Chapter 3

Constant-Time Algorithms for Continuous Optimization Problems

Yuichi Yoshida

Abstract In this chapter, we consider constant-time algorithms for continuous optimization problems. Specifically, we consider quadratic function minimization and tensor decomposition, both of which have numerous applications in machine learning and data mining. The key component in our analysis is *graph limit theory*, which was originally developed to study graphs analytically.

3.1 Introduction

In this chapter, we turn our attention to constant-time algorithms for continuous optimization problems. Specifically, we consider quadratic function minimization and tensor decomposition, both of which have numerous applications in machine learning and data mining. The key component in our analysis is *graph limit theory*, which was originally developed to study graphs analytically.

We introduce graph limit theory in Sect. 3.2, and then discuss quadratic function minimization and tensor decomposition in Sects. 3.3 and 3.4, respectively. Throughout this chapter, we assume the real RAM model, in which we can perform basic algebraic operations on real numbers in one step. For a positive integer n , let $[n]$ denote the set $\{1, 2, \dots, n\}$. For real values $a, b, c \in \mathbb{R}$, $a = b \pm c$ is used as shorthand for $b - c \leq a \leq b + c$. The algorithms and analysis presented in this chapter are based on [5, 6].

3.2 Graph Limit Theory

This section reviews the basic concepts of graph limit theory. For further details, refer to the book by Lovász [7].

Y. Yoshida (✉)

National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
e-mail: yyoshida@nii.ac.jp

We call a (measurable) function $\mathcal{W} : [0, 1]^K \rightarrow \mathbb{R}$ a *dikernel of order K* . We define

$$|\mathcal{W}|_F = \sqrt{\int_{[0,1]^K} \mathcal{W}(\mathbf{x})^2 d\mathbf{x}}, \quad (\text{Frobenius norm})$$

$$|\mathcal{W}|_{\max} = \max_{\mathbf{x} \in [0,1]^K} |\mathcal{W}(\mathbf{x})|, \quad (\text{Max norm})$$

$$|\mathcal{W}|_{\square} = \sup_{S_1, \dots, S_K \subseteq [0,1]} \left| \int_{S_1 \times \dots \times S_K} \mathcal{W}(\mathbf{x}) d\mathbf{x} \right| \quad (\text{Cut norm})$$

We note that these norms satisfy the triangle inequality. For two dikernels $\mathcal{W}$ and $\mathcal{W}'$, we define their *inner product* as $\langle \mathcal{W}, \mathcal{W}' \rangle = \int_{[0,1]^K} \mathcal{W}(\mathbf{x}) \mathcal{W}'(\mathbf{x}) d\mathbf{x}$. For a dikernel $\mathcal{W} : [0, 1]^2 \rightarrow \mathbb{R}$ and a function $f : [0, 1] \rightarrow \mathbb{R}$, we define a function $\mathcal{W}f : [0, 1] \rightarrow \mathbb{R}$ as $(\mathcal{W}f)(x) = \langle \mathcal{W}(x, \cdot), f \rangle$.

Let λ be a Lebesgue measure. A map $\pi : [0, 1] \rightarrow [0, 1]$ is said to be *measure-preserving* if the pre-image $\pi^{-1}(X)$ is measurable for every measurable set X , and $\lambda(\pi^{-1}(X)) = \lambda(X)$. A *measure-preserving bijection* is a measure-preserving map whose inverse map exists and is also measurable (and, in turn, also measure-preserving). For a measure-preserving bijection $\pi : [0, 1] \rightarrow [0, 1]$ and a dikernel $\mathcal{W} : [0, 1]^K \rightarrow \mathbb{R}$, we define a dikernel $\pi(\mathcal{W}) : [0, 1]^K \rightarrow \mathbb{R}$ as $\pi(\mathcal{W})(x_1, \dots, x_K) = \mathcal{W}(\pi(x_1), \dots, \pi(x_K))$.

A partition $\mathcal{P} = (V_1, \dots, V_p)$ of the interval $[0, 1]$ is called an *equipartition* if $\lambda(V_i) = 1/p$ for every $i \in [p]$. For a dikernel $\mathcal{W} : [0, 1]^K \rightarrow \mathbb{R}$ and an equipartition $\mathcal{P} = (V_1, \dots, V_p)$ of $[0, 1]$, we define $\mathcal{W}_{\mathcal{P}} : [0, 1]^K \rightarrow \mathbb{R}$ as the dikernel obtained by averaging each $V_{i_1} \times \dots \times V_{i_K}$ for $i_1, \dots, i_K \in [p]$. More formally, we define

$$\mathcal{W}_{\mathcal{P}}(\mathbf{x}) = \frac{1}{\prod_{k \in [K]} \lambda(V_{i_k})} \int_{V_{i_1} \times \dots \times V_{i_K}} \mathcal{W}(\mathbf{x}') d\mathbf{x}' = p^K \int_{V_{i_1} \times \dots \times V_{i_K}} \mathcal{W}(\mathbf{x}') d\mathbf{x}',$$

where i_k is the unique index such that $x_k \in V_{i_k}$ for each $k \in [K]$. The following lemma states that any dikernel $\mathcal{W} : [0, 1]^K \rightarrow \mathbb{R}$ can be well approximated by $\mathcal{W}_{\mathcal{P}}$ for some equipartition $\mathcal{P}$ into a small number of parts.

Lemma 3.1 (Weak regularity lemma for dikernels [4]) *Let $\mathcal{W}^1, \dots, \mathcal{W}^T : [0, 1]^K \rightarrow \mathbb{R}$ be dikernels. Then, for any $\epsilon > 0$, there exists an equipartition $\mathcal{P}$ into $|\mathcal{P}| \leq 2^{O(T/\epsilon^{2K})}$ parts, such that for every $t \in [T]$,*

$$|\mathcal{W}^t - \mathcal{W}_{\mathcal{P}}^t|_{\square} \leq \epsilon |\mathcal{W}^t|_F.$$

We can construct the dikernel $\mathcal{X} : [0, 1]^K \rightarrow \mathbb{R}$ from a tensor $X \in \mathbb{R}^{N_1 \times \dots \times N_K}$ as follows. For an integer $n \in \mathbb{N}$, let $I_1^n = [0, \frac{1}{n}]$, $I_2^n = (\frac{1}{n}, \frac{2}{n}]$, $\dots$, $I_n^n = (\frac{n-1}{n}, 1]$. For $x \in [0, 1]$, we define $i_n(x) \in [n]$ as the unique integer such that $x \in I_{i_n}^n$. We then define $\mathcal{X}(x_1, \dots, x_K) = X_{i_{N_1}(x_1) \dots i_{N_K}(x_K)}$. The main motivation of creating a dikernel from a tensor is that, in doing so, we can define the distance between two tensors

X and Y of different sizes via the cut norm—that is, $|\mathcal{X} - \mathcal{Y}|_\square$, where $\mathcal{X}$ and $\mathcal{Y}$ are dikernels corresponding to X and Y , respectively.

