Sparse additive function decompositions facing basis transforms

Fatima Antarou Ba * Oleh Melnyk* Christian Wald*
Gabriele Steidl*

March 29, 2024

High-dimensional real-world systems can often be well characterized by a small number of simultaneous low-complexity interactions. The analysis of variance (ANOVA) decomposition and the anchored decomposition are typical techniques to find sparse additive decompositions of functions. In this paper, we are interested in a setting, where these decompositions are not directly spare, but become so after an appropriate basis transform. Noting that the sparsity of those additive function decompositions is equivalent to the fact that most of its mixed partial derivatives vanish, we can exploit a connection to the underlying function graphs to determine an orthogonal transform that realizes the appropriate basis change. This is done in three steps: we apply singular value decomposition to minimize the number of vertices of the function graph, and joint block diagonalization techniques of families of matrices followed by sparse minimization based on relaxations of the zero "norm" for minimizing the number of edges. For the latter one, we propose and analyze minimization techniques over the manifold of special orthogonal matrices. Various numerical examples illustrate the reliability of our approach for functions having, after a basis transform, a sparse additive decomposition into summands with at most two variables.

1. Introduction

The approximation of high-dimensional functions is a classical topic of mathematical analysis with rich real-world applications. Due to new numerical techniques arising from (stochastic) Fourier- and wavelet [7, 42, 35, 36] as well as kernel [39] methods and deep learning approaches, the topic has recently attained increasing attention. For

*Technische Universität Berlin, Institute of Mathematics, Straße des 17. Juni 136, D-10623 Berlin, Germany

example, including information about the structure of the function class of interest into a deep neural network architecture can improve its approximation quality [3, 17, 14]. In particular, in [17] block sparse additive neural network architectures are developed with sparsity patterns estimated from data. These block sparse additive neural networks show increased training speed, better memory efficiency and improved generalization.
In this paper, we are interested in high-dimensional functions admitting an additive decomposition into functions depending only on a few variables, i.e.,

$$f(x) = \sum_{\mathbf{u}\in\mathcal{S}} f_{\mathbf{u}}(x_{\mathbf{u}}), \quad x_{\mathbf{u}} \coloneqq (x_i)_{i\in\mathbf{u}}, \tag{1}$$

where $\mathcal{S}$ consists of small subsets of $\{1,\dots,d\}$, and $d \gg 1$ is the dimension of the problem. Recently, we dealt with such decompositions in the context of multimarginal optimal transport, where only special structured cost functions can be treated in an efficient way [5, 8].

In general, decompositions of functions of the form (**??**) are not unique, but there exist prominent examples in the literature having special desirable properties. So the analysis of variance (ANOVA) decomposition [13, 43] and anchored [24, 25] decompositions and their generalizations [32] arise in finance for option pricing, bond valuation and the pricing of collateral mortgage-backed securities problems [18]. The ANOVA decomposition is uniquely determined and the anchored decomposition up to a so-called anchored point. Both decompositions make it possible to analyze the different dimensions and their interactions, and to perform high dimensional integration [16, 19] using quadrature methods as well as infinite-dimensional integration [6, 19, 31]. Further, in [23], we proposed, inspired by the analysis of variance ANOVA decomposition of functions, a Gaussian-Uniform mixture model on the high-dimensional torus which relies on the assumption that the function we wish to approximate can be well explained by limited variable interactions.

A further motivation for considering additive function decompositions comes from probabilistic graphical models [47, Chapter 13]. According to the Hammersley-Clifford theorem, the logarithm of a continuous density function ϕ of a random vector can be decomposed as (1) for a certain class of sets $\mathcal{S}$ defined by conditional dependency between the components of the random vector. Finding an appropriate additive decomposition allows for inference on the interactions between the observed data, which naturally leads to applications related to causality inference and complex systems that frequently appear in computational biology [48], climate change [40], speech recognition [9, 10], predictive modelling [28], energy systems [34] and many more.

One of the key features of ANOVA, anchored and Hammersley-Clifford decompositions is their minimality in the number of elements $\mathbf{u} \in \mathcal{S}$. That is, for $\mathbf{u} \in \mathcal{S}$ it is not possible to find further decomposition $f_{\mathbf{u}}(x_{\mathbf{u}}) = \sum_{\mathbf{v}\subsetneq\mathbf{u}} f_{\mathbf{v}}(x_{\mathbf{v}})$ consisting only of smaller subsets of $\mathbf{u}$. A composition with this minimality property is closely related to the structure of the graph associated with f. For twice continuously differentiable f, an undirected graph $G(f)$ is defined by vertices and edges

$$\mathcal{V}(f) = \{i \in [d] : \partial_{x_i} f \neq 0\} \quad \text{and} \quad \mathcal{E}(f) = \{(i,j) \in [d]^2 : \partial^2_{x_i,x_j} f \neq 0 \text{ and } i \neq j\},$$

respectively. Then, $\mathcal{S}$ for both ANOVA and anchored decomposition is a subset of the cliques in $\mathcal{C}(f)$ of the graph. Furthermore, $\mathcal{S}$ in the Hammersley-Clifford decomposition coincides with maximal cliques in $\mathcal{C}(f)$. We will use the relation between function graphs and sparse additive function decompositions in Section 2.

In this paper, we are interested in the setting, where the function does only admit a sparse additive decomposition after a basis transform. In other words, given (noisy) values of the gradient and the Hessian of f, we aim to find an orthogonal matrix $U \in O(d)$ such that the graph of the function

$$f_U \coloneqq f(U\cdot)$$

has the smallest number of cliques. We propose a three-step procedure. First, we find an orthogonal matrix $U \in O(d)$ such that the graph of f_U has the smallest number of vertices $|\mathcal{V}(f_U)|$. We will see that this can be done by the singular value decomposition of the matrix having the gradient of f at different values as columns. Then, in the second and third steps, we aim to minimize the number of edges $|\mathcal{E}(f_U)|$ by joint block diagonal decomposition of the Hessians of f at the given points and subsequent sparsification of the individual blocks. While we rely on results in [38, 37] for finding the joint blockdiagonalization, we apply recent sparse optimization methods with a relaxed 0 "norm" to further sparsify the single blocks. The third step requires the solution of a nonconvex optimization problem over the manifold of special orthogonal matrices for which the Riemannian gradient descent [11] or the Landing method [1] is employed in combination with grid search.

Outline of the paper. In Section 2, we consider sparse additive decompositions. We are interested in so-called minimal decompositions and recall that both the ANOVA and anchored decomposition are of such minimal type. We address the relation between minimal additive decompositions, first and second derivatives of functions and corresponding function graphs. Finally, Theorem 2.6 gives an estimate of the influence of $\|\partial_{\mathbf{v}} f\|_\infty$ to summands $f_{\mathbf{u}}$ in the anchored and ANOVA decompositions for $\mathbf{u} \supseteq \mathbf{v}$. In Section 3, we detail the three steps to obtain the desired basis transform towards a function which has a sparse additive decomposition. Section 4, deals with the optimization problem arising in Step 3, which requires optimization techniques over the special orthogonal group. Numerical results for functions admitting, after an appropriate basis transform, a sparse additive decomposition into summands depending at most on two variables are given in Section 5. The appendix contains technical proofs, details on the algorithms and gives additional numerical results.

2. Sparse additive decompositions and mixed derivatives

In the following, let $[d] \coloneqq \{1, \dots, d\}$. We set $D \coloneqq \bigotimes_{i\in[d]} I_i$, where $I_i \coloneqq [a_i, b_i]$, $a_i < b_i$. For a subset $\mathbf{u} \subseteq [d]$, we use $D_{\mathbf{u}} \coloneqq \bigotimes_{i\in\mathbf{u}} I_i$. By $\lambda_{\mathbf{u}}$, we denote the normalized Lebesgue measure on $D_{\mathbf{u}}$ and $\lambda \coloneqq \lambda_{[d]}$. Further, for $\mathbf{u} = \{i_1, \dots, i_k\} \subseteq [d]$, we use the abbreviation

$x_{\mathbf{u}} \coloneqq (x_{i_1}, \dots, x_{i_k})$ and

$$\partial_{\mathbf{u}} \coloneqq \partial^{|\mathbf{u}|}_{x_{i_1},\dots,x_{i_k}}.$$

By $\mathcal{F}$, we denote an appropriate subspace of function $f : D \to \mathbb{R}$ which will specified later. We define projection operators on $\mathcal{F}$ fulfilling the following assumption:

Assumption I: The operators P_j, $j \in [d]$ are commuting projections on $\mathcal{F}$, i.e.,

$$P_i P_j f = P_j P_i f \quad \text{for all} \quad f \in \mathcal{F},$$

$P_j f$ is independent of the j-th component x_j and $P_j f = f$ if f is independent of x_j.

We set $P_{\mathbf{u}} \coloneqq \prod_{j\in\mathbf{u}} P_j$. Based on the above projection operators, Kuo et al. [32] defined an additive decomposition of functions in $\mathcal{F}$ which fulfills a certain minimization property.

Theorem 2.1. *Let P_j, $j \in [d]$ fulfill Assumption I. Then any $f \in \mathcal{F}$ admits a decomposition*

$$f(x) = \sum_{\mathbf{u}\subseteq[d]} f_{\mathbf{u}}(x_{\mathbf{u}}) \tag{2}$$

with

$$f_{\mathbf{u}} \coloneqq \prod_{j\in\mathbf{u}} (Id - P_j) P_{[d]\setminus\mathbf{u}} f = \sum_{\mathbf{v}\subseteq\mathbf{u}} (-1)^{|\mathbf{u}|-|\mathbf{v}|} P_{[d]\setminus\mathbf{u}} f. \tag{3}$$

This decomposition is minimal in the following sense: if for an arbitrary decomposition

$$f(x) = \sum_{\mathbf{u}\subseteq[d]} \tilde{f}_{\mathbf{u}}(x_{\mathbf{u}}),$$

we have $\tilde{f}_{\mathbf{v}} = 0$ for all $\mathbf{v} \supseteq \mathbf{u}$ and some $\mathbf{u} \in [d]$, then also $f_{\mathbf{v}} = 0$ in (2).

To illustrate the role of the indices, we give an example.

Example 2.2. For $d = 3$, we have

$$f = f_0 + f_1 + f_2 + f_3 + f_{1,2} + f_{1,3} f + f_{2,3} + f_{1,2,3}$$

with $f_0 \coloneqq f_\emptyset = P_1 P_2 P_3 f = \text{const}$ and

$$\begin{aligned}
f_1 &= (\mathrm{Id} - P_1) P_2 P_3 f = P_2 P_3 f - P_1 P_2 P_3 f, \\
f_{1,2} &= (\mathrm{Id} - P_1)(\mathrm{Id} - P_2) P_3 f = P_3 f - P_1 P_3 f - P_2 P_3 f + P_1 P_2 P_3 f \\
f_{1,2,3} &= (\mathrm{Id} - P_1)(\mathrm{Id} - P_2)(\mathrm{Id} - P_3) f \\
&= f - P_1 f - P_2 f - P_3 f + P_1 P_2 f + P_1 P_3 f + P_2 P_3 f - P_1 P_2 P_3 f
\end{aligned}$$

and similarly for the other summands.

The concrete decomposition (2) depends on the chosen projection which must be well defined on the space $\mathcal{F}$. Two frequently addressed decompositions are the anchored and the ANOVA decompositions, where both projections fulfill Assumption I:

i) the *anchored decomposition* with anchor point $c = (c_1, \ldots, c_d) \in D$ is determined for any $f : D \to \mathbb{R}$ by

$$P_i f = P_{i,c} f \coloneqq f(x_1, \cdots, x_{i-1}, c_i, x_{i+1}, \cdots, x_d), \quad i \in [d], \tag{4}$$

and we denote the corresponding decomposition (2) by

$$f(x) = \sum_{\mathbf{u} \subseteq [d]} f_{\mathbf{u},c}(x_{\mathbf{u}}). \tag{5}$$

ii) the *analysis of variance* (ANOVA) *decomposition* is given for absolutely integrable functions $f : D \to \mathbb{R}$ by

$$P_i f = P_{i,A} \coloneqq \frac{1}{b_i - a_i} \int_{I_i} f \, \mathrm{d}x_i, \quad i \in [d],$$

and we use the notation

$$f(x) = \sum_{\mathbf{u} \subseteq [d]} f_{\mathbf{u},A}(x_{\mathbf{u}}).$$

The relation between the anchor and ANOVA summands follows immediately by their definition.

Proposition 2.3. *For $f : D \to \mathbb{R}$ be absolutely integrable and $\mathbf{u} \subseteq [d]$, we have*

$$f_{\mathbf{u},A} = \int_D f_{\mathbf{u},c}(x_{\mathbf{u}}) \, \mathrm{d}\lambda(c).$$

Proof. By (3), we obtain

$$\begin{aligned}
\int_D f_{\mathbf{u},c}(x_{\mathbf{u}}) \mathrm{d}\lambda(c) &= \int_D \sum_{\mathbf{v} \subseteq \mathbf{u}} (-1)^{|\mathbf{u}|-|\mathbf{v}|} (P_{[d]\setminus\mathbf{v},c} f)(x_{\mathbf{v}}) \, \mathrm{d}\lambda(c) \\
&= \sum_{\mathbf{v} \subseteq \mathbf{u}} (-1)^{|\mathbf{u}|-|\mathbf{v}|} \int_D (P_{[d]\setminus\mathbf{v},c} f)(x_{\mathbf{v}}) \, \mathrm{d}\lambda(c) \\
&= \sum_{\mathbf{v} \subseteq \mathbf{u}} (-1)^{|\mathbf{u}|-|\mathbf{v}|} \int_{D_{[d]\setminus\mathbf{v}}} f(x_{\mathbf{v}}, c_{[d]\setminus v}) \mathrm{d}\lambda_{[d]\setminus\mathbf{v}}(c) \\
&= \sum_{\mathbf{v} \subseteq \mathbf{u}} (-1)^{|\mathbf{u}|-|\mathbf{v}|} P_{[d]\setminus\mathbf{v},A} f = f_{\mathbf{u},A}(x_{\mathbf{u}}).
\end{aligned}$$

□

Next, we are interested in the relation between first- and second-order derivatives of $f \in C^2(D)$ and minimal additive decompositions (2). For this, it appears to be useful to consider the undirected graphs of functions.

Definition 2.4 (Graph of a functions)**.** To $f \in C^2(D)$, we assign the graph $\mathcal{G}(f)$ with vertices

$$\mathcal{V}(f) \coloneqq \{i \in [d] : \partial_i f \neq 0\} \tag{6}$$

and edges

$$\mathcal{E}(f) \coloneqq \{(i,j) \in [d]^2 : \partial_{i,j} f \neq 0 \text{ for } i \neq j\}.$$

A *clique* $\mathcal{C}(f)$ of $\mathcal{G}(f)$ is a subset of vertices of $\mathcal{V}(f)$ such that every two distinct vertices are connected.

Then we have the following relation.

Theorem 2.5. *Let $f \in C^2(D)$. Assume that $\partial_{i,j} f = 0$ for some $i \neq j$. Then we have in (2) that $f_{\mathbf{u}} = 0$ for all $\mathbf{u} \in [d]$ with $\{i,j\} \subseteq \mathbf{u}$. Moreover, it holds*

$$f(x) = \sum_{\mathbf{u} \subseteq \mathcal{C}(f)} f_{\mathbf{u}}(x_{\mathbf{u}}).$$

Proof. We will show that f has a decomposition

$$f = \sum_{\mathbf{u} \subseteq [d]} \tilde{f}_{\mathbf{u}} = P_{i,c} f + P_{j,c} f - P_{\{i,j\},c} f. \tag{7}$$

Then this decomposition fulfills $\tilde{f}_{\mathbf{u}} = 0$ whenever $\{i,j\} \subseteq \mathbf{u}$, and Theorem 2.1 implies that $f_{\mathbf{u}} = 0$. This gives the first part of the assertion. As a consequence, $f_{\mathbf{u}}$ is zero unless all pairs $\{i,j\} \subseteq \mathbf{u}$ admit $(i,j) \in \mathcal{E}(f)$, i.e., $\mathbf{u}$ is a clique and we are done.

It remains to prove (7). Without loss of generality, let $\partial_{1,2} f = 0$. Then the Taylor expansion at $\tilde{x} \coloneqq (c_1, c_2, x_3, \ldots, x_d)$ reads as

$$\begin{aligned} f(x) = f(\tilde{x}) &+ \partial_1 f(\tilde{x})(x_1 - c_1) + \partial_2 f(\tilde{x})(x_2 - c_2) \\ &+ R_{1,1}(x_1 - c_1)^2 + R_{2,2}(x_2 - c_2)^2 + 2R_{1,2}(x_1 - c_1)(x_2 - c_2), \end{aligned}$$

where

$$R_{i,j} \coloneqq \int_0^1 (1-t) \partial_{i,j} f(\tilde{x} + t(x - \tilde{x}))\, \mathrm{d}t, \quad i,j \in \{1,2\}.$$

By the assumption, $R_{1,2} f = 0$. Further, the Taylor expansion of $P_{2,c} f$ at $\tilde{x}$ is

$$P_{2,c} f(x) = f(\bar{x}) = f(\tilde{x}) + \partial_1 f(\tilde{x})(x_1 - c_1) + \bar{R}_{1,1}(x_1 - c_1)^2$$

with $\bar{x} \coloneqq (x_1, c_2, x_3, \ldots, x_d)$ and

$$\bar{R}_{1,1} \coloneqq \int_0^1 (1-t) \partial_{1,1} f(\tilde{x} + t(\bar{x} - \tilde{x}))\, \mathrm{d}t,$$

and similarly for $P_{1,c} f$. Since $\partial_1 f$ is constant with respect to x_2, the same holds true for $\partial_{1,1} f$. Hence we conclude by their definition that $R_{1,1} = \bar{R}_{1,1}$. Consequently, we obtain

$$f(x) = P_{1,c} f(x) + P_{2,c} f(x) - f(\tilde{x}) = P_{1,c} f(x) + P_{2,c} f(x) - P_{\{1,2\},c} f(x).$$

This finishes the proof. □

We have seen that $\partial_{i,j} f = 0$ implies that all summands $f_{\mathbf{u}}$ with $\mathbf{u} \supseteq \{i,j\}$ vanish in (2). Next, we want to estimate the general influence of $\|\partial_{\mathbf{v}} f\|_\infty$ to summands $f_{\mathbf{u},\cdot}$ in the anchored and ANOVA decompositions for $\mathbf{u} \supseteq \mathbf{v}$.

Theorem 2.6. *Let $\mathbf{v} \subseteq [d]$ and $f \in C^{|\mathbf{v}|}(D)$. Then, for $\mathbf{u} \supseteq \mathbf{v}$, the following estimates hold true in the anchor and ANOVA decompositions:*

i) $\|f_{\mathbf{u},c}\|_\infty \leq 2^{|\mathbf{u}|-|\mathbf{v}|} \|\partial_{x_\mathbf{v}} f\|_\infty \, \lambda_\mathbf{v}(D_\mathbf{v})$,

ii) $\|f_{\mathbf{u},A}\|_\infty \leq 2^{|\mathbf{u}|-|\mathbf{v}|} \|\partial_{x_\mathbf{v}} f\|_\infty \, \lambda(D)\lambda_\mathbf{v}(D_\mathbf{v})$,

iii) $\big\|f_{\mathbf{u},A}\big\|_1 \leq 2^{|\mathbf{u}|-|\mathbf{v}|} \|\partial_{x_\mathbf{v}} f\|_1 \, \lambda_\mathbf{u}(D_\mathbf{u})\lambda_\mathbf{v}(D_\mathbf{v})$.

