The Inverse Problem of Positive Autoconvolution

Lorenzo Finesso and Peter Spreij

***Abstract*— We pose the problem of approximating optimally a given nonnegative signal with the scalar autoconvolution of a nonnegative signal. The I-divergence is chosen as the optimality criterion being well suited to incorporate nonnegativity constraints. After proving the existence of an optimal approximation, we derive an iterative descent algorithm of the alternating minimization type to find a minimizer. The algorithm is based on the lifting technique developed by Csiszár and Tusnádi and exploits the optimality properties of the related minimization problems in the larger space. We study the asymptotic behavior of the iterative algorithm and prove, among other results, that its limit points are Kuhn-Tucker points of the original minimization problem. Numerical experiments confirm the asymptotic results and exhibit the fast convergence of the proposed algorithm.**

***Index Terms*— Autoconvolution, inverse problem, positive system, I-divergence, alternating minimization.**

I. Introduction

INVERSE problems in system modeling and identification have a long tradition and have been the subject of a vast technical literature in applied mathematics, engineering, and specialized applied fields. The focus of this paper is on the subclass of inverse problems for which the models are of *autoconvolution* type. In linear time-invariant systems, inputs are transformed into outputs by convolution with a kernel representing the system's impulse response. Autoconvolution systems produce the output by convolution of the input signal with itself.

A lot of work has been dedicated to the inverse problem of autoconvolution for functions on the real line, emphasizing the functional analytic aspects and motivating its interest in a variety of applications in physics and engineering. Most of the contributions analyse special cases, where exact solutions to the inverse problem exist, and propose different theoretical approaches for their construction. For example, [9] focuses on inversion of autoconvolution integrals using spline functions. In [22], inversion is studied based on the application of the FFT algorithm and digital signal processing concepts. Special cases arise when dealing with autoconvolution of probability density functions, as in [16]. In [10], the autoconvolution has been introduced for continuous-time processes as an alternative to autocorrelation.

The purpose of this paper is threefold. First, we pose the problem of time-domain approximation of a discrete-time nonnegative input/output system by finite autoconvolutions, when the output observations are available. Following the choice made in other optimization problems for nonnegative system, we opt for the I-divergence, which as argued in [7] (see also [28]), is the natural choice under nonnegativity constraints. We provide a result on the existence of the minimizer of the approximation criterion. Then we propose an iterative algorithm to find the best approximation, and finally we study the asymptotic behavior of the algorithm.

We employ techniques that have already been used in [12] to analyse a nonnegative matrix factorization problem and the approach is similar to the one in [13], [14], but differs from the latter references as they treat linear convolutional problems, whereas the autoconvolution is inherently nonlinear. The algorithm that we propose is of the alternating minimization type, and the optimality conditions (the Pythagorean relations) are satisfied at each step.

The inherent nonconvexity and nonlinearity of the problem make the analysis of the asymptotic behavior challenging. The main result in this respect is contained in Proposition IV.8, which states that all limit points of the algorithm satisfy the Kuhn-Tucker optimality conditions. This should be compared with other known results on the convergence of alternating minimization algorithms. In some cases, it is possible to show convergence to a (unique) limit, which is also the minimizer of the criterion. This happens, in particular, when dealing with a convex criterion. Contributions in this direction are e.g. [5], [28], and [13], [14], [30]. On the other hand, for nonconvex, nonlinear problems, to the best of our knowledge, there are no asymptotic results comparable with the present Proposition IV.8.

The nonparametric approach to the inverse problem we follow in this paper is different from the one followed in identification or realization of nonnegative and linear systems; see [2] for a survey, and for instance [1], [11].

The main differences between the cited literature and this paper are that we consider approximation problems, rather than looking for exact solutions which exist only exceptionally, and that our (time) domain is discrete rather than the real line. Moreover, the nonnegativity constraint, which we impose on all signals, is a crucial feature of the present work. Some earlier work shares, at least in part, our point of view, e.g. the papers [3], [4] dealing with image processing and 2D

Manuscript received 9 June 2022; revised 29 November 2022; accepted 31 January 2023. Date of publication 13 February 2023; date of current version 19 May 2023. This work was supported in part by the STM-2020 CNR Program. *(Corresponding author: Peter Spreij.)*

Lorenzo Finesso is with the Institute of Electronics, Information Engineering and Telecommunications, National Research Council, (CNR-IEIIT), 35131 Padua, Italy (e-mail: lorenzo.finesso@unipd.it).

Peter Spreij is with the Korteweg-de Vries Institute for Mathematics, Universiteit van Amsterdam, 1012 WX Amsterdam, The Netherlands, and also with the Institute for Mathematics, Astrophysics and Particle Physics, Radboud University, 6525 XZ Nijmegen, The Netherlands (e-mail: spreij@uva.nl).

Communicated by R. Talmon, Associate Editor for Signal Processing and Source Coding.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2023.3244407.

Digital Object Identifier 10.1109/TIT.2023.3244407

systems contain an algorithm of the same type as ours and an analysis of its behavior. In [26], an algorithm similar to ours is setup to solve a problem of signal recovery using auto and cross-correlations instead of autoconvolutions. A few further remarks on the principal features of this paper and how it deviates from earlier research follow. Although, as said above, [3] already contains the algorithm studied in the present paper, albeit in a different context, there are remarkable differences. First, in that paper the specific form of the algorithm lacks motivation, whereas we *derive* the algorithm as the consequence of a lifting procedure, which forms the *raison d'être* of the present paper. Optimality properties of the algorithm, as well as many other ones, come almost for free as a bonus from our setup, and do not have to be derived by unnecessary complicated arguments. Second, we show that the limit points of the algorithm are Kuhn-Tucker points of the minimization problem. This result has been claimed in [3] too (Theorem 2) but unfortunately their proof contains an essential error, see Remark IV.9. The present paper thus sheds light on the origin of the algorithm, contributes to the understanding of its properties from a new perspective and gives original proofs of the main results.

A brief summary of the paper follows. Section II states the problem, shows the existence of a solution, and gives some of its properties. In Section III, the original problem is lifted into a higher dimensional setting, thus making it amenable to alternating minimization. The optimality properties (Pythagoras rules) of the ensuing partial minimization problems are discussed here. After that, we derive in Section IV the iterative minimization algorithm combining the solutions of the partial minimizations, and analyse the convergence properties. In particular, we show that limit points of the algorithm are Kuhn-Tucker points of the original optimization problem. In the concluding Section V, we present numerical experiments that show the quick convergence of the algorithm and corroborate the theoretical results on its asymptotic behaviour.

II. Problem Statement and Initial Results

We consider real valued signals $x : \mathbb{Z} \to \mathbb{R}$, mapping $i \mapsto x_i$, that vanish for $i < 0$, i.e., $x_i = 0$ for $i < 0$. The *support* of x is the discrete time interval $[0, n]$, where $n = \inf\{\, k : \ x_i = 0, \ \text{ for } i > k \,\}$, if the infimum is finite (and then a minimum), and $[0, \infty)$ otherwise. The autoconvolution of x is the signal $x * x$, vanishing for $i < 0$, and satisfying,

$$(x * x)_i = \sum_{j=-\infty}^{\infty} x_{i-j} x_j = \sum_{j=0}^{i} x_{i-j} x_j \,, \qquad i \geq 0. \tag{1}$$

Notice that if the support of x is finite $[0, n]$, the support of $x * x$ is $[0, 2n]$. In this case, when computing $(x * x)_i$ for $i > n$, the summation in (1) has non zero addends only in the range $i - n \leq j \leq n$, as $x_{i-j} = 0$ and $x_j = 0$ for $i - j > n$ and $j > n$ respectively. If the signal x is nonnegative, i.e. $x_i \geq 0$ for all $i \in \mathbb{Z}$, the autoconvolution (1) is too. Given a finite *nonnegative* data sequence

$$y = (y_0, \ldots, y_n),$$

the problem is finding a *nonnegative* signal x whose autoconvolution $x * x$ best approximates y. Since the signals involved are nonnegative, the approximation criterion is chosen to be the I-divergence, see [6], [7], the natural criterion in such a situation. Alternatively, one could opt for a different criterion, like least squares. But then nonnegativity should be added as a constraint, which could possibly lead to a range of technical complications when combined with the lifting device in Section III, that are absent with I-divergence optimization. The analytic tractability of using I-divergence additionally motivates to use this criterion. The I-divergence between two nonnegative vectors u and v of equal length is

$$\mathcal{I}(u, v) = \sum_i u_i \log \frac{u_i}{v_i} - u_i + v_i \,,$$

if $u_i = 0$ whenever $v_i = 0$, and $\mathcal{I}(u, v) = \infty$ if there exist an index i with $u_i > 0$ and $v_i = 0$. It is known that $\mathcal{I}(u, v) \geq 0$, with equality iff $u = v$.

Depending on the constraints imposed on the support of x, the basic problem splits into two different cases. The first case involves a full length signal $x = (x_0, \ldots, x_n)$ and produces the approximation problem specified below, where we write $x * x \in \mathbb{R}^{n+1}$ for the restriction to $[0, n]$ of the convolution $x * x$ defined in (1).

Problem II.1: Given $y \in \mathbb{R}_+^{n+1}$ minimize, over $x \in \mathbb{R}_+^{n+1} = [0, \infty)^{n+1}$,

$$\mathcal{I} = \mathcal{I}(x) := \mathcal{I}(y || x * x) = \sum_{i=0}^{n} \Big(y_i \log \frac{y_i}{(x * x)_i} - y_i + (x * x)_i \Big). \tag{2}$$

As an alternative, recalling that, in the finite case, the support of $x * x$ is twice the support of x, one can consider, when $n = 2m$, the problem of approximating the given data $y = (y_0, \ldots, y_{2m})$ with the autoconvolution $x * x$ of a signal of half length, $x = (x_0, \ldots, x_m)$. This leads to the following approximation problem.

Problem II.2: Given $y \in \mathbb{R}_+^{2m+1}$ minimize, over $x \in \mathbb{R}_+^{m+1} = [0, \infty)^{m+1}$,

$$\mathcal{I} = \mathcal{I}(x) := \mathcal{I}(y || x * x) = \sum_{i=0}^{2m} \Big(y_i \log \frac{y_i}{(x * x)_i} - y_i + (x * x)_i \Big). \tag{3}$$

Notice that if the given data are $y = (y_0, \ldots, y_n)$ with n odd, i.e. $n = 2m - 1$ for some integer $m \geq 1$, one can still pose Problem II.2 with $x \in \mathbb{R}_+^{m+1}$, simply by introducing the fictitious data point $y_{2m} = 0$. Hence in Problem II.2, without loss of generality, the number of data points will always be assumed odd, that is we assume n is even, $n = 2m$.

Note that Problem II.1, under the constraint that the support of x is $[0, m]$, where $m = \lfloor \frac{n+1}{2} \rfloor$, reduces to Problem II.2. Although the latter is a constrained version of the former problem and the approaches to their solutions are similar, the analysis and the results are very different. This paper concentrates on Problem II.2 which is easier to analyse and produces an algorithm with a much simpler structure. Problem II.1 will be investigated in a future publication.

