Global Optimality Bounds for ICA Algorithms

Nicolo Colombo[1]*, Johan Thunberg[1] and Jorge Goncalves[1]

Abstract— **Independent Component Analysis is a popular statistical method for separating a multivariate signal into additive components. It has been shown that the signal separation problem can be reduced to the joint diagonalization of the matrix slices of some higher-order cumulants of the signal. In this approach, the unknown mixing matrix can be computed directly from the obtained joint diagonalizer. Various iterative algorithms for solving the non-convex joint diagonalization problem exist, but they usually lack global optimality guarantees. In this paper, we introduce a procedure for computing an optimality gap for local optimal solutions. The optimality gap is then used to obtain an empirical error bound for the estimated mixing matrix. Finally, a class of simultaneous matrix decomposition problems that admit such relaxation procedure is identified.**

I. INTRODUCTION

A large class of algorithms for Independent Components Analysis (ICA) is based on the simultaneous diagonalisation of several symmetric matrices, obtained from the higher order cumulants of the empirical data [Cardoso, 1999]. In the orthogonal approach, where the signal is spatially white, the matrix slices of the fourth-order cumulant tensor can be diagonalised by an orthogonal mixing matrix. The ICA is then equivalent to a simultaneous diagonalisation problem, over the set of orthogonal matrices O(n). In general, for non spatially white signals, the ICA is either cast into a non-orthogonal simultaneous digaonalization problem [Yeredor, 2002] [Afsari and Krishnaprasad, 2004] [Ziehe et al., 2004], or reduced to the orthogonal case via the empirical whitening of the signal. See [Souloumiac, 2009] for a comparison of these two approaches. In the presence of noise, the problem of finding the approximate joint diagonalizer of a set of matrices is usually reformulated as a minimisation problem over a suitable set matrix. In the orthogonal case, the existence of a global optimal solution for this minimisation problem is guaranteed by the compactness of the $O(n)$. Due to the non-convexity of the associated objective function, finding a global optimizer is in general hard and should be approximately solved in an iterative fashion. Two popular classes of iterative methods have been proposed in the literature: Jacobi-like algorithms based on Givens rotation updates [Bunse-Gerstner et al., 1993] [Cardoso and Souloumiac, 1996] and matrix-manifold approaches where the descent steps are computed from a gradient flow equation [Manton, 2002] [Yamada and Ezaki, 2003] [Afsari and Krishnaprasad, 2004]. Other approaches have also been considered, see for example [Wax and Sheinvald, 1997] [Van der Veen, 2001] [Afsari, 2006]. A drawback of all these iterative methods is the lack of global optimality guarantees and some, often obscure, dependence on the initialisation. However, most of these iterative algorithms have been proven to achieve local convergence. The 'distance' between the obtained local optimal solution and the global optimal solution depends on the cost function landscape which in general is not straightforward to analyse. The optimality gap of such solutions, *i.e.* the distance between a particular local optimum and the global minimum of the objective function, can depend on the initialisation but also on the unknown noise level of the data. This paper provides a procedure based on a spectral relaxation of the non-convex objective function, to compute such an optimality gap.

A series of numerical experiment show that the obtained optimality gap can be very small in relation to the global minimum of the relaxed objective function. Two main conclusions can be drawn: i) for the class of simultaneous matrix decomposition related to ICA, the spectral relaxation can be used as a good approximation of the true non-convex objective function and ii) in most of the cases, the solutions provided by the iterative algorithms tested here are practically equivalent to the theoretical global optimal solution . In the case of the orthogonal joint diagonalization, we show how the optimality gap can be used to characterize a given sub-optimal solution. In particular, it is possible to estimate the distance between the sub-optimal solution and the closest exact diagonalizer of the exactly joint diagonalizable ground-truth matrices. Under the reasonable assumption that an estimation of the noise level is available such a bound involves only 'empirical' quantities that can computed from the input matrices and the optimality gap. More practically, in the ICA framework these bounds can be used to estimate the error in the recovery of the ICA mixing matrices by using only the observed signal and a prior knowledge of the noise level of the signal. Finally, we show that the method is quite general and can be used to compute the optimality gap in a broader class of problem. Under certain rank-dimension conditions, this class includes a large number of simultaneous matrix decomposition problems where the optimization problem is harder that in the ICA setting and subject to stronger initialisation issues.

This paper proceeds as follows: Section II contains a brief review of the link between ICA and simultaneous diagonalisation; Section III is dedicated to the spectral relaxation in the case of the orthogonal simultaneous diagonalisation; the empirical bound and a sketch on how this is obtained from the optimality gap is given in Section IV while all

[1] Luxembourg Centre for System Biomedicine

* nicolo.colombo@uni.lu

This work was supported by Fonds National de la Recherche Luxemourg FNR8864515, I1R-SCG-PFN-14CRCL.

the details about the derivation of the empirical bound are provided in Section V; finally, Section VII includes possible generalisations and Section VII a series of numerical simulations and experiments on synthetic data.

II. ICA and Simultaneous Diagonalisation

Consider the ICA model

$$x = U_0 s \tag{1}$$

where x is the observed n-dimensional signal, U_0 is the mixing matrix and the components of s are the n independent sources. In the orthogonal approach, x and s are assumed to be unit norm spatially white vectors and U_0 an orthogonal matrix. The general case where x is not spatially white can be reconnected to the orthogonal case via an opportune pre whitening of the signal [Souloumiac, 2009]. It can be shown that the matrix slices of the fourth-order cumulant tensor of x can be diagonalised by the orthogonal matrix U_0 [Cardoso, 1999]. Given four random variables u_1, u_2, u_3, u_4, their fourth order cumulant tensor is defined by

$$Cum(u_1, u_2, u_3, u_4) = \tag{2}$$
$$E(\bar{u}_1\bar{u}_2\bar{u}_3\bar{u}_4) - \bar{R}_{12}\bar{R}_{34} - \bar{R}_{13}\bar{R}_{24} - \bar{R}_{14}\bar{R}_{23}$$

where $\bar{u} = u - E(u)$, with $E(u)$ being the expectation value of u, and $\bar{R}_{ij} = E(\bar{u}_i\bar{u}_j)$. The (assumed) independence of the sources implies $Cum(s_i, s_j, s_k, s_l) = k(s_i)\delta(i,j,k,l)$, where the kurtosis is defined as $k(u) = Cum(u,u,u,u)$, and hence

$$Cum(x_i, x_j, x_k, x_l) \tag{3}$$
$$= \sum_{a=1}^{n} k(s_a)[U_0]_{ia}[U_0]_{ja}[U_0]_{ka}[U_0]_{la}$$