The proof is given in Appendix A.

3. Basis transforms towards sparse function decomposition

Decompositions of the form (2) can be applied for an efficient integration over high-dimensional data if

i) the number of non-vanishing components $\mathbf{u} \in [d]$, and/or

ii) their cardinalities $|\mathbf{u}|$

are small. In Theorem 2.5, we observed that these properties depend on the structure of the underlying function graph $\mathcal{G}(f)$ and in particular on the set of cliques $\mathcal{C}(f)$. More precisely $\mathcal{G}(f)$ should not contain large cliques.

In this section, we are interested in the case when the function f admits only a sparse additive decomposition after an appropriate basis transform. We aim to determine such a basis transform given the first and second partial derivatives of f at various points in $D \subset \mathbb{R}^d$. Later, in our numerical experiments, we will assume that the summands depend at most on two of the transformed variables.

Let $\mathbb{B}_r(d) \coloneqq \{x \in \mathbb{R}^d : \|x\| \leq r\}$. If the dimension is clear, we just write $\mathbb{B}_r = \mathbb{B}_r(d)$. By $M_d(\mathbb{R})$, we denote the space of $d \times d$ real-valued matrices. Let $O(d)$ denote the Lie group of orthogonal $d \times d$ matrices and $SO(d)$ its subgroup of rotation matrices having determinant 1. We deal with functions $f : \mathbb{B}_r(d) \to \mathbb{R}$. For $U \in O(d)$, we set $f_U \coloneqq f(U\cdot)$. Clearly, $\mathbb{B}_r(d)$ is invariant under the action of orthogonal matrices, so that $f_U : \mathbb{B}_r(d) \to \mathbb{R}$ is also well-defined. Further, we have for twice differentiable f that

$$\nabla f_U = U^{\mathrm{T}} \nabla f(U\cdot) \quad \text{and} \quad \nabla^2 f_U = U^{\mathrm{T}} \nabla^2 f(U\cdot) U. \tag{8}$$

Example 3.1. Consider the orthogonal matrix

$$U = (u_1, u_2, u_3, u_4) = \frac{1}{2} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 \end{pmatrix}.$$

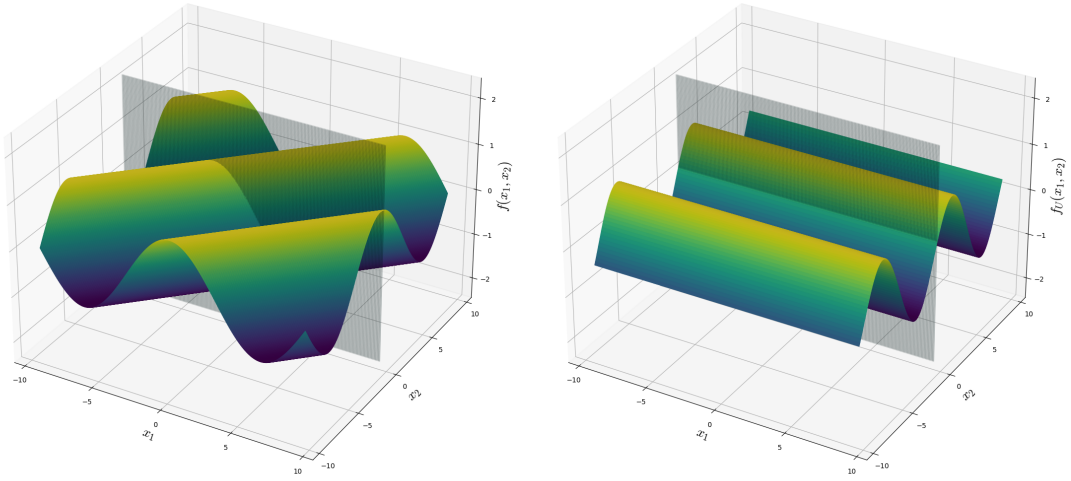

Figure 1: Left: $f(x) \coloneqq \sin(u_1^{\mathrm{T}} x\sqrt{2}/2)$ where $U = (u_1, u_2)$ is a 2-dimensional rotation matrix of rotation angle $\pi/4$. Right: Plot of f_U. The left-hand diagram shows that f depends on both variables x_1 and x_2 while f_U is constant in x_1.

Then the function

$$f(x) = h_1(u_1^{\mathrm{T}} x, u_2^{\mathrm{T}} x) + h_2(u_1^{\mathrm{T}} x, u_3^{\mathrm{T}} x), \quad x = (x_1, x_2, x_3, x_4)$$

has in general no sparse decomposition in the components of x, but

$$f_U(x) = f(Ux) = h_1(u_1^{\mathrm{T}} Ux, u_2^{\mathrm{T}} Ux) + h_2(u_1^{\mathrm{T}} Ux, u_3^{\mathrm{T}} Ux) = h_1(x_1, x_2) + h_2(x_1, x_3)$$

admits such a decomposition. This phenomenon is also illustrated in Figures 1,2,3.

Based on the given values of the gradient $\nabla f(x^{(n)})$ and the Hessian $\nabla^2 f(x^{(n)})$ at $x^{(n)} \in \mathbb{B}_r(d)$, $n \in [N]$, we are searching for those matrices $U \in O(d)$ for which the graph $\mathcal{G}(f_U)$ has the smallest number of vertices and edges. We will find (an approximation of) such $U \in O(d)$ in two steps which are detailed in the following subsections:

1. **Vertex minimization**: Find a vertex minimizing matrix

$$U_{\mathcal{V}} \in \operatorname*{argmin}_{U \in O(d)} \mathcal{V}(f_U). \tag{9}$$

2. **Edge minimization**: Using only the relevant vertices appearing in $\mathcal{G}(f_{U_{\mathcal{V}}})$, say wlog related to $x = (x_1, \ldots, x_{d_1})$, $d_1 \leq d$, we reduce our attention to $g(x) \coloneqq f_{U_{\mathcal{V}}}(x, 0)$, $x \in \mathbb{B}_r(d_1)$ and search for an edge-minimizing matrix, i.e.

$$U_{\mathcal{E}} \in \operatorname*{argmin}_{U \in O(d_1)} \mathcal{E}(g) \tag{10}$$

in two steps:

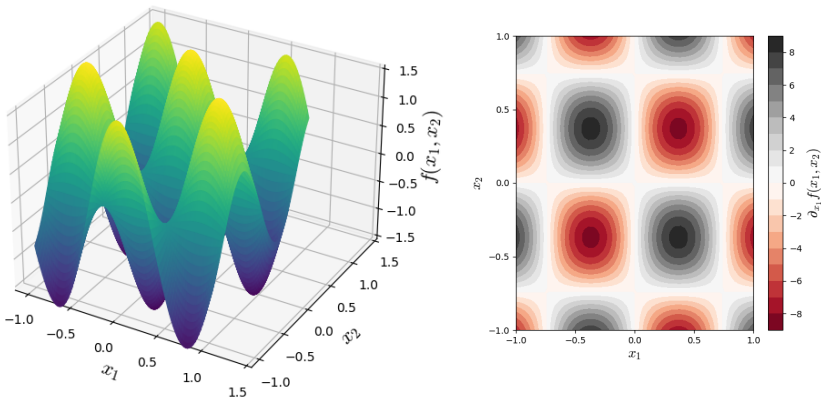

Figure 2: Left: $f(x) = \sin(5u_1^{\mathrm{T}}x) + \sin(5u_2^{\mathrm{T}}x)$ where $U = (u_1, u_2)$ is a 2-dimensional rotation matrix of angle $\pi/4$. Right: Image of the partial derivative $\partial_{x_1} f(x)$. The partial derivative $\partial_{x_1} f(x)$ depends on x_1 and x_2 since its values vary in both variables and thus f cannot be decomposed as a sum of two univariate functions.

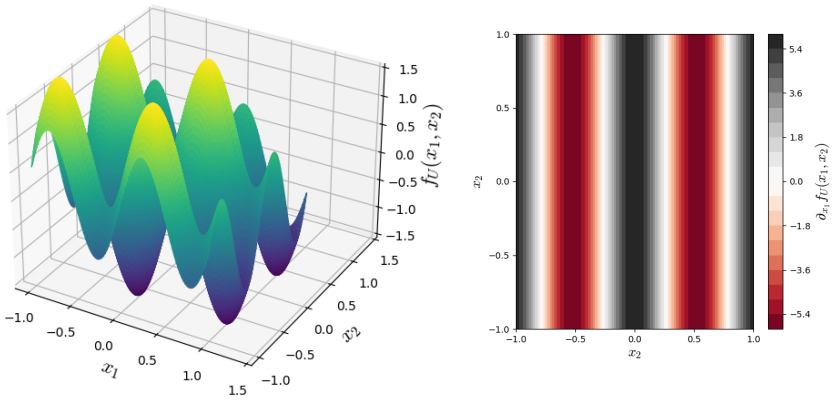

Figure 3: Left: f_U where f, U are as in Figure 2. Right: Image of the partial derivative $\partial_{x_1} f_U$ which is constant with respect to x_2. Therefore f_U can be decomposed as a sum of two univariate functions (see Theorem 2.5).

2.1. **Finest connected component decomposition**: Find $U \in O(d_1)$ such that $\mathcal{G}(g_U)$ provides the "finest" decomposition into connected components.

2.2. **Sparse component decomposition**: For each of the connected components, find an orthogonal matrix that transforms a connected component into one with the smallest number of edges.

3.1. Vertex minimization

First, we deal with the minimization problem (9). By (6), this is equivalent to the fact that most of the directional derivatives of f in direction u_j, $j = 1, \dots, d$ vanish, i.e.

$$\partial_j f_U = u_j^{\mathrm{T}} \nabla f(U\cdot) = 0. \tag{11}$$

Let

$$V_{\nabla f} \coloneqq \operatorname{span}\{\nabla f(x) : x \in \mathbb{B}_r\} \subseteq \mathbb{R}^d.$$

The following lemma allows us to describe vertices of the graph of f_U in terms of the gradient of f at a finite number of points.

Lemma 3.2. *Let $f \in C^1(\mathbb{B}_r)$ and $U \in O(d)$ and assume that $V_{\nabla f} = \operatorname{span}\{b_1, \dots, b_N\}$. Then $\partial_j f_U = 0$ if and only if $(U^{\mathrm{T}} b_n)_j = 0$ for all $n \in [N]$.*

Proof. By (12), we have that $\partial_j f_U = 0$ if and only if $u_j^{\mathrm{T}} v = 0$ for all $b \in V_{\nabla f}$. This is the case if and only if $u_j^{\mathrm{T}} b_n = (U^{\mathrm{T}} b_n)_j = 0$ for all $n \in [N]$. □

By the following proposition, the vertex minimizing transform in (9) can be obtained via singular value decomposition.

Proposition 3.3. *For $x^{(n)} \in \mathbb{B}_r$, $n \in [N]$, let*

$$B \coloneqq \left(\nabla f(x^{(1)}), \dots, \nabla f(x^{(N)})\right) = (b_1, \dots, b_N) \in \mathbb{R}^{d,N}$$

such that $V_{\nabla f} = \operatorname{span}\{b_1, \dots, b_N\}$. Then the left singular matrix $U \in O(d)$ of B is a minimizer of (9).

Proof. By (12), we have that $U \in O(d)$ is a minimizer of (9) if and only if $U^{\mathrm{T}} B$ has the largest number of zero rows. This is, if and only of U contains the largest number of columns which are orthogonal to all columns of B. This is exactly given by a left singular matrix of B. □

Once we have found a minimizer $U_{\mathcal{V}}$ by an SVD of B, say wlog.

$$U_{\mathcal{V}} = (\underbrace{u_1, \dots, u_{d_1}}_{U_1}, \underbrace{u_{d_1+1} \dots u_d}_{U_2}) \in O(d)$$

such that $\operatorname{span}\{u_1,\dots,u_{d_1}\} = \operatorname{span}\{b_1,\dots,b_N\}$ and $U_2^{\mathrm{T}} B = 0$, we know that $f_{U_\mathcal{V}}$ does only depend on the first d_1 components, so that we can restrict our attention to

$$g(\underbrace{x_1,\dots,x_{d_1}}_{x_{[d_1]}}) \coloneqq f_{U_\mathcal{V}}(x_1,\dots,x_{d_1},0\dots,0) = f_{U_\mathcal{V}}(x_1,\dots,x_d).$$

It follows immediately that

$$\nabla^2 f_{U_\mathcal{V}}(x) = \begin{pmatrix} H(x_{[d_1]}) & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad H(x_{[d_1]}) \coloneqq \nabla^2 g(x_{[d_1]}) \in \mathbb{R}^{d_1 \times d_1}.$$

Further, we have by (8) that

$$\nabla^2 f_{U_\mathcal{V}}(U_\mathcal{V}^{\mathrm{T}} x) = U_\mathcal{V}^{\mathrm{T}} \nabla^2 f(x) U_\mathcal{V} = \begin{pmatrix} H\left((U_\mathcal{V}^{\mathrm{T}} x)_{[d_1]}\right) & 0 \\ 0 & 0 \end{pmatrix}.$$

Thus, given $\nabla^2 f(x^{(n)})$, $n \in [N]$, we obtain the values $\nabla^2 g\left((U_\mathcal{V}^{\mathrm{T}} x^{(n)})_{[d_1]}\right)$ by

$$U_\mathcal{V}^{\mathrm{T}} \nabla^2 f(x^{(n)}) U_\mathcal{V} = \begin{pmatrix} \nabla^2 g\left((U_\mathcal{V}^{\mathrm{T}} x^{(n)})_{[d_1]}\right) & 0 \\ 0 & 0 \end{pmatrix}.$$

Remark 3.4. *The minimizing matrix $U_\mathcal{V}$ is not unique. However, it follows immediately that for any other left singular matrix U of B we would have*

$$\nabla^2 f_U(x) = P^{\mathrm{T}} \begin{pmatrix} W^{\mathrm{T}} H(x_{[d_1]}) W & 0 \\ 0 & 0 \end{pmatrix} P$$

with some $W \in O(d_1)$ and a permutation matrix P. In particular, the choice of the singular matrix does not influence the results in the next subsection.

3.2. Edge minimization

Next, we are interested in the minimization problem (10) taking only the d_1 relevant variables from the previous subsection into account. For simplicity of notation, we reset $d_1 \to d$ and $(U_\mathcal{V}^{\mathrm{T}} x^{(n)})_{[d_1]} \to x^{(n)}$, but keep the g. By the previous section, we can assume that $\nabla^2 g(x^{(n)})$, $n \in [N]$ are given. By (6), the edge minimization problem is equivalent to the fact that

$$\partial_{i,j} g_U = \partial_{j,i} g_U = u_i^{\mathrm{T}} \nabla^2 g(U\cdot) u_j = 0 \tag{12}$$

holds true for the largest number of pairs $\{i,j\}$. Let

$$V_{\nabla^2 g} \coloneqq \operatorname{span}\{\nabla^2 g(x) : x \in \mathbb{B}_r\}.$$

Clearly, the Hessians of g at $x \in \mathbb{R}^d$ belong to the space of symmetric $d \times d$ has therefore a spectral (eigenvalue) decomposition. Then, similarly as in Lemma 3.2, we observe the following relation.

Lemma 3.5. *Let $g \in C^2(\mathbb{B}_r)$ and $U \in O(d)$ and assume that $V_{\nabla^2 g} = \operatorname{span}\{H_1, \dots, H_N\}$. Then $\partial_{i,j} g_U = 0$ if and only if $(U^T H_n U)_{ij} = 0$ for all $n \in [N]$.*

Now we proceed in two steps.

3.2.1. Finest connected component decomposition

We start with the definition of graphs with „finest connected components".

Definition 3.6 (Graph with finest components)**.** For $d \in \mathbb{N}$, we denote by $\mathcal{N}_d$ the set of all descending sequences $a = (a_1, \dots, a_{d_a}) \in \mathbb{N}_{>0}^{k_a}$ satisfying $\sum_{i=1}^{d_a} a_i = d$. On $\mathcal{N}_d$ we introduce a preordering by $a \preceq b$ if there exists a partition $\cup_i I_i = [d_a]$ such that for all $i = 1, \dots, d_b$ it holds

$$b_i = \sum_{j \in I_i} a_j.$$

On the set G_d of all graphs with d vertices, we define a function $\phi : G_d \to \mathcal{N}_d$ as follows: for a graph $\mathcal{G} \in G_d$, let $\mathcal{G} = \cup_{i=1}^K \mathcal{G}_i$ be its finest decomposition into disjoint connected components, i.e., the $\mathcal{G}_i$ cannot be further decomposed into disjoint connected components, and define $\phi(\mathcal{G}) \coloneqq (|\mathcal{V}(\mathcal{G}_1)|, \dots, |\mathcal{V}(\mathcal{G}_K|) \in \mathcal{N}_d$. Then a preordering on G_d is given by $\mathcal{G} \preceq \tilde{\mathcal{G}}$ if and only if $\phi(\mathcal{G}) \preceq \phi(\tilde{\mathcal{G}})$. Now we consider the subset $\mathcal{S} \subset G_d$ and define the set of *graphs with finest connected components* in $\mathcal{S}$ to be its minimal elements with respect to $\preceq$.

We are interested in the minimal elements of

$$\{\mathcal{G}(g_U) : U \in O(d)\} \subset G_d,$$

i.e., we are asking for $U \in O(d)$ such that $\mathcal{G}(g_U)$ is a graph with finest connected components. Assuming that

$$V_{\nabla^2 g} = \operatorname{span}\{H_n \coloneqq \nabla^2 g(x^{(n)}) : n \in [N]\},$$

we see by Definition 2.4, that this is equivalent to determining $U \in O(d)$ such that H_n, $n \in [N]$ admit the finest joint block diagonalization.

Definition 3.7 (Finest joint block diagonalization.)**.** For $H_n \in \mathbb{R}^{d \times d}$, $n \in [N]$, let

$$U^T H_n U = \operatorname{diag}(H_n^{U,1}, \dots, H_n^{U,K}) \tag{13}$$

denote a joint block diagonalization of the H_n, $n \in [N]$ into K blocks $H_n^{U,k} \in \mathbb{R}^{d_k^U \times d_k^U}$ which cannot be further splitted. Define $\phi : O(d) \to \mathcal{N}_d$ by $\phi(U) \coloneqq d_U = (d_1^U, \dots, d_K^U)$. Then $U \preceq \tilde{U}$ if and only if $\phi(U) \preceq \phi(\tilde{U})$ is a preordering on $O(d)$. We say that (13) is the *finest joint block diagonalization* if U is minimal with respect to $\preceq$.

Remark 3.8. *In [37] the term finest joint block diagonalization is defined in terms of matrix $*$-algebras. It is shown Remark B.5 that both definitions are equivalent.*

To find such a finest block diagonalization, we apply a technique based on the following observation.

Proposition 3.9 (Joint block diagonalization)**.** *For $c \in \mathbb{R}^N$ randomly sampled from the uniform distribution on the unit sphere $\mathbb{S}^{N-1} \in \mathbb{R}^N$, let*

$$H \coloneqq \sum_{n=1}^{N} c_n H_n. \tag{14}$$

Let $U \in O(d)$ be a matrix that diagonalizes H. Then, with probability one, U provides a finest joint block diagonalization of the H_n, $n \in [N]$. More precisely, the set of all randomly sampled $c \in \mathbb{S}^{N-1}$ from the uniform distribution such that there exists $U \in O(d)$ that diagonalizes H but does not provides to a finest block diagonalization has measure zero.

