} = \begin{bmatrix} 1 & 5 & 2 & 6 \\ 3 & 7 & 4 & 8 \end{bmatrix}, \qquad \mathcal{T}^{\{2\}} = \begin{bmatrix} 1 & 5 & 3 & 7 \\ 2 & 6 & 4 & 8 \end{bmatrix}.$$

Uvažujme první sloupec matice $\mathcal{T}^{\{1,2\}}$. Ten lze považovat za vektorizaci $\mathrm{vec}(\mathcal{C})$ *tenzoru druhého řádu*

$$\mathcal{C} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \in \mathbb{R}^{2\times 2},$$

jeho rozvoj do matice $\mathcal{C}^{\mathscr{C}_\mathfrak{L}} \equiv \mathcal{C}^{\{1\}}$ je roven přímo $\mathcal{C}$. V důkazu využíváme toho, že libovolný sloupec matice $\mathcal{C}^{\mathscr{C}_\mathfrak{L}}$ je také sloupcem matice $\mathcal{T}^{\mathscr{C}_\mathfrak{L}} \equiv \mathcal{T}^{\{1\}}$. Podobně sloupce matice $\mathcal{C}^{\mathscr{C}_\mathfrak{R}} \equiv \mathcal{C}^{\{2\}} = \mathcal{C}^T$ jsou zároveň sloupci matice $\mathcal{T}^{\mathscr{C}_\mathfrak{R}} \equiv \mathcal{T}^{\{2\}}$.

4.2.2 Matice přenosu

Obor hodnot libovolné matice je roven oboru hodnot matice levých singulárních vektorů, které odpovídají nenulovým vlastním číslům. Zřejmě tvrzení z věty 1 ihned ukazuje platnost vztahu (4.5).

Nechť $U_{\mathscr{C}}, U_{\mathscr{C}_\mathfrak{L}}, U_{\mathscr{C}_\mathfrak{R}}$ jsou matice, jejichž sloupce tvoří libovolné báze sloupcových prostorů (tedy oborů hodnot) matic $\mathcal{T}^{\mathscr{C}}, \mathcal{T}^{\mathscr{C}_\mathfrak{L}}, \mathcal{T}^{\mathscr{C}_\mathfrak{R}}$. Z věty 1 a následujícího důkazu vyplývá, že existuje matice $\mathcal{B}_{\mathscr{C}}$ taková, že

$$U_{\mathscr{C}} = (U_{\mathscr{C}_\mathfrak{R}} \otimes U_{\mathscr{C}_\mathfrak{L}}) \cdot B_{\mathscr{C}}, \qquad B_{\mathscr{C}} \in \mathbb{R}^{(r_{\mathscr{C}_\mathfrak{L}} \cdot r_{\mathscr{C}_\mathfrak{R}}) \times r_{\mathscr{C}}}. \tag{4.14}$$

Speciálně, jsou-li sloupce matic $U_{\mathscr{C}}, U_{\mathscr{C}_\mathfrak{L}}, U_{\mathscr{C}_\mathfrak{R}}$ levé singulární vektory matic $\mathcal{T}^{\mathscr{C}}$, $\mathcal{T}^{\mathscr{C}_\mathfrak{L}}$, $\mathcal{T}^{\mathscr{C}_\mathfrak{R}}$, budeme tuto matici $B_{\mathscr{C}}$ nazývat *matice přenosu* (anglicky transfer matrix). Poznamenejme, že čísla $r_{\mathscr{C}_\mathfrak{L}}, r_{\mathscr{C}_\mathfrak{R}}, r_{\mathscr{C}}$ značí hodnosti odpovídajících rozvojů tenzoru $\mathcal{T}$, viz (4.6)–(4.7). Matice přenosu je prostředek, který nám umožní „rozbít“ tenzor do předepsaného tvaru sítě. Matici $B_{\mathscr{C}}$ můžeme chápat jako rozvoj tenzoru $\mathcal{B}_{\mathscr{C}}$ třetího řádu tak, že

$$B_{\mathscr{C}} = \mathcal{B}_{\mathscr{C}}{}^{\{1,2\}}, \qquad \text{kde} \qquad \mathcal{B}_{\mathscr{C}} \in \mathbb{R}^{r_{\mathscr{C}_\mathfrak{L}} \times r_{\mathscr{C}_\mathfrak{R}} \times r_{\mathscr{C}}}, \tag{4.15}$$

viz (4.7) a podrobněji také viz příklad 2. Tyto tenzory jsou tedy vrcholy třetího stupně v tenzorové síti představující HTD, viz obrázek 4.2.

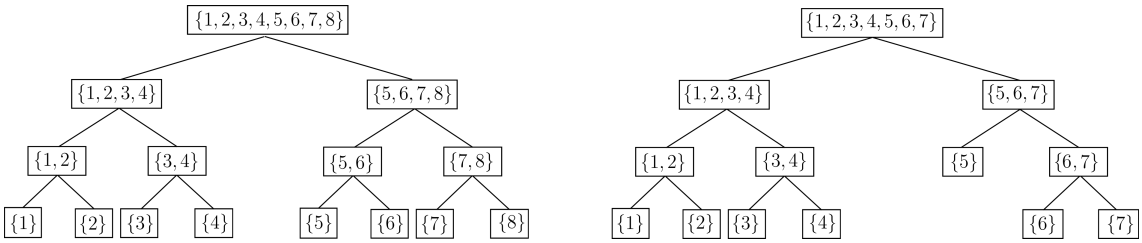
4.3 Shrnutí konstrukce hierarchického Tuckerova rozkladu

Předně poznamenejme, že postup při hledání hierarchického Tuckerova rozkladu, tak jak ho vysvětlujeme, je pouze náznakem výpočtu. Slouží zde zejména pro pochopení struktury rozkladu a jako důkaz jeho existence. Praktický výpočet rozkladu a implementace příslušného algoritmu není jednoduchá. Jeden z možných postupů výpočtu bude naznačen v kapitole 6.

4.3.1 Větvení binárního stromu a tzv. dimension tree

Uvědomme si, že podmnožiny množiny indexů lze vybírat libovolně. Právě způsob, jakým rozdělujeme množinu indexů (a postupně její podmnožiny), vytváří strom odpovídajícího tvaru. Pro dosažení potřebného tvaru sítě volíme v každém kroku

odpovídající rozdělení množiny indexů tenzoru. Tomuto rozdělení odpovídá tzv. *dimension tree*. Příklady rozdělení množiny indexů (konkrétně ty možnosti větvení, které dávají co nejvíce vyvážený binární strom) můžemem vidět na obrázku 4.3 pro tenzor osmého (vlevo) a sedmého řádu (vpravo).



Obrázek 4.3: Struktura rozdělení indexů, tzv. dimension tree, tenzoru osmého řádu z příkladu 2, převzato z [19, str.22], a tenzoru sedmého řádu při snaze o co největší vyváženost sítě.

Poznamenejme, že tensor train, jehož podobu můžeme vidět na obrázku 3.7 uprostřed, je rozkladem, který funguje úplně stejně jako HTD. Jediným rozdílem je způsob rozdělování množiny indexů. Jedna z podmnožin vždy obsahuje pouze jeden index a získáváme pro něj tedy přímo matici z klasického Tuckerova rozkladu a další krok algoritmu potom musíme provést vždy pouze pro jednu větev.

4.4 Efektivita uložení dat pomocí hierarchického Tuckerova rozkladu

Při práci s tenzory jsme často omezeni tím, že tenzor obsahující velké množství prvků nelze kvůli vysokým paměťovým nárokům uložit v počítači. Z tohoto důvodu vznikla celá řada různých algoritmů, tenzorových rozkladů, umožňující tenzor uložit pomocí menších objektů s významnou úsporou paměti. K nim zřejmě patří i Tuckerův rozklad a hierarchický Tuckerův rozklad. V této části chceme porovnat, kolik paměti ušetříme, budeme-li s takovými rozklady pracovat.

Označme

$$r = \max_{\mathscr{C}\subseteq\{1,\dots,k\}} \operatorname{rank}(\mathcal{T}^{\mathscr{C}}) \qquad \text{a} \qquad n = \max\{n_1,\dots,n_k\}. \tag{4.16}$$

Budeme porovnávat paměťové nároky, tj. počet reálných čísel, které je potřeba uložit, abychom získali tenzor, případně jeho dobrou aproximaci. Zřejmě tento počet můžeme odhadnout pomocí čísel r, n a k.

- V případě nerozloženého tenzoru je počet ukládaných reálných čísel shora omezen hodnotou n^k.
- V případě klasického Tuckerova rozkladu ukládáme k matic s rozměry nejvýše $n \times r$ a počet prvků Tuckerova jádra je omezen hodnotou r^k, tedy celkem $knr + r^k$ reálných čísel.

- V případě hierarchického Tuckerova rozkladu opět ukládáme k matic s rozměry nejvýše $n \times r$ (listy stromu). Je-li řád tenzoru mocninou dvou, tj. $k = 2^s$, pak zcela vyvážený binární strom Tuckerova jádra obsahuje právě jednu matici s rozměry nejvýše $r \times r$ (která je navíc diagonální; kořen stromu) a dále $k-2$ tenzorů třetího řádu s rozměry nejvýše $r \times r \times r$. Tedy celkem ukládáme $knr + (k-2) \cdot r^3 + r^2$ reálných čísel.
- Také v případě rozkladu typu tensor train (TT) ukládáme k matic s rozměry nejvýše $n \times r$, dále pak $(k-2)$ tenzorů třetího řádu s rozměry nejvýše $r \times r \times r$ a dvě matice s rozměry nejvýše $r \times r$. Tedy celkem ukládáme $knr + (k-2) \cdot r^3 + 2r^2$ reálných čísel.

Paměťové nároky jsou také shrnuty v tabulce 4.1 a ilustrovány na obrázku 4.4.