The matrix slices of these fourth-order tensors $[M_r]_{ij} = \sum_{kl}[\Theta_r]_{kl}C^{(4)}_{ijkl}$, where $C^{(4)}_{ijkl} = Cum(x_i, x_j, x_k, x_l)$, Θ_r, for $r = 1, \dots, R$, are $n \times n$ random matrices, can be simultaneously diagonalised by the matrix U_0. For any $r = 1, \dots, R$, the matrix slice M_r is a symmetric matrices of the form

$$M_r = U_0 \mathrm{diag}([\Lambda_{r1}, \dots, \Lambda_{rn}])U_0^T \qquad r = 1, \dots, R \tag{4}$$

where $\mathrm{diag}(a)$, for $a = [a_1, \dots, a_n]$, is a diagonal matrix with entries $a_1, \dots a_n$ and $\Lambda_{ra} = k(s_a)[U_0^T\Theta_r U_0]_{aa}$. In presence of white noise, ε, the ICA model (1) is $x = U_0 s + \varepsilon$ and the empirical cumulant matrices are no longer simultaneously diagonalisable. However, an estimation of the mixing matrix U_0 can be obtained by finding their approximate joint diagonalizer. The empirical matrix slices can be written as

$$\hat{M}_r \quad = \quad M_r + \sigma W_r \qquad r = 1, \dots, R \tag{5}$$

where M_r are the joint diagonalizable matrices defined in (4), σ is a positive scalar and W_r are symmetric matrices whose explicit form in terms of U_0, s and ε can be obtained from the definition of the fourth-order cumulant (3). With no loss of generality we can assume $\|W_r\| \leq 1$. The approximate orthogonal diagonalizer of the empirical matrices $\hat{M}_r$ can be found by solving the following non-convex optimization problem[1]

$$\begin{aligned} \text{minimize} \qquad & \mathcal{L}(U) = \sum_{r=1}^{R} \|\mathrm{off}(U^T\hat{M}_r U)\|^2 \\ \text{s.t.} \qquad & U^T U = 1 \end{aligned} \tag{6}$$

where $[\mathrm{off}(A)]_{ij} = 0$ if $i = j$ and $[\mathrm{off}(A)]_{ij} = A_{ij}$ if $i \neq j$ and $\|\cdot\|$ is the Frobenius norm, *i.e.* $\|A\| = \sqrt{\mathrm{Tr}(A^T A)}$.

III. Spectral Relaxation and Optimality Gap

Due to the non-convexity of $\mathcal{L}(U)$, solving (6) is not straightforward. However, $\mathcal{L}$ admits a convex relaxation that can be solved globally by means of a spectral decomposition of a positive semi-definite matrix computed from the data. The global optimum of the relaxed objective bounds from below the global optimum of $\mathcal{L}$. By letting U_* be a solution obtained from an iterative algorithm, a bound on the optimality gap is obtained as the distance between $\mathcal{L}(U_*)$ and global optimum of the relaxed objective. By construction it bounds from above the distance between the obtained local minimum and the global minimum of $\mathcal{L}(U)$.

A. Spectral Relaxation

Consider the function $L : O(n^2) \to \mathbf{R}$ defined by

$$L(V) \quad = \quad \mathrm{Tr}\left(\mathrm{Off} V\hat{m}\hat{m}^T V^T\right) \tag{7}$$

where the matrix Off is defined by Off $\mathrm{vec}(A) = \mathrm{vec}\ \mathrm{off}(A)$ and $\hat{m} = [\mathrm{vec}(\hat{M}_1), \dots, \mathrm{vec}(\hat{M}_R)]^T$. It easy to verify that $\mathcal{L}(U) = L(U \otimes U)$, where $\otimes$ is the Kronecker product. A 'relaxed' optimization problem is

$$\begin{aligned} \text{minimize} \qquad & L(V) = \mathrm{Tr}\left(\mathrm{Off} V\hat{m}\hat{m}^T V^T\right) \\ \text{s.t.} \qquad & V^T V = 1 \end{aligned} \tag{8}$$

where the relaxation is to drop the constraint $V = U \otimes U$. The objective function of the relaxed problem is quadratic in V and a global optimal solution can be found by computing the spectral decomposition of the positive semi-definite matrix $\hat{m}\hat{m}^T$. Since Off is diagonal with only zeros and ones on the diagonal, the solution to (8) is given by a column wise reordering of the matrix consisting of the singular vectors of $\hat{m}\hat{m}^T$. More precisely

$$V_* = Q\Gamma\bar{V}^T \qquad \hat{m}\hat{m}^T = \bar{V}\Sigma\bar{V}^T \tag{9}$$

where Σ is a diagonal matrix, Q is an orthogonal matrix such that Off$Q$ = Off and Γ is a permutation matrix that swaps the $n^2 - n$ smallest singular values of $\hat{m}\hat{m}^T$ to the positions defined by the non-zero entries of Off. In other words, Q is an orthogonal matrix that leave the subspace of the smallest singular values of mm^T invariant, up to permutations.

[1] This is a rather arbitrary but common choice and other more refined objective functions may be considered.

B. Optimality Gap

Due to the compactness of $O(n)$, the optimization proptimizoblem (6) has a globally optimal solution. Let $\bar{U}_{\rm opt}$ be such a globally optimal solution to (6) and U_* the solution computed by an iterative algorithm. Then

$$L(V_*) \leq L(\bar{U}_{\rm opt} \otimes \bar{U}_{\rm opt}) = \mathcal{L}(\bar{U}_{\rm opt}) \leq \mathcal{L}(U_*) \qquad (10)$$

In particular, for any suboptimal U_* we can bound the distance from the global optimum by computing the relaxed solution V_* and use

$$\mathcal{L}(U_*) - \mathcal{L}(\bar{U}_{\rm opt}) \leq \delta_{\rm opt}, \quad \delta_{\rm opt} = \mathcal{L}(U_*) - L(V_*) \quad (11)$$

IV. Error Bounds

In this section we provide bounds for the error in the estimation of the mixing matrix U_0 by a function of the optimality gap and other empirical quantities. The key idea is to combine a characterization of the global optimum analogous to the perturbation analysis of [Cardoso, 1994] with the optimality gap (11). In particular, the optimality gap is used to establish an upper bound on the distance between the obtained sub-optimal solution U_* and a provably good optimal solution $U_{\rm opt}$, where provably good means that it is possible to prove its closeness to a ground truth mixing matrix U_0. Both the analysis of [Cardoso, 1994] and all bounds derived here are linear approximations, that hold up to second order terms in the noise parameter. Our main result is the following theorem.