Let $\mathcal{W} : [0, 1]^K \rightarrow \mathbb{R}$ be a dikernel and $S_k = (x_1^k, \dots, x_s^k)$ for $k \in [K]$ be sequences of elements in $[0, 1]$. Then, we define a dikernel $\mathcal{W}|_{S_1, \dots, S_K} : [0, 1]^K \rightarrow \mathbb{R}$ as follows: We first extract a tensor $W \in \mathbb{R}^{s \times \dots \times s}$ by setting $W_{i_1 \dots i_K} = \mathcal{W}(x_{i_1}^1, \dots, x_{i_K}^K)$. Next, we define $\mathcal{W}|_{S_1, \dots, S_K}$ as the dikernel corresponding to $W|_{S_1, \dots, S_K}$. The following is the key technical lemma in the analysis of the algorithms given in the subsequent sections.

Lemma 3.2 *Let $\mathcal{W}^1, \dots, \mathcal{W}^T : [0, 1]^K \rightarrow [-L, L]$ be dikernels. Let $S_1, \dots, S_K$ be sequences of s elements uniformly and independently sampled from $[0, 1]$. Then, with probability at least $1 - \exp(-\Omega_K(s^2(T/\log s)^{1/K}))$, there exists a measure-preserving bijection $\pi : [0, 1] \rightarrow [0, 1]$ such that, for every $t \in [T]$, we have*

$$|\mathcal{W}^t - \pi(\mathcal{W}^t|_{S_1, \dots, S_K})|_\square = L \cdot O_K \left(\frac{T}{\log s} \right)^{1/2K},$$

where $O_K(\cdot)$ and $\Omega_K(\cdot)$ hide factors depending on K .

3.3 Quadratic Function Minimization

Background

Quadratic functions are one of the most important function classes in machine learning, statistics, and data mining. Many fundamental problems such as linear regression, k -means clustering, principal component analysis, support vector machines, and kernel methods can be formulated as a minimization problem of a quadratic function. See, e.g., [8] for more details.

In some applications, it is sufficient to compute the minimum value of a quadratic function rather than its solution. For example, Yamada et al. [13] proposed an efficient method for estimating the Pearson divergence, which provides useful information about data, such as the density ratio [10]. They formulated the estimation problem as the minimization of a squared loss and showed that the Pearson divergence can be estimated from the minimum value. Least-squares mutual information [9] is another example that can be computed in a similar manner.

Despite its importance, minimization of quadratic functions suffers from the issue of scalability. Let $n \in \mathbb{N}$ be the number of variables. In general, this kind of minimization problem can be solved by quadratic programming (QP), which requires $\text{poly}(n)$ time. If the problem is convex and there are no constraints, then the problem is reduced to solving a system of linear equations, which requires $O(n^3)$ time. Both methods easily become infeasible, even for medium-scale problems of, say, $n > 10000$.

Although several techniques have been proposed to accelerate quadratic function minimization, they require at least linear time in n . This is problematic when handling

Algorithm 1

Input: $n \in \mathbb{N}$, query access to a matrix $A \in \mathbb{R}^{n \times n}$ and to vectors $\mathbf{d}, \mathbf{b} \in \mathbb{R}^n$, and $\epsilon, \delta \in (0, 1)$.
 1: $S \leftarrow$ a sequence of $s = s(\epsilon, \delta)$ indices independently and uniformly sampled from $[n]$.
 2: **return** $\frac{n^2}{s^2} \min_{\mathbf{v} \in \mathbb{R}^n} p_{s, A|_S, \mathbf{d}|_S, \mathbf{b}|_S}(\mathbf{v})$.

large-scale problems, where even linear time is slow or prohibitive. For example, stochastic gradient descent (SGD) is an optimization method that is widely used for large-scale problems. A nice property of this method is that, if the objective function is strongly convex, it outputs a point that is sufficiently close to an optimal solution after a constant number of iterations [1]. Nevertheless, each iteration needs at least $\Omega(n)$ time to access the variables. Another popular technique is low-rank approximation such as Nystrom’s method [12]. The underlying idea is to approximate the input matrix by a low-rank matrix, which drastically reduces the time complexity. However, we still need to compute the matrix vector product of size n , which requires $\Omega(n)$ time. Clarkson et al. [2] proposed sublinear-time algorithms for special cases of quadratic function minimization. However, these are “sublinear” with respect to the number of pairwise interactions of the variables, which is $\Theta(n^2)$, and the algorithms require $O(n \log^c n)$ time for some $c \geq 1$.

Constant-time algorithm for quadratic function minimization

Let $A \in \mathbb{R}^{n \times n}$ be a matrix and $\mathbf{d}, \mathbf{b} \in \mathbb{R}^n$ be vectors. Then, we consider the following quadratic problem:

$$\underset{\mathbf{v} \in \mathbb{R}^n}{\text{minimize}} \quad p_{n, A, \mathbf{d}, \mathbf{b}}(\mathbf{v}), \quad \text{where } p_{n, A, \mathbf{d}, \mathbf{b}}(\mathbf{v}) = \langle \mathbf{v}, A\mathbf{v} \rangle + n\langle \mathbf{v}, \text{diag}(\mathbf{d})\mathbf{v} \rangle + n\langle \mathbf{b}, \mathbf{v} \rangle, \quad (3.1)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and $\text{diag}(\mathbf{d})$ denotes a diagonal matrix in which the diagonal entries are specified by $\mathbf{d}$. Note that although a constant term can be included in (3.1), it is omitted here because it is irrelevant when optimizing (3.1), and hence we omit it.

Let $z^* \in \mathbb{R}$ be the optimal value of (3.1) and let $\epsilon, \delta \in (0, 1)$ be parameters. Then, our goal is then to compute z with $|z - z^*| = O(\epsilon n^2)$ with probability at least $1 - \delta$ in constant time. We further assume that we have query access to A , $\mathbf{b}$, and $\mathbf{d}$, with which we can obtain their entry by specifying an index. We note that z^* is typically $\Theta(n^2)$ because $\langle \mathbf{v}, A\mathbf{v} \rangle$ consists of $\Theta(n^2)$ terms, and $\langle \mathbf{v}, \text{diag}(\mathbf{d})\mathbf{v} \rangle$ and $\langle \mathbf{b}, \mathbf{v} \rangle$ consist of $\Theta(n)$ terms. Hence, we can regard the error of $\Theta(\epsilon n^2)$ as an error of $\Theta(\epsilon)$ for each term, which is reasonably small in typical situations.

Let $\cdot|_S$ be an operator that extracts a submatrix (or subvector) specified by an index set $S \subset \mathbb{N}$. Our algorithm is then given by Algorithm 1, where the parameter $s := s(\epsilon, \delta)$ is determined later. In other words, we sample a constant number of indices from the set $[n]$, and then solve the problem (3.1) restricted to these indices.

Note that the number of queries and the time complexity are $O(s^2)$ and $\text{poly}(s)$, respectively.

The goal of the rest of this section is to show the following approximation guarantee of Algorithm 1.