The objective function (3) is nonconvex and nonlinear in x; the existence of a minimizer is therefore not immediately

clear. Our first result settles in the affirmative the question of the existence. The issue of uniqueness remains open, but we have evidence of the existence of multiple local minima of $\mathcal{I}(x)$. See Section V for numerical examples.

Proposition II.3: Problem II.2 admits a solution.

Proof: Let $x = x^0$ be an arbitrary vector in $\mathbb{R}_+^{m+1}$. Performing one step of Algorithm IV.1, introduced below, yields the iterate x^1 satisfying $\mathcal{I}(x^1) \leq \mathcal{I}(x)$ and $(\sum_{i=0}^m x_i^1)^2 = \sum_{i=0}^{2m} y_i$, by virtue of Proposition IV.3. The search for a minimizer can hence be limited to the compact subset $K_0 \subset \mathbb{R}_+^{m+1}$ of the x's satisfying $(\sum_{i=0}^m x_i^1)^2 = \sum_{i=0}^{2m} y_i$. Noting that $\mathcal{I}(x) = \sum_{i:y_i>0}(y_i \log \frac{y_i}{(x*x)_i} - y_i) + \sum_i (x*x)_i$, we can restrict attention even further to those x's for which $(x*x)_i \geq \varepsilon$ for all i such that $y_i > 0$, by choosing ε sufficiently small and positive. This implies that we restrict the finding of the minimizers to an even smaller compact set K_1 on which $\mathcal{I}$ is continuous. This proves the existence of a minimizer. □

A basic ingredient for the minimization of the cost (3) is its gradient, which is computed below. As a preliminary step, note that

$$\frac{\partial}{\partial x_j}(x*x)_i = \begin{cases} 2x_{i-j}, & \text{for } 0 \leq j \leq m,\ \ j \leq i \leq j+m \\ 0, & \text{otherwise}, \end{cases}$$

therefore

$$\begin{aligned} \nabla_j \mathcal{I}(x) := \frac{\partial \mathcal{I}(x)}{\partial x_j} &= \frac{\partial}{\partial x_j}\Big(\sum_{i=0}^{2m} -y_i \log(x*x)_i + (x*x)_i\Big) \\ &= 2\sum_{i=j}^{j+m}\Big(-x_{i-j}\frac{y_i}{(x*x)_i} + x_{i-j}\Big) \\ &= 2\sum_{\ell=0}^{m}\Big(-x_\ell \frac{y_{\ell+j}}{(x*x)_{\ell+j}} + x_\ell\Big). \end{aligned} \tag{4}$$

Equations (4) are highly nonlinear in x and solving the first order optimality conditions $\nabla\mathcal{I}(x) = 0$, where ∇ denotes the gradient vector, to find the stationary points of (3), will not result in analytic solutions except in trivial cases. This observation calls for a numerical approach to the optimization, which we will present in Section IV.

The following result shows a useful property of the minimizers of $\mathcal{I}(x)$.

Proposition II.4: For any $x \in \mathbb{R}^{m+1}$, it holds that

$$\sum_{i=0}^{2m}(x*x)_i = \Big(\sum_{i=0}^{2m} x_i\Big)^2. \tag{5}$$

Moreover, if $x^\star \in \mathbb{R}_+^{m+1}$ is a minimizer of Problem II.2,

$$\sum_{i=0}^{2m}(x^\star * x^\star)_i = \Big(\sum_{i=0}^{2m} x_i^\star\Big)^2 = \sum_{i=0}^{2m} y_i\,. \tag{6}$$

Proof: The identity (5) is a general property. Indeed, for any x,

$$\begin{aligned} \sum_{i=0}^{2m}(x*x)_i &= \sum_{i=0}^{2m}\sum_{j=0}^{i} x_{i-j}x_j = \sum_{j=0}^{2m}\sum_{i=j}^{2m} x_{i-j}x_j \\ &= \sum_{j=0}^{2m} x_j \sum_{i=j}^{2m} x_{i-j} = \Big(\sum_{j=0}^{m} x_j\Big)^2. \end{aligned}$$

To prove identity (6), let $x^\star$ be a minimizer of $\mathcal{I}(x)$ and define $f(\alpha) = \mathcal{I}(\alpha x^\star)$, for $\alpha > 0$. It follows that $f'(1) = 0$. A direct computation of $f'(\alpha)$ gives $f'(\alpha) = -\frac{2}{\alpha}\sum_{i=0}^{2m} y_i + 2\alpha \sum_{i=0}(x^\star * x^\star)_i$, hence $f'(1) = 0$ yields the wanted identity. □

Remark II.5: If y is strictly positive, the I-divergence in (3) vanishes if and only if $y_i = (x*x)_i$ for all $i \in [0, 2m]$. That is the (special) case where an exact solution to the deautoconvolution problem exists. Notice that this is a non generic case as the $2m+1$ equations $y_i = (x*x)_i$ in the $m+1$ variables x specify an (at most) $(m+1)$-dimensional submanifold in the data space $\mathbb{R}_+^{2m+1}$. See the example below for an illustration.

Example II.6: For $m = 1$, let $y = (y_0, y_1, y_2)$ be the given data. Setting the gradient $\nabla\mathcal{I}(x) = 0$, one gets the unique minimizer $x^\star = (x_0^\star, x_1^\star)$ as

$$x_0^\star = \frac{2y_0 + y_1}{2\sqrt{y_0+y_1+y_2}}, \quad x_1^\star = \frac{2y_2 + y_1}{2\sqrt{y_0+y_1+y_2}}.$$

One easily verifies that $x^\star$ satisfies property (6). Note that this solution, in general, does not give a perfect match; e.g. it should hold that $(x^\star * x^\star)_0 = (x_0^\star)^2 = y_0$. In fact, a necessary and sufficient condition on y that insures the existence of the exact solution, i.e. $\mathcal{I}(y||x^\star * x^\star) = 0$, is $y_1^2 = 4y_0y_2$.

Remark II.7: Problem II.2 has an interesting probabilistic interpretation when $\sum_{i=0}^{2m} y_i = 1$. The y_i can then be considered as the distribution of a random variable Y taking on $2m+1$ different values. The problem is then to find the optimal distribution of independent and identically distributed random variables X_1 and X_2 (assuming $m+1$ values) such that $Y = X_1 + X_2$. Note that Proposition II.4 guarantees that the optimal vector $x^\star$ indeed has the interpretation of a distribution. In Example II.6, with $y_0 + y_1 + y_2 = 1$, the optimal distribution is then $(x_0, x_1) = (y_0 + \frac{1}{2}y_1, y_2 + \frac{1}{2}y_1)$. As now one has $y_1 = 1 - y_0 - y_2$, it follows that $(x_0, x_1) = \frac{1}{2}(y_0 - y_2 + 1, y_2 - y_0 + 1)$ and the perfect match condition reduces to $\sqrt{y_0} + \sqrt{y_2} = 1$, in which case of course X_1 and X_2 can be thought of having a Bernoulli distribution and Y a binomial distribution.

III. Lifting and Partial Minimizations

In this section, Problem II.2 is recast as a double minimization problem by lifting it into a larger space. The ambient spaces for the lifted problem are the subsets $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{W}}$, defined below, of the set of matrices $\mathbb{R}_+^{(2m+1)\times(m+1)}$,

$$\boldsymbol{\mathcal{Y}} := \big\{\mathbf{Y} : \mathbf{Y}_{ij} = 0,\ 0 \leq i < j,\ i > j+m, \text{ and } \textstyle\sum_j \mathbf{Y}_{ij} = y_i\big\},$$

with $y = (y_0, \ldots, y_{2m}) \in \mathbb{R}_+^{2m+1}$, the given data vector, and

$$\begin{aligned} \boldsymbol{\mathcal{W}} := \big\{\mathbf{W} :\ & \mathbf{W}_{ij} = x_{i-j}x_j, \text{ if } 0 \leq j \leq m,\ j \leq i \leq j+m \\ & \mathbf{W}_{ij} = 0, \text{ otherwise}\big\}. \end{aligned}$$

The structure of the matrices in $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{W}}$ is shown below for $m = 3$,

$$\mathbf{Y}=\begin{bmatrix}\mathbf{Y}_{00} & 0 & 0 & 0\\ \mathbf{Y}_{10} & \mathbf{Y}_{11} & 0 & 0\\ \mathbf{Y}_{20} & \mathbf{Y}_{21} & \mathbf{Y}_{22} & 0\\ \mathbf{Y}_{30} & \mathbf{Y}_{31} & \mathbf{Y}_{32} & \mathbf{Y}_{33}\\ 0 & \mathbf{Y}_{41} & \mathbf{Y}_{42} & \mathbf{Y}_{43}\\ 0 & 0 & \mathbf{Y}_{52} & \mathbf{Y}_{53}\\ 0 & 0 & 0 & \mathbf{Y}_{63}\end{bmatrix},\ \mathbf{W}=\begin{bmatrix}x_0x_0 & 0 & 0 & 0\\ x_1x_0 & x_0x_1 & 0 & 0\\ x_2x_0 & x_1x_1 & x_0x_2 & 0\\ x_3x_0 & x_2x_1 & x_1x_2 & x_0x_3\\ 0 & x_3x_1 & x_2x_2 & x_1x_3\\ 0 & 0 & x_3x_2 & x_2x_3\\ 0 & 0 & 0 & x_3x_3\end{bmatrix}.$$

The interpretation is as follows. The matrices $\mathbf{Y}\in\boldsymbol{\mathcal{Y}}$ and $\mathbf{W}\in\boldsymbol{\mathcal{W}}$ have common support on the diagonal and first m subdiagonals of $\mathbb{R}_+^{(2m+1)\times(m+1)}$. The row marginal (i.e. the column vector of row sums) of any $\mathbf{Y}\in\boldsymbol{\mathcal{Y}}$ coincides with the given data vector y. The elements of the $\mathbf{W}$ matrices factorize, equivalently their row marginal is the autoconvolution of the column marginal rescaled by $1/\sum_i x_i$.

We introduce two partial minimization problems over the subsets $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{W}}$. Recall that the I-divergence between two nonnegative matrices of the same sizes $M,N\in\mathbb{R}_+^{p\times q}$ is defined as

$$\mathcal{I}(M||N):=\sum_{i,j}\Big(M_{ij}\log\frac{M_{ij}}{N_{ij}}-M_{ij}+N_{ij}\Big).$$

Problem III.1: For $\mathbf{W}\in\boldsymbol{\mathcal{W}}$, minimize $\mathcal{I}(\mathbf{Y}||\mathbf{W})$ over $\mathbf{Y}\in\boldsymbol{\mathcal{Y}}$.

Problem III.2: For $\mathbf{Y}\in\boldsymbol{\mathcal{Y}}$, minimize $\mathcal{I}(\mathbf{Y}||\mathbf{W})$ over $\mathbf{W}\in\boldsymbol{\mathcal{W}}$.