Theorem 1. *Let $\hat{M}_r$, for $r = 1, \dots, R$, be the nearly joint diagonalizable matrices defined in* (5). *Let U_* be any sub-optimal solution of the joint diagonalization problem* (6). *Then there is a U_0, which is an exact diagonalizer of the ground-truth matrices, such that U_* can be written as*

$$U_* = U_0 e^{\alpha_* X_*} \qquad \alpha_* > 0 \qquad X = -X^T \quad \|X\| = 1 \qquad (12)$$

and α_* obeys

$$\alpha_* \leq \frac{\sigma \hat{g}}{\hat{\Gamma}} \left(\hat{g} + 2 + \sqrt{\hat{g}^2 + 4\hat{g} + 1 + \frac{\delta_{\rm opt}}{2\sigma^2 \hat{g}}} \right) \qquad (13)$$

up to $O(\sigma^2)$ terms, where $\hat{g} = \frac{\hat{\Gamma}^2}{\hat{\gamma}}$,

$$\hat{\gamma} = \min_{i<i'} \sum_{r=1}^{R} ([U_*^T \hat{M}_r U_*]_{ii} - [U_*^T \hat{M}_r U_*]_{i'i'})^2 \qquad (14)$$

$$\hat{\Gamma} = \sum_{r=1}^{R} \sum_{i<i'} ([U_*^T \hat{M}_r U_*]_{ii} - [U_*^T \hat{M}_r U_*]_{i'i'})^2 \qquad (15)$$

and $\delta_{\rm opt}$ is defined in (11).

Proof's sketch

Let U_0 be a joint diagonalizer of the ground-truth matrices M_r, $r = 1, \dots, R$. Every orthogonal matrix can be written in the form $U = U_0 e^{\alpha X}$, for some $\alpha > 0$ and $X = -X^T$. Without loss of generality one can also assume $\|X\| = 1$. The empirical bound in Theorem 1 is an inequality on the perturbation parameter α. According to the expansion $U = U_0 e^{\alpha X}$, the perturbation parameter α is interpreted as the 'distance' between U and the ground-truth solution U_0. The proof of Theorem 1 consists of four steps:

(i) *Characterization of the optimal solutions*: an optimal solution $U_{\rm opt}$ of (6) is expanded around U_0, the closest joint diagonalizer of the ground-truth matrices, and characterized via a linear inequality on $\alpha_{\rm opt}$, the perturbation parameter defined by the expansion $U_{\rm opt} = U_0 e^{\alpha_{\rm opt} X_{\rm opt}}$;

(ii) *Distance from the optimal solution*: the distance between $U_{\rm opt}$ and the obtained sub-optimal solution U_* is estimates as a function of the ground-truth matrices and the optimality gap (11);

(iii) *Empirical estimation*: the theoretical inequalities obtained in the previous steps are converted into 'empirical' linear inequalities via a Taylor expansion in the parameter σ;

(iv) *Triangular inequality*: a bound on α_* in terms of the empirical bound on $\alpha_{\rm opt}$ and the obtained distance from the optimal solution is derived via the triangular inequality.

V. Proof of Theorem 1

The proof of Theorem 1 is organised in the four main steps outlined in the previous section. A detailed description of each step is provided in the followingsolution to.

A. Characterization of the optimal solutions

In this subsection we recall a result for the perturbation of joint diagonalizers that first appeared in [Cardoso, 1994]. However, the bound obtained here is slightly different from the one of [Cardoso, 1994] for the following reason. Since our goal is only to bound the 'distance' between the optimal solutions of (6), $U_{\rm opt}$, and the ground-truth mixing matrices, U_0, we focus on the magnitude of the perturbation parameter α_{opt} appearing in the expansion $U_{\rm opt} = U_0 e^{\alpha_{\rm opt} X_{\rm opt}}$, where $X_{\rm opt}$ is a unit norm skew-symmetric matrix. As in [Cardoso, 1994], a first order bound in the parameter $\alpha_{\rm opt}$ and the noise level σ defined in (5) is obtained by expanding around U_0 the stationarity equation $\nabla \mathcal{L}(U) = 0$, where $\nabla \mathcal{L}(U)$ is the gradient of the objective function (6) at U. The key difference between the derivation described here and [Cardoso, 1994] is that we do not provide an explicit form for $X_{\rm opt}$ in terms of the unperturbed matrices M_r and noise matrices W_r. In fact, these are ground-truth quantities that are in general not known and cannot appear in the empirical estimation.

The bound on $\alpha_{\rm opt}$ is obtained as follows. The gradient of (6) at U is computed by considering the directional derivative of (6) at U in a tangent space direction $Z = -Z^T$

$$\begin{aligned}
\langle Z, \nabla \mathcal{L} \rangle &= \frac{d}{dt} \mathcal{L}(U e^{tZ}) \Big|_{t=0} && (16) \\
&= \frac{d}{dt} \sum_{r=1}^{R} \| \mathrm{off}(e^{-tZ} U^T \hat{M}_r U e^{tZ}) \|^2 \Big|_{t=0} && (17) \\
&= 2 \sum_{r=1}^{R} \mathrm{Tr} \left(Z [\mathrm{off}(U^T \hat{M}_r U), U^T \hat{M}_r^T U] \right) && (18) \\
&= -2 \langle Z, \sum_{r=1}^{R} [\mathrm{off}(U^T \hat{M}_r U), U^T \hat{M}_r U] \rangle && (19)
\end{aligned}$$

where we have defined $\langle A, B\rangle = \mathrm{Tr}(A^T B)$, $[A, B] = AB - BA$, used the cliclic properties of the trace, $\hat{M}_r = \hat{M}_r^T$ and $Z = -Z^T$. The solutions of (6) are, by definition, stationary points of (6) and hence solutions of the stationarity equation above. Let U_{opt} be a minimizer of (6) in a neighbourhood of U_0, which is a joint diagonalizer of the ground-truth matrices. Since U_{opt} is an orthogonal matrix, it can be written as $U_{\mathrm{opt}} = U_0 e^{\alpha_{\mathrm{opt}} X_{\mathrm{opt}}}$, with $X_{\mathrm{opt}} = -X_{\mathrm{opt}}^T$, $\|X_{\mathrm{opt}}\| = 1$. Moreover, if the noise parameter is small enough, one can assume that U_{opt} is not too far from U_0 and consider a linear expansion of the stationarity condition around U_0. We have