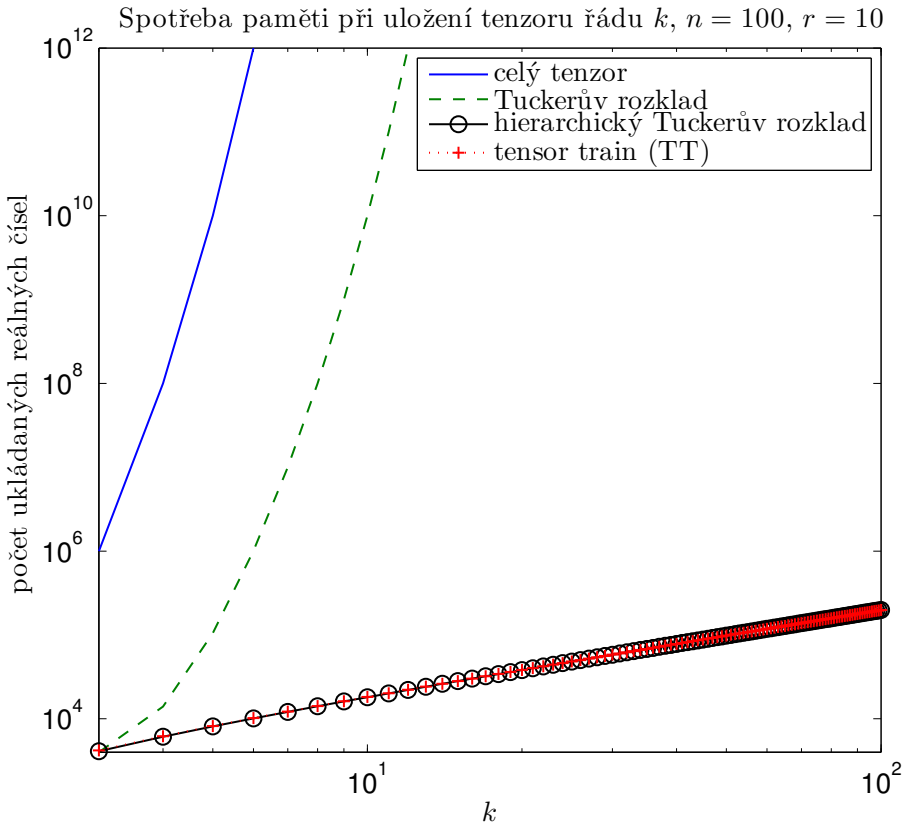
Ze způsobu konstrukce hierarchického Tuckerova rozkladu a tedy i tensor train (který se od HTD liší pouze způsobem větvení) tak, jak jsme popsali v kapitole 4.1.1, je zřejmé, že odhady u těchto dvou způsobů rozkladu můžeme navíc upřesnit, jelikož matice v kořeni stromu je diagonální, tj. obsahuje pouze r nenulových čísel. Počet reálných číslel potřebných k uložení je potom omezen na $knr + (k-2) \cdot r^3 + r$ čísel pro HTD a $knr + (k-2) \cdot r^3 + r^2 + r$ pro TT.

Poznamenejme dále, že vztah pro paměťové nároky hierarchického Tuckerova rozkladu je odvozen pro zcela vyvážený binární strom tenzoru řádu mocniny dvou, my ho ale budeme používat pro tenzor libovolného řádu. Můžeme si to dovolit proto, že TT odpovídá maximálně nevyváženému binárnímu stromu, přičemž vztah pro jeho paměťové nároky je odvozen pro tenzor libovolného řádu a dává prakticky stejný odhad.

Tabulka 4.1: Porovnání paměťových nároků při uložení tenzoru různými způsoby.

použitý rozklad	počet ukládaných reálných čísel
celý tenzor	n^k
Tuckerův rozklad	$knr + r^k$
hierarchický Tuckerův rozklad	$knr + (k-2)r^3 + r^2$
tensor train (TT)	$knr + (k-2)r^3 + 2r^2$

Z tabulky 4.1 vidíme, že zatímco paměťové nároky (počet ukládaných prvků) jsou u nerozloženého tenzoru exponenciální v k, pro hierarchický Tuckerův rozklad, příp. tesor train, jsou lineární v k a kubické v r. Případná úspora místa samozřejmě závisí na tom, jak malé může reálně být r pro daná data.



Obrázek 4.4: Porovnání paměťových nároků při uložení tenzoru různými způsoby.

5 Manipulace s tenzory ve tvaru HTD

Ukázali jsme už, jakým způsobem lze ukládat tenzory ve tvaru sítě. Dále nás bude zajímat, jakým způsobem lze s tenzory uloženými ve formátu HTD pracovat dále. Ukážeme si, jakým způsobem lze tenzory v HTD násobit maticí, sčítat i násobit mezi sebou. Budeme se navíc snažit, aby výsledný tenzor byl uložen opět v HTD a to v co nejúspornějším tvaru.

5.1 Součin tenzoru s maticí v ℓ-tém módu

První z operací, kterou popíšeme bude součin tenzoru s maticí v daném módu ℓ, viz definici 2. Mějme pro jednoduchost tenzor osmého řádu $\mathcal{T} \in \mathbb{R}^{n_1 \times \cdots \times n_8}$ (viz příklad 2, str. 33) a matici $M \in \mathbb{R}^{m \times n_\ell}$ a $\ell = 3$. Pro součin

$$\mathcal{D} = \mathcal{T} \times_3 M \in \mathbb{R}^{n_1 \times n_2 \times m \times n_4 \times \cdots \times n_8}$$

zřejmě platí

$$\operatorname{vec}(\mathcal{D}) = \operatorname{vec}(\mathcal{T} \times_3 M) = (I_{n_8} \otimes \cdots \otimes I_{n_4} \otimes M \otimes I_{n_2} \otimes I_{n_1}) \cdot \operatorname{vec}(\mathcal{T}), \tag{5.1}$$

kde vektorizaci tenzoru $\mathcal{T}$ lze pomocí vztahu (4.9) zapsat

$$\begin{aligned} \operatorname{vec}(\mathcal{T}) = {} & (U_{(8)} \otimes U_{(7)} \otimes U_{(6)} \otimes U_{(5)} \otimes U_{(4)} \otimes U_{(3)} \otimes U_{(2)} \otimes U_{(1)}) \\ & \cdot (B_{(7-8)} \otimes B_{(5-6)} \otimes B_{(3-4)} \otimes B_{(1-2)}) \cdot (B_{(5-8)} \otimes B_{(1-4)}) \cdot B_{(1-8)}. \end{aligned} \tag{5.2}$$

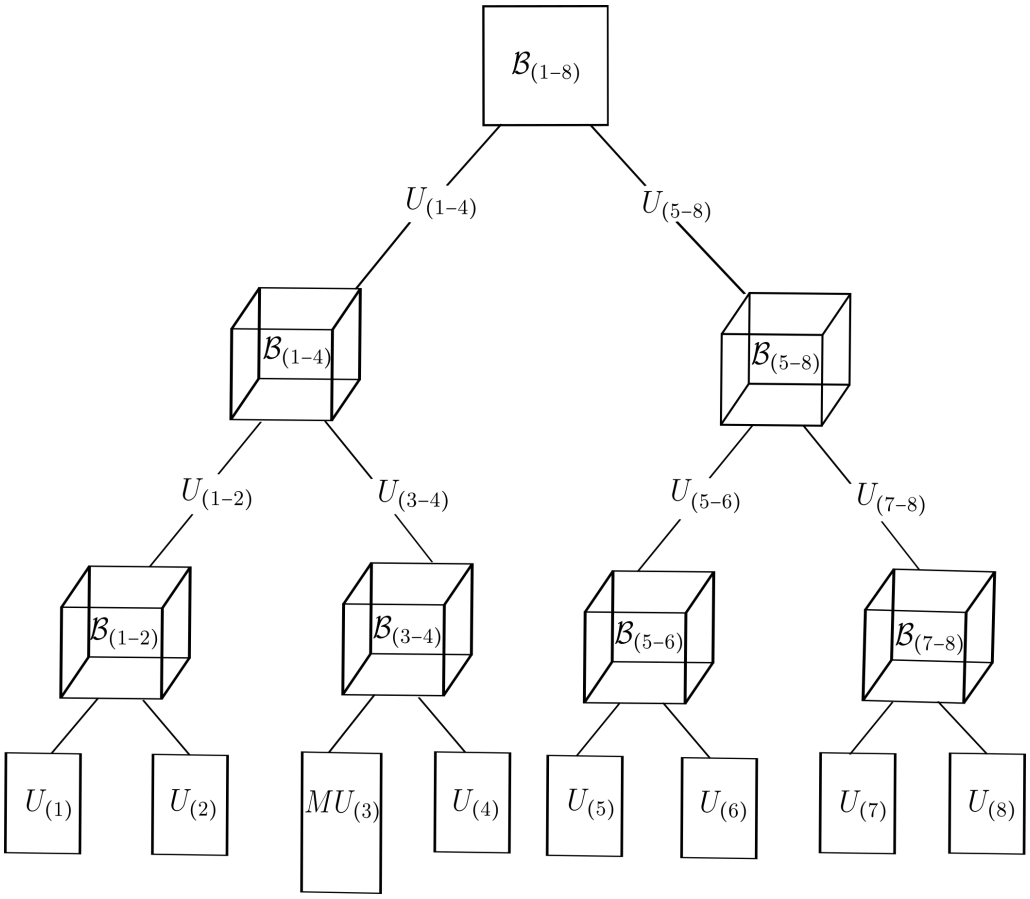
Kombinací vztahů (5.1) a (5.2) dostáváme tenzor $\mathcal{D}$, resp. jeho vektorizaci ve tvaru

$$\begin{aligned} \operatorname{vec}(\mathcal{D}) = {} & (I_{n_8} \otimes I_{n_7} \otimes I_{n_6} \otimes I_{n_5} \otimes I_{n_4} \otimes M \otimes I_{n_2} \otimes I_{n_1}) \\ & \cdot (U_{(8)} \otimes U_{(7)} \otimes U_{(6)} \otimes U_{(5)} \otimes U_{(4)} \otimes U_{(3)} \otimes U_{(2)} \otimes U_{(1)}) \\ & \cdot (B_{(7-8)} \otimes B_{(5-6)} \otimes B_{(3-4)} \otimes B_{(1-2)}) \cdot (B_{(5-8)} \otimes B_{(1-4)}) \cdot B_{(1-8)} \\ = {} & \Big(U_{(8)} \otimes U_{(7)} \otimes U_{(6)} \otimes U_{(5)} \otimes U_{(4)} \otimes (MU_{(3)}) \otimes U_{(2)} \otimes U_{(1)}\Big) \\ & \cdot (B_{(7-8)} \otimes B_{(5-6)} \otimes B_{(3-4)} \otimes B_{(1-2)}) \cdot (B_{(5-8)} \otimes B_{(1-4)}) \cdot B_{(1-8)}, \end{aligned}$$

kde $(MU_{(3)}) \in \mathbb{R}^{m \times r_{(3)}}$; s využitím vztahu mezi klasickým maticovým násobením a Kroneckerovým součinem matic, viz např. [26, poznámka 3].

Slovně vyjádřeno, pokud je tenzor $\mathcal{T}$ uložený v HTD, vynásobením listu $U_{(\ell)}$ maticí M získáme součin tenzoru $\mathcal{T}$ s maticí M v módu ℓ, který *formálně* vypadá jako

hierarchický Tuckerův rozklad, viz obrázek 5.1. Tedy je vyjádřený jako tenzorová síť, resp. binární strom se stejnou strukturou jako původní tenzor $\mathcal{T}$. Narozdíl od HTD ale ℓ-tý list stromu tenzoru $\mathcal{D}$, tj. matice $(MU_{(\ell)})$, obecně nemá navzájem ortonormální sloupce. Abychom HTD získali, je potřeba provést reortogonalizaci sloupců této matice a následně přepočítat ostatní dotčené tenzory sítě. Těmito kroky se budeme podrobněji zabývat v kapitole 5.3.



Obrázek 5.1: Ilustrace součinu tenzoru (z obrázku 4.2) osmého řádu s maticí M ve třetím módu.

