JMLR: Workshop and Conference Proceedings vol 49:1–42, 2016

Basis Learning as an Algorithmic Primitive*

Mikhail Belkin MBELKIN@CSE.OHIO-STATE.EDU
Luis Rademacher LRADEMAC@CSE.OHIO-STATE.EDU
James Voss VOSSJ@CSE.OHIO-STATE.EDU
The Ohio State University

Abstract

A number of important problems in theoretical computer science and machine learning can be interpreted as recovering a certain basis. These include symmetric matrix eigendecomposition, certain tensor decompositions, Independent Component Analysis (ICA), spectral clustering and Gaussian mixture learning. Each of these problems reduces to an instance of our general model, which we call a "Basis Encoding Function" (BEF). We show that learning a basis within this model can then be provably and efficiently achieved using a first order iteration algorithm (gradient iteration). Our algorithm goes beyond tensor methods while generalizing a number of existing algorithms—e.g., the power method for symmetric matrices, the tensor power iteration for orthogonal decomposable tensors, and cumulant-based FastICA—all within a broader function-based dynamical systems framework. Our framework also unifies the unusual phenomenon observed in these domains that they can be solved using efficient non-convex optimization. Specifically, we describe a class of BEFs such that their local maxima on the unit sphere are in one-to-one correspondence with the basis elements. This description relies on a certain "hidden convexity" property of these functions.

We provide a complete theoretical analysis of the gradient iteration even when the BEF is perturbed. We show convergence and complexity bounds polynomial in dimension and other relevant parameters, such as perturbation size. Our perturbation results can be considered as a non-linear version of the classical Davis-Kahan theorem for perturbations of eigenvectors of symmetric matrices. In addition we show that our algorithm exhibits fast (superlinear) convergence and relate the speed of convergence to the properties of the BEF. Moreover, the gradient iteration algorithm can be easily and efficiently implemented in practice.

1. Introduction

A good algorithmic primitive is a procedure which is simple, allows for theoretical analysis, and ideally for efficient implementation. It should also be applicable to a range of interesting problems. An example of an extremely successful and widely used primitive, both in theory and practice, is the diagonalization/eigendecomposition of symmetric matrices.

In this paper, we propose a more general basis recovery mechanism as an algorithmic primitive. We provide the underlying algorithmic framework and its theoretical analysis. Our approach can be viewed as a non-linear/non-tensorial generalization of the classical matrix diagonalization results and perturbation analyses. We will show that a number of problems and techniques of recent theoretical and practical interest can be viewed within our setting.

* Extended abstract. Full version appears as (Belkin et al., 2016, v4)

Let $\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ be a full or partial, unknown orthonormal basis in $\mathbb{R}^d$. Choosing a set of one-dimensional *contrast functions*[1] $g_i : \mathbb{R} \to \mathbb{R}$, we define the Basis Encoding Function (BEF) $F : \mathbb{R}^d \to \mathbb{R}$ as

$$F(\mathbf{u}) := \sum_{i=1}^{m} g_i(\langle \mathbf{u}, \mathbf{e}_i \rangle) \, . \tag{1}$$

Our goal will be to recover the set $\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ (fully or partially) through access to $\nabla F(\mathbf{u})$ (the exact setting), or to provide a provable approximation to these vectors given an estimate of $\nabla F(\mathbf{u})$ (the noisy/perturbation setting). We will see that for several important problems, the relevant information about the problem can be encoded as a BEF (Section 2.1).

Taking a dynamical systems point of view, we propose a fixed point method for recovering the hidden basis. The basic algorithm consists simply of replacing the point with the normalized gradient at each step using the "gradient iteration" map $\mathbf{u} \mapsto \nabla F(\mathbf{u})/\|\nabla F(\mathbf{u})\|$. This gradient iteration map arises as a generalization of the eigenvector problem for symmetric matrices and tensors: When F is properly constructed from a matrix or tensor, the fixed points[2] of this map may be taken as a definition of the matrix/tensor eigenvectors (Qi, 2005; Lim, 2006). For symmetric matrices, the classical power method for eigenvector recovery may be written as gradient iteration. We extend this idea to our BEF setting. We show for a BEF under certain "hidden convexity" assumptions, the desired basis directions are the only stable fixed points of the gradient iteration, and moreover that the gradient iteration converges to one of the basis vectors given almost any starting point. Further, we link this gradient iteration algorithm to optimization of F over the unit sphere by demonstrating that the hidden basis directions (that is, the stable fixed point of the gradient iteration) are also a complete enumeration of the local maxima of $|F(\mathbf{u})|$.

The gradient iteration also generalizes several influential fixed point methods for performing hidden basis recovery in machine learning contexts including cumulant-based FastICA (Hyvärinen, 1999) and the tensor power method (De Lathauwer et al., 1995; Anandkumar et al., 2012b) for orthogonally decomposable symmetric tensors. Our main conceptual contribution is to demonstrate that the success of such power iterations need not be viewed as a consequence of a linear or multilinear algebraic structure, but instead relies on a coordinate-wise decomposition of the function F combined with a more fundamental "hidden convexity" structure. In particular, our most important assumption is that the contrast functions g_i satisfy that $x \mapsto |g_i(\sqrt{x})|$ be convex[3].

Under our assumptions, we demonstrate that the gradient iteration exhibits superlinear convergence as opposed to the linear convergence of the standard power iteration for matrices but in line with some known results for ICA and tensor power methods (Hyvärinen, 1999; Nguyen and Regev, 2009; Anandkumar et al., 2012b). We provide conditions on the contrast functions g_i to obtain specific higher orders of convergence.

It turns out that a similar analysis still holds when we only have access to an approximation of ∇F (the noisy setting). In order to give polynomial run-time bounds we analyze gradient iteration

1. We call the g_is contrast functions following the Independent Component Analysis (ICA) terminology. Note, however, that in the ICA setting our "contrast functions" correspond to different scalings of the ICA contrast function.
2. We are considering fixed points in projective space here, i.e. antipodal points on the sphere are identified as a single equivalence class when defining fixed points.
3. For technical reasons our analysis requires strict convexity while convexity is sufficient for the classical matrix power method. The analysis of matrix power iteration is a limit case of our setting with slightly different properties.

with occasional random jumps[4]. The resulting algorithm still provably recovers an approximation to a hidden basis element. By repeating the algorithm we can recover the full basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$. We provide an analysis of the resulting algorithm's accuracy and running time under a general perturbation model. Our bounds involve low degree polynomials in all relevant parameters—e.g., the ambient dimension, the number of basis elements to be recovered, and the perturbation size—and capture the superlinear convergence speeds of the gradient iteration. Our accuracy bounds can be considered as a non-linear version of the classical perturbation theorem of Davis and Kahan (1970) for eigenvectors of symmetric matrices. Interestingly, to obtain these bounds we only require approximate access to ∇F and do not need to assume anything about the perturbations of the second derivatives of F or even F itself. We note that our perturbation results allow for substantially more general perturbations than those used in the matrix and tensor settings, where the perturbation of a matrix/tensor is still a matrix/tensor. In many realistic settings the perturbed model does not have the same structure as the original. For example, in computer computations, $A\mathbf{x}$ is not actually a linear function of $\mathbf{x}$ due to finite precision of floating point arithmetic. Our perturbation model for ∇F still applies in these cases.

Below in Section 2.1 we will show how a number of problems can be viewed in terms of hidden basis recovery. Specifically, we briefly discuss how our primitive can be used to recover clusters in spectral clustering, independent components in Independent Component Analysis (ICA), parameters of Gaussian mixtures and certain tensor decompositions. Finally, in Section 7 we apply our framework to obtain the first provable ICA recovery algorithm for arbitrary model perturbations.

Organization of the paper. In Section 2 we introduce the problem of basis recovery and show connections to spectral clustering, ICA, matrix and tensor decompositions and Gaussian Mixture Learning. We describe our framework and sketch the main theoretical results of the paper. In Section 3 we analyze the structure of the extrema of basis encoding functions. In Section 4 we show that the fixed points of gradient iteration are in one-to-one correspondence with the BEF's maxima and analyze convergence of gradient iteration in the exact case. In Section 5 we give an interpretation of the gradient iteration algorithm as a form of adaptive gradient ascent. In Section 6 we describe a robust version of our algorithm and give a complete theoretical analysis for arbitrary perturbations. In Section 7 we apply our framework to obtain the first ICA algorithm which is provably robust for arbitrary model perturbations.

2. Problem description and the main results

We consider a function optimization framework for hidden basis recovery. More formally, let $\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ be a non-empty set of orthogonal unit vectors in $\mathbb{R}^d$. These unit vectors form the unseen basis. A function on a closed unit ball $F : \overline{B(0,1)} \to \mathbb{R}$ is defined from "contrast functions" $g_i : [-1,1] \to \mathbb{R}$ as:

$$F(\mathbf{u}) := \sum_{i=1}^{m} g_i(\langle \mathbf{u}, \mathbf{e}_i \rangle) \,. \tag{2}$$

We call F a *basis encoding function (BEF)* with the associated tuples $\{(g_i, \mathbf{e}_i) \mid i \in [m]\}$. The goal is to recover the hidden basis vectors $\mathbf{e}_i$ for $i \in [m]$ up to sign given evaluation access to F and its

4. In a related work, Ge et al. (2015) use the standard gradient descent with random jumps to escape from saddle points in the context of online tensor decompositions.

gradient. We will assume that $d \geq 2$ since otherwise the problem is trivial. We consider contrast functions $g_i \in \mathcal{C}^{(2)}([-1,1])$ which satisfy the following assumptions:

A1. g_i is either an even or odd function.

A2. Strict convexity of $|g_i(\sqrt{x})|$: Either $\frac{d^2}{dx^2} g_i(\sqrt{x}) > 0$ on $(0,1]$ or $-\frac{d^2}{dx^2} g_i(\sqrt{x}) > 0$ on $(0,1]$.

A3. The right derivative at the origin $\frac{d}{dx} g_i(\sqrt{x})|_{x=0^+} = 0$.

A4. $g_i(0) = 0$.

The assumption A2 is slightly stronger than stating that one of $\pm g_i(\sqrt{x})$ is strictly convex on $(0,1]$. From now on F and the term BEF will refer to a BEF with associated $\mathbf{e}_i$s and g_is satisfying Assumptions A1–A4 unless otherwise stated.

Remark: The Assumption A4 is non-essential. If each g_i satisfies A1–A3, then $x \mapsto [g_i(x) - g_i(0)]$ satisfies A1–A4 making $[F(\mathbf{u}) - F(\mathbf{0})] = \sum_{i=1}^m [g_i(\langle \mathbf{u}, \mathbf{e}_i \rangle) - g_i(0)]$ a BEF of the desired form.

We shall see that BEFs arise naturally in a number of problems, and also that given a BEF, the directions $\mathbf{e}_1, \ldots, \mathbf{e}_m$ can be efficiently recovered up to sign.

2.1. Motivations for and Example of BEF Recovery

Before discussing our main results on BEF recovery, we first motivate why the BEF recovery problem is of interest through a series of examples. We first show how BEF recovery and the gradient iteration relate to ideas from the eigenvector analysis of matrices and tensors. Then, we will discuss several settings where the problem of BEF recovery arises naturally in machine learning.

Connections to matrix eigenvector recovery. Our algorithm can be viewed as a generalization of the classical power iteration method for eigendecomposition of symmetric matrices. Let A be a symmetric matrix. Put $F(\mathbf{u}) = \mathbf{u}^T A \mathbf{u}$. From the spectral theorem for matrices, we have $F(\mathbf{u}) = \sum_i \lambda_i \langle \mathbf{u}, \mathbf{e}_i \rangle^2$ where each λ_i is an eigenvalue of A with corresponding eigenvector $\mathbf{e}_i$. We see that $F(\mathbf{u})$ is a BEF[5] with the contrast functions $g_i(x) := \lambda_i x^2$. It is easy to see that our gradient iteration is an equivalent update to the power method update $\mathbf{u} \mapsto A\mathbf{u}/\|A\mathbf{u}\|$. As such, the fixed points[6] of the gradient iteration are eigenvectors of the matrix A. We also note that it is not necessary to know each $g_i(x)$ to have access to the BEF $F(\mathbf{u})$ or its derivative $\nabla F(\mathbf{u})$.

In addition, we note that the gradient iteration for a BEF may be written to look very much like the power iteration for matrices. Let $F(\mathbf{u}) = \sum_{i=1}^m g_i(\langle \mathbf{u}, \mathbf{e}_i \rangle)$ denote a BEF. In order to better capture the convexity Assumption A2, we may define functions $h_i(t) := g_i(\mathrm{sign}(t)\sqrt{|t|})$. To compress notation, we use $\pm$ to denote the $\mathrm{sign}(\langle \mathbf{u}, \mathbf{e}_i \rangle)$. Then, $F(\mathbf{u}) = \sum_{i=1}^m h_i(\pm \langle \mathbf{u}, \mathbf{e}_i \rangle^2)$. Taking derivatives, we obtain that

$$\nabla F(\mathbf{u}) = 2 \sum_{i=1}^m \pm h_i'(\pm \langle \mathbf{u}, \mathbf{e}_i \rangle^2) \langle \mathbf{u}, \mathbf{e}_i \rangle \mathbf{e}_i \, .$$

Note that in the matrix example above, the power iteration can be expanded as

$$A\mathbf{u} = \sum_i \lambda_i \langle \mathbf{u}, \mathbf{e}_i \rangle \mathbf{e}_i \, .$$

5. Note that condition A2 is not satisfied as in this case $g_i(\sqrt{x})$ is convex but not strictly convex.

6. These fixed points are fixed possibly up to a sign flip. Alternatively stated, these are fixed points in projective space.

We see that the formula for $\nabla F(\mathbf{u})$ is the same as the power iteration for matrices with the (constant) eigenvalues λ_i being replaced by the functional term $\pm h_i'(\pm\langle \mathbf{u}, \mathbf{e}_i\rangle^2)$. By the Assumption A2, $|h_i(t)|$ is strictly convex, and in particular each $|h_i'(t)|$ is strictly increasing as a function of $|t|$. The gradient iteration for general BEFs may be thought of as a power iteration where matrix eigenvalues are being replaced by functions whose magnitude grows with the magnitude of their respective coordinate values $\langle \mathbf{u}, \mathbf{e}_i\rangle$. The change in these "eigenvalues" by location allows each of the basis directions $\mathbf{e}_1, \ldots, \mathbf{e}_m$ to become an attractor locally since there is no single fixed "top eigenvalue" as in the matrix setting.

Connections to the tensor eigenvector problem. While in general not a special case of the BEF framework, there are also connections between the gradient iteration algorithm and the definition of an eigenvector of a symmetric tensor (Qi, 2005; Lim, 2006). In particular, given a symmetric tensor $T \in \mathbb{R}^{d\times\cdots\times d}$ (with r copies of d), we may treat T as an operator on $\mathbb{R}^d$ using the operation $T\mathbf{u}^r := \sum_{i_1,\ldots,i_r} T_{i_1\ldots i_r} u_{i_1}\cdots u_{i_r}$. We note that this formula encapsulates the matrix quadratic form $\mathbf{u}^T A\mathbf{u} = A\mathbf{u}^2$ as a special case. We also denote by $T\mathbf{u}^{r-1}$ the vector such that $[T\mathbf{u}^{r-1}]_j = \sum_{i_2,\ldots,i_r} T_{ji_2\ldots i_r} u_{i_2}\cdots u_{i_r}$. If we define the function $f(\mathbf{u}) = T\mathbf{u}^r$, then the Z-eigenvectors of T are defined to be vectors $\mathbf{u}$ for which there exists $\lambda \in \mathbb{R}$ such that $\nabla f(\mathbf{u}) = r\lambda\mathbf{u}$. Expanding this formula, we get the slightly more familiar looking form that the Z-eigenvectors of T are the points such that $T\mathbf{u}^{r-1} = \lambda\mathbf{u}$, or alternatively the fixed points[7] of the iteration $\mathbf{u} \mapsto \frac{T\mathbf{u}^{r-1}}{\|T\mathbf{u}^{r-1}\|}$. Note that this iteration may alternatively be written as $\mathbf{u} \mapsto \frac{\nabla f(\mathbf{u})}{\|\nabla f(\mathbf{u})\|}$. Replacing the function $f(\mathbf{u}) = T\mathbf{u}^r$ with a BEF F, the fixed points[7] of the gradient iteration $\mathbf{u} \mapsto \frac{\nabla F(\mathbf{u})}{\|\nabla F(\mathbf{u})\|}$ are like eigenvectors for our function F in this dynamical systems sense.

Orthogonal tensor decompositions. In a recent work (Anandkumar et al., 2012b), it was shown that the tensor eigenvector recovery problem for tensors with orthogonal decompositions[8] can be applied to a variety of problems including ICA and previous works on learning mixtures of spherical Gaussians (Hsu and Kakade, 2013), latent Dirichlet allocation (Anandkumar et al., 2012a), and learning hidden Markov models (Anandkumar et al., 2012c).

Their framework involves using the moments of the various models to obtain a tensor of the form $T = \sum_{k=1}^m w_k \boldsymbol{\mu}_k^{\otimes r}$ where (1) each $w_k \in \mathbb{R}\setminus\{0\}$, (2) each $\boldsymbol{\mu}_k \in \mathbb{R}^d$ is a unit vector, and (3) $\boldsymbol{\mu}_k^{\otimes r}$ is the tensor power defined by $(\boldsymbol{\mu}_k^{\otimes r})_{i_1\ldots i_r} = (\boldsymbol{\mu}_k)_{i_1}\ldots(\boldsymbol{\mu}_k)_{i_r}$. The $\boldsymbol{\mu}_k$s may be assumed to have unit norm by rescaling the w_ks appropriately. In the special case where the $\boldsymbol{\mu}_k$s are orthogonal, then the direction of each $\boldsymbol{\mu}_k$ can be recovered using tensor power methods (Anandkumar et al., 2012b). It can be shown that $T\mathbf{u}^r = \sum_{k=1}^m w_k\langle \mathbf{u}, \boldsymbol{\mu}_k\rangle^r$. In particular, the function $F(\mathbf{u}) = T\mathbf{u}^r$ is a BEF with the contrasts $g_i(x) := w_i x^r$ and hidden basis elements $\mathbf{e}_k := \boldsymbol{\mu}_k$. Further, the fixed point iteration $\mathbf{u} \mapsto \frac{T\mathbf{u}^{r-1}}{\|T\mathbf{u}^{r-1}\|}$ proposed by Anandkumar et al. (2012b) for eigenvector recovery in this setting can be equivalently written as the gradient iteration update $\mathbf{u} \mapsto \frac{\nabla F(\mathbf{u})}{\|\nabla F(\mathbf{u})\|}$.

Spectral clustering. Spectral clustering is a class of methods for multiway cluster analysis. We describe now a prototypical version of the method that works in two phases (Bach and Jordan, 2006; Ng et al., 2002; Shi and Malik, 2000; Yu and Shi, 2003). The first phase, spectral embedding, constructs a similarity graph based on the features of the data and then embeds the data in $\mathbb{R}^d$ (where

7. These fixed points are fixed possibly up to a sign flip. Alternatively stated, these are fixed points in projective space.
8. Another related work (Anandkumar et al., 2015) investigates properties of the tensor power method in certain settings where the symmetric tensor is not orthogonal decomposable and has symmetric rank exceeding d.

d is the number of clusters) using the bottom d eigenvectors of the Laplacian matrix of the similarity graph. The second phase clusters the embedded data using a variation of the k-means algorithm. A key aspect in the justification of spectral clustering is the following observation: If the graph has d connected components, then a pair of data points is either mapped to the same vector if they are in the same connected component or mapped to orthogonal vectors if they are in different connected components (Weber et al., 2004). If the graph is close to this ideal case, which can be interpreted as a realistic graph with d clusters, then the embedding is close to that ideal embedding.

This suggests the following alternate approach (introduced in Belkin et al., 2014) to the second phase of spectral clustering by interpreting it as a hidden basis recovery problem: Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be the embedded points. Let $g : \mathbb{R} \to \mathbb{R}$ be a function satisfying Assumptions A1–A4. Let

$$F(\mathbf{u}) = \sum_{i=1}^{n} g(\langle \mathbf{u}, \mathbf{x}_i \rangle). \tag{3}$$

In the ideal case, there exists an orthonormal basis $\mathbf{Z}_1, \ldots, \mathbf{Z}_d$ of $\mathbb{R}^d$ and positive scalars $b_1, \ldots, b_d$ such that $\mathbf{x}_i = b_j \mathbf{Z}_j$ for every i in the j^{th} connected component of the graph. Thus, in the ideal case we can write

$$F(\mathbf{u}) = \sum_{j=1}^{d} a_j g(b_j \langle \mathbf{u}, \mathbf{Z}_j \rangle)$$

where a_j is the number of points from the j^{th} connected component. Thus, F is a BEF in the ideal case with contrasts $g_j(t) := a_j g(b_j t)$. In the general case, it is a perturbed BEF and the hidden basis can be approximately recovered using our robust algorithm (Section 6). Note that via (3), F and its derivatives can be evaluated at any $\mathbf{u}$ just with knowledge of the $\mathbf{x}_i$s, and without knowing the hidden basis.

We note that for this spectral clustering application, the choice of g is arbitrary so long as it satisfies our Assumptions A1–A4. In particular, this is an example where the generality of the gradient iteration beyond the tensorial setting provides greater flexibility.

Independent component analysis (ICA). In the ICA model, one observes samples of the random vector $\mathbf{X} = A\mathbf{S}$ where $A \in \mathbb{R}^{d \times d}$ is a mixing matrix and $\mathbf{S} = (S_1, \ldots, S_d)$ is a latent random vector such that the S_is are mutually independent and non-Gaussian. The goal is to recover the mixing matrix $A = [A_1 | \cdots | A_d]$, typically with the goal of using A^{-1} to invert the mixing process and recover the original signals. This recovery is possible up to natural indeterminacies, namely the ordering of the columns of A and the choice of the sign of each A_i (Comon, 1994). ICA has a vast literature (see the books Comon and Jutten, 2010; Hyvärinen et al., 2001, for a broad overview) with numerous applications including speech separation (Makino et al., 2007), denoising of EEG/MEG brain recordings (Vigário et al., 2000), and various vision tasks (Bartlett et al., 2002; Bell and Sejnowski, 1997) to name a few.

To demonstrate that ICA fits within our BEF framework, we rely on the properties of the cumulant statistics.[9] Let $\kappa_r(X)$ denote the r^{th} cumulant of a random variable X. The cumulant $\kappa_r(X)$ satisfies the following: (1) Homogeneity: $\kappa_r(\alpha X) = \alpha^r \kappa_r(X)$ for any $\alpha \in \mathbb{R}$ and (2) Additivity: if X and

9. An important class of ICA methods with guaranteed convergence to the columns of A are based on the optimization of $\kappa_4(\langle \mathbf{u}, \mathbf{X} \rangle)$ over the unit sphere (see e.g., Arora et al., 2012; Delfosse and Loubaton, 1995; Hyvärinen, 1999). Other contrast functions are also frequently used in the practical implementations of ICA (see e.g., Hyvärinen and Oja, 1998). However, these non-cumulant functions can have spurious maxima (Wei, 2015).

Y are independent, then $\kappa_r(X+Y) = \kappa_r(X) + \kappa_r(Y)$. Given an ICA model $\mathbf{X} = A\mathbf{S}$, these properties imply that for all $\mathbf{u} \in \mathbb{R}^d$, $\kappa_r(\langle \mathbf{u}, \mathbf{X}\rangle) = \kappa_r(\sum_{i=1}^d \langle \mathbf{u}, A_i\rangle S_i) = \sum_{i=1}^d \langle \mathbf{u}, A_i\rangle^r \kappa_r(S_i)$. A preprocessing step called whitening (i.e., linearly transforming the observed data to have identity covariance) makes the columns of A into orthogonal unit vectors. Under whitening, the columns of A form a hidden basis of the space. In particular, defining the contrast functions $g_i(x) := x^r \kappa_r(S_i)$ and the basis encoding elements $\mathbf{e}_i := A_i$, then the function $F(\mathbf{u}) := \kappa_r(\langle \mathbf{u}, \mathbf{X}\rangle) = \sum_{i=1}^d g_i(\langle \mathbf{u}, \mathbf{e}_i\rangle)$ is a BEF so long as each $\kappa_r(S_i) \neq 0$. Further, these directional cumulants and their derivatives have natural sample estimates (see e.g., Kenney and Keeping, 1962; Voss et al., 2013, for the third and fourth order estimates), and as such this choice of F will be admissible to our algorithmic framework for basis recovery.

