Low-rank tensor recovery using sequentially optimal modal projections in iterative hard thresholding (SEMPIHT)

José Henrique de Morais Goulart, Gérard Favier

SIAM J. SCI. COMPUT.
Vol. 39, No. 3, pp. A860–A889

LOW-RANK TENSOR RECOVERY USING SEQUENTIALLY OPTIMAL MODAL PROJECTIONS IN ITERATIVE HARD THRESHOLDING (SEMPIHT)*

JOSÉ HENRIQUE DE MORAIS GOULART† AND GÉRARD FAVIER‡

Abstract. Iterative hard thresholding (IHT) is a simple and effective approach to parsimonious data recovery. Its multilinear rank (mrank)-based application to low-rank tensor recovery (LRTR) is especially valuable given the difficulties involved in this problem. In this paper, we propose a novel IHT algorithm for LRTR, choosing sequential per-mode SVD truncation as its thresholding operator. This operator is less costly than those used in existing IHT algorithms for LRTR, and often leads to superior performance. Furthermore, by exploiting the sequential optimality of the employed modal projections, we derive recovery guarantees relying on restricted isometry constants. Though these guarantees are suboptimal, our numerical studies indicate that a quasi-optimal number of Gaussian measurements suffices for perfect data reconstruction. We also investigate a continuation technique which yields a sequence of progressively more complex estimated models until attaining a target mrank. When recovering real-world data, this strategy stabilizes the estimation error and can also accelerate convergence. In tensor completion, in particular, it can cope with nonideal characteristics of the sensed tensors and so is crucial for achieving a satisfactory performance. Extensive numerical experiments are reported, including the completion of hyperspectral imaging data and comparisons with several other existing approaches.

Key words. low-rank tensor recovery, tensor completion, multilinear rank, iterative hard thresholding, sequentially optimal modal projections, hyperspectral image reconstruction

AMS subject classifications. 15A69, 90C59

DOI. 10.1137/16M1062089

1. Introduction.

We consider the recovery of tensors lying in $\mathbb{R}^{N_1\times\cdots\times N_P}$ (with $P>2$) from *undercomplete linear measurements*, assuming the corresponding tensors have low-rank properties. This problem, called *low-rank tensor recovery* (LRTR), is an extension of the well-studied low-rank matrix recovery (LRMR) problem [4]. Essentially, in the tensor setting, one wishes to exploit some joint low dimensionality along multiple *modes* (i.e., geometric dimensions) of a data tensor in order to reconstruct it from a few measurements.

We assume the reader is familiar with basic tensor algebra concepts and notation (see, e.g., [22, 29]). The following notational conventions are adopted: $[\![P]\!] \triangleq \{1,\ldots,P\}$, $\bar{N} \triangleq \prod_{p=1}^{P} N_p$, and $\bar{N}_p \triangleq \bar{N}/N_p$. Throughout the text, we shall identify tensors in $\mathbb{R}^{N_1}\otimes\cdots\otimes\mathbb{R}^{N_P}$, where $\otimes$ denotes the tensor product, with P-way arrays (hypermatrices) in $\mathbb{R}^{N_1\times\cdots\times N_P}$, assuming the coordinates are given with respect to known (given) bases.

The LRTR problem is formulated here as follows:

$$\min_{\boldsymbol{\mathcal{X}}\in\mathcal{L}_{\mathbf{r}}} \|\mathbf{y}-\mathscr{A}(\boldsymbol{\mathcal{X}})\|_2^2\,, \tag{1}$$

*Submitted to the journal's Methods and Algorithms for Scientific Computing section February 18, 2016; accepted for publication (in revised form) February 7, 2017; published electronically May 23, 2017.

http://www.siam.org/journals/sisc/39-3/M106208.html

Funding: The first author's work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, under the program Ciência sem Fronteiras.

†Univ. Grenoble Alpes, CNRS, GIPSA-Lab, F-38000 Grenoble, France (jose-henrique.de-morais-goulart@gipsa-lab.fr).

‡I3S Laboratory, CNRS, Univ. Côte D'Azur, 06900 Sophia-Antipolis, France (favier@i3s.unice.fr).

where $\mathbf{r} = (R_1, \ldots, R_P) \in \mathbb{Z}_+^P$, $\mathscr{A} : \mathbb{R}^{N_1 \times \cdots \times N_P} \mapsto \mathbb{R}^M$ is a linear *measurement operator* (MO), with $M < \bar{N}$, and

$$\mathcal{L}_{\mathbf{r}} = \left\{ \boldsymbol{\mathcal{X}} \in \mathbb{R}^{N_1 \times \cdots \times N_P} : \operatorname{rank}\left(\mathbf{X}_{\langle p \rangle}\right) \leq R_p,\ p \in [\![P]\!] \right\}, \tag{2}$$

with $\mathbf{X}_{\langle p \rangle} = (\boldsymbol{\mathcal{X}})_{\langle p \rangle} \in \mathbb{R}^{N_p \times \bar{N}_p}$ denoting the mode-p matrix unfolding[1] of $\boldsymbol{\mathcal{X}}$. The quantity $\rho \triangleq M/\bar{N}$ is called (measurement) undersampling. In its most general form, the vector of measurements $\mathbf{y} \in \mathbb{R}^M$ is given by $\mathbf{y} = \mathscr{A}(\boldsymbol{\mathcal{X}}^\star) + \mathbf{e}$ for some error vector $\mathbf{e}$ and a tensor of interest $\boldsymbol{\mathcal{X}}^\star$ satisfying either $\boldsymbol{\mathcal{X}}^\star \in \mathcal{L}_{\mathbf{r}}$ or $\boldsymbol{\mathcal{X}}^\star \approx \boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} \in \mathcal{L}_{\mathbf{r}}$ (where proximity is in the Euclidean distance sense). As in most works which deal with LRTR, our formulation (1) is based on the *multilinear rank*[2] (mrank), defined as [11, 22]

$$\operatorname{mrank}(\boldsymbol{\mathcal{X}}) = \left(\operatorname{rank}\left(\mathbf{X}_{\langle 1 \rangle}\right), \ldots, \operatorname{rank}\left(\mathbf{X}_{\langle P \rangle}\right)\right). \tag{3}$$

Note that the components of $\operatorname{mrank}(\boldsymbol{\mathcal{X}})$ are a generalization of the row and column ranks of a matrix. Yet, they do not necessarily have the same value.

The reader might wonder why one would rely on the mrank rather than on the tensor rank [25], which is often regarded as the most natural extension of matrix rank to higher-order tensors. This is basically due to computational and analytical difficulties which arise when dealing with the tensor rank [12, 24]. Nevertheless, it should be noted that ways of circumventing these difficulties have been recently studied in [7, 39, 50], pointing at interesting research directions. At any rate, since $\operatorname{rank}(\mathbf{X}_{\langle p \rangle})$ is majorized by the tensor rank for all $p \in [\![P]\!]$ [11], tensors with sufficiently[3] low rank necessarily have low mrank components.

A central feature of every parsimonious data recovery problem, such as compressive sensing (CS) [6] or LRMR, is its underlying model, characterized by a number of degrees of freedom (DOFs) smaller than its algebraic (ambient) dimension. Tensors having mrank $\mathbf{r}$ belong to a manifold of dimension [31]

$$\Phi(\mathbf{r}) \triangleq \textstyle\prod_p R_p + \sum_p R_p(N_p - R_p) = \mathcal{O}(\prod_p R_p + \sum_p R_p N_p), \tag{4}$$

which is much smaller than $\bar{N} = \dim(\mathbb{R}^{N_1 \times \cdots \times N_P})$ for low values of R_p. A tensor $\boldsymbol{\mathcal{X}}^\star = (x^\star_{n_1,\ldots,n_P})$ is in $\mathcal{L}_{\mathbf{r}}$ if and only if it can be written as a Tucker model [47],

$$\boldsymbol{\mathcal{X}}^\star = \boldsymbol{\mathcal{G}} \times_1 \mathbf{U}^{(1)} \times_2 \cdots \times_P \mathbf{U}^{(P)} \quad \Leftrightarrow \quad x^\star_{n_1,\ldots,n_P} = \sum_{r_1} \cdots \sum_{r_P} g_{r_1,\ldots,r_P} \prod_p u^{(p)}_{n_p,r_p}, \tag{5}$$

where $\boldsymbol{\mathcal{G}} = (g_{r_1,\ldots,r_P}) \in \mathbb{R}^{R_1 \times \cdots \times R_P}$ is the core tensor and $\mathbf{U}^{(p)} = (u^{(p)}_{n_p,r_p}) \in \mathbb{R}^{N_p \times R_p}$ is the pth matrix factor. Without loss of generality, one can constrain (5) similarly to the higher-order singular value decomposition (HOSVD) [11], requiring each $\mathbf{U}^{(p)}$ to belong to $\mathcal{V}_{R_p}(\mathbb{R}^{N_p})$, the Stiefel manifold of $N_p \times R_p$ matrices having orthonormal columns, and each mode-p unfolding of $\boldsymbol{\mathcal{G}}$ to have mutually orthogonal rows. Then, counting the parameters of (5) gives (4).