5.1.1 Lineární zobrazení ve tvaru Kroneckerova součinu

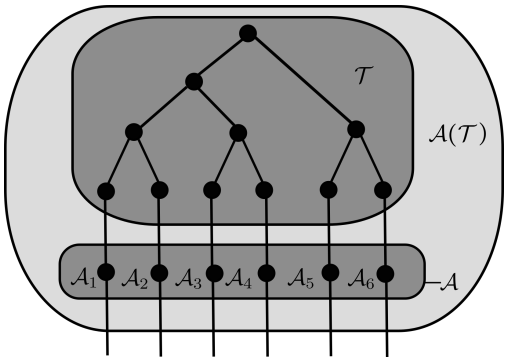
Speciálně pro lineární zobrazení, které lze zapsat ve tvaru Kroneckerova součinu, tj.

$$\mathcal{A} : \mathcal{T} \longmapsto \mathcal{D}, \qquad \text{kde} \qquad \mathcal{A} = A_k \otimes A_{k-1} \otimes \cdots \otimes A_1, \tag{5.3}$$

přičemž oba uvažované tenzory $\mathcal{T}$ i $\mathcal{D}$ jsou nyní řádu k (pro jednoduchost uvažujme $k = 2^\varsigma$, kde ς je přirozené číslo), zřejmě platí

$$\begin{aligned}\operatorname{vec}(\mathcal{D}) = \operatorname{vec}(\mathcal{A}(\mathcal{T})) = \Big((A_k U_{(k)}) \otimes (A_{k-1} U_{(k-1)}) \otimes \cdots \otimes (A_1 U_{(1)})\Big) \\ \cdot (B_{((k-1)-k)} \otimes \cdots \otimes B_{(1-2)}) \cdot (B_{((k-3)-k)} \otimes \cdots \otimes B_{(1-4)}) \\ \cdot \ \ldots \\ \cdot (B_{((k/2+1)-k)} \otimes B_{(1-(k/2))}) \cdot B_{(1-k)}.\end{aligned}$$

Schematicky lze součin vyjádřit pomocí tenzorové sítě na obrázku 5.2.



Obrázek 5.2: Lineární zobrazení ve tvaru Kroneckerova součinu.

5.2 Součet dvou tenzorů

Součet tenzorů stejného řádu a stejných rozměrů získáme jednoduše, bez jakýchkoli aritmetických operací pouhým zřetězením odpovídajících faktorů, obdobně jako je ukázáno v [26, kap. 1.3.1 a 3.2.1] pro matice ve tvaru singulárních rozkladů, resp. tenzory v Tuckerově rozkladu.

Mějme například dva tenzory $\mathcal{C}, \mathcal{D} \in \mathbb{R}^{n_1 \times \cdots \times n_4}$ v HTD popsaném stejným stromem, tj.

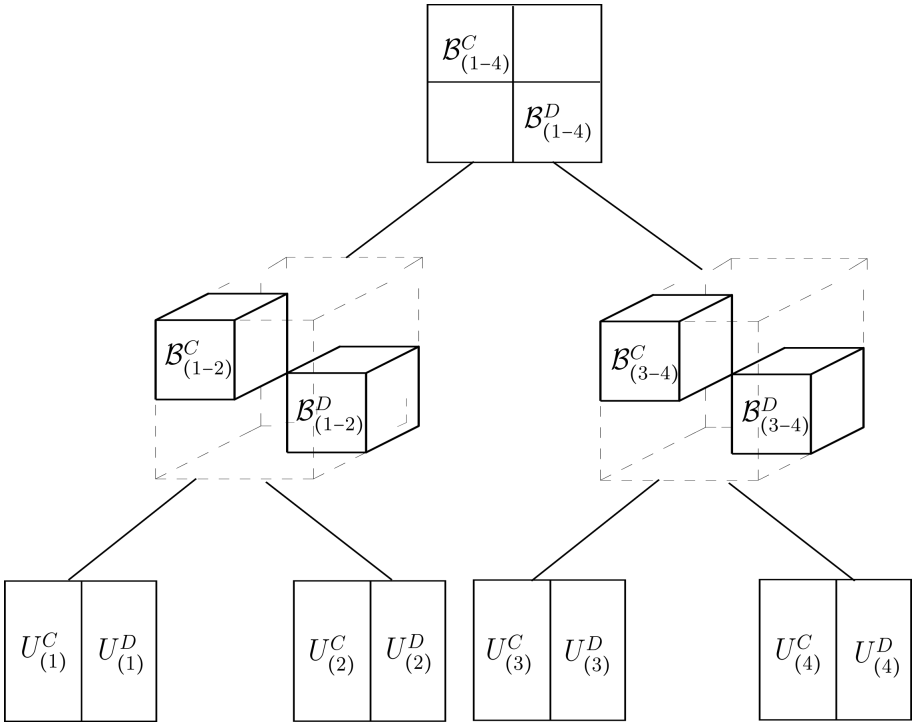
$$\begin{aligned}\operatorname{vec}(\mathcal{C}) &= (U^C_{(4)} \otimes U^C_{(3)} \otimes U^C_{(2)} \otimes U^C_{(1)}) \cdot (B^C_{(3-4)} \otimes B^C_{(1-2)}) \cdot B^C_{(1-4)}, \\ \operatorname{vec}(\mathcal{D}) &= (U^D_{(4)} \otimes U^D_{(3)} \otimes U^D_{(2)} \otimes U^D_{(1)}) \cdot (B^D_{(3-4)} \otimes B^D_{(1-2)}) \cdot B^D_{(1-4)}.\end{aligned} \tag{5.4}$$

Potom jejich součet $\mathcal{E} = \mathcal{C} + \mathcal{D}$ získáme seřazením příslušných matic $U^C_{(\ell)}$ a $U^D_{(\ell)}$ za sebe, tj. dostaneme matice

$$[U^C_{(\ell)}, U^D_{(\ell)}] \in \mathbb{R}^{n_{(\ell)} \times (r^C_{(\ell)} + r^D_{(\ell)})}, \qquad \ell = 1, \ldots 4,$$

a diagonálním zřetězením příslušných tenzorů $\mathcal{B}$ odpovídajících jednotlivým maticím přenosu. Pro pochopení nejlépe poslouží příklad na obrázku 5.3.

Takovýmto způsobem však získáme, stejně tak jako v případě součinu tenzoru s maticí, tenzorovou síť, která má *formálně* stejnou strukturu jako původní tenzory,



Obrázek 5.3: Ilustrace součtu dvou tenzorů $\mathcal{C}$ a $\mathcal{D}$ (5.4) čtvrtého řádu ve tvaru HTD.

ale matice odpovídající listům nemají ortogonální sloupce a tedy tato síť není hierarchickým Tuckerovým rozkladem tenzoru $\mathcal{E}$. Pro získání takové sítě musíme provést reortogonalizaci jako v předchozím případě; viz kap. 5.3.

Zde navíc, na rozdíl od součinu tenzoru s maticí, vzniká praktický problém s velikostí ukládaných dat. Vidíme, že takto zkonstruovaný tenzor $\mathcal{E} = \mathcal{C} + \mathcal{D}$ potřebuje circa dvakrát více paměťových prostředků než tenzory $\mathcal{C}$ nebo $\mathcal{D}$. To by samo o sobě nevadilo, pokud nebude potřeba takovou operaci provádět opakovaně, např. při řešení soustavy lineárních rovnic (tj. $\mathcal{A}(\mathcal{X}) = \mathcal{B}$ s tenzorovou pravou stranou a lineárním zobrazením ve tvaru Kroneckerova součinu) iterační metodou. Tehdy by rostly paměťové nároky exponenciálně s číslem iterace.

5.2.1 Lineární kombinace tenzorů

Víme-li jak tenzory ve tvaru hierarchického Tuckerova rozkladu sčítat, už není těžké pochopit, jak bude vypadat linerární kombinace tenzorů stejných řádů se stejnou strukturou binárního stromu HTD. Mějme tenzory $\mathcal{T}_i \in \mathbb{R}^{n_1 \times \cdots \times n_k}$ v HTD se stejným stromem, a koeficienty $\alpha_i \in \mathbb{R}$ tvořící lineární kombinaci

$$\mathcal{E} = \sum_i \alpha_i \mathcal{T}_i.$$

Zřejmě pro α-násobek tenzoru $\mathcal{T}$ (pro jednoduchost řádu $k = 2^s$) platí

$$\begin{aligned}\operatorname{vec}(\alpha\,\mathcal{T}) = \alpha\operatorname{vec}(\mathcal{T}) = \alpha\,&(U_{(k)}\otimes U_{(k-1)}\otimes\cdots\otimes U_{(1)})\\ &\cdot(B_{((k-1)-k)}\otimes\cdots\otimes B_{(1-2)})\cdot(B_{((k-3)-k)}\otimes\cdots\otimes B_{(1-4)})\\ &\cdot\ \cdots\\ &\cdot(B_{((k/2+1)-k)}\otimes B_{(1-(k/2))})\cdot B_{(1-k)}\\ = (&U_{(k)}\otimes U_{(k-1)}\otimes\cdots\otimes U_{(1)})\\ &\cdot(B_{((k-1)-k)}\otimes\cdots\otimes B_{(1-2)})\cdot(B_{((k-3)-k)}\otimes\cdots\otimes B_{(1-4)})\\ &\cdot\ \cdots\\ &\cdot(B_{((k/2+1)-k)}\otimes B_{(1-(k/2))})\cdot(\alpha\,B_{(1-k)}).\end{aligned}$$

Připomeňme, že kořen $\mathcal{B}_{(1-k)}$ binárního stromu představujícího HTD tenzoru je diagonální maticí se singulárními čísly rozvoje tenzoru $\mathcal{T}$ podle multiindexu daného větvením stromu; viz kap. 4.1.1 a obrázek 4.2.

Pro výpočet lineární kombinace $\mathcal{E} = \sum_i \alpha_i\mathcal{T}_i$ tedy potřebujeme

- nejprve získat jednotlivé sčítance $\alpha_i\mathcal{T}_i$, které dostaneme tak, že číslem α_i vynásobíme diagonální matici v kořeni stromu tenzoru $\mathcal{T}_i$, a
- provést součet tenzorů $\alpha_i\mathcal{T}_i$, který budeme získávat postupem, který je popsaný v předchozím textu.

Takto získáme tenzor $\mathcal{E}$ ve tvaru tenzorové sítě se stejnou strukturou jako měly tenzory $\mathcal{T}_i$, přičemž matice uložené jako listy binárního stromu opět nemají navzájem ortogonální sloupce. I v tomto případě budeme provádět reortogonalizaci.

5.3 Reortogonalizace a rekomprese

V předchozích částech textu (kapitoly 5.1 a 5.2) jsme popsali první kroky některých operací s tenzory ve tvaru hierarchického Tuckerova rozkladu, kdy výsledkem byly vždy tenzory, které se formálně strukturou podobaly původním tenzorům. Často však chceme i výsledný tenzor ukládat ve tvaru HTD, proto potřebujeme udělat ještě několik kroků, které nám toto umožní, konkrétně to bude

- ortogonalizace listů binárního stromu,
- ortogonalizace rozvojů tenzorů třetího řádu – výpočet nových matic přenosu,
- komprese rozměrů faktorů binárního stromu.

Princip bude analogický principu při operacích s maticemi uloženými v ekonomickém tvaru singulárního rozkladu, resp. s tenzory ve tvaru klasického Tuckerova rozkladu, viz [26, kap. 1.3], resp. [26, kap. 3.2.1].