$$\begin{aligned} 0 &= \sum_{r=1}^{R} [\mathrm{off}(U_{\mathrm{opt}}^T \hat{M}_r U_{\mathrm{opt}}), U_{\mathrm{opt}}^T \hat{M}_r U_{\mathrm{opt}}] & (20) \\ &= \sum_{r=1}^{R} [\mathrm{off}\left([\Lambda_r, \alpha_{\mathrm{opt}} X_{\mathrm{opt}}]\right), \Lambda_r] & (21) \\ &\quad + \sum_{r=1}^{R} [\mathrm{off}\left(U_0^T \sigma W_r U_0\right), \Lambda_r] + O((\alpha_{\mathrm{opt}} + \sigma)^2) \end{aligned}$$

where we have defined $\Lambda_r = \mathrm{diag}([\Lambda_{r1}, \ldots, \Lambda_{rn}])$. The idea is to isolate the terms that are linear in $\alpha_{\mathrm{opt}} X_{\mathrm{opt}}$ and obtain a linear equation of the form $\alpha_{\mathrm{opt}} \tilde{T} \mathrm{vec}(X_{\mathrm{opt}}) = w$, where $\tilde{T}$ is a $d^2 \times d^2$ matrix and and w is the vectorization of the term proportional to σ. The linear equation can be used to estimate α_{opt} by finding a lower bound of $\|\alpha_{\mathrm{opt}} \tilde{T} \mathrm{vec}(X_{\mathrm{opt}})\|$ under the constraint $\|X_{\mathrm{opt}}\| = 1$. It is easy to show that the linear operator $\tilde{T}$ is given by

$$\tilde{T} = \sum_{r=1}^{R} \tilde{t}_r^2 \quad \tilde{t}_r = (1 \otimes \Lambda_r - \Lambda_r \otimes 1)\mathrm{Off} \tag{22}$$

and that $\tilde{T}\mathrm{vec}(X) = 0$ for all $X = -X^T$. In order to use the linearized stationarity equation to bound $\alpha_{\mathrm{opt}} = \|\mathrm{vec}(\alpha_{\mathrm{opt}} X_{\mathrm{opt}})\|$ we need to project the equation into a subspace where the corresponding linear operator is invertible. This subspace is the subspace of strictly lower-diagonal matrices. The idea is to exploit the fact that $\alpha_{\mathrm{opt}} X_{\mathrm{opt}}$ is skew-symmetric. Since the lower-diagonal part equals the upper-diagonal part up to a sign flip, it is enough to characterize its lower-diagonal part.[2] First of all, it easy to verify that $\sum_{r=1}^{R} [\mathrm{off}\left([\Lambda_r, X]\right), \Lambda_r]$ is a skew symmetric matrix. For any skew symmetric matrix X one has $X = \mathrm{low}(X) - \mathrm{low}(X)^T$, where the operator $\mathrm{low}(\cdot)$ is defined by $[\mathrm{low}(A)]_{ij} = A_{ij}$ if $i > j$ and $[\mathrm{low}(A)]_{ij} = 0$ for $i \leq j$. It follows that $\sum_{r=1}^{R} [\mathrm{off}\left([\Lambda_r, X]\right), \Lambda_r] = 0$ implies $\mathrm{low}\left(\sum_{r=1}^{R} [\mathrm{off}\left([\Lambda_r, X]\right), \Lambda_r]\right) = 0$ and vice versa. In particular we observe that

$$\begin{aligned} \mathrm{low}\left(\sum_{r=1}^{R} [\mathrm{off}\left([\Lambda_r, X]\right), \Lambda_r]\right) = & \qquad (23) \\ \mathrm{low}\left(\sum_{r=1}^{R} [\mathrm{low}\left([\Lambda_r, \mathrm{low}(X)]\right), \Lambda_r]\right) & \end{aligned}$$

as it can be shown by using the fact that the commutator between a diagonal matrix and a strictly upper-diagonal matrix is strictly upper-diagonal. We consider the projected stationarity equation

$$\begin{aligned} \mathrm{low}\left(\sum_{r=1}^{R} [\mathrm{low}\left([\Lambda_r, \mathrm{low}(\alpha_{\mathrm{opt}} X_{\mathrm{opt}})]\right), \Lambda_r]\right) = & \qquad (24) \\ -\mathrm{low}\left(\sum_{r=1}^{R} [\mathrm{off}\left(U_0^T \sigma W_r U_0\right), \Lambda_r]\right) & \end{aligned}$$

that holds up to $O((\alpha_{\mathrm{opt}} + \sigma)^2)$ terms. The vectorization of the above equation reads

$$\begin{aligned} T\mathrm{vec}(\alpha_{\mathrm{opt}} X_{\mathrm{opt}}) = & \qquad (25) \\ -\mathrm{vec}\left(\sum_{r=1}^{R} \mathrm{low}\left([\mathrm{off}\left(U_0^T \sigma W_r U_0\right), \Lambda_r]\right)\right) & \end{aligned}$$

where T is a linear operator defined by

$$Tx = \mathrm{vec}\left(\mathrm{low}\left(\sum_{r=1}^{R} [\mathrm{low}\left([\Lambda_r, \mathrm{low}(\mathrm{mat}(x))]\right), \Lambda_r]\right)\right) \tag{26}$$

where x is a d^2-dimensional vector and $\mathrm{mat}(x)$ its column-wise matricization. The linear operator T can be written explicitly in terms of Kronecker products of the diagonal matrices Λ_r and a linear operator Low defined by $\mathrm{Low}\mathrm{vec}(A) = \mathrm{vec}(\mathrm{low}(A))$. Its explicit form is

$$T = \sum_{r=1}^{R} t_r^2 \qquad t_r = \mathrm{Low}(1 \otimes \Lambda_r - \Lambda_r \otimes 1)\mathrm{Low} \tag{27}$$

The matrices t_r are a diagonal matrices whose non-vanishing diagonal elements are

$$[t_r]_{ii} = \sum_{r=1}^{R} \Lambda_{rj} - \Lambda_{rj'} \qquad i = d(j-1) + j' \text{ and } j < j' \tag{28}$$