For a proof see [38, Proposition 3]. By the proposition, we can just look for the spectral decomposition of a matrix H of the form (14). However, this strategy, is susceptible to noise. Therefore, Maehara and Murota [37] suggested an extension using the space of matrices that commute with the H_n, $n \in [N]$.

Proposition 3.10. *[37, Proposition 3.8] Let $A_1, \dots, A_M$ be a basis of the matrix space*

$$\{A \in M_d(\mathbb{R}) : [A, H_n] \coloneqq AH_n - H_nA = 0 \text{ for all } n \in [N]\}. \tag{15}$$

For a randomly drawn $c \in \mathbb{S}^{M-1}$ from the uniform distribution, let $U \in O(d)$ diagonalize

$$A \coloneqq \sum_{m=1}^{M} c_m A_m.$$

Then, with probability one U provides a finest block diagonalization of H_n, $n \in [N]$.

An important observation in [37] is that if the condition in (15) is relaxed to

$$\|[A, H_n]\|_{\mathrm{F}} \leq \delta \quad \text{for all } n \in [N] \tag{16}$$

with small $\delta > 0$, then a matrix $U \in O(d)$ that diagonalizes A, also jointly block diagonalizes the H_n, $n \in [N]$ up to a small error. Here $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. More precisely, consider the linear operators $T_n : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ with $A \mapsto [A, H_n]$ and set

$$T \coloneqq \sum_{n=1}^{N} T_n^{\mathrm{T}} T_n. \tag{17}$$

With the columnwise vectorization operator $\mathrm{vec} : R^{d\times d} \to \mathbb{R}^{d^2}$ and the Kronecker product $\otimes$ of matrices we have that

$$\mathrm{vec}\left(T_n(A)\right) = \boldsymbol{T}_n \mathrm{vec}(A), \quad \boldsymbol{T}_n \coloneqq H_n^{\mathrm{T}} \otimes I_d - I_d \otimes H_n$$

and $\|T_n(A)\|_{\mathrm{F}} = \|\boldsymbol{T}_n \mathrm{vec}(A)\|_2$. Let $V_1, \dots, V_K \in \mathbb{R}^{d\times d}$ be orthonormal eigenvectors of T with eigenvalues λ_k smaller than δ^2. Then, we obtain for

$$V \coloneqq \sum_{k=1}^{K} c_k V_k$$

and $v \coloneqq \mathrm{vec}(V)$ that

$$\begin{aligned}\sum_{n=1}^{N} \|[V^{\mathrm{T}}, H_n]\|_F^2 = \sum_{n=1}^{N} \|[V, H_n]\|_F^2 = \sum_{n=1}^{N} \|T_n(V)\|_F^2 = \sum_{n=1}^{N} \|\boldsymbol{T}_n v\|_2^2 \\ = \sum_{n=1}^{N} \langle \boldsymbol{T}_n^{\mathrm{T}} \boldsymbol{T}_n v, v\rangle = \langle \boldsymbol{T} v, v\rangle = \sum_{k=1}^{K} c_k^2 \lambda_k \le \delta^2 \sum_{k=1}^{K} c_k^2.\end{aligned}$$

Thus, for any $c \in \mathbb{S}^{K-1}$, the matrix $\frac{1}{2}(V + V^{\mathrm{T}})$ fulfills (16). Then we know by [37, Lemma 4.1], for $U \in O(n)$ satisfying $\frac{1}{2}U^{\mathrm{T}}(V + V^{\mathrm{T}})U = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$, that

$$(U^{\mathrm{T}} H_n U)_{ij} |\lambda_i - \lambda_j| \le \delta.$$

Consequently, $U^{\mathrm{T}} H_n U$ has an almost block diagonal structure and the blocks correspond to different eigenvalues. The resulting error-controlled version of the block diagonalization is stated as Algorithm 1.

Algorithm 1 Error-controlled block diagonalization

Input: Hessian matrices H_n, $n \in [N]$, error tolerance $\delta > 0$.
Find an orthonormal basis $V_1, \dots, V_K$ corresponding to eigenvalues smaller than δ^2 of T in (17).
Sample $c \in \mathbb{S}^{K-1}$ randomly from the uniform distribution on $\mathbb{S}^{K-1}$ and set $V \coloneqq \sum_{k=1}^{K} c_k V_k$.
Compute $U \in O(d)$ that diagonalizes $\frac{1}{2}(V + V^{\mathrm{T}})$.
Output: U

While we know by Proposition 3.10 that the above algorithm is guaranteed to find the finest block diagonalization with probability 1 for $\delta = 0$, we did not find a uniqueness statement in the literature. The following theorem contains the desired result. Its proof requires a deeper look into theory of matrix-$*$ algebras and can be found in Appendix B.

Theorem 3.11. *Let U_1, U_2 correspond to finest joint block diagonalizations of H_n, $n \in [N]$ with block sizes $d_1^{U_1}, \dots, d_{K_1}^{U_1}$ and $d_1^{U_2}, \dots, d_{K_2}^{U_2}$ and*

$$U_1^T H_n U_1 = \mathit{blockdiag}(H_n^{U_1,1}, \dots, H_n^{U_1,K_1}), \quad U_2^T H_n U_2 = \mathit{blockdiag}(H_n^{U_2,1}, \dots, H_n^{U_2,K_2}).$$

Then, it holds $K_1 = K_2 =: K$. Further, there exists a permutation σ of $[K]$ and matrices $V_k \in O(d_k^{U_1})$, $k \in [K]$ such that for all $k \in [K]$ and all $n \in [N]$ we have

i) $d_k^{U_1} = d_{\sigma(k)}^{U_2}$ *and*

ii) $H_n^{U_2,\sigma(k)} = V_k^{\mathrm{T}} H_n^{U_1,k} V_k$.

Note that neither $U_{\mathcal{V}}$ from (9) nor U obtained from the finest joint block diagonalization algorithm is unique. However, the only ambiguity is a possible conjugation of the blocks by orthogonal matrices as Remark 3.4 and Theorem 3.11 show.

3.2.2. Sparse component decomposition

Having found a joint approximate finest block decomposition of the Hessians H_n, $n \in [N]$, i.e.,

$$U^{\mathrm{T}} H_n U = \mathrm{blockdiag}(H_n^{U,1}, \dots, H_n^{U,K}), \quad H_n^{U,k} \in \mathbb{R}^{d_k^{U,k} \times d_k^{U,k}} \tag{18}$$

it remains to to enforce the sparsity in each block. In particular, we can treat the blocks separately which reduces the dimension of the problem drastically. Therefore, we consider for an arbitrary $k \in [K]$, the k-th blocks $B_n = H_n^{U,k}$ in (18) and set $d \coloneqq d_k^{U,k}$ now. Moreover, we agree that the diagonal elements of B_n, $n \in [N]$ are zero. The 0-"norm" $\|B\|_0$ of a $d \times d$ matrix B is the number of its nonzero components. We would like to find a matrix $U \in O(d)$ that minimizes

$$\ell(U) \coloneqq \big\| \frac{1}{N} \sum_{n=1}^{N} (U^{\mathrm{T}} B_n U)^2 \big\|_0 \tag{19}$$

where the square of the matrix is taken compomentwise. Unfortunately, it is well-known that the 0-"norm" minimization is NP-hard. Therefore, we will instead minimize the relaxed differntiable loss function

$$\ell_\varepsilon(U) = \sum_{\substack{i,j=1 \\ i \neq j}}^{d} \Big(\frac{1}{N} \sum_{n=1}^{N} (U^T B_n U)_{i,j}^2 + \varepsilon \Big)^{\frac{1}{2}}, \quad \varepsilon > 0. \tag{20}$$

The following proposition provides a sufficient condition that the relaxed version coincides with original one.

Proposition 3.12. *Let $B_n \in \mathbb{R}^{d \times d}$, $n \in [N]$ with zero diagonal components. If $U \in O(d)$ fulfills*

$$\ell_\varepsilon(U) = \max \Big\{ \dim(\mathrm{span}\{B_n, n \in [N]\}), \max_{n \in [N]} \mathrm{rank}(B_n) \Big\},$$

then U is a minimizer of (19).

Proof. By the properties of the rank, we get for arbitrary $M_1, \dots, M_N$ that

$$\Big\| \big(\|((M_1)_{ij}, \dots, (M_N)_{ij})\|_\infty \big)_{i,j=1}^{d} \Big\|_0 \geq \|M_n\|_0 \geq \mathrm{rank}(M_n), \quad n \in [N]$$

and consequently

$$\left\| \left(\|((M_1)_{ij}, \dots, (M_N)_{ij})\|_\infty \right)_{i,j=1}^d \right\|_0 \geq \max_{n\in[N]} \operatorname{rank}(M_n).$$

Now let $\mathcal{B} := \{e_{i,j} : \exists M_k \text{ such that } (M_n)_{i,j} \neq 0\}$. Then, $\operatorname{span}(\mathcal{B}) \supseteq \operatorname{span}\{M_n, n \in [N]\}$ and $\dim(\operatorname{span}(\mathcal{B})) = \left\| \left(\|((M_1)_{ij}, \dots, (M_N)_{ij})\|_\infty \right)_{i,j\in[d]^2} \right\|_0$, which gives

$$\left\| \left(\|((M_1)_{ij}, \dots (M_N)_{ij})\|_\infty \right)_{i,j\in[d]^2} \right\|_0 \geq \dim(\operatorname{span}\{M_n, n \in [N]\}).$$

Therefore, for $M_n \coloneqq U^T B_n U$ the above inequalities yield

$$\begin{aligned} \ell_\varepsilon(U) &= \left\| \left(\|((M_1)_{ij}, \dots (M_N)_{ij})\|_\infty \right)_{i,j=1}^d \right\|_0 \\ &\geq \max\left\{ \dim(\operatorname{span}\{M_n, n \in [N]\}), \max_{n\in[N]} \operatorname{rank}(M_n) \right\}. \end{aligned}$$

Since $U \in O(d)$, it holds that $\dim(\operatorname{span}\{B_n, n \in [N]\}) = \dim(\operatorname{span}\{U^T B_n U, n \in [N]\})$ and $\operatorname{rank}(B_n) = \operatorname{rank}(U^T B_n U)$ for all $n \in [N]$. Thus, we obtained a constant lower bound for $\ell_{0,e}(U)$,

$$\ell_\varepsilon(U) \geq \max\left\{ \dim(\operatorname{span}\{B_n, n \in [N]\}), \max_{n\in[N]} \operatorname{rank}(B_n) \right\}$$

and if it is reached, U has to be the global minimizer of ℓ_ε. □

4. Minimization of the loss ℓ_ε over $SO(d)$

This section deals with the efficient minimization of ℓ_ε in (20). In Subsection 4.1, we propose a gradient descent algorithm on $SO(d)$ and also consider the Landing algorithm for a regularized version of ℓ_ε. Subsection 4.2 contains convergence results. The efficient computation requires further a grid search to have an appropriate starting point for the algorithm. The corresponding results can be found in Appendix C, where we also have a closer look to the computational complexity.

4.1. Gradient descent and Landing algorithm over $SO(d)$

If $U \in O(d)$ is a minimizer of (20), then $\operatorname{blockdiag}(-1, I_{d-1})U \in SO(d)$ is a minimizer as well so that the optimization can be reduced to those over $SO(d)$. Now, we turn to the minimization of the loss function (20) on $SO(d)$ using gradient-based manifold optimization techniques. Let us start with a short overview of the basic definitions and properties related to $SO(d)$ as a manifold. For further details, we refer to [1, 11, 26]. Recall that $SO(d)$ is a Riemannian manifold. For each $U \in SO(d)$ the tangent space to $SO(d)$ at U is given by

$$T_U = \{V \in \mathbb{R}^{d\times d} : VU^T + UV^T = 0\} = \{AU : A \in \mathbb{R}^{d\times d}, A + A^T = 0\}.$$

For a function $F : \mathbb{R}^{d\times d} \to \mathbb{R}$ differentiable in the neighborhood of $SO(d)$, its Riemannian gradient $\operatorname{grad} F(U)$ at point U is given by an orthogonal projection of $\nabla F(U)$ onto tangent space T_U. For $SO(d)$ specifically, it has a closed-form

$$\operatorname{grad} F(U) := \tfrac{1}{2}\nabla F(U) - \tfrac{1}{2}U\nabla F(U)^T U.$$

We denote by $TSO(d) \coloneqq \cup_{U\in SO(d)}\{U\} \times T_U$ the tangent bundle of $SO(d)$. A retraction operator is a smooth map of manifolds $R : TSO(d) \to SO(d)$ which satisfies for all $U \in SO(d)$ that

- $R(U, 0) = U$,
- $DR_U(0) = \mathrm{Id}_{T_U}$ where $R_U = R|_{\{U\}\times T_U} : T_U \to SO(d)$ and D is the differential.

Note that the last properties ensures that for a line $\gamma(t) = tV$ in the tangent space T_U we have that $\frac{d}{dt}R(U, \gamma(t))|_{t=0} = V$ i.e. a retraction is a first-order approximation of the exponential map. There are various choices of retraction operators for $SO(d)$ such as the exponential map, the Cayley transform, and the Polar decomposition [26, Chapter 3.3], see also [21]. In this work, we use the QR-factorization [26] based retraction operator defined as

$$\mathrm{Retr}(U, V) = \mathrm{QR}(U - V),$$

where $\mathrm{QR}(U - V)$ denotes the orthogonal matrix of the QR-factorization of $U - V$.

Among many optimization methods on arbitrary Riemannian manifolds [26], and $O(d)$ [1] and $SO(d)$ [49] in particular, we focus on the gradient descent and Landing methods. Given an initial guess $U^{(0)}$ and step sizes $\{\nu_r\}_{r\geq 0}$, the sequence $\{U^{(r)}\}_{r\geq 0}$ constructed via Riemannian gradient descent is given by

$$U^{(r+1)} = \mathrm{Retr}\left(U^{(r)}, -\nu_r \operatorname{grad} \ell_\varepsilon(U^{(r)})\right), \quad r \geq 0. \tag{21}$$

If the algorithm converges to a fixed point if $U^{(r+1)} = U^{(r)}$, the condition $\operatorname{grad} \ell_\varepsilon(U^{(r)}) = 0$ is satisfied. Computing the retraction operator can be very time-consuming when the space dimension d is large. This is why, for $O(d)$, an alternative method called the Landing algorithm was developed in [1]. Instead of performing the minimization of ℓ_ε on $SO(d)$, it minimizes the penalized objective

$$\ell_\varepsilon(U) + \tfrac{\lambda}{4}\left\|I_d - UU^T\right\|_F^2,$$

over $\mathbb{R}^{d\times d}$ with the regularization parameter $\lambda > 0$. Note that the penalty is zero if and only if $U \in O(d)$. The Landing update step is given by

$$U^{(r+1)} = U^{(r)} - \nu_r \left(\operatorname{grad} \ell_\varepsilon(U^{(r)}) + \lambda \left[U^{(r)}(U^{(r)})^T - I_d\right] U^{(r)}\right), \quad r \geq 0, \tag{22}$$

where the second term pushes $U^{(r+1)}$ in the direction of the manifold. Note that in general $U^{(r+1)} \notin O(d)$ and only after a certain number of iterations $U^{(r+1)}$ comes close to $O(d)$. Therefore, the fixed point U of the Landing algorithm admits

$$\operatorname{grad} \ell(U) = -\lambda \left[UU^T - I_d\right] U.$$

If $U \in O(d)$, once again the gradient vanishes.

4.2. Convergence analysis

In the previous subsection, we observed that if the algorithms converge, the fixed point $U \in SO(d)$ admits $\operatorname{grad}(U) = 0$. Yet, it may not necessarily be global minima, unless the function ℓ_ε is geodesically convex [11, Corollary 11.18]. The latter is not true as the following theorem shows, see [51].

Theorem 4.1. *Let $F : SO(d) \to \mathbb{R}$ be a continuous geodesically convex function on $SO(d)$. Then, F is constant.*

Consequently, convergence to a global minimum of either of the methods depends on the initial guess $U^{(0)}$. In the following, we focus on the Riemannian gradient descent and derive its local sublinear convergence. As for the Landing algorithm, its local convergence to the global minimizer remains an open problem. We start with the following sublinear convergence result.

Theorem 4.2. *There exist $L > 0$ such that the sequence $\{U^{(r)}\}_{r\geq 0}$ generated by Riemannian gradient descent for (20) with step sizes $\nu_r = \nu \leq 1/L$ admits*

$$\ell_\varepsilon(U^{(r+1)}) - \ell_\varepsilon(U^{(r)}) \leq -\tfrac{1}{2L}\left\|\operatorname{grad}\ell_\varepsilon(U^{(r)})\right\|_F^2, \quad r \geq 0. \tag{23}$$

Furthermore, $\{\ell_\varepsilon(U^{(r)})\}_{r\geq 0}$ converges and $\operatorname{grad}\ell_\varepsilon(U^{(r)}) \to 0$ as $r \to \infty$.

Proof. We first note that $\ell_\varepsilon(U) > 0$ for all $U \in \mathbb{R}^{d\times d}$. Then, the convergence results follow from (23) via [12, Theorem 2.5]. Hence, we only need to show that (23) holds, which is, in turn, guaranteed by [12, Lemma 2.7]. Therefore, let us check that all the conditions of [12, Lemma 2.7] are satisfied. The manifold $SO(d)$ is a compact Riemannian submanifold of $\mathbb{R}^{d\times d}$. Furthermore, ℓ_ε is a continuously-differentiable function on a compact set, and, hence, it is Lipschitz-continuous [46, Corollary 6.4.20]. □

Theorem 4.2 only ensures convergence of manifold optimization to a fixed point and not necessarily to a global minimum, which is a typical outcome in nonconvex optimization. Let us define

$$\ell^* := \min_{U\in SO(d)} \ell_\varepsilon(U), \quad \mathcal{M} := \operatorname*{argmin}_{U\in SO(d)} \ell_\varepsilon(U), \quad \text{and} \quad \mathcal{F} := \{U \in SO(d) : \operatorname{grad}\ell_\varepsilon(U) = 0\}.$$

The next result ensures that by initializing Riemannian gradient descent in a neighborhood of the global minima, it will converge to it.

Theorem 4.3. *There exists a level set $\{U \in SO(d) : \ell_\varepsilon(U) \leq \ell^* + q^*\}$ with $q^* > 0$ that contains all global minimizers $\mathcal{M}$ and no other fixed points $\mathcal{F}$ of ℓ. Consequently, if $U^{(0)} \in \{U : \ell_\varepsilon(U) < \ell^* + q^*\}$, the sequence generated by Riemannian gradient descent with step sizes as in Theorem 4.2 converges to a global minimum of ℓ_ε.*

Proof. The main idea of the proof is to show that there exists an open neighborhood of $\mathcal{M}$ such that it does not contain other critical points $\mathcal{F}\backslash\mathcal{M}$ or in other words that $\mathcal{M}$ is

isolated from $\mathcal{F}\backslash\mathcal{M}$. Then, we will show that ℓ_ε on $\mathcal{F}\backslash\mathcal{M}$ is strictly larger than ℓ^* and it is possible to find suitable $q^* > 0$.

