Musings on Multilinear Fitting*

Martin J. Mohlenkamp †

December 26, 2010

Abstract

We show that the problems of approximating tensors and multivariate functions as a sums of (tensor) products of vectors/functions can be considered in a unified framework, thus exposing their common multilinear structure. We study the alternating least squares algorithm within this framework from the orthogonal projection and gradient perspectives. We then use these perspectives to study its convergence behavior with and without regularization. Finally, we formulate the infinite dimensional version of this problem and an algorithm to compute in that context.

AMS Subject Classification: 15A69, 65D15
Keywords: Tensor approximation, Multilinear approximation, Alternating least squares

1 Introduction

The tensor approximation problem

$$F(j_1, j_2, \ldots, j_d) \approx \sum_{l=1}^{r} \prod_{i=1}^{d} v_i^l(j_i) \quad \text{for} \quad j_i = 1, 2, \ldots, M_i \tag{1}$$

and the multivariate function approximation problem

$$f(x_1, x_2, \ldots, x_d) \approx \sum_{l=1}^{r} \prod_{i=1}^{d} \phi_i^l(x_i) \quad \text{for} \quad x_i \in X_i \tag{2}$$

have becoming increasing important as uses for them develop. Yet, even after many years of study, our understanding of them and the main algorithm used to compute them is unsatisfactory. In this paper we attempt to clarify several aspects of these problems and the Alternating Least Squares (ALS) algorithm that tries to solve them. Although we include some mathematical "results", much of this work consists of very carefully formulating the problems in order to expose their essence. We also interpret this essence to suggest what should and should not be done when developing algorithms for and understanding of these problems.

The first goal of this paper is to show that the problems (1) and (2) are the same. More specifically, (1) is a special case of (2). This is not say that the usage, desired side constraints, etc. are always the same, but only that the essential multilinear nature is the same. To support this assertion, in Section 2 we develop a common framework for both (1) and (2). Thereafter, the discussion is independent of which problem we are considering, with the exception that when f in (2) requires $r = \infty$ to achieve equality we cannot use

*This material is based upon work supported by the National Science Foundation under Grant DMS-0545895.
†Department of Mathematics, Ohio University, 321 Morton Hall, Athens OH 45701; mohlenka@ohio.edu.

compactness arguments in the posedness and convergence discussion. We will use the function notation, since it is the most general (and cleanest). Since the essential multilinear structure of (1) and (2) is the same, if a general conceptual framework of multilinear approximations exists, then it must apply to both problems. In particular, it can not use partial derivatives such as $\frac{\partial^2}{\partial x_1 \partial x_2} f(x_1, x_2, \ldots, x_d)$ because they make no sense for (1). Unfortunately, this excludes the techniques developed for the closely-related sparse-grid methods (e.g. [6, 10, 11, 23, 24, 25]), which have produced the best results so far.

The "workhorse" algorithm for finding good approximations to (1) and (2) is Alternating Least Squares (ALS). The second goal of this paper is to clarify the nature of ALS. As a preliminary, in Section 3 we review ordinary linear least-squares in a setting and notation compatible with Section 2. We show how to solve a linear least-squares problem with a method based on orthogonal projectors and with a method based on gradients, and show that these two perspectives result in exactly the same method. In Section 4 we review the ALS algorithm and show how to solve it using the method(s) in Section 3. The ALS routine can therefore be viewed as either alternately performing orthogonal projections or as alternately solving to make partial gradients zero. The gradient perspective allows one to relate ALS to other gradient-based optimization methods. The orthogonal projectors perspective is cleaner, and also shows that ALS always produces factors ϕ_i^l in a minimal subspace determined by f. In Section 4.3 we consider the convergence properties of ALS. Without regularization the convergence is unsatisfactory since the norms of summands can diverge to infinity. With regularization the convergence is more satisfactory, but still has some subtleties. Even at their best, the convergence results do not guarantee finding a global minimum of the approximation error or of converging at any particular rate.

In Section 5 we formally pose the multilinear approximation problem in infinite dimensions. We show that an ALS-like algorithm can be performed in the infinite dimensional setting. Although the infinite dimensional setting may be purely academic, it provides a framework for thought experiments comparing ALS to alternative algorithms. If a proposed method cannot function in $d = \infty$ with minor modifications, then it likely will not scale well to large d.

Since most of the work here is based on classical definitions in analysis and linear algebra, we have gone rather light on the references. For a recent survey focusing on the tensor problem, see [16]. There is no such survey for the function problem, but we offer as examples [1, 2, 3, 15, 9, 13, 14, 7, 6, 22, 23, 24].

2 The Multilinear Setting

2.1 Basic Hilbert Spaces

Consider some variable x which can take values in some set X, on which there is a measure dx. For any two complex-valued functions of x, we can define the inner product by

$$\langle f, g \rangle = \int_X f(x)\bar{g}(x)dx \tag{3}$$

where $\bar{g}$ denotes complex conjugation. From this inner product, we can define a norm by $\|f\| = \sqrt{\langle f, f \rangle}$. The set of functions with finite norm defines a Hilbert space H. We need to consider two functions to be the same element of H if they differ pointwise only on a set of measure zero with respect to dx, which means we say $f \equiv g \Leftrightarrow \|f - g\| = 0$.

This function notation is quite general, and includes some special cases of note. If $X = \{1, 2, \ldots, M\}$ and dx is the counting measure, then a function f is specified by its values $(f(1), f(2), \ldots, f(M))$. We can then refer to f as a vector, although nothing is gained by doing so. If $X = \{1, 2, \ldots\}$ and dx is the counting measure, then a function f is specified by the infinite vector (i.e. sequence) $(f(1), f(2), \ldots)$. For f to have finite norm, we need $\sum_{j=1} |f(j)|^2 < \infty$. If H has a countable orthonormal basis, then f is similarly

specified by the infinite vector of coefficients $(\hat{f}(1), \hat{f}(2), \dots)$ and the inner product can also be computed by $\langle f, g\rangle = \sum_{j=1} \hat{f}(j)\overline{\hat{g}}(j)$.

2.2 Tensor Product Hilbert Space

Consider a set of such Hilbert spaces $\{H_i\}_{i=1}^d$. When it is not clear from the context, we will use the subscript i to indicate which domain X_i, inner product $\langle\cdot,\cdot\rangle_i$, etc. we are talking about. Now take a set of elements $f_i \in H_i$. The tensor product $f = f_1 \otimes f_2 \otimes \cdots \otimes f_d = \bigotimes_{i=1}^d f_i$ defines a function of $\mathbf{x} = (x_1, x_2, \dots, x_d)$ by $f(\mathbf{x}) = \prod_{i=1}^d f_i(x_i)$. We can see from this definition that scalars can be moved between factors f_i without changing f. We can define an inner product of two such objects by

$$\langle f, g\rangle = \left\langle \bigotimes_{i=1}^d f_i, \bigotimes_{i=1}^d g_i \right\rangle = \prod_{i=1}^d \langle f_i, g_i\rangle_i \,, \tag{4}$$

and the corresponding norm $\|f\| = \sqrt{\langle f, f\rangle}$. Using linearity, the inner product is defined for finite linear combination of such objects. The tensor product space $H = \bigotimes_{i=1}^d H_i$ is defined as the completion of the set of finite linear combinations of these objects. The inner product can also be written as

$$\langle f, g\rangle = \int f(\mathbf{x})\bar{g}(\mathbf{x})d\mathbf{x} = \iint \cdots \int f(x_1, x_2, \dots, x_d)\bar{g}(x_1, x_2, \dots, x_d)dx_1 dx_2 \cdots dx_d \,. \tag{5}$$

As a special case, if each $X_i = \{1, 2, \dots, M_i\}$ and dx_i is the counting measure, then f is specified by its values on all combinations of these x_i. These values can be arranged in a d-dimensional array and f referred to as a tensor. Nothing in particular is gained by doing so.

For notational convenience we allow the inner product on H_i to apply to functions involving other variables as well, as in

$$\langle f, g\rangle_i = \int f(x_1, x_2, \dots, x_d)\bar{g}(x_1, x_2, \dots, x_d)dx_i \,, \tag{6}$$

which in general yields of function of all x_j with $j \neq i$. Similarly, we define the complementary inner product

$$\langle f, g\rangle_{\setminus i} = \iint \cdots \int f(x_1, x_2, \dots, x_d)\bar{g}(x_1, x_2, \dots, x_d)dx_1 dx_2 \cdots dx_{i-1} dx_{i+1} \cdots dx_d \,, \tag{7}$$

which in general yields a function of x_i.

2.3 Approximation with Sums of Products

Let $f \in H$ be some fixed "target" element. We wish to approximate f with a function g that is a sum of separable functions, written as

$$g(x_1, x_2, \dots, x_d) = \sum_{l=1}^r g^l(x_1, x_2, \dots, x_d) = \sum_{l=1}^r \prod_{i=1}^d \phi_i^l(x_i) \,, \tag{8}$$

where the superscript l is an index rather than a power. We would like to minimize the error function

$$E(g) = \|f - g\|^2 = \langle (f - g), (f - g)\rangle \tag{9}$$

over all choices of $\{\phi_i^l\}$ for fixed r. Formally, we define

$$G_r = \left\{ \sum_{l=1}^{r} \prod_{i=1}^{d} \phi_i^l(x_i) \quad | \quad \phi_i^l \in H_i \right\}, \tag{10}$$

$$E^r = \{E(g) \quad | \quad g \in G_r\}, \quad \text{and} \tag{11}$$

$$\hat{E}^r = \inf E^r, \tag{12}$$

so the goal is to find $g \in G_r$ with $E(g) = \hat{E}^r$, if possible.