[1] For unambiguously referring to "the mode-p unfolding," one must establish an ordering convention for the columns of the resulting matrix. In our argument, such a choice is irrelevant but must be used consistently.

[2] Also called "n-rank" or "Tucker rank" [52, 17, 35].

[3] Recall that the rank of a tensor can exceed its dimensions. For instance, the smallest typical rank of an $8 \times 8 \times 8$ real tensor is 24 [9]; hence a low-rank tensor might still have high modal ranks if $8 < \operatorname{rank}(\boldsymbol{\mathcal{X}}^\star) < 24$.

Ideally, one would like to come up with a computationally efficient algorithm provably capable of recovering any $\boldsymbol{\mathcal{X}}^\star \in \mathcal{L}_\mathbf{r}$ from $M \approx \Phi(\mathbf{r})$ (sufficiently informative) measurements. Given an instance of the LRTR problem (1) with an associated pair $(\mathbf{r}, M)$, we refer to the ratio $\theta \triangleq \Phi(\mathbf{r})/M$ as its *regime*. In general, as θ decreases, successful recovery becomes more likely, and thus small values of θ correspond to favorable regimes. Conversely, the recovery performance of an algorithm typically degrades as $\theta \to 1$. Given M random measurements of a certain class (e.g., Gaussian or Bernoulli), the interval $]0, \theta_0]$ of regimes for which perfect recovery is achieved with high probability using a given algorithm is called its *recovery regime* for M with respect to this class.

The practical relevance of the LRTR problem stems from the fact that many real-world tensors can be well approximated by elements of $\mathcal{L}_\mathbf{r}$, such as three-dimensional medical images [17, 34], seismic data [30], video sequences [34, 49], and hyperspectral images [17, 42]. Of particular interest is the frequent problem of reconstructing partially observed low-mrank tensors. This problem, called *tensor completion* (TC) in analogy with matrix completion (MC) [5], is a particular case of (1) where $\mathscr{A}$ is a sampling operator (SO) which reveals only some entries of $\boldsymbol{\mathcal{X}}^\star$. Nonetheless, practical applicability of LRTR is not restricted to TC, as other measurement schemes (such as, e.g., subsampling in the frequency domain) can be implemented for acquiring a few data from which a large low-mrank tensor can then be recovered.

1.1. Overview of the state of the art. Unlike the LRMR setting, no provably efficient (in terms of sampling requirements) and tractable convex approach is currently known for LRTR. Recovery results were derived in [51] for the tensor nuclear norm, but this norm is intractable [16]. Nonetheless, several tractable LRTR approaches have been developed in recent years. In the following, we briefly describe some of them and their corresponding recovery guarantees. For simplicity of exposition, we now consider the model (5) with $N_p = N$ for all $p \in [\![P]\!]$ and mrank $\mathbf{r} = (R, \ldots, R)$, implying $\Phi(\mathbf{r}) = \mathcal{O}(R^P + PNR)$.

The first proposed LRTR algorithms [17, 33, 45] relied on minimizing or bounding a weighted sum of the nuclear norms (SNN) of the modal unfoldings. This idea was motivated by the effectiveness of nuclear norm minimization in LRMR and yields convex formulations. For these reasons, it was later employed many times, as in [34, 42, 46]. It was shown in [46] that SNN minimization succeeds when at least $\mathcal{O}(RN^{P-1})$ Gaussian measurements are taken. As argued by [35], this bound is actually sharp, while a certain (intractable) nonconvex formulation permits, in principle, perfect recovery (with $\boldsymbol{\mathcal{X}}^\star \in \mathcal{L}_\mathbf{r}$ and $\mathbf{e} = \mathbf{0}$) by taking no more than $\mathcal{O}(R^P + PNR)$ Gaussian measurements. In an attempt to reduce this gap, [35] proposed minimizing the nuclear norm of a single matrix unfolding having "more balanced" dimensions. This leads to recovery guarantees with $\mathcal{O}(R^{\lfloor \frac{P}{2} \rfloor} N^{\lceil \frac{P}{2} \rceil})$ Gaussian measurements. Despite the progress, this bound still grows much faster than $\Phi(\mathbf{r})$ and only brings improvement for $P > 3$. Still in the realm of convex SNN-based approaches, robust principal component analysis (PCA) techniques were extended to a TC setting in [26], relying on an underlying model which consists of a sum of a low-mrank tensor plus a sparse one. With this approach, [26] stated the first recovery guarantees for TC, which apply with $\mathcal{O}(\mu RN^{P-1}P^2 \log^2(N^{P-1}))$ measurements, where μ is a measure of coherence of the tensor. In [50], a (convex) TC formulation based on tensor rank is tackled by means of a greedy Frank–Wolfe scheme, which updates the estimate at each iteration by a rank-one term. This approach relies on a constrained least-squares formulation where the nuclear norm of the sought tensor is bounded by a

positive constant β.

Other existing methods are predominantly based on nonconvex formulations. For instance, a joint low-rank matrix factorization of all modal unfoldings is sought in [49] by minimizing a weighted sum of quadratic errors. For the TC problem, Riemannian optimization techniques have been used in [28, 31] by exploiting the smooth manifold structure of sets of low-mrank tensors. Finally, the TC problem is also addressed in [21] by relying on the so-called tensor train (TT) model, which has its own definition of rank, the TT rank. When the TT rank has components bounded by R, the number of DOFs of this model grows as $\mathcal{O}(PR^2N)$, making it attractive for large P and small R.

In the rest of this paper, we focus on iterative hard thresholding (IHT) algorithms for LRTR, which build upon ideas used in CS and LRMR [1, 2, 27, 43]. The first proposed one was tensor IHT (TIHT) [36], whose thresholding operator is the truncated HOSVD, a standard technique for computing a quasi-optimal low-mrank tensor approximation. An accelerated variant called ISS-TIHT (where ISS stands for "improved step size") was later proposed in [19], relying on a step size selection heuristic to increase convergence speed. However, recovery guarantees based solely on bounding restricted isometry constants (henceforth abbreviated as RICs; see section 2 for a definition) are still lacking for TIHT, though partial results have been recently delivered in [38] and an RIC-based local convergence result was derived in [37, Th. 3]. The minimum n-rank approximation (MnRA) algorithm [52], in its turn, uses a convex combination of truncated SVDs in lieu of the hard thresholding operator. This approach enjoys recovery guarantees based on RIC conditions. However, the RICs exploited in [52] apply to the sensing of tensors having only one low-rank mode. Consequently, the tightest possible sampling bound implied by these results for achieving recovery with high probability is $M \geq M_{\min} = \mathcal{O}(RN^{P-1})$.

1.2. Our contributions and paper organization. We propose an IHT algorithm relying on the low-mrank approximation technique developed in [48], which we call[4] *sequentially optimal modal projections* (SeMP). This technique is significantly less costly than the thresholding operators used by TIHT and MnRA, especially for very low mrank, and often leads to better performance. Our algorithm is named SeMPIHT. At the theoretical level, we derive recovery guarantees for SeMPIHT under a certain RIC condition by exploiting the sequential optimality of the modal projections which constitute SeMP. In particular, we show that SeMPIHT converges to the true tensor in the ideal case (i.e., when $\boldsymbol{\mathcal{X}}^\star \in \mathcal{L}_{\mathbf{r}}$ and $\mathbf{e} = \mathbf{0}$). In light of [38, Th. 2], for fixed P the derived RIC condition is met with high probability when $M \geq M_{\min} = \mathcal{O}(RN^{P-1})$ Gaussian measurements are taken, similarly to the result of [52]. Thus, our theoretical results unfortunately do not improve upon previous sampling bounds. Nevertheless, our simulation results suggest that the bound of SeMPIHT actually scales as $M_{\min} = \mathcal{O}(R^P + PNR)$, which is order-optimal with respect to $\Phi(\mathbf{r})$. The same optimality was also observed in our experiments for TIHT and MnRA, which achieved good results when coupled with the ISS heuristic (see subsection 3.5).

We also propose a gradual rank increase (GRI) technique akin to those of [21, 31], consisting in estimating a sequence of increasingly more complex models (in terms of mrank). Our systematic numerical experiments show that, when dealing with data

[4]Though [48] uses the name "sequentially truncated HOSVD," we prefer to adopt "sequentially optimal modal projections," because the resulting projection operators are not associated with the original dominant modal subspaces.

having fast decaying modal singular spectra, such a GRI heuristic mitigates or avoids degradation of the results when $\mathbf{r}$ is set beyond the recovery regime. Moreover, it is decisive for satisfactorily recovering tensors of that kind in TC, where their nonideal coherence properties bring severe difficulties even under a highly favorable regime. We extensively compare SeMPIHT with other algorithms in the recovery of two classes of synthetic tensors, one of which has fast decaying modal spectra, similarly to many real-world data tensors. These simulations involve Gaussian sensing and also the TC setting. Finally, the completion of a hyperspectral imaging data tensor is also performed, validating the usefulness of our contributions.