Pro vysvětlení mechanismu výpočtu nových matic přenosu připomeňme nejprve vztah (4.14), tj.

$$U_{\mathscr{C}} = (U_{\mathscr{C}_{\mathfrak{R}}}\otimes U_{\mathscr{C}_{\mathfrak{L}}})\cdot B_{\mathscr{C}},\qquad \text{kde}\qquad B_{\mathscr{C}}\in\mathbb{R}^{(r_{\mathscr{C}_{\mathfrak{L}}}\cdot r_{\mathscr{C}_{\mathfrak{R}}})\times r_{\mathscr{C}}},\tag{5.5}$$

je matice přenosu a kde $U_{\mathscr{C}}, U_{\mathscr{C}_\mathfrak{R}}, U_{\mathscr{C}_\mathfrak{L}}$ jsou matice levých singulárních vektorů tvořící ortonormální báze oborů hodnot příslušných rozvojů tenzoru. Tedy platí

$$U_{\mathscr{C}}^T U_{\mathscr{C}} = I, \qquad U_{\mathscr{C}_\mathfrak{R}}^T U_{\mathscr{C}_\mathfrak{R}} = I, \qquad U_{\mathscr{C}_\mathfrak{L}}^T U_{\mathscr{C}_\mathfrak{L}} = I,$$

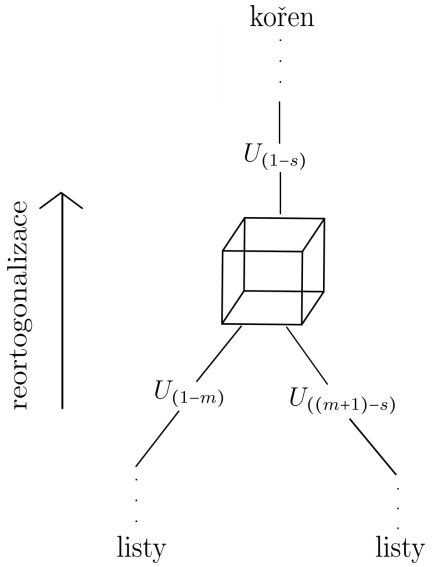
kde jednotkové matice na pravých stranách jsou vhodných (obecně různých) řádů. Pak zřejmě také platí

$$\begin{aligned}(U_{\mathscr{C}_\mathfrak{R}} \otimes U_{\mathscr{C}_\mathfrak{L}})^T (U_{\mathscr{C}_\mathfrak{R}} \otimes U_{\mathscr{C}_\mathfrak{L}}) &= (U_{\mathscr{C}_\mathfrak{R}}^T \otimes U_{\mathscr{C}_\mathfrak{L}}^T)(U_{\mathscr{C}_\mathfrak{R}} \otimes U_{\mathscr{C}_\mathfrak{L}}) \\ &= (U_{\mathscr{C}_\mathfrak{R}}^T U_{\mathscr{C}_\mathfrak{R}}) \otimes (U_{\mathscr{C}_\mathfrak{L}}^T U_{\mathscr{C}_\mathfrak{L}}) = I \otimes I = I,\end{aligned} \tag{5.6}$$

jak plyne z vlastností Kroneckerova součinu. Kombinací předchozích rovnic získáme vztah

$$I = U_{\mathscr{C}}^T U_{\mathscr{C}} = \mathcal{B}_{\mathscr{C}}^T (U_{\mathscr{C}_\mathfrak{R}} \otimes U_{\mathscr{C}_\mathfrak{L}})^T (U_{\mathscr{C}_\mathfrak{R}} \otimes U_{\mathscr{C}_\mathfrak{L}}) \mathcal{B}_{\mathscr{C}} = \mathcal{B}_{\mathscr{C}}^T \mathcal{B}_{\mathscr{C}}, \tag{5.7}$$

tedy také matice přenosu má ortonormální sloupce. Této vlastnosti budeme využívat při reortogonalizaci. Budeme vždy postupovat od listů ke kořeni, tak jak naznačuje obrázek 5.4. Postup pro jednotlivé operace rozebereme v samostatných podkapitolách.



Obrázek 5.4: Znázornění postupu reortogonalizace.

5.3.1 Reortogonalizace součinu tenzoru s maticí

V prvním kroku součinu tenzoru v HTD s maticí, který jsme popsali na příkladu v kapitole 5.1, tj. součinu

$$\mathcal{D} = \mathcal{T} \times_3 M, \qquad \mathcal{T} \in \mathbb{R}^{n_1 \times \cdots \times n_8}, \qquad M \in \mathbb{R}^{m \times n_3},$$

jsme na místě třetího listu získali součin

$$MU_{(3)} \in \mathbb{R}^{m \times r_{(3)}},$$

viz obrázek 5.1. Abychom získali výsledný tenzor $\mathcal{D}$ v hierarchickém Tuckerově rozkladu, potřebujeme nejdříve zajistit, aby matice – listy binárního stromu měly ortogonální sloupce. V našem příkladu je nutné ortogonalizovat pouze sloupce třetího listu, tedy matice $MU_{(3)}$.

Reortogonalizace listu

Ortogonální sloupce zajistíme pomocí QR rozkladu (viz např. [4, kap. 3]) této matice, tj. dostaneme

$$MU_{(3)} = Q_{(3)}R_{(3)}, \qquad \text{kde} \qquad Q_{(3)} \in \mathbb{R}^{m\times\widetilde{r}_{(3)}} \quad \text{a} \quad R_{(3)} \in \mathbb{R}^{\widetilde{r}_{(3)}\times r_{(3)}}, \tag{5.8}$$

kde

$$\widetilde{r}_{(3)} = \operatorname{rank}(MU_{(3)}) \le r_{(3)}. \tag{5.9}$$

Matice $Q_{(3)}$ má ortogonální sloupce a bude tedy listem binárního stromu tak, jak je naznačeno na obrázku 5.5.

Poznamenejme, že $\widetilde{r}_{(3)} \le r_{(3)}$ způsobí, že matice $R_{(3)}$ obecně není ryze trojúhelníková (anglicky proper upper triangular), ale je v tzv. *horním schodovitém tvaru* (anglicky row echelon form). Zde je mimo jiné prostor pro **kompresi** – zanedbáváním vhodně určených malých prvků matice R z QR rozkladu listu (resp. listů) se můžeme cíleně snažit o snížení hodnoty $\widetilde{r}_{(3)}$.

Násobení trojúhelníkovým faktorem

Nyní když jsme zajistili, že list má vzájemně ortogonální sloupce, nás bude zajímat, jak se projeví matice $R_{(3)}$ z QR rozkladu ve zbytku tenzorové sítě. Dalším krokem tedy bude násobení tenzoru $\mathcal{B}_{(3\text{–}4)}$ maticí $R_{(3)}$.

Připomeňme vztah (4.14) vyjadřující vztah matic $U_{(\ell)}$ a matic přenosu. Aplikujeme-li tento vztah pro tenzor $\mathcal{T}$ z našeho příkladu, platí

$$U_{(3\text{–}4)} = (U_{(4)} \otimes U_{(3)}) \cdot B_{(3\text{–}4)}, \qquad \text{kde} \qquad B_{(3\text{–}4)} = \mathcal{B}^{\{1,2\}}_{(3\text{–}4)}. \tag{5.10}$$

Pro součin $\mathcal{T} \times_3 M$, potom kombinací vztahů (5.10) a (5.8) dostáváme

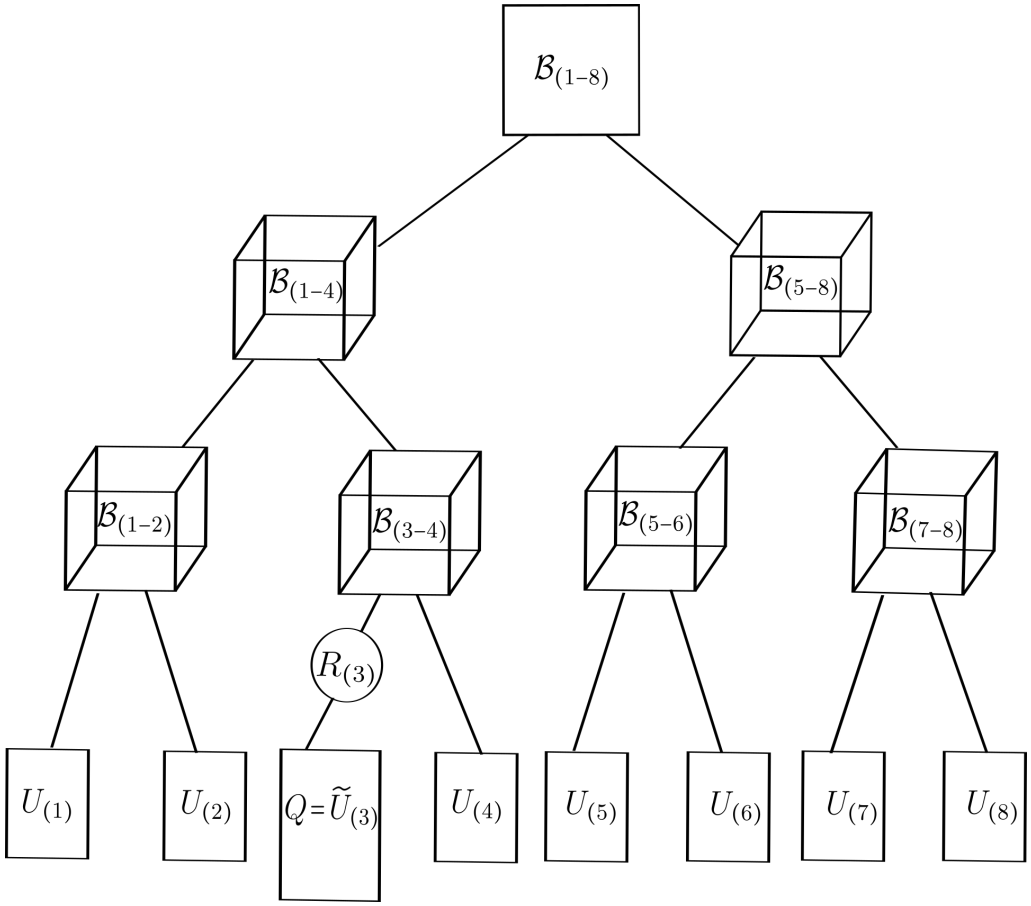
$$\begin{aligned} \widetilde{U}_{(3\text{–}4)} &= (U_{(4)} \otimes Q_{(3)}R_{(3)}) \cdot B_{(3\text{–}4)} \\ &= (U_{(4)} \otimes Q_{(3)})(I \otimes R_{(3)}) \cdot \mathcal{B}^{\{1,2\}}_{(3\text{–}4)} = (U_{(4)} \otimes Q_{(3)}) \cdot \widehat{\mathcal{B}}^{\{1,2\}}_{(3\text{–}4)}, \end{aligned} \tag{5.11}$$

tj.

$$\widehat{\mathcal{B}}_{(3\text{–}4)} = \mathcal{B}_{(3\text{–}4)} \times_1 R_{(3)} \in \mathbb{R}^{\widetilde{r}_{(3)}\times r_{(4)}\times r_{(3\text{–}4)}}. \tag{5.12}$$

Poznamenejme, že vztah (5.9) zaručuje, že násobení maticí R lze vždy provést, pokud nastane $\widetilde{r}_{(3)} < r_{(3)}$, stačí doplnit matici R nulovými prvky do potřebných rozměrů.

Matice $\widehat{\mathcal{B}}^{\{1,2\}}_{(3\text{–}4)}$ nyní ale není maticí přenosu v pravém slova smyslu (tj. jak jsme ji zavedli na str. 37; viz (4.14)), nemá ortonormální sloupce. Abychom z ní matici přenosu vytvořili, musíme zortogonalizovat sloupce této matice.