Theorem 3.1 *Let $\mathbf{v}^*$ and z^* be an optimal solution and the optimal value, respectively, of problem (3.1). By choosing $s(\epsilon, \delta) = 2^{\Theta(1/\epsilon^2)} + \Theta(\log \frac{1}{\delta} \log \log \frac{1}{\delta})$, with probability at least $1 - \delta$, a sequence S of s indices independently and uniformly sampled from $[n]$ satisfies the following: Let $\tilde{\mathbf{v}}^*$ and $\tilde{z}^*$ be an optimal solution and the optimal value, respectively, of the problem $\min_{\mathbf{v} \in \mathbb{R}^s} p_{s, A|_S, \mathbf{d}|_S, \mathbf{b}|_S}(\mathbf{v})$. Then, we have*

$$\left| \frac{n^2}{s^2} \tilde{z}^* - z^* \right| \leq \epsilon L M^2 n^2,$$

where

$$L = \max \left\{ \max_{i,j} |A_{ij}|, \max_i |d_i|, \max_i |b_i| \right\} \text{ and } M = \max \left\{ \max_{i \in [n]} |v_i^*|, \max_{i \in [n]} |\tilde{v}_i^*| \right\}.$$

We can show that M is bounded when A is symmetric and full rank. To see this, we first note that we can assume $A + n \text{diag}(\mathbf{d})$ is positive-definite, as otherwise $p_{n, A, \mathbf{d}, \mathbf{b}}$ is not bounded and the problem is uninteresting. Then, for any set $S \subseteq [n]$ of s indices, $(A + n \text{diag}(\mathbf{d}))|_S$ is again positive-definite because it is a principal submatrix. Hence, we have $\mathbf{v}^* = (A + n \text{diag}(\mathbf{d}))^{-1} n \mathbf{b} / 2$ and $\tilde{\mathbf{v}}^* = (A|_S + n \text{diag}(\mathbf{d}|_S))^{-1} n \mathbf{b}|_S / 2$, which means that M is bounded.

3.3.1 Proof of Theorem 3.1

To use dikernels in our analysis, we first introduce a continuous version of $p_{n, A, \mathbf{d}, \mathbf{b}}$. The real-valued function $P_{n, A, \mathbf{d}, \mathbf{b}}$ on the functions $f : [0, 1] \rightarrow \mathbb{R}$ is defined as

$$P_{n, A, \mathbf{d}, \mathbf{b}}(f) = \langle f, \mathcal{A}f \rangle + \langle f^2, \mathcal{D}1 \rangle + \langle f, \mathcal{B}1 \rangle,$$

where $\mathcal{D}$ and $\mathcal{B}$ are the dikernels corresponding to $\mathbf{d} \mathbf{1}^\top$ and $\mathbf{b} \mathbf{1}^\top$, respectively, $f^2 : [0, 1] \rightarrow \mathbb{R}$ is a function such that $f^2(x) = f(x)^2$ for every $x \in [0, 1]$ and $1 : [0, 1] \rightarrow \mathbb{R}$ is a constant function that has a value of 1 everywhere. The following lemma states that the minimizations of $p_{n, A, \mathbf{d}, \mathbf{b}}$ and $P_{n, A, \mathbf{d}, \mathbf{b}}$ are equivalent:

Lemma 3.3 *Let $A \in \mathbb{R}^{n \times n}$ be a matrix and $\mathbf{d}, \mathbf{b} \in \mathbb{R}^{n \times n}$ be vectors. Then, we have*

$$\min_{\mathbf{v} \in [-M, M]^n} p_{n, A, \mathbf{d}, \mathbf{b}}(\mathbf{v}) = n^2 \cdot \inf_{f : [0, 1] \rightarrow [-M, M]} P_{n, A, \mathbf{d}, \mathbf{b}}(f).$$

for any $M > 0$.

Proof First, we show that $n^2 \cdot \inf_{f:[0,1] \rightarrow [-M,M]} P_{n,A,d,b}(f) \leq \min_{\mathbf{v} \in [-M,M]^n} p_{n,A,d,b}(\mathbf{v})$. Given a vector $\mathbf{v} \in [-M,M]^n$, we define $f:[0,1] \rightarrow [-M,M]$ as $f(x) = v_{i_n(x)}$. Then,

$$\begin{aligned} \langle f, \mathcal{A}f \rangle &= \sum_{i,j \in [n]} \int_{I_i^n} \int_{I_j^n} A_{ij} f(x) f(y) dx dy = \frac{1}{n^2} \sum_{i,j \in [n]} A_{ij} v_i v_j = \frac{1}{n^2} \langle \mathbf{v}, \mathbf{A} \mathbf{v} \rangle, \\ \langle f^2, \mathcal{D}1 \rangle &= \sum_{i,j \in [n]} \int_{I_i^n} \int_{I_j^n} d_{ij} f(x)^2 dx dy = \sum_{i \in [n]} \int_{I_i^n} d_i f(x)^2 dx \\ &= \frac{1}{n} \sum_{i \in [n]} d_i v_i^2 = \frac{1}{n} \langle \mathbf{v}, \text{diag}(\mathbf{d}) \mathbf{v} \rangle, \\ \langle f, \mathcal{B}1 \rangle &= \sum_{i,j \in [n]} \int_{I_i^n} \int_{I_j^n} b_{ij} f(x) dx dy = \sum_{i \in [n]} \int_{I_i^n} b_i f(x) dx = \frac{1}{n} \sum_{i \in [n]} b_i v_i = \frac{1}{n} \langle \mathbf{v}, \mathbf{b} \rangle. \end{aligned}$$

Hence, we have $n^2 P_{n,A,d,b}(f) \leq p_{n,A,d,b}(\mathbf{v})$.

Next, we show that $\min_{\mathbf{v} \in [-M,M]^n} p_{n,A,d,b}(\mathbf{v}) \leq n^2 \cdot \inf_{f:[0,1] \rightarrow [-M,M]} P_{n,A,d,b}(f)$. Let $f:[0,1] \rightarrow [-M,M]$ be a measurable function. For $x \in [0,1]$, we then have

$$\begin{aligned} &\frac{\partial P_{n,A,d,b}(f(x))}{\partial f(x)} \\ &= \sum_{i \in [n]} \int_{I_i^n} A_{i i_n(x)} f(y) dy + \sum_{j \in [n]} \int_{I_j^n} A_{i_n(x) j} f(y) dy + 2d_{i_n(x)} f(x) + b_{i_n(x)}. \end{aligned}$$

Note that the form of this partial derivative depends on only $i_n(x)$. Hence, in the optimal solution $f^*:[0,1] \rightarrow [-M,M]$, we can assume $f^*(x) = f^*(y)$ if $i_n(x) = i_n(y)$. In other words, f^* is constant on each of the intervals $I_1^n, \dots, I_n^n$. For such f^* , we define the vector $\mathbf{v} \in \mathbb{R}^n$ as $v_i = f^*(x)$, where $x \in [0,1]$ is any element in I_i^n . Then, we have

$$\begin{aligned} \langle \mathbf{v}, \mathbf{A} \mathbf{v} \rangle &= \sum_{i,j \in [n]} A_{ij} v_i v_j = n^2 \sum_{i,j \in [n]} \int_{I_i^n} \int_{I_j^n} A_{ij} f^*(x) f^*(y) dx dy = n^2 \langle f^*, \mathcal{A} f^* \rangle, \\ \langle \mathbf{v}, \text{diag}(\mathbf{d}) \mathbf{v} \rangle &= \sum_{i \in [n]} d_i v_i^2 = n \sum_{i \in [n]} \int_{I_i^n} d_i f^*(x)^2 dx = n \langle (f^*)^2, \mathcal{D}1 \rangle, \\ \langle \mathbf{v}, \mathbf{b} \rangle &= \sum_{i \in [n]} b_i v_i = n \sum_{i \in [n]} \int_{I_i^n} b_i f^*(x) dx = n \langle f^*, \mathcal{B}1 \rangle. \end{aligned}$$

Hence, we have $p_{n,A,d,b}(\mathbf{v}) \leq n^2 P_{n,A,d,b}(f^*)$.