The solutions to both problems can be given in closed form.

Lemma III.3: Problem III.1 has the explicit minimizer $\mathbf{Y}^\star=\mathbf{Y}^\star(\mathbf{W})$ given by

$$\mathbf{Y}^\star_{ij}=\frac{\mathbf{W}_{ij}}{\sum_j\mathbf{W}_{ij}}\,y_i=\begin{cases}\frac{x_{i-j}x_j}{(x*x)_i}\,y_i & \text{if}\quad 0\le j\le i\le j+m\\ 0 & \text{otherwise.}\end{cases}\tag{7}$$

Moreover the *Pythagorean identity*

$$\mathcal{I}(\mathbf{Y}||\mathbf{W})=\mathcal{I}(\mathbf{Y}||\mathbf{Y}^\star)+\mathcal{I}(\mathbf{Y}^\star||\mathbf{W}),\tag{8}$$

holds for any $\mathbf{Y}\in\boldsymbol{\mathcal{Y}}$, and

$$\mathcal{I}(\mathbf{Y}^\star||\mathbf{W})=\mathcal{I}(y||x*x).\tag{9}$$

Proof: Proceed by direct computation. The Lagrangian is

$$L=\sum_{i,j}\big(\mathbf{Y}_{ij}\log\mathbf{Y}_{ij}-\mathbf{Y}_{ij}\log\mathbf{W}_{ij}-\mathbf{Y}_{ij}+\mathbf{W}_{ij}\big)-\sum_i\lambda_i\big(\sum_j\mathbf{Y}_{ij}-y_i\big),$$

therefore

$$\frac{\partial L}{\partial\mathbf{Y}_{ij}}=\log\mathbf{Y}_{ij}-\log\mathbf{W}_{ij}-\lambda_i=0,$$

yields $\mathbf{Y}_{ij}=\mathbf{W}_{ij}e^{\lambda_i}$, and imposing the marginal constraint $\sum_j\mathbf{Y}_{ij}=y_i$, one gets the asserted minimizer (7). Next, introducing the notation $\mathbf{W}_{i\,\cdot}=\sum_j\mathbf{W}_{ij}$ substitution of (7) into the RHS of (8) gives

$$\begin{aligned}&\mathcal{I}(\mathbf{Y}||\mathbf{Y}^\star)+\mathcal{I}(\mathbf{Y}^\star||\mathbf{W})\\ &=\sum_{i,j}\mathbf{Y}_{ij}\log\frac{\mathbf{Y}_{ij}}{\mathbf{Y}^\star_{ij}}-\mathbf{Y}_{ij}+\mathbf{Y}^\star_{ij}+\mathbf{Y}^\star_{ij}\log\frac{\mathbf{Y}^\star_{ij}}{\mathbf{W}_{ij}}-\mathbf{Y}^\star_{ij}+\mathbf{W}_{ij}\\ &=\sum_{i,j}\Big(\mathbf{Y}_{ij}\log\frac{\mathbf{Y}_{ij}}{\mathbf{W}_{ij}}-\mathbf{Y}_{ij}\log\frac{y_i}{\mathbf{W}_{i\,\cdot}}-\mathbf{Y}_{ij}\Big)\\ &\qquad+\Big(\frac{y_i}{\mathbf{W}_{i\,\cdot}}\mathbf{W}_{ij}\log\frac{y_i}{\mathbf{W}_{i\,\cdot}}+\mathbf{W}_{ij}\Big)\\ &=\mathcal{I}(\mathbf{Y}||\mathbf{W}),\end{aligned}$$

thus proving (8). As a byproduct of the Pythagorean identity, one finds that $\mathbf{Y}^\star$ is indeed a minimizer for Problem III.1. Finally, using $\mathbf{W}_{ij}=x_{i-j}x_j$ and $\mathbf{W}_{i\,\cdot}=(x*x)_i$, one finds that the optimal value of Problem III.1 coincides with (9). Indeed,

$$\begin{aligned}\mathcal{I}(\mathbf{Y}^\star||\mathbf{W})&=\sum_{i,j}\Big(\mathbf{W}_{ij}\frac{y_i}{\mathbf{W}_{i\,\cdot}}\log\frac{y_i}{\mathbf{W}_{i\,\cdot}}-\mathbf{W}_{ij}\frac{y_i}{\mathbf{W}_{i\,\cdot}}+\mathbf{W}_{ij}\Big)\\ &=\sum_i\Big(y_i\log\frac{y_i}{\mathbf{W}_{i\,\cdot}}-y_i+\mathbf{W}_{i\,\cdot}\Big)\\ &=\sum_i\Big(y_i\log\frac{y_i}{(x*x)_i}-y_i+(x*x)_i\Big)\\ &=\mathcal{I}(y||x*x).\end{aligned}$$

□

Remark III.4: Note that the minimizer $\mathbf{Y}^\star$ in (7) always exhibits the following symmetry

$$\mathbf{Y}^\star_{j+\ell,\ell}=\mathbf{Y}^\star_{j+\ell,j},\quad\text{for all }\ \ell,j=0,\dots,m,\tag{10}$$

i.e., for all $j=0,\dots,m$, the j-th subdiagonal of $\mathbf{Y}^\star$ and the $(\mathbf{Y}_{j,j},\dots,\mathbf{Y}_{j+m,j})^\top$ subvector of its j-th column coincide.

Lemma III.5: Problem III.2 has explicit minimizer $\mathbf{W}^\star=\mathbf{W}^\star(\mathbf{Y})$ corresponding to $x^\star_j$ as follows,

$$x^\star_j=\frac{\widehat{\mathbf{Y}}_j}{2\sqrt{\sum_{i=0}^{2m}y_i}},\qquad j=0,\dots,m,\tag{11}$$

where

$$\widehat{\mathbf{Y}}_j:=\sum_{i=0}^{m}\mathbf{Y}_{i+j,i}+\sum_{i=j}^{j+m}\mathbf{Y}_{ij},\qquad j=0,\dots,m.\tag{12}$$

Moreover the *Pythagorean identity*

$$\mathcal{I}(\mathbf{Y}||\mathbf{W})=\mathcal{I}(\mathbf{Y}||\mathbf{W}^\star)+\mathcal{I}(\mathbf{W}^\star||\mathbf{W}),\tag{13}$$

holds for any $\mathbf{W}\in\boldsymbol{\mathcal{W}}$.

Proof: Minimizing the I-divergence

$$\mathcal{I}(\mathbf{Y}||\mathbf{W})=\sum_{j=0}^{m}\sum_{i=j}^{j+m}\Big(\mathbf{Y}_{ij}\log\frac{\mathbf{Y}_{ij}}{\mathbf{W}_{ij}}-\mathbf{Y}_{ij}+\mathbf{W}_{ij}\Big),$$

with respect to $\mathbf{W}\in\boldsymbol{\mathcal{W}}$, is equivalent, since $\mathbf{W}_{ij}=x_{i-j}x_j$, to minimizing

$$\begin{aligned}F(x)&:=\sum_{j=0}^{m}\sum_{i=j}^{j+m}\Big(-\mathbf{Y}_{ij}\log(x_{i-j}x_j)+x_{i-j}x_j\Big)\\ &=\sum_{j=0}^{m}\sum_{i=j}^{j+m}\Big(-\mathbf{Y}_{ij}\log x_{i-j}-\mathbf{Y}_{ij}\log x_j+x_{i-j}x_j\Big).\end{aligned}\tag{14}$$

Applying to the first and third double sums in (14) the identity

$$\sum_{j=0}^{m}\sum_{i=j}^{j+m} a(i,j) = \sum_{j=0}^{m}\sum_{\ell=0}^{m} a(\ell+j,\ell)\,, \tag{15}$$

and recalling the definition (12), one easily finds

$$F(x) = -\sum_{j=0}^{m} \widehat{\mathbf{Y}}_j \log x_j + \Big(\sum_{j=0}^{m} x_j\Big)^2 . \tag{16}$$

The partial derivatives of F immediately follow from (16) as

$$\frac{\partial F}{\partial x_j} = -\frac{\widehat{\mathbf{Y}}_j}{x_j} + 2\sum_{\ell=0}^{m} x_\ell\,, \qquad j=0,\dots,m\,.$$

Setting $\frac{\partial F}{\partial x_j} = 0$ gives

$$x^\star_j = \frac{\widehat{\mathbf{Y}}_j}{2\sum_{\ell=0}^{m} x^\star_\ell}\,, \qquad j=0,\dots,m\,,$$

and hence by summation

$$\Big(\sum_{j=0}^{m} x^\star_j\Big)^2 = \frac{1}{2}\sum_{j=0}^{m}\widehat{\mathbf{Y}}_j = \sum_{i=0}^{2m} y_i\,. \tag{17}$$

To prove the last identity, it is sufficient to observe that (12) defines $\widehat{\mathbf{Y}}_j$ as the sum of the j-th subdiagonal and j-th column of the matrix $\mathbf{Y} \in \boldsymbol{\mathcal{Y}}$. This completes the proof of (11). To prove the Pythagorean identity (13), it is convenient to prove that $\mathcal{I}(\mathbf{Y}||\mathbf{W}) - \mathcal{I}(\mathbf{Y}||\mathbf{W}^\star) = \mathcal{I}(\mathbf{W}^\star||\mathbf{W})$, which is equivalent to

$$\sum_{j=0}^{m}\sum_{i=j}^{j+m} \mathbf{Y}_{ij} \log\frac{x^\star_{i-j}x^\star_j}{x_{i-j}x_j} = \sum_{j=0}^{m}\sum_{i=j}^{j+m} x^\star_{i-j}x^\star_j \log\frac{x^\star_{i-j}x^\star_j}{x_{i-j}x_j}\,.$$

The last identity is easily verified by direct substitution of (11) and (12) to express $x^\star_j$, and using the identity (15). Again, as a byproduct of the Pythagorean identity, one finds that $\mathbf{W}^\star$ is indeed a minimizer for Problem III.2. □

Remark III.6: Problem III.2 admits an interesting interpretation as a symmetric (constrained) rank one approximation of a given nonnegative matrix. We introduce the square matrices $\overline{Y}, \overline{W} \in \mathbb{R}^{(m+1)\times(m+1)}$, as 'rectifications' of the $\mathbf{Y}$ and $\mathbf{W}$ matrices, defined as

$$\overline{Y}_{ij} = \mathbf{Y}_{i+j,j}\,, \qquad \overline{W}_{ij} = W_{i+j,j} = x_i x_j\,.$$

Problem III.2 can be rephrased as

$$\min_{x\in\mathbb{R}^{m+1}_+} D(\overline{Y}\;||\;xx^\top)\,,$$

whose solution is attained at

$$x^\star_i = \frac{1}{2}\,\frac{\overline{Y}_{\cdot\, i} + \overline{Y}_{i\,\cdot}}{\sqrt{\sum_{i,j}\overline{Y}_{ij}}}\,.$$

In the probabilistic case ($\sum_{i,j}\overline{Y}_{ij} = 1$), the interpretation is that the best approximation of a two-dimensional distribution ($\overline{Y}$) by an i.i.d. product distribution ($xx^\top$) is attained at $x^\star$ equal to the average of the row and column marginals of $\overline{Y}$.