Obrázek 5.5: Schéma ortogonalizace listu binárního stromu při součinu tenzoru s maticí. Matice Q z QR rozkladu je uložena jako list binárního stromu, maticí $R_{(3)}$ budeme násobit příslušný tenzor třetího řádu.

Reortogonalizace matice přenosu

Pro získání ortonormální báze sloupcového prostoru matice $\widehat{\mathcal{B}}^{\{1,2\}}_{(3-4)}$ provedeme opět její QR rozklad, tj.

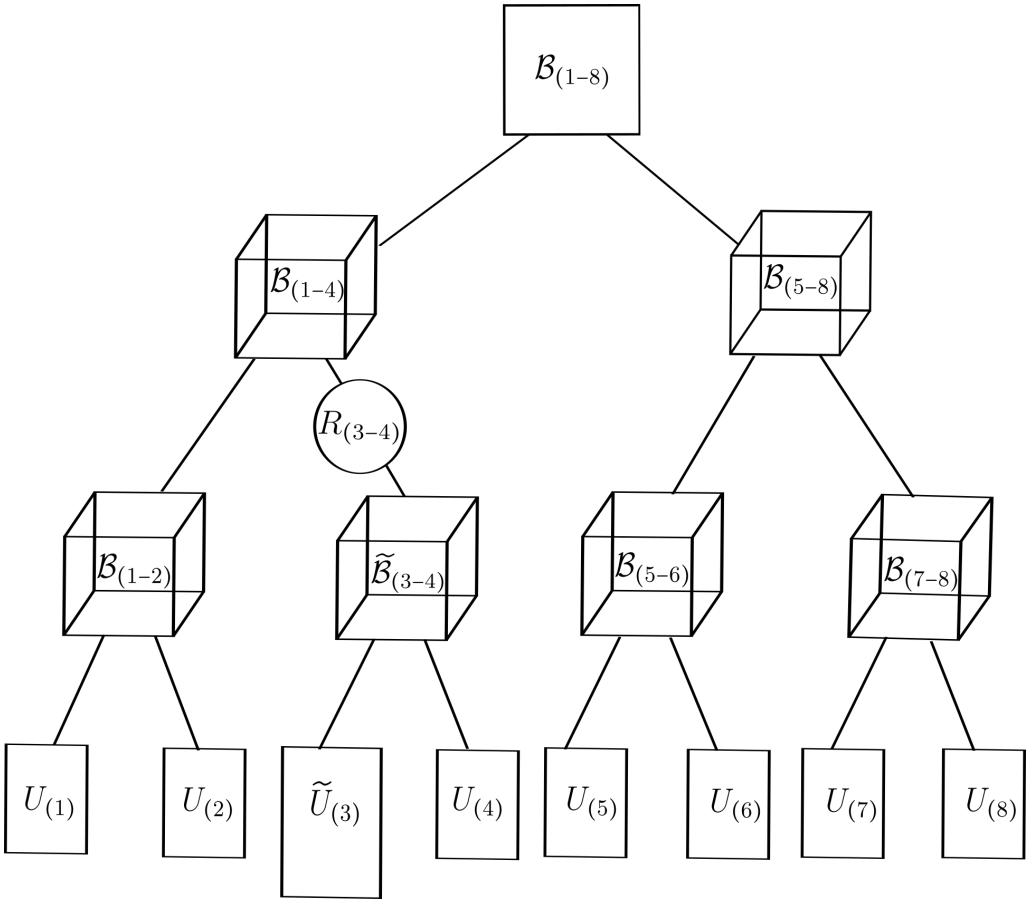
$$\widehat{\mathcal{B}}^{\{1,2\}}_{(3-4)} = Q_{(3-4)} R_{(3-4)}, \quad Q_{(3-4)} \in \mathbb{R}^{(\widetilde{r}_{(3)} \cdot r_{(4)}) \times \widetilde{r}_{(3-4)}}, \quad R_{(3-4)} \in \mathbb{R}^{\widetilde{r}_{(3-4)} \times r_{(3-4)}}, \tag{5.13}$$

kde $\widetilde{r}_{(3-4)} = \operatorname{rank}(\widehat{\mathcal{B}}^{\{1,2\}}_{(3-4)})$. Zde matice $Q_{(3-4)}$ má ortonormální sloupce a je nově vypočítanou maticí přenosu. Označíme formálně $Q_{(3-4)} \equiv \widetilde{B}_{(3-4)}$, tj. rozvoj tenzoru $\widetilde{\mathcal{B}}^{\{1,2\}}_{(3-4)}$, který bude uložen v binárním stromu HTD výsledného tenzoru; pro ilustraci viz obrázek 5.6.

Stejným způsobem postupujeme dále binárním stromem, a tedy dalším krokem je součin

$$\widehat{\mathcal{B}}_{(1-4)} = \mathcal{B}_{(1-4)} \times_2 R_{(3-4)},$$

QR rozklad rozvoje $\widehat{\mathcal{B}}^{\{1,2\}}_{(1-4)} = Q_{(1-4)} R_{(1-4)}$ atd., dokud se nedostaneme ke kořeni binárního stromu. Vektorizaci tenzoru $\mathcal{D} = \mathcal{T} \times_3 M$ tedy po těchto krocích dostáváme



Obrázek 5.6: Schéma postupu reortogonalizace, kdy jsme získali reortogonalizovanou matici přenosu, uložili ji do binárního stromu jako tenzor třetího řádu a maticí $R_{(3-4)}$ z QR rozkladu budeme násobit dále.

v podobě

$$\begin{aligned}\operatorname{vec}(\mathcal{D}) = {} & (U_{(8)} \otimes U_{(7)} \otimes U_{(6)} \otimes U_{(5)} \otimes U_{(4)} \otimes \widetilde{U}_{(3)} \otimes U_{(2)} \otimes U_{(1)}) \\ & \cdot (B_{(7-8)} \otimes B_{(5-6)} \otimes \widetilde{B}_{(3-4)} \otimes B_{(1-2)}) \cdot (B_{(5-8)} \otimes \widetilde{B}_{(1-4)}) \cdot \widehat{B}_{(1-8)}.\end{aligned} \tag{5.14}$$

Formálně ale ještě nemáme HTD tenzoru $\mathcal{D}$, protože kořen stromu – matice $\widehat{\mathcal{B}}_{(1-8)}$ (pozn. $\widehat{B}_{(1-8)} = \operatorname{vec}(\widehat{\mathcal{B}}_{(1-8)})$) – není diagonální; tím se budeme zabývat později, viz kap. 5.3.3.

5.3.2 Reortogonalizace součtu dvou tenzorů

V případě součtu dvou tenzorů budeme postupovat analogicky jako v případě součinu tenzoru s maticí. Jediným rozdílem je, že ortogonalitu sloupců nebudeme potřebovat zajistit jen pro jednu matici (jeden list binárního stromu), ale pro všechny listy binárního stromu, tj. matice $[U^C_{(\ell)}, U^D_{(\ell)}]$, sestaveného tak, jak jsme popsali v kapitole 5.2. Stejně tak budeme chít zajistit, aby tenzory třetího řádu v binárním stromu odpovídaly maticím přenosu, a tedy jejich rozvoje měly ortonormální sloupce.

Reortogonalizace listů

Spočítáme tedy QR rozklady listů stromu, tj.

$$[U^C_{(\ell)}, U^D_{(\ell)}] = Q_{(\ell)} R_{(\ell)}, \qquad \text{kde} \qquad Q_{(\ell)} \in \mathbb{R}^{n_{(\ell)} \times \widetilde{r}_{(\ell)}} \quad \text{a} \quad R_{(\ell)} \in \mathbb{R}^{\widetilde{r}_{(\ell)} \times (r^C_{(\ell)} + r^D_{(\ell)})}$$

(srovnej s (5.8)). Každá matice $Q_{(\ell)}$ má ortogonální sloupce a tedy všechny tyto matice budou uloženy jako listy binárního stromu HTD tenzoru $\mathcal{E}$ formálně označené $\widetilde{U}^E_{(\ell)} \equiv Q_{(\ell)}$. Matice $R_{(\ell)}$ potom musíme vynásobit příslušné tenzory binárního stromu.

Násobení trojúhelníkovými faktory

V našem příkladu z kapitoly 5.2 označme jako $\mathcal{B}^E_{(1-2)}$ diagonální tenzor sestavený z tenzorů $\mathcal{B}^C_{(1-2)}$ a $\mathcal{B}^D_{(1-2)}$. Provedeme tedy součin

$$\widehat{\mathcal{B}}^E_{(1-2)} = \mathcal{B}^E_{(1-2)} \times_1 R_{(1)} \times_2 R_{(2)} \in \mathbb{R}^{\widetilde{r}_{(1)} \times \widetilde{r}_{(2)} \times r_{(1-2)}}$$

(srovnej s (5.12)). Analogicky budeme provádět další součiny, tenzorů $\mathcal{B}$ a matic R. Takto získané tenzory $\widehat{\mathcal{B}}$ ale opět nereprezentují matice přenosu, protože nemají ortonormální sloupce.

Reortogonalizace matic přenosu

Pro matice $\widehat{\mathcal{B}}^{\{1,2\}}$ tedy vždy provedeme QR rozklad, tj. v našem příkladu získáme

$$(\widehat{\mathcal{B}}^E_{(1-2)})^{\{1,2\}} = Q_{(1-2)} R_{(1-2)}, \quad Q_{(1-2)} \in \mathbb{R}^{(\widetilde{r}_{(1)} \cdot \widetilde{r}_{(2)}) \times \widetilde{r}_{(1-2)}}, \quad R_{(1-2)} \in \mathbb{R}^{\widetilde{r}_{(1-2)} \times r_{(1-2)}},$$

kde $\widetilde{r}_{(1-2)} = \operatorname{rank}((\widehat{\mathcal{B}}^E_{(1-2)})^{\{1,2\}})$ (srovnej s (5.13)); atd. Matice Q mají ortonormální sloupce a proto odpovídají maticím přenosu. Označíme-li formálně $\widetilde{B}^E_{(1-2)} \equiv Q_{(1-2)}$ atd., tyto matice jsou rozvoji tenzorů třetího řádu v binárním stromu HTD tenzoru $\mathcal{E}$, tj. například $\widetilde{B}^E_{(1-2)} = (\widetilde{\mathcal{B}}^E_{(1-2)})^{\{1,2\}}$. Tímto způsobem postupujeme dále směrem ke kořeni stromu. Připomeňme znovu, že v případě součtu tenzorů se ortogonalizace bude týkat všech tenzorů v binárním stromu tenzoru $\mathcal{E}$, tj. nakonec dostáváme

$$\operatorname{vec}(\mathcal{E}) = (\widetilde{U}^E_{(4)} \otimes \widetilde{U}^E_{(3)} \otimes \widetilde{U}^E_{(2)} \otimes \widetilde{U}^E_{(1)}) \cdot (\widetilde{B}^E_{(3-4)} \otimes \widetilde{B}^E_{(1-2)}) \cdot \widehat{B}^E_{(1-4)}. \tag{5.15}$$

Formálně ale ještě nemáme HTD tenzoru $\mathcal{D}$, protože kořen stromu – matice $\widehat{\mathcal{B}}^E_{(1-4)}$ (pozn. $\widehat{B}^E_{(1-4)} = \operatorname{vec}(\widehat{\mathcal{B}}^E_{(1-4)})$) – není diagonální.