The proof is based on the Łojasiewicz inequality. We say that function ℓ_ε satisfies Łojasiewicz inequality at point $U \in SO(d)$ if there exists $\delta > 0$ such that for all $V \in SO(d)$, $\|U - V\|_F \leq \delta$ the inequality

$$\left\|\operatorname{grad} \ell_\varepsilon(V)\right\|_F \geq c|\ell_\varepsilon(U) - \ell_\varepsilon(V)|^{1-\zeta}$$

holds for some $c > 0$ and $\zeta \in [0, 1/2)$. By [45, Proposition 2.2 and Remark 1], for an analytic function on an analytic manifold[1], Łojasiewicz inequality is satisfied at every point on the manifold. $SO(d)$ is analytic and $\ell_\varepsilon \in C^\infty(r)$ is analytic as a superposition of a square root and a positive polynomial. Consequently, for each $U \in \mathcal{M}$ we can find an $\delta = \delta(U) > 0$ from Łojasiewicz inequality.

Next, we show by contradiction that there exists a threshold $q^* > 0$ such that $\ell_\varepsilon(U) > \ell^* + q^*$ on $\mathcal{F}\backslash\mathcal{M}$. Assume that the opposite holds and for every $q > 0$ there exists $U = U(q) \in \mathcal{F}\backslash\mathcal{M}$ such that $\ell_\varepsilon(U(q)) \leq \ell^* + q$. Then, let us consider a sequence $\{U(1/k)\}_{k\geq 1} \subseteq \mathcal{F}\backslash\mathcal{M}$. Since $\mathcal{F}\backslash\mathcal{M} \subseteq SO(d)$ and $SO(d)$ is compact, $\{U(1/k))\}_{k\geq 1}$ is bounded and there exists a convergent subsequence $\{U(1/k_j)\}_{j\geq 1}$ with limit U^*. By construction, $U^* \in \mathcal{F}$ and it admits

$$\ell^* \leq \ell_\varepsilon(U^*) = \ell_\varepsilon\left(\lim_{j\to\infty} U(1/k_j)\right) = \lim_{j\to\infty} \ell_\varepsilon\left(U(1/k_j)\right) \leq \lim_{j\to\infty} \ell^* + \frac{1}{k_j} = \ell^*,$$

so that $U^* \in \mathcal{M}$. However, from the convergence it follows that there exists $j_0 \in \mathbb{N}$ such that for all $j \geq j_0$ we have $\left\|U^* - U(1/k_j)\right\|_F < \delta(U^*)$. The Łojasiewicz inequality then gives

$$0 = \left\|\operatorname{grad} \ell_\varepsilon(U(1/k_j))\right\|_F \geq c|\ell_\varepsilon(U^*) - \ell_\varepsilon(U(1/k_j))|^{1-\zeta} \geq 0,$$

which is only possible if $U(1/k_j) \in \mathcal{M}$ and $\ell_\varepsilon(U(1/k_j)) = \ell^*$. Yet, it contradicts $U(1/k_j) \in \mathcal{C}\backslash\mathcal{M}$. Therefore, we obtain the contradiction. We also note that the scenario of a global minimum at infinity is impossible as $SO(d)$ is compact.

Thus, there exists $q^* > 0$ such that $\ell_\varepsilon(U) > \ell^* + q^*$ for all $U \in \mathcal{F}\backslash\mathcal{M}$. Consequently, all fixed points in the set $\mathcal{L} := \{U \in SO(d) : \ell_\varepsilon(U) \leq \ell^* + q^*\}$ are global minimizers of ℓ_ε. If $U^{(0)} \in \mathcal{L}$, by Theorem 4.2, the sequence $\{U^{(r)}\}_{r\geq 0}$ generated by Riemannian gradient descent will remain in $\mathcal{L}$ and converge to a point in $\mathcal{L} \cap \mathcal{F} = \mathcal{M}$. □

Theorem 4.3 has a flavor of standard convergence results based on Łojasiewicz-type inequalities, e.g., [2, Theorem 2.2] or [4, Theorem 3.3]. Although commonly a linear convergence is expected in a neighborhood of the accumulation point, a stronger assumption on the function is required. For instance, the exponent ζ in the Łojasiewicz inequality has to be $1/2$, which is known as the Polyak-Łojasiewicz inequality. For C^2 functions, it is equivalent to a number of other well-known conditions [44]. We refer the interested reader to Section 1.3 of [44] for an extensive literature overview on the topic. While there are some rules on the computation of Łojasiewicz exponents [33, 52], the establishment of local linear convergence in the case of (20) remains a topic of future research.

[1] According to Definition 2.7.1 in [30]

5. Numerical results

In the following section, we will first investigate the performance of the manifold optimization method for sparsifying a set of symmetric matrices. Furthermore by applying the three-step algorithm consisting of vertex minimization, finest connected component decomposition and sparse component decomposition in order to find an optimal $U \in SO(d)$ for a set of functions, we demonstrate that the gradients and Hessians admit the optimal sparsity patterns. For two test functions, we show that our algorithm finds $U \in SO(d)$ such that f_U exhibits the correct sparse ANOVA decomposition. [2]

To run the numerical experiments we have used the following libraries: *pytorch* [41], *numpy* [20] for the general computations, *RiemannianSGD* from *geoopt* [29] to perform the Riemannian gradient descent and *LandingSGD* from [1] to execute the Landing procedure. To compute the ANOVA-terms in section 5.2, we have used the *tntorch* library [50]. All computations were performed on a NVIDIA RTX A6000 48GB graphics card.

5.1. Manifold optimization on $SO(d)$ for jointly sparsifying a set of symmetric matrices

Creating jointly sparsifiable symmetric matrices

In order to test the feasibility of the manifold optimization methods for jointly sparsifying matrices we create sets of nonsparse matrices where we know that they can be made jointly sparse through conjugation with an orthogonal matrix. Let $J \subseteq [d] \times [d]$ be a set of jointly nonsparse entries. The matrices $\tilde{H}_n$, $1 \leq n \leq N$, are constructed by sampling $(\tilde{H}_n)_{i,j} = (\tilde{H}_n)_{j,i} \sim \mathrm{Unif}[-1,1]$ whenever $(i,j) \in J$ and setting the rest of the entries to zero. Then, with randomly drawn $R \in SO(d)$, we take

$$\mathcal{H}_R(J) := \left\{ H_n := R^T \tilde{H}_n R : 1 \leq n \leq N \right\}$$

as input data for edge minimization.

Remark 5.1. *For generically chosen $\tilde{H}_n$ and R we have that* $\dim \mathrm{span}(\mathcal{H}_R(J)) = \dim \mathrm{span}(\mathcal{H}(J)) = |J|$. *Thus by Lemma 3.12 we know that for any minimizer $U \in SO(d)$ of ℓ_ε it holds that $U^T \mathcal{H}_R(J) U$ has exactly $|J|$ nonzero entries.*

Creating noisy data. Additionally, to investigate whether the manifold optimization procedures are robust towards noise, we define the set

$$\mathcal{H}_R(J, \sigma) := \{ H + \zeta_H : H \in \mathcal{H}_R(J) \}$$

where each entry $(\zeta_H)_{i,j} \sim \mathcal{N}(0, \sigma^2)$ is drawn from a Gaussian distribution with mean 0 and variance σ^2.

[2] The code for our examples is available at https://github.com/fatima0111/Sparse-Function-Decomposition-via-Orthogonal-Transformation.

Details of the manifold optimization

Reducing complexity. To reduce the time complexity of the algorithm an approximate basis of the span($\mathcal{H}_R(J)$) resp. span($\mathcal{H}_R(J,\sigma)$) is computed via SVD. Choose a threshold τ and use SVD on $\mathcal{H}_R(J)$ resp. $\mathcal{H}_R(J,\sigma)$ to find all orthonormal vectors corresponding to singular values $s \geq \tau$. We denote the set of these orthonormal vectors by $H_R(J)$ resp. $H_R(J,\sigma)$. Note that even for $\mathcal{H}_R(J)$ the threshold τ is necessary since numerical errors may occur. Then, for $\mathcal{H} = \mathcal{H}_R(J)$ or $\mathcal{H} = \mathcal{H}_R(J,\sigma)$, we minimize

$$\ell_{\mathcal{H}}(U) \coloneqq \frac{1}{\sqrt{|\mathcal{H}|}} \sum_{(i,j)\in[d]^2} \Big(\sum_{H\in\mathcal{H}} (U^T H U)_{i,j}^2 + \varepsilon \Big)^{\frac{1}{2}}. \tag{24}$$

Note that we included the diagonal entries as they provide information on whether the function depends only linearly on the variables.

Random initialization. Since the loss function is not convex, an appropriate initialization is needed. The random initialization method consists of generating randomly 5 angles $\alpha_1, \cdots, \alpha_5 \in \Theta_d$ from the uniform distribution and to compute the corresponding rotation matrices $R_{\alpha_k}, k = 1, \cdots, 5$ according to (33). Each rotation matrix will be used to initialize the Riemannian gradient descent or Landing method for only $5 \cdot 10^3$ iterations to obtain the matrices $\hat{R}_{\alpha_k}$. Let

$$\ell_{\frac{1}{2},2}(U) := \frac{1}{\sqrt{|\mathcal{H}|}} \Big(\sum_{i,j=1}^{d} \big(\sum_{H\in\mathcal{H}} (U^T H U)_{i,j}^2 \big)^{\frac{1}{4}} \Big)^2$$

and use as the *random initializer*

$$R_{RI}^0 := \operatorname*{argmin}_{1\leq k\leq 5} \ell_{\frac{1}{2},2}\left(\hat{R}_{\alpha_k}\right).$$

Note that we used $\ell_{\frac{1}{2},2}$ here because it better approximates the sparsity norm $\|\cdot\|_{0,\infty}$.
Grid search initialization. For the grid search in Section C.1 and for $h \in \{1, 0.5, 0.25, 0.125, 0.1\}$ and $\Gamma(h)$ a grid as in (34) we let

$$R_h^0 := \operatorname*{argmin}_{R\in\Gamma(h)} \ell_{\frac{1}{2},2}(R).$$

For $d = 5$ we only use this method for $h = 1$ due to high computational complexity.

Convergence and performance of the manifold optimization methods

For $d = 2, 3, 4, 5$ we applied the manifold optimization methods for minimizing $\ell_{\mathcal{H}}$, see (24). For every dimension d we created 100 sets of jointly sparsifiable symmetric matrices $\mathcal{H}_R(J)$ resp. $\mathcal{H}_R(J,\sigma)$ where J was chosen randomly with the constraint that
Convergence. Note that the minimum of our objective (24) is unknown but we know that $R^T\mathcal{H}_R(J)R$ admits the optimal sparsity pattern. Let $U^{(r)}$ be the output of the manifold optimization at iteration r. We then compare $\ell_{\mathcal{H}}(U^{(r)})$ with $\ell_{\mathcal{H}}(R^T)$ Figure 4

	$d=2$	$d=3$	$d=4$	$d=5$
$1 \leq \lvert J \rvert \leq$	3	6	7	11

in order to analyze the convergence of the manifold optimization methods. It shows that especially for $d \in \{4, 5\}$ initialization with grid search admits a better convergence than random initialization which is in line with Corollary C.3. The Landing method and the Riemannian gradient descent quantitatively give similar results both in objective value and runtime, see Figure 4 and Table 8 in Appendix D.

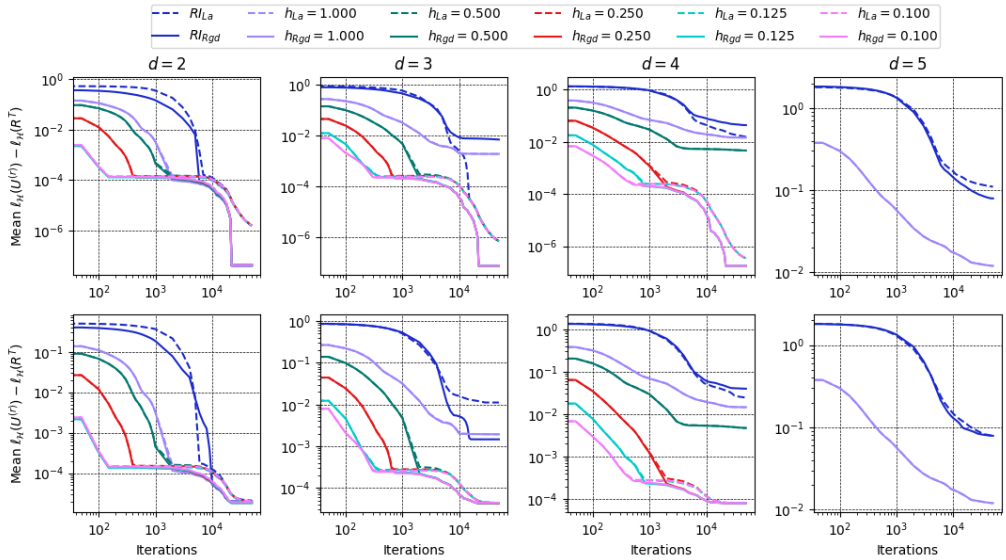

Figure 4: Mean optimality gap $\ell_{\mathcal{H}}\big(U^{(r)}\big) - \ell_{\mathcal{H}}\big(R^{\mathrm{T}}\big)$ over 100 experiments. Top: Noise-free matrices. Bottom: Noisy matrices. RI denotes random initialization and h uses the grid search with the corresponding grid density value. Subscripts Rgd and La stand for Riemannian gradient descent and Landing algorithm, respectively.

Sparsifying performance. Since ℓ_ε approximates ℓ, we will only obtain approximate sparsity and use an upper threshold instead of the $\|\cdot\|_{0,\infty}$ norm as follows. For $H \in \mathbb{R}^{d\times d}$, we define $|H| = (|H_{i,j}|)_{(i,j)\in[d]^2}$ and

$$\bar{H} := \frac{1}{n}\sum_{H\in\mathcal{H}} |H| \in \mathbb{R}^{d\times d}, \quad \big(\bar{H}_\eta\big)_{ij} := \begin{cases} 0, & \text{if } \bar{H}_{ij} \leq \eta, \\ \bar{H}_{ij}, & \text{otherwise,} \end{cases}$$

where $\mathcal{H} = U^T\mathcal{H}_R(J)U$. Thus we can compare the sparsity up to a level of η by

$$\chi(U,\eta) := \big\|\bar{H}_\eta\big\|_0 - |J|. \tag{25}$$

The quantity $\chi(U,\eta)$ measures how far from the optimal sparsity $U^T\mathcal{H}_R(J)U$ is, details about the results can be found in Appendix D.

To highlight the impact of the thresholding parameter η on the sparsity, we consider a failure ratio $\mathcal{R}$ as a mean

$$\text{Ratio } \mathcal{R} := \frac{1}{100} \sum_{j=1}^{100} \min\{1, \chi(U_j, \eta)\} \tag{26}$$

of 100 experiments with resulting matrices U_j. To evaluate how well our rotation matrices U obtained from noisy data sparsify the noiseless matrices, we apply the matrices U to $\mathcal{H}(J)$. The results are displayed in Figure 5 where the ratio of the algorithm not finding the optimal sparsity is plotted over the level η. It shows that initialization with grid search for small h is beneficial especially in dimension $d \in \{4, 5\}$ while the comparison of manifold optimization by means of Landing or Riemannian gradient descent is inconclusive.

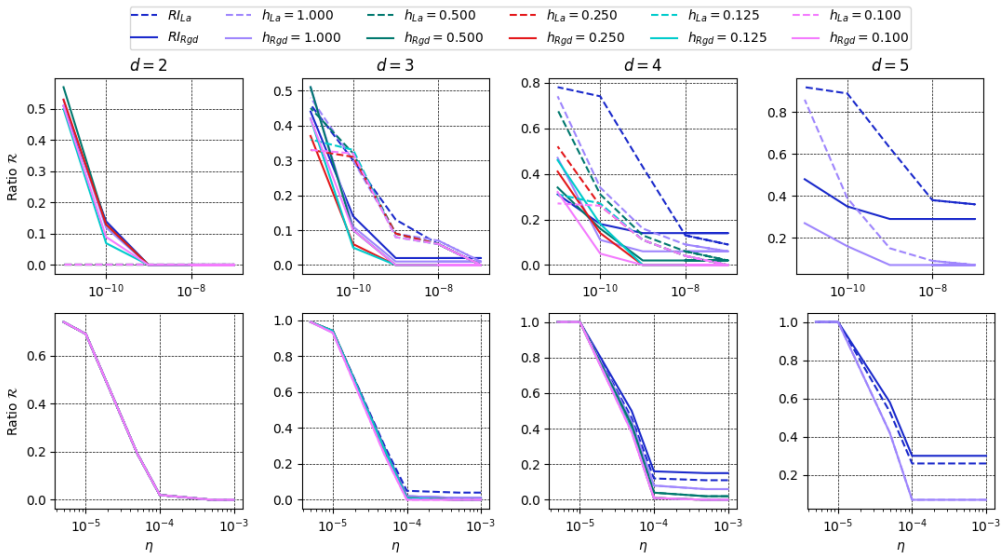

Figure 5: Failure ratio $\mathcal{R}$ (26) of suboptimal joint sparsity reconstructions for a given thresholding parameter η. First row: Clean data. Second row: Noisy data with additive random Gaussian noise $\mathcal{N}(0, \sigma), \sigma = 10^{-3}$.

5.2. Performance of the sparsifying algorithm on test functions

To illustrate the performance of our three-step algorithm from Section 3 consisting of vertex minimization, finest connected component decomposition and sparse component decomposition we apply it to 50 functions. Each test function $f : \mathbb{B}_r(d) \to \mathbb{R}$ with $10 \leq d \leq 15$ arguments is determined in the following way. First, a function $\tilde{f}$ with sparse $\mathcal{E}(\tilde{f})$ is constructed by randomly partitioning $[d]$ into connected components with 2 to 4 vertices. For each of the components, the edges (j, k) are picked at random. Then, we set $\tilde{f}$ as

$$\tilde{f}(x) := \sum_{(j,k) \in \mathcal{E}(\tilde{f})} c_{jk} g_{jk,1}(x_j) g_{jk,2}(x_k)$$

where the coefficients c_{jk} are drawn from the interval $[5, 20]$ and functions $g_{jk,1}, g_{jk,2}$ are drawn from the set

$$S := \left\{ x + t,\, x^t,\, \sqrt[3]{x^2 + t^2},\, \sin(tx),\, \cos(tx),\, e^{-(x-t)^2} : t \in \{1, 2, 3\} \right\}.$$

At last, a random rotation of the coordinates $f(x) = \tilde{f}(Rx)$ is applied. Moreover, to show that our three-step algorithm is robust towards additive noise, consider the noise function

$$N(x) := \frac{1}{2000} \sum_{\mu \in G_d} \exp\left(-\frac{1}{2}(x - \mu)^T Z^{-1}(x - \mu) \right), \tag{27}$$

where $G_d = \{(x_1, \ldots, x_d) : x_i \in \{-1/2,\ 3/2\}\}, Z = 0.5\mathbf{I}_d$. The noisy test functions are then

$$f_n := f + N.$$

We sample for each function $N = 100d$ points where we compute the gradient and Hessians in order to apply the algorithm. We compare the convergence of the different manifold optimization as follows. For a function f resp. f_n we apply the first two steps, the vertex minimization is done via SVD and the finest component decomposition (18) which always resulted in the correct number of relevant variables resp. block sizes. Then, for each collection of blocks $\mathcal{H}_k = \{H_n^{U,k} :\ 1 \leq n \leq N\}$, $1 \leq k \leq K$, we use either random initialization or grid search initialization followed by either Landing or Riemannian gradient descent for r steps on each block. Afterwards, we compute the loss (24) of each block, sum over all blocks

$$\bar{\ell}_{\mathcal{H}} := \sum_{k=1}^{K} \ell_{\mathcal{H}_k}. \tag{28}$$

and take the mean over 50 trials. The results shown in Figure 6 and Figure 7 indicate that grid search with smaller h converges faster and to lower function values than random initialization. For comparing the sparsifying performance, we compute the ratio of functions where the algorithm does not find the optimal sparsity of Hessians. It is smaller for grid search initialization with small h compared to random initialization as can be seen in the second column of Figure 6 and Figure 7. This is true for both the clean and the noisy functions. There is no substantial difference for the comparison of the Landing method and the Riemannian gradient descent.