Interestingly, it has been noted in several places (Hyvärinen, 1999; Nguyen and Regev, 2009; Zarzoso and Comon, 2010) that cubic convergence rates can be achieved using optimization techniques for recovering the directions A_i, particularly when performing ICA using the fourth cumulant or the closely related fourth moment. One explanation as to why this is possible arises from the dual interpretation (discussed in section 5) of the gradient iteration algorithm as both an optimization technique and as a power method. In the ICA setting, the gradient iteration algorithm for cumulants was introduced by Voss et al. (2013). This paper provides a significant generalization of those ideas as well as a theoretical analysis.

Parameter estimation in a spherical Gaussian Mixture Model. A Gaussian Mixture Model (GMM) is a parametric family of probability distributions. A spherical GMM is a distribution whose density can be written in the form $f(\mathbf{x}) = \sum_{i=1}^k w_i f_i(\mathbf{x})$, where $w_i \geq 0$, $\sum_i w_i = 1$ and f_i is a d-dimensional Normal density with mean $\boldsymbol{\mu}_i$ and covariance matrix $\sigma_i^2 \mathcal{I}$, for $\sigma_i > 0$. The parameter estimation problem is to estimate $w_i, \boldsymbol{\mu}_i, \sigma_i$ given i.i.d. samples of random vector $\mathbf{x}$ with density f. For clarity of exposition, we only discuss the case $k = d$ and $\sigma_i = \sigma$ for some fixed, unknown σ. Our argument is a variation of the moment method of Hsu and Kakade (2013). As in their work, similar ideas should work for the case $k < d$ and non-identical σ_is.

We explain how to recover the different parameters from observable moments. Firstly, σ^2 is the smallest eigenvalue of the covariance matrix of $\mathbf{x}$. This recovers σ. Let $\mathbf{v}$ be any unit norm eigenvector corresponding to the eigenvalue σ^2. Define $M_2 := \mathbb{E}(\mathbf{x}\mathbf{x}^T) - \sigma^2 \mathcal{I} \in \mathbb{R}^{d \times d}$. Then we have $M_2 = \sum_{i=1}^d w_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T$. Denote $D = \mathrm{diag}(w_1, \ldots, w_d)$, $A = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_d) \in \mathbb{R}^{d \times d}$. With this notation we have $M_2 = ADA^T$. Let $M = M_2^{1/2}$ (symmetric). This implies $M = AD^{1/2}R$, where R is some orthogonal matrix.

We have $\mathbb{E}(\langle \mathbf{x}, \mathbf{u}\rangle^3) = \sum_{i=1}^d w_i \langle \boldsymbol{\mu}_i, \mathbf{u}\rangle^3 + 3\sigma^2 \|\mathbf{u}\|^2 \mathbb{E}(\langle \mathbf{x}, \mathbf{u}\rangle)$. Then,

$$F(\mathbf{u}) := \mathbb{E}(\langle \mathbf{x}, M^{-1}\mathbf{u}\rangle^3) - 3\sigma^2 \|M^{-1}\mathbf{u}\|^2 \mathbb{E}(\langle \mathbf{x}, M^{-1}\mathbf{u}\rangle) = \sum_{i=1}^d w_i \langle \boldsymbol{\mu}_i, M^{-1}\mathbf{u}\rangle^3$$

$$= \sum_{i=1}^d w_i (\mathbf{u}^T R^T D^{-1/2} \mathbf{e}_i)^3 = \sum_{i=1}^d w_i^{-1/2} \langle \mathbf{u}, \mathbf{R}_{i\cdot}\rangle^3$$

is a BEF encoding the rows of R, with basis vectors $\mathbf{z}_i = \mathbf{R}_{i\cdot}$ and contrasts $g_i(t) = w_i^{-1/2} t^3$. The recovery of the rows of R allows the recovery of the directions of the columns of A, that is, the directions of $\boldsymbol{\mu}_i$s. The actual $\boldsymbol{\mu}_i$s then can be recovered from the identity $\langle \boldsymbol{\mu}_i, \mathbf{v}\rangle = \langle \mathbb{E}(\mathbf{x}), \mathbf{v}\rangle$. Finally, denoting $\mathbf{w} = (w_1, \ldots, w_d)$ we have $\mathbb{E}(\mathbf{x}) = A\mathbf{w}$ and we recover $\mathbf{w} = A^{-1}\mathbb{E}(\mathbf{x})$.

2.2. Summary of the main results

In what follows it will be convenient to append arbitrary orthonormal directions $\mathbf{e}_{m+1}, \ldots, \mathbf{e}_d$ to our hidden "basis" to obtain a full basis. For the remainder of this paper, we simplify our notation by indexing vectors in $\mathbb{R}^d$ with respect to this hidden basis $\mathbf{e}_1, \ldots, \mathbf{e}_d$. That allows us to introduce the notation $u_i := \langle \mathbf{u}, \mathbf{e}_i \rangle$ for $\mathbf{u} \in \mathbb{R}^d$. Thus, $F(\mathbf{u}) = \sum_{i=1}^m g_i(u_i)$.

We now state the first result indicating that a BEF encodes the basis $\mathbf{e}_1, \ldots, \mathbf{e}_m$. We use $S^{d-1} := \{\mathbf{u} \mid \|\mathbf{u}\| = 1\}$ to denote the unit sphere in $\mathbb{R}^d$.

Theorem 2.1 *The set $\{\pm \mathbf{e}_i \mid i \in [m]\}$ is a complete enumeration of the local maxima of $|F|$ with respect to the domain S^{d-1}.*

Theorem 2.1 implies that a form of gradient ascent can be used to recover maxima of $|F|$ and hence the hidden basis[10]. However, the performance of gradient ascent is dependent on the choice of a learning rate parameter. We propose a simple and practical parameter-free fixed point method, *gradient iteration*, for finding the hidden basis elements $\mathbf{e}_i$ in this setting.

The proposed method is based on the *gradient iteration function* $G : S^{d-1} \to S^{d-1}$ defined by

$$G(\mathbf{u}) := \frac{\nabla F(\mathbf{u})}{\|\nabla F(\mathbf{u})\|}$$

with the convention that $G(\mathbf{u}) = \mathbf{u}$ if $\nabla F(\mathbf{u}) = \mathbf{0}$. We use the map G as a fixed point iteration for recovering the hidden basis elements[11].

However, there is a difficulty: at any given step, the derivative $\partial_i F(\mathbf{u})$ can be of a different sign than u_i causing $\operatorname{sign}(u_i) \neq \operatorname{sign}(G_i(\mathbf{u}))$. Note that we do not know which coordinates flip their signs as the coordinates are hidden. As it turns out, this does not affect the algorithm, but the analysis is more transparent in a space of equivalence classes[12]. We divide S^{d-1} into equivalence classes using the equivalence relation $\mathbf{v} \sim \mathbf{u}$ if $|v_i| = |u_i|$ for each $i \in [d]$. Given $\mathbf{v} \in S^{d-1}$, we denote by $[\mathbf{v}]$ its corresponding equivalence class. The resulting quotient space $S^{d-1}/\!\sim$ may be identified with the positive orthant of the sphere $Q_+^{d-1} := \{\mathbf{u} \in S^{d-1} \mid u_i \geq 0 \text{ for all } i \in [d]\}$. There is a bijection $\phi : S^{d-1}/\!\sim \to Q_+^{d-1}$ given by $\phi([\mathbf{u}]) = \sum_{i=1}^d |u_i| \mathbf{e}_i$. We treat $S^{d-1}/\!\sim$ as a metric space with the metric $\mu([\mathbf{u}], [\mathbf{v}]) = \|\phi([\mathbf{v}]) - \phi([\mathbf{u}])\|$. Under Assumption A1, if $\mathbf{u} \sim \mathbf{v}$ then $G(\mathbf{u}) \sim G(\mathbf{v})$. As such, sequences are consistently defined modulo this equivalence class, and we consider the fixed points of $G/\!\sim$.

We will use the following terminology. A class $[\mathbf{v}]$ is a *fixed point* of $G/\!\sim$ if $G(\mathbf{v}) \sim \mathbf{v}$. We will consider sequences of the form $\{\mathbf{u}(n)\}_{n=0}^{\infty}$ defined recursively by $\mathbf{u}(n) = G(\mathbf{u}(n-1))$. In addition, by abuse of notation, we will sometimes refer to a vector $\mathbf{v} \in S^{d-1}$ as a fixed point of $G/\!\sim$.

We demonstrate that the attractors of $G/\!\sim$ are precisely the hidden basis elements, and that all other fixed points of $G/\!\sim$ are non-attractive (unstable hyperbolic). Further, convergence to a hidden basis element is guaranteed given almost any starting point $\mathbf{u}(0) \in S^{d-1}$.

10. We note that assumption A1 is stronger than what is actually required in Theorem 2.1. In particular, we could replace Assumption A1 with the assumption that $x \mapsto g_i(-\sqrt{|x|})$ is either strictly convex or strictly concave on $[-1, 0]$ for each $i \in [m]$.
11. A special case of this iteration was introduced in the context of ICA (Voss et al., 2013).
12. Alternative approaches to fixing the sign issue include analyzing the fixed points of the double iteration $\mathbf{u} \to G(G(\mathbf{u}))$ or working in projective space.

Theorem 2.2 (Gradient iteration stability) *The hidden basis elements $\{[\mathbf{e}_i] \mid i \in [m]\}$ are attractors of the dynamical system $G/\sim$. Further, there is a full measure set $\mathcal{X} \subset S^{d-1}$ such that for all $\mathbf{u}(0) \in \mathcal{X}$, $[\mathbf{u}(n)] \to [\mathbf{e}_i]$ for some $\mathbf{e}_i$ as $n \to \infty$.*

One implication of Theorem 2.2 is that given a $\mathbf{u}(0) \in S^{d-1}$ drawn uniformly at random, then with probability 1, $\mathbf{u}(n)$ converges (up to $\sim$) to one of the hidden basis elements.

Moreover, the rate of convergence to the hidden basis elements is fast (superlinear).

Theorem 2.3 (Gradient iteration convergence rate) *If $[\mathbf{u}(n)] \to [\mathbf{e}_i]$ as $n \to \infty$, then the convergence is superlinear. Specifically, if $x \mapsto g_i(x^{1/r})$ is convex on $[0, 1]$ for some $r > 2$, then the rate of convergence is at least of order $r - 1$.*

The above Theorems suggest the following practical algorithm for recovering the hidden basis elements: First choose a vector $\mathbf{u} \in \mathcal{S}^{d-1}$ and perform the iteration $\mathbf{u} \leftarrow G(\mathbf{u})$ until convergence is achieved to recover a single hidden basis direction. In practice, one may threshold $\min(\|G(\mathbf{u}) - \mathbf{u}\|,\ \|-G(\mathbf{u}) - \mathbf{u}\|)$ to determine if convergence is achieved. Then, to recover an additional hidden basis direction, one may repeat the procedure with a new starting vector $\mathbf{u}$ in the orthogonal complement to previously found hidden basis elements. We refer to this process as the gradient iteration algorithm.

From a practical standpoint, the fast and guaranteed convergence properties of the gradient iteration make it an attractive algorithm for hidden basis recovery. We also demonstrate that the gradient iteration is robust to a perturbation. Specifically, we modify the gradient iteration algorithm by occasionally performing a small random jump of size σ on the sphere. We call this algorithm ROBUSTGI-RECOVERY and show that it approximately recovers all hidden basis elements. More precisely, we consider the following notion of a perturbation of ∇F: If for every $\mathbf{u} \in \overline{B(0,1)}$, $\|\nabla F(\mathbf{u}) - \widehat{\nabla F}(\mathbf{u})\| \leq \epsilon$, then we say that $\widehat{\nabla F}$ is an ϵ-approximation of F. Further, if F satisfies a strong version of assumption A2, namely that there exists positive constants $\alpha \geq \beta$ and $\gamma \leq \delta$ such that for each $i \in [m]$, $\beta x^{\delta-1} \leq |\frac{d^2}{dx^2} g_i(\sqrt{x})| \leq \alpha x^{\gamma-1}$ for all $x \in (0, 1]$, then our perturbation result can be summarized as follows.

Theorem 2.4 (simplified) *Treating γ and δ as constants, if $\sigma \leq \operatorname{poly}^{-1}(\frac{\alpha}{\beta}, d, m)$ and if $\epsilon \leq \sigma\beta \operatorname{poly}^{-1}(\frac{\alpha}{\beta}, m, d)$, then with probability $1 - p$,* ROBUSTGI-RECOVERY *takes*

$$\operatorname{poly}(\frac{1}{\sigma}, \frac{\alpha}{\beta}, m, d) \log(\frac{1}{p}) + \operatorname{poly}(d, m) \log_{1+2\gamma}(\log(\frac{\beta}{\epsilon}))$$

time to recover $O(\epsilon/\beta)$ approximations of each $\mathbf{e}_i$ up to a sign. Specifically, ROBUSTGI-RECOVERY *returns vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m$ such that there exists a permutation π of $[m]$ such that $\|\pm\boldsymbol{\mu}_i - \mathbf{e}_{\pi(i)}\| \leq O(\epsilon/\beta)$ for all $i \in [m]$.*

Several observation are now in order:

1. We note that we only need a zero-order error bound for $\nabla F(\mathbf{u})$ for the perturbation analysis and do not need to assume anything about the perturbations of the second derivatives of F or even F itself. This perhaps surprising fact is due to the convexity conditions.
2. Our perturbation results allow for substantially more general perturbations than those used in the matrix and tensor settings, where the perturbation of a tensor is still a tensor. In our setting

the perturbation of a BEF corresponding to a tensor does not have to be tensorial in structure. This situation is very common whenever an observation of an object is not exact. For example, $A\mathbf{x}$ is not a linear function of $\mathbf{x}$ on a finite precision machine. The same phenomenon occurs in the tensor case.

3. $\log_{1+2\gamma}(\log(\frac{\beta}{\epsilon}))$ above corresponds to the superlinear convergence from Theorem 2.3 in the unperturbed setting.

The full algorithm and analysis for ROBUSTGI-RECOVERY, complete with more precise bounds, can be found in section 6.

Finally, in section 7, we show how to apply ROBUSTGI-RECOVERY to cumulant-based ICA under an arbitrary perturbation from the ICA model. In this setting, ROBUSTGI-RECOVERY provides an algorithm for robustly recovering the approximate ICA model.

3. Extrema structure of Basis Encoding Functions

In this section, we investigate the maximum structure of $|F|$ on the unit sphere and prove Theorem 2.1.

The optima structure of F relies on the hidden convexity implied by Assumption A2. To capture this structure, we define $h_i : [-1,1] \to \mathbb{R}$ as $h_i(x) := g_i(\text{sign}(x)\sqrt{|x|})$ for $i \in [m]$ and $h_i := 0$ for $i \in [d] \setminus [m]$. Thus,

$$F(\mathbf{u}) = \sum_{i=1}^{m} h_i(\text{sign}(u_i)u_i^2) \; . \tag{4}$$

These h_i functions capture the convexity from Assumption A2. Indeed, the functions h_i have the following properties:

Lemma 3.1 *The following hold for all $i \in [m]$:*

1. *The magnitude function $|h_i(t)|$ is strictly convex.*
2. *$h_i'(0) = 0$.*
3. *h_i is continuously differentiable.*
4. *The derivative's magnitude function $|h_i'(t)|$ is strictly increasing as a function of $|t|$. In particular, $|h_i'(t)| > 0$ for all $t \neq 0$.*
5. *Fix I to be one of the intervals $(0,1]$ or $[-1,0)$. If h_i is strictly convex on I, then $\text{sign}(t)h_i'(t) > 0$ for all $t \in I$, and otherwise $\text{sign}(t)h_i'(t) < 0$ for all $t \in I$.*

Proof We first show parts 2 and 3. We compute the derivative of h_i to see

$$h_i'(x) = \begin{cases} \frac{1}{2}g_i'(\text{sign}(x)\sqrt{|x|})/\sqrt{|x|} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

where the derivative at the origin is due to Assumptions A1 and A3. Since the derivative $h_i'(t)$ exists for all t, and since one of $\pm h_i$ is convex on either of the intervals $[0,1]$ and $[-1,0]$, it follows that h_i' is continuous (see Hiriart-Urruty and Lemaréchal, 1996, Corollary 4.2.3).

To see part 4, we note that $h_i'(0) = 0$ and apply Assumption A2 to see that h_i' is strictly monotonic on $[0,1]$. As such, $|h_i'(t)|$ is strictly increasing on $[0,1]$. The symmetries of Assumption A1 imply that $|h_i'(t)|$ is strictly increasing more generally as a function of $|t|$.

To see part 5, we note that $h_i'(t) = h_i'(0) + \int_0^t h_i''(x)\,dx = \int_0^t h_i''(x)\,dx$. Then, we use Assumption A2 to obtain the stated correspondence between $\text{sign}(h_i''(x))$ (which is +1 on I if h_i is convex and -1 otherwise) and $\text{sign}(h_i'(t))$.

To see that $|h_i|$ is strictly convex, it suffices to use that $|h_i|$ is continuously differentiable and to show that $\frac{d}{dt}|h_i(t)|$ is strictly increasing. Note that $\frac{d}{dt}|h_i(t)| = \operatorname{sign}(h_i(t))h_i'(t)$, and also that $\operatorname{sign}(h_i(t)) = \operatorname{sign}(\int_0^t h_i'(t)) = \operatorname{sign}(th_i'(t))$ by part 5. It follows that $\operatorname{sign}(\frac{d}{dt}|h_i(t)|) = \operatorname{sign}(t)$. Taking this sign into account, part 4 implies part 1. ■

In order to avoid dealing with unnecessary sign values, we restrict ourselves to analyzing the optima structure of $|F|$ over the domain Q_+^{d-1} (the all positive orthant of the sphere). Due to the symmetries of the of the problem (Assumption A1), it is actually sufficient to analyze the maxima structure of $|F|$ on Q_+^{d-1} in order to fully characterize the maxima of $|F|$ on the entire sphere S^{d-1}.

To characterize the extrema structure of the restriction of $|F|$ to Q_+^{d-1}, we will use its derivative structure expanded in terms of the h_i functions. It will be useful to establish some relationships between the g_i and h_i functions. We denote by $\mathbb{1}_{[\bullet]}$ the indicator function, and we use the convention that any summand containing a $\mathbb{1}_{[\text{FALSE}]}$ coefficient is 0 even if the term is indeterminant (e.g., $\mathbb{1}_{[\text{FALSE}]}/0 = 0$ and $\infty \cdot \mathbb{1}_{[\text{FALSE}]} = 0$).

Lemma 3.2 *The following hold for each $i \in [m]$:*

1. *For $x \in [0,1]$, $g_i'(x) = 2h_i'(x^2)x$ and $h_i'(x^2) = \frac{g_i'(x)}{2x}\mathbb{1}_{[x\neq 0]}$.*
2. *For $x \in [0,1]$, $g_i''(x) = \mathbb{1}_{[x\neq 0]}[4h_i''(x^2)x^2 + 2h_i'(x^2)]$*
3. *Fox $x \in (0,1]$, $h_i''(x^2) = \frac{1}{4}[g_i''(x)/x^2 - g_i'(x)/x^3]$.*

Proof By construction, $h_i(x^2) = g_i(x)$. Taking derivatives, we obtain $2h_i'(x^2)x = g_i'(x)$. Since $h_i'(0) = 0$ by the Assumptions A3, $h_i'(x^2) = \frac{g_i'(x)}{2x}\mathbb{1}_{[x\neq 0]}$.

Taking a second derivative away from $x = 0$, we see that $g_i''(x) = 4h_i''(x^2)x^2 + 2h_i'(x^2)$. At $x = 0$,

$$\begin{aligned} g_i''(0) &= \lim_{c\to 0} \frac{g_i'(c) - g_i'(0)}{c} = 2\lim_{c\to 0^+} \frac{1}{2}\frac{g_i'(\sqrt{c})}{\sqrt{c}} \\ &= 2\lim_{c\to 0^+} \Big(\frac{d}{dx} g_i(\sqrt{x})\Big)\Big|_{x=c} = 2\Big(\frac{d}{dx} g_i(\sqrt{x})\Big)\Big|_{x=0} = 0\,. \end{aligned}$$

In the above, the second equality uses that $g_i'(0) = 0$, a fact which is implied by Assumption A3 (in particular, $|g_i'(0)| \leq |\lim_{h\to 0^+} \frac{g(\sqrt{h})-g(0)}{\sqrt{h}}| \leq |\lim_{h\to 0^+} \frac{g(\sqrt{h})-g(0)}{h}| = 0$ since $h \leq \sqrt{h}$ in a neighborhood of the origin). The fourth equality uses that $\frac{d}{dx} g_i(\sqrt{x})$ is continuous due to the convexity of $g_i(\sqrt{x})$ (see Hiriart-Urruty and Lemaréchal, 1996, Corollary 4.2.3). The final equality uses Assumption A3.

As $g_i''(0) = 0$, we obtain the formula on $[-1,1]$ of

$$g_i''(x) = \mathbb{1}_{[x\neq 0]}[4h_i''(x^2)x^2 + 2h_i'(x^2)]$$

as desired.

When $x \neq 0$, we may rearrange terms to obtain:

$$h_i''(x^2) = \frac{g_i''(x) - 2h_i'(x^2)}{4x^2} = \frac{g_i''(x)}{4x^2} - \frac{g_i'(x)}{4x^3}\,.$$

■

As $F(\mathbf{u}) = \sum_{i=1}^m g_i(u_i)$ has first and second order derivatives of $\nabla F(\mathbf{u}) = \sum_{i=1}^m g_i'(u_i)\mathbf{e}_i$ and $\mathcal{H}F(\mathbf{u}) = \sum_{i=1}^m g_i''(u_i)\mathbf{e}_i\mathbf{e}_i^T$, we obtain the following derivative formulas for $F(\mathbf{u})$ in terms of the h_i functions for any $\mathbf{u} \in Q_+^{d-1}$:

$$\nabla F(\mathbf{u}) = 2\sum_{i=1}^m h_i'(u_i^2)u_i\mathbf{e}_i \qquad \mathcal{H}F(\mathbf{u}) = \sum_{i=1}^m \mathbb{1}_{[u_i\neq 0]}[4h_i''(u_i^2) + 2h_i'(u_i^2)]\mathbf{e}_i\mathbf{e}_i^T \tag{5}$$

The first derivative necessary condition for $\mathbf{u} \in S^{d-1}$ to be an extrema of F over Q_+^{d-1} can be obtained using the Lagrangian function $\mathcal{L} : \overline{B(0,1)} \times \mathbb{R}$ defined as $\mathcal{L}(\mathbf{u}, \lambda) := F(\mathbf{u}) - \lambda[\|\mathbf{u}\|^2 - 1]$. In particular, a point $\mathbf{u} \in Q_+^{d-1}$ is a critical point of F with respect to Q_+^{d-1} (that is, it satisfies the first order necessary conditions to be a local maximum of F with respect to Q_+^{d-1}) if and only if there exists $\lambda \in \mathbb{R}$ such that $(\mathbf{u}, \lambda)$ is a critical point of $\mathcal{L}$. The following result then enumerates the critical points of F with respect to the Q_+^{d-1}.

Lemma 3.3 *Let $\mathbf{u} \in Q_+^{d-1}$ and $\lambda \in \mathbb{R}$. The pair $(\mathbf{u}, \lambda)$ is a critical point of $\mathcal{L}$ if and only if $\lambda\mathbb{1}_{[u_i\neq 0]} = h_i'(u_i^2)$ for all $i \in [d]$.*

Proof We set the derivative

$$\frac{\partial}{\partial u_i}\mathcal{L}(\mathbf{u}, \lambda) = \partial_i F(\mathbf{u}) - 2\lambda u_i = 2h_i'(u_i^2)u_i - 2\lambda u_i \tag{6}$$

equal to 0 to obtain $h_i'(u_i^2)u_i = \lambda u_i$. If $u_i = 0$, then $h_i'(u_i^2) = h_i'(0) = 0$ by Assumption A3. Otherwise, $h_i'(u_i^2) = \lambda$. ■

While there are exponentially many (with respect to m) critical points of F as a function on the sphere, it turns out that only the hidden basis directions correspond to maxima of F on the sphere. The proof of the following statements uses the convexity structure from Lemma 3.1.