5.3.3 Aktualizace kořene stromu

Při operacích s tenzory a následné ortogonalizaci faktorů binárního stromu získáváme v kořeni stromu matici $\widehat{B}_{(1-k)}$ (resp. $\widehat{B}^E_{(1-k)}$) $\in \mathbb{R}^{\widetilde{r}_{(1-s)} \times \widetilde{r}_{((s+1)-k)}}$, která obecně není diagonální, jak bychom od HTD požadovali.

Pro získání diagonální matice v kořeni binárního stromu HTD tenzoru stačí provést singulární rozklad matice $\widehat{\mathcal{B}}_{(1-k)}$, tj.

$$\widehat{\mathcal{B}}_{(1-k)} = U_K \Sigma_K V_K^T, \qquad \text{kde} \qquad U_K \in \mathbb{R}^{\widetilde{r}_{(1-s)} \times r_K}, \quad \Sigma_K \in \mathbb{R}^{r_K \times r_K}, \quad V_K \in \mathbb{R}^{\widetilde{r}_{((s+1)-k)} \times r_K}$$

přičemž matice Σ_K je diagonální maticí se singulárními čísly matice $\widetilde{\mathcal{B}}_{(1-s)}$ na diagonále, která bude novým kořenem reortogonalizovaného binárního stromu, formálně ho označíme $\widetilde{\mathcal{B}}_{(1-k)} \equiv \Sigma_K$; $r_K = \operatorname{rank}(\widehat{\mathcal{B}}_{(1-k)})$. Dále zbývá už jen vynásobit maticemi U_K, V_K příslušné tenzory v síti HTD, tj. provét součiny

$$\widetilde{\widetilde{\mathcal{B}}}_{(1-s)} = \widetilde{\mathcal{B}}_{(1-s)} \times_3 U_K \qquad \text{a} \qquad \widetilde{\widetilde{\mathcal{B}}}_{((s+1)-k)} = \widetilde{\mathcal{B}}_{((s+1)-k)} \times_3 V_K. \tag{5.16}$$

Jelikož tenzory násobíme maticemi s ortogonálními sloupci, zůstanou ortogonální i rozvoje těchto tenzorů, tj. výsledné tenzory budou maticemi přenosu. Stačí si uvědomit, že předchozí rovnici lze přepsat

$$\widetilde{\widetilde{\mathcal{B}}}^{\{1,2\}}_{(1-s)} = \widetilde{B}_{(1-s)} U_K \qquad \text{a} \qquad \widetilde{\widetilde{\mathcal{B}}}^{\{1,2\}}_{((s+1)-k)} = \widetilde{B}_{((s+1)-k)} V_K \tag{5.17}$$

a tedy

$$\begin{aligned} (\widetilde{\widetilde{\mathcal{B}}}^{\{1,2\}}_{(1-s)})^T \widetilde{\widetilde{\mathcal{B}}}^{\{1,2\}}_{(1-s)} &= U_K^T \widetilde{B}^T_{(1-s)} \widetilde{B}_{(1-s)} U_K = I, \\ (\widetilde{\widetilde{\mathcal{B}}}^{\{1,2\}}_{((s+1)-k)})^T \widetilde{\widetilde{\mathcal{B}}}^{\{1,2\}}_{((s+1)-k)} &= V_K^T \widetilde{B}^T_{((s+1)-k)} \widetilde{B}_{((s+1)-k)} V_K = I. \end{aligned}$$

Tyto součiny jsou posledními kroky pro získání hierarchického Tuckerova rozkladu tenzoru, který je výsledkem některé z předchozích operací.

Zde je také další prostor pro **kompresi** – malá nově vypočtená singulární čísla můžeme v některých případech (v závislosti na aplikaci) zanedbat (položit rovny nule) a získat tak aproximaci původního tenzoru tenzorem nižší hodnosti r_K a tedy i s nižšími paměťovými nároky při jeho ukládání.

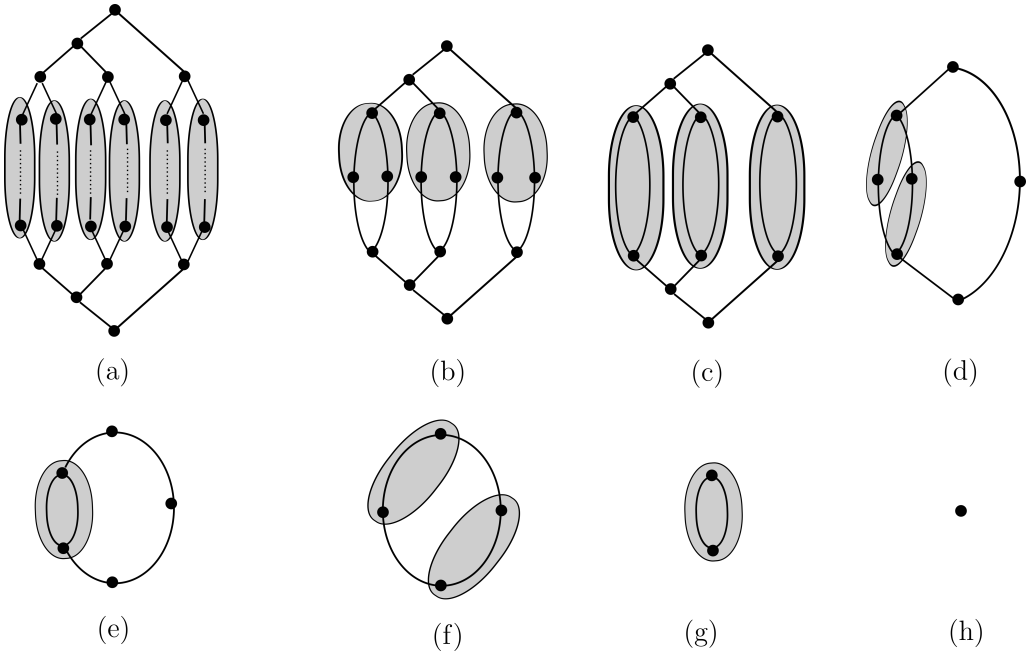
5.4 Skalární součin dvou tenzorů

V kapitole 3.3.3 jsme poukázali na možnosti tenzorových sítí pro zápis méně obvyklých tenzorových součinů, jejichž příklady jsme ilustrovali na obrázku 3.4. Zřejmě takto můžeme interpretovat i skalární součin tenzorů (uložených ve tvaru HTD).

Mějme tenzory $\mathcal{T}, \mathcal{S} \in \mathbb{R}^{n_1 \times \cdots \times n_k}$. Potom pro jejich skalární součin platí

$$\langle \mathcal{T}, \mathcal{S} \rangle = \langle \operatorname{vec}(\mathcal{T}), \operatorname{vec}(\mathcal{S}) \rangle = \sum_{i_1=1}^{n_1} \cdots \sum_{i_k=1}^{n_k} t_{i_1,\ldots,i_k} \cdot s_{i_1,\ldots,i_k}. \tag{5.18}$$

Jsou-li tenzory $\mathcal{T}$ a $\mathcal{S}$ uloženy ve tvaru sítě HTD (se stejnou strukturou stromu), lze jejich skalární součin interpretovat, resp. spočítat způsobem, jaký je vidět např. na obrázku 5.7.



Obrázek 5.7: *Příklad* výpočtu skalárního součinu dvou tenzorů $\mathcal{T}$ a $\mathcal{S}$ šestého řádu. Násobení v příslušných módech (viz kap. 3.2, speciálně obrázek 3.3) v jednotlivých krocích je naznačeno šedým podkladem. V jednotlivých krocích postupně počítáme/získáváme: (a) součin matic $U_{\mathcal{T}}^T U_{\mathcal{S}}$ (šestkrát); (b) součiny tenzoru s maticí $\mathcal{B}_{\mathcal{T}} \times_1 M_1 \times_2 M_2$ (třikrát); (c) součin tenzorů $\mathcal{B}_{\mathcal{T}} \times_{(1,2),(1,2)} \mathcal{B}_{\mathcal{S}}$ (třikrát); (d) součin tenzoru s maticí $\mathcal{B} \times_1 M$ (dvakrát); (e) součin tenzorů $\mathcal{B}_{\mathcal{T}} \times_{(1,2),(1,2)} \mathcal{B}_{\mathcal{S}}$ (jednou); (f) součin matic $M_{\mathcal{T}}^T M_{\mathcal{S}}$ (dvakrát); (g) skalární součin matic $\langle M_{\mathcal{T}}, M_{\mathcal{S}} \rangle$ (viz kap. 3.3.2; jednou); (h) výsledný skalár, tj. hledaný skalární součin tenzorů.

5.5 Výpočetní náročnost operací

Na závěr této kapitoly shrneme náročnost operací s tenzory uloženými v hierarchickém Tuckerově rozkladu. Zkoumat budeme přitom druh a počet operací, které musíme provést, pokud chceme i výsledný tenzor získat v podobě hierarchického Tuckerova rozkladu. Připoměňme značení (4.16)

$$r = \max_{\mathscr{C} \subseteq \{1,\dots,k\}} \operatorname{rank}(\mathcal{T}^{\mathscr{C}}) \qquad \text{a} \qquad n = \max\{n_1, \dots, n_k\}.$$

Detailněji rozebereme jen výpočet součinu tenzoru s maticí v jednom módu a součet dvou tenzorů ve tvaru HTD se stejným binárním stromem. Další operace jsou zmíněny v tabulce 5.1.

Je dobré si uvědomit, že vyvážený binární strom tenzoru řádu $k = 2^\varsigma$, $\varsigma \in \mathbb{N}$, obsahuje $\varsigma + 1 = (\log_2 k) + 1$ „pater“, přičemž nejspodnější (listy) a nejvrchnější (kořen) jsou matice. Obecně (řád tenzoru nemusí být mocninou dvou), budeme-li uvažovat „nejméně nevyvážený“ binární strom, HTD tenzoru obsahuje nejvýše

$$\lceil \log_2 k \rceil - 1$$

„pater“ tenzorů třetího řádu (viz např. obrázek 4.2), kde $\lceil \cdot \rceil$ značí horní celou část reálného čísla. Celkem bude takový strom obsahovat právě $2k-1$ objektů, přičemž k z nich jsou listy (matice), jeden je kořen (matice) a tedy

$$k-2$$

z nich jsou tenzory třetího řádu.

5.5.1 Náročnost součinu tenzoru s maticí

Pro součin tenzoru k-tého řádu s maticí $M \in \mathbb{R}^{m \times n_\ell}$, tj. $\mathcal{T} \times_\ell M$, ve tvaru HTD musíme provádět následující kroky:

- *jeden* součin matic s rozměry $m \times n$ a $n \times r$ (v ℓ-tém listu);
- *jeden* QR rozklad matice s rozměry $m \times r$ (v ℓ-tém listu);
- $(\lceil \log_2 k \rceil - 1)$-*krát* součin tenzoru třetího řádu s rozměry $r \times r \times r$ a matice s rozměry $r \times r$ (v prvním, nebo druhém módu);
- $(\lceil \log_2 k \rceil - 1)$-*krát* QR rozklad matice (resp. rozvoje tenzoru třetího řádu) s rozměry $r^2 \times r$;
- *jeden* součin tří matic se stejnými rozměry $r \times r$, přičemž prostřední matice je diagonální (aktulizace kořene);
- *jeden* SVD rozklad matice s rozměry $r \times r$ (aktualizace kořene); a
- *dvakrát* součin tenzoru s rozměry $r \times r \times r$ a matice s rozměry $r \times r$ (ve třetím módu; poslední krok aktualizace kořene; viz 5.16).