Remark III.7: In the next section, when considering Problem III.2, the given $\mathbf{Y} \in \boldsymbol{\mathcal{Y}}$ will always exhibit symmetry (10). When this is the case, Equation (11) for the optimal $x^\star$ simplifies considerably. Indeed, under symmetry (10), (12) becomes

$$\widehat{\mathbf{Y}}_j = \sum_{\ell=0}^{m}\mathbf{Y}_{\ell+j,\ell} + \sum_{i=j}^{j+m}\mathbf{Y}_{ij} = 2\sum_{i=j}^{j+m}\mathbf{Y}_{ij}\,, \quad j=0,\dots,m\,,$$

Equation (11) then reduces to

$$x^\star_j = \frac{1}{c}\sum_{i=j}^{j+m}\mathbf{Y}_{ij} = \frac{1}{c}\sum_{\ell=0}^{m}\mathbf{Y}_{\ell+j,j}\,, \tag{18}$$

where

$$c := \sqrt{\textstyle\sum_{i=0}^{2m} y_i} = \sum_{j=0}^{m} x^\star_j\,. \tag{19}$$

The connection between the original Problem II.2 and the two lifted minimization problems is explained in the next proposition.

Proposition III.8: The minimum value of the original Problem II.2 coincides with the double minimization Problems III.1 and III.2, i.e.

$$\min_{x\in\mathbb{R}^{m+1}_+} \mathcal{I}(y||x*x) = \min_{\mathbf{Y}\in\boldsymbol{\mathcal{Y}},\mathbf{W}\in\boldsymbol{\mathcal{W}}} \mathcal{I}(\mathbf{Y}||\mathbf{W})\,.$$

Proof: For given $x \in \mathbb{R}^{m+1}_+$, with corresponding $\mathbf{W} \in \boldsymbol{\mathcal{W}}$, and $\mathbf{Y} \in \boldsymbol{\mathcal{Y}}$ consider the optimizers $\mathbf{Y}^\star$ and $\mathbf{W}^\star$ from Lemmas III.3 and III.5 and recall Equation (9). Then $\mathcal{I}(\mathbf{Y}||\mathbf{W}) \geq \mathcal{I}(\mathbf{Y}^\star||\mathbf{W}) = \mathcal{I}(y|x*x) \geq \min_x \mathcal{I}(y||x*x)$, where the use of the minimum is justified by Proposition II.3. Taking the joint minimum on the left hand side over $\mathbf{Y}$ and $\mathbf{W}$, justified by the just cited lemmas, leads to $\min_{\mathbf{Y},\mathbf{W}} \mathcal{I}(\mathbf{Y}||\mathbf{W}) \geq \min_x \mathcal{I}(y||x*x)$. Conversely, for given $x \in \mathbb{R}^{m+1}_+$ with corresponding $\mathbf{W} \in \boldsymbol{\mathcal{W}}$, recalling again (9), one obtains

$$\begin{aligned}\mathcal{I}(y||x*x) &= \mathcal{I}(\mathbf{Y}^\star||\mathbf{W}) \\ &= \min_{\mathbf{Y}} \mathcal{I}(\mathbf{Y}||\mathbf{W}) \geq \min_{\mathbf{W}}\min_{\mathbf{Y}} \mathcal{I}(\mathbf{Y}||\mathbf{W})\,,\end{aligned}$$

which, taking the minimum x, shows that $\min_x \mathcal{I}(y||x*x) \geq \min_{\mathbf{Y},\mathbf{W}} \mathcal{I}(\mathbf{Y}||\mathbf{W})$, thus concluding the proof. □

IV. The Algorithm

This section is the core of the paper. It contains an algorithm aiming at finding a minimizer of Problem II.2, which we know to exist in view of Proposition II.3, and an analysis of its behavior.

A. Construction of the Algorithm and Basic Properties

Starting at an initial $\mathbf{W}^0 \in \boldsymbol{\mathcal{W}}$ and combining the two partial minimization problems, one produces a classic alternating minimization sequence,

$$\cdots\ \mathbf{W}^t \xrightarrow{1} \mathbf{Y}^t \xrightarrow{2} \mathbf{W}^{t+1} \xrightarrow{1} \mathbf{Y}^{t+1}\ \cdots\,, \tag{20}$$

where the superscript $t \in \mathbb{N}$ denotes the iteration step. The arrow $\xrightarrow{1}$ denotes the partial minimization Problem III.1, the matrix at the tail of the arrow is the given input, and the matrix at the head $\mathbf{Y}^t = \mathbf{Y}^\star(\mathbf{W}^t)$, is the optimal solution. The meaning of $\xrightarrow{2}$ is analogous, and represents the partial minimization Problem III.2, and $\mathbf{W}^{t+1} = \mathbf{W}^\star(\mathbf{Y}^t)$. Note

that, at each iteration, $\mathbf{W}^t$ is completely specified by the fixed data y and by the vector $x^t = (x_0^t, \dots, x_m^t) \in \mathbb{R}_+^{m+1}$. An iterative algorithm for the minimization Problem II.2, solely in terms of x^t, can be extracted from the sequence (20) as it immediately follows combining Lemmas III.3 and III.5. The update equation, say $x^{t+1} = I(x^t)$, is given below.

Algorithm IV.1. Starting from an arbitrary vector $x^0 \in \mathbb{R}_+^{m+1}$ the update equation $x^{t+1} = I(x^t)$ is given componentwise by

$$x_j^{t+1} = x_j^t \, \frac{1}{c} \sum_{\ell=0}^{m} \frac{x_\ell^t \, y_{\ell+j}}{(x^t * x^t)_{\ell+j}} \,, \qquad j = 0, \dots, m \,. \tag{21}$$

To verify (21), it is enough to shunt the $\mathbf{Y}^t$ step in the chain (20) and directly concatenate $\mathbf{W}^t$ to $\mathbf{W}^{t+1}$. Starting with (18) and recalling the expression of $\mathbf{Y}^t$ given by (7), one has

$$x_j^{t+1} = \frac{1}{c} \sum_{\ell=0}^{m} \mathbf{Y}_{\ell+j,j}^t = \frac{1}{c} \sum_{\ell=0}^{m} \frac{x_\ell^t x_j^t}{(x^t * x^t)_{\ell+j}} y_{\ell+j} \,, \tag{22}$$

which coincides with (21).

Remark IV.2: Application of Algorithm IV.1 to Example II.6 gives the exact solution in one step, starting from any initial $x_j^0 > 0$, as is easily verified. This is an exceptional case.

The portmanteau proposition below summarizes some useful properties of the algorithm.

Proposition IV.3: The iterates x^t, $t \geq 0$, of Algorithm IV.1 satisfy the following properties.

(i) If $x^0 > 0$ componentwise, then $x^t > 0$ componentwise, for all $t > 0$.

(ii) x^t belongs to the simplex $\mathcal{S} = \{x \in \mathbb{R}_+^{m+1} : \sum_{i=0}^m x_i = c\}$ for all $t > 0$.

(iii) $\mathcal{I}(y||x^t * x^t)$ decreases at each iteration, in fact one has

$$\begin{aligned} \mathcal{I}(y||x^t * x^t) - \mathcal{I}(y||x^{t+1} * x^{t+1}) \\ = \mathcal{I}(\mathbf{Y}^t||\mathbf{Y}^{t+1}) + \mathcal{I}(\mathbf{W}^{t+1}||\mathbf{W}^t) \geq 0 \,, \end{aligned} \tag{23}$$

and, as a corollary, $\mathcal{I}(\mathbf{W}^{t+1}||\mathbf{W}^t)$ vanishes asymptotically.

(iv) If $y = x^t * x^t$ then $x^{t+1} = x^t$, i.e. perfect matches are fixed points of the algorithm.

(v) The update equation (21) can be written in the form

$$x_j^{t+1} = x_j^t \Big(1 - \frac{1}{2c} \nabla_j \mathcal{I}(x^t)\Big) \,. \tag{24}$$

(vi) If $\nabla_j \mathcal{I}(x^t) = 0$ then $x_j^{t+1} = x_j^t$, and if $\nabla \mathcal{I}(x) = 0$ then $x^{t+1} = x^t$, i.e. stationary points of $\mathcal{I}(x)$ are fixed points of the algorithm.

(vii) If $\mathcal{I}(x^t)$ is increasing (decreasing) in x_j^t, then $x_j^{t+1} < x_j^t$ ($x_j^{t+1} > x_j^t$).

Proof:

(i) Obvious from (21).

(ii) Consider the first equality in (22). Summing over j gives

$$\sum_{j=0}^{m} x_j^{t+1} = \frac{1}{c} \sum_{j=0}^{m} \sum_{\ell=0}^{m} \mathbf{Y}_{\ell+j,j}^t = c \,,$$

in view of the two equalities in (18) and (19).

(iii) Combining the Pythagorean identities (8), (13) for the chain (20) one finds

$$\mathcal{I}(\mathbf{Y}^t||\mathbf{W}^t) = \mathcal{I}(\mathbf{Y}^t||\mathbf{Y}^{t+1}) + \mathcal{I}(\mathbf{Y}^{t+1}||\mathbf{W}^{t+1}) + \mathcal{I}(\mathbf{W}^{t+1}||\mathbf{W}^t),$$

from which (23) follows applying (9). The corollary is proved noting that the decreasing sequence $\mathcal{I}(y||x^t * x^t)$ certainly has a limit, therefore the LHS of the equation vanishes asymptotically and so do the terms on the RHS which are nonnegative for all $t > 0$.

(iv) In view of (ii), under the assumption (21) reduces to

$$x_j^{t+1} = x_j^t \, \frac{1}{c} \sum_{\ell=0}^{m} x_\ell^t = x_j^t \,.$$

(v) From (4) one gets

$$\nabla_j \mathcal{I}(x) = -2 \sum_{\ell=0}^{m} \frac{x_\ell \, y_{\ell+j}}{(x * x)_{\ell+j}} + 2 \sum_{\ell=0}^{m} x_\ell \,,$$

and recalling that $x^t \in \mathcal{S}$ it follows that $\nabla_j \mathcal{I}(x^t) = -2 \sum_{\ell=0}^m \frac{x_\ell^t \, y_{\ell+j}}{(x^t * x^t)_{\ell+j}} + 2\,c$. Hence, the update equation (21) can be written as in (24).

(vi), (vii) follow immediately from (v). □

The decrease of the divergence, as follows from Proposition IV.3(iii) is shared by many algorithms of alternating minimization type, also known as majorization-minimization (MM) algorithms. A very important instance is the classical EM algorithm, see [8] for the orginal, [24], or [15] for a more recent exposition. The descent property follows as soon as one can use what is a so-called *auxiliary* function. This concept has been used in a different context, in e.g. [27] and in [12] in nonnegative matrix factorization. Other, more general, references for MM-type problems are [17], [18], [25], [29] and the survey [21].