5.3. ANOVA decomposition

To illustrate the dependency of the higher-order ANOVA terms on the first- and second-order derivatives discussed in Section 2 we will consider two 7-dimensional sparse additive test functions $f : \mathbb{B}_r(7) \to \mathbb{R}$ defined as

1. $\tilde{f}^1(x) = 5e^{-(x_1-1)^2}(x_4 + 1) + 7\sin(2x_1)x_7^3 + 10\cos(2x_2)(x_5 + 3)$ where the connected components of $\mathcal{G}(f)$ correspond to

$$\mathcal{B}^1 = \{\{x_1, x_4, x_7\}, \{x_2, x_5\}\}$$

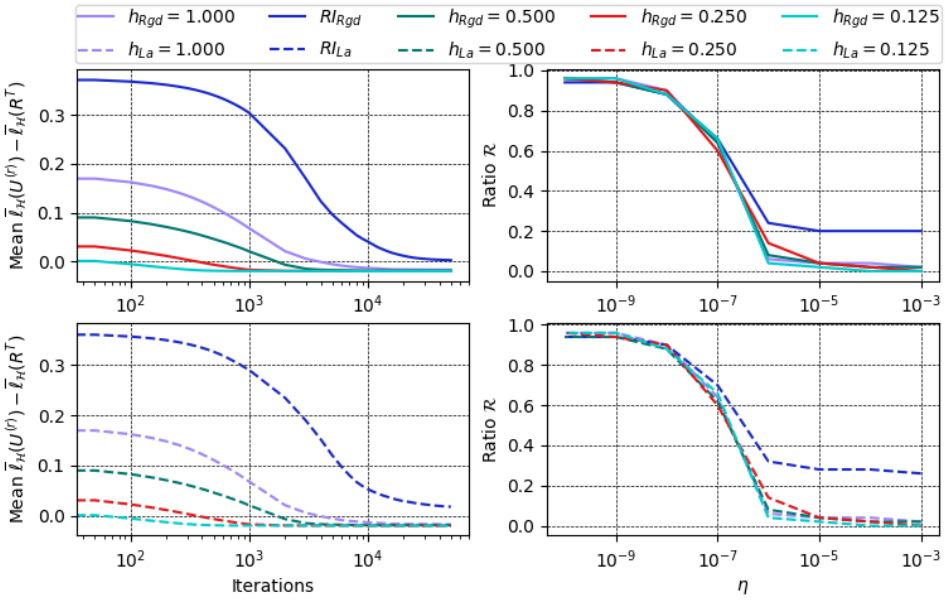

Figure 6: Mean optimality gap $\bar{\ell}_{\mathcal{H}}(U^{(r)}) - \bar{\ell}_{\mathcal{H}}(R^T)$ and failure ratio $\mathcal{R}$ for 50 noise-free functions, see (28) and (26) respectively. RI denotes random initialization and h uses the grid search with the corresponding grid density value. Subscripts Rgd and La stand for Riemannian gradient descent and Landing algorithm, respectively.

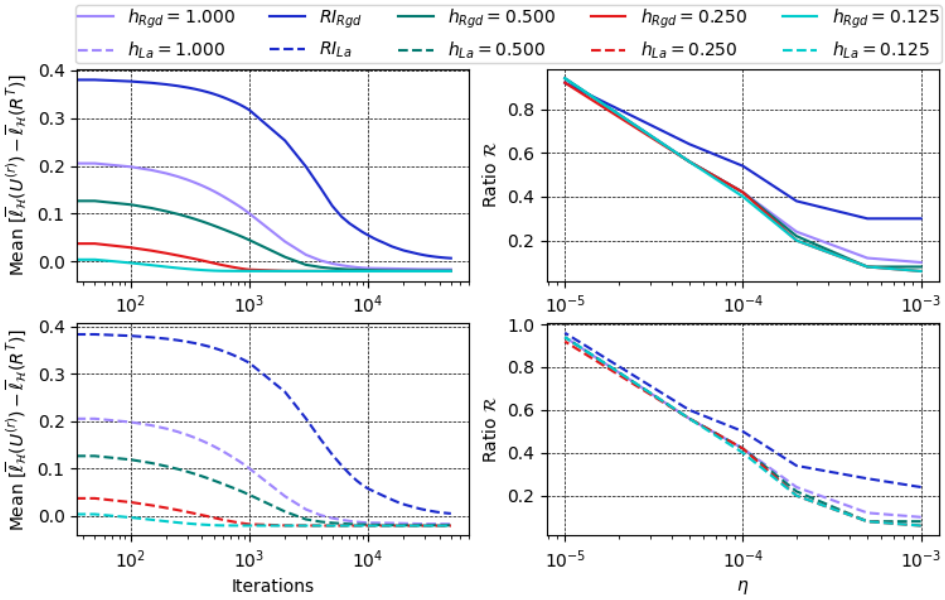

Figure 7: Reconstructions for noisy test functions f_n with notation as in Figure 6.

2. $\tilde{f}^2(x) = 5e^{-(x_1-1)^2}\cos(3x_4)+10x_1x_7^3+8\sin(x_2)\cos(x_7)+12\cos(2x_3)\sin(3x_5)+6x_5x_6$ with connected components corresponding to

$$\mathcal{B}^2 = \left\{\{x_1, x_2, x_4, x_7\}, \{x_3, x_5, x_6\}\right\}.$$

For each test function $\tilde{f}^i, i = 1, 2$ we randomly draw an orthogonal matrix $R^i \in O(7)$ such that $f^i := \tilde{f}^i_{R^i}$ are not sparse anymore. Additionally, we add the noise function N from equation (27) to obtain the noisy functions $f^{i,n} = f^i + N$. As Table 1 shows f^i has no first- or second-order vanishing ANOVA terms anymore.

$f =$	f^1		f^2	
$d_{sp} =$	1	2	1	2
$\mathcal{A}(f, d_{sp}, \infty)$	0.6788	0.2465	0.3421	0.5783
$\mathcal{A}(f, d_{sp}, 1)$	0.3614	0.0453	0.2102	0.1203

Table 1: Minimal values of the L^p-norm, $p \in \{1, \infty\}$ among all first- and second-order ANOVA terms, $\mathcal{A}(f, d_{sp}, p) := \min_{\substack{\mathbf{u}\subseteq[d] \\ |\mathbf{u}|=d_{sp}}} \left\|f_{\mathbf{u},A}\right\|_p$.

We apply our three-step algorithm with the Riemannian gradient descent and a grid-search initialization procedure with step-size $h = 1$ to the functions $f^i, f^{i,n}$ to obtain the orthogonal matrices $U_i, U_{i,n}$. Table 2 shows that the matrices $U_i, U_{i,n}$ both applied to f^i yield the correct number of first- and second-order derivatives vanish approximately in the sense that the empirical L_1 resp. L_∞ norm is small. Then by Proposition 2.6 all ANOVA terms $(f^i_{U_i})_{\mathbf{u}}, (f^i_{U_{i,n}})_{\mathbf{u}}$ have to be approximately zero for all $\mathbf{u} \subseteq [7]$ which contains a first- or second-order term that vanishes. This is confirmed empirically in Figures 8,9,10 where the ANOVA decomposition and its p-norm was computed empirically.

$f =$	$\tilde{f}^1$	$f^1_{U_1}$	$f^1_{U_{1,n}}$	$\tilde{f}^2$	$f^2_{U_2}$	$f^2_{U_{2,n}}$
$G(f, \infty)$	2	2	2	0	0	0
$G(f, 1)$	2	2	2	0	0	0
$H(f, \infty)$	18	18	17	16	15	14
$H(f, 1)$	18	18	18	16	15	15

Table 2: Number $G(f, p) := \left|\{i \in [d] : \|\partial_{\{i\}} f\|_p \leq 10^{-4}\}\right|$ of small first-order derivatives and number $H(f, p) := \left|\{\{i, j\} \subset [d] : i \neq j \text{ and } \|\partial_{\{i,j\}} f\|_p \leq 10^{-4}\}\right|$ of small second-order derivatives for the ground truth functions $\tilde{f}^i$ and for $f^i_{U_i}, f^i_{U_{i,n}}$ where $U_i, U_{i,n}$ is the output of the sparsifying algorithm.

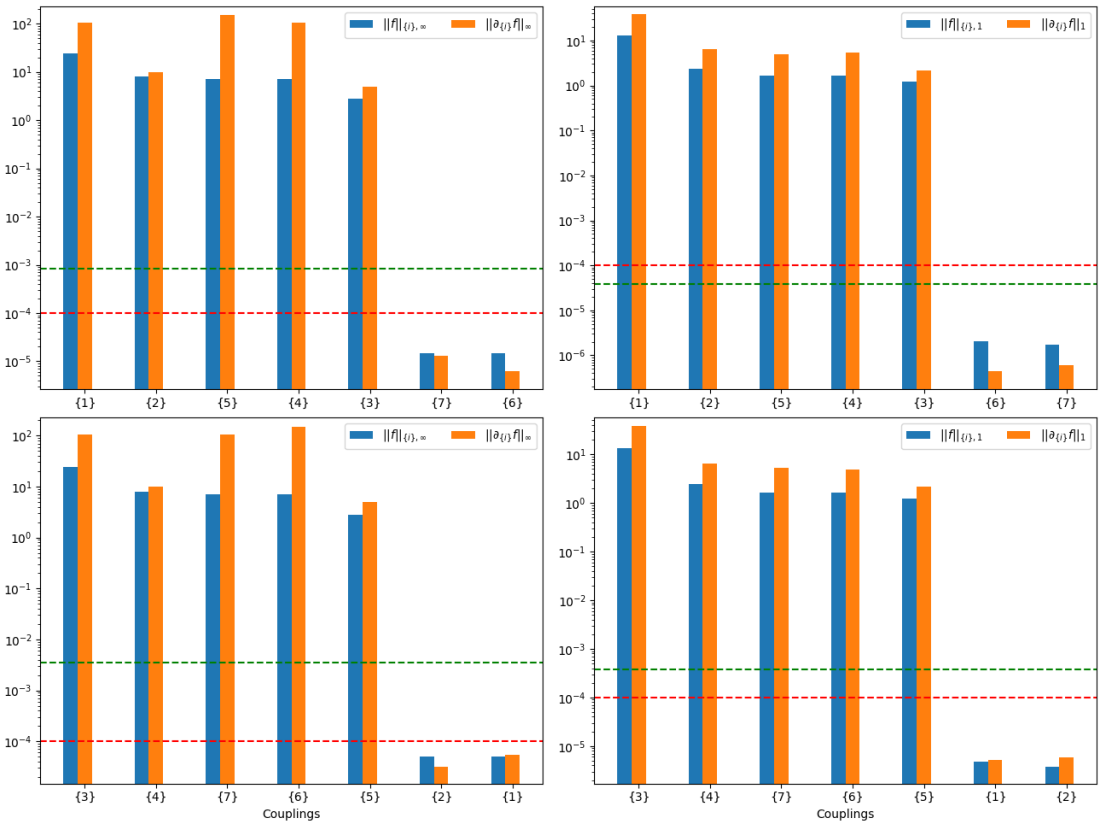

Figure 8: First row: $f = f^1_{U_1}$. Second row: $f = f^1_{U_{1,n}}$. First column: $p = \infty$. Second column: $p = 1$. Blue bars: $\|f\|_{\{i\},p} := \max_{\{i\}\subseteq \mathbf{u}\subseteq[7]} \left\|f_{\mathbf{u},A}\right\|_p$, in decreasing order. Orange bars: L^p-norm of the corresponding first-order partial derivative $\|\partial_{\{i\}} f\|_p$. The green dashed line represents $2^6 \max\{\|\partial_{\{i\}} f\|_p : \|\partial_{\{i\}} f\|_p \leq 10^{-4}\}$ corresponding to Theorem 2.6 and the red dashed line represents the truncation value 10^{-4}.

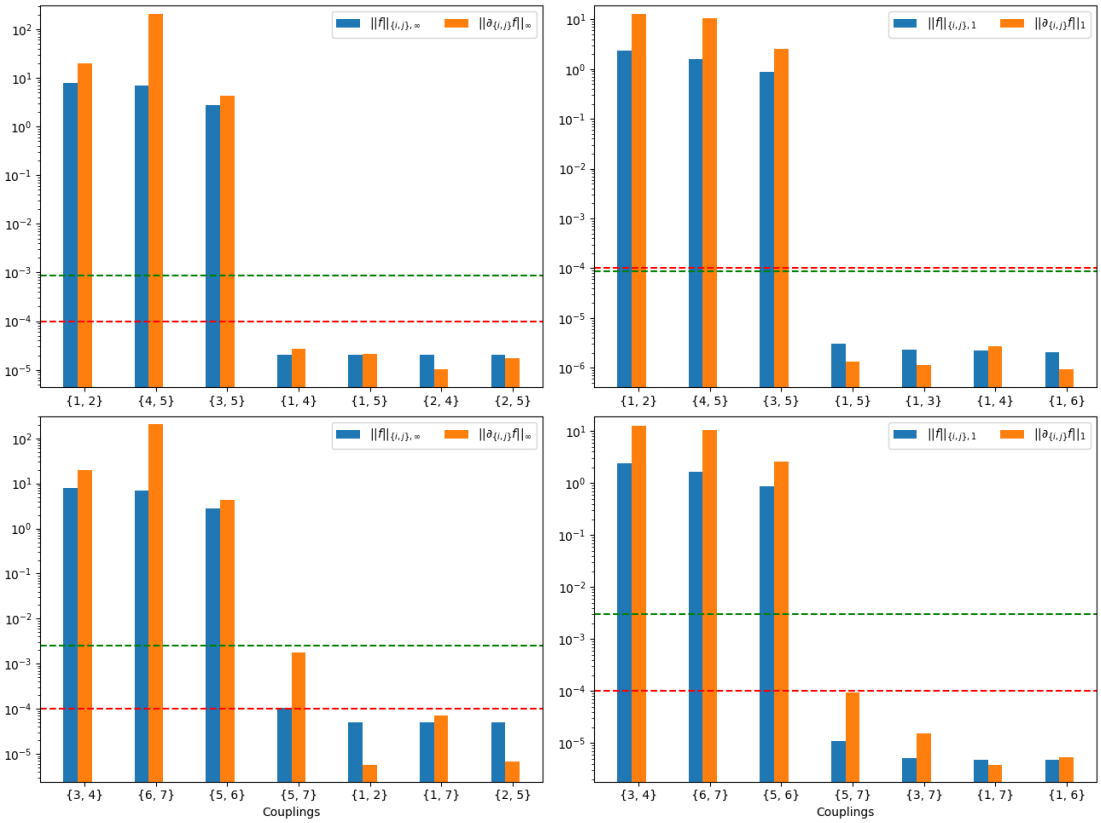

Figure 9: First row: $f = f^1_{U_1}$. Second row: $f = f^1_{U_{1,n}}$. First column: $p = \infty$. Second column: $p = 1$. Blue bars: $\|f\|_{\{i,j\},p} := \max_{\{i,j\}\subseteq\mathbf{u}\subseteq[7]} \left\|f_{\mathbf{u},A}\right\|_p$ in decreasing order. Orange bars: corresponding L^p-norm of the second-order partial derivative $\|\partial_{\{i,j\}}f\|_p$. Only the largest 8 terms are displayed. The green dashed line represents $2^5 \max\{\|\partial_{\{i,j\}}f\|_p : \|\partial_{\{i,j\}}f\|_p \leq 10^{-4}\}$ corresponding to Theorem 2.6 and the red dashed line represents the truncation value 10^{-4}.

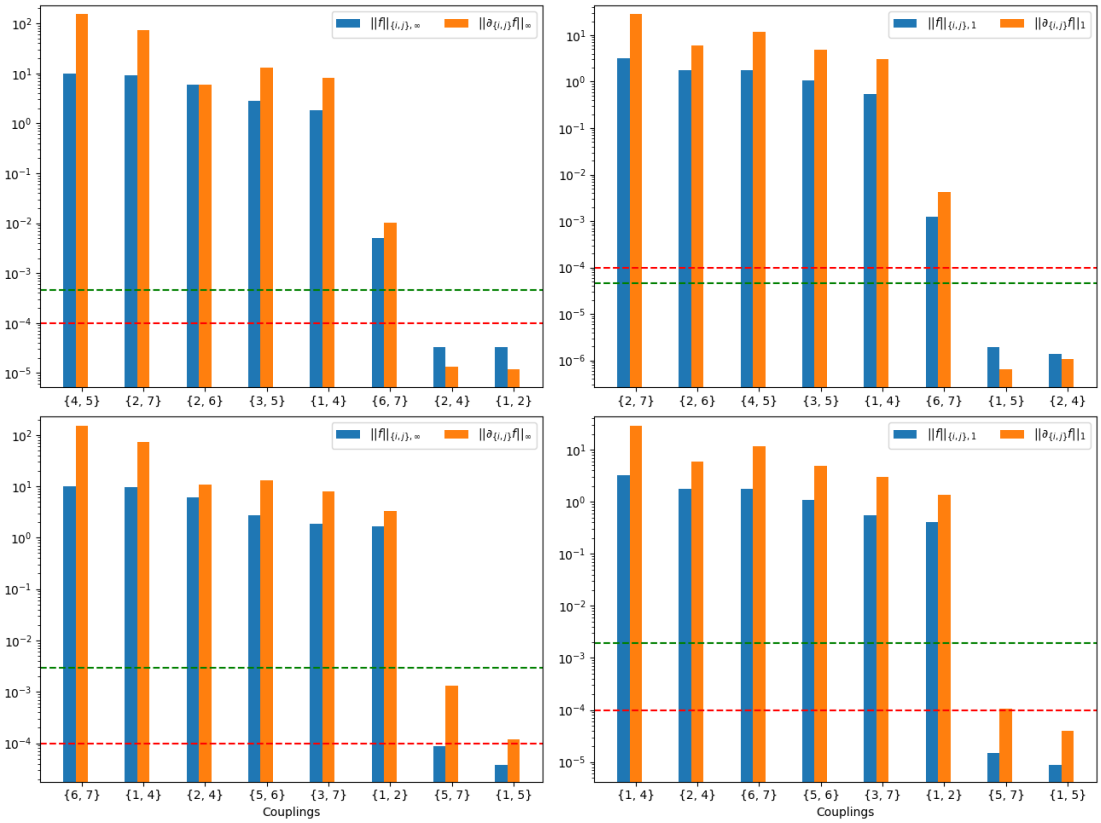

Figure 10: First row: $f = f^2_{U_2}$. Second row: $f = f^n_{U_{2,n}}$. First column: $p = \infty$. Second column: $p = 1$. Blue bars: $\|f\|_{\{i,j\},p} := \max_{\{i,j\}\subseteq\mathbf{u}\subseteq[7]} \left\|f_{\mathbf{u},A}\right\|_p$ in decreasing order. Orange bars: corresponding L^p-norm of the second-order partial derivative $\|\partial_{\{i,j\}}f\|_p$. Only the largest 8 terms are displayed. The green dashed line represents $2^5 \max\{\|\partial_{\{i,j\}}f\|_p : \|\partial_{\{i,j\}}f\|_p \leq 10^{-4}\}$ corresponding to Theorem 2.6 and the red dashed line represents the truncation value 10^{-4}.