Proof (of Theorem 3.1) We instantiate Lemma 3.2 with $s = 2^{\Theta(1/\epsilon^2)} + \Theta(\log \frac{1}{\delta} \log \log \frac{1}{\delta})$ and the dikernels $\mathcal{A}$, $\mathcal{D}$, and $\mathcal{B}$. Then, with probability at least $1 - \delta$, there exists a measure-preserving bijection $\pi:[0,1] \rightarrow [0,1]$ such that

$$\max \left\{ |\langle f, (\mathcal{A} - \pi(\mathcal{A}|_S)) f \rangle|, |\langle f^2, (\mathcal{D} - \pi(\mathcal{D}|_S)) 1 \rangle|, |\langle f, (\mathcal{B} - \pi(\mathcal{B}|_S)) 1 \rangle| \right\} \leq \frac{\epsilon L M^2}{3}$$

for any function $f : [0, 1] \rightarrow [-M, M]$. Conditioned on this event, we have

$$\begin{aligned}
\tilde{z}^* &= \min_{\mathbf{v} \in \mathbb{R}^s} p_{s, A|_S, d|_S, b|_S}(\mathbf{v}) = \min_{\mathbf{v} \in [-M, M]^s} p_{s, A|_S, d|_S, b|_S}(\mathbf{v}) \\
&= s^2 \cdot \inf_{f: [0, 1] \rightarrow [-M, M]} P_{s, A|_S, d|_S, b|_S}(f) \quad (\text{By Lemma 3}) \\
&= s^2 \cdot \inf_{f: [0, 1] \rightarrow [-M, M]} \left(\langle f, (\pi(\mathcal{A}|_S) - \mathcal{A})f \rangle + \langle f, \mathcal{A}f \rangle + \langle f^2, (\pi(\mathcal{D}|_S) - \mathcal{D})1 \rangle + \right. \\
&\quad \left. \langle f^2, \mathcal{D}1 \rangle + \langle f, (\pi(\mathcal{B}|_S) - \mathcal{B})1 \rangle + \langle f, \mathcal{B}1 \rangle \right) \\
&\leq s^2 \cdot \inf_{f: [0, 1] \rightarrow [-M, M]} \left(\langle f, \mathcal{A}f \rangle + \langle f^2, \mathcal{D}1 \rangle + \langle f, \mathcal{B}1 \rangle \pm \epsilon L M^2 \right) \\
&= \frac{s^2}{n^2} \cdot \min_{\mathbf{v} \in [-M, M]^n} p_{n, A, d, b}(\mathbf{v}) \pm \epsilon L M^2 s^2. \quad (\text{By Lemma 3}) \\
&= \frac{s^2}{n^2} \cdot \min_{\mathbf{v} \in \mathbb{R}^n} p_{n, A, d, b}(\mathbf{v}) \pm \epsilon L M^2 s^2 = \frac{s^2}{n^2} z^* \pm \epsilon L M^2 s^2.
\end{aligned}$$

Rearranging the inequality, we obtain the desired result.

3.4 Tensor Decomposition

Background

We say that a tensor (or a multidimensional array) is of *order* K if it is a K -dimensional array. Each dimension is called a *mode* in tensor terminology. Tensor decomposition, which approximates the input tensor by a number of smaller tensors, is a fundamental tool for dealing with large tensors because it drastically reduces memory usage.

Among the many existing tensor decomposition methods, *Tucker decomposition* [11] is a popular choice. To some extent, Tucker decomposition is analogous to singular-value decomposition (SVD). Whereas SVD decomposes a matrix into left and right singular vectors that interact via singular values, Tucker decomposition of an order- K tensor consists of K factor matrices that interact via the so-called core tensor. The key difference between SVD and Tucker decomposition is that, in the latter, the core tensor does not need to be diagonal and its “rank” can differ for each mode. We refer to the size of the core tensor, which is a K -tuple, as the *Tucker rank* of a Tucker decomposition.

We are usually interested in obtaining factor matrices and a core tensor to minimize the *residual error*—the error between the input and low-rank approximated tensors. Sometimes, however, knowing the residual error itself is a task of interest. The residual error tells us how suitable a low-rank approximation is to approximate the input tensor in the first place, and is also useful to predetermine the Tucker rank. In real applications, Tucker ranks are not explicitly given, and we must select them by considering the tradeoff between space usage and approximation accuracy. For

example, if the selected Tucker rank is too small, we risk losing essential information in the input tensor, whereas if the selected Tucker rank is too large, the computational cost of computing the Tucker decomposition (even if we allow for approximation methods) increases considerably along with space usage. As with the case of the matrix rank, one might think that a reasonably good Tucker rank can be found using a grid search. Unfortunately, grid search for an appropriate Tucker rank is challenging because, for an order- K tensor, the Tucker rank consists of K free parameters and the search space grows exponentially in K . Hence, we want to evaluate each grid point as quickly as possible.

Although several practical algorithms have been proposed, such as the higher order orthogonal iteration (HOOI) [3], they are not sufficiently scalable. For each mode, HOOI iteratively applies SVD to an unfolded tensor—a matrix that is reshaped from the input tensor. Given an $N_1 \times \cdots \times N_K$ tensor, the computational cost is hence $O(K \max_k N_k \cdot \prod_k N_k)$, which crucially depends on the input size $N_1, \dots, N_K$. Although there are several approximation algorithms, their computational costs are still intensive.

Constant-time algorithm for the Tucker fitting problem

The problem of computing the residual error is formalized as the following *Tucker fitting* problem: Given an order- K tensor $X \in \mathbb{R}^{N_1 \times \cdots \times N_K}$ and integers $R_k \leq N_k$ ($k = 1, \dots, K$), we want to compute the following normalized residual error:

$$\ell_{R_1, \dots, R_K}(X) := \min_{G \in \mathbb{R}^{R_1 \times \cdots \times R_K}, \{U^{(k)} \in \mathbb{R}^{N_k \times R_k}\}_{k \in [K]}} \frac{\|X - \llbracket G; U^{(1)}, \dots, U^{(K)} \rrbracket\|_F^2}{\prod_{k \in [K]} N_k}, \quad (3.2)$$

where $\llbracket G; U^{(1)}, \dots, U^{(K)} \rrbracket \in \mathbb{R}^{N_1 \times \cdots \times N_K}$ is an order- K tensor, defined as

$$\llbracket G; U^{(1)}, \dots, U^{(K)} \rrbracket_{i_1 \cdots i_K} = \sum_{r_1 \in [R_1], \dots, r_K \in [R_K]} G_{r_1 \cdots r_K} \prod_{k \in [K]} U_{i_k r_k}^{(k)}$$

for every $i_1 \in [N_1], \dots, i_K \in [N_K]$. Here, G is the core tensor, and $U^{(1)}, \dots, U^{(K)}$ are the factor matrices. Note that we are not concerned with computing the minimizer, but only want to compute the minimum value. In addition, we do not need the exact minimum. Indeed, a rough estimate still helps to narrow down promising rank candidates. The question here is how quickly we can compute the normalized residual error $\ell_{R_1, \dots, R_K}(X)$ with moderate accuracy.