We also consider the regularized error for $\lambda > 0$ of

$$E_\lambda(g) = E_\lambda(g^1, \ldots, g^r) = \|f - g\|^2 + \lambda \sum_{l=1}^{r} \|g^l\|^2 = \langle (f-g), (f-g) \rangle + \lambda \sum_{l=1}^{r} \prod_{i=1}^{d} \langle \phi_i^l, \phi_i^l \rangle \,. \tag{13}$$

The approximation of $f \in H$ with g using E_λ can be viewed as the approximation of $(f, 0, \ldots, 0) \in H^{r+1}$ with $(\sum_{l=1}^{r} g^l, g^1, \ldots, g^r)$ using $r+1$ copies of E with last r of them scaled by λ. Thus we still have a least-squares problem. Formally, we define

$$E_\lambda^r = \{E_\lambda(g) \quad | \quad g \in G_r\} \quad \text{and} \tag{14}$$

$$\hat{E}_\lambda^r = \inf E_\lambda^r, \tag{15}$$

so the goal is to find $g \in G_r$ with $E_\lambda(g) = \hat{E}_\lambda^r$, if possible.

A special case of this problem is when f itself is a sum of separable functions, written as

$$f(x_1, x_2, \ldots, x_d) = \sum_{l=1}^{R} \prod_{i=1}^{d} \theta_i^l(x_i) \,. \tag{16}$$

The goal then is to approximate $f \in G_R$ using $g \in G_r$ with $r < R$. By the definition of H, any $f \in H$ can be arbitrarily well approximated by choosing R large enough.

Remark 2.1 Since each $g^l = \prod_{i=1}^{d} \phi_i^l$, knowing g^l only defines $\{\phi_i^l\}$ up to scalars, since e.g. $\phi_i^l \phi_j^l = (c\phi_i^l)(c^{-1}\phi_j^l)$. When our analysis requires us to go from $\|g^l\|$ to $\{\|\phi_i^l\|\}$ we will impose a normalization convention that $\|\phi_i^l\| = \|\phi_j^l\|$, which implies $\|\phi_i^l\| = \|g^l\|^{1/d}$.

2.4 Posedness

Above, we set ourselves the goal of finding $g \in G_r$ with $E(g) = \hat{E}^r$. Unfortunately, this problem can be ill-posed.

Theorem 2.2 *There exist H, r, and f such that $\hat{E}^r$ is an infimum but not a minimum, and thus the minimization problem of finding $g \in G_r$ with $E(g) = \hat{E}^r$ is ill-posed.*

This fact has been observed several times; see [8] for a proof. The prototype example from [8] is

$$f = (0,1) \otimes (1,1) \otimes (1,1) + (1,1) \otimes (0,1) \otimes (1,1) + (1,1) \otimes (1,1) \otimes (0,1) \tag{17}$$

with $r = 2$ and H the space of $2 \times 2 \times 2$ tensors. For this $f \notin G_2$ there exists a sequence $g^{(k)} \in G_2$ with $E(g^{(k)}) \to \hat{E}^r = 0$ such that $\lim_{k\to\infty} g^{(k)} = f \notin G_2$. Thus there is no best approximation of that f for that r. The simplest illustration (from [2]) of what happens is the function $f(x_1, \ldots, x_d) = \sum_{i=1}^{d} \theta(x_i)$ for any

non-constant θ and $d > 2$. (This example includes (17) as a special case.) Defining the auxiliary function $q(t) = \prod_{i=1}^{d}(1 + t\theta(x_i))$, we see that $q'(0) = f$ by the product rule. Expressing the derivative as a limit (using a centered difference), we have the identity

$$f(x_1, \ldots, x_d) = \sum_{i=1}^{d} \theta(x_i) = \lim_{h \to 0} \left[\frac{1}{2h} \prod_{i=1}^{d}(1 + h\theta(x_i)) - \frac{1}{2h} \prod_{i=1}^{d}(1 - h\theta(x_i)) \right]. \tag{18}$$

Distributing the scalar $1/(2h)$ among the factors, this identity expresses f as a limit of elements of G_2. One notable aspect of (18) is that the sizes of the summands in the elements of G_2 go to infinity since $1/h$ does.

Theorem 2.3 *If there exists a sequence $g^{(k)} \in G_r$ with $E(g^{(k)}) \to \hat{E}^r$ such that the magnitudes $\|g^{l,(k)}\|$ remain bounded, then there exists $g \in G_r$ with $E(g) = \hat{E}^r$, and thus the minimization problem is well-posed for that f and r.*

In the general case, including infinite dimensional H, this theorem was established in [20]. That proof is too involved to include here, so instead for illustration we assume H is finite dimensional and give an argument adapted from [8]. *Proof (assuming H finite dimensional):* Using the normalization convention in Remark 2.1, $\|\phi_i^{l,(k)}\| = \|g^{l,(k)}\|^{1/d}$ and thus $\|\phi_i^{l,(k)}\|$ is also bounded. The assumption that H is finite-dimensional then implies that for each l and i the sequence $\{\phi_i^{l,(k)}\}$ lies in a bounded, finite-dimensional, and therefore compact set, and thus has a convergent subsequence. Thus we can select an infinite set $K_{1,1}$ of k so that the subsequence it defines has $\{\phi_1^{1,(k)}\}$ convergent. We can then select an infinite set $K_{1,2} \subset K_{1,1}$ such that $\{\phi_2^{1,(k)}\}$ is also convergent. Continuing in this way rd times we obtain a subsequence such that $\{\phi_i^{l,(k)}\}$ converges for all l and i. Using this final subsequence we define $\phi_i^l = \lim_{k\to\infty} \phi_i^{l,(k)}$, which thus yields $g \in G_r$ with $E(g) = \hat{E}^r$. Thus $\hat{E}^r$ is in fact a minimum, and the approximation problem is well-posed. □

The regularized error E_λ was introduced to prevent the norm-divergent behavior observed above for E. (It also controls the numerical loss-of-precision error due to ill-conditioning [2].) Since $E_\lambda(g) \geq \lambda \max_l\{\|g^l\|^2\}$, we know $\|g^l\|^2 \leq \lambda^{-1} E_\lambda(g)$ for all l. Thus as $E_\lambda(g^{(k)}) \to \hat{E}^r$ the norms $\|g^{l,(k)}\|$ remain bounded. We can then essentially apply the logic of the proof of Theorem 2.3. A rigorous proof, including infinite dimensional H, of the following theorem was established in [21] based on [20].

Theorem 2.4 *For all H, f, r, and $\lambda > 0$, there exists $g \in G_r$ with $E_\lambda(g) = \hat{E}^r_\lambda$, and thus the minimization problem of finding $g \in G_r$ with $E_\lambda(g) = \hat{E}^r_\lambda$ is well-posed.*

2.5 Uniqueness

Suppose we find $g \in G_r$ with $\|f - g\| = 0$. There are two trivial ways in which g is not unique. First, when $r > 1$ we can permute the summation index l. Second, we can move constants among the factors ϕ_i^l. There can also be non-trivial non-uniqueness. The simplest example (see [2, 17]) is illustrated by the identity

$$\sin\left(\sum_{i=1}^{d} x_i\right) = \sum_{l=1}^{d} \sin(x_l) \prod_{i=1, i\neq l}^{d} \frac{\sin(x_i + c_i - c_l)}{\sin(c_i - c_l)}, \tag{19}$$

which is valid for all choices of $\{c_i\}$ such that $\sin(c_i - c_l) \neq 0$ for all $i \neq l$. If our goal is just to find a good approximation of f then nonuniqueness may make our task easier. If we also want to be able to interpret elements of g, then nonuniqueness is annoying. See [16] for discussion of conditions under which the representation is essentially unique.

3 Linear Least Squares Minimization

In this section we pose the linear least squares minimization problem in a very general setting. We then translate standard solution methods to this setting. We will use these linear results within algorithms to solve the multilinear fitting problem. We assume that the order of integrals can be exchanged, that derivatives can pass through integrals, and that all integrals are finite.

Let H be a Hilbert space. Fix some "target" $f \in H$. Let α be some set of parameters; it may be discrete, continuous, or a mixture. We will need to distinguish between the identity of a parameter and the value which that parameter takes. We will use $s \in \alpha$ to specify which parameter we are talking about, and $t(s)$ to give the value of the parameter "indexed" by s. Suppose $g \in H$ depends linearly on the parameters in α with no constant term. With these assumptions and notation, we can write

$$g = \int_\alpha a(s)t(s)ds \tag{20}$$

for some fixed $a(s)$. We would like to minimize the error function

$$E(\alpha) = \|f - g\|^2 = \langle (f-g), (f-g) \rangle \tag{21}$$

over the values that the parameters in α can take.

For example, if $\alpha = \{1, 2, \ldots, r\}$ and $t(s)$ are scalars, then $a(s)$ should be elements of H and (20) becomes

$$g = \sum_{s=1}^{r} t(s)a(s)\,, \tag{22}$$

which is a simple linear combination. We then wish to determine the coefficients $\{t(s)\}_{s=1}^r$ to minimize (21).