Proposition 3.4 *If $j \in [m]$, then $\mathbf{e}_j$ is a strict local maximum of $|F|$ with respect to Q_+^{d-1}.*

Proof We will prove the case where h_j is strictly convex on $[0, 1]$ and note that the case h_j is strictly concave is exactly the same when replacing F with $-F$.

We first note that $F(\mathbf{e}_j) = h_j(1) > 0$ since h_j' is strictly increasing (see Lemma 3.1). In particular, using continuity of each g_i, it follows that $F(\mathbf{u}) > 0$ on a neighborhood of $\mathbf{e}_j$, and it suffices to demonstrate that F takes on a maximum with respect to S^{d-1} at $\mathbf{e}_j$. Letting $D_{\mathbf{u}}$ denote the derivative operator with respect to the variable $\mathbf{u}$ and continuing from equation (6), we obtain

$$D_{\mathbf{u}}^2\mathcal{L}(\mathbf{u}, \lambda) = \mathcal{H}F(\mathbf{u}) - 2\lambda D_{\mathbf{u}}\mathbf{u} = \sum_{i=1}^m \mathbb{1}_{[u_i\neq 0]}[4h_i''(u_i^2)u_i^2 + 2h_i'(u_i^2)]\mathbf{e}_i\mathbf{e}_i^T - 2\lambda I\,. \tag{7}$$

We now use the Lagrangian criteria for constrained extrema (see e.g., (Luenberger and Ye, 2008, chapter 11) for a discussion of the first order necessary and second order sufficient conditions for constrained extrema) to show that $\mathbf{e}_j$ is a maximum of $F|_{Q_+^{d-1}}$. From Lemma 3.3, we see that $(\mathbf{e}_j, h_j'(1))$ is a critical point of $\mathcal{L}$. Further, for any non-zero $\mathbf{v}$ such that $\mathbf{v} \perp \mathbf{e}_j$, we obtain

$\mathbf{v}^T(D^2_{\mathbf{u}}\mathcal{L})(\mathbf{e}_j, h_j'(1))\mathbf{v} = -2h_j'(1)\|\mathbf{v}\|^2$. As $h_j'(1) > 0$, it follows that $\mathbf{v}^T(D^2_{\mathbf{u}}\mathcal{L})(\mathbf{e}_j, h_j'(1))\mathbf{v} < 0$. Thus, $\mathbf{e}_j$ is a local maximum of F. ■

Proposition 3.5 *If $\mathbf{v} \in Q_+^{d-1}$ is not contained in the set $\{\mathbf{e}_i \mid i \in [m]\}$, then $\mathbf{v}$ is not a local maximum of $|F|$ with respect to Q_+^{d-1}.*

Proof We first consider the case in which $v_i = 0$ for all but at most one $i \in [m]$. We will call this $i \in [m]$ for which $v_i \neq 0$ as j if it exists and otherwise let $j \in [m]$ be arbitrary. Fix any $\mathbf{w} \in Q_+^{d-1}$ such that $w_j > v_j$ and $w_i = 0$ for $i \in [m] \setminus \{j\}$. Such a choice is possible since $\mathbf{v} \neq \mathbf{e}_j$ implies $v_j < 1$. Then, $|F(\mathbf{v})| = |h_j(v_j^2)|$ and $|F(\mathbf{w})| = |h_j(w_j^2)|$. Since $|h_j(t)|$ is a strictly increasing function on $[0, 1]$ from $|h_j(0)| = 0$ (see Lemma 3.1), it follows that $|F(\mathbf{w})| > |F(\mathbf{v})|$. Since $\mathbf{w}$ can be constructed in any open neighborhood of $\mathbf{v}$, $\mathbf{v}$ is not a local maximum of $|F|$ on Q_+^{d-1}.

Now, we consider the case where $\mathbf{v}$ is an extremum (either a maximum or a minimum) of $|F|$ with respect to Q_+^{d-1} such that there exists $j, k \in [m]$ distinct such that $v_j > 0$ and $v_k > 0$. We will demonstrate that this implies that $\mathbf{v}$ is a minimum of $|F|$.

We use the notation for a vector $\mathbf{u}$, $\mathbf{u}^{\langle k \rangle} := \sum_i u_i^k \mathbf{e}_i$ is the coordinate-wise power. Fix $\eta > 0$ sufficiently small that for all $\delta \in (-\eta, \eta)$ we have that $\mathbf{w}(\delta) := (\mathbf{v}^{\langle 2 \rangle} + \delta\mathbf{e}_j - \delta\mathbf{e}_k)^{\langle 1/2 \rangle} \in Q_+^{d-1}$. We now consider the difference $F(\mathbf{w}(\delta)) - F(\mathbf{v})$ for a non-zero choice of $\delta \in (-\eta, \eta)$:

$$\begin{aligned} F(\mathbf{w}(\delta)) - F(\mathbf{v}) &= h_j(w_j(\delta)^2) - h_j(v_j^2) + h_k(w_k(\delta)^2) - h_k(v_k^2) \\ &= h_j'(x_j(\delta)^2)[w_j(\delta)^2 - v_j^2] + h_k'(x_k(\delta)^2)[w_k(\delta)^2 - v_k^2] \\ &= \delta[h_j'(x_j(\delta)^2) - h_k'(x_k(\delta)^2)] \,, \end{aligned}$$

where $x_i(\delta) \in (v_j, w_j(\delta))$ and $x_i(\delta) \in (w_k(\delta), v_k)$ under the mean value theorem.

As $\mathbf{v}$ must be an extremum of F in order to be an extremum of $|F|$, there exists λ such that the pair $(\mathbf{v}, \lambda)$ is a critical point of $\mathcal{L}$. Let $\mathcal{S} = \{i \mid v_i \neq 0\}$. Lemma 3.3 implies that $\lambda = h_i'(v_i^2)$ for all $i \in \mathcal{S}$. In particular, $\operatorname{sign}(h_i'(v_i^2))$ is the same for each $i \in \mathcal{S}$, and we will call this sign value s. Under equation (4), we have $F(\mathbf{v}) = \sum_{i \in \mathcal{S}} h_i(v_i^2)$. By Lemma 3.1, sh_i is strictly increasing from $sh_i(0) = 0$ on $[0, 1]$ for each $i \in \mathcal{S}$. As such, $F(\mathbf{v})$ is separated from 0 and $\operatorname{sign}(F(\mathbf{v})) = s$. Further,

$$s[F(\mathbf{w}(\delta)) - F(\mathbf{v})] = s\delta[h_j'(x_j(\delta)^2) - h_k'(x_k(\delta)^2)] < s\delta[\lambda - \lambda] = 0$$

holds by noting that each sh_i' is strictly increasing on $[0, 1]$ (by Lemma 3.1). Thus, $\mathbf{v}$ is a minimum of $|F|$. ■

Theorem 2.1 follows by combining Propositions 3.4 and 3.5 and using the symmetries of F from Assumption A1.

4. Stability and convergence of gradient iteration

In this section we will sketch the analysis for the stability and convergence of gradient iteration (Theorems 2.2, 2.3). It turns out that a special form of basis encoding function is sufficient for our analysis.

Definition 4.1 *A BEF $F(\mathbf{u}) = \sum_{i=1}^m g_i(u_i)$ is called a* positive basis encoding function (PBEF) *if $x \mapsto g_i(\text{sign}(x)\sqrt{|x|})$ is strictly convex for each $i \in [m]$.*

A PBEF has several nice properties not shared by all BEFs. Its name is justified by the fact that for a PBEF F and for all $\mathbf{u} \in S^{d-1}$, $F(\mathbf{u}) \geq 0$. Further, when we expand $F(\mathbf{u}) = \sum_{i=1}^m h_i(\text{sign}(u_i)u_i^2) = \sum_{i=1}^m h_i(u_i^2)$ under equation (4), we see that each h_i is strictly convex over its entire domain. Finally, given a BEF F, we construct a PBEF $\bar{F}(\mathbf{u}) := \sum_{i=1}^m \bar{g}_i(u_i)$ where $\bar{g}_i(x) = |g_i(x)|$. We call $\bar{F}$ the *PBEF associated with* F.

We first establish that for PBEFs, the gradient iteration G is a true fixed point method on $\mathcal{S}^{d-1}$ without the need to consider equivalence classes (as in Section 2.2). Let ϕ and μ be defined as in Section 2.2. We identify each orthant of S^{d-1} by a sign vector $\mathbf{v}$ where each $v_i \in \{+1, -1\}$ by defining $Q_{\mathbf{v}}^{d-1} := \{\mathbf{u} \in S^{d-1} \mid v_i u_i \geq 0 \text{ for each } i \in [d]\}$ as the orthant of S^{d-1} containing $\mathbf{v}$.

Lemma 4.2 *Let $\mathbf{v} \in \mathbb{R}^d$ be a sign vector (that is, $v_i \in \{\pm 1\}$ for each $i \in [d]$). If $\mathbf{u}, \mathbf{w} \in Q_{\mathbf{v}}^{d-1}$, then $\mu([\mathbf{u}], [\mathbf{w}]) = \|\mathbf{u} - \mathbf{w}\|$.*

Proof By direct calculation we see:

$$\mu([\mathbf{u}], [\mathbf{w}])^2 = \left\| \sum_{i=1}^d |u_i| \mathbf{e}_i - \sum_{i=1}^d |w_i| \mathbf{e}_i \right\|^2 = \sum_{i=1}^d (|u_i| - |w_i|)^2 = \sum_{i=1}^d (u_i - w_i)^2 = \|\mathbf{u} - \mathbf{w}\|^2 .$$

The first equality uses the definition of μ, and the third equality uses that $\mathbf{u}, \mathbf{w} \in Q_{\mathbf{v}}^{d-1}$, i.e., u_i and w_i share the same sign (up to the possibility of being 0) for each $i \in [d]$. ■

In Proposition 4.3 below, we see that $\bar{G}$ is orthant preserving, and that the iterations $G/\!\sim$ and $\bar{G}|_{Q_+^{d-1}}$ are equivalent under the isometry ϕ. These iterations thus have equivalent fixed point properties. It will suffice to analyze $\bar{G}|_{Q_+^{d-1}}$ in place of $G/\!\sim$.

Proposition 4.3 *Let $\mathbf{v}$ be a sign vector in $\mathbb{R}^d$. Then, $\bar{G}$ has the following properties:*

1. *If $\mathbf{u} \in Q_{\mathbf{v}}^{d-1}$, then $\bar{G}(\mathbf{u}) \in Q_{\mathbf{v}}^{d-1}$.*
2. *If $\mathbf{u}, \mathbf{w} \in S^{d-1}$ are such that $\mathbf{u} \sim \mathbf{w}$, then $G(\mathbf{u}) \sim \bar{G}(\mathbf{w})$.*

Proof We first demonstrate property 1 holds. Let $\bar{h}_1, \ldots, \bar{h}_d$ be defined for $\bar{F}$ in the same way that $h_1, \ldots, h_d$ are defined for F in section 3. Then, $\partial_i \bar{F}(\mathbf{u}) = 2\bar{h}_i'(u_i^2)u_i$ for all $i \in [d]$. Under Lemma 3.1, $\text{sign}(x)\bar{h}_i'(x) \geq 0$ on for all $x \in \mathbb{R}$ and all $i \in [m]$. As $\bar{h}_i := 0$ for all $i \in [d] \setminus [m]$, it follows that $\text{sign}(u_i)\partial_i \bar{F}(\mathbf{u}) \geq 0$ for all $i \in [d]$. Thus, $\bar{G}(\mathbf{u}) \in Q_{\mathbf{v}}^{d-1}$.

We now demonstrate that property 2 holds. Since $\mathbf{u} \sim \mathbf{w}$, there exist sign values $s_i \in \{+1, -1\}$ such that $u_i = s_i w_i$. By Assumption A1 (i.e., g_i and hence its derivative is either an even or odd function), we see that $|\partial_i F(\mathbf{u})| = |g_i'(u_i)| = |g_i'(w_i)| = |\partial_i \bar{F}(\mathbf{w})|$. In particular, it follows that $\|\nabla \bar{F}(\mathbf{w})\| = \|\nabla F(\mathbf{u})\|$, and that $|\bar{G}_i(\mathbf{w})| = |G_i(\mathbf{u})|$ for each $i \in [d]$. Thus, $\bar{G}(\mathbf{w}) \sim G(\mathbf{u})$. ■

Throughout this section, we will assume that $F(\mathbf{u}) = \sum_{i=1}^m g_i(u_i)$ is a PBEF. The functions h_i are defined as in section 3. We will analyze the associated gradient iteration function G on the domain Q_+^{d-1}. It suffices to analyze PBEFs on Q_+^{d-1}, and the results can be easily extended to general BEFs on S^{d-1} due to Proposition 4.3. Unless otherwise stated, we will also assume in this section that $\{\mathbf{u}(n)\}_{n=0}^\infty$ is a sequence in Q_+^{d-1} satisfying $\mathbf{u}(n) = G(\mathbf{u}(n-1))$ for all $n \geq 1$.

We now proceed with the formal analysis of the global stability structure and the rate of convergence of our dynamical system $G/\sim$. It will be seen in section 4.3 that the fast convergence properties of the gradient iteration are due to the strict convexity in assumption A2. However, we will spend most of our time characterizing the stability of fixed points of $G/\sim$, in particular demonstrating that the hidden basis elements $\mathbf{e}_1, \ldots, \mathbf{e}_m$ are attractors, and that for almost any starting point $\mathbf{u}(0)$, $\mathbf{u}(n)$ converges to one of the hidden basis elements as $n \to \infty$.

We now give a brief outline of the argument for the global attraction of the hidden basis elements. For simplicity, we provide this sketch for the case where $d = m$. However, we will later provide all statements and proofs necessary to obtain the global stability in full generality. This argument has four main elements.

1. Enumeration of the fixed points of the gradient iteration (section 4.1). We enumerate the fixed points of G and see that, including the hidden basis elements $\mathbf{e}_1, \ldots, \mathbf{e}_d$, the dynamical system G actually has $2^d - 1$ fixed points in Q_+^{d-1}. In particular, we will see that for any subset $\mathcal{S} \subset [d]$, there exists exactly one fixed point $\mathbf{v}$ of G in Q_+^{d-1} such that $v_i \neq 0$ iff $i \in \mathcal{S}$. The proof of this enumeration of fixed points is based on the expansion $G(\mathbf{u}) = \frac{\nabla F(\mathbf{u})}{\|\nabla F(\mathbf{u})\|}$ where $\nabla F(\mathbf{u}) = \sum_{i=1}^m h_i'(u_i^2)\mathbf{e}_i$ and the monotonicity of the h_i' functions from Lemma 3.1. The proof also uses an observation that the fixed points of G are exactly the critical points of F on S^{d-1} arising in the optimization view.

2. Hyperbolic fixed point structure and stability/instability implications (section 4.2.2). We show that all fixed points of G are hyperbolic, i.e. the eigenvalues of the Jacobian matrix are different from 1 in absolute value (Lemma 4.11). As such, the stability properties of the fixed points of G can be inferred from the eigenvalues of its Jacobian.

We denote by $\mathrm{D}G_{\mathbf{u}}$ the Jacobian of G evaluated at $\mathbf{u}$, and we let $\mathbf{p}$ be a fixed point of G outside of the set $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$. Then, we show that as a linear operator $\mathrm{D}G_{\mathbf{p}} : \mathbf{p}^\perp \to \mathbf{p}^\perp$, $\mathrm{D}G_{\mathbf{p}}$ has at least one eigenvalue with magnitude strictly greater than 1. This implies that $\mathbf{p}$ is locally repulsive for the discrete dynamical system G except potentially on a low dimensional manifold called the local stable manifold of $\mathbf{p}$ (Lemma 4.12). As the local stable manifold of $\mathbf{p}$ is low dimensional, it is also of measure zero. By analyzing the measure of repeated compositions of G^{-1} applied to the local stable manifold of $\mathbf{p}$, we are able to demonstrate that globally on the sphere, the set of starting points $\mathbf{u}(0)$ such that $\mathbf{u}(n) \to \mathbf{p}$ is measure zero (Theorem 4.17).

We will also see that at a hidden basis element $\mathbf{e}_i$, $\mathrm{D}G_{\mathbf{e}_i} : \mathbf{e}_i^\perp \to \mathbf{e}_i^\perp$ is the zero map. In particular, $\mathbf{e}_i$ is an attractor of the dynamical system G. Taken together, these results show that the hidden basis directions $\mathbf{e}_i$ are the attractors of the gradient iteration, and that all other fixed points are unstable.

3. The big become bigger, and the small become smaller (section 4.2.1). We show that coordinates of $\mathbf{u}(n)$ go to zero as $n \to \infty$ under certain conditions. In particular, let $\mathcal{S} \subset [d]$ and let $\mathbf{v}$ be the fixed point of G such that $v_i \neq 0$ if and only if $i \in \mathcal{S}$. An implication of the convexity assumption A2 is that if $u_i > v_i$, then $\partial_i F(\mathbf{u})/u_i > \partial_i F(\mathbf{v})/v_i$, and similarly if $u_i < v_i$, then $\partial_i F(\mathbf{u})/u_i < \partial_i F(\mathbf{v})/v_i$. To see that these orderings hold, we use the expansion $F(\mathbf{u}) = \sum_{i=1}^m h_i(u_i^2)$ to see $\partial_i F(\mathbf{u})/u_i = 2h_i'(u_i^2)$ and we recall (from Lemma 3.1) that each h_i' is an increasing function. Using this monotonicity, we show that each gradient iteration update has the effect of increasing the gap (as a ratio) between $\max_{i\in\mathcal{S}} \partial_i F(\mathbf{u})/u_i$ and $\min_{i\in\mathcal{S}} \partial_i F(\mathbf{u})/u_i$. This implies a divergence between the coordinates of $\mathbf{u}(n)$ under the gradient iteration.

In particular, we show that if there exists an $i \in \mathcal{S}$ and $k \in \mathbb{N}$ such that $u_i(k) > v_i$, then the ratio between maximum magnitude and minimum magnitude coordinate values of $\mathbf{u}(n)$ within $\mathcal{S}$ goes to infinity as $n \to \infty$. In particular, there will exist an $i \in \mathcal{S}$ such that $u_i(n) \to 0$ as $n \to \infty$.

4. Global attraction of the hidden basis (Theorem 4.19). We alternate between applying parts 2 and 3 of this sketch in order to demonstrate that for almost any $\mathbf{u}(0)$, all but one of the coordinates of $\mathbf{u}(n)$ go to zero as n goes to infinity. Part 3 of the sketch allows us to force coordinates of $\mathbf{u}(n)$ to approach 0. By part 2, the trajectory never converges to one of the unstable fixed points of G. This guarantees for any particular unstable fixed point $\mathbf{v}$ that a coordinate of $\mathbf{u}(n)$ eventually exceeds the corresponding non-zero coordinate of $\mathbf{v}$ due to the interplay with part 3. As all but one of the hidden coordinates of $\mathbf{u}(n)$ must eventually go to 0, it follows that $\mathbf{u}(n) \to \mathbf{e}_i$ for some $i \in [m]$ as $n \to \infty$.

4.1. Enumeration of fixed points

We now begin the process of enumerating the fixed points of G. First, we observe that the fixed points of G are very closely related to the maxima structure of F.

Observation 4.4 *A vector $\mathbf{v} \in Q_+^{d-1}$ is a stationary point of G if and only if there exists λ^* such that $(\mathbf{v}, \lambda^*)$ is a critical point of the Lagrangian[13] function $\mathcal{L}(\mathbf{u},\ \lambda) = F(\mathbf{u}) - \lambda[\|u\|^2 - 1]$. In particular, if $\mathbf{v}$ is a stationary point of G, then $\lambda^* \mathbb{1}_{[v_i \neq 0]} = h_i'(v_i^2)$ for each $i \in [d]$.*

Proof This is a result of Lemmas 5.1 and 3.3. ■

With this characterization, we are actually able to enumerate the fixed points G. Note that if $v_i = 0$ for each $i \in [m]$, then by the definition of G, $\mathbf{v}$ is a stationary point. The remaining stationary points are enumerated by the following Lemma.

Lemma 4.5 *Let $\mathcal{S} \subset [m]$ be non-empty. Then there exists exactly one stationary point $\mathbf{v}$ of $G|_{Q_+^{d-1}}$ such that $v_i \neq 0$ for each $i \in \mathcal{S}$ and $v_i = 0$ for each $i \in [m] \setminus \mathcal{S}$. Further, $v_i = 0$ for each $i \in [d] \setminus \mathcal{S}$.*

Proof We prove this in two parts. First, we show that a $\mathbf{v}$ exists with all of the desired properties. Then, we show uniqueness.

Claim 4.5.1 *There exists $\mathbf{v}$ a stationary point of $G|_{Q_+^{d-1}}$ such that $v_i \neq 0$ if and only if $i \in \mathcal{S}$.*

Proof of claim We will construct $\mathbf{v}$ as the limit of a sequence. Consider the following construction of an approximation to $\mathbf{v}$ whose precision depends on the magnitude of $\frac{1}{N}$ where $N \in \mathbb{N}$.

function APPROXFIXPT(N)
$\mathbf{u} \leftarrow \mathbf{0}$
for $i \leftarrow 1$ to N **do**
  $j \leftarrow \arg\min_{k \in \mathcal{S}} h_k'(u_k^2)$
  $u_j \leftarrow \sqrt{u_j^2 + \frac{1}{N}}$
end for
return $\mathbf{u}$

13. This is the same Lagrangian function which arose in Section 3. Its critical points $(\mathbf{u}, \lambda)$ give the locations $\mathbf{u}$ where F satisfies the first order conditions for a constrained extrema on the sphere.

Let $\epsilon_0 > 0$ be fixed. Let $\epsilon_k = \frac{1}{k}\epsilon_0$ for each $k \in \mathbb{N}$. Since $[0,1]$ is a compact space, the h_i's are uniformly equicontinuous on this domain. Thus for each $k \in \mathbb{N} \cup \{0\}$, there exists $\delta_k > 0$ such that for $x, y \in [0,1]$, $|x - y| \leq \delta_k$ implies that $|h_i'(x) - h_i'(y)| \leq \epsilon_k$ for each $i \in \mathcal{S}$. We fix constants $N_k \in \mathbb{N} \cup \{0\}$ such that (1) $\frac{1}{N_k} \leq \delta_k$ for each k, (2) for each $k \geq 1$, N_k is an integer multiple of N_0, and (3) $N_0 \geq |\mathcal{S}|$. Then we construct a sequence $\{\mathbf{u}(k)\}_{k=0}^{\infty}$ by setting $\mathbf{u}(k) = \text{APPROXFIXPT}(N_k)$ for each $k \in \mathbb{N} \cup \{0\}$. It follows by construction that $|h_i'(u_i^2(k)) - h_j'(u_j^2(k))| \leq \epsilon_k$ for each $i, j \in \mathcal{S}$.