Let us now explain the concept of auxiliary function and see how to use it in the present context. Suppose one wants to minimize $f(x)$ over x in a certain domain. A function g of two variables x, x' satisfying the conditions $f(x) = g(x, x)$ for all x (tangency condition) and $f(x') \leq g(x, x')$ for all x, x' (dominance condition) is called an auxiliary function. It turns out that for $f(x) = \mathcal{I}(y||x * x)$ the function $g(x, x') = \mathcal{I}(\mathbf{Y}^*(\mathbf{W})||\mathbf{W}')$ is an auxiliary function when $\mathbf{W}$ is related to x as in the definition of the set $\boldsymbol{\mathcal{W}}$ in Section III, as we shall show now. Indeed, in view of (9) one has $\mathcal{I}(y||x * x) = g(x, x) = \mathcal{I}(\mathbf{Y}^*(\mathbf{W})||\mathbf{W})$. And by the optimizing property of $\mathbf{Y}^*(\mathbf{W})$ one has $\mathcal{I}(\mathbf{Y}^*(\mathbf{W})||\mathbf{W}') \leq \mathcal{I}(\mathbf{W}'||\mathbf{W}')$. Then the structure of the alternating minimization as presented in (20) at the beginning of Section IV-A gives

$$\begin{aligned} \mathcal{I}(y||x^{t+1} * x^{t+1}) &= \mathcal{I}(\mathbf{Y}^*(\mathbf{W}^{t+1})||\mathbf{W}^{t+1}) \\ &\leq \mathcal{I}(\mathbf{Y}^*(\mathbf{W}^t)||\mathbf{W}^{t+1}) \\ &= \mathcal{I}(\mathbf{Y}^t||\mathbf{W}^*(\mathbf{Y}^{t+1})) \\ &\leq \mathcal{I}(\mathbf{Y}^t||\mathbf{W}^t) = \mathcal{I}(Y||x^t * x^t). \end{aligned}$$

The first and last equalities in the above display follow from (9), the inequalities follow from the properties of the

optimizers $\mathbf{Y}^*$ and $\mathbf{W}^*$. Now we rewrite in terms of f and g.

$$\begin{aligned} f(x^{t+1}) &= g(x^{t+1}, x^{t+1}) \\ &\le g(x^t, x^{t+1}) \\ &\le g(x^t, x^t) = f(x^t), \end{aligned}$$

The added value of Proposition IV.3(iii) is that it gives, in addition to merely a decrease of the criterion, a quantification of how big this decrease is.

B. Convergence Analysis

The aim of this section is to investigate the behaviour of Algorithm IV.1 for large values of the iteration index t. We start with a technical lemma.

Lemma IV.4: For the iterates x^t and their corresponding $\mathbf{W}^t$, it holds that

(i) $\mathcal{I}(\mathbf{W}^{t+1}||\mathbf{W}^t) = 2c\mathcal{I}(x^{t+1}||x^t)$,
(ii) $\sum_i |x_i^{t+1} - x_i^t| \le \left(\mathcal{I}(\mathbf{W}^{t+1}||\mathbf{W}^t)\right)^{1/2}$,
(iii) $\lim_{t\to\infty} \mathcal{I}(x^{t+1}||x^t) = 0$, and hence $\sum_i |x_i^{t+1} - x_i^t| \to 0$.

Proof: To prove (i), a direct computation gives

$$\begin{aligned} \mathcal{I}(\mathbf{W}^{t+1}||\mathbf{W}^t) &= \sum_{j=0}^{m}\sum_{i=j}^{j+m}\left(\mathbf{W}_{ij}^{t+1}\log\frac{\mathbf{W}_{ij}^{t+1}}{\mathbf{W}_{ij}^t} - \mathbf{W}_{ij}^{t+1} + \mathbf{W}_{ij}^t\right) \\ &= \sum_{j=0}^{m}\sum_{i=j}^{j+m} x_{i-j}^{t+1}x_j^{t+1}\log\frac{x_{i-j}^{t+1}x_j^{t+1}}{x_{i-j}^t x_j^t} \\ &= \sum_{j=0}^{m}\sum_{\ell=0}^{m} x_\ell^{t+1}x_j^{t+1}\log\frac{x_\ell^{t+1}x_j^{t+1}}{x_\ell^t x_j^t} \\ &= 2\left(\sum_{\ell=0}^{m} x_\ell^{t+1}\right)\sum_{j=0}^{m} x_j^{t+1}\log\frac{x_j^{t+1}}{x_j^t} = 2\,c\mathcal{I}(x^{t+1}||x^t), \end{aligned} \tag{25}$$

where the last identity follows from (17).

To prove (ii), recall Pinsker's inequality which states, for probability vectors p, q, that $\sum_i |p_i - q_i| \le (2\mathcal{I}(p||q))^{1/2}$. The iterates x^t and x^{t+1} are not probability vectors in general, but both belong to the simplex $\mathcal{S}$. Therefore, by an easy corollary to Pinsker's inequality, $\sum_i |x_i^{t+1} - x_i^t| \le \left(2c\mathcal{I}(x^{t+1}||x^t)\right)^{1/2}$, from which, by direct application of (i), one gets (ii).

Finally, (iii) descends from the fact that $\mathcal{I}(\mathbf{W}^{t+1}||\mathbf{W}^t)$ vanishes asymptotically, as proved by the corollary to (23), and therefore, applying (i) again, so does $\mathcal{I}(x^{t+1}||x^t)$ and by Pinsker's inequality also $\sum_i |x_i^{t+1} - x_i^t|$. □

The existence of limit points of the sequence (x^t) of the iterates of the algorithm is obvious, as all x^t belong to the simplex $\mathcal{S}$; see Proposition IV.3, which is a compact set. Note that the sequence (x^t) depends on the initial point x^0. Changing x^0, the sequence (x^t) changes, and so do, in general, its limit points. To avoid cluttered notation, the dependence of the limit points on x^0 will not be evidenced. We continue with establishing some properties of the limit points of x^t.

Lemma IV.5: If x^∞ is a limit point of the sequence (x^t), then it is a fixed point of the algorithm, i.e.

$$x^\infty = I(x^\infty).$$

Proof: Let x^∞ be a limit point of the sequence (x^t). The map $x^{t+1} = I(x^t)$, given componentwise in (21), is continuous. Likewise, the I-divergence $\mathcal{I}(u||v)$ is jointly continuous in (u, v) for all $v > 0$. It follows that $\mathcal{I}(I(x^\infty)||x^\infty)$ is a limit point of $\mathcal{I}(x^{t+1}||x^t)$ which, by Lemma IV.4 (iii), vanishes asymptotically, implying that $\mathcal{I}(I(x^\infty)||x^\infty) = 0$, which yields $x^\infty = I(x^\infty)$, i.e. x^∞ is a fixed point of the algorithm. □

Proposition IV.6: The I-divergence $\mathcal{I}(y||x^\infty * x^\infty)$ is constant over the set of all limit points x^∞ of (x^t).

Proof: Iteration of (23) gives, for $t \le T$,

$$\begin{aligned} &\mathcal{I}(y||x^t * x^t) - \mathcal{I}(y||x^T * x^T) \\ &\quad = \sum_{k=t}^{T-1}\left(\mathcal{I}(\mathbf{Y}^{k+1}||\mathbf{Y}^k) + \mathcal{I}(\mathbf{W}^{k+1}||\mathbf{W}^k)\right). \end{aligned} \tag{26}$$

Suppose that the x^T converge along a subsequence to x^∞. Then, we also have

$$\begin{aligned} &\mathcal{I}(y||x^t * x^t) - \mathcal{I}(y||x^\infty * x^\infty) \\ &\quad = \sum_{k=t}^{\infty}\left(\mathcal{I}(\mathbf{Y}^{k+1}||\mathbf{Y}^k) + \mathcal{I}(\mathbf{W}^{k+1}||\mathbf{W}^k)\right). \end{aligned} \tag{27}$$

Suppose x' is another limit point and x^t converges to x' along a suitable subsequence indexed by t'. Taking the limit for $t = t' \to \infty$ in (27), one sees that the RHS vanishes, whereas the LHS gives $\mathcal{I}(y||x' * x') - \mathcal{I}(y||x^\infty * x^\infty)$, which is thus zero. □

Remark IV.7: Proposition IV.6 makes it clear that all limit points of x^t are equivalent, in the sense that their autoconvolutions have the same informational distance to the target y in Problem II.2. In particular, if one limit point is a minimizer, so are all other limit points.

One can show that the set of limit points of the sequence (x^t) is compact and connected. Compactness follows from Proposition IV.6 (the set of limit points is closed and contained in the simplex $\mathcal{S}$, hence bounded), whereas connectedness is essentially a consequence of Lemma IV.4. A similar statement can be found in [5].

Proposition IV.8: Limit points of the sequence (x^t) are Kuhn-Tucker points of the minimization Problem II.2.

Proof: Recall the version of the update equation of the algorithm as in (24). By Lemma IV.5, if x^∞ is a limit point of the x^t, then it is a fixed point of the algorithm, and (24) reduces to

$$x_j^\infty = x_j^\infty\left(1 - \frac{1}{2c}\nabla_j\mathcal{I}(x^\infty)\right),$$

showing that, if $x_j^\infty > 0$, then $\nabla_j\mathcal{I}(x^\infty) = 0$. To complete the verification that x^∞ satisfies the Kuhn-Tucker conditions for $\mathcal{I}(x)$, one has to check that if $x_j^\infty = 0$, then $\nabla_j\mathcal{I}(x^\infty) \ge 0$. So, we proceed with investigating limit points on the boundary. For a given initial condition x^0, let (x^t) be the sequence of iterates of the algorithm and define $O = \{x \in \mathbb{R}_+^{m+1} : \nabla_j\mathcal{I}(x) < 0\}$. Put $L_0 = 0$ and let $U_0 = \inf\{t > 0 : x^t \in O^c\}$. If $U_0 = \infty$, then all x^t belong to O and the x_j^t form an increasing sequence in view of (24), so certainly all $x_j^t > x_j^0 > 0$ and a limit point with $x_j^\infty = 0$ cannot occur. If U_0 is finite, we put $L^1 = \inf\{t > U_0 : x^t \in O\}$. If $L_1 = \infty$, then $x^t \in O^c$ for all $t \ge U_0$, so the x_j^t form a decreasing sequence, converging to some $x_j^\infty \ge 0$. With $\nabla_j\mathcal{I}(x^t) \ge 0$ for all t, then necessarily also $\nabla_j\mathcal{I}(x^\infty) \ge 0$, hence x^∞ satisfies

the Kuhn-Tucker conditions. In case $L^1 < \infty$ continue by alternating definitions, $U_1 = \inf\{t > L_1 : x^t \in O^c\}$, $L_2 = \inf\{t > U_1 : x^t \in O\}$, etc. As soon as some L_k or U_k is infinite, we are in either of the situations just described and in a limit point one necessarily has $x_j^\infty > 0$ or $x_j^\infty \geq 0$ and $\nabla_j \mathcal{I}(x^\infty) \geq 0$, satisfying the Kuhn-Tucker conditions.