In this section, we consider the following simple sampling algorithm, and show that it can be used to approximately solve the Tucker fitting problem. First, given an order- K tensor $X \in \mathbb{R}^{N_1 \times \cdots \times N_K}$, Tucker rank $(R_1, \dots, R_K)$, and sample size $s \in \mathbb{N}$, we sample a sequence of indices $S_k = (x_1^k, \dots, x_s^k)$ uniformly and independently from $[N_k]$ for each mode $k \in [K]$. We then construct a mini-tensor $X|_{S_1, \dots, S_K} \in \mathbb{R}^{s \times \cdots \times s}$, where $(X|_{S_1, \dots, S_K})_{i_1, \dots, i_K} = X_{x_{i_1}^1, \dots, x_{i_K}^K}$. Finally, we compute $\ell_{R_1, \dots, R_K}(X|_{S_1, \dots, S_K})$ using an arbitrary solver, such as HOOI, and output the obtained value. The details are provided in Algorithm 2. Note that the time complex-

Algorithm 2 Sampling algorithm for the Tucker fitting problem

Input: $N_1, \dots, N_K \in \mathbb{N}$, query access to a tensor $X \in \mathbb{R}^{N_1 \times \dots \times N_K}$, Tucker rank $(R_1, \dots, R_K)$, and $\epsilon, \delta \in (0, 1)$.

- 1: **for** $k = 1$ to K **do**
- 2: $S_k \leftarrow$ a sequence of $s = s(\epsilon, \delta)$ indices uniformly and independently sampled from $[N_k]$.
- 3: Construct a mini-tensor $X|_{S_1, \dots, S_K}$.
- 4: **return** $\ell_{R_1, \dots, R_K}(X|_{S_1, \dots, S_K})$.

ity for computing $\ell_{R_1, \dots, R_K}(X|_{S_1, \dots, S_K})$ does not depend on the input size $N_1, \dots, N_K$ but rather on the sample size s , meaning that the algorithm runs in *constant time*, regardless of the input size.

The goal of the rest of this section is to show the following approximation guarantee of Algorithm 2.

Theorem 3.2 *Let $X \in \mathbb{R}^{N_1 \times \dots \times N_K}$ be a tensor, $R_1, \dots, R_K$ be integers, and $\epsilon, \delta \in (0, 1)$. For $s(\epsilon, \delta) = 2^{\Theta(1/\epsilon^{2K-2})} + \Theta(\log \frac{1}{\delta} \log \log \frac{1}{\delta})$, we have the following. Let $S_1, \dots, S_K$ be sequences of indices as defined in Algorithm 2. Let $(G^*, U_1^*, \dots, U_K^*)$ and $(\tilde{G}^*, \tilde{U}_1^*, \dots, \tilde{U}_K^*)$ be minimizers of problem (3.2) on X and $X|_{S_1, \dots, S_K}$ for which the factor matrices are orthonormal, respectively. Then we have*

$$\ell_{R_1, \dots, R_K}(X|_{S_1, \dots, S_K}) = \ell_{R_1, \dots, R_K}(X) \pm O(\epsilon L^2(1 + 2MR)),$$

with probability at least $1 - \delta$, where $L = |X|_{\max}$, $M = \max\{|G^*|_{\max}, |\tilde{G}^*|_{\max}\}$, and $R = \prod_{k \in [K]} R_k$.

We remark that, for the matrix case (i.e., $K = 2$), $|G^*|_{\max}$ and $|\tilde{G}^*|_{\max}$ are equal to the maximum singular values of the original and sampled matrices, respectively.

3.4.1 Preliminaries

Let $X \in \mathbb{R}^{N_1 \times \dots \times N_K}$ be a tensor. We define

$$|X|_F = \sqrt{\sum_{i_1, \dots, i_K} X_{i_1 \dots i_K}^2}, \quad (\text{Frobenius norm})$$

$$|X|_{\max} = \max_{i_1 \in [N_1], \dots, i_K \in [N_K]} |X_{i_1 \dots i_K}|, \quad (\text{Max norm})$$

$$|X|_{\square} = \max_{S_1 \subseteq [N_1], \dots, S_K \subseteq [N_K]} \left| \sum_{i_1 \in S_1, \dots, i_K \in S_K} X_{i_1 \dots i_K} \right|. \quad (\text{Cut norm})$$

We note that these norms satisfy the triangle inequality.

For a vector $\mathbf{v} \in \mathbb{R}^n$ and a sequence $S = (x_1, \dots, x_s)$ of indices in $[n]$, we define the *restriction* $\mathbf{v}|_S \in \mathbb{R}^s$ of $\mathbf{v}$ as $(\mathbf{v}|_S)_i = v_{x_i}$ for $i \in [s]$. Let $X \in \mathbb{R}^{N_1 \times \dots \times N_K}$ be a

tensor, and $S_k = (x_1^k, \dots, x_s^k)$ be a sequence of indices in $[N_k]$ for each mode $k \in [K]$. Then, we define the *restriction* $X|_{S_1, \dots, S_K} \in \mathbb{R}^{s \times \dots \times s}$ of X to $S_1 \times \dots \times S_K$ as $(X|_{S_1, \dots, S_K})_{i_1 \dots i_K} = X_{x_{i_1}^1, \dots, x_{i_K}^K}$ for each $i_1 \in [N_1], \dots, i_K \in [N_K]$.

For a tensor $G \in \mathbb{R}^{R_1 \times \dots \times R_K}$ and vector-valued functions $\{F^{(k)} : [0, 1] \rightarrow \mathbb{R}^{R_k}\}_{k \in [K]}$, we define an order- K dikernel $\llbracket G; F^{(1)}, \dots, F^{(K)} \rrbracket : [0, 1]^K \rightarrow \mathbb{R}$ as

$$\llbracket G; F^{(1)}, \dots, F^{(K)} \rrbracket(x_1, \dots, x_K) = \sum_{r_1 \in [R_1], \dots, r_K \in [R_K]} G_{r_1, \dots, r_K} \prod_{k \in [K]} F^{(k)}(x_k)_{r_k}$$

We note that $\llbracket G; F^{(1)}, \dots, F^{(K)} \rrbracket$ is a continuous analogue of Tucker decomposition.

3.4.2 Proof of Theorem 3.2

To prove Theorem 3.2, we first consider the dikernel counterpart to the Tucker fitting problem, in which we want to minimize the following:

$$\ell_{R_1, \dots, R_K}(X) := \inf_{G \in \mathbb{R}^{R_1 \times \dots \times R_K}, \{f^{(k)} : [0, 1] \rightarrow \mathbb{R}^{R_k}\}_{k \in [K]}} \left| X - \llbracket G; f^{(1)}, \dots, f^{(K)} \rrbracket \right|_F^2, \quad (3.3)$$

The following lemma, which is proved in Sect. 3.4.3, states that the Tucker fitting problem and its dikernel counterpart have the same optimum values.

Lemma 3.4 *Let $X \in \mathbb{R}^{N_1 \times \dots \times N_K}$ be a tensor, and let $R_1, \dots, R_K \in \mathbb{N}$ be integers. Then, we have*

$$\ell_{R_1, \dots, R_K}(X) = \ell_{R_1, \dots, R_K}(X).$$

For a set of vector-valued functions $F = \{f^{(k)} : [0, 1] \rightarrow \mathbb{R}^{R_k}\}_{k \in [K]}$, we define $|F|_{\max} = \max_{k \in [K], r \in [R_k], x \in [0, 1]} f_r^{(k)}(x)$. For a dikernel $\mathcal{X} : [0, 1]^K \rightarrow \mathbb{R}$, we define a dikernel $\mathcal{X}^2 : [0, 1]^K \rightarrow \mathbb{R}$ as $\mathcal{X}^2(\mathbf{x}) = \mathcal{X}(\mathbf{x})^2$ for every $\mathbf{x} \in [0, 1]^K$. The following lemma, which is proved in Sect. 3.4.4, states that if $\mathcal{X}$ and $\mathcal{Y}$ are close in the cut norm, then the optimum values when the Tucker fitting problem is applied to them are also close.