It can be seen that $\min_{i \in \mathcal{S}} h_i'(u_i^2(k)) \geq \min_{i \in \mathcal{S}} h_i'(u_i^2(0)) > 0$ for each $k \in \mathbb{N}$. To see the second inequality $\min_{i \in \mathcal{S}} h_i'(u_i^2(0)) > 0$, we note that the h_i's are strictly increasing from 0 by Lemma 3.1, and in particular during the first $|\mathcal{S}|$ iterations of the loop in APPROXFIXPT, a new coordinate of $\mathbf{u}$ will be incremented. To see the first inequality $\min_{i \in \mathcal{S}} h_i'(u_i^2(k)) \geq \min_{i \in \mathcal{S}} h_i'(u_i^2(0))$ for each $k \in \mathbb{N}$, we argue by contradiction. Let $j = \arg\min_{i \in \mathcal{S}} h_i'(u_i^2(k))$. If $h_j'(u_j^2(k)) < \min_{\min i \in \mathcal{S}} h_i'(u_i^2(0))$, then $u_j^2(k) < \min_{i \in \mathcal{S}} u_i^2(0)$, and thus there exists $\ell \in \mathcal{S}$ with $\ell \neq j$ such that $u_\ell^2(k) > u_\ell^2(0)$. However, for this to be true, then during course of the execution of APPROXFIXPT(N_k) the decision must be made at line 4 that $\ell = \arg\min_{k \in \mathcal{S}} h_k'(u_k^2)$ when $u_\ell^2 = u_\ell^2(0)$ (since N_k is an integer multiple of N_0). During this update, strict monotonicity of h_i' implies that $h_j'(u_j^2) \leq h_j'(u_j^2(k)) < \min_{\min i \in \mathcal{S}} h_i'(u_i^2(0)) \leq h_\ell'(u_\ell^2)$. But this contradicts that $\ell = \arg\min_{k \in \mathcal{S}} h_k'(u_k^2)$ at line 4. It follows that there exists a $\Delta > 0$ such that for each $i \in \mathcal{S}$ and each $k \in \{0, 1, 2, \dots\}$ we have $h_i'(u_i^2(k)) > \Delta$, and in particular that $u_i^2(k) \geq \min_{j \in \mathcal{S}} (h_j')^{-1}(\Delta) > 0$.

Since S^{d-1} is compact, there exists a subsequence $i_1, i_2, i_3, \dots$ of $0, 1, 2, \dots$ such that the sequence $\{\mathbf{u}(i_k)\}_{k=1}^{\infty}$ converges to a vector $\mathbf{v} \in S^{d-1}$. Since each $\mathbf{u}(i_k) \in Q_+^{d-1}$, $\mathbf{v} \in Q_+^{d-1}$. Further, since the $u_j^2(i_k)$s are bounded from below by a constant $\Delta' = \min_{j \in \mathcal{S}} (h_j')^{-1}(\Delta) > 0$ for each $j \in \mathcal{S}$, we see that $v_j^2 \geq \Delta' > 0$ for each $j \in \mathcal{S}$. That is, $v_i = 0$ if and only if $i \in \mathcal{S}$. Further, for any $j, \ell \in \mathcal{S}$, $h_\ell'(v_\ell^2) - h_j'(v_j^2) = \lim_{k \to \infty} [h_\ell'(u_\ell^2(i_k)) - h_j'(u_j^2(i_k))] = 0$, and in particular $h_\ell'(v_\ell^2) = h_j'(v_j^2)$. By Observation 4.4, $\mathbf{v}$ is a stationary point of G. ▲

Claim 4.5.2 *There exists only one stationary point* $\mathbf{v}$ *of* $G|_{Q_+^{d-1}}$ *such that the following hold: (1)* $v_i \neq 0$ *if* $i \in \mathcal{S}$ *and (2)* $v_i = 0$ *if* $i \in [m] \setminus \mathcal{S}$.

Proof of Claim. We first show that if $\mathbf{v}$ is a stationary point of $G|_{Q_+^{d-1}}$ meeting the conditions of the claim, then $v_i = 0$ for each $i \in [d] \setminus [m]$. To see this, we use Observation 4.4, and we note that for each $i, j \in [d]$ such that $u_i \neq 0$ and $u_j \neq 0$, then $h_i'(u_i^2) = h_j'(u_j^2)$. In particular, choosing $i \in \mathcal{S}$, we see that $h_i'(u_i^2) > 0$. But for each $i \in [d] \setminus [m]$, $h_i := 0$ implies that $h_i'(u_i^2) = 0$. In particular, for $i \in [d] \setminus [m]$, $u_i = 0$.

Now suppose that there are two stationary points $\mathbf{v}$ and $\mathbf{w}$ meeting the requirements of this Claim. By Observation 4.4, there exists $\lambda_\mathbf{v}$ and $\lambda_\mathbf{w}$ such that $h_i'(v_i^2) = \lambda_\mathbf{v}$ and $h_i'(w_i^2) = \lambda_\mathbf{w}$ for each $i \in \mathcal{S}$. If $\lambda_\mathbf{v} < \lambda_\mathbf{w}$, then strict monotonicity of each h_i' implies that $v_i^2 < w_i^2$ for each $i \in \mathcal{S}$. But this contradicts that $\sum_{i \in \mathcal{S}} v_i^2 = 1 = \sum_{i \in \mathcal{S}} w_i^2$. By similar reasoning, it cannot be that $\lambda_\mathbf{w} < \lambda_\mathbf{v}$. As such, $\lambda_\mathbf{v} = \lambda_\mathbf{w}$, and further for each $i \in \mathcal{S}$ it follows that $h_i'(v_i^2) = h_i'(w_i^2)$. Using strict monotonicity of the h_i's, we see that $\mathbf{v} = \mathbf{w}$.

Note that the $\mathbf{v}$ constructed in Claim 4.5.1 gives the unique solution to this claim. ■

4.2. Convergence to the hidden basis directions

So far, we have enumerated the fixed points of the dynamical G on Q_+^{d-1}. We now analyze the stability properties of these fixes points. In section 4.2.1, we create a divergence criteria from the fixed points of G excluding the hidden basis elements $\mathbf{e}_1, \ldots, \mathbf{e}_m$. This divergence criterion sets up a natural manner under which the large coordinates of $\mathbf{u}(0)$ can increase in magnitude while other coordinates are driven rapidly towards 0. Then, in section 4.2.2, we demonstrate that the set of hidden basis elements of G are essentially global attractors of the dynamical system. In particular, it is seen that each $\mathbf{e}_i$ is locally an attractor, and that for $\mathbf{u}(0)$ drawn from a set of full measure on S^{d-1}, the sequence $\mathbf{u}(n)$ converges to one of the hidden basis elements.

Notation. Throughout this subsection, we will make use of the following notations. Given a $\mathcal{S} \subset [d]$, we define the projection matrix $P_{\mathcal{S}} := \sum_{i \in \mathcal{S}} \mathbf{e}_i \mathbf{e}_i^T$. In particular, this implies $P_{\mathcal{S}}\mathbf{u} := \sum_{i \in \mathcal{S}} u_i \mathbf{e}_i$. We will denote the set complement by $\bar{\mathcal{S}} := [d] \setminus \mathcal{S}$. Two projections will be of particular interest: the projection onto the distinguished basis elements $P_{[m]}\mathbf{u} := \sum_{i=1}^m u_i \mathbf{e}_i$ and its complement projection which we will denote by $P_0\mathbf{u} := \sum_{i=m+1}^d u_i \mathbf{e}_i$. In addition, if $\mathcal{X}$ is a subspace of $\mathbb{R}^d$, we will denote by $P_{\mathcal{X}}$ the orthogonal projection operator onto the subspace $\mathcal{X}$.

We denote by vol_{k-1} the volume measure on the unit sphere S^{k-1}. When the value of k is clear, we suppress it from the notation and simply write vol for the volume measure on the unit sphere ("surface area measure"). Finally, if $f : M \to N$ (with M and N manifolds), we denote by $\mathrm{D}f_{\mathbf{x}}$ the Jacobian (or transposed derivative) of f evaluated at $\mathbf{x}$. We also treat $\mathrm{D}f_{\mathbf{x}}$ as linear operator between tangent spaces: $\mathrm{D}f_{\mathbf{x}} : T_{\mathbf{x}}M \to T_{f(\mathbf{x})}N$, where $T_{\mathbf{x}}M$ denotes the tangent space of M at $\mathbf{x}$. See the book of do Carmo Valero (1992) for an overview of Romanian manifolds and the definition of volume on manifold surfaces.

4.2.1. DIVERGENCE CRITERIA FOR UNSTABLE FIXED POINTS

Proposition 4.6 *There exists $\epsilon > 0$ such that the following holds. Let $\mathbf{v} \in Q_+^{d-1}$ be a stationary point of G such that $\mathcal{S}_{\mathbf{v}} \subset [m]$. Suppose $\|P_{\bar{\mathcal{S}}_{\mathbf{v}}}\mathbf{u}(0)\| \leq \epsilon$ and there exists $i \in \mathcal{S}_{\mathbf{v}}$ such that $u_i(0) > v_i$, then there exists $j \in \mathcal{S}_{\mathbf{v}}$ such that $u_j(n) \to 0$ as $n \to \infty$.*

We now proceed with the proof of Proposition 4.6. We will need a couple of facts about the behavior of small coordinates of $\{\mathbf{u}(n)\}_{n=0}^{\infty}$ under the gradient iteration. In particular, we need to show that $G(\mathbf{u})$ is generally well behaved (i.e., $\|\nabla F(\mathbf{u})\|$ is typically separated from 0), and that the small coordinates of $\mathbf{u}(0)$ are attracted to 0.

Lemma 4.7 *Let F be a fixed BEF. Given $\Delta \in [0, 1)$, there exists $L > 0$ such that the following holds: For all $\mathbf{u} \in Q_+^{d-1}$ such that $\|P_0\mathbf{u}\| \leq \Delta$, $\|\nabla F(\mathbf{u})\| > L$.*

Proof Since $\sum_{i \in [m]} u_i^2 = 1 - \|P_0\mathbf{u}\|^2 \geq 1 - \Delta^2$, there exists $j \in [m]$ such that $u_j \geq \sqrt{\frac{1-\Delta^2}{m}}$. It follows that

$$\begin{aligned}\|\nabla F(\mathbf{u})\|^2 &= \sum_{i=1}^m (2h_i'(u_i^2)u_i)^2 \geq \max_{i \in [m]} 4h_i'(u_i^2)^2 u_i^2 \\ &\geq 4h_j'(u_j^2)^2 u_j^2 \geq \min_{i \in [m]} 4h_i'\Big(\frac{1-\Delta^2}{m}\Big) \cdot \frac{1-\Delta^2}{m} > 0\,.\end{aligned}$$

For the last inequality, we use that each h_i' is strictly increasing on $[0,1]$ from 0. ■

Lemma 4.8 *Let F be a PBEF, let $C > 0$, and let $\Delta \in [0,1)$. There exists $\epsilon > 0$ such that the following holds: Let $\mathbf{u} \in Q_+^{d-1}$ be such that $\|P_0\mathbf{u}\| \leq \Delta$. Define $A_\epsilon := \{i \mid u_i \leq \epsilon\}$. For all $i \in A_\epsilon$, $G_i(\mathbf{u}) < Cu_i$.*

Proof For all $i \in [m]$, h_i' is continuously increasing from $h_i'(0) = 0$. Given any $L > 0$, there exists $\epsilon > 0$ such that for all $i \in [m]$, $u_i \leq \epsilon$ implies that $2h_i'(u_i^2) < CL$. With the choice of L from Lemma 4.7 and the above construction of ϵ, we obtain the following: For all $i \in A_\epsilon$, $G_i(\mathbf{u}) = \frac{2h_i'(u_i^2)u_i}{\|\nabla F(\mathbf{u})\|} < \frac{CLu_i}{L} = Cu_i$. ■

Corollary 4.9 *Let F be a BEF. There exists $\epsilon > 0$ such that the following holds: Let $\{\mathbf{u}(n)\}_{n=0}^\infty$ be a sequence in Q_+^{d-1} defined recursively by $\mathbf{u}(n) = G(\mathbf{u}(n-1))$ such that $\|P_0\mathbf{u}(0)\| \neq 1$. Let $A_\epsilon(n) := \{i \mid u_i(n) \leq \frac{1}{2^n}\epsilon\}$. Then, $A_\epsilon(0) \subset A_\epsilon(1) \subset A_\epsilon(2) \subset \cdots$.*

Proof We then apply Lemma 4.8 with the choice of $C = \frac{1}{2}$ in order to choose ϵ. With this choice of ϵ, we see that $A_\epsilon(n) \supset A_\epsilon(n-1)$ for all $n \in \mathbb{N}$ by Lemma 4.8. ■

In the following Lemma, we identify a useful notion of progress for the gradient iteration.

Lemma 4.10 *The function $n \mapsto \max_{i\in[m]} |h_i'(u_i(n)^2)|$ is a non-decreasing function of n.*

Note that when given a stationary point $\mathbf{v} \in Q_+^{d-1}$, Observation 4.4 implies the existance of $\lambda > 0$ such that $h_i'(v_i^2) = \lambda$ for all $i \in \mathcal{S}_\mathbf{v}$. As the h_i's are strictly functions, we note that for an $i \in \mathcal{S}_\mathbf{v}$ $u_i(k) > v_i$ if and only if each $h_i'(u_i(k)^2) > h_i'(v_i^2)$. This criterion will be useful in demonstrating that once there exists

Proof [of Lemma 4.10] Let $A := \{i \mid u_i(0) \neq 0\} \cap [m]$. We may assume that $A \neq \emptyset$ as otherwise $\{\mathbf{u}(n)\}_{n=0}^\infty$ is a constant sequence, leaving nothing to prove. We only need consider the indices in A since for all $i \in \bar{A}$, $G_i(\mathbf{u}(n+1)) \propto h_i'(u_i(n)^2)u_i(n) = 0$. We note that for $i, j \in A$,

$$\frac{G_i(\mathbf{u}(n+1))}{G_j(\mathbf{u}(n+1))} = \frac{h_i'(u_i(n)^2)}{h_j'(u_j(n)^2)} \cdot \frac{u_i(n)}{u_j(n)} .$$

Fixing $i^* = \arg\max_{i\in A} |h_i'(u_i(n)^2)|$, we see that the ratio $\frac{|G_j(\mathbf{u}(n+1))|}{|G_{i^*}(\mathbf{u}(n+1))|} \leq \frac{|u_j(n)|}{|u_{i^*}(n)|}$ for all $j \in A$. In particular,

$$\frac{1}{G_{i^*}(\mathbf{u}(n+1))^2} = \sum_{j\in A} \frac{G_j(\mathbf{u}(n+1))^2}{G_{i^*}(\mathbf{u}(n+1))^2} \leq \sum_{j\in A} \frac{u_j(n)^2}{u_{i^*}(n)^2} = \frac{1}{u_{i^*}(n)^2}$$

implies that $|G_{i^*}(\mathbf{u}(n+1))| \geq |u_{i^*}|$. As each h_i' is a monotone function on $[0,1]$, it follows:

$$\max_{i\in[m]} |h_i'(u_i(n+1)^2)| \geq |h_{i^*}'(u_{i^*}(n+1)^2)| \geq |h_{i^*}'(u_{i^*}(n)^2)| = \max_{i\in[m]} |h_i'(u_i(n)^2)| .$$

■

We now proceed with the proof of Proposition 4.6.

Proof [of Proposition 4.6] We set $\lambda = h_i'(v_i^2)$ for any $i \in \mathcal{S}_{\mathbf{v}}$. Using Observation 4.4, we see that $\lambda = h_i'(v_i^2)$ for all $i \in \mathcal{S}_{\mathbf{v}}$. We choose $\epsilon > 0$ sufficiently small such that $u_i(0) \leq \epsilon$ implies that $h_i'(u_i(0)) < \lambda$, and also such that ϵ satisfies the conditions of Corollary 4.9.

We will assume that $u_i(0) \neq 0$ for each $i \in \mathcal{S}_{\mathbf{v}}$, since otherwise $u_i(n) = 0$ for all $n \in \mathbb{N}$ (for this choice of i), leaving nothing to prove. We will make use of the following claims.

Claim 4.6.1 *For any $\mathbf{w} \in Q_+^{d-1}$, there exists $j \in \mathcal{S}_{\mathbf{v}}$ such that $w_j \leq v_j$.*

Proof of claim As $\|P_{\mathcal{S}_{\mathbf{v}}}\mathbf{w}\|^2 = \sum_{i\in\mathcal{S}_{\mathbf{v}}} w_i^2 \leq 1 = \sum_{i\in\mathcal{S}_{\mathbf{v}}} v_i^2$, it must hold that for some $j \in \mathcal{S}_{\mathbf{v}}$, $w_j \leq v_j$. Otherwise, we would reverse the inequality, i.e. $\sum_{i\in\mathcal{S}_{\mathbf{v}}} w_i(n)^2 > \sum_{i\in\mathcal{S}_{\mathbf{v}}} v_i^2 = 1$, which yields a contradiction. ▲

Claim 4.6.2 *Given a fixed $\eta > 0$, there exists a choice of $\Delta > 0$ such that the following holds: If $\mathbf{w} \in Q_+^{d-1}$ satisfies that $w_i \neq 0$ for all $i \in \mathcal{S}_{\mathbf{v}}$, that there exists $i \in \mathcal{S}_{\mathbf{v}}$ such that $w_i > v_i$, and that $\max_{i,j\in\mathcal{S}_{\mathbf{v}}} \frac{w_i/v_i}{w_j/v_j} \geq 1+\eta$, then $\max_{i,j} \frac{G_i(\mathbf{w})/v_i}{G_j(\mathbf{w})/v_j} \geq (1+\Delta)\max_{i,j\in\mathcal{S}_{\mathbf{v}}} \frac{w_i/v_i}{w_j/v_j}$.*

Proof of claim Using Observation 4.4, there exists λ such that $\lambda = h_i'(v_i^2)$ for each $i \in \mathcal{S}_{\mathbf{v}}$. Since h_i' is strictly increasing on $[0,1]$ for each $i \in [m]$, there exists a $\Delta > 0$ satisfying the following for each $i \in \mathcal{S}_{\mathbf{v}}$:

1. Whenever $x > v_i + \eta/4$, then $\frac{h_i'(x^2)}{\lambda} > 1+\Delta$ for each $i \in \mathcal{S}_{\mathbf{v}}$.
2. Whenever $x < v_i - \eta/4$, then $\frac{h_i'(x^2)}{\lambda} < \frac{1}{1+\Delta}$ for each $i \in \mathcal{S}_{\mathbf{v}}$.

Further, whenever $\frac{w_i/v_i}{w_j/v_j} \geq 1+\eta$, either $w_i > v_i + \eta/4$ or $w_j < v_j - \eta/4$ holds. This can be seen by arguing via the contrapositive: If neither condition holds, then

$$\frac{w_k/v_k}{w_\ell/v_\ell} \leq \frac{1+\eta/4}{1-\eta/4} = 1 + \frac{\eta/2}{1-\eta/4} < 1+\eta\,,$$

where the last inequality uses that $1 - \eta/4 < \frac{1}{2}$.

Choosing $(i,j) = \arg\max_{i,j\in\mathcal{S}_{\mathbf{v}}} \frac{w_i/v_i}{w_j/v_j}$, we write:

$$\frac{G_i(\mathbf{w}))/v_i}{G_j(\mathbf{w})/v_k} = \frac{h_i'(w_i^2)w_i/v_i}{h_k'(w_j^2)w_j/v_k} = \frac{h_i'(w_i^2)/\lambda}{h_j'(w_j^2)/\lambda}\cdot\frac{w_i/v_i}{w_j/v_j}\,.$$

But by the construction of Δ, we see that one of $h_i'(w_i^2)/\lambda > 1+\Delta$ or $[h_j'(w_j^2)/\lambda]^{-1} > 1+\Delta$. Using that h_i' is strictly increasing we obtain that $h_i'(w_i^2)/\lambda \geq 1+\Delta$ and $[h_j'(w_j^2)/\lambda]^{-1} \geq 1$. Combining these results yields

$$\frac{G_i(\mathbf{w})/v_i}{G_j(\mathbf{w})/v_k} > (1+\Delta)\frac{w_i/v_i}{w_j/v_j}\,.$$

▲

Claim 4.6.3 *Suppose there exists $i_0 \in \mathcal{S}_\mathbf{v}$ such that $u_{i_0}(0) > v_{i_0}$. Then there exists $\Delta > 0$ such that the following holds: Defining $M_n := \max_{i,j \in \mathcal{S}_\mathbf{v}} \frac{u_i(n)/v_i}{u_j(n)/v_j}$, then $M_n \geq (1+\Delta)^n$.*

Proof of claim Setting $\eta = u_{i_0}/v_{i_0} - 1$, we construct Δ as in Claim 4.6.2. We define $i_n := \arg\max_{i \in \mathcal{S}_\mathbf{v}} u_i(n)/v_i$ and $j_n := \arg\min_{j \in \mathcal{S}_\mathbf{v}} u_j(n)/v_j$.

We proceed by induction on n with the following inductive hypothesis: For all $n \in \mathbb{N}$, $M_n \geq (1+\eta)(1+\Delta)^n$ and $u_{i_n}(n) \geq v_{i_n}$.

Base case $n = 0$**.** By Claim 4.6.1, there exists $j \in \mathcal{S}_\mathbf{v}$ such that $u_j(0) \leq v_j$. Thus, $u_{j_0} \leq v_j$. It follows that $\frac{u_{i_0}(0)/v_{i_0}}{u_{j_0}(0)/v_{j_0}} \geq 1 + \eta$. Note that $u_{i_0}(0)/v_{i_0} \geq 1 + \eta$ by the construction of η.

Inductive case. We assume the inductive hypothesis for n We apply Claim 4.6.2 to see the final inequality in:

$$\frac{u_{i_{n+1}}(n+1)/v_i}{u_{j_{n+1}}(n+1)/v_j} \geq \frac{G_{i_n}(\mathbf{u}(n))/v_{i_n}}{G_{j_n}(\mathbf{u}(n))/v_{i_n}} > (1+\Delta)\frac{u_{i_n}(n)/v_i}{u_{j_n}(n)/v_j}\,.$$

To see that $u_{i_{n+1}}(n+1) > v_i$, we use that $h_i'(v_i^2) = \lambda$, strict monotonicity of the h_i's, and Lemma 4.10 to see that $\max_{i \in \mathcal{S}_\mathbf{v}} h_i'(u_i(n+1)^2) \geq h_{i_n}'(u_{i_n}(n)^2) > h_{i_n}'(v_{i_n}^2) = \lambda$. It follows that there exists $i \in \mathcal{S}_\mathbf{v}$ such that $u_i(n+1) > v_i$, and in particular $u_{i_{n+1}}(n+1) > v_{i_{n+1}}$. ▲

Note that as a consequence of Claim 4.6.3, $\min_{i \in \mathcal{S}_\mathbf{v}} u_i(n) \to 0$ as $n \to \infty$. Choose $\epsilon > 0$ according to Corollary 4.9. There exists $j \in \mathcal{S}_\mathbf{v}$ and $N > 0$ such that $u_j(N) < \epsilon$. Applying Corollary 4.9 on the sequence $\{\mathbf{u}(n+N)\}_{n=0}^\infty$, we obtain that $u_j(n+N) \leq \frac{1}{2^n}\epsilon$ for all $n \in \mathbb{N}$, and in particular $u_j(n) \to 0$ as $n \to \infty$. ■

4.2.2. ALMOST EVERYWHERE ATTRACTION OF THE HIDDEN BASIS

In this subsection, we demonstrate that given a generic starting point $\mathbf{u}(n)$, then $\mathbf{u}(n) \to \mathbf{e}_i$ as $n \to \infty$ for some i.

We first demonstrate (in Lemma 4.11 below) that the fixed points of G are hyperbolic. As a direct implication[14], locally to any fixed point $\mathbf{v}$ of G besides the hidden basis elements, there is a manifold M of low dimension such that $\mathbf{v}$ is locally repulsive except on M (Lemma 4.12). In what follows, we will make use of $T_\mathbf{v}S^{d-1}$ the tangent space (or tangent plane) of the sphere S^{d-1} at $\mathbf{v}$ with $\mathbf{v}$ treated as the origin. This may alternatively be defined as $T_\mathbf{v}S^{d-1} := \mathbf{v}^\perp = \{\mathbf{u} \in \mathbb{R}^d \mid \mathbf{u} \perp \mathbf{v}\}$.