Acknowledgement

F.B. acknowledges funding by the BMBF 01|S20053B project SAℓE. is gratefully acknowledged. Ch.W. acknowledges funding by the DFG within the SFB "Tomography Across the Scales" (STE 571/19-1, project number: 495365311).

References

[1] P. Ablin and G. Peyré. Fast and accurate optimization on the orthogonal manifold without retraction. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5636–5657. PMLR, 2022.

[2] P. A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

[3] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton. Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems*, 34:4699–4711, 2021.

[4] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[5] F. A. Ba and M. Quellmalz. Accelerating the sinkhorn algorithm for sparse multimarginal optimal transport via fast Fourier transforms. *Algorithms*, 15(9), 2022.

[6] J. Baldeaux and M. Gnewuch. Optimal randomized multilevel algorithms for infinite-dimensional integration on function spaces with ANOVA-type decomposition. *SIAM Journal on Numerical Analysis*, 52(3):1128–1155, 2014.

[7] F. Bartel, D. Potts, and M. Schmischke. Grouped transformations and regularization in high-dimensional explainable ANOVA approximation. *SIAM Journal on Scientific Computing*, 44(3):A1606–A1631, 2022.

[8] F. Beier, J. von Lindheim, S. Neumayer, and G. Steidl. Unbalanced multi-marginal optimal transport. *Journal of Mathematical Imaging and Vision*, 65(3):394–413, Jun 2023.

[9] J. Bilmes and C. Bartels. Graphical model architectures for speech recognition. *IEEE Signal Processing Magazine*, 22(5):89–100, 2005.

[10] J. Bilmes and G. Zweig. The graphical models toolkit: An open source software system for speech and time-series processing. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3916–IV–3919, 2002.

[11] N. Boumal. *An Introduction to Optimization on Smooth manifolds*. Cambridge University Press, 2023.

[12] N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 02 2018.

[13] R. E. Caflisch. Valuation of morgage backed securities using Brownian bridges to reduce effective dimension. *The Journal of Computational Finance*, 1:27–46, 1997.

[14] C.-H. Chang, R. Caruana, and A. Goldenberg. Node-gam: Neural generalized additive model for interpretable deep learning. *arXiv:2106.01613*, 2021.

[15] P. M. Cohn. *Basic Algebra: Groups, Rings and Fields*. Springer Science & Business Media, 2012.

[16] J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.

[17] J. Enouen and Y. Liu. Sparse interaction additive networks via feature interaction detection and sparse selection. *Advances in Neural Information Processing Systems*, 35:13908–13920, 2022.

[18] M. Griebel and M. Holtz. Dimension-wise integration of high-dimensional functions with applications to finance. *Journal of Complexity*, 26(5):455–489, 2010. SI: HDA 2009.

[19] M. Griebel, F. Y. Kuo, and I. H. Sloan. The ANOVA decomposition of a non-smooth function of infinitely many variables can have every term smooth. *Mathematics of Computations*, 86(306):1855–1876, 2017.

[20] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020.

[21] M. Hasannasab, J. Hertrich, S.Neumayer, G. Plonka, S. Setzer, and G. Steidl. Parseval proximal neural networks. *Journal of Fourier Analysis and Applications*, 26(59):1–36, 2020.

[22] M. Hefter, K. Ritter, and G. W. Wasilkowski. On equivalence of weighted anchored and ANOVA spaces of functions with mixed smoothness of order one in l_1 or l_∞. *Journal of Complexity*, 32(1):1–19, 2016.

[23] J. Hertrich, F. A. Ba, and G. Steidl. Sparse mixture models inspired by ANOVA decompositions. *Electronic Transactions on Numerical Analysis*, pages 142–168, 2022.

[24] F. Hickernell, I. Sloan, and G. Wasilkowski. On tractability of weighted integration over bounded and unbounded regions in $\mathbb{R}^s$. *Mathematics of Computation*, 73(248):1885–1901, 2004.

[25] F. J. Hickernell, I. H. Sloan, and G. W. Wasilkowski. The strong tractability of multivariate integration using lattice rules. In H. Niederreiter, editor, *Monte Carlo and Quasi-Monte Carlo Methods 2002*, pages 259–273, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[26] J. Hu, X. Liu, Z.-W. Wen, and Y.-X. Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8:199–248, 2020.

[27] A. Hurwitz. Über die Erzeugung der Invarianten durch Integration. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1897:71–2, 1897.

[28] M. G. Kapteyn, J. V. R. Pretorius, and K. E. Willcox. A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nature Computational Science*, 1(5):337–347, 2021.

[29] M. Kochurov, R. Karimov, and S. Kozlukov. Geoopt: Riemannian optimization in pytorch. *arXiv:2005.02819*, 2020.

[30] S. G. Krantz and H. R. Parks. *A primer of real analytic functions*. Birkhäuser Advanced Texts Basler Lehrbücher. Birkhäuser Boston, MA, 2nd ed. edition, 2002.

[31] F. Kuo, D. Nuyens, L. Plaskota, I. Sloan, and G. Wasilkowski. Infinite-dimensional integration and the multivariate decomposition method. *Journal of Computational and Applied Mathematics*, 326:217–234, 2017.

[32] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Wozniakowski. On decompositions of multivariate functions. *Mathematics of Computation*, 79(270):953–966, 2010.

[33] G. Li and T. K. Pong. Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, 18(5):1199–1232, 2018.

[34] T. Li, Y. Zhao, K. Yan, K. Zhou, C. Zhang, and X. Zhang. Probabilistic graphical models in energy systems: A review. *Building Simulation*, 15(5):699–728, 2022.

[35] L. Lippert and D. Potts. Variable transformations in combination with wavelets and ANOVA for high-dimensional approximation. *arXiv:2207.12826*, 2022.

[36] L. Lippert, D. Potts, and T. Ullrich. Fast hyperbolic wavelet regression meets ANOVA. *Numerische Mathematik*, 154(1):155–207, Jun 2023.

[37] T. Maehara and K. Murota. Algorithm for error-controlled simultaneous block-diagonalization of matrices. *SIAM Journal on Matrix Analysis and Applications*, 32(2):605–620, 2011.

[38] K. Murota, Y. Kanno, M. Kojima, and S. Kojima. A numerical algorithm for block-diagonal decomposition of matrix $*$-algebras with application to semidefinite programming. *Japan Journal of Industrial and Applied Mathematics*, 27(1):125–160, 2010.

[39] F. Nestler, M. Stoll, and T. Wagner. Learning in high-dimensional feature spaces using ANOVA-based fast matrix-vector multiplication. *Foundations of Data Science*, 4(3):423–440, 2022.

[40] T. J. O'Kane, D. Harries, and M. A. Collier. Bayesian structure learning for climate model evaluation. *Earth and Space Science Open Archive*, 2023.

[41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[42] D. Potts and M. Schmischke. Approximation of high-dimensional periodic functions with Fourier-based methods. *SIAM Journal on Numerical Analysis*, 59(5):2393–2429, 2021.

[43] D. Potts and M. Schmischke. Interpretable transformed ANOVA approximation on the example of the prevention of forest fires. *Frontiers in Applied Mathematics and Statistics*, 8:795250, 2022.

[44] Q. Rebjock and N. Boumal. Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions. *arXiv 2303.00096*, 2023.

[45] R. Schneider and A. Uschmajew. Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality. *SIAM Journal on Optimization*, 25(1):622–646, 2015.

[46] H. H. Sohrab. *Basic Real Analysis*. Birkhäuser New York, NY, 2nd ed. edition, 2014.

[47] S. Sullivant. *Algebraic Statistics*. Graduate Studies in Mathematics. American Mathematical Society, Providence, Rhode Island, 2018.

[48] Trilla-Fuertes et al. Multi-omics characterization of response to pd-1 inhibitors in advanced melanoma. *Cancers*, 15(17), 2023.

[49] K. Usevich, J. Li, and P. Comon. Approximate matrix and tensor diagonalization by unitary transformations: Convergence of jacobi-type algorithms. *SIAM Journal on Optimization*, 30(4):2998–3028, 2020.

[50] M. Usvyatsov, R. Ballester-Ripoll, and K. Schindler. tntorch: Tensor network learning with PyTorch. *Journal of Machine Learning Research*, 23(208):1–6, 2022.

[51] S.-T. Yau. Non-existence of continuous convex functions on certain riemannian manifolds. *Mathematische Annalen*, 207(4):269–270, 1974.

[52] P. Yu, G. Li, and T. K. Pong. Kurdyka-Łojasiewicz Exponent via Inf-projection. *Foundations of Computational Mathematics*, 22(4):1171–1217, 2022.

A. Proof of Theorem 2.6

To prove the theorem, we need an auxiliary lemma. The anchored decomposition can also be described via integrals of mixed partial derivatives as mentioned in [32, Example 2.3] and for Sobolev spaces in [22, Lemma 6].

Lemma A.1. *Let $f : D \to \mathbb{R}$ such that $f \in C^{|\mathbf{u}|}(D)$. Then the summands in anchored decomposition* (5) *can be represented as*

$$f_{\mathbf{u},c} = \int_{D_{\mathbf{u}}} \partial_{t_{\mathbf{u}}} f(t_{\mathbf{u}}, c_{[d]\setminus\mathbf{u}}) M(t_{\mathbf{u}}, x_{\mathbf{u}}) \,\mathrm{d}\lambda_{\mathbf{u}}(t_{\mathbf{u}}),$$

where

$$M_{\mathbf{u}}(t_{\mathbf{u}}, x_{\mathbf{u}}) \coloneqq \prod_{i\in\mathbf{u}} M_i(t_i, x_i), \quad M_i(t_i, x_i) \coloneqq \begin{cases} 1, & c_i < t_i < x_i, \\ -1, & x_i < t_i < c_i, \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. We show the assertion by induction over $|\mathbf{u}|$ for $\mathbf{u} \subseteq [d]$. Without loss of generality assume that $c_i < x_i$ for all $i \in [d]$. If $\mathbf{u} = \emptyset$ then the assertion follows directly from the definition of the projection operator. Let $\emptyset \neq \mathbf{v} \subset [d]$ and let $j \in [d] \setminus \mathbf{v}$ be arbitrary but fixed. Define $c^{t_j} \coloneqq (c_1, \cdots, c_{j-1}, t_j, c_{j+1}, \cdots, c_d)$. Assume that the assertion holds for $\mathbf{v}$. We show that it also holds for $\mathbf{u} = \mathbf{v} \cup \{j\}$. The Leibniz integral rule [19, Theorem 6.2] implies on the one hand

$$\begin{aligned} h_{\mathbf{u},c}(x_{\mathbf{u}}) &= \int_{D_{\mathbf{u}}} \partial_j \partial_{\mathbf{v}} f(t_{\mathbf{v}}, t_j, c_{[d]\setminus\mathbf{u}}) M_{\mathbf{v}}(t_{\mathbf{v}}, x_{\mathbf{v}}) M_j(t_j, x_j) \,\mathrm{d}\lambda_{\mathbf{v}}(t_{\mathbf{v}}) \,\mathrm{d}\lambda_j(t_j) \\ &= \int_{I_j} \partial_j \left(\int_{D_{\mathbf{v}}} \partial_{\mathbf{v}} f(t_{\mathbf{v}}, c^{t_j}_{[d]\setminus\mathbf{v}}) M_{\mathbf{v}}(t_{\mathbf{v}}, x_{\mathbf{v}}) \,\mathrm{d}\lambda_{\mathbf{v}}(t_{\mathbf{v}}) \right) M_j(t_j, x_j) \,\mathrm{d}\lambda_j(t_j), \end{aligned}$$

and on the other hand the induction step yields

$$\begin{aligned}
h_{\mathbf{u},c}(x_{\mathbf{u}}) &= \int_{-1}^{1} \partial_j \underbrace{\left(\Big(\prod_{i\in\mathbf{v}} \big(\mathrm{Id} - P_{i,c^{t_j}}\big)\Big) P_{[d]\setminus\mathbf{v},c^{t_j}} f(x) \right)}_{=\tilde{f}(x_{\mathbf{v}}, c^{t_j}_{[d]\setminus\mathbf{v}})} M_j(t_j, x_j)\, \mathrm{d}\lambda_j(t_j) \\
&= \int_{c_j}^{x_j} \partial_j \tilde{f}(x_{\mathbf{v}}, c^{t_j}_{[d]\setminus\mathbf{v}}) \mathrm{d}\lambda_j(t_j) = \int_{c_j}^{x_j} \partial_j \tilde{f}(x_{\mathbf{v}}, t_j, c_{[d]\setminus\{j\}})\, \mathrm{d}\lambda_j(t_j) \\
&= \tilde{f}(x_{\mathbf{v}}, x_j, c_{[d]\setminus\{j\}}) - \tilde{f}(x_{\mathbf{v}}, c_j, c_{[d]\setminus\{j\}}) = (\mathrm{Id} - P_{j,c})\tilde{f}(x_{\mathbf{v}}, x_j, c_{[d]\setminus\{j\}}) \\
&= (\mathrm{Id} - P_{j,c})\tilde{f}(x_{\mathbf{v}}, c^{x_j}_{[d]\setminus\mathbf{v}}) = (\mathrm{Id} - P_{j,c})\Big(\prod_{i\in\mathbf{v}} \big(\mathrm{Id} - P_{i,c^{x_j}}\big)\Big) P_{[d]\setminus\mathbf{v},c^{x_j}} f(x).
\end{aligned}$$

Since $j \notin \mathbf{v}$, we have

$$\prod_{i\in\mathbf{v}} \big(\mathrm{Id} - P_{i,c^{x_j}}\big) = \prod_{i\in\mathbf{v}} \big(\mathrm{Id} - P_{i,c}\big),$$

and the definition of the projection operator and of c^{x_j} imply that

$$P_{[d]\setminus\mathbf{v},c^{x_j}} f(x) = f(x_{\mathbf{v}}, c^{x_j}_{[d]\setminus\mathbf{v}}) = f(x_{\mathbf{v}}, x_j, c_{[d]\setminus(\mathbf{v}\cup\{j\})}) = P_{[d]\setminus\mathbf{u},c} f(x)$$

which concludes the proof. □

Proof of Theorem 2.6. i) By Lemma A.1, we have for sufficiently smooth f that

$$\|f_{\mathbf{v},c}\|_\infty = \left\| \int_{D_{\mathbf{v}}} \partial_{t_{\mathbf{v}}} f(t_{\mathbf{v}}, c_{[d]\setminus\mathbf{v}}) M(t_{\mathbf{v}}, x_{\mathbf{v}})\, \mathrm{d}\lambda_{\mathbf{v}}(t_{\mathbf{v}}) \right\| \leq \|\partial_{t_{\mathbf{v}}} f\|_\infty \lambda_{\mathbf{v}}(D_{\mathbf{v}}).$$

For $\mathbf{w} \subseteq [d]$ let $f^{x_{\mathbf{w}}} : \mathbb{R}^{d-|\mathbf{w}|} \to \mathbb{R}$ be the family of functions defined by $f^{x_{\mathbf{w}}}(x_{[d]\setminus\mathbf{w}}) := f(x_{\mathbf{w}}, x_{[d]\setminus\mathbf{w}})$. For $\mathbf{w} := \mathbf{u} \setminus \mathbf{v}$ we have that

$$\begin{aligned}
f_{\mathbf{u},c} &= \prod_{i\in\mathbf{u}} (\mathrm{Id} - P_{i,c}) P_{[d]\setminus\mathbf{u},c} f = \prod_{i\in\mathbf{w}} (\mathrm{Id} - P_{i,c}) \prod_{j\in\mathbf{v}} (\mathrm{Id} - P_{j,c}) P_{([d]\setminus\mathbf{w})\setminus\mathbf{v}} f \qquad (29) \\
&= \prod_{i\in\mathbf{w}} (\mathrm{Id} - P_{i,c}) f^{x_{\mathbf{w}}}_{\mathbf{v}}(x_{\mathbf{v}}).
\end{aligned}$$

Combining with $\Big|(\mathrm{Id} - P_{i,c}) g(x_{[d]\setminus i})\Big| \leq 2\|g\|_\infty$ for functions g, we obtain

$$\begin{aligned}
\|f_{\mathbf{u},c}\|_\infty &= \left\| \prod_{i\in\mathbf{w}} (\mathrm{Id} - P_{i,c}) f^{x_{\mathbf{w}}}_{\mathbf{v}}(x_{\mathbf{v}}) \right\|_\infty \leq 2^{|\mathbf{w}|} \sup_{x_{\mathbf{w}}\in D_{\mathbf{w}}} \{\|f^{x_{\mathbf{w}}}_{\mathbf{v}}\|_\infty\} \\
&\leq 2^{|\mathbf{u}|-|\mathbf{v}|} \|\partial_{x_{\mathbf{v}}} f\|_\infty \lambda_{\mathbf{v}}(D_{\mathbf{v}}).
\end{aligned}$$

ii) This follows from part i) and the fact that we have by by Proposition 2.3 for the ANOVA summands

$$f_{\mathbf{u},A} = \int_D f_{\mathbf{u},c}\,\mathrm{d}\lambda(c).$$

iii) For $\mathbf{w} \subseteq [d]$, let $f^{x_\mathbf{w}}$ be defined as above and for any disjoint subsets $\mathbf{s}, \mathbf{t} \subseteq [d]$ let $f^{x_\mathbf{s}, c_\mathbf{t}}(x_{[d]\setminus(\mathbf{s}\cup\mathbf{t})}) = f(x_\mathbf{s}, c_\mathbf{t}, x_{[d]\setminus(\mathbf{s}\cup\mathbf{t})})$. For $\mathbf{w} \subseteq [d]$ and any commuting linear operators $a_j, b_j, j \in \mathbf{w}$ the following holds

$$\prod_{j\in\mathbf{w}}(a_j - b_j) = \sum_{\mathbf{s}\subseteq\mathbf{w}} \left(\prod_{j\in\mathbf{s}} a_j\right)\left(\prod_{j\in\mathbf{w}\setminus\mathbf{s}} b_j\right). \tag{30}$$

If we define $\mathbf{w} := \mathbf{u} \setminus \mathbf{v}$ then equations (29), (30) and (4) imply that

$$f_{\mathbf{u},c} = \prod_{i\in\mathbf{w}}(\mathrm{Id} - P_{i,c}) f^{x_\mathbf{w}}_{\mathbf{v},c_{[d]\setminus\mathbf{w}}} = \sum_{\mathbf{s}\subseteq\mathbf{w}} (-1)^{|\mathbf{w}|-|\mathbf{s}|} P_{\mathbf{w}\setminus\mathbf{s}} f^{x_\mathbf{w}}_{\mathbf{v},c_{[d]\setminus\mathbf{w}}} = \sum_{\mathbf{s}\subseteq\mathbf{w}} (-1)^{|\mathbf{w}|-|\mathbf{s}|} f^{x_\mathbf{s}, c_{\mathbf{w}\setminus\mathbf{s}}}_{\mathbf{v},c_{[d]\setminus\mathbf{w}}}.$$