Lemma 3.5 *Let $\mathcal{X}, \mathcal{Y} : [0, 1]^K \rightarrow \mathbb{R}$ be dikernels with $|\mathcal{X} - \mathcal{Y}|_{\square} \leq \epsilon$ and $|\mathcal{X}^2 - \mathcal{Y}^2|_{\square} \leq \epsilon$. For $R_1, \dots, R_K \in \mathbb{N}$, we have*

$$\ell_{R_1, \dots, R_K}(\mathcal{X}) = \ell_{R_1, \dots, R_K}(\mathcal{Y}) \pm 2\epsilon \left(1 + R(|G_{\mathcal{X}}|_{\max} |F_{\mathcal{X}}|_{\max}^K + |G_{\mathcal{Y}}|_{\max} |F_{\mathcal{Y}}|_{\max}^K) \right),$$

where $(G_{\mathcal{X}}, F_{\mathcal{X}} = \{f_{\mathcal{X}}^{(k)}\}_{k \in [K]})$ and $(G_{\mathcal{Y}}, F_{\mathcal{Y}} = \{f_{\mathcal{Y}}^{(k)}\}_{k \in [K]})$ are solutions to problem (3.3) on $\mathcal{X}$ and $\mathcal{Y}$, respectively, which have objective values exceeding the infima by at most ϵ , and $R = \prod_{k \in [K]} R_k$.

Proof (of Theorem 3.2) We apply Lemma 3.2 to $\mathcal{X}$ and $\mathcal{X}^2$. Thus, with probability at least $1 - \delta$, there exists a measure-preserving bijection $\pi : [0, 1] \rightarrow [0, 1]$ such that

$$|\mathcal{X} - \pi(\mathcal{X}|_{S_1, \dots, S_K})|_{\square} \leq \epsilon L \quad \text{and} \quad |\mathcal{X}^2 - \pi(\mathcal{X}^2|_{S_1, \dots, S_K})|_{\square} \leq \epsilon L^2.$$

In the following, we assume that this has happened. By Lemma 3.5 and the fact that $\ell_{R_1, \dots, R_K}(\mathcal{X}|_{S_1, \dots, S_K}) = \ell_{R_1, \dots, R_K}(\pi(\mathcal{X}|_{S_1, \dots, S_K}))$, we have

$$\ell_{R_1, \dots, R_K}(\mathcal{X}|_{S_1, \dots, S_K}) = \ell_{R_1, \dots, R_K}(\mathcal{X}) \pm \epsilon L^2 \left(1 + 2R(|G|_{\max}|F|_{\max}^K + |\tilde{G}|_{\max}|\tilde{F}|_{\max}^K) \right),$$

where $(G, F = \{f^{(k)}\}_{k \in [K]})$ and $(\tilde{G}, \tilde{F} = \{\tilde{f}^{(k)}\}_{k \in [K]})$ are as in the statement of Lemma 3.5. From the proof of Lemma 3.4, we can assume that $|G|_{\max} = |G^*|_{\max}$, $|\tilde{G}|_{\max} = |\tilde{G}^*|_{\max}$, $|F|_{\max} \leq 1$, and $|\tilde{F}|_{\max} \leq 1$ (owing to the orthonormality of $U_1^*, \dots, U_K^*$ and $\tilde{U}_1^*, \dots, \tilde{U}_K^*$). It follows that

$$\ell_{R_1, \dots, R_K}(\mathcal{X}|_{S_1, \dots, S_K}) = \ell_{R_1, \dots, R_K}(\mathcal{X}) \pm \epsilon L^2 \left(1 + 2R(|G^*|_{\max} + |\tilde{G}^*|_{\max}) \right). \quad (3.4)$$

Then, we have

$$\begin{aligned} \ell_{R_1, \dots, R_K}(\mathcal{X}|_{S_1, \dots, S_K}) &= \ell_{R_1, \dots, R_K}(\mathcal{X}|_{S_1, \dots, S_K}) && \text{(By Lemma 4)} \\ &= \ell_{R_1, \dots, R_K}(\mathcal{X}) \pm \epsilon L^2 \left(1 + 2R(|G^*|_{\max} + |\tilde{G}^*|_{\max}) \right) && \text{(By 4)} \\ &= \ell_{R_1, \dots, R_K}(\mathcal{X}) \pm \epsilon L^2 \left(1 + 2R(|G^*|_{\max} + |\tilde{G}^*|_{\max}) \right). && \text{(By Lemma 4)} \end{aligned}$$

Hence, the proof is complete.

3.4.3 Proof of Lemma 3.4

We say that a vector-valued function $f : [0, 1] \rightarrow \mathbb{R}^R$ is *orthonormal* if $\langle f_r, f_r \rangle = 1$ for every $r \in [R]$ and $\langle f_r, f_{r'} \rangle = 0$ if $r \neq r'$. First, we calculate the partial derivatives of the objective function. We omit the proof because it is a straightforward (but tedious) calculation.

Lemma 3.6 *Let $\mathcal{X} \in [0, 1]^K \rightarrow \mathbb{R}$ be a dikernel, $G \in \mathbb{R}^{R_1 \times \dots \times R_K}$ be a tensor, and $\{f^{(k)} : [0, 1] \rightarrow \mathbb{R}^{R_k}\}_{k \in [K]}$ be a set of orthonormal vector-valued functions. Then, we have*

$$\begin{aligned}
& \frac{\partial}{\partial f_{r_0}^{(k_0)}(x_0)} \left\| X - \llbracket G; f^{(1)}, \dots, f^{(K)} \rrbracket \right\|_F^2 \\
&= 2 \sum_{r_1, \dots, r_K : r_{k_0} = r_0} G_{r_1 \dots r_K} \int_{[0,1]^K : x_{k_0} = x_0} \mathcal{X}(\mathbf{x}) \prod_{k \in [K] \setminus \{k_0\}} f_{r_k}^{(k)}(x_k) d\mathbf{x} \\
&\quad - 2 \sum_{r_1, \dots, r_K} G_{r_1 \dots r_K} G_{r_1 \dots r_{k_0-1} r_0 r_{k_0+1} \dots r_K} f_{r_{k_0}}^{(k_0)}(x_0).
\end{aligned}$$

Proof (of Lemma 3.4) First, we show that (LHS) $\leq$ (RHS). Consider a sequence of solutions for the continuous problem (3.3) for which the objective values attain the infimum. For Tucker decompositions, it is well known that there exists a minimizer for which the factor matrices $U^{(1)}, \dots, U^{(K)}$ are orthonormal. By similar reasoning, we can show that the vector-valued functions $f^{(1)}, \dots, f^{(K)}$ in each solution of the sequence are orthonormal. As the objective function is coercive with respect to tensor G , we can take a subsequence for which G converges. Let G^* be the limit. Now, for any $\delta > 0$, we can create a matrix $\tilde{G}$ by perturbing G^* so that (i) by fixing G to $\tilde{G}$ in the continuous problem, the infimum increases only by δ , and (ii) a matrix constructed from $\tilde{G}$ is invertible and has a condition number at least $\delta' = \delta'(\delta) > 0$.