3.1 Solution using Orthogonal Projectors

A projector $\mathcal{P}$ is a linear operator such that $\mathcal{P}^2 = \mathcal{P}$. Its complement $\mathcal{I}-\mathcal{P}$ is also a projector since $(\mathcal{I}-\mathcal{P})^2 = \mathcal{I}^2 - \mathcal{I}\mathcal{P} - \mathcal{P}\mathcal{I} + \mathcal{P}^2 = \mathcal{I} - \mathcal{P}$. The nullspace of $\mathcal{P}$ is exactly the range of $\mathcal{I} - \mathcal{P}$ since $\mathcal{P}v = 0 \Rightarrow v = (\mathcal{I}-\mathcal{P})v$ and $y = (\mathcal{I}-\mathcal{P})v \Rightarrow \mathcal{P}y = \mathcal{P}(\mathcal{I}-\mathcal{P})v = (\mathcal{P}-\mathcal{P}^2)v = 0$. For a linear operator $\mathcal{L}$, its adjoint $\mathcal{L}^*$ is defined to be the linear operator such that $\langle \mathcal{L}f, g\rangle = \langle f, \mathcal{L}^* g\rangle$. (In the case of matrices and vectors, the adjoint is the complex conjugate of the transpose of the matrix.) An orthogonal projector is a projector such that the range of $\mathcal{P}$ is orthogonal to the nullspace of $\mathcal{P}$. Since the nullspace of $\mathcal{P}$ is the range of $\mathcal{I}-\mathcal{P}$, a projector is orthogonal if and only if $\langle \mathcal{P}v, (\mathcal{I}-\mathcal{P})u\rangle = 0$ for all u and v. By the definition of the adjoint this is equivalent to $\langle v, \mathcal{P}^*(\mathcal{I}-\mathcal{P})u\rangle = 0$. If $\mathcal{P} = \mathcal{P}^*$ (i.e. $\mathcal{P}$ is self-adjoint) then we have $\langle v, (\mathcal{P}-\mathcal{P}^2)u\rangle = \langle v, 0\rangle = 0$ and so $\mathcal{P}$ is an orthogonal projector. One can show that the condition $\mathcal{P} = \mathcal{P}^*$ is also necessary.

Suppose we can construct an orthogonal projector onto functions of the form (20), where the $a(s)$ are fixed but $t(s)$ are allowed to vary. In other words, we have a linear operator $\mathcal{P}$ such that $\mathcal{P}^2 = \mathcal{P}$, $\mathcal{P} = \mathcal{P}^*$, and

$$g = \int_\alpha a(s)t(s)ds \Rightarrow \mathcal{P}g = g\,. \tag{23}$$

The element g of the given form (20) that minimizes $E = \|f-g\|^2$ is then given precisely by $\mathcal{P}f$. For any other g in the range of $\mathcal{P}$, we know $\mathcal{P}f - g$ is in the range of $\mathcal{P}$ and hence is orthogonal to $(\mathcal{I}-\mathcal{P})f$, which is in the nullspace of $\mathcal{P}$. Thus we have $\|f-g\|^2 = \|f-\mathcal{P}f\|^2 + \|\mathcal{P}f - g\|^2$ and so $\|f-g\| > \|f-\mathcal{P}f\|$.

We now set about constructing this projector. First let us endow the set of parameter values t with a Hilbert space structure by defining an inner product with respect to s by

$$\langle t, v\rangle_s = \int_\alpha t(s)\bar{v}(s)ds\,. \tag{24}$$

Define the linear operator $\mathcal{L}$ by

$$\mathcal{L}t(s') = \int_\alpha \langle a(s), a(s')\rangle t(s) ds \,. \tag{25}$$

The operator $\mathcal{L}$ is positive semi-definite since

$$\langle \mathcal{L}t, t\rangle_s = \int_\alpha \int_\alpha \langle a(s), a(s')\rangle t(s) ds\, \overline{t(s')} ds' = \left\langle \int_\alpha a(s)t(s)ds, \int_\alpha a(s')t(s')ds' \right\rangle \geq 0 \,. \tag{26}$$

If there exists $\tau_1 \geq \tau_2 > 0$ such that

$$\tau_1 \|t\| \geq \left\| \int_\alpha a(s)t(s)ds \right\| \geq \tau_2 \|t\| \tag{27}$$

then $\mathcal{L}$ is invertible. (If α is a finite set, this condition is equivalent to $\{a(s)\}_{s\in\alpha}$ being linearly independent, in which case (26) has a strict inequality when $\|t\| \neq 0$ and $\mathcal{L}$ is positive definite.) Notice that

$$\langle \mathcal{L}t, v\rangle_s = \int_\alpha \int_\alpha \langle a(s), a(s')\rangle t(s) ds\, \bar{v}(s') ds' = \int_\alpha t(s) \overline{\left(\int_\alpha \langle a(s'), a(s)\rangle v(s') ds' \right)} ds = \langle t, \mathcal{L}v\rangle_s \tag{28}$$

so $\mathcal{L} = \mathcal{L}^*$ and thus $\mathcal{L}$ is self-adjoint. If $\mathcal{L}$ is invertible, then $\mathcal{L}^{-1}$ will thus also be self-adjoint.

The projector $\mathcal{P}$ is defined by

$$\mathcal{P}f = \int_\alpha a(s) \left(\mathcal{L}^{-1} \langle f, a(\cdot)\rangle(s) \right) ds \,. \tag{29}$$

To verify that $\mathcal{P}^2 = \mathcal{P}$ we check

$$\mathcal{P}\mathcal{P}f = \int_\alpha a(s) \left(\mathcal{L}^{-1} \left\langle \int_\alpha a(s') \left(\mathcal{L}^{-1} \langle f, a(\cdot)\rangle(s') \right) ds', a(\cdot) \right\rangle (s) \right) ds \tag{30}$$

$$= \int_\alpha a(s) \left(\mathcal{L}^{-1} \int_\alpha \langle a(s'), a(s)\rangle \left(\mathcal{L}^{-1} \langle f, a(\cdot)\rangle(s') \right) ds' \right) ds \tag{31}$$

$$= \int_\alpha a(s) \left(\mathcal{L}^{-1} \mathcal{L} \left(\mathcal{L}^{-1} \langle f, a(\cdot)\rangle \right)(s) \right) ds = \mathcal{P}f \,. \tag{32}$$

To verify that $\mathcal{P}^* = \mathcal{P}$ we check

$$\langle \mathcal{P}f, v\rangle = \left\langle \int_\alpha a(s) \left(\mathcal{L}^{-1} \langle f, a(\cdot)\rangle(s) \right) ds, v \right\rangle = \int_\alpha \left(\mathcal{L}^{-1} \langle f, a(\cdot)\rangle(s) \right) \langle a(s), v\rangle ds \tag{33}$$

$$= \left\langle \mathcal{L}^{-1} \langle f, a(\cdot)\rangle, \langle v, a(\cdot)\rangle \right\rangle_s = \left\langle \langle f, a(\cdot)\rangle, \mathcal{L}^{-1} \langle v, a(\cdot)\rangle \right\rangle_s \tag{34}$$

$$= \int_\alpha \langle f, a(s)\rangle \overline{(\mathcal{L}^{-1} \langle v, a(\cdot)\rangle(s))} ds = \left\langle f, \int_\alpha a(s) \left(\mathcal{L}^{-1} \langle v, a(\cdot)\rangle(s) \right) ds \right\rangle = \langle f, \mathcal{P}v\rangle \,. \tag{35}$$

To verify (23), we check

$$\mathcal{P}g = \int_\alpha a(s) \left(\mathcal{L}^{-1} \left\langle \int_\alpha a(s')t(s')ds', a(\cdot) \right\rangle (s) \right) ds \tag{36}$$

$$= \int_\alpha a(s) \left(\mathcal{L}^{-1} \int_\alpha \langle a(s'), a(s)\rangle \, t(s')ds' \right) ds = \int_\alpha a(s) \left(\mathcal{L}^{-1} \mathcal{L}t(s) \right) ds = g \,. \tag{37}$$

Comparing (29) and (20), we see that the error function E in (21) is minimized by choosing $t(s) = \mathcal{L}^{-1}\langle f, a(\cdot)\rangle(s)$ in (20).

In the example where g is the simple sum (22), the definition (25) for the operator $\mathcal{L}$ becomes

$$\mathcal{L}t(s') = \sum_{s=1}^{r} \langle a(s), a(s') \rangle t(s) \,, \tag{38}$$

which means $\mathcal{L}$ is an $r \times r$ matrix $\mathbb{L}$ with entries $L(s', s) = \langle a(s), a(s') \rangle$. The definition (29) for the projector $\mathcal{P}$ becomes

$$\mathcal{P}f = \sum_{s=1}^{r} a(s) \sum_{s'=1}^{r} L^{-1}(s, s') \langle f, a(s') \rangle = \overline{\mathbf{a}}^* \mathbb{L}^{-1} \langle f, \mathbf{a} \rangle \,. \tag{39}$$

The optimal vector $\mathbf{t}$ of coefficients is given by $\mathbf{t} = \mathbb{L}^{-1} \langle f, \mathbf{a} \rangle$.

Remark 3.1 If $\mathcal{L}$ in (25) is not invertible, then the orthogonal projection is still well-defined, but there may not be a unique representation of it in terms of $a(s)$. For purposes of minimizing E, any representation will do. For example, in the discrete case, if $\mathbb{L}$ is not invertible then one can use a pseudo-inverse $\mathbb{L}^{\dagger}$ in place of $\mathbb{L}^{-1}$.

3.2 Solution using Gradients

In this section we assume H is real so that we can avoid the problem that $\bar{z}$ is not a differentiable function of z without resorting to writing $z = a + bi$ and differentiating separately in a and b.

Inserting (20) into (21), we have

$$E(\alpha) = \langle f, f \rangle + \langle g, g \rangle - 2\langle f, g \rangle \tag{40}$$

$$= \langle f, f \rangle + \left\langle \int_\alpha a(s')t(s')ds', \int_\alpha a(s)t(s)ds \right\rangle - 2 \left\langle f, \int_\alpha a(s)t(s)ds \right\rangle \tag{41}$$

$$= \langle f, f \rangle + \int_\alpha \int_\alpha \langle a(s'), a(s) \rangle t(s')t(s)ds'\, ds - 2 \int_\alpha \langle f, a(s) \rangle t(s) ds \,, \tag{42}$$

under the assumption that we can exchange the order of integration. Under the assumption that we can differentiate through an integral, we can explicitly compute the partial derivative

$$\frac{\partial}{\partial t(s')} E(\alpha) = \left(2 \int_\alpha \langle a(s), a(s') \rangle t(s) ds \right) - 2 \langle f, a(s') \rangle \,. \tag{43}$$

Assembling these partial derivatives into the gradient, we have

$$\nabla E(\alpha) = 2 \int \langle a(s), a(s') \rangle t(s) ds - 2 \langle f, a(s') \rangle = 2(\mathcal{L}t(s') - \langle f, a(s') \rangle) \,, \tag{44}$$

using $\mathcal{L}$ from (25). Setting the gradient equal to zero, we obtain the normal equations

$$\mathcal{L}t(s) = \langle f, a(s) \rangle \,. \tag{45}$$

Expressed as $t(s) = \mathcal{L}^{-1} \langle f, a(\cdot) \rangle(s)$, these are the same equations we obtained using orthogonal projectors in Section 3.1.