By proposition 2.3 the ANOVA term fulfills

$$\begin{aligned}
\left\|f_{\mathbf{u},A}\right\|_1 &\leq \int_{D_\mathbf{u}} \int_D \left|f_{\mathbf{u},c}\right| d\lambda(c) dx_\mathbf{u} = \int_D \int_{D_\mathbf{u}} \left|\sum_{\mathbf{s}\subseteq\mathbf{w}} (-1)^{|\mathbf{w}|-|\mathbf{s}|} f^{x_\mathbf{s}, c_{\mathbf{w}\setminus\mathbf{s}}}_{\mathbf{v},c_{[d]\setminus\mathbf{w}}}\right| d\lambda_\mathbf{u}(x_\mathbf{u}) d\lambda(c) \\
&\leq \sum_{\mathbf{s}\subseteq\mathbf{w}} \lambda(D_\mathbf{w}) \int_{D_{[d]\setminus\mathbf{s}}} \int_{D_{\mathbf{v}\cup\mathbf{s}}} \left|f^{x_\mathbf{s}, c_{\mathbf{w}\setminus\mathbf{s}}}_{\mathbf{v},c_{[d]\setminus\mathbf{w}}}\right| d\lambda_{\mathbf{v}\cup\mathbf{s}}(x_{\mathbf{v}\cup\mathbf{s}}) d\lambda_{[d]\setminus\mathbf{s}}(c_{[d]\setminus\mathbf{s}}) \\
&= \lambda(D_\mathbf{w}) \sum_{\mathbf{s}\subseteq\mathbf{w}} \int_{D_{[d]\setminus\mathbf{w}}} \int_{D_\mathbf{v}} \int_{D_\mathbf{w}} \left|f^{x_\mathbf{w}}_{\mathbf{v},c_{[d]\setminus\mathbf{w}}}\right| d\lambda_\mathbf{w}(x_\mathbf{w}) d\lambda_\mathbf{v}(x_\mathbf{v}) d\lambda_{[d]\setminus\mathbf{w}}(c_{[d]\setminus\mathbf{w}}) \\
&= 2^{|\mathbf{w}|}\lambda(D_\mathbf{w}) \underbrace{\int_{D_{[d]\setminus\mathbf{w}}} \int_{D_\mathbf{v}} \int_{D_\mathbf{w}} \left|f^{x_\mathbf{w}}_{\mathbf{v},c_{[d]\setminus\mathbf{w}}}\right| d\lambda_\mathbf{w}(x_\mathbf{w}) d\lambda_\mathbf{v}(x_\mathbf{v}) d\lambda_{[d]\setminus\mathbf{w}}(c_{[d]\setminus\mathbf{w}})}_{=:I}.
\end{aligned}$$

Finally, the triangle inequality yields

$$\begin{aligned}
I &= \int_{D_{[d]\setminus\mathbf{w}}} \int_{D_\mathbf{u}} \left|\int_{D_\mathbf{v}} \partial_{t_\mathbf{v}} f(t_\mathbf{v}, c_{([d]\setminus\mathbf{w})\setminus\mathbf{v}}, x_\mathbf{w}) M(t_\mathbf{v}, x_\mathbf{v}) d\lambda_\mathbf{v}(t_\mathbf{v})\right| d\lambda(x_\mathbf{u}) d\lambda(c_{[d]\setminus\mathbf{w}}) \\
&\leq \int_{D_{[d]\setminus\mathbf{w}}} \int_{D_\mathbf{u}} \int_{D_\mathbf{v}} \left|\partial_{t_\mathbf{v}} f(t_\mathbf{v}, c_{([d]\setminus\mathbf{w})\setminus\mathbf{v}}, x_\mathbf{w})\right| d\lambda_\mathbf{v}(t_\mathbf{v}) d\lambda(x_\mathbf{u}) d\lambda(c_{[d]\setminus\mathbf{w}}) \\
&= \lambda_\mathbf{v}(D_\mathbf{v}) \int_{D_{[d]\setminus\mathbf{w}}} \int_{D_\mathbf{w}} \int_{D_\mathbf{v}} \left|\partial_{t_\mathbf{v}} f(t_\mathbf{v}, c_{([d]\setminus\mathbf{w})\setminus\mathbf{v}}, x_\mathbf{w})\right| d\lambda_\mathbf{v}(t_\mathbf{v}) d\lambda(x_\mathbf{w}) d\lambda(c_{[d]\setminus\mathbf{w}}) \\
&= \lambda^2_\mathbf{v}(D_\mathbf{v}) \int_{D_{[d]\setminus\mathbf{u}}} \int_{D_\mathbf{w}} \int_{D_\mathbf{v}} \left|\partial_{t_\mathbf{v}} f(t_\mathbf{v}, c_{[d]\setminus\mathbf{u}}, x_\mathbf{w})\right| d\lambda_\mathbf{v}(t_\mathbf{v}) d\lambda(x_\mathbf{w}) d\lambda(c_{[d]\setminus\mathbf{u}}) \\
&= \|\partial_{t_\mathbf{v}} f\|_1 \lambda^2_\mathbf{v}(D_\mathbf{v}).
\end{aligned}$$

By combining all together, we obtain

$$\left\|f_{\mathbf{u},A}\right\|_1 \leq 2^{|\mathbf{u}|-|\mathbf{v}|}\|\partial_{x_\mathbf{v}} f\|_1 \lambda_\mathbf{u}(D_\mathbf{u})\lambda_\mathbf{v}(D_\mathbf{v}).$$

□

B. Proof of Theorem 3.11

The proof requires a lot of preliminary results. In particular, the uniqueness up to permutation of blocks and conjugation by orthogonal matrices relies on the theory of matrix $*$ algebras.

A *matrix $*$ algebra* $\mathcal{A}$ is a subalgebra of the $\mathbb{R}$ algebra of $n \times n$ matrices $M_n(\mathbb{R})$ which is closed under taking the transpose i.e. $A \in \mathcal{A}$ implies $A^T \in \mathcal{A}$. Let $A_i \in M_{n_i}(\mathbb{R})$ and denote by $\mathrm{diag}(A_1, \ldots, A_n)$ the block diagonal matrix

$$\begin{pmatrix} A_1 & 0 & 0 & 0 \\ 0 & A_2 & 0 & 0 \\ \ldots & \ldots & \ddots & 0 \\ 0 & 0 & 0 & A_n \end{pmatrix}.$$

a subset $\mathcal{A} \subseteq \{\mathrm{diag}(A_1, \ldots, A_k) : A_i \in M_{\tilde{n}_i}\}$ we denote by π_j the projection onto the j-th block i.e. $\pi_j(\mathrm{diag}(A_1, \ldots, A_k)) = A_j$. If $\mathcal{A}$ is a matrix $*$ algebra the map $\pi_j : \mathcal{A} \to M_{\tilde{n}}(\mathbb{R})$ is a *morphism of matrix $*$ algebras* which means that π_j is a $\mathbb{R}$ algebra morphism fullfiling $\pi_j(A^T) = \pi_j(A)^T$.
Recall that an $\mathbb{R}$ algebra is *simple* if it does not contain any proper ideals. A matrix algebra $\mathcal{A} \subseteq M_n(\mathbb{R})$ has a canonical representation $\rho : \mathcal{A} \to End(\mathbb{R}^n)$ given by $\rho(x) = x$ which is called the *regular representation* of $\mathcal{A}$. We call $\mathcal{A}$ *irreducible* if the regular representation is irreducible i.e. the regular representation contains no non-trivial subrepresentation.

Remark B.1. *An important decomposition of a semisimple algebra is the Artin-Wedderburn decomposition and we will describe its close relationship to the block diagonalization. While subalgebras of $M_n(\mathbb{R})$ are not semisimple in general, matrix $*$ algebras subalgebras of $M_n(\mathbb{R})$ are semisimple $\mathbb{R}$ algebras [38, Lemma A.3]. Thus by the Artin-Wedderburn Theorem [15, Theorem 5.2.4] for every matrix $*$-subalgebra S of $M_n(\mathbb{R})$ there exists division algebras D_i over $\mathbb{R}$ and $n_i \in \mathbb{N}$ such that*

$$S \simeq M_{n_1}(D_1) \times \cdots \times M_{n_j}(D_j)$$

where up to permutation of factors the n_i are unique and the D_i are unique up to isomorphism. Note that the individual factors $M_{n_i}(D_i)$ are simple $\mathbb{R}$-algebras and simple S modules.

Remark B.2. *Note that $U^T H_i U, 1 \leq i \leq n$ has a common block diagonal form if and only if this is true for every element of the matrix $*$ algebra generated by $U^T H_i U, 1 \leq i \leq n$.*

This follows from the fact that for $U \in O(d)$ we have that

$$U^T(A+B)U = U^TAU + U^TBU, \qquad U^TABU = U^TAUU^TBU$$
$$(U^TAU)^T = U^TA^TU.$$

In particular if $H_1, \dots, H_n$ generate $V_{\nabla^2 f}$ and $U \in O(d)$ corresponds to a finest block diagonalization of $H_1, \dots, H_n$ then $f_U \in \mathcal{MK}(\mathcal{GS}(f))$. Furthermore up to a permutation matrix every element of $\mathcal{MK}(\mathcal{GS}(f))$ is of this form. Thus we can use block diagonalization results for matrix $*$ algebras in order to find a suitable $U \in O(d)$.

While there is algorithms about block diagonalization of matrix $*$ algebras to our knowledge the uniqueness of the blocks up to orthogonal conjugation of the blocks is not explicitly stated in the corresponding papers which is why we will briefly discuss it in the following.
The general strategy for proving the uniqueness of a finest block diagonalization up to conjugation of the blocks by orthogonal matrices is as follows. Let $H_1, \dots, H_n \in M_d(\mathbb{R})$ and let $U_1, U_2 \in O(d)$ correspond to finest block diagonalizations. Denote by B_i the matrix $*$-algebra generated by $U_i^TH_1U_i, \dots, U_i^TH_nU_i$. Then

- Show that there exists simple matrix $*$-subalgebras $\mathcal{T}_{i,j} \subseteq B_i$ such that

  $$B_i = \{\mathrm{diag}(T_{i,1}, \dots, T_{i,n_i}) : T_{i,j} \in \mathcal{T}_{i,j}\}.$$

  Show that $n_1 = n_2$.

- Denote by $\varphi : B_1 \to B_2$ the matrix $*$-isomorphism $\varphi : x \mapsto U_2^TU_1xU_1^TU_2$ and by $\varphi|_{\mathcal{T}_{1,j}} : \mathcal{T}_{1,j} \to B_2$ the restriction to $\mathcal{T}_{1,j}$. Show up to permutation of the indices j it holds $\varphi|_{\mathcal{T}_{1,j}}(\mathcal{T}_{1,j}) = \mathcal{T}_{2,j}$. Consequently

  $$\varphi|_{\mathcal{T}_{1,j}} : \mathcal{T}_{1,j} \to \mathcal{T}_{2,j}$$

  is a matrix $*$-isomorphism.

- To conclude the argument we show that a matrix $*$-algebra isomorphism between simple matrix $*$-algebras $S_1, S_2 \in M_n(\mathbb{R}^d)$ which are of finest decomposition can be described as conjugation of the blocks by orthogonal matrices.
  To be more precise let $S_1 \subseteq \{\mathrm{diag}(A_1, \dots, A_n) : A_j \in M_m(\mathbb{R})\} \supseteq S_2$ be simple matrix $*$-subalgebras such that the image of the projection π_j onto the j-th block is irreducible. Then for every matrix $*$-isomorphism $\phi : S_1 \to S_2$ there exists $V_j \in O(d)$ such that

  $$\phi(x) = \mathrm{diag}(V_1, \dots, V_n)^T x \mathrm{diag}(V_1, \dots, V_n).$$

Overall we see that up to a permutation of blocks finest block diagonalizations are unique up to a orthogonal conjugation of the individual blocks.
The following theorem guarantees the existence of a finest block diagonalization.

Theorem B.3. *[38, Theorem 1] Let $\mathcal{T}$ be a $*$ subalgebra of $M_n(\mathbb{R})$. Then there exists $U \in O(n)$ and simple subalgebras $\mathcal{T}_j \subseteq M_{n_j}$ s.t.*

$$U^T \mathcal{T} U = \{\mathrm{diag}(T_1, \dots, T_k) : T_i \in \mathcal{T}_i\}.$$

Furthermore for a simple subalgebra $\mathcal{T} \subseteq M_n(\mathbb{R})$ there exists an irreducible subalgebra $\mathcal{T}^i \subseteq M_{\hat{n}}(\mathbb{R})$ and a $V \in O(n)$ s.t.

$$V^T \mathcal{T} V = \{\mathrm{diag}(A, \dots, A) : A \in \mathcal{T}^i\}.$$

Remark B.4. *Theorem B.3 also illustrates that for two isomorphic matrix $*$-algebras $S_1 \simeq S_2$ which are both contained in $M_n(\mathbb{R})$ it is not necessarily true that they are isomorphic as vector spaces. Take e.g. the irreducible matrix $*$-algebra $S = M_m(\mathbb{R})$ and $S_1 = \{\mathrm{diag}(A, 0) : A \in S\} \subset M_{2m}(\mathbb{R})$ and $S_2 = \{\mathrm{diag}(A, A) : A \in S\} \subset M_{2m}(\mathbb{R})$.*

Remark B.5. *All finest block diagonalizations correspond to a matrix $*$-algebra decomposition as in Theorem B.3 i.e. the block of a finest block diagonalization are irreducible matrix $*$-algebras. Otherwise they could be further refined. By Lemma B.7 these irreducible blocks can be grouped into simple algebras algebras.*

We start the rigorous proof with the following auxiliary lemma.

Lemma B.6. *Let U correspond to a finest block diagonalization of $H_1, \dots, H_n$ with block sizes $(d_1, \dots, d_k)$. Let A be the matrix $*$-subalgebra of $M_d(\mathbb{R})$ generated by $H_1, \dots, H_n$. Then also $U^T A U$ is a matrix $*$-algebra and every $a \in A$ is blockdiagonal with block sizes $(d_1, \dots, d_k)$. Furthermore the projection $\pi_i : U^T A U \to M_{d_i}(\mathbb{R})$ is a $*$-algebra morphism onto an irreducible subalgebra of $M_{d_i}(\mathbb{R})$.*

Proof. It is immediate that $U^T A U$ is a matrix $*$-algebra which is block diagonal and that π_i is a $*$-algebra morphism, see also Remark B.2. Assume that the image of π_i was not irreducible. Then according to Theorem B.3 this block could be further refined by an orthogonal conjugation which would contradict that U corresponds to a finest block diagonalization. □

Next we make sure that the simple components of the Artin-Wedderburn decomposition show up as, not necessarily finest, blocks in the finest block decomposition.

Lemma B.7. *Let $H_1, \dots, H_n \in \mathbb{R}^{d \times d}$ and assume that $U \in O(d)$ corresponds to a finest block diagonalization. Let A be the matrix $*$-algebra generated by $H_1, \dots, H_n$ and let $d_1, \dots, d_k$ be the size of the blocks. Then there exists unique (up to block permutation) simple matrix $*$-algebras $\mathcal{T}_i \subseteq U^T A U$ such that*

$$U^T A U = \{\mathrm{diag}(T_1, \dots, T_k) : T_i \in \mathcal{T}_i\}$$

Proof. Let $B = U^T A U$ and let $B \simeq \oplus_{i=1}^k \mathcal{R}_i$ be the Artin-Wedderburn decomposition into simple B modules which are also subalgebras (without 1), which exists by Remark B.1. By abuse of notation denote by $\mathcal{R}_i$ also the corresponding subalgebra of B. Let π_j be the

projection of B onto the j-th block. By Corollary B.6 the image of π_j is irreducible and hence in particular simple. Thus Schur's lemma implies that $\pi_j\Big|_{\mathcal{R}_i}$ is either zero or an isomorphism onto the block. Since $\mathcal{R}_i$ are orthogonal, we have that if for indices n, m we have that both $\pi_j\big|_{\mathcal{R}_n} \neq 0$ and $\pi_j\big|_{\mathcal{R}_m} \neq 0$, then $n = m$ and the claim follows. □

Lemma B.8. *Let U_1, U_2 correspond to finest block diagonalizations and denote by B_i the matrix $*$-algebras $U_i^T A U_i$. Denote by $\mathcal{T}_{1,1}, \dots, \mathcal{T}_{1,k_1}$ resp. $\mathcal{T}_{2,1}, \dots, \mathcal{T}_{2,k_2}$ be the simple matrix $*$ algebras from Lemma B.7. Then $k_1 = k_2 =: k$ and for $\varphi : B_1 \to B_2$ defined by $\varphi(x) = U_2^T U_1 x U_1^T U_2$ there exists a permutation $\sigma \in \Sigma(k)$ such that $\varphi(\mathcal{T}_{1,j}) = \mathcal{T}_{2,\sigma(j)}$ and $\varphi : \mathcal{T}_{1,j} \to \mathcal{T}_{\sigma(j),2}$ is an isomorphism of matrix $*$ algebras.*

Proof. By Artin-Wedderburn $k_1 = k_2$. By Schur's lemma we have that $\pi_{T_{2,m}} \circ \varphi\Big|_{\mathcal{T}_{1,n}}$ is either an isomorphism or zero. Since the $T_{1,l}$ are orthogonal we have that if for m, n it holds that both $\pi_{T_{2,j}} \circ \varphi\Big|_{\mathcal{T}_{1,m}} \neq 0$ and $\wedge \pi_{T_{2,j}} \circ \varphi\Big|_{\mathcal{T}_{1,n}} \neq 0$ then $m = n$. Hence there exists a partition $[k_2] = \cup_{j=1}^{k_1} I_j$ with $I_j = \{n \in [k_2] : \pi_{T_{2,n}} \circ \varphi\Big|_{\mathcal{T}_{1,j}} \neq 0\}$. Since φ is an isomporhism we know that $|I_j| \geq 1$. Using that $k_1 = k_2$ we can conclude that $|I_j| = 1$ and thus the claim. □

Proposition B.9. *Let $\mathcal{S} \subseteq M_n(\mathbb{R})$ be an irreducible matrix $*$ subalgebra and let $\mathcal{T} \subseteq M_{\tilde{n}}$ be a simple matrix $*$ subalgebra isomporhic to $\mathcal{S}$ via a matrix $*$ morphism $\varphi : \mathcal{S} \to \mathcal{T}$. Assume furthermore that $\mathcal{T} \subseteq \{\mathrm{diag}(A_1, \dots, A_k) : A_i \in M_n(\mathbb{R})\}$. Then there exists $V_1, \dots, V_k \in O(n)$ such that for $V = \mathrm{diag}(V_1, \dots, V_k)$ we have that $\varphi(x) = V^T \mathrm{diag}(x, \dots, x) V$.*

Proof. By Skolem-Noether there exists matrices $W_1, \dots, W_k \in GL_n(\mathbb{R})$ such that for $W = \mathrm{diag}(W_1, \dots, W_k)$ we have that $\varphi(x) = W^{-1}\mathrm{diag}(x, \dots, x)W$. This implies that for π_i the projection of the k-th block we have that $\pi_i \circ \varphi$ is an isomorphism of matrix $*$ representations via the map $W_i : \mathbb{R}^n \to \mathbb{R}^n$. Thus by [38, Lemma A.4] there exists $V_i \in O(n)$ such that $\pi_i \circ \varphi(x) = V^T x V$ and thus the claim. □

Corollary B.10. *Let $\mathcal{T}_1 \subseteq M_n(\mathbb{R})$ and $\mathcal{T}_2 \subseteq M_m(\mathbb{R})$ be simple isomorphic matrix $*$ algebra with isomorphism φ. Assume furthermore that $\mathcal{T}_i \subseteq \{\mathrm{diag}(A_1, \dots, A_{k_i}) : A_j \in M_{\tilde{n}}(\mathbb{R})\}$ such that the image of the projection onto the ℓ-th block is an irreducible representation. Then φ induces an isomorphism $\varphi_{i,j} : \pi_i(\mathcal{T}_1) \to \pi_j(\mathcal{T}_2)$ and there exists a $V \in O(\tilde{n})$ such that $\varphi_{i,j}(x) = V^T x V$.*

Proof. There exists an irreducible matrix $*$ algebra $\mathcal{S}$ and an isomorphism $\tau : \mathcal{S} \to \mathcal{T}_1$. By Proposition B.9 we have that there exists V_1, V_2 such that $\pi_i \circ \tau(x) = V_1^T x V_1$ and $\pi_j \circ \varphi \circ \tau(x) = V_2^T x V_2$. Thus we can use $V := V_1^{-1} V_2$. □

Proposition B.11. *Let U_1, U_2 correspond to finest block diagonalizations and denote by B_1, B_2 the matrix $*$ algebras $U_1^T A U_1$ resp. $U_2^T A U_2$. Let $d_{1,j}, \dots, d_{k_j,j}$ be the blocks sizes of B_j and let $\varphi : B_1 \to B_2$ be defined by $\varphi(x) = U_2^T U_1 x U_1^T U_2$. Then $k_1 = k_2 =: k$*

and there exists a permutation $\sigma \in \Sigma(k)$ such that $d_{i,1} = d_{\sigma(i),2}$ and a block permutation matrix P_σ and $V_1, \dots, V_k, V_i \in O(d_{i,1})$ such that $\varphi(x) = V^T P_\sigma^T x P_\sigma V$.