Lemma 4.11 (Hyperbolicity of fixed points) *Let $\mathbf{v} \in Q_+^{d-1}$ be a fixed point of G and suppose that $\mathcal{S}_\mathbf{v} := \{i \mid v_i \neq 0\}$ is contained in $[m]$. Let $\phi : T_\mathbf{v}S^{d-1} \to S^{d-1}$ be the exponential[15] map. We let $R = \mathcal{R}(P_{\mathcal{S}_\mathbf{v}})$ and $K = \mathcal{R}(P_{\bar{\mathcal{S}}_\mathbf{v}})$. Then, $\mathrm{D}[\phi \circ G \circ \phi^{-1}]_{\phi(\mathbf{v})}$ is a symmetric matrix which satisfies:*

1. *$[\mathrm{D}[\phi \circ G \circ \phi^{-1}]_{\phi(\mathbf{v})}]|_K$ is the 0 map.*
2. *$[\mathrm{D}[\phi \circ G \circ \phi^{-1}]_{\phi(\mathbf{v})} - \mathcal{I}]|_{R \cap \mathbf{v}^\perp}$ is strictly positive definite. In particular, there exists $\lambda > 0$ such that for any $\mathbf{w} \in R \cap \mathbf{v}^\perp$, $\mathbf{w}^T[\mathrm{D}[\phi \circ G \circ \phi^{-1}]_\mathbf{v} - P_\mathcal{S}]\mathbf{w} \geq \lambda$.*

14. Luo (2012) provides a characterization of the relationship between hyperbolic fixed points and their local stable and unstable manifolds (see in particular his Theorem 2.2).

15. The exponential map for a point on the sphere $\exp_\mathbf{v} : T_\mathbf{v}S^{d-1} \to S^{d-1}$ is defined by $\exp_\mathbf{v}(\mathbf{x}) = \mathbf{v}\cos(\|\mathbf{x}\|) + \frac{\mathbf{x}}{\|\mathbf{x}\|}\sin(\|\mathbf{x}\|)$. For our purposes, we only use that $\exp_\mathbf{v}$ is a coordinate system ϕ of S^{d-1} containing $\mathbf{v}$ such that $\mathrm{D}\phi_\mathbf{v} = \mathrm{D}\phi_\mathbf{v}^{-1} = P_{\mathbf{v}^\perp}$.

Proof We expand the formula $\mathrm{D}[\phi \circ G \circ \phi^{-1}]_{\mathbf{v}}$ to obtain:

$$\mathrm{D}[\phi \circ G \circ \phi^{-1}]_{\mathbf{v}} = \mathrm{D}\phi_{G\circ\phi^{-1}(\mathbf{v})}\mathrm{D}G_{\phi^{-1}(\mathbf{v})}\mathrm{D}\phi_{\mathbf{v}}^{-1} = P_{\mathbf{v}^\perp}\mathrm{D}G_{\mathbf{v}}P_{\mathbf{v}^\perp} \tag{8}$$

Since $G(\mathbf{u}) = \frac{\nabla F(\mathbf{u})}{\|\nabla F(\mathbf{u})\|}$, the Jacobian of G is

$$\mathrm{D}G_{\mathbf{u}} = \frac{\mathcal{H}F(\mathbf{u})}{\|\nabla F(\mathbf{u})\|} - \frac{\nabla F(\mathbf{u})\nabla F(\mathbf{u})^T\mathcal{H}F(\mathbf{u})}{\|\nabla F(\mathbf{u})\|^3} = \frac{P_{G(\mathbf{u})^\perp}\mathcal{H}F(\mathbf{u})}{\|\nabla F(\mathbf{u})\|}\,. \tag{9}$$

As $\mathbf{v}$ is a fixed point of G, equation (9) implies that $\mathrm{D}G_{\mathbf{v}} = \frac{P_{\mathbf{v}^\perp}\mathcal{H}F(\mathbf{v})}{\|\nabla F(\mathbf{v})\|}$. As such, equation (8) becomes

$$\mathrm{D}[\phi \circ G \circ \phi^{-1}]_{\mathbf{v}} = \frac{1}{\|\nabla F(\mathbf{v})\|}P_{\mathbf{v}^\perp}\mathcal{H}F(\mathbf{v})P_{\mathbf{v}^\perp}$$

which is a symmetric map.

Since $\mathbf{v} = G(\mathbf{v}) = \frac{\nabla F(\mathbf{v})}{\|\nabla F(\mathbf{v})\|} = \frac{\sum_{i\in\mathcal{S}} 2h_i'(v_i^2)v_i\mathbf{e}_i}{\|\nabla F(\mathbf{v})\|}$, we see that $2h_i'(v_i^2) = \|\nabla F(\mathbf{v})\|$ for each $i \in \mathcal{S}$. Expanding $\mathcal{H}F(\mathbf{v})$, we thus obtain:

$$\frac{\mathcal{H}F(\mathbf{v})}{\|\nabla F(\mathbf{v})\|} = \sum_{i\in\mathcal{S}} \frac{4h_i''(v_i^2)v_i^2 + 2h_i'(v_i^2)}{\|\nabla F(\mathbf{v})\|}\mathbf{e}_i\mathbf{e}_i^T = \sum_{i\in\mathcal{S}} \frac{4h_i''(v_i^2)v_i^2}{\|\nabla F(\mathbf{v})\|}\mathbf{e}_i\mathbf{e}_i^T + P_{\mathcal{S}}\,.$$

Notice that the first summand is strictly positive definite on $\mathcal{R}(P_{\mathcal{S}})$, and that the second term is the identity map on $\mathcal{R}(P_{\mathcal{S}})$. Careful inspection of the resulting equation

$$\mathrm{D}[\phi \circ G \circ \phi^{-1}]_{\mathbf{v}} = P_{\mathbf{v}^\perp}\Big[\sum_{i\in\mathcal{S}} \frac{4h_i''(v_i^2)v_i^2}{\|\nabla F(\mathbf{v})\|}\mathbf{e}_i\mathbf{e}_i^T + P_{\mathcal{S}}\Big]P_{\mathbf{v}^\perp}$$

gives all of the claimed results. In particular if $\mathbf{x} \in K$, we note that $\mathbf{x} \in \mathbf{v}^\perp$ and $\mathbf{x} \perp \mathbf{e}_i$ for each $i \in \mathcal{S}$; thus, $\Big[\sum_{i\in\mathcal{S}} \frac{4h_i''(v_i^2)v_i^2}{\|\nabla F(\mathbf{v})\|}\mathbf{e}_i\mathbf{e}_i^T + P_{\mathcal{S}}\Big]P_{\mathbf{v}^\perp}\mathbf{x} = 0$. Further, for a non-zero $\mathbf{x} \in R \cap \mathbf{v}^\perp$, we have that the non-zero coordinates of $\mathbf{x}$ are contained in $\mathcal{S}$. Thus, using that the coefficients $4h_i''(v_i^2)v_i^2$ are strictly positive for $i \in \mathcal{S}$, we obtain:

$$\mathbf{x}^T[\mathrm{D}[\phi \circ G \circ \phi^{-1}]_{\mathbf{v}} - \mathcal{I}]\mathbf{x} = \mathbf{x}^T\Big[\sum_{i\in\mathcal{S}} \frac{4h_i''(v_i^2)v_i^2}{\|\nabla F(\mathbf{v})\|}\mathbf{e}_i\mathbf{e}_i^T\Big]\mathbf{x} > 0\,.$$

■

Lemma 4.12 (Local stable manifold) *Suppose that $\mathbf{v} \in Q_+^{d-1}$ is a stationary point of G. Let $\mathcal{S}_{\mathbf{v}} = \{i \mid v_i \neq 0\}$. Suppose that $\mathcal{S}_{\mathbf{v}} \subset [m]$. In a neighborhood U of $\mathbf{v}$ on S^{d-1}, there is a manifold $M_K \subset U$ such that*

1. *$\mathbf{v} \in M_K$.*
2. *$\dim(M_K) = \dim(\mathcal{R}(P_{\bar{\mathcal{S}}})) = d - |\mathcal{S}_{\mathbf{v}}|$.*
3. *There exists a $\delta > 0$ such that if $\mathbf{u}(0) \in U \setminus M_K$, then for some $N \in \mathbb{N}$, $\|\mathbf{u}(N) - \mathbf{v}\| \geq \delta$.*
4. *If $\mathbf{u}(0) \in M_K$, then $\mathbf{u}(N) \to \mathbf{v}$ as $n \to \infty$.*

In Lemma 4.12, M_K is called the local stable manifold of $\mathbf{v}$ (see appendix A.1 for the formal definition).

Proof [of Lemma 4.12] Notice in Lemma 4.11, $K := \mathcal{R}(P_{\bar{\mathcal{S}}})$ is the 0-eigenspace of $[\mathrm{D}[\phi \circ G \circ \phi^{-1}]_{\mathbf{v}}]|_{v^\perp}$, and $R := \mathcal{R}(P_{\mathcal{S}})$ is the span of non-zero eigenvectors of $[\mathrm{D}[\phi \circ G \circ \phi^{-1}]_{\mathbf{v}}]|_{v^\perp}$, with each eigenvalue of R being strictly greater than 1. Further, $\dim(K) = d - |\mathcal{S}_{\mathbf{v}}|$.

Applying Theorem A.3, we obtain the existance of a locally stable manifold M_K for the discrete dynamical system G with $\dim(M_K) = \dim(K) = (d-1)$ (that is property 2). The construction from Theorem A.3 also implies that M_K satisfies the properties 1, 3, and 4. ■

In Lemma 4.12, $\dim(M_K) = d - |\mathcal{S}_{\mathbf{v}}|$ implies a number of things. If $\mathbf{v}$ is one of the hidden basis elements $\mathbf{e}_i$, then $|\mathcal{S}_{\mathbf{e}_i}| = 1$ implies that $\dim(M_K) = \dim(S^{d-1})$. In this case, M_K is an open neighborhood of $\mathbf{e}_i$. Thus, the hidden basis elements are stable attractors.

Proposition 4.13 *The directions $\mathbf{e}_1, \ldots, \mathbf{e}_m$ are attractors of $G|_{Q_+^{d-1}}$.*

Also under Lemma 4.12, if $\mathbf{v} \notin \{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$, then $|\mathcal{S}_{\mathbf{v}}| \geq 2$ and $\dim(M_K) \leq d-2$. In this case, M_K has volume measure 0 on the sphere's surface, and in particular $\mathbf{v}$ is an unstable fixed point of G.

We now wish to demonstrate that the set $\mathcal{X} := \{\mathbf{u}(0) \in S^{d-1} \mid \mathbf{u}(n) \to \mathbf{v} \text{ as } n \to \infty\}$ has measure 0 globally on S^{d-1}. We will proceed first in the setting in which $d = m$. In this setting, we will see that G^{-1} is a well defined function which maps measure 0 sets to measure 0 sets (Lemma 4.14 and Lemma 4.16). Using that $\mathcal{X}$ can alternatively be viewed as the set of preimages of M_K under repeated application of G^{-1} we will obtain that $\mathrm{vol}(\mathcal{X}) = 0$ as desired (Theorem 4.17).

Lemma 4.14 *Suppose that $d = m$ and that F is a PBEF. Then $G : S^{d-1} \to S^{d-1}$ is a continuous bijection.*

Proof Since $F(\mathbf{u}) = \sum_{i=1}^d h_i(u_i^2)$, we obtain

$$G(\mathbf{u}) = \frac{\sum_{i=1}^d 2h_i'(u_i^2)u_i\mathbf{e}_i}{\|\nabla F(\mathbf{u})\|}. \tag{10}$$

To see that G is continuous, we note that Lemma 4.7 implies that $\|\nabla F(\mathbf{u})\| \neq 0$ on its entire domain (since $d = m$). As both the numerator and denominator of equation (10) are continuous, G is continuous.

To see that G is one-to-one, we fix $\mathbf{x}, \mathbf{y} \in S^{d-1}$ and suppose that $G(\mathbf{x}) = G(\mathbf{y})$. Then, $G(\mathbf{x}) = G(\mathbf{y})$ implies that $2h_i'(x_i^2)x_i \propto 2h_i'(y_i^2)y_i$, and in particular there exists $\lambda > 0$ (positive since G is orthant preserving by Proposition 4.3) such that $2h_i'(x_i^2)x_i = \lambda 2h_i'(y_i^2)y_i$ for all $i \in S^{d-1}$. If $\lambda < 1$, then $|h_i'(x_i^2)x_i| < |h_i'(y_i^2)y_i|$ implies (by monotonicity of each h_i') that $x_i^2 < y_i^2$ for all i, which contradicts that $\|\mathbf{x}\|^2 = \|\mathbf{y}\|^2 = 1$. Similarly, it cannot happen that $\lambda > 1$. Thus, $\lambda = 1$, and that $\mathbf{x} = \mathbf{y}$.

We now argue that G is onto. Fix $\mathbf{u} \in S^{d-1}$. We will show that there exists $\mathbf{w}$ such that $G(\mathbf{w}) = \mathbf{u}$. By the symmetries of the problem, we may assume without loss of generality that $u_i \geq 0$ for each $i \in [d]$.

We let $\alpha_1, \ldots, \alpha_\ell$ be an enumeration of $\mathcal{S} := \{i \mid u_i \neq 0\}$. For each $k \in [\ell]$, we define $\Gamma^{(k)} : (0,1] \to \mathbb{R}^k$ by $\Gamma^{(k)}(C) = (x_1, \ldots, x_k)$ such that $h_{\alpha_i}'(x_i^2)x_i/(h_{\alpha_j}'(x_j^2)x_j) = u_{\alpha_i}/u_{\alpha_j}$ for

each $i, j \in [k]$, $\|\Gamma^{(k)}(C)\| = C$, and $x_i > 0$ for all $i \in [k]$. We proceed by induction on k in proving that $\Gamma^{(k)}$ is well defined. In the base case, $\Gamma^{(1)}(C) = (C)$. We now consider the inductive step.

Suppose the inductive hypothesis holds for k. Define $\beta(C,t) := (\Gamma^{(k)}(\sqrt{C-t^2}), t)$. As the functions $x \mapsto h_i'(x^2)x$ are continuous and strictly increasing from 0 when $x_i = 0$, it follows that

$$\rho_C(t) := \frac{h'_{\alpha_{k+1}}(\beta_{k+1}(C,t)^2)\beta_{k+1}(C,t)}{h'_{\alpha_k}(\beta_k(C,t)^2)\beta_k(C,t)}$$

satisfies $\lim_{t\to 0^+} \rho_C(t) = 0$ and $\lim_{t\to C^+} \rho_C(t) = +\infty$. Since ρ_C is a continuous function on $(0, C)$, there exists $t_0 \in (0, C)$ such that $\rho_C(t_0) = \frac{u_{\alpha_{k+1}}}{u_{\alpha_k}}$. In particular, defining $\Gamma^{(k+1)}(C) = (\Gamma^{(k)}(\sqrt{C-t_0^2}), t_0)$ according to this construction, it can be verified that $\|\Gamma^{(k+1)}(C)\| = C$ and that

$$\frac{h'_{\alpha_i}(\Gamma_i^{(k+1)}(C)^2)\Gamma_i^{(k+1)}(C)}{h'_{\alpha_j}(\Gamma_j^{(k+1)}(C)^2)\Gamma_j^{(k+1)}(C)} = \frac{u_{\alpha_i}}{u_{\alpha_j}}$$

for all $i, j \in [k]$ as desired.

By construction, $G\big(\sum_{i=1}^{\ell} \Gamma_i^{(\ell)}(1)\mathbf{e}_{\alpha_i}\big) = \mathbf{u}$. ■

Lemma 4.15 *Let $A := \{\mathbf{u} \in S^{d-1} \mid u_i \neq 0 \text{ for all } i \in [d]\}$. If $d = m$ and if F is a PBEF, then G has the following properties:*

1. *$G(A) = A$.*
2. *For all $\mathbf{p} \in A$, $\mathrm{D}G_{\mathbf{p}} : T_{\mathbf{p}}S^{d-1} \to T_{G(\mathbf{p})}S^{d-1}$ is full rank (invertible).*
3. *$G(\bar{A}) = \bar{A}$.*

Proof We first prove parts 1 and 3. Since each $h_i'(u_i^2)u_i = 0$ if and only if $u_i = 0$ (by Lemma 3.1 and by anti-symmetry of h_i'), it follows from equation (10) both that $\mathbf{u} \in A$ implies $G(\mathbf{u}) \in A$ and that $\mathbf{u} \in \bar{A}$ implies $G(\mathbf{u}) \in \bar{A}$. Thus, $G(A) \subset A$ and $G(\bar{A}) \subset \bar{A}$. Since $G(S^{d-1}) = S^{d-1}$ (by Lemma 4.14), it follows that $G(A) = A$ (since otherwise, $A \not\subset G(A)$ and $A \cap G(\bar{A}) = \emptyset$ implies that $A \not\subset G(A \cup \bar{A}) = G(S^{d-1})$). By similar reasoning, $G(\bar{A}) = G(\bar{A})$.

We now prove part 2. Fix $\mathbf{p} \in A$. Without loss of generality, we assume that $p_i > 0$ for all $i \in [d]$. Fix a non-zero $\mathbf{x} \in T_{\mathbf{p}}S^{d-1}$. Since $\langle \mathbf{p}, \mathbf{x}\rangle = \sum_{i\in d} p_i x_i = 0$ and $\mathbf{x} \neq 0$, there exists $j, k \in [d]$ such that $x_j < 0$ and $x_k > 0$. Note that

$$\mathrm{D}G(\mathbf{p}) = \frac{P_{G(\mathbf{p})^\perp}\mathcal{H}F(\mathbf{p})}{\|\nabla F(\mathbf{p})\|} = \frac{1}{\|\nabla F(\mathbf{p})\|} P_{G(\mathbf{p})^\perp} \sum_{i=1}^{d} [4h_i''(p_i^2)p_i^2 + 2h_i'(p_i^2)]\mathbf{e}_i\mathbf{e}_i^T$$

satisfies (by Lemma 3.1 and Assumption A2) that each $[\mathcal{H}F(\mathbf{u})]_{ii} > 0$; it follows that $[\mathcal{H}F(\mathbf{u})\mathbf{x}]_j < 0$ and $[\mathcal{H}F(\mathbf{u})\mathbf{x}]_k > 0$. Since $G(\mathbf{p}) \in A$ satisfies $G_i(\mathbf{p}) > 0$ for all $i \in [d]$, we see that $\mathcal{H}F(\mathbf{u})\mathbf{x} \nparallel G(\mathbf{p})$, and thus $\mathrm{D}G(\mathbf{p})\mathbf{x} \neq \mathbf{0}$. ■

Lemma 4.16 *Suppose $d = m$ and that $B \subset S^{d-1}$ has volume measure 0. Then, $\mathrm{vol}(G^{-1}(B)) = 0$.*

Proof We let the set A be as in Lemma 4.15. Since $G(\bar{A}) = \bar{A}$ (by Lemma 4.15), then $G^{-1}(\bar{A}) = \bar{A}$. In particular, $G^{-1}(B \cap \bar{A}) \subset \bar{A}$ implies that $\text{vol}(G^{-1}(B \cap \bar{A}) \leq \text{vol}(\bar{A}) = 0$.

On the open set $A = G(A)$, Lemma 4.15 combined with the inverse function theorem implies that G^{-1} exists and is continuously differentiable function. As $B \cap A \subset A$ is a measure 0 set, Theorem A.1 implies that $G^{-1}(B \cap A)$ is a measure 0 set by using an appropriate choice of coordinate atlas for S^{d-1}. For instance, we fix $\mathbf{p} \in \bar{A}$ and let $\phi : \mathbb{R}^{d-1} \to S^{d-1} \setminus \{\mathbf{p}\}$ denote the coordinates arising from the stereographic projection through $\mathbf{p}$. Then, consider the map $\phi^{-1} \circ G^{-1} \circ \phi : \phi^{-1}(A) \to \phi^{-1}(A)$. As the canonical Riemannian metric on the sphere has everywhere positive determinant, $\text{vol}(B \cap A) = 0$ implies that $\phi^{-1}(B \cap A)$ has Lebesgue measure 0. By Theorem A.1, it follows that $\phi^{-1}(G^{-1}(B \cap A)) = \phi^{-1} \circ G^{-1} \circ \phi(\phi^{-1}(B \cap A))$ has Lebesgue measure 0, and hence $\text{vol}(G^{-1}(B \cap A)) = 0$.

Combining these results, we see $\text{vol}(G^{-1}(B)) = \text{vol}(G^{-1}(B \cap A)) + \text{vol}(G^{-1}(B \cap \bar{A})) = 0$. ■

Theorem 4.17 *Suppose that $d = m$. Let $\mathcal{S} \subset [m]$ be such that $|\mathcal{S}| \geq 2$, and let $\mathbf{v}$ be the stationary point of G such that $v_i \neq 0$ if and only if $i \in \mathcal{S}$. Define $\mathcal{X}_{\mathbf{v}} := \{\mathbf{u}(0) \in S^{d-1} \mid \mathbf{u}(n) \to \mathbf{v}$ as $n \to \infty\}$. The set $\mathcal{X}_{\mathbf{v}}$ has volume measure 0 on S^{d-1}.*

Proof In this proof, we denote repeated applications of the gradient iteration and its inverse by

$$G^{(k)} = \underbrace{G \circ \cdots \circ G}_{k \text{ times}} \qquad \text{and} \qquad G^{(-k)} = \underbrace{G^{-1} \circ \cdots \circ G^{-1}}_{k \text{ times}}$$

with $G^{(0)}$ being the identity map.

Let U, M_K, and $\delta > 0$ be as in Lemma 4.12. For each $\mathbf{u}(0) \in \mathcal{X}$, there exists $N > 0$ such that for all $n \geq N$, $\mathbf{u}(n) \in U \cap B(\mathbf{v}, \delta)$. Lemma 4.12 implies that $\mathbf{u}(n) \in M_K$ for all $n \geq N$. In particular, it follows that $\mathbf{u}(0) \in G^{(-n)}(M_K)$ for all $n \geq N$. As such, $\mathcal{X}_{\mathbf{v}} \subset \bigcup_{n=0}^{\infty} G^{(-n)}(M_K)$.

Since $\text{vol}(M_K) = 0$, Lemma 4.16 implies $\text{vol}(G^{(-n)}(M_K)) = 0$ for all $n \in \mathbb{N}$. As such, $\text{vol}(\mathcal{X}) \leq \text{vol}(\bigcup_{n=0}^{\infty} G^{(-n)}(M_K)) \leq \sum_{n=0}^{\infty} \text{vol}(G^{(-n)}(M_K)) = 0$. ■

We now proceed in showing (in the case where $d = m$) that for almost any starting point $\mathbf{u}(0) \in Q_+^{d-1}$, there exists $i \in [d]$ such that $\mathbf{u}(n) \to \mathbf{e}_i$ as $n \to \infty$. The essential ingredients are the preceding measure 0 argument from Theorem 4.17 combined with Proposition 4.6 and Lemma 4.18 below. In particular, Theorem 4.17 implies non-convergence to the unstable fixed points of the dynamical system G from almost any starting point, Proposition 4.6 provides criteria under which coordinates of $\mathbf{u}(n)$ can be driven towards 0, and Lemma 4.18 below will serve as a bridge between the non-convergence to unstable fixed points of G and the preconditions of Proposition 4.6 for demonstrating that all coordinates are driven to 0.