Now, consider a sequence of solutions for the continuous problem (3.3) with G fixed to $\tilde{G}$ for which the objective values attain the infimum. We can show that the partial derivatives converge to zero almost everywhere. For any $\epsilon > 0$, there then exists a solution $(\tilde{G}, f^{(1)}, \dots, f^{(K)})$ in the sequence such that the partial derivatives are at most ϵ almost everywhere.

Then by Lemma 3.6, for any $k_0 \in [K]$, $r_0 \in [R_{k_0}]$, and almost all $x \in [0, 1]$, we have

$$\begin{aligned}
& \sum_{r_1, \dots, r_K} \tilde{G}_{r_1 \dots r_K} \tilde{G}_{r_1 \dots r_{k_0-1} r_0 r_{k_0+1} \dots r_K} f_{r_{k_0}}^{(k_0)}(x_0) \\
&= \sum_{r_1, \dots, r_K : r_{k_0} = r_0} \tilde{G}_{r_1 \dots r_K} \int_{[0,1]^K : x_{k_0} = x_0} \mathcal{X}(\mathbf{x}) \prod_{k \in [K] \setminus \{k_0\}} f_{r_k}^{(k)}(x_k) d\mathbf{x} \pm \epsilon(k_0, r_0, x), \quad (3.5)
\end{aligned}$$

where $\epsilon(k_0, r_0, x) = O(\epsilon)$. Now, we consider a system of linear equations consisting of (3.5) for $r_0 = 1, \dots, R_{k_0}$, where the variables are $f_1^{k_0}(x_0), \dots, f_{R_{k_0}}^{k_0}(x_0)$. We can assume that the matrix involved in this system is invertible and has a positive condition number. For any $k \in [K]$, $r \in [R_k]$ and almost every pair $x, x' \in [0, 1]$ with $i_{N_k}(x) = i_{N_k}(x')$, we then have $f_{r_0}^{(k_0)}(x) = f_{r_0}^{(k_0)}(x') \pm O(\epsilon/\delta')$. For each $k \in [K]$, we can define a matrix $U^{(k)} \in \mathbb{R}^{N_k \times R_k}$ as $U_{ir}^{(k)} = f_r^{(k)}(x)$, where $x \in [0, 1]$ is an arbitrary value with $i_{N_k}(x) = i$. Then, we have

$$\begin{aligned}
\frac{1}{N} \left\| X - \llbracket \tilde{G}; U^{(1)}, \dots, U^{(K)} \rrbracket \right\|_F^2 &= \frac{1}{N} \sum_{i_1, \dots, i_K} \left(X_{i_1 \dots i_K} - \llbracket \tilde{G}; U^{(1)}, \dots, U^{(K)} \rrbracket_{i_1 \dots i_K} \right)^2 \\
&= \sum_{i_1, \dots, i_K} \int_{I_{i_1}^{N_1} \times \dots \times I_{i_K}^{N_K}} \left(\mathcal{X}(\mathbf{x}) - \llbracket \tilde{G}; f^{(1)}, \dots, f^{(K)} \rrbracket(\mathbf{x}) \pm O(\epsilon/\delta') \right)^2 d\mathbf{x} \\
&= \left\| \mathcal{X} - \llbracket \tilde{G}; f^{(1)}, \dots, f^{(K)} \rrbracket \right\|_F^2 \pm O(\epsilon^2 N / (\delta')^2)
\end{aligned}$$

for $N = \prod_{k \in [K]} N_k$. As the choice of ϵ and δ are arbitrary, we obtain (LHS) $\leq$ (RHS).

Second, we show that (RHS) $\leq$ (LHS). Let $U^{(k)} \in \mathbb{R}^{N_k \times R_k}$ ($k \in [K]$) be matrices. We define a vector-valued function $f^{(k)} : [0, 1] \rightarrow \mathbb{R}^{R_k}$ as $f_r^{(k)}(x) = U_{i_{N_k}(x)r}^{(k)}$ for each $k \in [K]$ and $r \in [R_k]$. Then, we have

$$\begin{aligned}
\left\| \mathcal{X} - \llbracket G; f^{(1)}, \dots, f^{(K)} \rrbracket \right\|_F^2 &= \int_{[0,1]^K} \left(\mathcal{X}(\mathbf{x}) - \llbracket G; f^{(1)}, \dots, f^{(K)} \rrbracket(\mathbf{x}) \right)^2 d\mathbf{x} \\
&= \sum_{i_1, \dots, i_K} \int_{\prod_{k \in [K]} I_{i_k}^{N_k}} \left(\mathcal{X}(\mathbf{x}) - \llbracket G; f^{(1)}, \dots, f^{(K)} \rrbracket(\mathbf{x}) \right)^2 d\mathbf{x} \\
&= \frac{1}{N} \sum_{i_1, \dots, i_K} \left(X_{i_1 \dots i_K} - \llbracket G; U^{(1)}, \dots, U^{(K)} \rrbracket_{i_1 \dots i_K} \right)^2 \\
&= \frac{1}{N} \left\| X - \llbracket G; U^{(1)}, \dots, U^{(K)} \rrbracket \right\|_F^2,
\end{aligned}$$

from which the claim follows.

3.4.4 Proof of Lemma 3.5

For a sequence of functions $f^{(1)}, \dots, f^{(K)}$, we define their tensor product $\bigotimes_{k \in [K]} f^{(k)} \in [0, 1]^K \rightarrow \mathbb{R}$ as $\bigotimes_{k \in [K]} f^{(k)}(x_1, \dots, x_K) = \prod_{k \in [K]} f^{(k)}(x_k)$, which is a dikernel of order- K .

The cut norm is useful for bounding the absolute value of the inner product between a tensor and a tensor product:

Lemma 3.7 *Let $\epsilon \geq 0$ and $\mathcal{W} : [0, 1]^K \rightarrow \mathbb{R}$ be a dikernel with $|\mathcal{W}|_{\square} \leq \epsilon$. Then, for any functions $f^{(1)}, \dots, f^{(K)} : [0, 1] \rightarrow [-L, L]$, we have $|\langle \mathcal{W}, \bigotimes_{k \in [K]} f^{(k)} \rangle| \leq \epsilon L^K$.*

Proof For $\tau \in \mathbb{R}$ and the function $h : [0, 1] \rightarrow \mathbb{R}$, let $L_\tau(h) := \{x \in [0, 1] \mid h(x) = \tau\}$ be the level set of h at τ . For $f^{(i)} = f^{(i)}/L$, we have

$$\begin{aligned}
\left| \left\langle \mathcal{W}, \bigotimes_{k \in [K]} f^{(k)} \right\rangle \right| &= L^K \left| \left\langle \mathcal{W}, \bigotimes_{k \in [K]} f'^{(k)} \right\rangle \right| \\
&= L^K \left| \int_{[-1,1]^K} \prod_{k \in [K]} \tau_k \int_{\prod_{k \in [K]} L_{\tau_k}(f'^{(k)})} \mathcal{W}(\mathbf{x}) d\mathbf{x} d\boldsymbol{\tau} \right| \\
&\leq L^K \int_{[-1,1]^K} \prod_{k \in [K]} |\tau_k| \left| \int_{\prod_{k \in [K]} L_{\tau_k}(f'^{(k)})} \mathcal{W}(\mathbf{x}) d\mathbf{x} d\boldsymbol{\tau} \right| \\
&\leq \epsilon L^K \int_{[-1,1]^K} \prod_{k \in [K]} |\tau_k| d\boldsymbol{\tau} = \epsilon L^K.
\end{aligned}$$