Proof. By Lemma B.7 we know that there exists a decomposition of B_1, B_2 into simple algebras which have disjoint block form. By Lemma B.8 there exists a permutation of simple blocks such that φ maps a simple block isomorphically to exactly one simple block. Since φ does preserve the rank of matrices the corresponding simple blocks must be of same dimension. By Corollary B.10 the claim follows. □

C. Algorithmic Details

C.1. Grid search and initialization close to a global minimum

According to Lemma 4.1, the loss function ℓ_ε is not geodesically convex, and, thus, convergence to a global solution is not guaranteed. By Theorem 4.3, the Riemannian gradient descent method converges locally with suitable step sizes. To find a good starting point $U \in SO(d)$, we establish a grid search procedure on $SO(d)$. Its main idea is to compute the loss on the discrete subset $\Gamma \subset SO(d)$ and initialize the gradient-descent method with the minimizer of

$$U = \operatorname*{argmin}_{V \in \Gamma} \ell_\varepsilon(V). \tag{31}$$

The benefit of working with $SO(d)$ is that it can be parametrized by $d(d-1)/2$ angles. Let us consider the Jacobi rotation matrix with parameters $r \in [d-1]$ and $\alpha \in [0, 2\pi)$ given by

$$\begin{aligned} R_{r,r}(r,\alpha) &= \cos(\alpha) \quad R_{r,r+1}(r,\theta) = -\sin(\alpha) \qquad R_{j,j}(r,\alpha) = 1,\ j \notin \{r, r+1\}, \\ R_{r+1,r}(r,\alpha) &= \sin(\alpha), \quad R_{r+1,r+1}(r,\alpha) = \cos(\alpha), \qquad R_{i,j}(r,\alpha) = 0 \text{ otherwise.} \end{aligned} \tag{32}$$

Then, every $SO(d)$ matrix factorizes into a product of $d(d-1)/2$ Jacobi matrices.

Proposition C.1 ([27])**.** *Let $d \geq 2$. For every $U \in SO(d)$, there exist vectors*

$$\alpha^r = (\alpha_1^r, \cdots, \alpha_r^r) \in [0, 2\pi) \times [0, \pi)^{r-1}, \quad r \in [d-1],$$

such that

$$U = U(\alpha) = \prod_{r=1}^{d-1} \prod_{j=1}^{r} R(d-1-r+j, \alpha_j^r). \tag{33}$$

Proposition C.1 provides a parametrization of $U \in SO(d)$ by a vector of angles $\alpha \in [0, 2\pi)^{d-1} \times [0, \pi)^{(d-1)(d-2)/2} \subset \mathbb{R}^{d(d-1)/2}$. In the context of the grid search, the main benefit of the angle representation of $SO(d)$ is the possibility to construct lattices efficiently. Let

$$\Theta(a, h) := \{kh : k \in \lceil a/h \rceil\} \quad \text{and} \quad \Theta(h) := \Theta(2\pi, h)^{d-1} \times \Theta(\pi, h)^{(d-1)(d-2)/2}$$

be one- and multidimensional lattices, respectively. Then, we define Γ in (31) as

$$\Gamma := \Gamma(h) = \{U(\alpha) \in SO(d) : \alpha \in \Theta(h)\}. \tag{34}$$

Then, for such Γ we have the following results.

Proposition C.2. *Let $h \in (0,1)$. Then, for every $V \in SO(d)$ there exists $\alpha \in \Theta(h)$ such that*

$$\big\|V - U(\alpha)\big\|_F \leq d(d-1)h.$$

Proof. Proposition C.1 allows to parametrize V by angles $\beta \in [0,2\pi)^{d-1} \times [0,\pi)^{(d-1)(d-2)/2}$ such that

$$V = V(\beta) = \prod_{r=1}^{d-1} \prod_{j=1}^{r} R(d-1-r+j, \beta_j^r) =: \prod_{k=1}^{d(d-1)/2} R_k(\beta),$$

where in the last equality we introduced a short single indexed notation. Then, for any $\alpha \in \Theta(h)$, using the telescopic sum, we get

$$\begin{aligned}
\big\|U(\alpha) - V(\beta)\big\|_F &= \Big\| \prod_{k=1}^{d(d-1)/2} R_k(\alpha) - \prod_{k=1}^{d(d-1)/2} R_k(\beta) \Big\|_F \\
&= \Big\| \sum_{n=1}^{d(d-1)/2} \Big[\prod_{k=1}^{n-1} R_k(\alpha)\Big] \big(R_n(\alpha) - R_n(\beta)\big) \Big[\prod_{k=n+1}^{d(d-1)/2} R_k(\beta)\Big] \Big\|_F \\
&\leq \sum_{n=1}^{d(d-1)/2} \Big\| \Big[\prod_{k=1}^{n-1} R_k(\alpha)\Big] (R_n(\alpha) - R_n(\beta)) \Big[\prod_{k=n+1}^{d(d-1)/2} R_k(\alpha)\Big] \Big\|_F \\
&= \sum_{n=1}^{d(d-1)/2} \big\|R_n(\alpha) - R_n(\beta)\big\|_F.
\end{aligned}$$

The last equality follows from the fact that the Frobenius norm is invariant under left and right multiplication with orthogonal matrices. Next, recall that for every $n \in [d(d-1)/2]$ there exists a unique pair $r \in [d-1]$ and $j \in [r]$ such that

$$\begin{aligned}
\big\|R_n(\alpha) - R_n(\beta)\big\|_F^2 &= \Big\|R(d-1-r+j, \alpha_j^r) - R(d-1-r+j, \beta_j^r)\Big\|_F^2 \\
&= 2(\cos(\alpha_j^r) - \cos(\beta_j^r))^2 + 2(\sin(\alpha_j^r) - \sin(\beta_j^r))^2 \leq 4|\alpha_j^r - \beta_j^r|^2,
\end{aligned}$$

where used definition (32) of R and that cosine and sine are Lipschitz-1 functions. By construction of $\Theta(h)$ there exists $\alpha \in \Theta(h)$, such that $|\alpha_j^r - \beta_j^r| < h$ for all r, j. Thus, we conclude that

$$\big\|U(\alpha) - V(\beta)\big\|_F \leq 2 \sum_{n=1}^{d(d-1)/2} |\alpha_j^r - \beta_j^r| < 2hd(d-1)/2 = d(d-1)h.$$

□

In other words, Γ is a $hd(d-1)$-net over $SO(d)$. Hence, for sufficiently small h, the minimizer of (31) will be in the neighborhood where local convergence of Riemannian gradient descent is guaranteed.

Corollary C.3. *For $0 < h < 1$, let $U(h)$ denote the minimizer of (31) with $\Gamma(h)$ given by (34). Then, there exists $h > 0$, such that $U(h)$ belongs to a level set defined in Theorem 4.3. Consequently, the sequence generated by Riemannian gradient descent with step sizes as in Theorem 4.2 and initial guess $U^{(0)} = U(h)$ converges to a global minimum of ℓ_ε.*

Proof. By Theorem 4.3, the all fixed points in the set $\mathcal{L} := \{U \in SO(d) : \ell_\varepsilon(U) < \ell^* + q^*\}$ are global minimizers of ℓ_ε over $SO(d)$. It is open as a preimage $\ell_\varepsilon^{-1}((-\infty, \ell^* + q^*))$ of an open set via continuous function. Let $U_* \in \mathcal{L}$ be an arbitrary global minimizer of ℓ_ε over $SO(d)$. Since $\mathcal{L}$ is open, the open ball $\{U \in SO(d) : \|U - U_*\|_F < \delta\}$ is contained in $\mathcal{L}$ for some $\delta > 0$. Then, by Proposition C.2 with $h = \delta/d(d-1)$, we can find $U_b \in \Gamma$ satisfying $\|U_b - U^*\|_F < \delta$ and, hence, $U_b \in \mathcal{L}$. Consequently, the minimizer $U(h)$ of (31) admits

$$\ell_\varepsilon(U(h)) = \min_{U \in \Gamma} \ell_\varepsilon(U) \le \ell_\varepsilon(U_b) < \ell^* + q^*,$$

so that $U(h) \in \mathcal{L}$. The rest of the claim follows from Theorem 4.3. □

We summarize the manifold optimization methods for in Algorithm 2.

Algorithm 2 Manifold optimization for edge minimization

Input Diagonal blocks B_n as in (18), $n \in [N]$, step sizes $\{\nu_r\}_{r \ge 0}$, grid search density $h \in (0, 1)$, smoothing $\varepsilon > 0$.
Construct $\Gamma(h)$ as in (34) and compute $U^{(0)}$ as the minimizer of (31).
for r = 0,1,... **do**
  Compute $U^{(r+1)}$ via (21) or (22) with step sizes ν_r.
  Check the stopping criteria.
end for
Return: $\{U^{(r)}\}_{r \ge 0}$.

C.2. Computational complexity

The main drawback of computation complexity is the necessity to evaluate the function value for every $U \in \Gamma$. By construction (34), it leads to an exponential number of elements,

$$|\Theta(h)| = \left(\left\lceil \frac{2\pi}{h} \right\rceil\right)^{d-1} \left(\left\lceil \frac{\pi}{h} \right\rceil\right)^{\frac{(d-1)(d-2)}{2}}.$$

Therefore, the grid search is only available for small dimension d and requires efficient computation of matrices $U \in \Gamma$ and corresponding values $\ell_\varepsilon(U)$.

Given that Algorithm 2 is applied to blocks $B^{U,k}$ in (18), the dimension of the problems is small as a result of splitting the graph into the finest connected components and optimizing the edges for each of them.

Furthermore, to avoid possible memory issues, the grid $\Theta(h)$ is split into blocks. First, the minimizer of $\ell(U)$ within the blocks is computed, and then U in (31) is taken as the minimum among the blocks.

As for the computation of U, the construction of rotation matrices (32) leads to a fast multiplication with other matrices. Let $A = (a_1 \dots\ a_d) \in \mathbb{R}^{d\times d}$. Then, for arbitrary $r \in [d-1], \alpha \in [0, 2\pi)$, the product $AR(r, \alpha)$ has is given by

$$\begin{pmatrix} a_1 & \dots & a_{r-1} & \cos(\alpha)a_r + \sin(\alpha)a_{r+1} & \cos(\alpha)a_{r+1} - \sin(\alpha)a_r & a_{r+1} & \dots & a_d \end{pmatrix},$$

and requires only $4d$ operations. Using this, any $U \in \Gamma$ can be efficiently computed using $\mathcal{O}(d^3)$ operations.

The computational complexity of Algorithm 2 is summarized in Tabular 3.

Grid search	Number of grid points $\Theta(h)$	$\mathcal{O}\big((\lceil 2\pi/h\rceil)^d(\lceil \pi/h\rceil)^{d^2}\big)$
	Matrix $U \in \Gamma$	$\mathcal{O}(d^3)$
	Loss function $\ell(U)$	$\mathcal{O}(Nd^3)$
	Total	$\mathcal{O}\big(Nd^3(\lceil 2\pi/h\rceil)^d(\lceil \pi/h\rceil)^{d^2}\big)$
Manifold Optimiztion	Riemannian/Euclidean gradient	$\mathcal{O}(Nd^3)$
	Riemannian gradient descent (21)	$\mathcal{O}(KNd^3)$
	Landing (22)	$\mathcal{O}(KNd^3)$

Table 3: Number of operations for the grid search and for K iterations of the manifold optimization method. Here d is the dimension, N the number of Hessians and h the density of the grid.

D. Additional numerical results

A more detailed analysis for the results of Figure 5 are given for $\eta = 10^{-9}$ (noise free Hessians), and $\eta = 10^{-4}$ (noisy Hessians) in tables 4, 5, 6, 7. They indicate that the sparsifying performance is better for the grid search initialization for small h compared to random initialization for dimensions $d \in \{4, 5\}$ for both clean and noisy data. Riemannian gradient descent performs better than the Landing method on clean data but the comparison on noisy data is inconclusive. Table 8 shows that when the underlying dimension d is small, there is no significant difference between the two methods regarding the time complexity. The only case where we can get improvement from Landing is runtime for large d [1]. In addition, as expected, the runtime effort for the calculation of the rotation matrices $U \in \Gamma(h)$ and the losses $\ell(U)$ increases with the dimension d, the step size h and the number of Hessians N. Furthermore, we see that the runtime of the grid search method for the step sizes considered in our numerics is still low and the grid-search initialization provides better results than the random initialisation method. To apply the grid search method more efficiently to more than 1 set of matrices, it is beneficial to calculate the rotation matrices only once, and reuse then for any set of matrices.

		Clean data								Noisy data $\sigma = 10^{-3}$							
		Rgd				Landing				Rgd				Landing			
i		0	1	2	≥ 3	0	1	2	≥ 3	0	1	2	≥ 3	0	1	2	≥ 3
Random initialization		100	0	0	0	100	0	0	0	98	2	0	0	98	2	0	0
Grid-Search initialization	$h = 1$	100	0	0	0	100	0	0	0	98	2	0	0	98	2	0	0
	$h = 0.5$	100	0	0	0	100	0	0	0	98	2	0	0	98	2	0	0
	$h = 0.25$	100	0	0	0	100	0	0	0	98	2	0	0	98	2	0	0
	$h = 0.125$	100	0	0	0	100	0	0	0	98	2	0	0	98	2	0	0
	$h = 0.1$	100	0	0	0	100	0	0	0	98	2	0	0	98	2	0	0

Table 4: Number of set of matrices such that diff $\chi(U, \eta) = i$ (see (25)) and input dimension $d = 2$, $\eta = 10^{-9}, 10^{-4}$ for clean resp. noisy data. Rgd: Riemannian gradient descent.

		Clean data								Noisy data $\sigma = 10^{-3}$							
		Rgd				Landing				Rgd				Landing			
i		0	1	2	≥ 3	0	1	2	≥ 3	0	1	2	≥ 3	0	1	2	≥ 3
Random initialization		98	0	0	2	87	0	5	8	98	1	1	0	95	2	3	0
Grid-Search initialization	$h = 1$	99	1	0	0	91	1	0	8	98	1	1	0	98	1	1	0
	$h = 0.5$	100	0	0	0	91	0	0	9	99	0	1	0	99	0	1	0
	$h = 0.25$	100	0	0	0	91	0	0	9	99	0	1	0	99	0	1	0
	$h = 0.125$	100	0	0	0	92	0	0	8	99	0	1	0	99	0	1	0
	$h = 0.1$	100	0	0	0	92	0	0	8	100	0	0	0	100	0	0	0

Table 5: Same as table 4 for input dimension $d = 3$.

		Clean data								Noisy data $\sigma = 10^{-3}$							
		Rgd				Landing				Rgd				Landing			
i		0	1	2	≥ 3	0	1	2	≥ 3	0	1	2	≥ 3	0	1	2	≥ 3
Random initialization		86	1	8	5	57	2	21	20	84	6	7	3	88	1	11	0
Grid-Search initialization	$h = 1$	94	3	2	1	84	1	7	8	92	5	2	1	92	5	2	1
	$h = 0.5$	98	0	1	1	87	0	7	6	96	2	2	0	96	2	2	0
	$h = 0.25$	100	0	0	0	89	0	6	5	99	1	0	0	99	1	0	0
	$h = 0.125$	100	0	0	0	89	0	6	5	99	1	0	0	99	1	0	0
	$h = 0.1$	100	0	0	0	89	0	6	5	99	1	0	0	99	1	0	0

Table 6: Same as table 4 for input dimension $d = 4$

		Clean data								Noisy data $\sigma = 10^{-3}$							
		Rgd				Landing				Rgd				Landing			
i		0	1	2	≥ 3	0	1	2	≥ 3	0	1	2	≥ 3	0	1	2	≥ 3
Random initialization		71	4	13	12	37	7	13	43	70	9	9	12	74	5	8	13
Grid-Search initialization	$h = 1$	93	2	4	1	85	2	8	5	93	2	4	1	93	2	4	1

Table 7: Same as table 4 for input dimension $d = 5$

Input dimension	Random Init.		Grid-search Init.									
	Rgd	La	Rotation $U \in \Gamma(h)$					Losses $\ell(U), U \in \Gamma(U)$				
			1	0.5	0.25	0.125	0.1	1	0.5	0.25	0.125	0.1
$d = 2$	29.36	33.36	0.0010	0.0009	0.0008	0.0008	0.0008	0.0003	0.0003	0.0003	0.0003	0.0003
								0.0009	0.0009	0.0009	0.0009	0.0009
$d = 3$	29.39	32.18	0.0211	0.0211	0.0211	0.0213	0.0215	0.0032	0.0032	0.0032	0.0035	0.0041
								0.0194	0.0193	0.0193	0.0210	0.0244
$d = 4$	29.66	32.50	0.6105	0.7062	2.0015	11.3731	168.4473	0.0343	0.0761	0.1732	5.9366	12.4121
								0.3434	0.7612	1.7323	59.3655	124.1206
$d = 5$	29.77	32.55	23.5289	–	–	–	–	1.1344	–	–	–	–
								17.0158	–	–	–	–

Table 8: Mean runtime in seconds(s) for the 5-time random initialisation method with $K = 5 \cdot 10^3$ iterations and the grid search initialisation (step size $h = 1, 0.5, 0.25, 0.125, 0.1$) procedure over 10 random chosen examples out of the 100 randomly generated examples for dimension $d = 2, 3, 4, 5$. For each dimension $d = 2, 3, 4, 5$ the first row in the subtable losses $\ell(U), U \in \Gamma(U)$ denotes the minimal runtime for $N = 1$ Hessians and the second row the maximal runtime for $N = d(d + 1)/2$ Hessians. Rgd: Riemannian gradient descent. La: Landing method.