Lemma 4.18 *Let $\mathbf{v} \in Q_+^{d-1}$ be a fixed point of G, and let $\mathcal{S} := \{i \mid v_i \neq 0\}$. Let $\mathbf{u} \in Q_+^{d-1}$ be such that $\|P_{\mathcal{S}}\mathbf{u}\| < \frac{1}{2}\eta$ and such that $\|\mathbf{u} - \mathbf{v}\| > \eta$. Then there exists $i \in \mathcal{S}$ such that $u_i > (1 + \frac{1}{4}\eta^2)v_i$ and $j \in \mathcal{S}$.*

Proof Expanding $\|\mathbf{u} - \mathbf{v}\|^2 > \eta^2$ yields $\|\mathbf{u}\|^2 - 2\langle \mathbf{u}, \mathbf{v}\rangle + \|\mathbf{v}\|^2 > \eta^2$. Hence, $\sum_{i \in \mathcal{S}} u_i v_i < 1 - \frac{1}{2}\eta^2$ (since $\|\mathbf{u}\|^2 = \|\mathbf{v}\|^2 = 1$). Assume for the sake of contradiction that $u_i \leq (1 + \epsilon)v_i$ for all $i \in \mathcal{S}$

where $\epsilon \geq 0$ is arbitrary to be chosen later. Then, $\sum_{i\in\mathcal{S}} u_i v_i \geq \frac{1}{1+\epsilon}\sum_{i\in\mathcal{S}} u_i^2 = \frac{1}{1+\epsilon}(1-\|P_{\bar{\mathcal{S}}}\mathbf{u}\|^2)$. In particular, we obtain:

$$\frac{1}{1+\epsilon}(1-\|P_{\bar{\mathcal{S}}}\mathbf{u}\|^2) < 1-\frac{1}{2}\eta^2$$
$$1 < (1+\epsilon)(1-\frac{1}{2}\eta^2) + \|P_{\bar{\mathcal{S}}}\mathbf{u}\|^2 \,.$$

In particular, with the choices of $\epsilon < \frac{1}{4}\eta^2$ and $\|P_{\bar{\mathcal{S}}}\mathbf{u}\|^2 < \frac{1}{4}\eta^2$, we obtain that

$$\begin{aligned} 1 &< (1+\epsilon)(1-\frac{1}{2}\eta^2) + \|P_{\bar{\mathcal{S}}}\mathbf{u}\|^2 \\ &< (1+\frac{1}{4}\eta^2)(1-\frac{1}{2}\eta^2) + \frac{1}{4}\eta^2 \\ &= 1-\frac{1}{4}\eta^2 - \frac{1}{8}\eta^4 + \frac{1}{4}\eta^2 < 1\,, \end{aligned}$$

which is a contradiction. ■

Theorem 4.19 (Global attraction of the hidden basis) *Suppose that $d = m$. There exists a set $\mathcal{X} \subset Q_+^{d-1}$ with $\mathrm{vol}(\mathcal{X}) = 0$ and the following property: If $\mathbf{u}(0) \in Q_+^{d-1} \setminus \mathcal{X}$, then there exists $i \in [m]$ such that $\mathbf{u}(n) \to \mathbf{e}_i$ as $n \to \infty$.*

Proof Let $\boldsymbol{\mu} : 2^{[m]} \to Q_+^{d-1}$ (denoting by $2^{[m]}$ the power set of $[m]$) be the map which takes $\mathcal{S} \subset [m]$ to $\boldsymbol{\mu}(\mathcal{S})$ the stationary point of G in Q_+^{d-1} such that $\mu_i(\mathcal{S}) \neq 0$ if and only if $i \in \mathcal{S}$. We define $\mathcal{X}_{\boldsymbol{\mu}(\mathcal{S})}$ as in Theorem 4.17. Let $\mathcal{X} := \bigcup\{\mathcal{X}_{\boldsymbol{\mu}(\mathcal{S})} \mid \mathcal{S} \subset [m], |\mathcal{S}| \geq 2\}$. Using Theorem 4.17, we see that $\mathrm{vol}(\mathcal{X}) \leq \sum_{\mathcal{S}\subset[m],|\mathcal{S}|\geq 2} \mathrm{vol}(\boldsymbol{\mu}(\mathcal{S})) = 0$. It remains to be seen that $\mathbf{u}(0) \notin \mathcal{X}$ implies the existance of $i \in [m]$ such that $\mathbf{u}(n) \to \mathbf{e}_i$ as $n \to \infty$. The main idea behind the proof is to demonstrate various coordinates of $\mathbf{u}(n)$ approach 0 until only one coordinate remains separated from 0. We will recurse on the following Claim.

Claim 4.19.1 *Let $\mathcal{S} \subset [m]$ be such that $|\mathcal{S}| \geq 2$. If $u_i(n) \to 0$ for all $i \in \bar{\mathcal{S}}$ as $n \to \infty$, then there exists $j \in \mathcal{S}$ such that $u_j(n) \to 0$ as $n \to \infty$.*

Proof of claim Fix $\mathbf{v} = \boldsymbol{\mu}(\mathcal{S})$. Since $\mathbf{u}(0) \notin \mathcal{X}_{\mathbf{v}}$, there exists $\eta > 0$ and an infinite subsequence $n_0, n_1, n_2, n_3, \ldots$ of $\mathbb{N}$ such that $\|\mathbf{u}(n_i) - \mathbf{v}\| \geq \eta$ for each $i \in \mathbb{N}$. Further, since $\|P_{\bar{\mathcal{S}}}\mathbf{u}(n)\| \to 0$ as $n \to \infty$, there exists $N \in \mathbb{N}$ such that $\|P_{\bar{\mathcal{S}}}\mathbf{u}(n)\| \leq \frac{1}{2}\eta$ for all $n \geq N$. Choose $i \in \mathbb{N}$ such that $n_i \geq N$. By Lemma 4.18, there exists $j \in \mathcal{S}$ such that $u_j(n_i) > v_j$. Thus, Proposition 4.6 implies the existance of $k \in \mathcal{S}$ such that $u_k(n) \to 0$ as $n \to \infty$. ▲

We set $\mathcal{S}_0 = [m]$. Using Claim 4.19.1, we see that there exists $i \in [m]$ such that $u_i(n) \to 0$ as $n \to \infty$. We construct $\mathcal{S}_1 = \mathcal{S}_0 \setminus \{i\}$.

By repeating this application of Claim 4.19.1, we can construct a strictly decreasing sequence $\mathcal{S}_0 \supset \mathcal{S}_1 \supset \cdots \supset \mathcal{S}_{m-1}$ such that for each k, $|\mathcal{S}_k| = m - k$ and for all $i \in \mathcal{S}_k$ $u_i(n) \to 0$ as $n \to \infty$. As $\|P_{\mathcal{S}_{m-1}}\mathbf{u}(n)\|^2 + \|P_{\bar{\mathcal{S}}_{m-1}}\mathbf{u}(n)\|^2 = 1$ with $P_{\bar{\mathcal{S}}_{m-1}}\mathbf{u}(n) \to \mathbf{0}$ as $n \to \infty$, it follows that $\|P_{\mathcal{S}_{m-1}}\mathbf{u}(n)\|^2 \to 1$ as $n \to \infty$. Letting j be the lone element in $\mathcal{S}_{m-1}$, we see that $\mathbf{u}(n) \to \mathbf{e}_j$ as $n \to \infty$. ■

We now extend our result from Theorem 4.19 to the general setting in which $d \geq m$.

Theorem 4.20 *Suppose that $1 \leq m \leq d$. There exists a set $\mathcal{X} \subset Q_+^{d-1}$ with $\mathrm{vol}_{d-1}(\mathcal{X}) = 0$ which has the following property: If $\mathbf{u}(0) \notin \mathcal{X}$, then there exists $i \in [m]$ such that that $\mathbf{u}(0) \to \mathbf{e}_i$ for some $i \in [m]$.*

Proof We note that the case that $m = 1$ is trivial as $G(\mathbf{u}) = \mathbf{e}_1$ for all $\mathbf{u} \in Q_+^{d-1} \setminus \mathbf{e}_1^\perp$, and since $\mathbf{e}_1^\perp$ has volume 0. We assume without loss of generality that $m \geq 2$, thus making S^{m-1} a smooth manifold.

Throughout this proof, we will treat $\mathbb{R}^m$ as a subset of $\mathbb{R}^d$ within $\mathrm{span}\{\mathbf{e}_i \mid i \in [m]\}$ by mapping $(x_1, \ldots, x_m) \mapsto (x_1, \ldots, x_m, 0, \ldots, 0)$ so that we can abuse notation and have $\mathbf{x} \in \mathbb{R}^m$ also part of the domain $\mathbb{R}^d$. In particular, we also will view $S^{m-1} \subset S^{d-1}$ in this fashion.

We first construct a new family of basis encoding functions. In particular, we let $A := B(\mathbf{0}, 1) \cap \mathrm{span}\{\mathbf{e}_i \mid i \notin [m]\}$ (with $B(\mathbf{0}, 1)$ the open ball of radius 1 in $\mathbb{R}^d$). We define the functions $\mathfrak{g}_i : A \times \mathbb{R}$ by $\mathfrak{g}_i(\mathbf{p}, t) := g_i(t\sqrt{1 - \|\mathbf{p}\|^2})$, $\mathfrak{F} : A \times \mathbb{R}^m$ by $\mathfrak{F}(\mathbf{p}, \mathbf{u}) = \sum_{i=1}^m \mathfrak{g}_i(\mathbf{p}, u_i)$, and $\mathfrak{G} : A \times Q_+^{m-1} \to Q_+^{m-1}$ such that $\mathfrak{G}(\mathbf{p}, \bullet)$ is the gradient iteration function associated with $\mathfrak{F}(\mathbf{p}, \bullet)$. Notice that the functions $\mathfrak{F}(\mathbf{p}, \bullet)$ are BEFs. Further, it can be verified that $\mathfrak{G}(\mathbf{p}, \mathbf{u}) = G(\mathbf{p} + \mathbf{u}\sqrt{1 - \|\mathbf{p}\|^2})$. It will sometimes be more convenient to use a more pure function notation, and we thus define $\mathfrak{G}_\mathbf{p} := \mathfrak{G}(\mathbf{p}, \bullet)$.

Define $\mathcal{X}_m$ as $\mathcal{X}$ from Theorem 4.19 for the function $\mathfrak{G}(\mathbf{0}, \bullet) = G|_{\mathrm{span}\{\mathbf{e}_i | i \in [m]\}}$. We note that $\mathrm{vol}_{m-1}(\mathcal{X}_m) = 0$. By Lemma 4.16, we see that $\mathrm{vol}_{m-1}(\mathfrak{G}_\mathbf{p}^{-1}(\mathcal{X}_m)) = 0$ for any $\mathbf{p} \in A$. As such,

$$\mathrm{vol}_{d-1}(G^{-1}(\mathcal{X}_m)) = \int_{\mathbf{p} \in A} (1 - \|\mathbf{p}\|^2)^{m/2} \, \mathrm{vol}_{m-1}(\mathfrak{G}_\mathbf{p}^{-1}(\mathcal{X}_m)) d\mathbf{p} = 0 \, .$$

We define $\mathcal{X} := G^{-1}(X_m) \cup \{\mathbf{u} \mid u_i = 0 \text{ for all } i \in [m]\}$. Note that

$$\mathrm{vol}_{d-1}(\mathcal{X}) \leq \mathrm{vol}_{d-1}(G^{-1}(\mathcal{X}_m)) + \mathrm{vol}_{d-1}(\{\mathbf{u} \mid u_i = 0 \text{ for all } i \in [m]\}) = 0 \, .$$

Also note that for any $\mathbf{u}(0) \notin \mathcal{X}$, $\mathbf{u}(1) \in Q_+^{m-1}$ and $\mathbf{u}(1) \notin \mathcal{X}_m$. Applying Theorem 4.19 to the sequence $\{\mathbf{u}(n)\}_{n=1}^\infty$ with gradient iteration function $G|_{Q_+^{m-1}}$, we obtain that $\mathbf{u}(n) \to \mathbf{e}_i$ for some $i \in [m]$. ■

Using the symmetries of the gradient iteration (Proposition 4.3), the Theorem 2.2 is implied by the combination of Proposition 4.13 and Theorem 4.20.

4.3. Fast convergence of the gradient iteration

We now proceed with the proof of Theorem 2.3. The stability analysis relied on the change of variable $\mathbf{u} \mapsto (u_i^2)$ (which gave rise to the definitions of h_i for $i \in [d]$) due the fact that for each $i \in [m]$, $g_i(x^{1/2})$ is convex on $[0, 1]$. The fast convergence of the gradient iteration algorithm relies on a more general change of variable $\mathbf{u} \mapsto (u_i^r)$ where $r \geq 2$, and in particular it is assumed that $g_i(x^{1/r})$ is convex on $[0, 1]$ for each $i \in [m]$. We encode this potentially stronger convexity constraint within our PBEF by extending the definition of the h_is from section 3 to the more general family of maps $\gamma_{ir} : [0, 1] \to \mathbb{R}$ defined by $\gamma_{ir}(x) := g_i(x^{\frac{1}{r}})$ for $i \in [m]$ and $\gamma_{ir} = 0$ for $i \notin [m]$. We note that $h_i = \gamma_{i2}$ on $[0, 1]$ for each $i \in [d]$. We then write

$$F(\mathbf{u}) = \sum_{i=1}^m g_i(u_i) = \sum_{i=1}^m \gamma_{ir}(u_i^r) \, , \tag{11}$$

where each γ_{ir} is a convex function.

Lemma 4.21 *For all $i \in [m]$, $\gamma'_{ir}(x) = \frac{2}{r}\gamma'_{i2}(x^{\frac{2}{r}})x^{\frac{2-r}{r}}$ on the domain $(0,1]$.*

Proof This is by direct computation. We have the formulas:

$$\gamma'_{i2}(x) = \frac{1}{2}g'_i(x^{\frac{1}{2}})x^{-\frac{1}{2}} \qquad\qquad \gamma'_{ir}(x) = \frac{1}{r}g'_i(x^{\frac{1}{r}})x^{\frac{1-r}{r}}$$

We may rewrite $\gamma'_{ir}(x)$ as follows:

$$\gamma'_{ir}(x) = \frac{2}{r}\cdot\frac{1}{2}g'_i((x^{\frac{2}{r}})^{\frac{1}{2}})(x^{\frac{2}{r}})^{-\frac{1}{2}}x^{\frac{2-r}{r}} = \frac{2}{r}\gamma'_{i2}(x^{\frac{2}{r}})x^{\frac{2-r}{r}} \ .$$

■

Proposition 4.22 *Suppose that $\{\mathbf{u}(n)\}_{n=0}^{\infty}$ is a sequence in Q_+^{d-1} defined recursively by $\mathbf{u}(n) = G(\mathbf{u}(n-1))$ which converges to a $\mathbf{e}_j$ for some $j \in [m]$. Then, the following hold:*

1. *The sequence $\{\mathbf{u}(n)\}_{n=0}^{\infty}$ converges to $\mathbf{e}_j$ at a super-linear rate.*
2. *Fix $r \geq 2$. If $x \mapsto g_i(x^{\frac{1}{r}})$ is convex for every $i \in [m]$, then $\{\mathbf{u}(n)\}_{n=0}^{\infty}$ converges to $\mathbf{e}_j$ with order of convergence at least $r-1$.*

Proof It is sufficient to consider a sequence converging to $\mathbf{e}_1$. If there exists n_0 such that $\mathbf{u}(n_0) = \mathbf{e}_1$, then there is nothing to prove as $\mathbf{e}_1$ is a stationary point of G. So, we assume that $\mathbf{u}(n) \neq \mathbf{e}_1$ for all $n \in \mathbb{N}$.

Taking derivatives of F from equation (11), we get: $\partial_i F(\mathbf{v}) = r\gamma'_{ir}(v_i^r)v_i^{r-1}$. We will make use of the following ratios in analyzing the rate of convergence of $\mathbf{u}(n)$:

$$\rho(i,j;n) := \frac{u_i(n)}{u_j(n)} = \frac{\gamma'_{ir}(u_i(n-1)^r)u_i(n-1)^{r-1}}{\gamma'_{jr}(u_j(n-1)^r)u_j(n-1)^{r-1}} \ .$$

Define $U = \gamma'_{1r}(1)$ and $L = \max_{j\neq 1}\{\lim_{x\to 0^+}\gamma'_{jr}(x)\}$. We note that the strict convexity of $x \mapsto g_i(\sqrt{x})$ (for $i \in [m]$) implies that $\gamma'_{i2}(1) > 0$, and since Lemma 4.21 implies $\gamma'_{ir}(1) = \frac{2}{r}\gamma'_{i2}(1) > 0$, it follows that $U > 0$. Since γ_{ir} is convex, γ'_{jr} is a non-decreasing function. It follows that L is well defined and is also equal to $\max_{j\neq 1}\{\inf_{x>0}\gamma'_{jr}(x)\}$. Finally, noting that γ'_{i2} is non-negative on $[0,1]$ (indeed, γ'_{i2} is increasing from $\gamma'_{i2}(0) = 0$ by Lemma 3.1), it follows from Lemma 4.21 that $\gamma'_{ir}(x) \geq 0$ for all $x > 0$, and in particular $L \geq 0$.

Fix $\epsilon \in (0, \frac{1}{2}U)$. There exists $\delta > 0$ such that:

1. If $\mathbf{v} \in Q_+^{d-1}$ is such that $1 - v_1 < \delta$, then $\gamma'_{1r}(u_1) > U - \epsilon$. The existence of such a choice for δ is implied by the continuity of g'_1 and hence γ'_{1r} near 1.

2. If $\mathbf{v} \in Q_+^{d-1}$ is such that $v_j < \delta$ for some $j \neq 1$, then $\gamma'_{jr}(u_j) < L + \epsilon$. The existence of such a δ follows from the characterization of L as $\max_{j\neq 1}\{\inf_{x>0}\gamma'_{jr}(x)\}$ and γ'_{jr} being monotonic on $[0,1]$.

Fix N sufficiently large that for each $n \geq N$, $\|\mathbf{e}_1 - \mathbf{u}(n)\|_1 < \delta$. With any fixed $j \neq 1$ and $n \geq N+1$, it follows that

$$\rho(j,1;n) = \frac{\gamma'_{jr}(u_j(n-1)^r)u_j(n-1)^{r-1}}{\gamma'_{1r}(u_i(n-1)^r)u_1(n-1)^{r-1}} < \frac{L+\epsilon}{U-\epsilon} \cdot \frac{u_j(n-1)^{r-1}}{u_1(n-1)^{r-1}} \,. \tag{12}$$

Denote by $\mathbf{u}'$ the vector $\sum_{i=2}^d u_i \mathbf{e}_i$. Then,

$$\begin{aligned}\|\mathbf{e}_1 - \mathbf{u}(n)\| &= \|\mathbf{e}_1(1-u_1(n)) - (\mathbf{u}(n) - u_1(n)\mathbf{e}_1)\| \\ &\leq \|\mathbf{e}_1(1-u_1(n))\| + \|\mathbf{u}'(n)\| = 1 - u_1(n) + \|\mathbf{u}'(n)\| \,.\end{aligned}$$

Since $\mathbf{u}$ is a unit vector, we see that $u_1(n) + \|\mathbf{u}'(n)\| \geq u_1(n)^2 + \|\mathbf{u}'(n)\|^2 = 1$. It follows that $1 - u_1(n) \leq \|\mathbf{u}'(n)\|$. Thus,

$$\begin{aligned}\|\mathbf{e}_1 - \mathbf{u}(n)\| &\leq 2\|\mathbf{u}'(n)\| \leq 2\|\mathbf{u}'(n)\|_1 = 2\sum_{i=2}^d u_i(n) \\ &\leq 2\sum_{i=2}^d \rho(i,1;n) < 2 \cdot \frac{L+\epsilon}{U-\epsilon} \cdot \frac{\sum_{i=2}^d u_i(n-1)^{r-1}}{u_1(n-1)^{r-1}}\end{aligned}$$

where the second to last inequality uses that $\mathbf{u}(n)$ is a unit vector making $u_1(n) \leq 1$, and the last inequality uses equation (12). Continuing (with $n \geq N+1$), we see $u_1(n-1) \geq 1 - \|\mathbf{e}_1 - \mathbf{u}(n-1)\|_1 \geq 1 - \delta$. Hence,

$$\|\mathbf{e}_1 - \mathbf{u}(n)\| < 2 \cdot \frac{L+\epsilon}{(U-\epsilon)(1-\delta)^{r-1}} \cdot \sum_{i=2}^d u_i(n-1)^{r-1} \,.$$

Since for each $i \geq 2$ we have $u_i(n-1) \leq \|\mathbf{e}_1 - \mathbf{u}(n-1)\|$

$$\frac{\|\mathbf{e}_1 - \mathbf{u}(n)\|}{\|\mathbf{e}_1 - \mathbf{u}(n-1)\|^{r-1}} < 2d \cdot \frac{L+\epsilon}{(U-\epsilon)(1-\delta)^{r-1}} \,.$$

As the right hand side is a finite constant, the sequence has order of convergence at least $r-1$. In the case where $r = 2$, Lemma 3.1 combined with the fact that $\gamma'_{i2} = 0$ for each $i \in [d] \setminus [m]$ implies that $\lim_{x \to 0^+} \gamma'_{i2}(x) = 0$ for each $i \in [d]$; and in particular, $L = 0$. Since ϵ can be chosen arbitrarily small, the sequence $\{\mathbf{u}(n)\}_{n=0}^\infty$ has super-linear convergence even when $r = 2$. ■

Under Proposition 4.3, part 1 of Theorem 2.3 is implied by Proposition 4.22. Part 2 of Theorem 2.3 follows from the fact that for any i such that $\mathbf{u} \perp \mathbf{e}_i$, then $\partial_i F(\mathbf{u}) = 0$ implies that $G(\mathbf{u}) \perp \mathbf{e}_i$. In particular, it can be seen by induction on n that for a sequence defined recursively by $\mathbf{u}(n) = G(\mathbf{u}(n-1))$ and $\mathbf{u}(0) \perp \mathbf{e}_i$, then $\mathbf{u}(n) \perp \mathbf{e}_i$ for all $n \in \mathbb{N}$ and in particular $\mathbf{u}(n) \not\to \mathbf{e}_i$.

5. Connections of gradient iteration to gradient ascent and power methods

In this section, we briefly interpret the gradient iteration as a form of adaptive, projected gradient ascent. As the gradient iteration is also a generalized power iteration, these dual interpretations

closely link the gradient iteration and other power methods with hill climbing techniques for finding the maxima of a function[16]. In particular, this connection gives a conceptual explanation of the relationship between the fixed points of the gradient iteration and the maxima structure of a BEF F on the unit sphere. For the remainder of this section, we take F to be a PBEF.

The projected gradient ascent update (with learning rate η) is given in the function GRADASCENTUPDATE below.

Algorithm 1 A single projected gradient ascent step for function maximization over S^{d-1}.

1: **function** GRADASCENTUPDATE($\mathbf{u}$, η)
2: $\mathbf{u}' \leftarrow \mathbf{u} + \eta P_{\mathbf{u}^\perp} \nabla F(\mathbf{u})$
3: **return** $\frac{\mathbf{u}'}{\|\mathbf{u}'\|}$

The update in GRADASCENTUPDATE differs from the standard gradient ascent in two ways. First, the update occurs in the direction $P_{\mathbf{u}^\perp} \nabla F(\mathbf{u})$ rather than $\nabla F(\mathbf{u})$. This takes into account the geometry structure of S^{d-1} by updating within the plane tangent to S^{d-1} at $\mathbf{u}$. This arises naturally when treating S^{d-1} as a manifold with the local coordinate system defined by the projective space centered at $\mathbf{u}$. Then, $\mathbf{u}'$ is projected back onto the sphere in order to stay within S^{d-1}.

We now compare the update rules $\mathbf{u} \leftarrow$ GRADASCENTUPDATE($\mathbf{u}$, η) and $\mathbf{u} \leftarrow G(\mathbf{u})$. If $P_{\mathbf{u}^\perp} \nabla F(\mathbf{u}) = \mathbf{0}$, then both updates are the identity map and are thus identical. If $P_{\mathbf{u}^\perp} \nabla F(\mathbf{u}) \neq \mathbf{0}$, then

$$G(\mathbf{u}) = \frac{\nabla F(\mathbf{u})}{\|\nabla F(\mathbf{u})\|} = \frac{\langle \nabla F(\mathbf{u}), \mathbf{u} \rangle \mathbf{u} + P_{\mathbf{u}^\perp} \nabla F(\mathbf{u})}{\|\nabla F(\mathbf{u})\|} = \frac{\mathbf{u} + P_{\mathbf{u}^\perp} \nabla F(\mathbf{u}) / \langle \nabla F(\mathbf{u}), \mathbf{u} \rangle}{\|\nabla F(\mathbf{u})\| / \langle \nabla F(\mathbf{u}), \mathbf{u} \rangle} . \quad (13)$$

The numerator of the rightmost fraction can be interpreted as line 2 of GRADASCENTUPDATE($\mathbf{u}$, η) using the choice $\eta = \langle \mathbf{u}, \nabla F(\mathbf{u}) \rangle^{-1}$. Lemma 3.1 implies that $u_i > 0$ if and only if $\partial_i F(\mathbf{u}) = 2h_i'(u_i^2)u_i > 0$. More generally, the symmetries from Assumption A1 imply that $\operatorname{sign}(u_i) = \operatorname{sign}(\partial_i F(\mathbf{u}))$ for all $i \in [m]$. As such, $\eta = \langle \mathbf{u}, \nabla F(\mathbf{u}) \rangle^{-1} > 0$ is a valid learning rate generically (whenever $\nabla F(\mathbf{u}) \neq \mathbf{0}$). The denominator of the rightmost fraction in equation (13) gives the normalization to project back onto the unit sphere (line 3 of GRADASCENTUPDATE). We obtain the following relationship between gradient ascent and gradient iteration.