Note that T is always rank-deficient and hence non-invertible. However, since $X_{\mathrm{opt}} = -X_{\mathrm{opt}}^T$ and $\|X_{\mathrm{opt}}\| = 1$, a lower bound on the perturbation parameter α_{opt} can be obtained by taking the norm of both sides of (25). Let

$$\begin{aligned} \gamma &= \min\{\|T\mathrm{vec}(X)\|, \|X\| = 1, X = -X^T\} & (29) \\ &= \min_{j<j'} \sum_{r=1}^{R} (\Lambda_{rj} - \Lambda_{rj'})^2 & (30) \end{aligned}$$

then it easy to see that $\gamma > 0$ if, for all $j \neq j'$ there exists at least one $r \in \{1, \ldots, R\}$ such that $\Lambda_{rj} \neq \Lambda_{rj'}$. This is the non-degeneracy condition required in [Cardoso, 1994] for (6) to be well posed. Note that one has $\gamma > 0$ because all vanishing elements on the diagonal of T multiply the vanishing elements of $\mathrm{vec}(X)$, if $X = -X^T$. From (25)

[2] A similar splitting operator technique has been used for example in [Konstantinov et al., 1994] for the perturbation analysis of a slightly different matrix factorisation problem.

one has

$$(\alpha_{\rm opt}\gamma)^2 \leq \|T\mathrm{vec}(\alpha_{\rm opt}X_{\rm opt})\|^2 \tag{31}$$

$$\leq \left\|\mathrm{low}\left(\sum_{r=1}^{R}[\mathrm{low}\left(U_0^T\sigma W_r U_0\right), \Lambda_r]\right)\right\|^2 \tag{32}$$

$$= \left\|\sum_{r=1}^{R} t_r \mathrm{vec}(U_0^T\sigma W_r U_0)\right\|^2 \tag{33}$$

$$\leq \sum_{r=1}^{R}\|t_r\|^2 \sum_{r=1}^{R}\|\sigma W_r\|^2 \tag{34}$$

$$= \Gamma^2\mathcal{W}^2 \tag{35}$$

where we have defined

$$\Gamma^2 = \sum_{r=1}^{R}\|t_r\|^2 = \sum_{r=1}^{R}\sum_{i<i'}(\Lambda_{ri}-\Lambda_{ri'})^2 \tag{36}$$

$$\mathcal{W}^2 = \sum_{r=1}^{R}\|\sigma W_r\|^2 \tag{37}$$

and used $\|A\| = \|\mathrm{vec}(A)\|$, $\mathrm{low}([\mathrm{off}(A),\Lambda]) = \mathrm{low}([\mathrm{low}(A),\Lambda])$, for any diagonal Λ and $\|U_0^TAU_0\| = \|A\|$. This implies

$$\alpha_{\rm opt} \leq \frac{\Gamma\mathcal{W}}{\gamma} + O((\alpha_{\rm opt}+\sigma)^2) \tag{38}$$

B. Distance From the Optimal Solution

Aim of this subsection is to estimate the distance between a sub-optimal solution and the closest optimal solution characterized by (38). Again, since the sub-optimal solution is an orthogonal matrix we can write $U_* = U_0e^{\alpha_*}X_*$, with $\|X_*\| = 1$ and $\alpha_* > 0$. The goal is to obtain a bound for $\|\alpha_*X_* - \alpha_{opt}X_{opt}\|$ as a function of the optimality gap (11). We consider the inequality

$$\delta_{\rm opt} \geq \sum_{r=1}^{R}\|\mathrm{off}(U_*^T\hat{M}_rU_*)\|^2 - \sum_{r=1}^{R}\|\mathrm{off}(U_{\rm opt}^T\hat{M}_rU_{\rm opt})\|^2 \tag{39}$$

$$= 2\sum_{r=1}^{R}\|\mathrm{low}(U_*^T\hat{M}_rU_*)\|^2 - 2\sum_{r=1}^{R}\|\mathrm{low}(U_{\rm opt}^T\hat{M}_rU_{\rm opt})\|^2 \tag{40}$$

where the second equality holds because the matrices $U_*^T\hat{M}_rU_*$ and $U_{\rm opt}^T\hat{M}_rU_{\rm opt}$ are symmetric. Expanding in $\alpha_{\rm opt}$, α_* and σ and neglecting higher order terms we obtain

$$\frac{\delta_{\rm opt}}{2} \geq \sum_{r=1}^{R}\left(y^Tt_r^Tt_ry + 2y^Tt_r^T(t_rz+a_r) - 2z^Tt_r^Tt_rz\right) \tag{41}$$

where we have defined $y = \mathrm{vec}(\alpha_*X_* - \alpha_{opt}X_{opt})$, $z = \mathrm{vec}(\alpha_{opt}X_{opt})$, $t_r = \mathrm{Low}(1\otimes\Lambda_r - \Lambda_r\otimes 1)\mathrm{Low}$ and $a_r = \mathrm{vec}(\mathrm{low}(U_0^T\sigma W_rU_0))$. This implies

$$y^TTy \leq \frac{\delta_{\rm opt}}{2} + 2z^TTz - 2y^T\sum_{r=1}^{R}t_r^T(t_rz+a_r) \tag{42}$$

where $T = \sum_{r=1}^{R}t_r^Tt_r$. Letting $\tilde{y} = \frac{y}{\|y\|}$ and using $\gamma \leq \tilde{y}^TT\tilde{y}$ one has

$$\|y\|^2\gamma \leq \left|\frac{\delta_{\rm opt}}{2} + 2z^TTz - 2y^T\sum_{r=1}^{R}t_r^T(t_rz+a_r)\right| \tag{43}$$

$$\leq C + 2B\|y\| \tag{44}$$

where

$$C = \frac{\delta_{\rm opt}}{2} + 2z^TTz \leq \frac{\delta_{\rm opt}}{2} + 2\alpha_{\rm opt}^2\Gamma^2 \tag{45}$$

$$B = \|\sum_{r=1}^{R}t_r^T(t_rz+a_r)\| \leq \alpha_{\rm opt}\Gamma^2 + \Gamma\mathcal{W} \tag{46}$$