Pro celkový přehled viz také tabulku 5.1.

5.5.2 Náročnost součtu dvou tenzorů

Pro součet dvou tenzorů ve tvaru HTD, musíme pro získání výsledného tenzoru opět ve tvaru HTD provést tyto kroky:

- zřetězit odpovídající listy, tenzory třetího řádu a kořeny, což je provedeno s nulovým počtem aritmetických operací;
- k-*krát* QR rozklad matice s rozměry $n \times 2r$ (ve všech listech);
- $(k-2)$-*krát* součin tenzoru třetího řádu s rozměry $2r \times 2r \times 2r$ a dvou matic s rozměry $r \times 2r$ (v prvním a druhém módu);
- $(k-2)$-*krát* QR rozklad matice (resp. rozvoje tenzoru třetího řádu) s rozměry $r^2 \times 2r$;
- *jeden* součin tří matic s rozměry $r \times 2r$, $2r \times 2r$ a $2r \times r$, přičemž prostřední matice je diagonální (aktulizace kořene);
- *jeden* SVD rozklad matice s rozměry $r \times r$ (aktualizace kořene); a

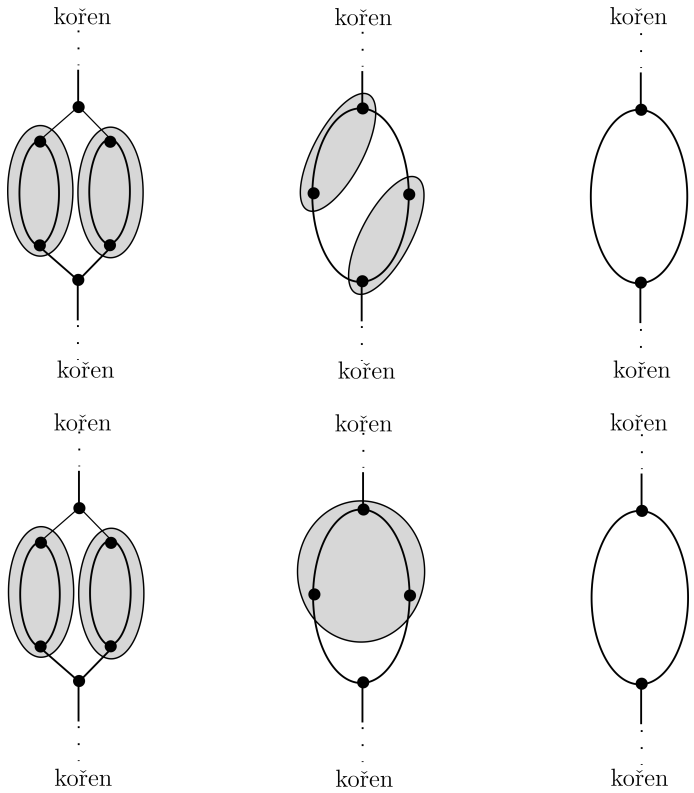
- *dvakrát* součin tenzoru s rozměry $r \times r \times r$ a matice s rozměry $r \times r$ (ve třetím módu; poslední krok aktualizace kořene; viz 5.16).

Pro celkový přehled viz také tabulku 5.1.

5.5.3 Náročnost výpočtu skalárního součinu

Skalární součin dvou tenzorů k-tého řádu ve tvaru HTD se stejnou strukturou stromu, provádíme postupně v krocích, které jsou ilustrovány na obrázku 5.7. Ze schématu na obrázku je zřejmé, že po vynásobení odpovídajících listů tenzorů postupujememe dále směrem ke kořenům obou stromů, kdy postupně násobením tenzorů s maticemi *eliminujeme* mezilehlé uzly tenzorové sítě. Tuto *eliminaci* můžeme provést dvěma způsoby, tak jak je ukázáno na obrázku 5.8. Pro výpočet skalárního součinu tedy provádíme:

- k-*krát* součin matic $r \times n$ a $n \times r$ – listů binárních stromů;
- *eliminací* provádíme tolik, kolik je v síti součinu odpovídajících si párů tenzorů třetího řádu, tedy $(k-2)$-*krát*. Přitom eliminace obsahuje (postupujeme-li



Obrázek 5.8: Dva způsoby eliminace mezilehlých uzlů sítě při výpočtu skalárního součinu. Výpočetní náročnosti jsou ekvivalentní, neboť součin tenzoru se dvěma maticemi (v různých módech; viz druhý řádek uprostřed) lze nahradit dvěma součiny tenzoru stejného řádu s maticí, jelikož platí $\mathcal{T} \times_\ell M_\ell \times_t M_t = (\mathcal{T} \times_\ell M_\ell) \times_t M_t$; viz např. [26, kap. 2.2].

podle prvního řádku obrázku 5.8; postup podle druhého řádku je okomentován v popisku obrázku):

- *dvakrát* součin dvou tenzorů s rozměry $r \times r \times r$ ve dvou módech (v prvním a druhém);
- *dvakrát* součin tenzoru s rozměry $r \times r \times r$ a matice s rozměry $r \times r$ (v prvním, nebo druhém módu).

Nakonec skalární součin ještě vyžaduje provést (viz obrázek 5.7(f–h)):

- *dvakrát* součin matic s rozměry $r \times r$ (po *eliminaci* tenzorů třetího řádu); a
- *jeden* skalární součin matic s rozměry $r \times r$.

Pro celkový přehled viz také tabulku 5.1.

Z tabulky 5.1 vidíme, že všechny operace jsou lineární (nikoliv exponenciální) v k – tj. v řádu původního tenzoru – a kvartické v r – veličině související s hodností tenzorů. Případná úspora operací (a tedy i výpočetního času) samozřejmě závisí na tom, jak malé reálně může být r pro daná data.

Na závěr poznamenejme, že operacemi (součet a skalární součin) jsme se zabývali jen v případě, že dvojice tenzorů v HTD, která vstupovala do operace, měla stejnou strukturu stromu včetně indexů. Pokud bychom chtěli provést některou z těchto operací na dvojici tenzorů s různým stromem, museli bychom nejprve jeden z tenzorů přepočítat – tedy adaptovat tak, aby struktura obou byla před operací shodná.

Podobně u lineárního zobrazení jsme potřebovali, aby mělo kroneckerovskou strukturu odpovídající rozměrům tenzoru. Tento požadavek je však přirozený, i když zobrazení může být i komplikovanější, než jak je naznačeno v (5.3); může se jednat o součet několika takových Kroneckerových součinů.

Tabulka 5.1: Porovnání výpočetní složitosti jednotlivých operací: MMp značí součin dvou matic (z anglického matrix-matrix product). TMp součin tenzoru třetího řádu a matice (z anglického tensor-matrix product); pozn. že součin tenzoru se dvěma maticemi v různých módech lze přepsat jako dva součiny tenzoru a matice. Výpočetní složitost QR rozkladu uvažujeme r^3 pro matici $r \times r$ a r^4 pro matici $r^2 \times r$; složitost SVD rozkladu uvažujeme r^3 pro matici $r \times r$; viz např. [4, kap. 3.5.6] a [6, kap. 8.6.4]. Složitost jedné sub-operace *eliminace* u skalárního součinu tenzorů je $4r^4$. Při odvozování složitostí jsme uvažovali nejméně nevyvážený strom, s výjimkou prvního řádku však budou odhady platit pro libovolný strom (např. i pro tensor train (TT)).

operace	počty elementárních sub-operací	složitost
součin tenzoru s maticí	3 MMp + $(\lceil \log_2 k \rceil + 1)$ TMp + $\lceil \log_2 k \rceil$ QR + SVD	$\sim 2\lceil \log_2 k \rceil r^4 + 5r^3$
aplikace lin. zobrazení (5.3)	$(k+2)$ MMp + $(2k-2)$ TMp + $(2k-2)$ QR + SVD	$\sim (9k+16)r^4 + (k+3)r^3$
součet dvou tenzorů	2 MMp + $(2k-2)$ TMp + $(2k-2)$ QR + SVD	$\sim (2k-2)r^4 + (2k+3)r^3$
skalární součin tenzorů	$k+2$ MMp + $(k-2)$ *eliminací* + skalární součin matic	$\sim (4k-8)r^4 + (k+2)r^3$

6 Náznak praktického výpočtu HTD

Jak jsme již konstatovali, hierarchický Tuckerův rozklad je vhodný zejména pro ukládání a manipulaci s tenzory vysokých řádů, protože jak paměťové nároky, tak operace s tenzorem jsou linární, nikoliv exponenciální, v řádu tenzoru. Obecně tenzory vysokých řádů nelze v prostém tvaru (ani v obyčejném Tuckerově rozkladu) vůbec v počítači uložit právě z důvodů enormní spotřeby paměti; viz obrázek 4.4. Je tedy zřejmé, že tenzor obecně nemůže do HTD dostat postupem, který by se nabízel z výkladu v kapitole 4. Naznačíme proto jednu z možností, kde se s tenzory v HTD tvaru můžeme setkat.

Uvažujme rovnici

$$\mathcal{A}(\mathcal{X}) = \mathcal{B}, \qquad \text{kde} \qquad \mathcal{X},\ \mathcal{B} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k} \tag{6.1}$$

jsou tenzor neznámých a tenzor pravých stran a kde zobrazení

$$\mathcal{A} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k} \longrightarrow \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_k} \tag{6.2}$$

je invertibilní. Rovnici tedy lze přepsat jako soustavu rovnic

$$A \operatorname{vec}(\mathcal{X}) = \operatorname{vec}(\mathcal{B}) \tag{6.3}$$

s regulární maticí A řádu $N \equiv n_1 \cdot n_2 \cdot \cdots \cdot n_k$.

Už samotný fakt, že nám je rovnice dána k řešení, znamená, že oba známé objekty A a $\mathcal{B}$ musí být nějak úsporně uloženy. Často má právě matice soustavy tvar Kroneckerova součinu (viz 5.3), nebo součtu několika málo (např. L) Kroneckerových součinů, tj.

$$A = \sum_{\ell=1}^{L} A_{\ell,k} \otimes A_{\ell,k-1} \otimes \cdots \otimes A_{\ell,1}, \qquad \text{kde} \qquad A_{\ell,j} \in \mathbb{R}^{n_j \times n_j}, \tag{6.4}$$

a tenzor pravých strany $\mathcal{B}$ je např. nízké hodnosti; např. hodnosti jedna. Tedy, je vnějším součinem k vektorů $b_j \in \mathbb{R}^{n_j}$, $j = 1, \ldots, k$; viz např. [13].