4 The Alternating Least Squares (ALS) Strategy

Suppose we wish to solve some least-squares problem that depends nonlinearly on some set of parameters. Assume we are given some initial guess at the correct parameter values. In its most abstract form, the ALS strategy is:

- Iterate until happy:
  - Select some subset of the parameters.
  - Solve the least-squares problem using those parameters while keeping the remaining parameters fixed.

To have any chance of solving the full problem, the subsets must be chosen so that each parameter is allowed to vary many times. Since we still must solve the inner least-squares problem, it is very convenient if the subsets chosen yield linear least-squares problems, which can then be solved with the methods in Section 3.

Within the multilinear setting of Section 2, we are interested in the least-squares problem of approximation with sums of products described in Section 2.3. Within this context there is a natural way to choose the subsets to yield linear problems. The ALS strategy becomes:

- Iterate until happy:
  - Loop through directions $k = 1, 2, \ldots, d$:
    - Fix $\{\phi_i^l\}$ for $i \neq k$ and solve the linear least-squares problem to minimize $E(g)$ with respect to $\{\phi_k^l\}_{l=1}^r$, thereby updating $\{\phi_k^l\}_{l=1}^r$.

We now consider how to formulate the one-directional problem so that we can use the methods in Section 3 to solve it. For notational convenience we consider the $k = 1$ case within the ALS loop, so the functions in $i = 2, 3, \ldots, d$ are fixed and $\{\phi_1^l\}_{l=1}^r$ are modified. The set of parameters α has elements of the form $s = (l, x_1)$ with $l = 1, \ldots, r$ and x_1 in the domain of definition of dx_1. The value of parameter s is $t(s) = \phi_1^l(x_1)$. The elements $a(s)$ in (20) should essentially be $\prod_{i=2}^d \phi_i^l(x_i)$, but we must be careful in their definition so that integration with them yields the desired results. In particular, we need an auxiliary integration variable $\hat{s} = (l, \hat{x}_1)$. We define

$$a(\hat{s}) = a(l, \hat{x}_1) = \delta(x_1 - \hat{x}_1) \prod_{i=2}^{d} \phi_i^l(x_i)\,, \tag{46}$$

where the delta function satisfies $\int \delta(x_1 - \hat{x}_1)\phi(\hat{x}_1)d\hat{x}_1 = \phi(x_1)$. Using $\hat{s}$ as the variable of integration, we then have that g in (20) reduces to g in (8) via

$$(20) \;=\; \int_\alpha a(\hat{s})t(\hat{s})d\hat{s} \;=\; \sum_{l=1}^{r} \int \delta(x_1 - \hat{x}_1) \left(\prod_{i=2}^{d} \phi_i^l(x_i) \right) \phi_1^l(\hat{x}_1) d\hat{x}_1 \;=\; \sum_{l=1}^{r} \prod_{i=1}^{d} \phi_i^l(x_i) \;=\; (8). \tag{47}$$

4.1 One-Directional Problem Solution using Orthogonal Projectors

With our definitions above for s, $t(s)$, and $a(s)$, the operator $\mathcal{L}$ in (25) becomes

$$\mathcal{L}t(l', x_1') = \sum_{l=1}^{r} \int \left\langle \delta(x_1 - \hat{x}_1) \prod_{i=2}^{d} \phi_i^l(x_i), \delta(x_1 - x_1') \prod_{i=2}^{d} \phi_i^{l'}(x_i) \right\rangle \phi_1^l(\hat{x}_1) d\hat{x}_1 \tag{48}$$

$$= \sum_{l=1}^{r} \phi_1^l(x_1') \prod_{i=2}^{d} \langle \phi_i^l, \phi_i^{l'} \rangle_i\,. \tag{49}$$

The projector $\mathcal{P}$ in (29) becomes

$$\mathcal{P}f = \sum_{l=1}^{r} \int \delta(x_1 - \hat{x}_1) \prod_{i=2}^{d} \phi_i^l(x_i) \left(\mathcal{L}^{-1} \left\langle f, \delta(x_1 - \cdot) \prod_{i=2}^{d} \phi_i^{(\cdot)}(x_i) \right\rangle (l, \hat{x}_1) \right) d\hat{x}_1 \tag{50}$$

$$= \sum_{l=1}^{r} \left(\mathcal{L}^{-1} \left\langle f, \prod_{i=2}^{d} \phi_i^{(\cdot)}(x_i) \right\rangle_{\setminus 1} (l, x_1) \right) \prod_{i=2}^{d} \phi_i^l(x_i) \,. \tag{51}$$

Thus we see that we should choose

$$t(s) = \phi_1^l(x_1) = \mathcal{L}^{-1} \left\langle f, \prod_{i=2}^{d} \phi_i^{(\cdot)}(x_i) \right\rangle_{\setminus 1} (l, x_1) \,. \tag{52}$$

We can simplify these expressions somewhat by noting that the operator $\mathcal{L}$ in (49) is separable. In mapping the discrete index, it acts as matrix multiplication from l to l'. In the continuous index it simply renames the variable to x_1'; by adjusting its name throughout, we can ignore this action. Define the matrix $\mathbb{L}$ with entries

$$L(l', l) = \prod_{i=2}^{d} \langle \phi_i^l, \phi_i^{l'} \rangle_i \,. \tag{53}$$

Define a column vector $\mathbf{q}$ whose entries are functions of x_1 by

$$q(l) = \phi_1^l(x_1) \,. \tag{54}$$

We can then write (49) as $\mathcal{L}t(l', x_1) = \mathbb{L}\mathbf{q}$. Define a column vector $\mathbf{p}$ whose entries are functions of $x_2, \ldots, x_d$ by

$$p(l) = \prod_{i=2}^{d} \phi_i^l(x_i) \tag{55}$$

and a column vector $\mathbf{b}$ whose entries are functions of x_1 by

$$b(l') = \left\langle f, \prod_{i=2}^{d} \phi_i^{l'}(x_i) \right\rangle_{\setminus 1} . \tag{56}$$

We can then write (51) as $\mathcal{P}f = \bar{\mathbf{p}}^* \mathbb{L}^{-1} \mathbf{b}$ and obtain the optimal $t(s) = \phi_1^l(x_1)$ in (52) as $\mathbf{t} = \mathbb{L}^{-1}\mathbf{b}$. Since each entry in $\mathbf{b}$ is in $\text{span}_{x_2,\ldots,x_d}\{f(\cdot, x_2, \ldots, x_d)\}$, we have $\phi_1^l \in \text{span}_{x_2,\ldots,x_d}\{f(\cdot, x_2, \ldots, x_d)\}$ for all l.

If f is of the form (16), then

$$b(l') = \left\langle \sum_{l=1}^{R} \prod_{i=1}^{d} \theta_i^l(x_i), \prod_{i=2}^{d} \phi_i^{l'}(x_i) \right\rangle_{\setminus 1} = \sum_{l=1}^{R} \theta_1^l(x_1) \prod_{i=2}^{d} \langle \theta_i^l, \phi_i^{l'} \rangle_i \tag{57}$$

and therefore the optimal $t(s) = \phi_1^l(x_1)$ in (52) are given as

$$\begin{bmatrix} \phi_1^1 \\ \phi_1^2 \\ \vdots \\ \phi_1^r \end{bmatrix} = \mathbb{L}^{-1}\mathbf{b} = \sum_{l=1}^{R} \theta_1^l(x_1) \mathbb{L}^{-1} \begin{bmatrix} \prod_{i=2}^{d} \langle \theta_i^l, \phi_i^1 \rangle_i \\ \prod_{i=2}^{d} \langle \theta_i^l, \phi_i^2 \rangle_i \\ \vdots \\ \prod_{i=2}^{d} \langle \theta_i^l, \phi_i^r \rangle_i \end{bmatrix} . \tag{58}$$

Thus we see $\phi_1^l \in \text{span}\{\theta_1^l\}_{l=1}^R$.

Remark 4.1 Since $g^l = \prod_{i=1}^d \phi_i^l(x_i)$, we could move constants around via $\phi_i^l \phi_j^l = (c\phi_i^l)(c^{-1}\phi_j^l)$. If we do so with $i \neq 1$ and $j \neq 1$, then $\mathbb{L}$, $\mathbf{b}$, and $\mathbf{t}$ are unchanged. If $j = 1 \neq i$, then column l and row l of $\mathbb{L}$ are multiplied by c, row l of $\mathbf{b}$ is multiplied by c, and thus row l of $\mathbf{t} = \mathbb{L}^{-1}\mathbf{b}$ is multiplied by c^{-1}. Since row l of $\mathbf{t}$ is the updated ϕ_1^l, the c^{-1} cancels with the c in ϕ_i^l, so the updated g^l is unaffected. Thus the normalization convention we use for the ϕ_i^l does not matter here.