This paper is organized as follows. In section 2, we review the IHT approach and some existing algorithms based on this technique for CS, LRMR, and LRTR. Section 3 recalls the SeMP technique and introduces our proposed algorithm, stating its recovery guarantees and comparing it with previous IHT schemes for LRTR. A description of our GRI continuation technique is then given in section 4. The effect of performing GRI is studied in detail in section 5 by means of numerical experiments, and then other simulations are presented for the purposes of evaluating our approach and comparing it with other LRTR algorithms. Finally, concluding remarks are drawn in section 6.

2. Iterative hard thresholding. IHT is a simple and effective technique for the recovery of parsimonious signals from undercomplete measurements, having been successfully applied in CS, LRMR, and LRTR [2, 27, 36, 43, 52]. Its rationale is as follows. In an arbitrary finite-dimensional inner product space $\mathcal{H}$ endowed with a scalar product $\langle \cdot, \cdot \rangle$, one poses

$$\min_{x \in \mathcal{S}} J(x) = \min_{x \in \mathcal{S}} \|\mathbf{y} - \mathscr{A}(x)\|_2^2 , \tag{6}$$

where $\mathscr{A} : \mathcal{H} \mapsto \mathbb{R}^M$ is a linear operator, $\|x\|_2^2 \triangleq \langle x, x \rangle$, and the set $\mathcal{S} \subset \mathcal{H}$ contains the parsimonious elements of interest. This set is typically nonconvex, closed, and nonempty. The basic idea of IHT is then to generate iterates of the form

$$x_k \in \mathcal{P}_{\mathcal{S}} \left(x_{k-1} - \frac{\mu_k}{2} \nabla J(x_{k-1}) \right), \qquad \text{with} \quad \nabla J(x) = -2\, \mathscr{A}^\dagger \left(\mathbf{y} - \mathscr{A}\left(x\right) \right), \tag{7}$$

where $\mu_k > 0$ is some chosen step size, $\mathscr{A}^\dagger$ is the adjoint of $\mathscr{A}$, and $\mathcal{P}_{\mathcal{S}}$ denotes[5] the (orthogonal) projector onto $\mathcal{S}$. Because $\mathcal{S}$ is possibly nonconvex, $\mathcal{P}_S(x) = \arg\min_{z \in \mathcal{S}} \|x - z\|_2^2$ generally yields a set (which is nonempty by the extreme value theorem, since $\mathcal{S}$ is closed and nonempty). In practice, whichever x_k (satisfying (7)) is chosen, convergence and recovery results usually remain the same.

The iterates in (7) resemble the projected gradient (or projected Landweber) algorithm, which is a convex optimization method [8]. Interestingly, it turns out that they apply to (6) even for nonconvex $\mathcal{S}$, due to the form of $J(x)$. The explanation relies on the majorization-minimization technique [1], which consists in minimizing at iteration k the functional

$$J_k(x) = \mu_k J(x) + \|x - x_{k-1}\|_2^2 - \mu_k \left\|\mathscr{A}\left(x - x_{k-1}\right)\right\|_2^2 \tag{8}$$

over $\mathcal{S}$ for some value of μ_k such that $J_k(x) > \mu_k J(x)$ for all $x \neq x_{k-1}$. Such a μ_k always exists: as $\mathcal{H}$ is finite-dimensional and thus $\|\mathscr{A}\|$ is bounded,[6] one can take

[5] This notation will be repeatedly used throughout the paper.

[6] $\|\mathscr{A}\|$ denotes the operator norm of $\mathscr{A}$.

$\mu_k < \|\mathscr{A}\|^{-1}$. Clearly, if $x_k \in \arg\min_{x\in\mathcal{S}} J_k(x)$ and $x_k \neq x_{k-1}$, then $\mu_k J(x_k) < J_k(x_k) \leq J_k(x_{k-1}) = \mu_k J(x_{k-1})$, thus achieving objective function reduction. So, the question is how to compute such an x_k. Expanding $J(x)$ in (8), we have

$$J_k(x) = \|x - x_{k-1}\|_2^2 - 2\mu_k\langle \mathscr{A}^\dagger(\mathbf{y} - \mathscr{A}(x_{k-1})), x\rangle - \mu_k\|\mathscr{A}(x_{k-1})\|_2^2 + \mu_k\|\mathbf{y}\|_2^2. \tag{9}$$

The expression in (9) is strictly convex, and hence its (unique) unconstrained minimum is straightforwardly obtained by solving $J_k'(x) = 0$, which gives

$$x_k^\star \triangleq \arg\min_{x\in\mathcal{H}} J_k(x) = x_{k-1} + \mu_k\mathscr{A}^\dagger(\mathbf{y} - \mathscr{A}(x_{k-1})) = x_{k-1} - \frac{\mu_k}{2}\nabla J(x_{k-1}). \tag{10}$$

The crucial point is that, because the quadratic term in x of $J(x)$ is canceled out in $J_k(x)$, the latter has circular level curves, and thus $\arg\min_{x\in\mathcal{S}} J_k(x) = \mathscr{P}_{\mathcal{S}}(x_k^\star)$ for any nonempty closed set $\mathcal{S}$ [19, Proposition 3.1]. Such simplicity is precisely the benefit of iteratively minimizing $J_k(x)$ rather than $J(x)$.

The effectiveness of IHT algorithms is typically demonstrated on the basis of RICs, which we now introduce by generalizing the definitions given in [6, 36, 40].

Definition 1 (restricted isometry constants (RICs)). *A linear operator $\mathscr{A} : \mathcal{H} \mapsto \mathbb{R}^M$ is said to satisfy the restricted isometry property (RIP) over $\mathcal{S} \subset \mathcal{H}$ if there exists a (minimal) constant $\delta_{\mathcal{S}} < 1$, called restricted isometry constant (RIC) of $\mathscr{A}$ with respect to $\mathcal{S}$, such that*

$$\forall\ x \in \mathcal{S}, \quad (1 - \delta_{\mathcal{S}})\|x\|_{\mathcal{H}}^2 \leq \|\mathscr{A}(x)\|_2^2 \leq (1 + \delta_{\mathcal{S}})\|x\|_{\mathcal{H}}^2. \tag{11}$$

2.1. Application to compressive sensing and low-rank matrix recovery. In Table 1, the main ingredients of IHT are particularized for CS, LRMR, and LRTR.

Formulation (6) applies to CS with $\mathcal{H} = \mathbb{R}^N$ and $\mathcal{S} = \mathcal{S}_s$, as defined in Table 1. Note that $\mathcal{S}_s$ is not convex, since $\mathbf{u}, \mathbf{v} \in \mathcal{S}_s$ generally implies $\alpha\mathbf{u} + (1-\alpha)\mathbf{v} \in \mathcal{S}_{2s}$ for $\alpha \in (0, 1)$. The iterates thus read [1]

$$\mathbf{x}_{k+1} = \mathscr{H}_s\left(\mathbf{x}_k + \mu_k\mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x}_k)\right), \tag{12}$$

where $\mathscr{H}_s : \mathbb{R}^N \mapsto \mathcal{S}_s$ projects a vector onto its best s-sparse approximation by zeroing all but its components of largest magnitude. As multiple best approximations may exist, an arbitrary $\mathscr{H}_s(\mathbf{x}) \in \mathscr{P}_{\mathcal{S}_s}(\mathbf{x})$ is picked. $\mathscr{H}_s$ is called a *hard thresholding operator*. RIC-based recovery guarantees for this algorithm have been given, e.g., in [15, Th. 6.18], and hold provided $\mathcal{O}(s\log(N/s))$ measurements are taken. This exceeds the number of DOFs of the model only by a logarithmic factor.

In analogy with (12), IHT can be applied to LRMR with $\mathcal{H} = \mathbb{R}^{N_1\times N_2}$ and $\mathcal{S} = \mathcal{L}_R$ by computing

$$\mathbf{X}_{k+1} = \mathscr{H}_R\Big(\mathbf{X}_k + \mu_k\mathscr{A}^\dagger(\mathbf{y} - \mathscr{A}(\mathbf{X}_k))\Big), \tag{13}$$

where $\mathscr{H}_R : \mathbb{R}^{N_1\times N_2} \mapsto \mathcal{L}_R$ delivers a best rank-R approximation of a matrix. From the Eckart–Young theorem [13], it can be computed through $\mathscr{H}_R(\mathbf{X}) = \sum_{r=1}^R \sigma_r\mathbf{u}_r\mathbf{v}_r^T$, where $\mathbf{X} = \sum_{n=1}^{\min\{N_1,N_2\}} \sigma_n\mathbf{u}_n\mathbf{v}_n^T$ is the SVD of $\mathbf{X}$, with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{N_1,N_2\}}$. If $\sigma_R = \cdots = \sigma_{R+d}$, with $d \in [\![\min\{N_1, N_2\} - R]\!]$, then $\mathscr{H}_R$ delivers one of the multiple best approximations of $\mathbf{X}$. RIC-based performance bounds are derived, e.g., in [32]. Similarly to the CS setting, certain random MOs have small RICs with high probability as long as $\mathcal{O}(R(N_1 + N_2 - R))$ measurements are taken [7]. This precisely matches the number of DOFs of a rank-R matrix. Unfortunately, though, these recovery guarantees do not apply to matrix completion.

TABLE 1
Particularization of the setting described in section 2 for some structured data recovery problems.