Tenzor $\mathcal{B}$, který je vnějším součinem k vektorů snadno přepíšeme do strukury, která bude rámcově odpovídat HTD. Stačí si vzít jednotlivé vektory b_j jako listy, vytvořit binární strom s tenzory třetího řádu $(1) \in \mathbb{R}^{1\times1\times1}$ a s kořenem $(1) \in \mathbb{R}^{1\times1}$. Pokud chceme mít tenzor ve tvaru HTD i fakticky, stačí zortogonalizovat (tedy pouze znormalizovat) listy, tj. vzít normalizované vektory $b_j/\|b_j\|$, matice přechodu svázané s tenzory třetího řádu ortonormální sloupce měly, zbývá tedy jako kořen vzít matici $[\|b_1\| \cdot \|b_2\| \cdot \cdots \cdot \|b_k\|] \in \mathbb{R}^{1\times1}$.

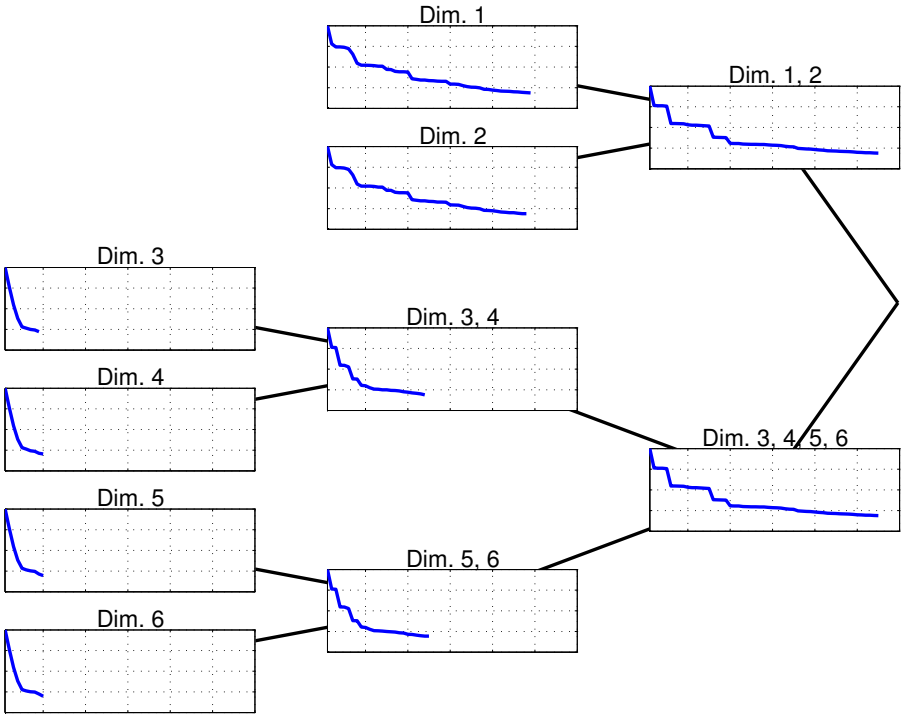
Nyní si stačí uvědomit, že pro řešení soustav rovnic existuje celá řada *iteračních metod* (viz např. [4, kapitoly 8 a 9], nebo [24], [25], [1], [8], a mnoho dalších) které zpravidla používají jen následující operace:

- násobení matice A vektorem,
- lineární kombinace dvou vektorů a
- skalární součin vektorů.

Všechny tyto operace ale umíme s tenzory v HTD tvaru provádět, přičemž výsledkem prvních dvou je opět tenzor v HTD tvaru. Zbývá tedy nějak zkonstruovat počáteční aproximaci tenzoru neznámých. Nejsnazší bude vzít počáteční odhad nulový, tj. $\mathcal{X}_0 = 0$. Takový počáteční odhad můžeme snadno zapsat do ve tvaru HTD se stejnou strukutrou, jako již máme zvolenou pro $\mathcal{B}$, protože $\mathcal{X}_0$ je fakticky také vnějším součinem k – tentokrát nulových – vektorů $[0, \ldots, 0]^T \in \mathbb{R}^{n_j}$, $j = 1, \ldots, k$ (tenzory třetího řádu i matici v kořenu stromu lze opět volit $(1) \in \mathbb{R}^{1\times1\times1}$, resp. $(1) \in \mathbb{R}^{1\times1}$; v případě nulového tenzoru mohou vznikout technické problémy s převodem do skutečného HTD protože zde budou figurovat nulové hodnosti, tedy např. matice beze sloupců, atd.).

V principu jsme tedy schopni nalézt tenzor řešení $\mathcal{X}$ rovnice (6.1) ve tvaru HTD. Pro jeho úspěšné nalezení zbývá jen doufat (nebo dokázat), že všechny hodnosti r všech tenzorů – meziproduktů iteračního algoritmu, ale také hledaného řešení (viz např. [13, kap. 3.2, str. 676–678]), budou rozumně malé.

Na závěr ještě poznamenejme, že mezi různými programy a toolboxy vyvinutými pro práci s tenzory, viz přehled v [11] nebo [26, příloha A], existuje software přímo navržený pro práci s tenzory v HTD. Na obrázku 6.1 vidíme tenzor šestého řádu v HTD v `htucker` toolboxu, viz [14], který jsme převzali z [12].



Obrázek 6.1: Podoba HTD tenzoru šestého řádu v `htucker` toolboxu (viz [14]) v MatLabu®, převzato z [12]. Modré čáry vyznačují singlulární čísla odpovídajících rozvojů tenzoru.

Závěr

Tato práce má sloužit jako studijní text uvádějící čtenáře, který se již setkal se základními pojmy multilineární algebry, do problematiky reprezentace tenzorů pomocí tenzorových sítí. V textu jsme zopakovali některé důležité pojmy týkající se tenzorů, se kterými pracujeme jako s vícerozměrnými poli čísel. Dále jsme definovali některé pojmy z teorie grafů a zavedli tzv. multigraf s volně visícími hranami a smyčkami, který jsme později použili pro reprezentaci tenzorové sítě. Tenzor pomocí grafů znázorňujeme jako vrchol s volně visícími hranami, jejichž počet odpovídá řádu tenzoru. Klasické hrany potom v tenzorové sítí představují násobení tenzorů v odpovídajících módech.

Hlavním cílem práce bylo zavedení hierarchického Tuckerova rozkladu tenzoru (HTD), objasnění principu na kterém je založen a tím i důkazu jeho existence. Grafickou reprezentací tohoto rozkladu je tenzorová síť v podobě (co nejméně nevyváženého) binárního stromu. V textu jsme ale poukázali na možnosti libovolné volby binárního stromu – např. tzv. tensor train je konstruován principiálně stejným způsobem jen s jiným tvarem binárního stromu. Výhodou hierarchického Tuckerova rozkladu je především paměťová úspora při ukládání tenzorů, za předpokladu, že hodnosti rozvojů tenzoru do matice odpovídající větvení binárního stromu jsou malé. V textu jsme proto nabídli porovnání náročnosti na paměť počítače, ukládáme-li tenzor různými způsoby.

Věnovali jsme se také principu a náročnosti některých operací s tenzory uloženými v hierarchickém Tuckerově rozkladu, přičemž jsme vždy chtěli i výsledný tenzor získat ve tvaru HTD. Na závěr jsme nabídli i náznak praktického výpočtu HTD v jedné konkrétní situaci.

Literatura

[1] O. Axelsson, *Iterative solution methods*, Cambridge University Press, Cambridge, 1994.

[2] B. W. Bader and T. G. Kolda, *Algorithm 862: Matlab tensor classes for fast algorithm prototyping*, ACM Transactions on Mathematical Software, 32 (2006), pp. 635–653.

[3] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A multilinear singular value decomposition*, SIAM Journal on Matrix Analysis and Applications, 21 (2000), pp. 1253–1278.

[4] E. J. Duintjer Tebbens, I. Hnětynková, M. Plešinger, Z. Strakoš, and P. Tichý, *Analýza metod pro maticové výpočty: základní metody*, Matfyzpress, Praha, 2012.

[5] M. Fiedler, *Speciální matice a jejich použití v numerické matematice*, TKI, Teoretická knižnice inženýra, SNTL, Státní nakladatelství technické literatury, Praha, 1981.

[6] G. H. Golub and C. F. Van Loan, *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 4th ed., 2013.

[7] L. Grasedyck, *Hierarchical singular value decomposition of tensors*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 2029–2054.

[8] A. Greenbaum, *Iterative methods for solving linear systems*, vol. 17 of Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[9] B. N. Khoromskij, *Tensors-structured numerical methods in scientific computing: Survey on recent advances*, Chemometrics and Intelligent Laboratory Systems, 110 (2012), pp. 1–19.

[10] H. A. L. Kiers, *Towards a standardized notation and terminology in multiway analysis*, Journal of Chemometrics, 14 (2000), pp. 105–122.

[11] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Review, 51 (2009), pp. 455–500.

[12] D. Kressner, M. Plešinger, and C. Tobler, *A preconditioned low-rank cg method for parameter-dependent lyapunov matrix equations*, tech. rep., EPFL MATHICSE Technical Report Nr. 18.2012, 2012. Preprint and extended version of [13]. Available at: http://mathicse.epfl.ch/files/contant/sites/mathicse/files/Mathicse reports 2012/18.2012_DK-MP-CT.pdf.

[13] ——, *A preconditioned low-rank CG method for parameter-dependent Lyapunov matrix equations*, Numerical Linear Algebra with Applications, 21 (2014), pp. 666–684.

[14] D. Kressner and C. Tobler, *htucker — a Matlab toolbox for tensors in hierarchical Tucker format*, Tech. Rep. 2012-02, ETH Zürich, Zürich, 2012. Available at: http://anchp.epfl.ch/htucker.

[15] K. Kuchař, *Základy obecné teorie relativity*, Academia, Praha, 1968.

[16] J. Matoušek and J. Nešetřil, *Kapitoly z diskrétní matematiky*, Karolinum, Praha, 2009.

[17] I. V. Oseledets, *Tensor-train decomposition*, SIAM Journal on Scientific Computing, 33 (2011), pp. 2295–2317.

[18] L. Sorber, M. Van Barel, and L. De Lathauwer, *Tentensor v2.0*, 2014. Available at: http://www.tensorlab.net.

[19] C. Tobler, *Low-rank tensor methods for linear systems and eigenvalue problems*, PhD thesis, ETH Zürich, Zürich, 2012. Available at: http://sma.epfl.ch/~ctobler/diss.pdf.

[20] L. R. Tucker, *Implications of factor analysis of three-way matrices for measurement of change*, in Problems in Measuring Change, C. W. Harris, ed., University of Wisconsin Press, 1963, pp. 122–137.

[21] ——, *The extension of factor analysis to three-dimensional matrices*, in Contributions to Mathematical Psychology, H. Gulliksen and N. Frederiksen, eds., New York, 1964, Holt, Rinehardt & Winston, pp. 110–127.

[22] ——, *Some mathematical notes on three-mode factor analysis*, Psychometrika, 31 (1966), pp. 279–311.

[23] D. A. Turkington, *Gerenalized vectorization, cross-products, and matrix calculus*, Cambridge University Press, Cambridge, 2013.

[24] R. S. Varga, *Matrix iterative analysis*, Prentice-Hall Inc., Englewood Cliffs, N.J., 1962.

[25] D. M. Young, *Iterative solution of large linear systems*, Academic Press, New York, 1971.

[26] J. Žáková, *Tenzory a kanonické tenzorové rozklady: Tuckerův rozklad*, Bachelor thesis, TU v Liberci, Liberec, 2015.