4.1.1 Including Regularization

As noted in Section 2.3, when we replace the error measure E from (9) with E_λ from (13), we still have a least-squares problem, but now we approximate $(f, 0, \dots, 0) \in H^{r+1}$ with $(\sum_{l=1}^r g^l, g^1, \dots, g^r)$. The procedure is the same but the formulas change somewhat. We change (46) to

$$a(\hat{s}) = a(l, \hat{x}_1) = \left(\delta(x_1 - \hat{x}_1) \prod_{i=2}^d \phi_i^l(x_i) \right) (1, 0, \dots, 0, 1, 0, \dots), \tag{59}$$

where the second 1 is in position $l + 1$. The operator $\mathcal{L}$ in (49) becomes

$$\mathcal{L}t(l', x_1') = \sum_{l=1}^r \phi_1^l(x_1') \prod_{i=2}^d \langle \phi_i^l, \phi_i^{l'} \rangle_i + \lambda \phi_1^{l'}(x_1') \prod_{i=2}^d \langle \phi_i^{l'}, \phi_i^{l'} \rangle_i . \tag{60}$$

The final formula (51) for the projector $\mathcal{P}$ is unchanged as is the solution formula (52). The matrix $\mathbb{L}$ defined in (53) is changed to

$$L(l', l) = \prod_{i=2}^d \langle \phi_i^l, \phi_i^{l'} \rangle_i + \lambda \delta(l - l') \prod_{i=2}^d \langle \phi_i^{l'}, \phi_i^{l'} \rangle_i = \left(\prod_{i=2}^d \langle \phi_i^l, \phi_i^{l'} \rangle_i \right) (1 + \lambda \delta(l - l')) . \tag{61}$$

Thus we merely take the original $\mathbb{L}$ and inflate its diagonal by a factor of $1 + \lambda$.

Remark 4.2 If we had assumed $\|\phi_i^l\| = 1$ for $i > 1$ then including regularization in this way is equivalent to modifying the original $\mathbb{L}$ by adding $\lambda \mathbb{I}$ to it. Such normalization is advisable numerically, since it acts as a preconditioner on $\mathbb{L}$ and thus makes the linear system involving $\mathbb{L}$ easier to solve.

4.2 One-Directional Problem Solution using Gradients

As we saw in Section 3.2, solving a linear least squares problem using gradients results in the same set of equations as the method using orthogonal projectors. The gradient (44) becomes

$$\nabla E(\alpha) = 2(\mathbb{L}\mathbf{t} - \mathbf{b}) \tag{62}$$

and the normal equations (45) become

$$\mathbb{L}\mathbf{t} = \mathbf{b} , \tag{63}$$

which is equivalent to the equation $\mathbf{t} = \mathbb{L}^{-1}\mathbf{b}$ obtained above. Recall that the set α has elements of the form $s = (l, x_1)$, so the gradient is with respect to the parameters $t(s) = \phi_1^l(x_1)$ with $l = 1, \dots, r$ and x_1 in the domain of definition of dx_1. To use E_λ instead of E, we need only change the definition of $\mathbb{L}$ to (61).

Remark 4.3 As noted in Remark 4.1, the normalization convention does not affect the updated g^l obtained. It does however affect the gradient. If we move a constant c via $\phi_i^l \phi_1^l = (c\phi_i^l)(c^{-1}\phi_1^l)$, then row l of ∇E is multiplied by c.

4.3 Convergence of ALS

The ALS algorithm produces a sequence $g^{(k)} \in G_r$ such that $E(g^{(0)}) \geq E(g^{(k)}) \geq E(g^{(k+1)}) \geq \hat{E}^r \geq 0$ or $E_\lambda(g^{(0)}) \geq E_\lambda(g^{(k)}) \geq E_\lambda(g^{(k+1)}) \geq \hat{E}^r_\lambda \geq 0$, depending on whether E or E_λ was used. In this section we examine the behavior of these sequences. We organize our discussion around a set of questions.

4.3.1 Can the sequence $g^{(k)}$ diverge because $\|g^{(k)}\|$ becomes unbounded?

No. Using the triangle inequality, we have $\|g^{(k)}\| \leq \|f\| + \|f - g^{(k)}\|$, and then can use $\|f - g^{(k)}\| = E(g^{(k)})^{1/2} \leq E(g^{(0)})^{1/2} \leq E_\lambda(g^{(0)})^{1/2}$.

4.3.2 Does the magnitude of the change $\|g^{(k)} - g^{(k+1)}\|$ converge to zero?

Yes. The update in direction 1 as described in Section 4.1 replaces $\{\phi_1^{l,(k)}\}$ by $\{\phi_1^{l,(k+1)}\}$ and gives a partially updated function

$$g^{(k,1)} = \sum_{l=1}^{r} g^{l,(k,1)} = \sum_{l=1}^{r} \phi_1^{l,(k+1)}(x_1) \prod_{i=2}^{d} \phi_i^{l,(k)}(x_i)\,. \tag{64}$$

Similarly, when we next update in direction 2 we obtain $g^{(k,2)}$, and so on; for consistency we denote $g^{(k,0)} = g^{(k)}$ and $g^{(k,d)} = g^{(k+1)}$. Writing $g^{(k)} - g^{(k+1)}$ as a telescoping sum, applying the triangle inequality, and then applying Jensen's inequality, we obtain

$$\|g^{(k)} - g^{(k+1)}\|^2 \leq \left(\sum_{i=1}^{d} \|g^{(k,i-1)} - g^{(k,i)}\| \right)^2 \leq d \sum_{i=1}^{d} \|g^{(k,i-1)} - g^{(k,i)}\|^2\,. \tag{65}$$

We first consider the case for E. In Section 4 we showed that each step of the ALS algorithm finds the linear least squares fit to f in the subspace spanned by the $a(\hat{s})$ in (46). In Section 4.1 we showed how to find this fit by doing an orthogonal projection onto this subspace, obtaining $g^{(k,1)}$. Since $g^{(k)}$ is also in this subspace, $g^{(k)}$, $g^{(k,1)}$, and f form a right triangle and we have

$$\|f - g^{(k)}\|^2 = \|f - g^{(k,1)}\|^2 + \|g^{(k)} - g^{(k,1)}\|^2\,. \tag{66}$$

Rearranging, we have

$$E(g^{(k)}) - E(g^{(k,1)}) = \|g^{(k)} - g^{(k,1)}\|^2\,. \tag{67}$$

Summming (67) we obtain $E(g^{(k)}) - E(g^{(k+1)}) = \sum_{i=1}^{d} \|g^{(k,i-1)} - g^{(k,i)}\|^2$, which combined with (65) yields

$$\|g^{(k)} - g^{(k+1)}\|^2 \leq d(E(g^{(k)}) - E(g^{(k+1)}))\,. \tag{68}$$

Since $E(g^{(k)})$ is decreasing and bounded below, we know $\big(E(g^{(k)}) - E(g^{(k+1)})\big) \to 0$ and the sequence $\{\big(E(g^{(k)}) - E(g^{(k+1)})\big)\}$ is in ℓ^1 (has finite sum). Thus $\|g^{(k)} - g^{(k+1)}\|^2 \to 0$ and is in ℓ^1.

When using E_λ, the right triangle is formed by $(g^{(k)}, g^{1,(k)}, \ldots, g^{r,(k)})$, $(g^{(k,1)}, g^{1,(k,1)}, \ldots, g^{r,(k,1)})$, and $(f, 0, \ldots, 0)$. We have

$$E_\lambda(g^{(k)}) - E_\lambda(g^{(k,1)}) = \|g^{(k)} - g^{(k,1)}\|^2 + \lambda \sum_{l=1}^{r} \|g^{l,(k)} - g^{l,(k,1)}\|^2 \tag{69}$$

and thus

$$E_\lambda(g^{(k)}) - E_\lambda(g^{(k+1)}) = \sum_{i=1}^{d} \|g^{(k,i-1)} - g^{(k,i)}\|^2 + \sum_{i=1}^{d} \lambda \sum_{l=1}^{r} \|g^{l,(k,i-1)} - g^{l,(k,i)}\|^2\,. \tag{70}$$

We thus obtain

$$\|g^{(k)} - g^{(k+1)}\|^2 \leq d(E_\lambda(g^{(k)}) - E_\lambda(g^{(k+1)}))\,. \tag{71}$$

4.3.3 Does $\|g^{(k)} - g^{(k+1)}\|$ converging to zero imply $g^{(k)}$ converges?

No. In order to conclude $g^{(k)}$ converges we need the increments $\|g^{(k)} - g^{(k+1)}\|$ to be in ℓ^1. We have shown $\|g^{(k)} - g^{(k+1)}\|^2$ to be in ℓ^1 but that means only $\|g^{(k)} - g^{(k+1)}\|$ is in ℓ^2, which is not sufficient. Thus it may be possible to have cycles or other bounded forms of divergence, although we have not found an example where this occurs.

4.3.4 What can we say about accumulation points of $g^{(k)}$?

We first assume H is finite dimensional. We know from Section 4.3.1 that $g^{(k)}$ remains bounded, and thus wehave a bounded sequence in a finite dimensional space, which must have an accumulation point. The set K of all accumulation points is closed and bounded and therefore compact. In Section 4.3.2 we showed that $\|g^{(k)} - g^{(k+1)}\| \to 0$ and thus $g^{(k)}$ uses steps of size going to zero. We will use this to show that K is also connected. If K is not connected, then we can write $K = A \cup B$ with $A \cap B = \emptyset$ and A and B nonempty and both relatively open and relatively closed within K. Since A and B are relatively closed subsets of a compact set, they are also compact. Since they are disjoint and compact, the minimum distance between them is some $\eta > 0$. Let N_A and N_B be open $\eta/4$ neighborhoods of A and B. The sequence $g^{(k)}$ must visit both N_A and N_B infinitely often. Once the step size is smaller than $\eta/2$ then the sequence must have an element outside $N_A \cup N_B$. Since $g^{(k)}$ must have such an element each time it moves from N_A to N_B, there are an infinite number of such elements outside $N_A \cup N_B$. These points must then have an accumulation point that is not in $N_A \cup N_B$ and thus not in K. This contradicts the definition of K and thus K must be connected. In particular, if K contains more than one point then it contains an uncountably infinite number of points.

If $f \in G_R$ for some finite R as in (16), then Section 4.1 showed that after one ALS pass we have $\phi_i^l \in \text{span}\{\theta_i^l\}_{l=1}^R$. Since this span is finite dimensional, both f and g can be embedded into a finite dimensional subspace of H, and it is as if H were finite dimensional.