Probl.	Ambient space $\mathcal{H}$	Parsimon. elem. set $\mathcal{S}$	Obj. funct. $J(x)$	RIC notation
CS	$\mathbb{R}^N$	$\mathcal{S} = \mathcal{S}_s \triangleq \{\mathbf{x} : \text{supp}(\mathbf{x}) \leq s\}$ (‡)	$J(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$	δ_s: order s
LRMR	$\mathbb{R}^{N_1 \times N_2}$	$\mathcal{S} = \mathcal{L}_R \triangleq \{\mathbf{X} : \text{rank}(\mathbf{X}) \leq R\}$	$J(\mathbf{X}) = \|\mathbf{y} - \mathscr{A}(\mathbf{X})\|_2^2$	δ_R: order R
LRTR	$\mathbb{R}^{N_1 \times \cdots \times N_P}$	$\mathcal{S} = \mathcal{L}_\mathbf{r}$, $\mathbf{r} = (R_1, \ldots, R_P)$	$J(\boldsymbol{\mathcal{X}}) = \|\mathbf{y} - \mathscr{A}(\boldsymbol{\mathcal{X}})\|_2^2$	$\delta_\mathbf{r}$: order $\mathbf{r}$

(‡): $\text{supp}(\mathbf{x})$ denotes the support of $\mathbf{x}$, i.e., its number of nonzero components.

2.2. Application to tensor recovery based on multilinear rank. Consider now the tensor recovery setting, with $\mathcal{H} = \mathbb{R}^{N_1 \times \cdots \times N_P}$ and $\mathcal{S} = \mathcal{L}_\mathbf{r}$, as defined by (2). Although computing projections onto $\mathcal{L}_\mathbf{r}$ is NP-hard, efficient approximate methods exist. A widely adopted one consists in truncating the HOSVD at mrank $\mathbf{r}$ [11]. Denoting the corresponding operator by $\mathscr{H}_\mathbf{r} : \mathbb{R}^{N_1 \times \cdots \times N_P} \mapsto \mathcal{L}_\mathbf{r}$, we have

$$\mathscr{H}_\mathbf{r}(\boldsymbol{\mathcal{X}}) = \boldsymbol{\mathcal{X}} \times_1 \mathbf{U}^{(1)}\mathbf{U}^{(1)^T} \times_2 \cdots \times_P \mathbf{U}^{(P)}\mathbf{U}^{(P)^T}, \tag{14}$$

where $\mathbf{U}^{(p)} \in \mathcal{V}_{R_p}(\mathbb{R}^{N_p})$ contains as columns the first R_p left singular vectors of the unfolding $\mathbf{X}_{\langle p \rangle}$. Therefore, $\mathbf{U}^{(p)}\mathbf{U}^{(p)^T}$ is an orthogonal projector onto the dominant subspace of dimension R_p associated with the pth mode of $\boldsymbol{\mathcal{X}}$. The truncated HOSVD (THOSVD) operator defined in (14) is easy to implement, as it requires only standard numerical linear algebra routines. Moreover, it can be shown to be quasi-optimal by a factor of $\sqrt{P}$, in the sense that [48]

$$\|\boldsymbol{\mathcal{X}} - \mathscr{H}_\mathbf{r}(\boldsymbol{\mathcal{X}})\|_F \leq \sqrt{P} \min_{\boldsymbol{\mathcal{Z}} \in \mathcal{L}_\mathbf{r}} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Z}}\|_F . \tag{15}$$

Due to the above properties, $\mathscr{H}_\mathbf{r}$ is employed by the TIHT algorithm [36], whose iterates read

$$\boldsymbol{\mathcal{X}}_{k+1} = \mathscr{H}_\mathbf{r}(\boldsymbol{\mathcal{V}}_k), \text{ where } \boldsymbol{\mathcal{V}}_k \triangleq \boldsymbol{\mathcal{X}}_k + \mu_k \mathscr{A}^\dagger (\mathbf{y} - \mathscr{A}(\boldsymbol{\mathcal{X}}_k)) \text{ and } \mu_k = \frac{\|\nabla J(\boldsymbol{\mathcal{X}}_k)\|_F^2}{\|\mathscr{A}(\nabla J(\boldsymbol{\mathcal{X}}_k))\|_F^2}. \tag{16}$$

More recently, the same authors have proposed the normalized TIHT (NTIHT) algorithm [38], which is also based on the THOSVD method but uses the step size formula

$$\mu_k = \frac{\|\boldsymbol{\mathcal{G}}_k\|_F^2}{\|\mathscr{A}(\boldsymbol{\mathcal{G}}_k)\|_F^2}, \quad \boldsymbol{\mathcal{G}}_k = \nabla J(\boldsymbol{\mathcal{X}}_k) \times_1 \mathbf{U}_k^{(1)}\mathbf{U}_k^{(1)^T} \times_2 \cdots \times_P \mathbf{U}_k^{(P)}\mathbf{U}_k^{(P)^T}, \tag{17}$$

where the orthogonal matrices $\mathbf{U}_k^{(p)}$ are bases for the modal unfoldings of $\boldsymbol{\mathcal{X}}_k$ obtained with the use of THOSVD at iteration $k-1$.

Although the effectiveness of TIHT and NTIHT was experimentally shown, recovery results based solely on typical RIP conditions are still lacking. The best one in this sense is as follows.

THEOREM 2 (performance bound of NTIHT [38, Th. 1]). *Put $a \in (0,1)$, and let $\mathscr{A}$ be an MO possessing a $3\mathbf{r}$-RIC satisfying $\delta_{3\mathbf{r}} < a/(a+8)$, where $3\mathbf{r} = (3R_1, \ldots 3R_P)$. Let $\boldsymbol{\mathcal{X}}^\star \in \mathcal{L}_\mathbf{r}$. Then, given measurements $\mathbf{y} = \mathscr{A}(\boldsymbol{\mathcal{X}}^\star) + \mathbf{e}$, if*

$$\|\boldsymbol{\mathcal{X}}_k - \boldsymbol{\mathcal{V}}_k\|_F \leq (1 + \varepsilon(a)) \|\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{V}}_k\|_F , \tag{18}$$

where $\varepsilon(a) = a^2(1-\delta_{3\mathbf{r}})^2(17(1-\delta_{3\mathbf{r}} + \sqrt{1+\delta_{2\mathbf{r}}}\|\mathscr{A}\|))^{-2}$, then for all k we have

$$\|\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}_k\|_F \leq a^{k-1}\|\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}_0\|_F + \frac{b(a)}{1-a}\|\mathbf{e}\|_2,$$

where $b(a) = 2\frac{\sqrt{1+\delta_{3\mathbf{r}}}}{1-\delta_{3\mathbf{r}}} + \sqrt{4\varepsilon(a) + 2\varepsilon(a)^2}\,\frac{1}{1-\delta_{3\mathbf{r}}}\,\|\mathscr{A}\|$.

Though a heuristic justification is given in [38] for condition (18), it cannot be guaranteed in general because the THOSVD is quasi-optimal by a factor $\sqrt{P}$, whereas $\varepsilon(a) \approx 0$. We point out that a local convergence result based solely on RIC assumptions was derived in [37, Th. 3]. However, it is rather restrictive, as it applies only in a sufficiently small neighborhood of the desired global minimum.

A similar scheme called MnRA is proposed in [52]. However, as it uses a convex combination of truncated SVDs in lieu of the hard thresholding operator, no projection onto $\mathcal{L}_{\mathbf{r}}$ is performed. Given nonnegative weights w_p satisfying $\sum_{p\in[\![P]\!]} w_p = 1$, this operator, denoted here by $\mathscr{C}_{\mathbf{r}}$, is defined as

$$\mathscr{C}_{\mathbf{r}}(\boldsymbol{\mathcal{X}}) = \sum_{p=1}^{P} w_p \boldsymbol{\mathcal{Z}}_p \quad \text{such that} \quad (\boldsymbol{\mathcal{Z}}_p)_{\langle p\rangle} = \mathscr{H}_{R_p}\left(\mathbf{X}_{\langle p\rangle}\right), \tag{19}$$

in which $\mathscr{H}_{R_p}$ is the same as in (13) with $R = R_p$. The step size of MnRA is fixed, thus yielding iterates

$$\boldsymbol{\mathcal{X}}_{k+1} = \mathscr{C}_{\mathbf{r}}\left(\boldsymbol{\mathcal{X}}_k + \mu\, \mathscr{A}^{\dagger}\left(\mathbf{y} - \mathscr{A}(\boldsymbol{\mathcal{X}}_k)\right)\right).$$

Although $\boldsymbol{\mathcal{X}}_k \notin \mathcal{L}_{\mathbf{r}}$ in general, convergence to the true estimate in the ideal case has been shown in [52] under RIP conditions. For convenience, this result is reproduced below.