This is a scalar second order inequality in $\|y\| = \|\alpha_*X_* - \alpha_{opt}X_{opt}\|$. One obtains

$$\|\alpha_*X_* - \alpha_{\rm opt}X_{\rm opt}\| \leq \tag{47}$$
$$\frac{B+\sqrt{B^2+C\gamma}}{\gamma} + O((\alpha_{\rm opt}+\alpha_*+\sigma)^2)$$

C. Empirical Estimations

In order to obtain an empirical error bound, all quantities in (38) and (47) should be expressed as functions of the empirical matrices $\hat{M}_r$, the obtained solution and the optimality gap. The formers are characterized by the empirical joint eigenvalues, given by

$$\hat{\Lambda}_{ri} = U_*^T\hat{M}_rU_* \qquad r = 1,\ldots,R \quad i = 1,\ldots,n \tag{48}$$

where U_* is the obtained sub-optimal solution. For all $r = 1,\ldots,R$ and all $i = 1,d\ldots,n$, the $\hat{\Lambda}_{ri}$ obey

$$\hat{\Lambda}_{ri} = \Lambda_{ri} + \left[[U_0^TM_rU_0, \alpha_*X_*] + U_0^T\sigma W_rU_0\right]_{ii} \tag{49}$$

$$= \Lambda_{ri} + O(\alpha_*) + O(\sigma) \tag{50}$$

that implies

$$\hat{\gamma} = \min_{i\neq i'}\sum_{r=1}^{R}(\hat{\Lambda}_{ri}-\hat{\Lambda}_{ri'})^2 = \gamma + O(\alpha_*+\sigma) \tag{51}$$

with γ defined in (29), and

$$\hat{\Gamma}^2 = \sum_{r=1}^{R}\sum_{i<i'}(\hat{\Lambda}_{ri}-\hat{\Lambda}_{ri'})^2 = \Gamma^2 + O((\alpha_*+\sigma)^2) \tag{52}$$

with Γ defined in (36). Finally, by assuming $\|\sigma W_r\|^2 \leq 1$ one has $\mathcal{W}^2 = \sigma R$. Alternatively, one could assume all W_r to be relatizations of a matrix random variable $\mathfrak{W}$ defined by $\{W \sim \mathfrak{W} : W = W^T, [W]_{ij} = \mathcal{N}(0,1)\}$. In this case the matrices W_r are unbounded but a probabilistic constraint on $\frac{1}{R}\sum_r\|W_r\|^2$ can be obtained by computing the expectation and the variance of the the random variable $\|\mathfrak{W}\|^2$ and then applying the Chebyshev's inequality $\Pr(|x-E(x)| > t) < \frac{E((x-E(x))^2)}{t^2}$, with $t > 0$ and x an arbitrary scalar random variable. By choosing for simplicity the former assumption, the following empirical bounds hold up to higher order terms in $(\alpha_{\rm opt}+\alpha_{\rm opt}+\sigma)$

$$C \leq \frac{\delta_{\rm opt}}{2} + 2\frac{\sigma^2\hat{\Gamma}^4}{\hat{\gamma}^2} \qquad B \leq \sigma\hat{\Gamma}(1+\frac{\hat{\Gamma}^2}{\hat{\gamma}}) \tag{53}$$

where we have used $\alpha_{\rm opt} \leq \frac{\sigma\hat{\Gamma}}{\hat{\gamma}} + O((\alpha_{\rm opt}+\alpha_{\rm opt}+\sigma)^2)$.

D. *Triangular inequality*

The result obtained in Section V-B can be combined with the error bound obtained in Section V-A and the empirical estimations given in Section V-C to obtain an empirical inequality on the distance between the sub-optimal solution U_* and the closest ground-truth mixing matrix U_0. This is done via the triangular inequality

$$\alpha_* = \|\alpha_* X_*\| \leq \|\alpha_* X_* - \alpha_{\rm opt} X_{\rm opt}\| + \|\alpha_{\rm opt} X_{\rm opt}\| \tag{54}$$

Up to second order terms in the perturbation parameters $\alpha_{\rm opt}$, α_* and σ one obtains

$$\begin{aligned} \alpha_* &\leq \frac{B + \sqrt{B^2 + C\gamma}}{\gamma} + \alpha_{opt} && (55) \\ &\leq \frac{\sigma \hat{g}}{\hat{\Gamma}} \left(\hat{g} + 2 + \sqrt{\hat{g}^2 + 4\hat{g} + 1 + \frac{\delta_{\rm opt}}{2\sigma^2 \hat{g}}} \right) && (56) \end{aligned}$$

where we have defined $\hat{g} = \frac{\hat{\Gamma}^2}{\hat{\gamma}}$ and $\delta_{\rm opt}$ is given in (11).

VI. Non Orthogonal ICA and Other Extensions

The method used to compute the optimality gap in Section III can be generalised to various simultaneous matrix decomposition problems. More precisely, a spectral relaxation of the type (8) exists for all simultaneous decomposition problems that can be associated to an objective function of the form

$$\begin{aligned} L(X, Y) &= \sum_r^R \|p(X\hat{M}_r Y)\|^2 && (57) \\ &= {\rm Tr}(PP^T (X \otimes Y)^T \hat{m}\hat{m}^T (X \otimes Y)) && (58) \end{aligned}$$

where P is a linear projector defined by $P{\rm vec}(A) = {\rm vec}(p(A))$, with $p(\cdot)$ being any generalisation of the off operator defined in Section II, and $\hat{m} = [{\rm vec}(\hat{M}_1), \ldots, {\rm vec}(\hat{M}_R)]$. It should be noticed that the obtained optimality gap trivialises if ${\rm rank}(X \otimes Y) < n^2 - {\rm rank}(\hat{m}\hat{m}^T)$ because in that case all independent columns of the variable $(X \otimes Y)$ can be chosen in the null space of $\hat{m}\hat{m}^T$ and the minimum of the relaxed objective is 0. When such rank-condition is satisfied, one obtains a non-vanishing optimality gap, which may be used to compute empirical bounds similar to the result of Section IV. However, this would require a characterisation of the optimal solution analogous to (38) that is not available for the general case. First requirement for such an analysis is the existence of a ground-truth solution. This is guaranteed if the matrices $\hat{M}_r$ are assumed to be in the form $\hat{M}_r = X_0^{-1} T_r Y_0^{-1} + \sigma W_r$, with $p(T_r) = 0$ for all $r = 1, \ldots, R$. More generally, all steps involved in the derivation of Theorem 1 would be related on the specific properties of the expansion of the relevant variables, *i.e.*

$$X = Z_0 + t \left(\frac{d}{dt} X(Z_0 + tE)|_{t=0} \right) + O(t^2) \tag{59}$$

where $X(t = 0) = Z_0$ and E is any direction. We leave the explicit derivation of the empirical bound in the case of more general objectives for future work.