When H is infinite dimensional and $f \notin G_R$ for any finite R, then we cannot use these arguments. We note that the set of $f \in G_R$ for finite R is dense in H.

4.3.5 Can $g^{(k)}$ converge to a local minimum?

Yes. Since the error measures E and E_λ are not convex, we expect to encounter local minima. The simplest example (inspired by [4]) is

$$f = 2 \cdot (1,0) \otimes (1,0) \otimes (1,0) + (0,1) \otimes (0,1) \otimes (0,1) \tag{72}$$

using $r = 1$ and $g^{(0)} \approx (0,1) \otimes (0,1) \otimes (0,1)$. The ALS algorithm (using E) will converge to $(0,1) \otimes (0,1) \otimes (0,1)$ even though $E((0,1) \otimes (0,1) \otimes (0,1)) = 4 > 1 = E(2 \cdot (1,0) \otimes (1,0) \otimes (1,0))$. To demonstrate the behavior of ALS on this example while avoiding excess parameters, we write our approximation of the form $g = s \cdot (a,1) \otimes (b,1) \otimes (c,1)$. We will show that for a, b, c sufficiently small the ALS algorithm preserves this form, $(a,b,c) \to (0,0,0)$, and $s \to 1$. Updating in direction one, we see $\mathbb{L} = [(1+b^2)(1+c^2)]$ and $\mathbf{b} = [(2bc,1)]$ and thus our update is $s \cdot (a,1) = ((1+b^2)(1+c^2))^{-1} \cdot (2bc,1)$. Assuming $1/4 > |b|$ and $1/8 > |c|$, we then have $|a| = |2bc| < |c|/2 < 1/16$. Updating in direction two we obtain $|b| = |2ac| < 1/64$ and then in direction three we obtain $|c| = |2ab| < 1/512$. Continuing, we have $(a,b,c) \to (0,0,0)$ and then $s \to 1$. We omit the proof that this point is in fact a local minimum. Using E_λ instead of E does not prevent convergence to a local minimum; on this example it changes $\mathbb{L}$ to $[(1+b^2)(1+c^2)(1+\lambda)]$ and the final approximation to $(1+\lambda)^{-1} \cdot (0,1) \otimes (0,1) \otimes (0,1)$.

4.3.6 Is the convergence of $g^{(k)}$ the correct question?

No. Since $g^{(k)} = \sum_l g^{l,(k)}$, we could have the representation (8) diverge because some of the $g^{l,(k)}$ diverge, but these divergences cancel allowing $g^{(k)}$ to converge to some $g \notin G_r$. A better question is the convergence of all the $g^{l,(k)}$. We collect them into a vector $\mathbf{u}^{(k)} = (g^{1,(k)}, \ldots, g^{r,(k)})$ and define its norm by $\|\mathbf{u}^{(k)}\|^2 = \sum_l \|g^{l,(k)}\|^2$.

4.3.7 Can the representation diverge because $\|\mathbf{u}^{(k)}\|$ becomes unbounded?

Yes, when using E. As stated in Theorem 2.2, the minimization problem of finding $g \in G_r$ with $E(g) = \hat{E}^r$ is ill-posed in general. Since the problem of finding $\hat{E}^r$ is ill-posed, any "good" algorithm *should* diverge, in the sense that some of the $g^{l,(k)}$ diverge. To illustrate this behavior, we consider

$$f = (0,1) \otimes (1,0) \otimes (1,0) + (1,0) \otimes (0,1) \otimes (1,0) + (1,0) \otimes (1,0) \otimes (0,1), \tag{73}$$

which is of the form (18) and has the same behavior as (17). We consider $r = 2$ approximations with initial guess of the form

$$g = s \cdot (1,a) \otimes (1,b) \otimes (1,c) - s \cdot (1,-a) \otimes (1,-b) \otimes (1,-c) \tag{74}$$

with a, b, c, and s positive, which is the form used in (18). We will show that this form is preserved under the ALS updates, $(a,b,c) \to (0,0,0)$, and $s \to \infty$. Updating in direction one, we can compute

$$\mathbb{L} = \begin{bmatrix} (1+b^2)(1+c^2) & (1-b^2)(1-c^2) \\ (1-b^2)(1-c^2) & (1+b^2)(1+c^2) \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} ((b+c),1) \\ (-(b+c),1) \end{bmatrix}, \tag{75}$$

and thus after some algebra

$$\mathbb{L}^{-1}\mathbf{b} = \begin{bmatrix} s \cdot (1,a) \\ -s \cdot (1,-a) \end{bmatrix} \quad \text{with} \quad s = \frac{b+c}{2(b^2+c^2)} \quad \text{and} \quad a = \frac{b^2+c^2}{(b+c)(1+b^2c^2)}\,. \tag{76}$$

Updating b and c in turn, we see that a, b, and c are determined by the recurrence

$$x_n = \frac{x_{n-1}^2 + x_{n-2}^2}{(x_{n-1}+x_{n-2})(1+x_{n-1}^2x_{n-2}^2)} = x_{n-2}\frac{(x_{n-1}/x_{n-2})^2+1}{(x_{n-1}/x_{n-2}+1)(1+x_{n-1}^2x_{n-2}^2)}\,, \tag{77}$$

so it suffices to show this converges to zero. From the second form of (77) we can see that $0 < x_{n-1} < x_{n-2}$ implies $0 < x_n < x_{n-2}$. Since the first form of (77) is symmetric, we in fact have that $0 < x_{n-1}$ and $0 < x_{n-2}$ implies $0 < x_n < \max\{x_{n-1}, x_{n-2}\}$. Applying the recurrence again, we have $0 < x_{n+1} < \max\{x_n, x_{n-1}\}$ and thus $0 < \max\{x_{n+1}, x_n\} < \max\{x_{n-1}, x_{n-2}\}$. Since this pairwise maximum is decreasing and bounded below by 0, it must converge to some $p \geq 0$. Now consider the two-dimensional sequence (x_{2n}, x_{2n+1}) defined by double application of (77). Since the coordinates are positive and bounded by $\max\{x_0, x_1\}$, this sequence must have a nonempty, closed set of accumulation points. Since these points must have one coordinate equal to p, they must be of the form (p,q) or (q,p) with $0 \leq q \leq p$. A basic result in dynamical systems (see e.g. the texts [5, 19]) is that the continuity of the defining recurrence implies that the set of accumulation points is a closed invariant set of the (double) recurrence. If $q > 0$ then applying the double recurrence to (p,q) or (q,p) results in a point with both coordinates strictly less than p, and thus neither (p,q) nor (q,p) can be in the invariant set. Thus $q = 0$ and the invariant set can only contain $(0,p)$ and/or $(p,0)$. The point $(0,p)$ leads to $(p, p/(1+p^4))$ and the point $(p,0)$ leads to (p,p), which then must be of the form $(0,p)$ or $(p,0)$, and thus $p = 0$. Therefore, we have $x_n \to 0$ and then (76) shows $s \to \infty$. We note that $x_{n-2} \approx x_{n-1}$ implies $x_n \approx x_{n-1}/(1+x_{n-1}^4)$, so the convergence $x_n \to 0$ is extremely slow.

No, when using E_λ. As noted in the well-posedness discussion before Theorem 2.4, we have $\|g^{l,(k)}\|^2 \leq \lambda^{-1}E_\lambda(g^{(0)})$ for all l and k. In Figure 1 we illustrate the difference in behavior between E and E_λ for the example (73) approximated by (74).

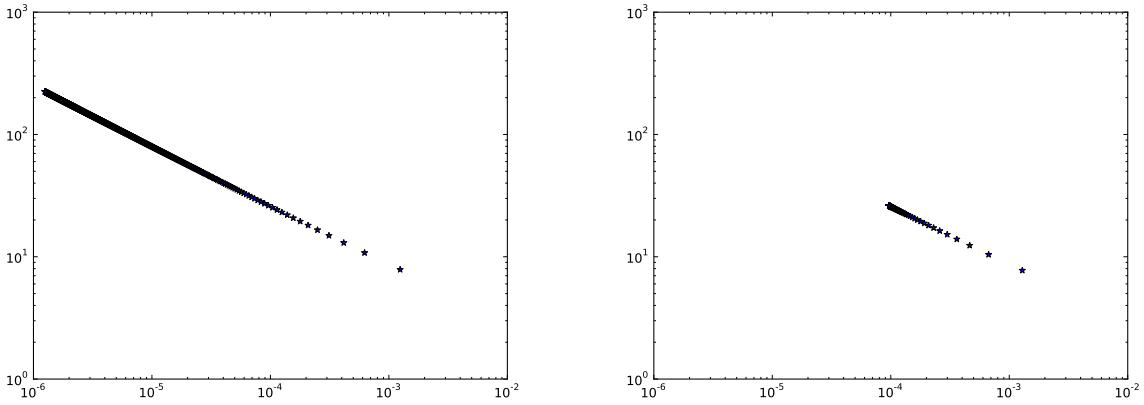

Figure 1: Behavior of the ALS algorithm for (73) approximated by (74) starting with $s = b = c = 1$. The horizontal axis is $\|f - g^{(k)}\|^2$ and the vertical is $\max_l\{\|g^{l,(k)}\|^2\}$, both on a log scale. A point is plotted every one hundred k up to $k = 100000$. The left plot is using E, and the norms of terms would continue to grow and so divergence. The right plot is using E_λ with $\lambda = 10^{-4}$, and indicates convergence of both error and norms.

4.3.8 Does the magnitude of the change $\|\mathbf{u}^{(k)} - \mathbf{u}^{(k+1)}\|$ converge to zero, and if so does this imply $\mathbf{u}^{(k)}$ converges?