THEOREM 3 (performance bound of MnRA [52, Th. 4.2]). *Let $\mathscr{A}$ be an MO with RICs $\delta_{\bar{\mathbf{r}}_p} < 1$ for all $p \in [\![P]\!]$, where $\bar{\mathbf{r}}_p \triangleq (N_1, \ldots, N_{p-1}, 3R_p, N_{p+1}, \ldots, N_P)$. Also let $\boldsymbol{\mathcal{X}}^{\star} \in \mathcal{L}_{\mathbf{r}}$ and $\mathbf{y} = \mathscr{A}(\boldsymbol{\mathcal{X}}^{\star}) + \mathbf{e}$, and assume $3/4 < \mu < 5/4$. If $\max_p \delta_{\bar{\mathbf{r}}_p} < \tau$, then MnRA satisfies*

$$\forall\, k, \qquad \|\boldsymbol{\mathcal{X}}^{\star} - \boldsymbol{\mathcal{X}}_k\|_F \le 2^{-k}\|\boldsymbol{\mathcal{X}}^{\star} - \boldsymbol{\mathcal{X}}_0\|_F + 2C\|\mathbf{e}\|_2,$$

where $C = 2\mu\sqrt{1+\tau}$ and $\tau = \frac{1/4-|1-\mu|}{\mu(1+\lceil \max_p N_p/R_p\rceil)}$.

Unlike Theorem 2, this result does not involve a restrictive assumption such as (18). Yet, it is not satisfying from a sampling efficiency standpoint, because $\mathscr{A}$ can only have an RIC $\delta_{\bar{\mathbf{r}}_p} < 1$ if $M \ge M_{\min} = \mathcal{O}(R_p\bar{N}_p)$, which grows much more quickly than $\Phi(\mathbf{r})$ given by (4). For instance, $\mathcal{O}(R_p\bar{N}_p) = \mathcal{O}(RN^{P-1})$ when $R_p = R$ and $N_p = N$ for all p.

We point out that [38] derives sampling bounds which ensure (with high probability) the RIP for subgaussian and for random partial Fourier MOs. They can be coupled with RIC-based recovery guarantees such as Theorems 2 and 3 in order to derive sampling requirements for the analyzed algorithms. Though random partial Fourier MOs require slightly more measurements than subgaussian ones (by a polylogarithmic factor), they are much more reasonable in practice because fast transform algorithms can be exploited to reduce both acquisition and recovery times.

3. The SeMPIHT algorithm. We propose an IHT scheme employing the SeMP technique for approximate projection onto $\mathcal{L}_{\mathbf{r}}$. The iterates of our SeMPIHT algorithm are thus computed as

$$\boldsymbol{\mathcal{X}}_k = \mathcal{S}_{\mathbf{r}}\left(\boldsymbol{\mathcal{X}}_{k-1} + \mu_k \mathscr{A}^{\dagger}\left(\mathbf{y} - \mathscr{A}\left(\boldsymbol{\mathcal{X}}_{k-1}\right)\right)\right), \tag{20}$$

where $\mathcal{S}_{\mathbf{r}} : \mathbb{R}^{N_1\times\cdots\times N_P} \mapsto \mathcal{L}_{\mathbf{r}}$ denotes the SeMP operator. Essentially, instead of computing the dominant subspaces of all modal unfoldings and then performing the

Algorithm 1 `sempiht(`$\boldsymbol{\mathcal{X}}_0$`,`$\mathbf{y}$`,`$\mathscr{A}$`,`$\mathbf{r}$`,`$K_{\max}$`,`$\epsilon$`)`.

Inputs: Initial solution $\boldsymbol{\mathcal{X}}_0$, measurement vector $\mathbf{y}$, measurement operator $\mathscr{A}$, target mrank $\mathbf{r}$, maximum number of iterations $K_{\max}$, tolerance ϵ

Output: Estimated tensor, $\hat{\boldsymbol{\mathcal{X}}}^{\star}$

1. **for** $k = 1, \ldots, K_{\max}$
   - (i) $\boldsymbol{\mathcal{G}}_k \leftarrow \mathscr{A}^* \left(\mathbf{y} - \mathscr{A}\left(\boldsymbol{\mathcal{X}}_{k-1}\right)\right)$
   - (ii) compute step size μ_k using either ISS (see subsection 3.5) or formula (17)
   - (iii) compute $\boldsymbol{\mathcal{X}}_k \leftarrow \mathcal{S}_{\mathbf{r}}\left(\boldsymbol{\mathcal{X}}_{k-1} + \mu_k \boldsymbol{\mathcal{G}}_k\right)$ using Algorithm 2
   - (iv) **if** criterion (32) is satisfied, **break**

   end
2. **return** $\hat{\boldsymbol{\mathcal{X}}}^{\star} \leftarrow \boldsymbol{\mathcal{X}}_k$

projection, SeMP proceeds by interleaving these operations in a sequential fashion. It is therefore rather similar in spirit to some techniques used in the context of hierarchical tensor representations; see [20] and references therein. The SeMPIHT algorithm is detailed in Algorithm 1.

In what follows we describe the SeMP operator and its properties.

3.1. The SeMP technique for approximate projection onto $\boldsymbol{\mathcal{L}}_{\mathbf{r}}$. The principle of multilinear orthogonal projection (14) which underlies the THOSVD operator $\mathscr{H}_{\mathbf{r}}$ can be more generally applied with other choices of modal projectors. Essentially, we seek an efficient way of computing P orthogonal projection matrices $\boldsymbol{\Pi}_p = \mathbf{V}^{(p)}\mathbf{V}^{(p)^T}$, with $\mathbf{V}^{(p)} \in \mathcal{V}_{R_p}(\mathbb{R}^{N_p})$, which approximate the solution of

$$\min_{\boldsymbol{\mathcal{Z}} \in \mathcal{L}_{\mathbf{r}}} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Z}}\|_F^2 = \min_{\substack{\boldsymbol{\Pi}_p = \mathbf{V}^{(p)}\mathbf{V}^{(p)^T} \\ \mathbf{V}^{(p)} \in \mathcal{V}_{R_p}(\mathbb{R}^{N_p})}} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}} \times_1 \boldsymbol{\Pi}_1 \times_2 \cdots \times_P \boldsymbol{\Pi}_P\|_F^2 .$$

In (14), each mode-p projector $\boldsymbol{\Pi}_p = \mathbf{U}^{(p)}\mathbf{U}^{(p)^T}$ is associated with the R_p-dimensional dominant subspace of $\mathbf{X}_{\langle p \rangle}$. This choice is motivated by the inequality [48]

$$\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}} \times_1 \boldsymbol{\Pi}_1 \times_2 \cdots \times_P \boldsymbol{\Pi}_P\|_F^2 \leq \sum_{p=1}^{P} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}} \times_p \boldsymbol{\Pi}_p\|_F^2 = \sum_{p=1}^{P} \left\|\boldsymbol{\mathcal{X}} \times_p \boldsymbol{\Pi}_p^{\perp}\right\|_F^2 ,$$

where $\boldsymbol{\Pi}_p^{\perp} = \mathbf{I} - \boldsymbol{\Pi}_p$ projects onto the orthogonal complement of the range of $\boldsymbol{\Pi}_p$. When each $\boldsymbol{\Pi}_p$ is associated with the dominant subspace of $\mathbf{X}_{\langle p \rangle}$, the above upper bound is minimized. In practice, note that applying $\mathscr{H}_{\mathbf{r}}$ requires computation of *all* P projectors (possibly in parallel) *before* they are applied.

The SeMP approximate projector proposed in [48], which we define next, is based on another choice for the modal projectors. Due to its sequential nature, an ordering must be specified for the modal projections. Such an ordering is denoted by a permutation $\pi = (p_1, p_2, \ldots, p_P)$ of $(1, \ldots, P)$, referred to as the modal projection ordering (MPO). For simplicity of exposition, we now assume $\pi = (1, \ldots, P)$.

DEFINITION 4. *Let us denote by* $\mathscr{H}_{R_p}^{(p)} : \mathbb{R}^{N_1 \times \cdots \times N_P} \mapsto \mathbb{R}^{N_1 \times \cdots \times N_P}$ *the operator which applies singular value hard thresholding to the pth mode of its argument, i.e.,* $(\mathscr{H}_{R_p}^{(p)}(\boldsymbol{\mathcal{X}}))_{\langle p \rangle} = \mathscr{H}_{R_p}(\mathbf{X}_{\langle p \rangle})$. *Then, the SeMP operator* $\mathcal{S}_{\mathbf{r}} : \mathbb{R}^{N_1 \times \cdots \times N_P} \mapsto \mathcal{L}_{\mathbf{r}}$ *is defined as*

$$\mathcal{S}_{\mathbf{r}}(\boldsymbol{\mathcal{X}}) = \mathscr{H}_{R_P}^{(P)} \mathscr{H}_{R_{P-1}}^{(P-1)} \ldots \mathscr{H}_{R_1}^{(1)}(\boldsymbol{\mathcal{X}}) . \tag{21}$$