As a last case, we investigate what happens if all L_k and U_k are finite and the interest is in possible boundary limit points x^∞ with $x_j^\infty = 0$. Observe that for t between the L_k and U_k, the x_j^t are increasing, and for t between the U_k and L_{k+1}, the x_j^t are decreasing. More precisely, for $L_k \leq t < U_k$, it holds that $x_j^{t+1} \leq x_j^t$, and for $U_k \leq t < L_{k+1}$, it holds that $x_j^{t+1} > x_j^t$. In particular, $x_j^{L_k} \leq x_j^{L_k - 1}$ and $x_j^{L_k} < x_j^{L_k+1}$, hence $x_j^{L_k}$ is a local minimum of the x_j^t. Suppose that x^∞ is a limit point, with $x_j^\infty = 0$. Then we have to consider the liminf of the x_j^t, which coincides with the liminf of the $x_j^{L_k}$. But, by Lemma IV.4, $x_j^{L_k - 1}$ also converges along a subsequence to the same liminf, and in these points one has $\nabla_j \mathcal{I}(x^{L_k - 1}) \geq 0$. Hence, along any convergent subsequence of the x^t with $\liminf x_j^t = 0$, one necessarily has $\nabla_j \mathcal{I}(x^\infty) \geq 0$. As a side remark, in this last case, since $\nabla_j \mathcal{I}(x^{L_k}) < 0$ for all k, one finds in fact $\nabla_j \mathcal{I}(x^\infty) = 0$. □

Remark IV.9: The assertion of Proposition IV.8 has also been claimed in [3, Theorem 2]. Unfortunately, their proof contains an essential flaw in the reasoning. It assumes in their Equation (39) that $\hat{x}^*$ is a *true limit* of the algorithm, not merely a limit point. This assumption is needed in their proof to conclude that, in their notation, $r^* \leq 1$ if $\hat{x}^* = 0$. The reasoning leading to that conclusion is only correct if indeed $\hat{x}^*$ is a limit, but not if it is only known that $\hat{x}^*$ is a limit point (a limit along a subsequence).

C. Convergence Properties, Further Considerations

All empirical examples suggest that the iterates of Algorithm IV.1 converge to a limit. Although a full proof cannot be given, a number of considerations make this result plausible, also from a theoretical point of view.

On a technical note, in order to prove that the algorithm converges, one would need to show that $\mathcal{I}(x^\infty || x^t)$ is decreasing in t, for any limit point x^∞. The proof of this property would go along the arguments of Lemma A.1 of [30] or Lemma 24 in [13], if one could prove that, in our notation, $\mathcal{I}(\mathbf{W}^\infty || \mathbf{W}^t) \leq c\mathcal{I}(x^\infty || x^t)$. Unfortunately it is only possible to prove the looser inequality $\mathcal{I}(\mathbf{W}^\infty || \mathbf{W}^t) \leq 2c\mathcal{I}(x^\infty || x^t)$. The factor 2 essentially appears as a consequence of the 'quadratic nature in x' of the autoconvolution terms $(x * x)_i$ whereas terms of type $(u * x)_i$, appearing in the context of e.g. [13] or [14], are linear in x. Consequently, one cannot conclude that the x^t converge to a global minimizer. For completeness we present, in Proposition IV.10, the proof of convergence of the algorithm under the proviso, empirically satisfied in all cases, that $\mathcal{I}(x^\infty || x^t)$ decreases in t. In the simulations in Section V we shall see an example where convergence of the x^t occurs, but not to a global minimizer of $\mathcal{I}(x)$.

Proposition IV.10: Let x^∞ be a limit point of the sequence (x^t) and assume that $\mathcal{I}(x^\infty || x^t)$ is decreasing in t. Then x^t converges to x^∞, which is the unique limit point of x^t.

Proof: By Proposition IV.3(ii), the x^t belong to $\mathcal{S}$ and therefore, along some subsequence, $x^{t_k} \to x^\infty$, for some limit point $x^\infty \in \mathcal{S}$. By continuity, $\mathcal{I}(x^\infty || x^{t_k}) \to 0$. On the other hand, as the divergences $\mathcal{I}(x^\infty || x^t)$ are decreasing, it must hold that $\mathcal{I}(x^\infty || x^t) \to 0$. Using Pinsker's inequality as in the proof of Lemma IV.4, $\sum_i |x_i^\infty - x_i^t| \leq (2c\mathcal{I}(x^\infty || x^t))^{1/2}$, one concludes that $x^t \to x^\infty$, and hence that x^∞ is the unique limit point. □

Next to the empirically observed behavior in Proposition IV.10, we present another argument for convergence based on an element of Morse theory, for which we need the Hessian of the criterion $\mathcal{I}(x)$. Differentiate $\frac{\partial \mathcal{I}(x)}{\partial x_j}$ as given by (4) w.r.t. x_i to get

$$H_{ij}(x) := -2\frac{y_{i+j}}{(x * x)_{i+j}} + 4\sum_{l=0}^{2m} \frac{y_{l+j}}{(x * x)_{l+j}^2} x_l x_{l+j-i} + 2.$$

Note that effectively the index l in the summation runs from $\max\{i-j, 0\}$ to $m + \min\{i-j, 0\}$, because of our convention $x_\ell = 0$ for $\ell < 0$ or $\ell > m$. The expression for $H_{ij}(x)$ can be rewritten as

$$H_{ij}(x) := -2\frac{y_{i+j}}{(x * x)_{i+j}} + 4\sum_{k=0}^{2m} \frac{y_k}{(x * x)_k^2} x_{k-j} x_{k-i} + 2,$$

with the same conventions as for the previous display. Effectively, the index k in the summation runs from $\max\{i, j\}$ to $m + \min\{i, j\}$.

Let $S^{(k)} \in \mathbb{R}^{(m+1)\times(m+1)}$, for $k \in \{0, \ldots, 2m\}$, be defined by $S_{ij}^{(k)} = \delta_{k,i+j}$ for $i, j \in \{0, \ldots, m\}$, where the δ's are Kronecker δ's. Let furthermore $x = (x_0, \ldots, x_m)^\top$ and $\xi_k = S^{(k)} x$. Define $P(x) \in \mathbb{R}^{(m+1)\times(m+1)}$ with elements $P_{ij}(x) = 4\sum_{k=0}^{2m} \frac{y_k}{(x*x)_k^2} x_{k-j} x_{k-i}$, then one can write

$$P(x) = 4\sum_{k=0}^{2m} \frac{y_k}{(x * x)_k^2} \xi_k \xi_k^\top.$$

Note that, if $x_0 > 0$, the $\{\xi_k\}_{k=0}^m$ form a basis of $\mathbb{R}^{m+1}$, therefore if $y_k > 0$ for $k \in [0, m]$, the matrix $P(x)$ is strictly positive definite. Alternatively, if $x_m > 0$, the $\{\xi_k\}_{k=m}^{2m}$ also form a basis of $\mathbb{R}^{m+1}$ and again, if $y_k > 0$ for $k \in [m, 2m]$, the matrix $P(x)$ is strictly positive definite. Furthermore, let $Q(x) \in \mathbb{R}^{(m+1)\times(m+1)}$, with elements $Q_{ij}(x) = 2 - 2\frac{y_{i+j}}{(x*x)_{i+j}}$. Hence, the Hessian $H(x)$ satisfies

$$H(x) = P(x) + Q(x).$$

Note that $Q(x)$ vanishes if $y_i = (x*x)_i$, for all $i \in [0, 2m]$, i.e. in the exact model case, making $H(x)$ strictly positive definite. To find a useful expression of the Hessian in the general case, introduce the matrix $R(x) \in \mathbb{R}^{(m+1)\times(m+1)}$ with elements $R_{ij}(x) = \frac{y_{i+j}}{(x*x)_{i+j}}$, and note that $Q(x) = 2\left(\mathbf{1}\mathbf{1}^\top - R(x)\right)$, then

$$H(x) = P(x) + 2\left(\mathbf{1}\mathbf{1}^\top - R(x)\right),$$

moreover the gradient $\nabla\mathcal{I}(x)$, written as a row vector, is

$$\nabla\mathcal{I}(x) = x^\top Q(x) = 2x^\top\left(\mathbf{1}\mathbf{1}^\top - R(x)\right).$$

Except in the special case of an exact model, it is not obvious that in an interior limit point x^∞ of the algorithm the

Hessian $H(x^\infty)$ is strictly positive definite. Even the weaker statement that $H(x^\infty)$ is non-singular is hard to prove, in spite of the rather explicit form of $H(x^\infty)$ and the fact that the gradient $\nabla\mathcal{I}(x^\infty)$ vanishes. The relevance of non singularity stems from the Morse lemma, Corollary 2.3 in [23], which states that the interior critical points of a function where the Hessian in is non singular are isolated.

Let us now look at a boundary (local) optimizer $x^\star$ of $\mathcal{I}(x)$. By the Kuhn-Tucker conditions, if $x^\star_j = 0$, then $\nabla_j\mathcal{I}(x^\star) \geq 0$, while if $x^\star_j > 0$, then $\nabla_j\mathcal{I}(x^\star) = 0$. Write the boundary optimizer $x^\star$ as $x^\star = (\underline{x}^\star, 0)$, possibly after a permutation of the coordinates, with all elements of $\underline{x}^\star$ strictly positive. We now look at optimization of $\mathcal{I}(x)$ under the constraint that $x = (\underline{x}, 0)$, so of $\underline{\mathcal{I}}(\underline{x}) := \mathcal{I}(\underline{x}, 0)$. The optimizing $\underline{x}^\star$ is now an interior point of the restricted domain, hence the gradient vanishes, $\nabla\underline{\mathcal{I}}(\underline{x}^\star) = 0$. The Hessian $\underline{H}(\underline{x}^\star)$ of $\underline{\mathcal{I}}(\underline{x})$ is strictly positive definite and certainly non-singular, and likely the same is true for $\underline{H}(\underline{x}^\infty)$ for any limit point $(\underline{x}^\infty, 0)$ of x^t. The arguments underlying this are similar to the above, although it is hard to give a proof. Again by the Morse lemma, the critical points of $\underline{\mathcal{I}}(\underline{x})$, which are now interior points of the restricted domain, will then be isolated.

Proposition IV.11: Let x^0 be a strictly positive starting point of the algorithm and let $L(x^0)$ be the set of interior limit points produced by the algorithm and assume that $H(x)$ is non-singular for all $x \in L(x^0)$. Then $L(x^0)$ is a singleton and thus the algorithm converges to a limit (possibly depending on the starting value x^0). The situation is analogous for boundary limit points. In both cases the limit is a Kuhn-Tucker point.