For E, our best bounds from Section 4.3.2 give us no control over $\|g^{l,(k)} - g^{l,(k+1)}\|$, and thus no control over $\|\mathbf{u}^{(k)} - \mathbf{u}^{(k+1)}\|$. For E_λ, we can use (70) to obtain

$$\|\mathbf{u}^{(k)} - \mathbf{u}^{(k+1)}\|^2 \leq \sum_{l=1}^{r} d \sum_{i=1}^{d} \|g^{l,(k,i-1)} - g^{l,(k,i)}\|^2 \leq d\lambda^{-1}(E_\lambda(g^{(k)}) - E_\lambda(g^{(k+1)}))\,. \tag{78}$$

As in Section 4.3.2, this means $\|\mathbf{u}^{(k)} - \mathbf{u}^{(k+1)}\|^2 \to 0$ and is in ℓ^1, but that is not sufficient to imply $\mathbf{u}^{(k)}$ converges.

4.3.9 What can we say about accumulation points of $\{\mathbf{u}^{(k)}\}$?

We can only say something when H is finite dimensional or $f \in G_R$, and $\|\mathbf{u}^{(k)}\|$ remains bounded. Then the argument of Section 4.3.4 shows $\mathbf{u}^{(k)}$ has accumulation points, the set of which must then be compact. For E_λ, we noted in Section 4.3.8 that $\mathbf{u}^{(k)}$ uses steps of size going to zero; by the same argument as used in Section 4.3.4, K is then also connected.

4.3.10 Is the convergence of $\mathbf{u}^{(k)}$ the correct question?

Almost. We need the representation of the form (8) to converge. That means we need convergence for all the parameters $\phi_i^{l,(k)}(x_i)$ with $l = 1, \ldots, r$, $i = 1, \ldots, d$, and x_i in the domain of definition of dx_i. We can collect these parameters into a vector $\mathbf{v}^{(k)} = (\phi_1^{1,(k)}, \ldots, \phi_1^{r,(k)}, \ldots, \phi_d^{1,(k)}, \ldots, \phi_d^{r,(k)})$ and define its norm by $\|\mathbf{v}^{(k)}\|^2 = \sum_l \sum_i \|\phi_i^{l,(k)}\|^2$. Knowing $\mathbf{u}^{(k)}$ and thus $g^{l,(k)} = \prod \phi_i^{l,(k)}$ leaves ambiguity in $\phi_i^{l,(k)}$ since scalars can be moved between factors. Once we remove this ambiguity, the convergence of $\mathbf{u}^{(k)}$ implies convergence of $\mathbf{v}^{(k)}$. We can remove the ambiguity in $\|\phi_i^{l,(k)}\|$ by applying the normalization convention

of Remark 2.1 that $\|\phi_i^{l,(k)}\| = \|\phi_j^{l,(k)}\|$. Removing the ambiguity in signs (or complex units) is a bit more cumbersome. Supposing $\mathbf{u}^{(k)}$ converges to $\mathbf{u}$, we can extract g^l from it, and then extract ϕ_i^l from it utilizing the normalization convention but allowing arbitrary sign. We can then define a sign convention on $\mathbf{v}^{(k)}$ by moving factors of -1 (or complex units) to $\phi_1^{l,(k)}$ to make $\langle \phi_i^l, \phi_i^{l,(k)} \rangle \geq 0$ for $i > 1$.

4.3.11 Can the representation diverge because $\|\mathbf{v}^{(k)}\|$ becomes unbounded?

Yes when using E, but no when using E_λ. The boundedness of $\|\mathbf{v}^{(k)}\|$ follows the boundedness of $\|\mathbf{u}^{(k)}\|$, which was determined in Section 4.3.7.

4.3.12 Does the magnitude of the change $\|\mathbf{v}^{(k)} - \mathbf{v}^{(k+1)}\|$ converge to zero, and if so does this imply $\mathbf{v}^{(k)}$ converges?

As with $\mathbf{u}^{(k)}$ discussed in Section 4.3.8, we know $\|\mathbf{v}^{(k)} - \mathbf{v}^{(k+1)}\| \to 0$ for E_λ but not for E. Even for E_λ this does not imply $\mathbf{v}^{(k)}$ converges.

4.3.13 What can we say about accumulation points of $\{\mathbf{v}^{(k)}\}$?

The results for $\{\mathbf{v}^{(k)}\}$ are the same as for $\{\mathbf{u}^{(k)}\}$ in Section 4.3.9.

When H is infinite dimensional and $f \notin G_R$ for any finite R, we could hope to use the methods of [21, 20], which provided the posedness results in Section 2.4 for the infinite dimensional case. However, those methods are non-constructive and do not yield the desired result.

4.3.14 If v is a fixed point of the ALS algorithm, is the gradient zero at v?

Yes. As shown in Section 4.2, the ALS algorithm works by solving to make the partial gradient $\nabla E(\alpha) = 0$, where for the current direction i the gradient is with respect to the parameters $\phi_i^l(x_i)$ with $l = 1, \ldots, r$ and x_i in the domain of definition of dx_i. We can denote this partial gradient $\nabla_i E(g)$ and note that the full gradient $\nabla E(g)$ is obtained by concatenating the $\nabla_i E(g)$. If g as defined by $\mathbf{v}$ is not changed by one full cycle of the ALS algorithm, then the total gradient $\nabla E(g)$ must have already been zero. Similarly, if g is a fixed point of the ALS algorithm using E_λ, then $\nabla E_\lambda(g) = 0$.

4.3.15 Is the gradient zero at every accumulation point of $\mathbf{v}^{(k)}$?

Each $\mathbf{v}^{(k)}$ defines $g^{(k)}$ and an accumulation point $\mathbf{v}$ defines $g \in G_r$. Since the gradients ∇_i are computed by (62), which involves no inverses or divisions, they are well-defined and continuous. By the definition of accumulation point, for every $\epsilon > 0$ there exists k such that $\|\mathbf{v} - \mathbf{v}^{(k)}\| < \epsilon$. Since the last step in forming $\mathbf{v}^{(k)}$ was to update in direction d, we know $\nabla_d E(g^{(k)}) = 0$. Thus for every $\epsilon > 0$ there exists a point within ϵ of $\mathbf{v}$ such that the partial gradient in direction d is zero. By continuity of the gradient, we thus have $\nabla_d E(g) = 0$. For E, it appears that this is all we can say, although we have no example where the full gradient $\nabla E(g)$ is not zero at an accumulation point.

Yes for E_λ, under the assumption that $\|g^l\| \neq 0$ for all l. As above, we already know $\nabla_d E_\lambda(g) = 0$. In Section 4.3.2 we updated $g^{(k)}$ in direction one to obtain $g^{(k,1)}$, whose components we denote $\mathbf{v}^{(k,1)}$. From (78) we have

$$\|g^{l,(k)} - g^{l,(k,1)}\| = \|\phi_1^{l,(k)} - \phi_1^{l,(k+1)}\| \prod_{i=2}^{d} \|\phi_i^{l,(k)}\| \to 0\,. \tag{79}$$

Assuming $\|g^{l,(k)}\| \not\to 0$, that implies $\|\phi_1^{l,(k)} - \phi_1^{l,(k+1)}\| \to 0$. We can then choose k so that $\|\mathbf{v} - \mathbf{v}^{(k)}\| < \epsilon/2$ and $\|\mathbf{v}^{(k)} - \mathbf{v}^{(k,1)}\| < \epsilon/2$, so that $\|\mathbf{v} - \mathbf{v}^{(k,1)}\| < \epsilon$. By construction, $\nabla_1 E(g^{(k,1)}) = 0$. Thus, for every $\epsilon > 0$

there exists a point within ϵ of $\mathbf{v}$ such that the partial gradient in direction 1 is zero. By continuity of the gradient, we thus have $\nabla_1 E(g) = 0$. By the same logic, when we update $g^{(k,1)}$ in direction two to obtain $g^{(k,2)}$, we also have $\|\phi_2^{l,(k)} - \phi_2^{l,(k+1)}\| \to 0$, we can find k with $\|\mathbf{v} - \mathbf{v}^{(k,2)}\| < \epsilon$, we have $\nabla_2 E_\lambda(g^{(k,2)}) = 0$, and thus $\nabla_2 E_\lambda(g) = 0$. Continuing this process through all d directions, we obtain $\nabla E_\lambda(g) = 0$.

Remark 4.4 The above proof was inspired by the proof in [12] that the gradient is zero at accumulation points of a block nonlinear Gauss-Seidel iteration, under certain assumptions. (See also [18, Chapter 14].) To apply the results in [12] one needs to show the objective function has strict componentwise quasiconvexity. In our context, this means assuming the $\mathbb{L}$ for all directions are nonsingular in an appropriate region. For E, checking the assumption that $\mathbb{L}$ stays nonsingular is no easier than checking the desired conclusion that the gradient is zero, so it appears nothing is gained. For E_λ, the matrix $\mathbb{L}$ given by (60) is nonsingular as long as $\|g^l\| \neq 0$ for all l, which is the same assumption we used above. (The proof in [12] was for a finite dimensional case, but it is not clear if that assumption was essential.)

4.3.16 What is left unresolved?

We left open the possibility of some bounded form of divergence for $g^{(k)}$ in Section 4.3.3, for $\mathbf{u}^{(k)}$ in Section 4.3.8, and for $\mathbf{v}^{(k)}$ in Section 4.3.12. Since we have no examples of this type, it is unclear if this is really possible. Numerically, it would be very difficult to discover such a case, because we must determine that the increments, which are in ℓ^2, are not in ℓ^1. Analytically, we can imagine a situation where the accumulation points lie on a circle and the sequence approaches the circle while rotating fast enough to avoid converging to any particular point on the circle. Such behavior occurs in dynamical systems (see e.g. the texts [5, 19]) so it may occur here. Since in our context we have a rather complicated iteration, it appears to be very difficult to show anything using techniques from dynamical systems.