From the Eckart–Young theorem, (21) amounts to choosing the modal projection matrices

$$\hat{\mathbf{\Pi}}_p = \arg\min_{\mathbf{\Pi}_p} \left\| \boldsymbol{\mathcal{X}} \times_1 \hat{\mathbf{\Pi}}_1 \times_2 \cdots \times_{p-1} \hat{\mathbf{\Pi}}_{p-1} \times_p \mathbf{\Pi}_p^{\perp} \right\|_F^2, \quad \text{subject to } \begin{cases} \mathbf{\Pi}_p = \mathbf{V}^{(p)} \mathbf{V}^{(p)^T}, \\ \mathbf{V}^{(p)} \in \mathcal{V}_{R_p}(\mathbb{R}^{N_p}), \end{cases} \tag{22}$$

so that

$$\mathscr{H}_{R_{p-1}}^{(p-1)} \ldots \mathscr{H}_{R_1}^{(1)} (\boldsymbol{\mathcal{X}}) = \boldsymbol{\mathcal{X}} \times_1 \hat{\mathbf{\Pi}}_1 \times_2 \cdots \times_{p-1} \hat{\mathbf{\Pi}}_{p-1}.$$

This choice can be justified by invoking the inequality [48]

$$\min_{\boldsymbol{\mathcal{Z}} \in \mathcal{L}_{\mathbf{r}}} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Z}}\|_F^2 \leq \sum_{p=1}^{P} = \min_{\substack{\mathbf{\Pi}_p = \mathbf{V}^{(p)} \mathbf{V}^{(p)^T} \\ \mathbf{V}^{(p)} \in \mathcal{V}_{R_p}(\mathbb{R}^{N_p})}} \left\| \boldsymbol{\mathcal{X}} \times_1 \hat{\mathbf{\Pi}}_1 \times_2 \cdots \times_{p-1} \hat{\mathbf{\Pi}}_{p-1} \times_p \mathbf{\Pi}_p^{\perp} \right\|_F^2 \tag{23}$$

$$\leq \sum_{p=1}^{P} = \min_{\substack{\mathbf{\Pi}_p = \mathbf{V}^{(p)} \mathbf{V}^{(p)^T} \\ \mathbf{V}^{(p)} \in \mathcal{V}_{R_p}(\mathbb{R}^{N_p})}} \left\| \boldsymbol{\mathcal{X}} \times_p \mathbf{\Pi}_p^{\perp} \right\|_F^2. \tag{24}$$

SeMP picks the minimizers of the upper bound in (23), while THOSVD picks those of (24). Another crucial difference exists in comparison with THOSVD: each $\hat{\mathbf{\Pi}}_p$ here depends on all previously calculated $\hat{\mathbf{\Pi}}_q$, with $q < p$. Hence, note that we *cannot* compute all the projectors $\hat{\mathbf{\Pi}}_p$ in parallel, since their computation and application must be interleaved.

For clarity, an algorithmic description of the computational procedure associated with (21) is given in Algorithm 2. Step 2(ii) of this procedure is equivalent to calculating

$$\bar{\boldsymbol{\mathcal{V}}}_p = \bar{\boldsymbol{\mathcal{V}}}_{p-1} \times_p \bar{\mathbf{U}}^{(p)^T} \in \mathbb{R}^{R_1 \times \cdots \times R_p \times N_{p+1} \times \cdots \times N_P}. \tag{25}$$

Therefore, the final outcome can be written as

$$\mathcal{S}_{\mathbf{r}} (\boldsymbol{\mathcal{X}}) = \boldsymbol{\mathcal{X}} \times_1 \bar{\mathbf{U}}^{(1)} \bar{\mathbf{U}}^{(1)^T} \times_2 \cdots \times_P \bar{\mathbf{U}}^{(P)} \bar{\mathbf{U}}^{(P)^T} = \boldsymbol{\mathcal{X}} \times_1 \hat{\mathbf{\Pi}}_1 \times_2 \cdots \times_P \hat{\mathbf{\Pi}}_P.$$

Note the similarity of the above expression to (14). The fact that the matrix $\bar{\mathbf{U}}^{(p)}$ calculated in Algorithm 2 satisfies $\bar{\mathbf{U}}^{(p)} \bar{\mathbf{U}}^{(p)^T} = \hat{\mathbf{\Pi}}_p$, with $\hat{\mathbf{\Pi}}_p$ defined by subsection 3.1, can be verified as follows. For brevity, let us denote $\boldsymbol{\mathcal{V}}_{p-1} \triangleq \mathscr{H}_{R_{p-1}}^{(p-1)} \ldots \mathscr{H}_{R_1}^{(1)} (\boldsymbol{\mathcal{X}})$, with $\boldsymbol{\mathcal{V}}_0 = \boldsymbol{\mathcal{X}}$. We need to show that $\bar{\mathbf{U}}^{(p)}$ contains the first left R_p singular vectors of $(\boldsymbol{\mathcal{V}}_{p-1})_{\langle p \rangle}$ as columns. For $p = 1$, this is clearly true, as $\boldsymbol{\mathcal{V}}_0 = \bar{\boldsymbol{\mathcal{V}}}_0 = \boldsymbol{\mathcal{X}}$. For $p > 1$, we proceed by induction. Assume the claim holds for all $q \in [\![p-1]\!]$, which implies $\bar{\mathbf{U}}^{(q)} \bar{\mathbf{U}}^{(q)^T} = \hat{\mathbf{\Pi}}_q$. Then, it is easy to verify that it holds also for p, as the left singular vectors of $(\boldsymbol{\mathcal{V}}_{p-1})_{\langle p \rangle} = (\boldsymbol{\mathcal{X}} \times_1 \hat{\mathbf{\Pi}}_1 \times_2 \cdots \times_{p-1} \hat{\mathbf{\Pi}}_{p-1})_{\langle p \rangle}$ are the same as those

Algorithm 2 Sequentially optimal projections (SeMP) for best low-mrank approximation [48].

Inputs: Tensor $\boldsymbol{\mathcal{X}}$ whose best approximation in $\mathcal{L}_{\mathbf{r}}$ is sought, target mrank $\mathbf{r} = (R_1, \ldots, R_P)$

Output: An approximate projection $\mathcal{S}_{\mathbf{r}}(\boldsymbol{\mathcal{X}})$ of $\boldsymbol{\mathcal{X}}$ onto $\mathcal{L}_{\mathbf{r}}$

1. set $\bar{\boldsymbol{\mathcal{V}}}_0 = \boldsymbol{\mathcal{X}}$
2. **for** $p = 1, \ldots, P$
   (i) compute the SVD: $(\bar{\boldsymbol{\mathcal{V}}}_{p-1})_{\langle p \rangle} = \begin{bmatrix} \bar{\mathbf{U}}^{(p)} & \tilde{\mathbf{U}}^{(p)} \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{\Sigma}}^{(p)} & \mathbf{0} \\ \mathbf{0} & \tilde{\boldsymbol{\Sigma}}^{(p)} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{W}}^{(p)} & \tilde{\mathbf{W}}^{(p)} \end{bmatrix}^T$, where $\bar{\mathbf{U}}^{(p)} \in \mathbb{R}^{N_p \times R_p}$, $\bar{\boldsymbol{\Sigma}}^{(p)} \in \mathbb{R}^{R_p \times R_p}$, $\bar{\mathbf{W}}^{(p)} \in \mathbb{R}^{L_p \times R_p}$, $L_p = (\prod_{q=1}^{p-1} R_q)(\prod_{q=p+1}^{P} N_q)$
   (ii) compute $\bar{\boldsymbol{\mathcal{V}}}_p$ through its mode-p unfolding: $(\bar{\boldsymbol{\mathcal{V}}}_p)_{\langle p \rangle} \leftarrow \bar{\boldsymbol{\Sigma}}^{(p)} \bar{\mathbf{W}}^{(p)T}$
   end
3. **return** $\mathcal{S}_{\mathbf{r}}(\boldsymbol{\mathcal{X}}) \leftarrow \bar{\boldsymbol{\mathcal{V}}}_P \times_1 \bar{\mathbf{U}}^{(1)} \times_2 \cdots \times_P \bar{\mathbf{U}}^{(P)}$

of $\left(\bar{\boldsymbol{\mathcal{V}}}_{p-1}\right)_{\langle p \rangle} = (\boldsymbol{\mathcal{X}} \times_1 \bar{\mathbf{U}}^{(1)T} \times_2 \cdots \times_{p-1} \bar{\mathbf{U}}^{(p-1)T})_{\langle p \rangle}$. Indeed,

$$\begin{aligned}
(\boldsymbol{\mathcal{V}}_{p-1})_{\langle p \rangle} &= \mathbf{X}_{\langle p \rangle} \left(\bar{\mathbf{U}}^{(1)} \bar{\mathbf{U}}^{(1)T} \otimes \cdots \otimes \bar{\mathbf{U}}^{(p-1)} \bar{\mathbf{U}}^{(p-1)T} \otimes \mathbf{I}_{N_{p+1}} \otimes \cdots \otimes \mathbf{I}_{N_P} \right)^T, \\
\left(\bar{\boldsymbol{\mathcal{V}}}_{p-1}\right)_{\langle p \rangle} &= \mathbf{X}_{\langle p \rangle} \left(\bar{\mathbf{U}}^{(1)T} \otimes \cdots \otimes \bar{\mathbf{U}}^{(p-1)T} \otimes \mathbf{I}_{N_{p+1}} \otimes \cdots \otimes \mathbf{I}_{N_P} \right)^T,
\end{aligned}$$

where $\mathbf{I}_N$ denotes the $N \times N$ identity matrix, which implies

$$\begin{aligned}
(\boldsymbol{\mathcal{V}}_{p-1})_{\langle p \rangle} (\boldsymbol{\mathcal{V}}_{p-1})_{\langle p \rangle}^T &= \mathbf{X}_{\langle p \rangle} \left(\bar{\mathbf{U}}^{(1)} \bar{\mathbf{U}}^{(1)T} \otimes \cdots \otimes \bar{\mathbf{U}}^{(p-1)} \bar{\mathbf{U}}^{(p-1)T} \otimes \mathbf{I}_{N_{p+1}} \otimes \cdots \otimes \mathbf{I}_{N_P} \right) \mathbf{X}_{\langle p \rangle}^T \\
&= \left(\bar{\boldsymbol{\mathcal{V}}}_{p-1}\right)_{\langle p \rangle} \left(\bar{\boldsymbol{\mathcal{V}}}_{p-1}\right)_{\langle p \rangle}^T.
\end{aligned}$$