Proof: By Remark IV.7, the set $L(x^0)$ is connected. By the above discussion the interior limit points are isolated and the same holds for the limit points on the boundary. The combination of these two properties yields that $L(x^0)$ has to be a singleton, and hence there is convergence of x^t to the (unique) limit. Its Kuhn-Tucker property follows from Proposition IV.8. □

Remark IV.12: In the literature, it is not uncommon to see situations where the limit points are isolated. For instance, along different lines, in [19] and [20] it is shown that in their setting the set of limit points of the iterates is finite, which is a consequence of the maximization of a concave objective function. As the objective function in our minimization problem is not convex, their arguments cannot be used here.

Remark IV.13: In principle, the algorithm produces limit points that depend on the initial values x^0. This has been observed in several numerical experiments. In fact, different x_0's may either result in a limit point in the interior or on the boundary (i.e. some of its coordinates are zero). The Kuhn-Tucker property was seen to be verified in these experiments.

To summarize the discussion of this section, it is plausible that Algorithm IV.1, given a starting value, converges to a limit. This conjecture is motivated by two considerations, for both of them there is ample numerical evidence. The first is a decreasing criterion, which Proposition IV.10 handles, and the second is non-singularity of the Hessian in limit points. Yet, a formal proof of the conjecture is lacking and we currently have to content ourselves with the Kuhn-Tucker property of limit points as in Proposition IV.8.

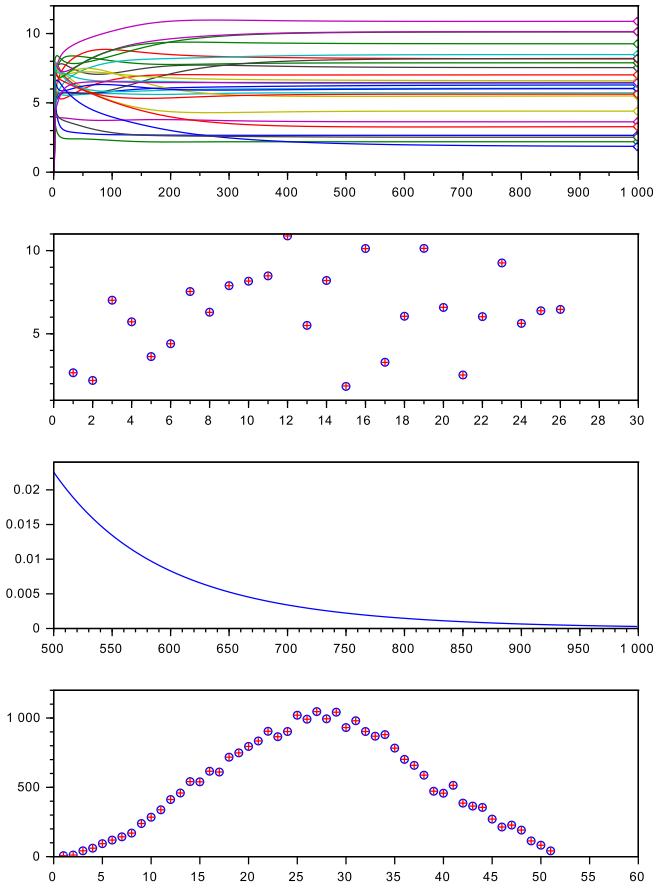

Fig. 1. True model, $m = 25$ and $T = 1000$. Top panel: $m+1$ components x_i^t against iteration t; the diamonds at $T = 1000$ are the true values x_i to which the x_i^t converge. Second panel: x_i^T (plusses) and true values x_i (circles) against i. Third panel: $\mathcal{I}(y||x^t * x^t)$ against t. Fourth panel: y_i (circles) and $(x^T * x^T)_i$ (plusses) against i.

V. Numerical Experiments

This section reviews the results of numerical experiments for three different data sets to illustrate the behaviour of Algorithm IV.1. For the first two data sets, with $m = 25$ and $m = 10$ respectively, we investigated whether the algorithm is capable of retrieving the true parameter vector x, when the data y are actually generated by the autoconvolution $y = x*x$. In the third data set the y are randomly generated, with $m = 10$.

To evaluate the performance of the algorithm we have generated, for each data set, one figure comprising three or four graphs. In all of the figures the top graph shows, in distinct colors, the trajectories of the iterates of the components, x_i^t, plotted against the iteration number $t \in [1, T]$.

In the exact model case, Figures 1, 2, 3, the diamonds at the right end of the top graph show the true x_i values. The second graph shows the superimposed plots of the data generating signal x, and of the reconstructed signal x^T, at the last iteration, both plotted against their component number (in the figure labelled $i = 1, 2, \ldots, m+1$ instead of $i = 0, 1, \ldots, m$).

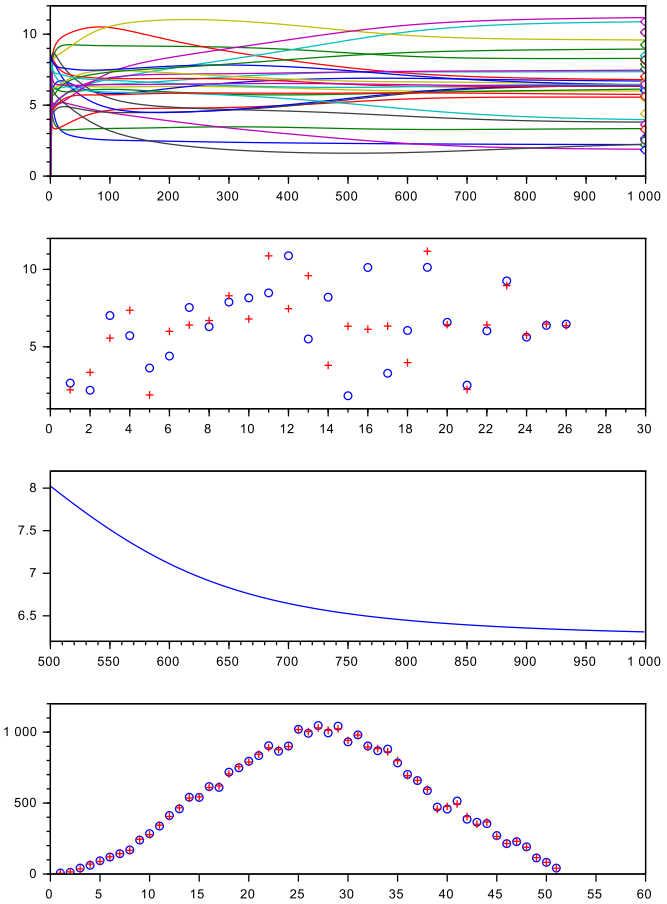

Fig. 2. The same data as in Figure 1, with different initial conditions x^0.

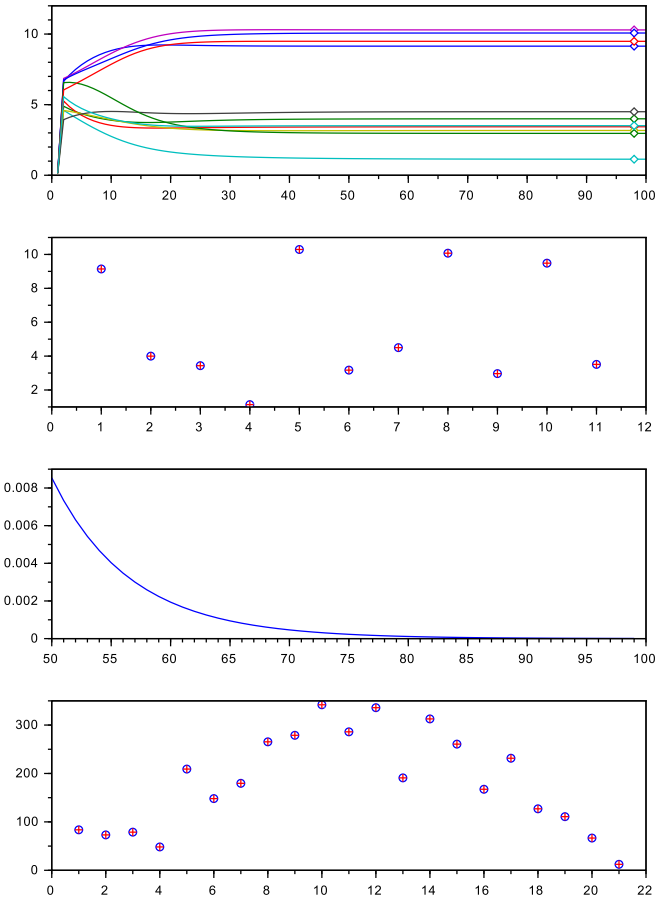

Fig. 3. A true model with $m = 10$ and $T = 100$. The panels are as in Figure 1 and the same conclusions can be drawn.

The third graph shows the decreasing sequence $\mathcal{I}(y||x^t * x^t)$. The fourth and last graph shows the superimposed plots of the data vector y and of the reconstructed convolution $x^T * x^T$, at the last iteration, both against the component number (labeled $i = 1, \ldots 2m + 1$).

Figures 4 and 5, relative to the randomly generated data set, contain only three graphs, as the graph of the data generating signal is meaningless in this case.

We have observed experimentally that the iterative algorithm always converges very fast. The precise features underlying the experiments are further detailed below.

A. True Autoconvolution Systems

For the first data set we have taken $m = 25$. The components of the true vector x (the target values of the algorithm) have been randomly generated from a uniform distribution on the interval $[1, 11]$, and the data computed as true autoconvolutions $y = x * x$. The algorithm has been initialized at a randomly chosen strictly positive x^0, with components generated from a uniform distribution in the interval $[0.1, 0.2]$ and run for $T = 1000$ iterations. Figures 1 and 2 show the results for two different runs (i.e. with the same true vector but different initial conditions) of the algorithm. In Figure 1, we see the desired behavior of the algorithm; the iterates converge to the true values and the divergence decreases to zero (because of the perfect match of $y = x * x$). This is the behavior that has been observed in a vast majority of numerical experiments of this kind. In Figure 2 we observe a different behavior. The iterates do not converge to the true values (see the second graph) and the divergence does not decrease to zero. On the other hand the convolution $x^T * x^T$ is always close to y (see the fourth graphs of both figures). In fact, the instance of running the algorithm that produced Figure 2 produced iterates that converged to a non-optimal local minimum of the objective function $\mathcal{I}(x)$. Indeed, we have verified that the gradient of $\mathcal{I}(x)$ at the final iteration vanished, whereas the Hessian turned out to be strictly positive definite. The conclusion of these two experiments is that it is wise to run the algorithm for the same data y, and same x, with different initial conditions and select the outcome with the lowest divergence. For the present example, the lowest divergence is of course zero, but the conclusion is also valid for any instance with any data vector y. The data set used to generate Figure 3 is again of the exact type, $y = x * x$, with $m = 10$ and consequently a lower number of iterations, $T = 100$. We see quick convergence of the algorithm; stabilization has already occurred at $t = 30$. The general behavior is identical to that observed in Figure 1.