We did not resolve the case of infinite dimensional H and $f \notin G_R$ in Sections 4.3.4, 4.3.9, and 4.3.13. In Section 4.3.15 when using E we left open the possibility that the gradient is not zero at an accumulation point, but we have no examples showing this actually occurs.

We did not discuss how to choose a good starting point to avoid local minima. We cannot say anything rigorous on this point, but experience has shown the strategy of "growing" r has a dramatically beneficial effect. First one fits using $r = 1$, then adds to that a small random separable term to obtain $r = 2$, fits again, and continues until either the desired r is obtained or the error is sufficiently low.

5 Infinite Dimensional Multilinear Setting

Recall the multilinear setting from Section 2 but let there be a countably infinite number of variables x_i for $i = 1, 2, \ldots$. Our goal is to show how simple modifications of the ALS algorithm can operate in this context, and challenge developers of other algorithms to make sure theirs do so as well. Since one cannot implement a truly infinite algorithm, the framework here is meant for thought-experiments, rather than actual computations.

Fix an element

$$f(\mathbf{x}) = \sum_{l=1}^{R} \prod_{i=1}^{\infty} \theta_i^l(x_i)\,. \tag{80}$$

We wish to approximate f with a function g of the form

$$g(\mathbf{x}) = \sum_{l=1}^{r} \prod_{i=1}^{\infty} \phi_i^l(x_i) \tag{81}$$

with some $1 \leq r < R$. We assume that we can find an initial g with convergent sums, products, and integrals and with $\langle f, g\rangle \neq 0$. We will assume all infinite products over the directions are absolutely convergent, so the order of multiplication does not matter.

We assume that we have one regular computational node per direction and that these can operate in parallel to do the same operation on all directions at once. We assume there is one central node that can communicate simultaneously with all directional nodes and can take infinite products. Directional node i has available to it complete information on $\{\theta_i^l(x_i)\}_{l=1}^R$ and $\{\phi_i^l(x_i)\}_{l=1}^r$ and whatever (scalar) infinite product information is passed from the central node. In particular it does not have access to θ_j^l or ϕ_j^l for $j \neq i$.

5.1 Simultaneous ALS Update Algorithm

The first algorithm computes the orthogonal projections for each direction and then applies all of them simultaneously.

- Iterate until happy:
  - In parallel, the directional nodes for all i:
    - Compute $A_i(l,l') = \langle \phi_i^l, \phi_i^{l'} \rangle_i$ for $l = 1, \ldots r$ and $l' = 1, \ldots, r$
    - Compute $B_i(l,l') = \langle \phi_i^l, \theta_i^{l'} \rangle_i$ for $l = 1, \ldots r$ and $l' = 1, \ldots, R$
    - Send $A_i(l,l')$ and $B_i(l,l')$ to the central node.
  - The central node:
    - Computes $A(l,l') = \prod_{i=1}^{\infty} A_i(l,l')$ for $l = 1, \ldots r$ and $l' = 1, \ldots, r$.
    - Computes $B(l,l') = \prod_{i=1}^{\infty} B_i(l,l')$ for $l = 1, \ldots r$ and $l' = 1, \ldots, R$.
    - Sends $A(l,l')$ and $B(l,l')$ to all directional nodes.
  - In parallel, the directional nodes for all i:
    - Compute $\mathbb{L}_i$ with entries
    $$L_i(l,l') = A(l,l')/A_i(l,l')\,. \tag{82}$$
    - Compute a column vector $\mathbf{b}_i$ whose entries are functions of x_1 by
    $$b_i(l) = \sum_{l'=1}^{R} \theta_i^{l'}(x_i) B(l,l')/B_i(l,l')\,. \tag{83}$$
    - Solve for $\mathbf{t}_i$ in
    $$\mathbb{L}_i \mathbf{t}_i = \mathbf{b}_i\,. \tag{84}$$
    - Update ϕ_i^l to be $\tilde{\phi}_i^l = t_i(l)$.

The algorithm is clean except for potential divisions by zero. (When such a number is computed as zero, a very small arbitrary number could be used instead.) If the simultaneous updates are sufficiently small it should converge, but for arbitrary starting points it (probably) can diverge. One could change the simultaneous parallel steps to asynchronous parallel steps. This should be more efficient, but even harder to analyze.

5.2 Greedy ALS Algorithm

This algorithm computes the orthogonal projections for each direction and then applies the one that appears most useful.

- Iterate until happy:
  - In parallel, the directional nodes for all i: Compute and send $A_i(l, l')$ and $B_i(l, l')$.
  - The central node: Compute and send $A(l, l')$ and $B(l, l')$.
  - In parallel, the directional nodes for all i:
    - Compute $\mathbb{L}_i$ and $\mathbf{b}_i$.
    - Solve $\mathbb{L}_i \mathbf{t}_i = \mathbf{b}_i$ for $\mathbf{t}_i$.
    - Compute $d_i = \sum_{l=1}^{r} \|\phi_i^l - t_i(l)\|^2$ send to the central node.
  - The central node selects the maximum d_i and sends the instruction for the node that produced it to update its $\phi_i^l(x_i)$ to be $t_i(l) = \tilde{\phi}_i^l(x_i)$.

The central node needs the additional ability to choose the maximum of a countably infinite set of numbers and send an instruction to the directional node that supplied that maximum. One could relax this condition to finding a direction within some fixed fraction of the maximum/supremum. Like ALS, this algorithm can never increase the error. The criterion used above to select which direction to update was the greatest change in $\phi_i^l(x_i)$. One could instead use the largest (directional) gradient, most improvement in $\|f - g\|$, or other criteria.

Acknowledgments

This work was inspired by participation in the workshop *Computational Optimization for Tensor Decompositions* at the American Institute of Mathematics in 2010. Thanks to Todd Young for his assistance with the dynamical system in Section 4.3.7.

References

[1] G. Beylkin and M. J. Mohlenkamp. Numerical operator calculus in higher dimensions. *Proc. Natl. Acad. Sci. USA*, 99(16):10246–10251, August 2002.

[2] G. Beylkin and M. J. Mohlenkamp. Algorithms for numerical analysis in high dimensions. *SIAM J. Sci. Comput.*, 26(6):2133–2159, July 2005.

[3] G. Beylkin, M. J. Mohlenkamp, and F. Pérez. Approximating a wavefunction as an unconstrained sum of Slater determinants. *Journal of Mathematical Physics*, 49(3):032107, 2008.

[4] Ryan Botts. *Recovery and Analysis of Regulatory Networks from Expression Data Using Sums of Separable Functions*. PhD thesis, Ohio University, Athens, OH, 2010.

[5] Michael Brin and Garrett Stuck. *Introduction to dynamical systems*. Cambridge University Press, Cambridge, 2002.

[6] Hans-Joachim Bungartz and Michael Griebel. Sparse grids. *Acta Numer.*, 13:147–269, 2004.

[7] Sambasiva Rao Chinnamsetty, Mike Espig, Boris N. Khoromskij, Wolfgang Hackbusch, and Heinz-Juergen Flad. Tensor product approximation with optimal rank in quantum chemistry. *Journal of Chemical Physics*, 127(8):Art. No. 084110, August 2007.

[8] Vin de Silva and Lek-Heng Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications, Special Issue on Tensor Decompositions and Applications*, 30(3):1084–1127, 2008.

[9] Ivan P. Gavrilyuk, Wolfgang Hackbusch, and Boris N. Khoromskij. Hierarchical tensor-product approximation to the inverse and related operators for high-dimensional elliptic problems. *Computing*, 74(2):131–157, 2005.

[10] M. Griebel and S. Knapek. Optimized general sparse grid approximation spaces for operator equations. *Math. Comp.*, 78(268):2223–2257, 2009.

[11] Michael Griebel and Jan Hamaekers. Sparse grids for the Schrödinger equation. *M2AN Math. Model. Numer. Anal.*, 41(2):215–247, 2007.

[12] L. Grippo and M. Sciandrone. On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Oper. Res. Lett.*, 26(3):127–136, 2000.

[13] W. Hackbusch and B. N. Khoromskij. Low-rank Kronecker-product approximation to multi-dimensional nonlocal operators. I. Separable approximation of multi-variate functions. *Computing*, 76(3-4):177–202, 2006.

[14] W. Hackbusch and B. N. Khoromskij. Low-rank Kronecker-product approximation to multi-dimensional nonlocal operators. II. HKT representation of certain operators. *Computing*, 76(3-4):203–225, 2006.

[15] W. Hackbusch, B. N. Khoromskij, and E. E. Tyrtyshnikov. Hierarchical Kronecker tensor-product approximations. *J. Numer. Math.*, 13(2):119–156, 2005.

[16] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[17] Martin J. Mohlenkamp and Lucas Monzón. Trigonometric identities and sums of separable functions. *The Mathematical Intelligencer*, 27(2):65–69, 2005.

[18] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Academic Press, New York, 1970.

[19] Lawrence Perko. *Differential equations and dynamical systems*, volume 7 of *Texts in Applied Mathematics*. Springer-Verlag, New York, third edition, 2001.

[20] André Uschmajew. Well-posedness of convex maximization problems on stiefel manifolds and orthogonal tensor product approximations. *Numerische Mathematik*, 115:309–331, 2010. 10.1007/s00211-009-0276-9.

[21] André Uschmajew. Regularity of tensor product approximations to square integrable functions. *Constructive Approximation*, to appear.

[22] Harry Yserentant. On the regularity of the electronic Schrödinger equation in Hilbert spaces of mixed derivatives. *Numer. Math.*, 98(4):731–759, 2004.

[23] Harry Yserentant. Sparse grid spaces for the numerical solution of the electronic Schrödinger equation. *Numer. Math.*, 101:381–389, 2005.

[24] Harry Yserentant. The hyperbolic cross space approximation of electronic wavefunctions. *Numer. Math.*, 105:659–690, 2007.

[25] Harry Yserentant. *Regularity and approximability of electronic wave functions*, volume 2000 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2010.