Thus, we have the following:

Lemma 3.8 *Let $\mathcal{X}, \mathcal{Y} : [0, 1]^K \rightarrow \mathbb{R}$ be kernels with $|\mathcal{X} - \mathcal{Y}|_{\square} \leq \epsilon$ and $|\mathcal{X}^2 - \mathcal{Y}^2|_{\square} \leq \epsilon$, where $\mathcal{X}^2(\mathbf{x}) = \mathcal{X}(\mathbf{x})^2$ and $\mathcal{Y}^2(\mathbf{x}) = \mathcal{Y}(\mathbf{x})^2$ for every $\mathbf{x} \in [0, 1]^K$. Then, for any tensor $G \in \mathbb{R}^{R_1 \times \dots \times R_K}$ and a set of vector-valued functions $F = \{f^{(k)} : [0, 1] \rightarrow \mathbb{R}^{R_k}\}_{k \in [K]}$, we have*

$$|\mathcal{X} - \llbracket G; f^{(1)}, \dots, f^{(K)} \rrbracket|_F^2 = |\mathcal{Y} - \llbracket G; f^{(1)}, \dots, f^{(K)} \rrbracket|_F^2 \pm \epsilon \left(1 + 2R|G|_{\max} |F|_{\max}^K\right),$$

where $R = \prod_{k \in [K]} R_k$.

Proof We have

$$\begin{aligned}
&\left| |\mathcal{X} - \llbracket G; f^{(1)}, \dots, f^{(K)} \rrbracket|_F^2 - |\mathcal{Y} - \llbracket G; f^{(1)}, \dots, f^{(K)} \rrbracket|_F^2 \right| \\
&= \left| \int_{[0,1]^K} \left(\mathcal{X}(\mathbf{x}) - \llbracket G; f^{(1)}, \dots, f^{(K)} \rrbracket(\mathbf{x}) \right)^2 d\mathbf{x} \right. \\
&\quad \left. - \int_{[0,1]^K} \left(\mathcal{Y}(\mathbf{x}) - \llbracket G; f^{(1)}, \dots, f^{(K)} \rrbracket(\mathbf{x}) \right)^2 d\mathbf{x} \right| \\
&= \left| \int_{[0,1]^K} (\mathcal{X}(\mathbf{x})^2 - \mathcal{Y}(\mathbf{x})^2) d\mathbf{x} - 2 \int_{[0,1]^K} (\mathcal{X}(\mathbf{x}) - \mathcal{Y}(\mathbf{x})) \llbracket G; f^{(1)}, \dots, f^{(K)} \rrbracket(\mathbf{x}) d\mathbf{x} \right| \\
&\leq |\mathcal{X}^2 - \mathcal{Y}^2|_{\square} + 2 \sum_{r_1 \in [R_1], \dots, r_K \in [R_K]} |G_{r_1 \dots r_K}| \cdot \left| \left\langle \mathcal{X} - \mathcal{Y}, \bigotimes_{k \in [K]} f_{r_k}^{(k)} \right\rangle \right| \\
&\leq \epsilon + 2\epsilon R |G|_{\max} |F|_{\max}^K
\end{aligned}$$

by Lemma 3.7.

Proof (of Lemma 3.5) By Lemma 3.8, we have

$$\begin{aligned}
\left| \mathcal{Y} - \llbracket G_{\mathcal{Y}}; f_{\mathcal{Y}}^{(1)}, \dots, f_{\mathcal{Y}}^{(K)} \rrbracket \right|_F^2 &\leq \left| \mathcal{Y} - \llbracket G_{\mathcal{X}}; f_{\mathcal{X}}^{(1)}, \dots, f_{\mathcal{X}}^{(K)} \rrbracket \right|_F^2 + \epsilon \\
&\leq \left| \mathcal{X} - \llbracket G_{\mathcal{X}}; f_{\mathcal{X}}^{(1)}, \dots, f_{\mathcal{X}}^{(K)} \rrbracket \right|_F^2 + \left(2\epsilon + 2\epsilon R \|G_{\mathcal{X}}\|_{\max} \|F_{\mathcal{X}}\|_{\max}^K \right).
\end{aligned}$$

Similarly, we have

$$\begin{aligned} \left\| \mathcal{X} - \llbracket G_{\mathcal{X}}; f_{\mathcal{X}}^{(1)}, \dots, f_{\mathcal{X}}^{(K)} \rrbracket \right\|_F^2 &\leq \left\| \mathcal{X} - \llbracket G_{\mathcal{Y}}; f_{\mathcal{Y}}^{(1)}, \dots, f_{\mathcal{Y}}^{(K)} \rrbracket \right\|_F^2 + \epsilon \\ &\leq \left\| \mathcal{Y} - \llbracket G_{\mathcal{Y}}; f_{\mathcal{Y}}^{(1)}, \dots, f_{\mathcal{Y}}^{(K)} \rrbracket \right\|_F^2 + \left(2\epsilon + 2\epsilon R \|G_{\mathcal{Y}}\|_{\max} \|F_{\mathcal{Y}}\|_{\max}^K \right). \end{aligned}$$

Hence, the claim follows.

References

1. L. Bottou, Stochastic learning, in *Advanced Lectures on Machine Learning* (2004), pp. 146–168
2. K.L. Clarkson, E. Hazan, D.P. Woodruff, Sublinear optimization for machine learning. *J. ACM* **59**(5), 23:1–23:49 (2012)
3. Lieven De Lathauwer, Bart De Moor, Joos Vandewalle, On the best rank-1 and rank- $(r_1, r_2, \dots, r_n)$ approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.* **21**(4), 1324–1342 (2000)
4. A. Frieze, R. Kannan, The regularity lemma and approximation schemes for dense problems, in *FOCS* (1996), pp. 12–20
5. K. Hayashi, Y. Yoshida, Minimizing quadratic functions in constant time, in *NIPS* (2016), pp. 2217–2225
6. K. Hayashi, Y. Yoshida, Fitting low-rank tensors in constant time, in *NIPS* (2017), pp. 2473–2481
7. L. Lovász, *Large Networks and Graph Limits* (American Mathematical Society, 2012)
8. K.P. Murphy, *Machine Learning: A Probabilistic Perspective* (The MIT Press, 2012)
9. Taiji Suzuki, Masashi Sugiyama, Least-squares independent component analysis. *Neural Comput.* **23**(1), 284–301 (2011)
10. M. Sugiyama, T. Suzuki, T. Kanamori, *Density Ratio Estimation in Machine Learning* (Cambridge University Press, 2012)
11. Ledyard R. Tucker, Some mathematical notes on three-mode factor analysis. *Psychometrika* **31**(3), 279–311 (1966)
12. K.I. Christopher, in *Using the Nyström Method to Speed up Kernel Machines*, *NIPS* eds. by C. Williams, M. Seeger (2001)
13. M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, M. Sugiyama, Relative density-ratio estimation for robust distribution comparison, in *NIPS* (2011)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.