As a closing remark to this section we emphasize that in the chosen examples the true autoconvolutional data have been observed without errors. In a more realistic situation it is conceivable that data are observed subject to noise. This would lead to a statistical analysis, possibly also pointing at

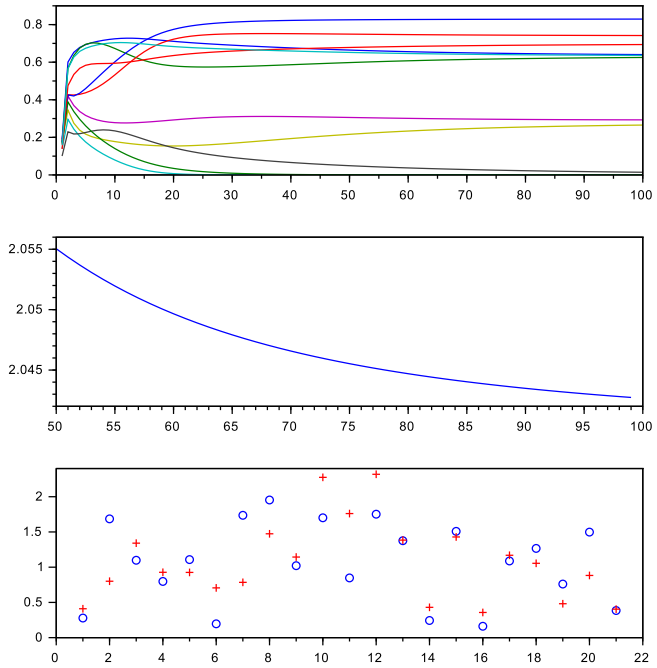

Fig. 4. Randomly generated y, with $m = 10$ and $T = 100$. Top panel: components x_i^t against iteration index t. Second panel: $\mathcal{I}(y||x^t * x^t)$ against t. Third panel: y_i (circles) and final autoconvolutions $(x^T * x^T)_i$ (plusses) against i. Third panel: The values of the y_i (circles) and the final autoconvolutions $(x^T * x^T)_i$ (plusses).

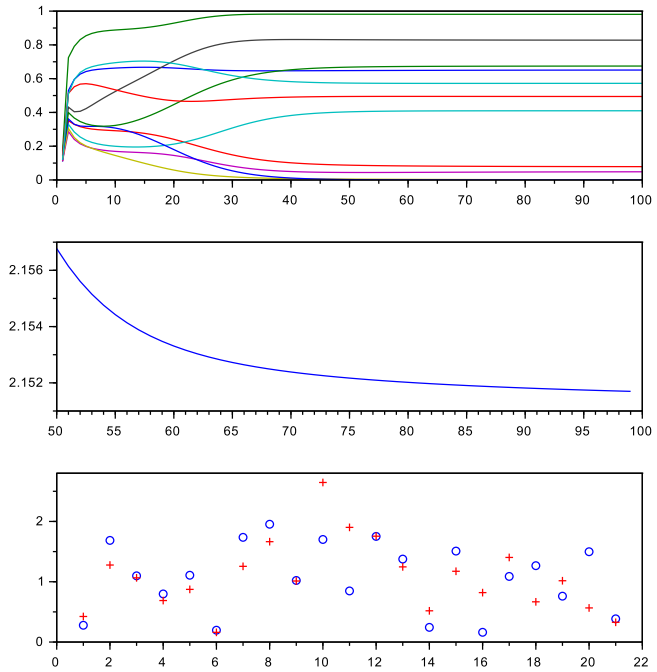

Fig. 5. The same data as in Figure 4, with different initial conditions x^0.

robustness issues. Investigations in these directions will be deferred to future research.

B. Approximation of Arbitrary Data

For the third data set there is no true input signal x such that $y = x * x$, rather the components of the data vector y, with $m = 10$, have been randomly generated from a uniform distribution on the interval $[0.1, 2]$. Thus, here we deal with a genuine approximation problem. Figures 4 and Figure 5 show the results of two runs of the algorithm, for $T = 100$ iterations, and are relative to the same y vector and different initial conditions x^0, both with components randomly generated from a uniform distribution in $[0.1, 0.2]$. The aim is to find the vector x which yields the best autoconvolutional approximation to y. Inspecting the figures we conclude that the algorithm quickly stabilises in both runs. The final values x^T of the iterates and the final divergences $\mathcal{I}(y||x^* x^T)$ differ in the two runs, indicating that (at least) in the second case (with divergence slightly higher than in the first case) the algorithm is trapped in a non-optimal local minimum. For the same y several other runs have produced results that were nearly identical to those in Figure 4, so we infer that this figure represents the optimal approximation of y. The observed behavior suggests again to run the algorithm with different initial conditions, possibly in parallel, and to select the best final approximation as the one with smallest divergence $\mathcal{I}(y||x^T * x^T)$.

ACKNOWLEDGMENT

The authors would like to thank the constructive criticism of the referees. Their useful suggestions have lead to improvements of their paper.

REFERENCES

[1] B. D. O. Anderson, M. Deistler, L. Farina, and L. Benvenuti, "Nonnegative realization of a linear system with nonnegative impulse response," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 43, no. 2, pp. 134–140, Feb. 1996.

[2] L. Benvenuti and L. Farina, "A tutorial on the positive realization problem," *IEEE Trans. Autom. Control*, vol. 49, no. 5, pp. 651–664, May 2004.

[3] K. Choi and A. D. Lanterman, "An iterative deautoconvolution algorithm for nonnegative functions," *Inverse Problems*, vol. 21, no. 3, p. 981, 2005.

[4] K. Choi, A. D. Lanterman, and R. Raich, "Convergence of the Schulz–Snyder phase retrieval algorithm to local minima," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 23, no. 8, pp. 1835–1845, 2006.

[5] T. M. Cover, "An algorithm for maximizing expected log investment return," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 2, pp. 369–373, Mar. 1984.

[6] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, 1975.

[7] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, vol. 19, no. 4, pp. 2032–2066, 1991.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B, Methodol.*, vol. 39, pp. 1–38, Feb. 1977.

[9] V. Dose, T. Fauster, and H.-J. Gossmann, "The inversion of autoconvolution integrals," *J. Comput. Phys.*, vol. 41, no. 1, pp. 34–50, May 1981.

[10] S. C. Douglas and D. P. Mandic, "Autoconvolution and panorama: Augmenting second-order signal analysis," in *Proc. IEEE ICASSP Conf.*, May 2014, pp. 384–388.

[11] L. Farina and L. Benvenuti, "Positive realizations of linear systems," *Syst. Control Lett.*, vol. 26, no. 1, pp. 1–9, Sep. 1995.

[12] L. Finesso and P. Spreij, "Nonnegative matrix factorization and I-divergence alternating minimization," *Linear Algebra Appl.*, vol. 416, nos. 2–3, pp. 270–287, Jul. 2006.

[13] L. Finesso and P. Spreij, "Approximation of nonnegative systems by finite impulse response convolutions," *IEEE Trans. Inf. Theory*, vol. 61, no. 8, pp. 4399–4409, Aug. 2015.

[14] L. Finesso and P. Spreij, "Approximation of nonnegative systems by moving averages of fixed order," *Automatica*, vol. 107, pp. 1–8, Sep. 2019.

[15] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.
[16] B. Hofmann and L. von Wolfersdorf, "On the determination of a density function by its autoconvolution coefficient," *Numer. Funct. Anal. Optim.*, vol. 27, nos. 3–4, pp. 357–375, Aug. 2006.
[17] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statist.*, vol. 58, no. 1, pp. 30–37, Feb. 2004.
[18] K. Lange, *MM Optimization Algorithms*. Philadelphia, PA, USA: SIAM, 2016.
[19] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comput. Assist. Tomogr.*, vol. 8, no. 2, pp. 306–316, 1984.
[20] K. Lange and J. A. Fessler, "Globally convergent algorithms for maximum *a posteriori* transmission tomography," *IEEE Trans. Image Process.*, vol. 4, no. 10, pp. 1430–1438, Oct. 1995.
[21] K. Lange, J.-H. Won, A. Landeros, and H. Zhou, "Nonconvex optimization via MM algorithms: Convergence theory," 2021, *arXiv:2106.02805*.
[22] V. Martinez, "Global methods in the inversion of a self-convolution," *J. Electron Spectrosc. Rel. Phenomena*, vol. 17, no. 1, pp. 33–43, Jan. 1979.
[23] J. Milnor, "Morse theory," in *Annals of Mathematics Studies*, no. 51. Princeton, NJ, USA: Princeton Univ. Press, 1963.
[24] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. Dordrecht, The Netherlands: Springer, 1998, pp. 355–368.
[25] H. D. Nguyen, "An introduction to majorization-minimization algorithms for machine learning and statistical estimation," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 7, no. 2, 2017, Art. no. e1198.
[26] T. J. Schulz and D. G. Voelz, "Signal recovery from autocorrelation and cross-correlation data," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 22, no. 4, pp. 616–624, 2005.
[27] H. S. Seung and D. D. Lee, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2001, pp. 556–562.
[28] D. L. Snyder, T. J. Schulz, and J. A. O'Sullivan, "Deblurring subject to nonnegativity constraints," *IEEE Trans. Signal Process.*, vol. 40, no. 5, pp. 1143–1150, May 1992.
[29] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.
[30] Y. Vardi, L. A. Shepp, and L. Kaufman, "A statistical model for positron emission tomography," *J. Amer. Stat. Assoc.*, vol. 80, no. 389, pp. 8–37, 1985.

Lorenzo Finesso received the laurea degree (cum laude) in electrical engineering from the University of Padua, in 1979, and the M.Sc. and Ph.D. degrees in EE from the University of Maryland (UMCP), in 1987 and 1990, respectively. Since 1984, he has been a Researcher with the Italian National Research Council, formerly with LADSEB, then with IEIIT. His main scientific interests include realization, approximation, and estimation of finite stochastic systems and of systems with positivity constraints. He and Peter Spreij have developed Csiszàr-Tusnàdy type algorithms to solve a wide variety of signal and system problems: nonnegative matrix factorizations (NMF), approximate realization of hidden Markov models (HMM), approximate factor analysis, and approximations of nonnegative systems, by FIR and by MA of fixed order.

Peter Spreij received the M.Sc. degree in mathematics from the Vrije Universiteit Amsterdam in 1979 and the Ph.D. degree from Twente University in 1987. He has been a Researcher with CWI, Amsterdam, and the Department of Econometrics, Vrije Universiteit. Since 1999, he has been a Researcher with the Korteweg-de Vries Institute for Mathematics, Universiteit van Amsterdam, and the Institute for Mathematics, Astrophysics and Particle Physics, Radboud University, Nijmegen, since 2015. He has published on a variety of topics in stochastic systems theory, asymptotic statistics, Bayesian statistics, theory of stochastic processes, matrix theory, mathematical finance, and information theory.