Let us calculate the resulting cost. Assuming it takes $\mathcal{O}(N_1 N_2 \min\{N_1, N_2\})$ operations to compute the SVD of an $N_1 \times N_2$ matrix,[7] the cost of Algorithm 2 is

$$c_{\text{SeMP}} = \mathcal{O}\left(\sum_{p=1}^{P} N_p L_p \min\{N_p, L_p\} \right) + \sum_{p=1}^{P} R_1 \ldots R_p N_{p+1} \ldots N_P + \mathcal{O}\left(\sum_{p=1}^{P} N_1 \ldots N_p R_p \ldots R_P \right), \tag{26}$$

where L_p is as defined in Algorithm 2. The first term corresponds to the computation of the SVD of $\bar{\boldsymbol{\mathcal{V}}}_0, \ldots, \bar{\boldsymbol{\mathcal{V}}}_{P-1}$, while the second and third terms comprise, respectively, the costs of steps 2(ii) and 3 of Algorithm 2.

Now, when $N_p \ll L_p$, instead of computing the SVD of $(\bar{\boldsymbol{\mathcal{V}}}_{p-1})_{\langle p \rangle} \in \mathbb{R}^{N_p \times L_p}$ as described by Algorithm 2, one can proceed as follows. First, the eigenvalue decomposition of $(\bar{\boldsymbol{\mathcal{V}}}_{p-1})_{\langle p \rangle} (\bar{\boldsymbol{\mathcal{V}}}_{p-1})_{\langle p \rangle}^T \in \mathbb{R}^{N_p \times N_p}$ provides (only) the left singular vectors of $(\bar{\boldsymbol{\mathcal{V}}}_{p-1})_{\langle p \rangle}$, which are then used for the projection stage (25). The goal is decomposing a much smaller matrix, at a cost of $\mathcal{O}(N_p^3)$ in lieu of $\mathcal{O}(N_p^2 L_p)$. Though

[7] In principle, the first R terms can be computed with $\mathcal{O}(R N_1 N_2)$ operations [23]. Yet, in our experience, optimized classical algorithms delivering the whole SVD, such as that of LAPACK, are usually faster.

the overall cost of the decomposition stage (i.e., the first term of (26)) remains $\mathcal{O}(N_p^2 L_p)$ because of the matrix product $(\bar{\boldsymbol{\mathcal{V}}}_{p-1})_{\langle p\rangle}(\bar{\boldsymbol{\mathcal{V}}}_{p-1})_{\langle p\rangle}^T$, it is generally much faster in practice, compensating for the increased effort of using (25), which costs $\mathcal{O}(\sum_{p=1}^P R_1 \dots R_p N_p \dots N_P)$, instead of step 2(ii) of Algorithm 2. So, even though this alternative implementation has an asymptotic cost of

$$\mathcal{O}\left(\sum_{p=1}^P N_p^2 L_p\right) + \mathcal{O}\left(\sum_{p=1}^P R_1 \dots R_p N_p \dots N_P\right) + \mathcal{O}\left(\sum_{p=1}^P N_1 \dots N_p R_p \dots R_P\right),$$

it is often quite advantageous, due to the reduced scale of the decomposition problem.

As a final comment, we mention that an important property of SeMP is its quasi-optimality by a factor of $\sqrt{P}$, i.e., its compliance to inequality (15), just like the THOSVD [48]. This can be easily shown from (24).

3.2. Computational cost per iteration. The computing effort involved in the use of $\mathcal{S}_{\mathbf{r}}$ is given by (26). Calculation of the argument of $\mathcal{S}_{\mathbf{r}}$ can be split into three stages: (i) computing the gradient of J, (ii) calculating the step size μ_k, and (iii) calculating the sum $\boldsymbol{\mathcal{V}}_0 = \boldsymbol{\mathcal{X}}_{k-1} - \frac{\mu_k}{2}\nabla J(\boldsymbol{\mathcal{X}}_{k-1})$. Stage (i) requires $\mathcal{O}(M\bar{N})$ operations for unstructured (e.g., Gaussian) operators, which can be alleviated by working with structured MOs. For instance, it requires $\mathcal{O}(M)$ in TC, while a cost of $\mathcal{O}(\bar{N}\log(\bar{N}))$ is achieved when random partial Fourier or noiselet measurements are taken by means of fast transform algorithms (see, e.g., [32, 41]). The cost of stage (ii) depends on the step size selection strategy, and thus we postpone its discussion to subsection 3.5. Finally, (iii) generally takes $\mathcal{O}(\bar{N})$ operations. In TC, this cost drops to $\mathcal{O}(M)$ because the gradient is sparse (due to the form of the SO).

3.3. Comparison with previous approaches. Clearly enough, the hard thresholding operator employed in an IHT algorithm has a major impact on its convergence speed, computing cost, and recovery effectiveness. We thus compare the operators of SeMPIHT, TIHT, and MnRA according to the following criteria.

(1) *Approximation accuracy.* As seen above, both $\mathscr{H}_{\mathbf{r}}$ and $\mathcal{S}_{\mathbf{r}}$ are quasi-optimal by a factor $\sqrt{P}$. In fact, our practical experience is consistent with the observations reported in [48], in that $\|\boldsymbol{\mathcal{X}} - \mathcal{S}_{\mathbf{r}}(\boldsymbol{\mathcal{X}})\|_F < \|\boldsymbol{\mathcal{X}} - \mathscr{H}_{\mathbf{r}}(\boldsymbol{\mathcal{X}})\|_F$ holds in most observed cases. MnRA's operator $\mathscr{C}_{\mathbf{r}}$, in its turn, satisfies

$$\begin{aligned}\|\boldsymbol{\mathcal{X}} - \mathscr{C}_{\mathbf{r}}(\boldsymbol{\mathcal{X}})\|_F &= \left\|\textstyle\sum_{p=1}^P w_p\,(\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Z}}_p)\right\|_F \le \textstyle\sum_{p=1}^P w_p \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Z}}_p\|_F \\ &\le \textstyle\sum_{p=1}^P w_p \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_{\mathbf{r}}\|_F = \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_{\mathbf{r}}\|_F\end{aligned}$$

for all $\boldsymbol{\mathcal{X}}_{\mathbf{r}} \in \mathscr{P}_{\mathcal{L}_{\mathbf{r}}}(\boldsymbol{\mathcal{X}})$, where the second inequality comes from the fact that $(\boldsymbol{\mathcal{Z}}_p)_{\langle p\rangle}$ is the best rank-R_p approximation of $\mathbf{X}_{\langle p\rangle}$ (see (19)). This perhaps surprising result is explained by the fact that $\mathscr{C}_{\mathbf{r}}$ is not really a projection onto $\mathcal{L}_{\mathbf{r}}$, due to the sum of terms which are low-rank only with respect to one mode.

(2) *Computing cost.* Applying $\mathscr{H}_{\mathbf{r}}$ requires

$$\mathcal{O}\left(\sum_{p=1}^P N_p \bar{N}_p \min\{N_p, \bar{N}_p\}\right) + \mathcal{O}\left(\sum_{p=1}^P R_1 \dots R_p N_p \dots N_P\right) + \mathcal{O}\left(\sum_{p=1}^P N_1 \dots N_p R_p \dots R_P\right) \tag{27}$$

operations, where the first sum is the cost of the P required SVDs and the others come from the projection onto the dominant modal subspaces (see (14)). The latter is broken down into two terms because it is faster to first compute the $R_1 \times \cdots \times R_P$ core of the THOSVD and then reconstruct the full tensor. Overall, the cost is dominated by the first sum of (27). Similarly, applying $\mathscr{C}_{\mathbf{r}}$ demands

$$\mathcal{O}\left(\sum_p N_p \bar{N}_p \min\{N_p, \bar{N}_p\}\right) + \mathcal{O}(P\bar{N}) \tag{28}$$

operations, where the first term is related to the SVDs of all modal unfoldings and the second to the convex combination of (19). Though (27) and (28) are asymptotically equivalent, $\mathscr{C}_{\mathbf{r}}$ is less costly in practice due to the difference between the second terms of these expressions.