Lemma 5.1 *The update $\mathbf{u} \leftarrow G(\mathbf{u})$ is an adaptive form of projected gradient ascent. Specifically,*

1. *If $\nabla F(\mathbf{u}) \neq \mathbf{0}$, then $G(\mathbf{u}) =$ GRADASCENTUPDATE$(\mathbf{u}, \langle \mathbf{u}, \nabla F(\mathbf{u}) \rangle^{-1})$.*
2. *If $\nabla F(\mathbf{u}) = \mathbf{0}$ and $\eta \in \mathbb{R}$ is arbitrary, then $G(\mathbf{u}) =$ GRADASCENTUPDATE$(\mathbf{u}, \eta)$.*

The step size chosen by the gradient iteration function is in several ways very good. By Proposition 4.3, $G(\mathbf{u})$ and hence $\nabla F(\mathbf{u})$ belong to the same orthant as $\mathbf{u}$. As such we never overshoot a basis direction $\mathbf{e}_i$ during the ascent procedure. Further, the gradient iteration has the fast convergence properties stated in Theorem 2.3.

6. Gradient iteration under a perturbation

16. We note that in a special setting of recovering a parallelopiped a closely related observation was made by Nguyen and Regev (2009).

In section 4, we saw that the hidden basis elements $\mathbf{e}_i$ are attractors, that convergence to this set of attractors is guaranteed except on a set of measure 0, and that the rate of convergence is super-linear. In this section, we provide a robust extension to the gradient iteration algorithm for recovering all of the hidden basis elements. We demonstrate that for a wide class of contrasts, the recovery process is robust to a perturbation, and that the hidden basis elements $\mathbf{e}_1, \ldots, \mathbf{e}_m$ can be efficiently recovered given approximate access to ∇F.

To provide quantifiable algorithmic bounds, we require quantifiable assumptions upon the hidden convexity (or concavity) of the h_i functions associated with F. For smooth functions, convexity is characterized by the second derivative of the function. In particular, we use the following notion of robustness.

Definition 6.1 *If for strictly positive constants α, β, γ, and δ, $\beta x^{\delta-1} \leq |h_i''(x)| \leq \alpha x^{\gamma-1}$ on $(0,1]$ for each $i \in [m]$, then F is said to be $(\alpha, \beta, \gamma, \delta)$-robust.*

This definition is designed to capture a broad class of functions of interest. For instance, we capture monomials of the form $p_{a,r}(x) = \frac{a}{(r+1)r} x^{2r+2}$ on $[0,1]$ where $r > 0$ and $a > 0$ are real (with either positive or negative reflections of this on $[-1,0]$. Indeed, Definition 6.1 may alternatively be stated as $\frac{d^2}{dt^2}(p_{\beta,\delta}(\sqrt{t}))\big|_{t=x} \leq |h_k''(x)| \leq \frac{d^2}{dt^2}(p_{\alpha,\gamma}(\sqrt{t}))\big|_{t=x}$. In particular, the monomial functions ax^r with $r \geq 3$ an integer which arise in the setting of orthogonal tensor decompositions are captured as a special case.

Definition 6.1 provides several natural condition numbers which arise in our analysis.

Remark 6.2 *If F is $(\alpha, \beta, \gamma, \delta)$-robust, then $\alpha \geq \beta$ and $\gamma \leq \delta$.*

Proof To see that $\alpha \geq \beta$, we note that $\alpha x^{\gamma-1} \geq \beta x^{\delta-1}$ holds at $x = 1$. To see that $\gamma \leq \delta$, we note that asymptotically as $x \to 0$ from the right, $\beta x^{\delta-1} = O(x^{\gamma-1})$. ■

Under Remark 6.2, we see that $\frac{\alpha}{\beta}$ and $\frac{\delta}{\gamma}$ are both lower bounded by 1. These ratios will act as condition numbers in our time and error bounds.

For the remainder of this section, we will assume that F is $(\alpha, \beta, \gamma, \delta)$-robust unless otherwise specified. Hatted objects such as $\widehat{\nabla F}$ and $\hat{G}$ will represent the natural estimates of un-hatted objects, and in particular

$$\hat{G}(\mathbf{u}) := \begin{cases} \widehat{\nabla F}(\mathbf{u}) / \|\widehat{\nabla F}(\mathbf{u})\| & \text{if } \widehat{\nabla F}(\mathbf{u}) \neq \mathbf{0} \\ \mathbf{u} & \text{otherwise} \end{cases}.$$

For $\epsilon > 0$, we say that $\widehat{\nabla F}$ is an *ϵ-approximation* of ∇F if $\|\widehat{\nabla F}(\mathbf{u}) - \nabla F(\mathbf{u})\| \leq \epsilon$ for all $\mathbf{u} \in \overline{B(0,1)}$. We assume (unless otherwise stated) throughout this section that $\widehat{\nabla F}$ is an ϵ-approximation of ∇F with any bounds on ϵ being made clear by context.

Algorithm 2 Perform the gradient iteration for a predetermined number of iterations. The inputs are $\mathbf{u}(0)$ (an initialization vector) and N (the number of iterations). The output is $\mathbf{u}(N)$ (the N^{th} element of the resulting gradient iteration sequence).

function GI-LOOP($\mathbf{u}(0), N$)
for $n \leftarrow 1$ to N **do**
  $\mathbf{u}(n) \leftarrow \hat{G}(\mathbf{u}(n-1))$
end for
return $\mathbf{u}(N)$

Algorithm 3 A robust extension to the gradient iteration algorithm for guaranteed recovery of a single hidden basis element.

Inputs:

$\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$ A (possibly empty) set of approximate hidden basis directions.

σ Positive parameter determining jump size to break stagnation of $\hat{G}$.

Outputs: An approximate basis element not estimated by any of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$.

function FINDBASISELEMENT($\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}, \sigma$)
// Find a starting vector sufficiently outside the subspace $\mathrm{span}(\mathbf{e}_{m+1}, \ldots, \mathbf{e}_d)$.
Let $\mathbf{x}_1, \ldots, \mathbf{x}_{d-k}$ be orthonormal vectors in $\mathrm{span}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)^{\perp}$.
$j \leftarrow \arg\max_{i \in [d-k]} \|\widehat{\nabla F}(\mathbf{x}_i)\|$
$\mathbf{u} \leftarrow \hat{G}(\mathbf{x}_j)$ // "Zero" the values of $u_{m+1}, \ldots, u_d$.
$\mathbf{u} \leftarrow$ GI-LOOP($\mathbf{u}, N_1$)
for $i \leftarrow 1$ to I **do**
  // Start of the main loop
  Draw $\mathbf{x}$ uniformly at random from $\sigma S^{d-1} \cap \mathbf{u}^{\perp}$
  $\mathbf{w} \leftarrow \mathbf{u}\cos(\|\mathbf{x}\|) + \frac{\|\mathbf{x}\|}{\mathbf{x}}\sin(\|\mathbf{x}\|)$ // A random jump from $\mathbf{u}$
  $\mathbf{u} \leftarrow$ GI-LOOP($\mathbf{w}, N_2$)
end for
return $\mathbf{u}$

Algorithm 4 A robust algorithm to recover approximations to all of the hidden basis elements.

Inputs:

$\hat{m}$ The desired number of basis elements to recover. It is required that $\hat{m} \geq m$.

σ Positive parameter determining jump size to break stagnation of $\hat{G}$.

Outputs:

$\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{\hat{m}}$ The first m of these are approximate hidden basis elements.

function ROBUSTGI-RECOVERY($\hat{m}, \sigma$)
for $i \leftarrow 1$ to $\hat{m}$ **do**
  $\boldsymbol{\mu}_i \leftarrow$ FINDBASISELEMENT($\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{i-1}\}, \hat{m}, \sigma$)
end for
return $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{\hat{m}}$

We will see that under these assumptions, FINDBASISELEMENT (Algorithm 3) robustly recovers a single hidden basis element $\pm\mathbf{e}_i$ using $\hat{G}$ and occasional random jumps. The recovery takes into account previously found basis elements. In particular, if $\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_k}$ are approximated by $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$, then FINDBASISELEMENT approximately recovers a hidden basis element $\mathbf{e}_{i_{k+1}}$ such that $i_{k+1} \in [m] \setminus \{i_1, \ldots, i_k\}$. In particular, FINDBASISELEMENT may be run repeatedly to recover all hidden basis elements. Formally, we have the following result.

For clarity, we will denote by $C_0, C_1, C_2, \ldots$ positive universal constants in the main theorem statements. These can represent different constant values in different theorem statements.

Theorem 6.3 *Suppose that*

- $\sigma \leq \frac{C_0}{\sqrt{d(1+\delta)}}\left[\frac{\beta\gamma}{16\alpha\delta}\right]^{\frac{1}{\gamma}} m^{-\frac{\delta}{\gamma}}$,
- $\epsilon \leq C_1 4^{-\frac{4+2\delta}{\gamma}} \frac{\sigma\beta}{\delta}\left[\frac{\beta\gamma}{\alpha\delta}\right]^{\frac{4\delta+7}{2\gamma}} m^{-\frac{\delta}{\gamma}(2\delta-\gamma+\frac{7}{2})} d^{-\frac{1}{2}-\delta}$,
- $N_1 \geq C_2\lceil \log_{1+2\gamma}(\log_2(\frac{\beta}{\delta\epsilon}))\rceil$, *and*
- $N_2 \geq C_3\lceil 4^{\frac{2}{\gamma}}\frac{\sqrt{d}}{\sigma}(\frac{\alpha\delta}{\beta\gamma})^{\frac{\delta+2}{\gamma}} m^{\frac{\delta}{\gamma}(\delta-\gamma+2)}[\frac{1}{\gamma}\log(\frac{\alpha\delta}{\beta\gamma}) + \frac{\delta}{\gamma}\log(m)]\rceil + C_2\lceil \log_{1+2\gamma}(\log_2(\frac{\beta}{\delta\epsilon}))\rceil$.

Let $p \in (0,1)$, and suppose that $I \geq C_4 m\lceil\log(m/p)\rceil$. Suppose that $\|\boldsymbol{\mu}_i - \mathbf{e}_i\| \leq C_5\delta\epsilon/\beta$. After executing $\boldsymbol{\mu}_{k+1} \leftarrow$ FINDBASISELEMENT$(\{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k\},\sigma)$, then with probability at least $1-p$, there exists $j \in [m]\setminus[k]$ such that $\|\pm\boldsymbol{\mu}_{k+1} - \mathbf{e}_j\| \leq C_5\delta\epsilon/\beta$.

FINDBASISELEMENT operates as follows. We first find a warm start $\mathbf{u}$ which is approximately contained in $\mathrm{span}(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k)$ for which $\|P_0\mathbf{u}\|$ is small. Then, we enter the main loop. There are three main ideas underlying the main loop and its analysis:

Small coordinates decay rapidly. There exists a threshold $\tau > 0$ such that if $i \in [m]$ satisfies that $|u_i| \leq \tau$, then when applying the gradient iteration $|\hat{G}_i(\mathbf{u})| \leq \frac{C}{|u_i|}$ (with $C < 1$) unless u_i is already on the order of ϵ. We call coordinates of $\mathbf{u}$ small if they are below such a threshold and large if they are above it. This constant C actually gets smaller as the u_i gets smaller, and we see super-exponential decay in the small coordinates of $\mathbf{u}$. This super-exponential decay is seen in the lower bound on N_1, which interestingly includes the only dependency on ϵ seen in the running time of FINDBASISELEMENT.

The big become bigger. During the execution of step 11, we may consider a fixed point $\mathbf{v}$ of $G/\!\sim$ such that $v_i \neq 0$ if and only if i corresponds to a large coordinate of $\mathbf{w}$. Similarly to what was seen before in Proposition 4.6 in the exact case, if there is an i such that $w_i > v_i$ with a sufficient gap, then the gradient iteration drives one of the large coordinates to become small. The remaining large coordinates become bigger to compensate. When finally only one hidden coordinate of $\mathbf{u}$ remains big, we have recovered an approximate hidden basis element.

Jumping out of stagnation. It is possible for the gradient iteration to stagnate. In particular, this can occur as follows. If $\mathcal{S} \subset [m]$ is the set of large coordinates, $\mathbf{v}$ is the fixed point of $G/\!\sim$ such that $v_i \neq 0$ if and only if $i \in \mathcal{S}$, and if $|u_i| \leq |v_i|$ (or under the perturbed setting, $|u_i|$ is not sufficiently larger than $|v_i|$ from the unperturbed setting), then the large coordinate progress from the preceding paragraph is not guaranteed. However, by taking a small random jump from $\mathbf{u}$ as is done in steps 9 and 10 of FINDBASISELEMENT, then with at least constant probability, we can make one of the large coordinates of $\mathbf{u}$ sufficiently greater than the corresponding coordinate of $\mathbf{v}$. Then, the large coordinate analysis from the preceding paragraph applies. It is from this interplay between the big becoming bigger and the jumping out of stagnation that we are able guarantee with probability $1-\Delta$ that $O(m\log(m/\Delta))$ iterations of the main loop suffice to drive all but one of the hidden coordinates of $\mathbf{u}$ to 0, and hence producing an approximation to one of the hidden basis elements.

Finally, in ROBUSTGI-RECOVERY (Algorithm 4), we run FINDBASISELEMENT until all hidden basis elements are well approximated. More formally, we have the following result. For clarity, we use $\pm$ to denote unknown signs.

Theorem 6.4 *Suppose that*

- $\sigma \leq \frac{C_0}{\sqrt{d(1+\delta)}}\left[\frac{\beta\gamma}{16\alpha\delta}\right]^{\frac{1}{\gamma}} m^{-\frac{\delta}{\gamma}}$,
- $\epsilon \leq C_1 4^{-\frac{4+2\delta}{\gamma}} \frac{\sigma\beta}{\delta}\left[\frac{\beta\gamma}{\alpha\delta}\right]^{\frac{4\delta+7}{2\gamma}} m^{-\frac{\delta}{\gamma}(2\delta-\gamma+\frac{7}{2})} d^{-\frac{1}{2}-\delta}$,
- $N_1 \geq C_2\lceil \log_{1+2\gamma}(\log_2(\frac{\beta}{\delta\epsilon}))\rceil$, *and*
- $N_2 \geq C_3\lceil 4^{\frac{2}{\gamma}}\frac{\sqrt{d}}{\sigma}(\frac{\alpha\delta}{\beta\gamma})^{\frac{\delta+2}{\gamma}} m^{\frac{\delta}{\gamma}(\delta-\gamma+2)}[\frac{1}{\gamma}\log(\frac{\alpha\delta}{\beta\gamma}) + \frac{\delta}{\gamma}\log(m)]\rceil + C_2\lceil \log_{1+2\gamma}(\log_2(\frac{\beta}{\delta\epsilon}))\rceil$.

Let $p \in (0,1)$, and suppose that $I \geq C_3 m\lceil \log(m/p)\rceil$.

If $\hat{m} \geq m$ and we execute $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{\hat{m}} \leftarrow$ ROBUSTGI-RECOVERY$(\hat{m}, \sigma)$, then with probability at least $1-p$, there a permutation π of $[m]$ such that $\|\pm\boldsymbol{\mu}_i - \mathbf{e}_{\pi(i)}\| \leq C_4\delta\epsilon/\beta$ for all $i \in [m]$.

The proofs of Theorems 6.3 and 6.4 are omitted from this extended abstract. They can be found in the full length version of this paper (Belkin et al., 2016, v4).

We now consider the running time of ROBUSTGI-RECOVERY. First, I, N_1, and N_2 can be viewed as parameters controlling the running time of the algorithm. More formally, we have the following result.

Theorem 6.5 *Suppose that we are working in a computation model supporting the following operations: Basic arithmetic operations, square roots, and trigonometric functions on scalars, branches on conditional; inner products in $\mathbb{R}^d$; and computations of $\widehat{\nabla F}(\mathbf{u})$. Then,* ROBUSTGI-RECOVERY *runs in $O(\hat{m}(N_1 + IN_2) + \hat{m}d^2)$ time.*

To see the $O(\hat{m}d^2)$ portion of the upper bound on scalar and vector operations in Theorem 6.5, we note that step 3 of FINDBASISELEMENT can be implemented using Gram-Schmidt orthogonalization involving the $\boldsymbol{\mu}_i$s and the canonical vectors in the ambient space. When the desired number of basis elements m is known, then $\hat{m}$ can be chosen as m. When the number of basis elements is unknown, then $\hat{m}$ may be chosen as d, and in a more practical setting the values of $\|\widehat{\nabla F}(\boldsymbol{\mu}_\ell)\|$ may be thresholded to determine which returned vectors correspond to hidden basis elements.

In addition, we note that ∇F is an ϵ-approximation to itself for any $\epsilon > 0$. As such, Theorem 6.4 also implies a polynomial time algorithm for recovering each hidden basis element within a preset but arbitrary precision η. In the following Corollary of Theorem 6.4, we characterize the running time of ROBUSTGI-RECOVERY as a function of the precision of the hidden basis estimate.

Corollary 6.6 *Suppose*

- $\sigma \leq \frac{C_0}{\sqrt{d(1+\delta)}}\left[\frac{\beta\gamma}{16\alpha\delta}\right]^{\frac{1}{\gamma}} m^{-\frac{\delta}{\gamma}}$,
- $\eta \leq C_1 4^{-\frac{4+2\delta}{\gamma}} \sigma\left[\frac{\beta\gamma}{\alpha\delta}\right]^{\frac{4\delta+7}{2\gamma}} m^{-\frac{\delta}{\gamma}(2\delta-\gamma+\frac{7}{2})} d^{-\frac{1}{2}-\delta}$,
- $N_1 \geq C_2\lceil \log_{1+2\gamma}(\log_2(\frac{1}{\eta}))\rceil$, *and*
- $N_2 \geq C_3\lceil 4^{\frac{2}{\gamma}}\frac{\sqrt{d}}{\sigma}(\frac{\alpha\delta}{\beta\gamma})^{\frac{\delta+2}{\gamma}} m^{\frac{\delta}{\gamma}(\delta-\gamma+2)}[\frac{1}{\gamma}\log(\frac{\alpha\delta}{\beta\gamma}) + \frac{\delta}{\gamma}\log(m)]\rceil + C_2\lceil \log_{1+2\gamma}(\log_2(\frac{1}{\eta}))\rceil$.

Let $p \in (0,1)$, and suppose that $I \geq C_4 m\lceil \log(m/p)\rceil$. Suppose further that $\widehat{\nabla F}$ is a $C_5\frac{\beta}{\delta}\eta$-approximation to ∇F.

If $\hat{m} \geq m$ and we execute $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{\hat{m}} \leftarrow$ ROBUSTGI-RECOVERY$(\hat{m}, \sigma)$, then with probability at least $1-p$, there exists a permutation π of m such that $\|\pm\boldsymbol{\mu}_i - \mathbf{e}_{\pi(i)}\| \leq \eta$ for all $i \in [m]$.

7. A provably robust algorithm for Independent Component Analysis

In addition to being a very popular technique for blind source separation, Independent Component Analysis (ICA) has been of recent interest in the computer science theory community. Frieze et al. (1996) gave an early analysis of ICA in the setting where the underlying source distributions are continuous uniform distributions. The analysis of this setting was simplified in a cryptographic context by Nguyen and Regev (2009). More recently, there have been a number of works which discuss provable ICA in the presence of additive Gaussian noise (Vempala and Xiao, 2011; Arora et al., 2012; Belkin et al., 2013; Goyal et al., 2014).

In this section, we show how our BEF framework can be used to analyze ICA. In so doing, we provide the first analysis of a general perturbed ICA model. We assume throughout this section that $\mathbf{X} = A\mathbf{S}$ is an ICA model where realizations of $\mathbf{X}$ and $\mathbf{S}$ are both in $\mathbb{R}^d$ (i.e., we consider the fully determined setting in which the number of latent sources equals the ambient dimension of the space). For a random variable Y, we denote its r^{th} moment $m_r(Y) := \mathbb{E}[Y^r]$ and its order r cumulant by $\kappa_r(Y)$. We make the following assumptions:

B1. $\mathbf{S}$ has identity covariance.

B2. For all $i \in [d]$, $|\kappa_4(S_i)| > 0$

B3. For all $i \in [d]$, $m_8(S_i) < \infty$.

B4. A is an orthogonal matrix and $\mathbf{S}$ has $\mathbf{0}$ mean.

Assumption B1 is commonly used within the ICA literature in order to minimize the ambiguities of the ICA model. Assumption B2 is commonly made for cumulant-based ICA algorithms which are used in practice. Assumption B3 will play an important role in our error analysis for cumulant estimation. We include Assumption B4 in order to simplify the exposition and more quickly highlight how our proposed BEF framework applies to ICA. It is common in many ICA algorithms to preprocess the data by placing the data in isotropic position (this is typically referred to as whitening) so that it has $\mathbf{0}$ mean and identity covariance. After this preprocessing step, A is of the desired form. By including the final assumption, we remove the necessity of analyzing the whitening step and propagating the resulting error. Our approach can be generalized to include an error analysis of the whitening step.

We first recall from the discussion on ICA in Section 2.1 that the function $F : S^{d-1} \to \mathbb{R}$ defined by $F(\mathbf{u}) := \kappa_4(\langle \mathbf{u}, \mathbf{X} \rangle)$ is a basis encoding function with associated contrasts $g_i(x) := x^4 \kappa_4(S_i)$ (for $i \in [d]$) and hidden basis elements $\mathbf{e}_i := A_i$ (for $i \in [d]$). We now see that this choice of F is actually a robust BEF.

Lemma 7.1 *Define* $\kappa_{\min} := \min_{i \in [d]} |\kappa_4(S_i)|$ *and* $\kappa_{\max} := \max_{i \in [d]} |\kappa_4(S_i)|$. *Let* $F : S^{d-1} \to \mathbb{R}$ *be defined by* $F(\mathbf{u}) := \kappa_4(\langle \mathbf{u}, \mathbf{X} \rangle)$. *Then,* F *is a* $(2\kappa_{\max}, 2\kappa_{\min}, 1, 1)$*-robust BEF.*

Proof Using the definition of h_i from section 3, we obtain for all $i \in [d]$ that

$$h_i(x) = g_i(\text{sign}(x)\sqrt{|x|}) = x^2 \kappa_4(S_i) \, .$$

Taking derivatives, we see that $h_i''(x) = 2\kappa_4(S_i)$, and hence that $2\kappa_{\min} \leq |h_i''(x)| \leq 2\kappa_{\max}$. Recalling Definition 6.1 with $(\alpha, \beta, \gamma, \delta) = (2\kappa_{\max}, 2\kappa_{\min}, 1, 1)$ completes the proof. ■

We do not have direct access to F. Instead, we will estimate F from samples. We note that for any $\mathbf{u} \in S^{d-1}$, $\mathrm{var}(\langle \mathbf{u}, \mathbf{X} \rangle) = 1$. For a 0-mean random variable Y with unit variance, the fourth cumulant is known to take on a very simple form: $\kappa_4(Y) = m_4(Y) - 3$. This provides a natural sample estimate for the fourth cumulant in our setting. Given samples $y(1), y(2), \ldots, y(N)$ of a random variable Y, we will estimate $\kappa_4(Y)$ by $\hat{\kappa}_4(y(i)) := \frac{1}{N}\sum_{i=1}^{N} y(i)^4 - 3$.