VII. Experiments

A set of numerical experiments has been designed to study the tightness of the spectral relaxation described in Section III and the empirical bound defined by Theorem 1 in Section IV. We considered the problem of the approximate joint diagonalisation of symmetric matrices of dimension $n = 5, 10$. For each n, distinct datasets of nearly joint diagonalizable matrices have been generated, with varying sample size (number of matrices) $R = 10, 50, 100, 500$ and noise level $\sigma = 10^{-12}, 10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}$. For each parameter setting, we have created 10 different datasets by choosing 10 random ground-truth matrices U_0, 10 joint eigenvalues matrices Λ and the corresponding noise matrices W_r. In particular, the noise matrices W_r have been chosen to be gaussian symmetric random matrices of zero mean and unit variance, *i.e.* $[W_r]_{ij} \sim \mathcal{N}(0, 1)$ for all $i \leq j$ $i, j = 1, \ldots, n$ and $W_r = W_r^T$. According to the assumption considered in our perturbation analysis, the noise matrices have been randomly renormalised to satisfy $\|W_r\| \leq 1$. The joint diagonalization of all datasets has been performed with the Jacobi algorithm of [Cardoso and Souloumiac, 1996]. For each datasets we have considered 10 different initial conditions. The convergence of the algorithm for all different initial conditions is shown in Figure 1. In all cases the values assumed by the objective function (6) at convergence within the same dataset are statistically equivalent, independently of the initial conditions. Moreover, a visual comparison with the global minimum of the corresponding convex relaxation shows that the obtained solutions are close to global optimality. To make this evaluation more quantitative we have defined a relative optimality gap

$$\tilde{\delta}_{\rm opt} = \frac{\mathcal{L}(U_*) - L(V_*)}{\mathcal{L}(U_*)} \tag{60}$$

where $\mathcal{L}(U)$ and $L(V)$ are given in (6) and (8) respectively, U_* is solution computed by the Jacobi algorithm and V_* the global optimum of the relaxed objective, computed as explained in Section III. For all R and σ, all datasets and all initial conditions we have computed the relative optimality gap and averaged over all simulations sharing the same parameters. The obtained values are reported in Table I. A synthetic view of the dependence of the relative optimality gap respect to the sample size, R, and the noise level, σ, is provided by Figure 2 Finally, we have used the same numerical simulations to investigate the tightness of the empirical bound described in Section IV. For each empirical solution U_* obtained by the Jacobi algorithm on a given dataset, we have computed the perturbation parameter α_* and the associated empirical bound, according to the definitions given in Theorem 1. Average values for each n, R and σ are shown in Figure 3. The intermediate lines in the plots correspond to the bound (38) for the provably good optimal solutions $U_{\rm opt}$. In general, even if the bound provided by the linear analysis of Section IV is expected to hold only up to second order terms, the inequalities are not violated for any value of the perturbation parameters α_*, σ. However, the bound does not seem to be sufficiently tight to be used

$n=5$	$\epsilon=10^{-12}$	$\epsilon=10^{-8}$	$\epsilon=10^{-4}$	$\epsilon=10^{-2}$	$\epsilon=10^{-1}$
$R=10$	0.41 ± 0.11	0.43 ± 0.09	0.45 ± 0.08	0.45 ± 0.09	0.42 ± 0.06
$R=50$	0.07 ± 0.01	0.08 ± 0.01	0.08 ± 0.02	0.08 ± 0.01	0.09 ± 0.01
$R=100$	0.04 ± 0.01	0.04 ± 0.01	0.04 ± 0.01	0.04 ± 0.00	0.04 ± 0.00
$R=500$	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00

$n=10$	$\epsilon=10^{-12}$	$\epsilon=10^{-8}$	$\epsilon=10^{-4}$	$\epsilon=10^{-2}$	$\epsilon=10^{-1}$
$R=10$	0.99 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
$R=50$	0.18 ± 0.01	0.18 ± 0.01	0.19 ± 0.01	0.18 ± 0.02	0.18 ± 0.01
$R=100$	0.09 ± 0.01	0.09 ± 0.01	0.09 ± 0.00	0.09 ± 0.01	0.09 ± 0.00
$R=500$	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00

TABLE I: Average relative optimality gaps for n=5,10, different sample sizes R and noise levels σ. The displayed values are obtained by averaging over all datasets and initial conditions for a given choice of the parameters R and σ. The reported uncertainty is the corresponding standard variation.

in practical applications. Improvements in this sense can be expected from a nonlinear extension of the first order analysis considered here.

VIII. Conclusions

This paper shows how to compute a globally defined optimality gap for a class of joint matrix decomposition problems. In the orthogonal setting (joint diagonalization of symmetric matrices) it is possible to use the obtained optimality gap to bound the error in the recovered orthogonal diagonalizer as a function of the input matrices and noise level. Applied to ICA, this provides global guarantees on the estimation of the mixing matrix that can be computed from the observed signal alone (by assuming a prior knowledge of the noise level). Numerical simulations show that the spectral relaxation used to compute the gap can be really tight. Moreover, the result allows us to conclude that, in most cases, the local solutions computed by the Jacobi iterative algorithm of [Cardoso and Souloumiac, 1996] are close to global optimality. As a future direction, we plan to extend the full error analysis to the non-orthogonal setting or more general simultaneous matrix decomposition problems, as for example the simultaneous Schur decomposition. More practical applications in various fields (signal processing, linear algebra, tensor factorisation..) will be also addressed in forthcoming follow-up of this work.

References

[Afsari, 2006] Afsari, B. (2006). Simple lu and qr based non-orthogonal matrix joint diagonalization. In *Independent Component Analysis and Blind Signal Separation*, pages 1–7. Springer.

[Afsari and Krishnaprasad, 2004] Afsari, B. and Krishnaprasad, P. S. (2004). Some Gradient Based Joint Diagonalization Methods for ICA. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Puntonet, C. G., and Prieto, A., editors, *Independent Component Analysis and Blind Signal Separation*, volume 3195, pages 437–444. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Bunse-Gerstner et al., 1993] Bunse-Gerstner, A., Byers, R., and Mehrmann, V. (1993). Numerical Methods for Simultaneous Diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 14(4):927–949.