Let $p_{\mathbf{Y}}$ denote the probability density function of a random vector $\mathbf{Y}$. In order to handle a perturbation away from the ICA model, we will consider metrics of the form $\mu_k(\mathbf{X}, \mathbf{Y}) := \int_{t \in \mathbb{R}^d} \|\mathbf{t}\|^8 |p_{\mathbf{X}}(\mathbf{t}) - p_{\mathbf{Y}}(\mathbf{t})| \mathrm{d}\mathbf{t}$ on the space of probability densities. We will assume sample access to a random variable $\hat{\mathbf{X}}$ such that $\mu_8(\mathbf{X}, \hat{\mathbf{X}})$ is sufficiently small (to be quantified later). Given samples $\hat{\mathbf{x}}(1), \hat{\mathbf{x}}(2), \ldots, \hat{\mathbf{x}}(N)$ i.i.d. from $\hat{\mathbf{X}}$, we estimate F by the function $\hat{F}(\mathbf{u}) := \frac{1}{N}\sum_{i=1}^{N} \langle \mathbf{u}, \hat{\mathbf{x}}(i) \rangle^4 - 3$. The gradient of $\hat{F}$ is easily computed as $\nabla \hat{F}(\mathbf{u}) = \frac{4}{N}\sum_{i=1}^{N} \langle \mathbf{u}, \hat{\mathbf{x}}(i) \rangle^3 \hat{\mathbf{x}}(i)$ and acts as an estimate of ∇F. As such, we have all of the information required to implement ROBUSTGI-RECOVERY using $\widehat{\nabla F} := \nabla \hat{F}$.

We now provide uniform bounds on the estimate errors for ∇F under this model.

Lemma 7.2 *Fix $\delta > 0$ and $\eta > 0$. Let $M_8 := \max_{i \in [d]} m_8(S_i)$. Let $\hat{\mathbf{X}}$ be a random vector in $\mathbb{R}^d$ such that $\mu_8(\mathbf{X}, \hat{\mathbf{X}})$ is finite. Suppose that $\hat{\mathbf{x}}(1), \hat{\mathbf{x}}(2), \ldots, \hat{\mathbf{x}}(N)$ are drawn i.i.d. from $\hat{\mathbf{X}}$ with $N \geq \frac{d^4[M_8 + \mu_8(\mathbf{X}, \hat{\mathbf{X}})]}{\eta^2 \delta}$. If $\hat{F}(\mathbf{u}) := \frac{1}{N}\sum_{i=1}^{d} \langle \mathbf{u}, \hat{\mathbf{x}}(i) \rangle^4 - 3$ and $F(\mathbf{u}) := \kappa_4(\langle \mathbf{u}, \mathbf{X} \rangle)$, then with probability $1 - \delta$ the following bounds hold for all $\mathbf{u} \in S^{d-1}$: (1) $|F(\mathbf{u}) - \hat{F}(\mathbf{u})| \leq (\eta + \mu_4(\mathbf{X}, \hat{\mathbf{X}}))d^2$ and (2) $\|\nabla F(\mathbf{u}) - \nabla \hat{F}(\mathbf{u})\| \leq 4(\eta + \mu_4(\mathbf{X}, \hat{\mathbf{X}}))d^2$.*

Proof In this proof, we proceed with the convention that we are indexing with respect to the hidden basis in which $\mathbf{e}_i := A_i$ for all $i \in [d]$. In particular, this implies $X_i = \langle A_i, \mathbf{X} \rangle = S_i$.

We use multi-index notation to compress our discussion as follows: $J \in [d]^k$ will denote a multi-index $J = (j_1, j_2, \ldots, j_k)$ such that each $j_\ell \in [d]$. For a vector $\mathbf{v}$, v_J denotes the product $\prod_{\ell=1}^{k} v_{j_\ell}$. Our objective function $\hat{F}(\mathbf{u})$ may be expanded as a polynomial of the u_js as follows:

$$\hat{F}(\mathbf{u}) = \frac{1}{N}\sum_{i=1}^{N} \langle \mathbf{u}, \hat{\mathbf{x}}(i) \rangle^4 - 3 = \frac{1}{N}\sum_{i=1}^{N} \sum_{J \in [d]^4} u_J \hat{x}_J(i) - 3 = \sum_{J \in [d]^4} u_J \left[\frac{1}{N}\sum_{i=1}^{N} \hat{x}_J(i) \right] - 3 \, .$$

By a similar argument, it can be shown that $F(\mathbf{u}) = \sum_{J \in [d]^4} u_J \mathbb{E}[X_J] - 3$. We obtain the error bound $|\hat{F}(\mathbf{u}) - F(\mathbf{u})| \leq \sum_{J \in [d]^4} u_J | \frac{1}{N}\sum_{i=1}^{d} \hat{x}_J(i) - \mathbb{E}[X_J]|$. Similarly, we can bound the error estimate for $\nabla F(\mathbf{u})$:

$$\|\nabla \hat{F}(\mathbf{u}) - \nabla F(\mathbf{u})\| = 4 \sum_{J \in [d]^3} u_J \left\| \frac{1}{N}\sum_{i=1}^{N} \hat{x}_J(i) \hat{\mathbf{x}}(i) - \mathbb{E}[X_J \mathbf{X}] \right\|.$$

With $J \in [d]^4$, we define $\varepsilon_J := \frac{1}{N}\sum_{i=1}^{N} \hat{x}_J(i) - \mathbb{E}[X_J]$ and $\varepsilon_{\max} := \max_{J \in [d]^4} |\varepsilon_J|$. Using that each $\mathbf{u}$ is a unit vector, we see that $|\sum_{J \in [d]^k} u_J| \leq \|\mathbf{u}\|_1^k \leq d^{k/2}$. Using the norm inequalities that for vector $\mathbf{v} \in \mathbb{R}^d$ and matrix $A \in \mathbb{R}^{d \times d}$, $\|\mathbf{v}\| \leq \max_{i \in [d]} |v_i| \sqrt{d}$ and $\|A\| \leq \max_{(i,j) \in [d]^2} |a_{ij}| d$, we are able to obtain the following bounds for all $\mathbf{u} \in S^{d-1}$: $|\hat{F}(\mathbf{u}) - F(\mathbf{u})| \leq d^2 \varepsilon_{\max}$ and $\|\nabla \hat{F}(\mathbf{u}) - \nabla F(\mathbf{u})\| \leq 4d^2 \varepsilon_{\max}$. All that remains is to bound $\varepsilon_{\max}$. To do so, we will bound each ε_J using Chebyshev's inequality.

For each $J \in [d]^4$, we obtain under the sampling process that

$$\begin{aligned}\operatorname{var}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{x}_J(i)\right) &= \frac{1}{N^2}\operatorname{var}\left(\sum_{i=1}^{N}\hat{x}_J(i)\right) = \frac{1}{N}\operatorname{var}(\hat{X}_J) \le \frac{1}{N}\mathbb{E}[(\hat{X}_J)^2] \\ &\le \frac{1}{N}\mathbb{E}[\hat{X}_{j_1}^4\hat{X}_{j_2}^4]^{\frac{1}{2}}\mathbb{E}[\hat{X}_{j_3}^4\hat{X}_{j_4}^4]^{\frac{1}{2}} \le \frac{1}{N}\left(\prod_{\ell=1}^{4}\mathbb{E}[\hat{X}_{j_\ell}^8]\right)^{\frac{1}{4}} \le \frac{1}{N}\max_{\ell\in[d]}\mathbb{E}[\hat{X}_\ell^8]\end{aligned}$$

where the first equality uses that variance is order-2 homogenous, the second equality uses independence, the first inequality follows from the formula $\operatorname{var}(\hat{X}_J) = \mathbb{E}[(\hat{X}_J)^2] - \mathbb{E}[\hat{X}_J]^2$, and the second and third inequalities use the Cauchy-Schwartz inequality. We bound $\max_{\ell\in[d]}\mathbb{E}[\hat{X}_\ell^8]$ as:

$$\mathbb{E}[\hat{X}_\ell^8] = \int_{\mathbf{t}\in\mathbb{R}^d} t_\ell^8 p_{\hat{\mathbf{X}}}(\mathbf{t})\mathrm{d}\mathbf{t} = \int_{\mathbf{t}\in\mathbb{R}^d} t_\ell^8 p_{\mathbf{X}}(\mathbf{t})\mathrm{d}\mathbf{t} + \int_{\mathbf{t}\in\mathbb{R}^d} t_\ell^8 (p_{\hat{\mathbf{X}}}(\mathbf{t}) - p_{\mathbf{X}}(\mathbf{t}))\mathrm{d}\mathbf{t} \le M_8 + \mu_8(\mathbf{X}, \hat{\mathbf{X}}) \,.$$

Thus, $\operatorname{var}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{x}_J(i)\right) \le \frac{1}{N}(M_8 + \mu_8(\mathbf{X}, \hat{\mathbf{X}}))$.

Chebyshev's inequality states that for any random variable Y and any $k > 0$, $\mathbb{P}[|Y - \mathbb{E}[Y]| \ge k\sqrt{\operatorname{var}(Y)}] \le \frac{1}{k^2}$. We fix any $J \in [d]^4$, choose $Y = \frac{1}{N}\sum_{i=1}^{N}\hat{x}_J(i)$, and choose $k = \frac{d^2}{\sqrt{\delta}}$. We obtain that with probability at least $1 - \delta/d^4$,

$$\left|\frac{1}{N}\sum_{i=1}^{N}\hat{x}_J(i) - \mathbb{E}[\hat{X}_J]\right| < \frac{d^2}{\sqrt{\delta}}\sqrt{\frac{1}{N}(M_8 + \mu_8(\mathbf{X}, \hat{\mathbf{X}}))} \le \eta \tag{14}$$

by our given bound on N. Taking a union bound, then with probability at least $1 - \delta$, the bound in equation (14) holds for all $J \in [d]^4$.

We then obtain the following bound on each ε_J for each $J \in [d]^4$ (with probability at least $1 - \delta$):

$$\begin{aligned}|\varepsilon_J| &= \left|\frac{1}{N}\sum_{i=1}^{N}\hat{x}_J(i) - \mathbb{E}[X_J]\right| \le \eta + |\mathbb{E}[\hat{X}_J] - \mathbb{E}[X_J]| \\ &= \eta + \left|\int_{\mathbf{t}\in\mathbb{R}^4} t_J[p_{\hat{\mathbf{X}}}(\mathbf{t}) - p_{\mathbf{X}}(\mathbf{t})]\mathrm{d}\mathbf{t}\right| \le \eta + \mu_4(\hat{\mathbf{X}}, \mathbf{X}) \,.\end{aligned}$$

To obtain the result, we use $\varepsilon_{\max} \le \eta + \mu_4(\hat{\mathbf{X}}, \mathbf{X})$ in our previously derived uniform bounds over all $\mathbf{u} \in S^{d-1}$ of $|\hat{F}(\mathbf{u}) - F(\mathbf{u})| \le d^2\varepsilon_{\max}$ and $\|\nabla\hat{F}(\mathbf{u}) - \nabla F(\mathbf{u})\| \le 4d^2\varepsilon_{\max}$. ■

We now state our result for ICA. We assume $\mathbf{X} = A\mathbf{S}$ is an ICA model satisfying Assumptions B1–B4 with associated constants $\kappa_{\min} := \min_{i\in[d]}|\kappa_4(S_i)|$, $\kappa_{\max} := \max_{i\in[d]}|\kappa_4(S_i)|$, and $M_8 := \max_{i\in[d]} m_8(S_i)$. We assume $\hat{\mathbf{X}}$ is a perturbed ICA model, and we approximate the BEF $F(\langle\mathbf{u}, \mathbf{X}\rangle) := \kappa_4(\langle\mathbf{u}, \mathbf{X}\rangle)$ from an i.i.d. sample $\hat{\mathbf{x}}(1), \ldots, \hat{\mathbf{x}}(\mathcal{N})$ of $\hat{\mathbf{X}}$. That is, we define $\hat{F}(\mathbf{u}) := \frac{1}{\mathcal{N}}\sum_{i=1}^{d}\langle\mathbf{u}, \hat{\mathbf{x}}(i)\rangle^4 - 3$ and compute its gradient as $\nabla\hat{F}(\mathbf{u}) := \frac{4}{\mathcal{N}}\sum_{i=1}^{d}\langle\mathbf{u}, \hat{\mathbf{x}}(i)\rangle^3\hat{\mathbf{x}}(i)$. We further assume that we are working in a computation model which can perform the following operations in $O(d)$ time: Inner products in $\mathbb{R}^d$, scalar operations including basic arithmetic operations, trigonometric functions, square roots, and branches on conditionals. In the following, $C_1, C_2, \ldots$ are positive universal constants.

Theorem 7.3 *Fix $\delta > 0$ and $\varepsilon > 0$. Suppose $\sigma \leq \frac{C_0}{d^2}\frac{\kappa_{\min}}{\kappa_{\max}}$, $\varepsilon \leq C_1\sigma\left(\frac{\kappa_{\min}}{\kappa_{\max}}\right)^{9/2}d^{-6}$, $\mu_4(\hat{\mathbf{X}}, \mathbf{X}) \leq C_2\frac{\kappa_{\min}}{d^2}\varepsilon$, and $\mathcal{N} \geq \frac{C_3 d^8[M_8+\mu_8(\hat{\mathbf{X}},\mathbf{X})]}{\kappa_{\min}^2\varepsilon^2\delta}$. Let $\hat{A}_1, \ldots, \hat{A}_d \leftarrow$ ROBUSTGI-RECOVERY(d, σ), where* FINDBASISELEMENT *is implemented with $I \geq C_6 d\log(d/\delta)$, $N_1 \geq C_4\lceil\log_2(\log_2(\frac{1}{\varepsilon}))\rceil$, and $N_2 \geq C_5\lceil\frac{d^{2.5}}{\sigma}\left(\frac{\kappa_{\max}}{\kappa_{\min}}\right)^3\log(d\cdot\frac{\kappa_{\max}}{\kappa_{\min}})\rceil + \lceil\log_2(\log_2(\frac{1}{\varepsilon}))\rceil$. Then, with probability at least $1-\delta$, there exists a permutation π of $[d]$ and sign values $s_i \in \{\pm 1\}$ such that $\|\hat{A}_i - s_i A_{\pi(i)}\| \leq \varepsilon$ for all $i \in [d]$.* ROBUSTGI-RECOVERY *recovers such $\hat{A}_i$s in $C_7\mathcal{N}[d^4 + d^2N_1 + d^2IN_2]$ time.*

Proof By Lemma 7.2 with the choice of $\eta = O(\frac{\kappa_{\min}}{d^2}\varepsilon)$, we obtain that with probablity at least $1 - \frac{\delta}{2}$,

$$\|\nabla F(\mathbf{u}) - \nabla\hat{F}(\mathbf{u})\| \leq 4(\eta + \mu_4(\mathbf{X}, \hat{\mathbf{X}}))d^2 \leq O\Big(\frac{\kappa_{\min}}{d^2}\varepsilon + \frac{\kappa_{\min}}{d^2}\varepsilon\Big)d^2 = O(\kappa_{\min}\varepsilon),$$

for all $\mathbf{u} \in S^{d-1}$. In particular, $\hat{F}$ is an $O(\kappa_{\min}\varepsilon)$-approximation to F.

We recall from Lemma 7.1 that F is an $(2\kappa_{\max}, 2\kappa_{\min}, 1, 1)$-robust BEF. As such, we may apply Corollary 6.6 to obtain that ROBUSTGI-RECOVERY returns vectors $\hat{A}_1, \ldots, \hat{A}_d$ of the desired form. Finally, we note that within our computational model for this theorem, computations of $\nabla\hat{F}(\mathbf{u})$ take $O(\mathcal{N}d)$ time. Thus, applying Theorem 6.5 with $\hat{m} = d$ yields the claimed time bound. ■

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. 1350870, 1422830, 15507576, and 1117707.

References

Anima Anandkumar, Dean P. Foster, Daniel Hsu, Sham Kakade, and Yi-Kai Liu. A spectral algorithm for latent Dirichlet allocation. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems.*, pages 926–934, 2012a. URL http://books.nips.cc/papers/files/nips25/NIPS2012_0441.pdf.

Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559, 2012b. URL http://arxiv.org/abs/1210.7559.

Animashree Anandkumar, Daniel Hsu, and Sham M. Kakade. A method of moments for mixture models and hidden Markov models. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, volume 23 of *JMLR Proceedings*, pages 33.1–33.34. JMLR.org, 2012c. URL http://www.jmlr.org/proceedings/papers/v23/anandkumar12/anandkumar12.pdf.

Animashree Anandkumar, Rong Ge, and Majid Janzamin. Learning overcomplete latent variable models through tensor methods. In *Conference on Learning Theory (COLT)*, 2015.

Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ICA with unknown Gaussian noise, with implications for Gaussian mixtures and autoencoders. In *NIPS*, pages 2384–2392, 2012.

Francis R. Bach and Michael I. Jordan. Learning spectral clustering, with application to speech separation. *Journal of Machine Learning Research*, 7:1963–2001, 2006.

Marian Stewart Bartlett, Javier R Movellan, and Terrence J Sejnowski. Face recognition by independent component analysis. *Neural Networks, IEEE Transactions on*, 13(6):1450–1464, 2002.

Mikhail Belkin, Luis Rademacher, and James Voss. Blind signal separation in the presence of Gaussian noise. In *JMLR W&CP*, volume 30: COLT, pages 270–287, 2013.

Mikhail Belkin, Luis Rademacher, and James R. Voss. The hidden convexity of spectral clustering. *CoRR*, abs/1403.0667, 2014. URL http://arxiv.org/abs/1403.0667.

Mikhail Belkin, Luis Rademacher, and James Voss. Basis Learning as an Algorithmic Primitive. *CoRR*, abs/1411.1420v4, 2016. URL http://arxiv.org/abs/1411.1420v4.

A.J. Bell and T.J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Pierre Comon and Christian Jutten, editors. *Handbook of Blind Source Separation*. Academic Press, 2010.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Lieven De Lathauwer, Pierre Comon, Bart De Moor, and Joos Vandewalle. Higher-order power method. *NOLTA Conference*, 1995.

Nathalie Delfosse and Philippe Loubaton. Adaptive blind separation of independent sources: A deflation approach. *Signal processing*, 45(1):59–83, 1995.

Manfredo Perdigao do Carmo Valero. *Riemannian geometry*. 1992.

Alan M. Frieze, Mark Jerrum, and Ravi Kannan. Learning linear transformations. In *FOCS*, pages 359–368. IEEE Computer Society, 1996.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. *arXiv preprint arXiv:1503.02101*, 2015.

Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier PCA and robust tensor decomposition. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 584–593, 2014. doi: 10.1145/2591796.2591875. URL http://doi.acm.org/10.1145/2591796.2591875.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms: Part 1: Fundamentals*, volume 1. Springer, 1996.

Daniel Hsu and Sham M Kakade. Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science (ITCS)*, pages 11–20. ACM, 2013.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. John Wiley & Sons, 2001.

Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

Aapo Hyvärinen and Erkki Oja. Independent component analysis by general nonlinear Hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

John Francis Kenney and Ernest Sydney Keeping. *Mathematics of Statistics, part 2*. van Nostrand, 1962.

Lek-Heng Lim. Singular values and eigenvalues of tensors: a variational approach. *arXiv preprint math/0607648*, 2006.

David G Luenberger and Yinyu Ye. *Linear and nonlinear programming*, volume 116. Springer, 2008.

Albert CJ Luo. *Regularity and complexity in dynamical systems*. Springer, 2012.

Shoji Makino, Te-Won Lee, and Hiroshi Sawada. *Blind speech separation*. Springer, 2007.

Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

Phong Q. Nguyen and Oded Regev. Learning a parallelepiped: Cryptanalysis of GGH and NTRU signatures. *J. Cryptology*, 22(2):139–160, 2009.

Liqun Qi. Eigenvalues of a real supersymmetric tensor. *Journal of Symbolic Computation*, 40(6): 1302–1324, 2005.

Walter Rudin. *Real and complex analysis (3rd)*. New York: McGraw-Hill Inc, 1986.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Santosh S. Vempala and Ying Xiao. Structure from local optima: Learning subspace juntas via higher order PCA. *CoRR*, abs/1108.3329, 2011. URL http://arxiv.org/abs/1108.3329.

Ricardo Vigário, Jaakko Sarela, V Jousmiki, Matti Hamalainen, and Erkki Oja. Independent component approach to the analysis of EEG and MEG recordings. *Biomedical Engineering, IEEE Transactions on*, 47(5):589–593, 2000.

James R Voss, Luis Rademacher, and Mikhail Belkin. Fast algorithms for gaussian noise invariant independent component analysis. In *Advances in Neural Information Processing Systems*, pages 2544–2552, 2013.

Marcus Weber, Wasinee Rungsarityotin, and Alexander Schliep. *Perron cluster analysis and its connection to graph partitioning for noisy data*. Konrad-Zuse-Zentrum für Informationstechnik Berlin, 2004.

Tianwen Wei. A study of the fixed points and spurious solutions of the deflation-based fastica algorithm. *Neural Computing and Applications*, pages 1–12, 2015. ISSN 0941-0643. doi: 10.1007/s00521-015-2033-6. URL http://dx.doi.org/10.1007/s00521-015-2033-6.

Stella X. Yu and Jianbo Shi. Multiclass spectral clustering. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on (ICCV)*, pages 313–319. IEEE Computer Society, 2003. ISBN 0-7695-1950-4.

Vicente Zarzoso and Pierre Comon. Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size. *Neural Networks, IEEE Transactions on*, 21(2):248–261, 2010.

Appendix A. Miscellaneous Facts

In this section, we collect a number useful results and statements which are used in the proofs of the theorems of this paper.

The following is a special case of Lemma 7.25 of Rudin (1986).

Theorem A.1 *Let $U \subset \mathbb{R}^k$ be an open set. Suppose $f : V \to \mathbb{R}^k$ is differentiable on its entire domain. If $E \subset V$ has Lebesgue measure 0, then $f(E)$ has Lebesgue measure 0.*

A.1. Locally Stable Manifold for Fixed Points of Discrete Dynamical Systems

The eigenspaces of a linear operator $T : \mathbb{R}^k \to \mathbb{R}^k$ may be decomposed into several subspaces. We let $(\lambda_1, \mathbf{v}_1), \ldots, (\lambda_k, \mathbf{v}_k)$ denote the eigenvalue-eigenvector pairs for T, and define the subspaces:

$$\begin{aligned}\mathcal{E}^S(T) &:= \operatorname{span}\{\mathbf{v}_i \mid |\lambda_i| \leq 1\} \\ \mathcal{E}^U(T) &:= \operatorname{span}\{\mathbf{v}_i \mid |\lambda_i| \geq 1\} \\ \mathcal{E}^C(T) &:= \operatorname{span}\{\mathbf{v}_i \mid |\lambda_i| = 1\}\,.\end{aligned}$$

It turns out that for a discrete dynamical system f with a fixed point $\mathbf{x}^*$, the dimensionality of $\mathcal{E}^S(\mathrm{D}f(\mathbf{x}^*))$ is locally related to the dimensionality of the space on which convergence to $\mathbf{x}^*$ is achieved. More precisely, letting $\{\mathbf{x}(n)\}_{n=0}^{\infty}$ denote arbitrary sequences defined recursively by $\mathbf{x}(n) = f(\mathbf{x}(n-1))$, there is the following notion of a locally stable manifold around $\mathbf{x}^*$.

Definition A.2 *Within a neighborhood U of $\mathbf{x}^*$, the manifold*

$$\mathcal{L}_{loc}(\mathbf{x}^*) := \{\mathbf{x}(0) \in U \mid \lim_{k\to\infty} \mathbf{x}(k) = \mathbf{x}^*, \mathbf{x}(k) \in U \; \forall k \in \mathbb{N}\}$$

is called the local stable manifold.

The following result is a special case of Theorem 2.2 of Luo (2012).

Theorem A.3 *Let $f : \mathcal{X} \to \mathcal{X}$ be a discrete dynamical system with a fixed point $\mathbf{x}^*$ such that (i) f is continuously differentiable on a neighborhood of $\mathbf{x}^*$ and (ii) the eigenspace of $\mathrm{D}f(\mathbf{x}^*)$ can be decomposed as $\mathcal{E}^S(\mathrm{D}f(\mathbf{x}^*)) \oplus \mathcal{E}^U(\mathrm{D}f(\mathbf{x}^*))$. Then, $\dim(\mathcal{L}_{loc}) = \dim(\mathcal{E}^S(\mathrm{D}f(\mathbf{x}^*)))$. Further, there exists $\delta > 0$ such that for all for all $\mathbf{x}(0) \notin L_{loc}$, there exists $N \in \mathbb{N}$ such that $\|\mathbf{x}(N) - \mathbf{x}^*\| > \delta$.*