[Cardoso, 1994] Cardoso, J.-F. (1994). Perturbation of joint diagonalizers. Ref# 94d027. Technical report, Tlcom Paris.

[Cardoso, 1999] Cardoso, J.-F. (1999). High-Order Contrasts for Independent Component Analysis. *Neural Computation*, 11(1):157–192.

[Cardoso and Souloumiac, 1996] Cardoso, J.-F. and Souloumiac, A. (1996). Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, 17(1):161–164.

[Konstantinov et al., 1994] Konstantinov, M., Petkov, P. H., and Christov, N. (1994). Nonlocal perturbation analysis of the schur system of a matrix. *SIAM Journal on Matrix Analysis and Applications*, 15(2):383–392.

[Manton, 2002] Manton, J. H. (2002). Optimization algorithms exploiting unitary constraints. *Signal Processing, IEEE Transactions on*, 50(3):635–650.

[Souloumiac, 2009] Souloumiac, A. (2009). Joint diagonalization: Is non-orthogonal always preferable to orthogonal? pages 305–308. IEEE.

[Van der Veen, 2001] Van der Veen, A.-J. (2001). Joint diagonalization via subspace fitting techniques. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 5, pages 2773–2776. IEEE.

[Wax and Sheinvald, 1997] Wax, M. and Sheinvald, J. (1997). A least-squares approach to joint diagonalization. *Signal Processing Letters, IEEE*, 4(2):52–53.

[Yamada and Ezaki, 2003] Yamada, I. and Ezaki, T. (2003). An orthogonal matrix optimization by dual Cayley parametrization technique. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*.

[Yeredor, 2002] Yeredor, A. (2002). Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. *Signal Processing, IEEE Transactions on*, 50(7):1545–1553.

[Ziehe et al., 2004] Ziehe, A., Laskov, P., Nolte, G., and Mller, K.-R. (2004). A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *The Journal of Machine Learning Research*, 5:777–800.

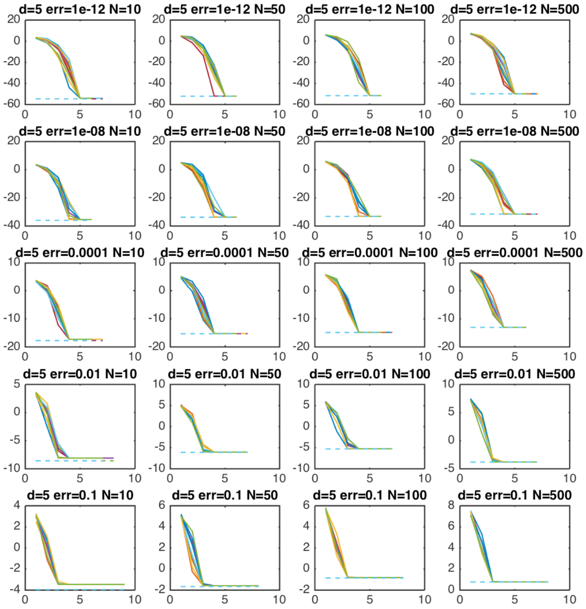

Fig. 1: The convergence of the Jacobi algorithm of [Cardoso and Souloumiac, 1996] for different initial conditions and different datasets. The orthogonal matrices used as initial conditions were generated by taking the left singular vectors of a random matrix. In all cases, the algorithm converges to points that can be considered statistically equivalent in terms of the corresponding objective values. For all parameters settings ($R = 10, 50, 100, 500$, $\sigma = 10^{-12}, 10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}$) only the results obtained on the first of the 10 experiments are considered. The plots show the value of the objective function (y-axis) at each iteration (x-axis), with different lines corresponding to different initial conditions. In each plot, the black dashed line represents the global minimum of the relaxed objective that is used to compute the optimality gap.

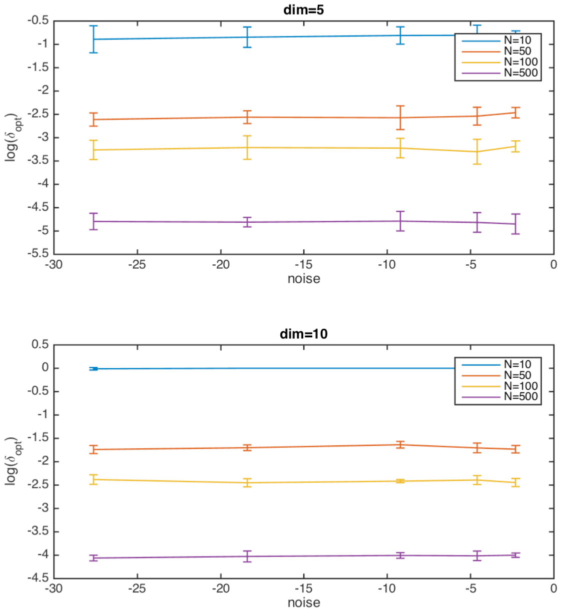

Fig. 2: Average relative gap for different sample size $R = 10, 50, 100, 500$ and noise level $\sigma = 10^{-12}, 10^{-8}, 10^{-4}, 10^{-2}, 10^{-1}$. For given R and σ, the average gap is over all corresponding datasets and initial conditions. Errorbars represent the associated standard deviations.

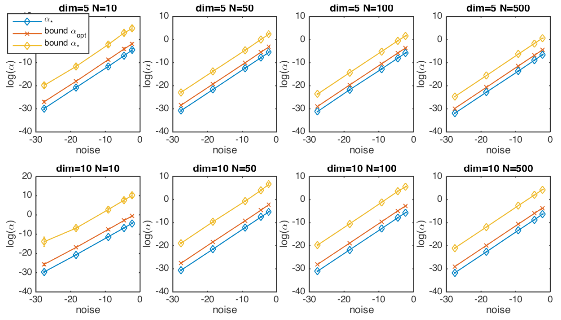

Fig. 3: Tightness of the empirical bound defined by Theorem 1 for different sample sizes and noise levels. The plots show average values for the empirical perturbation parameter α_*, defined by the expansion $U_* = U_0 e^{\alpha_* X_*}$ with $X_* = -X_*^T$ and $\|X_*\| = 1$, and the associated empirical bound. The empirical bound is computed from the obtained solution by applying the definitions of Theorem 1. The intermediate line is a bound on the perturbation parameter α_{opt}, associated with the optimal solutions of (6).