Comparing now the first term of (26) with those of (27) and (28), it is seen that $\mathscr{S}_{\mathbf{r}}$ is less costly than $\mathscr{H}_{\mathbf{r}}$ and $\mathscr{C}_{\mathbf{r}}$, which is due to the dimensionality reduction performed for each p in Algorithm 2.

(3) *Analytical tractability.* Theorem 2 states a partial recovery result which applies to TIHT. Unfortunately, it relies upon a condition which cannot be ensured a priori. MnRA, in its turn, enjoys the RIC-based performance bound of Theorem 3, despite the fact that in general $\mathscr{C}_{\mathbf{r}}(\boldsymbol{\mathcal{X}}) \notin \mathcal{L}_{\mathbf{r}}$. This result, however, leads to suboptimal sampling bounds. At this point, it is not clear whether a similar (suboptimal) result based only on RIC assumptions can be derived for TIHT. As for SeMP, the sequential optimality of its modal projections allows establishing RIC-based performance bounds, as we will show next.

3.4. Theoretical recovery results. This section establishes a performance bound for SeMPIHT under the standard assumption that $\mathscr{A}$ has sufficiently low RICs. Our main result, whose proof is inspired by (but is simpler than) that of [52], is as follows.

THEOREM 5. *Let $\boldsymbol{\mathcal{X}}^\star \in \boldsymbol{\mathcal{T}}$ and $\mathbf{y} = \mathscr{A}(\boldsymbol{\mathcal{X}}^\star) + \mathbf{e}$. If $\mathscr{A}$ has an RIC $\delta_{\bar{\mathbf{r}}_p} < 2^{-P}$, where $\bar{\mathbf{r}}_p = (N_1, \ldots, N_{p-1}, 3R_p, N_{p+1}, \ldots, N_P)$, then the iterates computed via (20) with fixed step size $\mu_k = 1$ and MPO given by*[8] *$\pi = (p, p_2, \ldots, p_P)$ satisfy after k iterations*

$$\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_k\|_F \leq \xi^k \|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_0\|_F + \frac{2^P\sqrt{1+\delta_{\bar{\mathbf{r}}_p}}}{1-\xi}\,\|\mathscr{A}(\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}^\star_{\mathbf{r}}) + \mathbf{e}\|_2\,, \tag{29}$$

where $\xi = 2^P\delta_{\bar{\mathbf{r}}_p} < 1$ and $\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} \in \mathscr{P}_{\mathcal{L}_{\mathbf{r}}}(\boldsymbol{\mathcal{X}}^\star) = \arg\min_{\boldsymbol{\mathcal{Z}} \in \mathcal{L}_{\mathbf{r}}} \|\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{Z}}\|_F$, with $\mathbf{r} = (R_1, \ldots, R_P)$. If the step size formula (17) is used, then (29) holds with $\delta_{\bar{\mathbf{r}}_p} < 1/(2^{P+1}+1)$ and $\xi = \sup_k 2^P(|1-\mu_k| + \mu_k\delta_{\bar{\mathbf{r}}_p}) < 1$.

Proof. See Appendix A. ∎

COROLLARY 6. *Let $\boldsymbol{\mathcal{X}}^\star \in \mathcal{L}_{\mathbf{r}}$ and $\mathbf{y} = \mathscr{A}(\boldsymbol{\mathcal{X}}^\star)$. If $\mathscr{A}$ has an RIC $\delta_{\bar{\mathbf{r}}_p} < 2^{-P}$, then the scheme (20) with fixed step size $\mu_k = 1$ and MPO $\pi = (p, p_2, \ldots, p_P)$ converges to $\boldsymbol{\mathcal{X}}^\star$. If the step size formula (17) is used, then the same result holds with $\delta_{\bar{\mathbf{r}}_p} < 1/(2^{P+1}+1)$.*

Proof. The proof follows from taking $k \to \infty$ in (29) with $\boldsymbol{\mathcal{X}}^\star = \boldsymbol{\mathcal{X}}^\star_{\mathbf{r}}$ and $\mathbf{e} = \mathbf{0}$. ∎

Ideally, mrank-based recovery results should assume a small RIC of order $(dR_1, \ldots, dR_P)$ for some constant d. But, just as in Theorem 3, our results rely

[8] The first component of π was chosen as $p_1 = p$ to simplify the writing of the theorem and its demonstration.

instead on an RIC of order $(N_1, \ldots, N_{p-1}, 3R_p, N_{p+1}, \ldots, N_P)$. Consequently, they unfortunately do not improve upon currently known sampling bounds. Indeed, applying [38, Th. 2] with $\delta = 2^{-P}$ for fixed P, the RIC condition in Theorem 5 is met with high probability provided that one takes[9]

$$M \geq M_{\min} = \mathcal{O}(R_p \bar{N}_p + R_p N_p + \textstyle\sum_{q \neq p} N_q^2) \tag{30}$$

subgaussian measurements, which grows much faster than the model complexity $\Phi(\mathbf{r})$ (see (4)). Nevertheless, our numerical simulations of subsection 5.3 will show that in practice $M_{\min} = \mathcal{O}(\Phi(\mathbf{r})) = \mathcal{O}(\prod_p R_p + \sum_p N_p R_p)$ Gaussian measurements are sufficient for achieving recovery with SeMPIHT. We note that the same is true also for both TIHT and MnRA. Formally demonstrating such an observed (near-)optimality of SeMPIHT remains an open problem.

It is also important to bear in mind that, since our recovery guarantees are RIC-based, they do not apply to TC, because sampling operators cannot possess small RICs (a simple counterexample for the matrix case is given in [4] which can be easily extended to TC). When using uniformly distributed SOs, the analysis typically requires imposing certain incoherence conditions (similar to, e.g., those in [5, 26]) on the target low-mrank tensors, in order to, roughly speaking, avoid a high concentration of the tensor energy in a small number of entries. The motivation is guaranteeing that any set of sampled entries be sufficiently informative, which is not the case, for instance, when a sparse tensor is uniformly sampled.

3.5. Step size selection and stopping criteria. As emphasized in [43], the issue of step size selection is of great importance when using IHT. On the one hand, μ_k should be sufficiently large to accelerate convergence and diminish the occurrence of convergence to local minima. In particular, the requirement $J_k(x) > \mu_k J(x)$ for all $x \neq x_{k-1}$ can be relaxed, since it is sufficient but not necessary for having objective function decrease. On the other hand, too large steps may cause the algorithm to diverge. In addition, invariance with respect to the scaling of the MO is desirable, which is not possible with a fixed step size. To pursue these requirements, some adaptive step size strategies have been proposed in the literature [2, 43, 19, 38].

Upon evaluation of the SeMPIHT iteration (20) with fixed step size $\mu_k = 1$ through computer experiments, one observes that the algorithm is sensitive to the scaling of the used MO, and recovery is only achieved under a highly favorable regime. Furthermore, convergence can be impractically slow.

Our first approach to overcoming these problems consists in employing the ISS heuristic proposed in [19]. This heuristic was motivated by the poor performance displayed by TIHT when the formula in (16) is employed. It consists in imposing a lower bound and an upper bound on μ_k, namely,

$$\alpha\omega(\mu_k) \leq \mu_k < \omega(\mu_k) \triangleq \frac{\|\boldsymbol{\mathcal{X}}_k - \boldsymbol{\mathcal{X}}_{k-1}\|_F^2}{\|\mathcal{A}(\boldsymbol{\mathcal{X}}_k - \boldsymbol{\mathcal{X}}_{k-1})\|_2^2}, \tag{31}$$

for $\alpha \in]0, 1[$. The upper bound is similar to that proposed in [2] for CS, aiming at achieving objective function decrease. In its turn, the lower bound is meant to avoid

[9] Reference [38, Th. 2] states that $\delta_{\mathbf{r}} \leq \delta$ if a bound of the form $M \geq \mathcal{O}(\delta^{-2}(R^P + PNR)\log(P))$ is met, where $\mathbf{r} = (R_1, \ldots, R_P)$, $R = \max_p R_p$, and $N = \max_p N_p$. Nonetheless, an inspection of its proof reveals that this bound can be refined as $M \geq \mathcal{O}(\delta^{-2}(\prod_p R_p + \sum_p N_p R_p)\log(P))$, which for fixed P and $\mathbf{r} = \bar{\mathbf{r}}_p = (N_1, \ldots, N_{p-1}, 3R_p, N_{p+1}, \ldots, N_P)$ implies (30). The refinement of the term PNR is mentioned in [38].

too small values for μ_k. As a first candidate step size, we use here the initial guess $\mu_k = 1$ (unlike [19], which uses TIHT's formula) and then keep it if it satisfies (31). Otherwise, a new candidate step size given by $\beta\omega(\mu_k)$ is generated, where $\beta \in]\alpha, 1[$, its corresponding estimate $\boldsymbol{\mathcal{X}}_k$ is computed, and the verification is repeated. This process is interrupted if a maximum number of generated candidates (denoted as L in [19]) is attained, and then the largest step satisfying at least the upper bound is kept. If none of the generated candidate step sizes satisfies that upper bound, then the smallest one is repeatedly divided by $\kappa > 1$ until it does, similarly to [2].

Based on the above description, we conclude that the extra cost depends on the number of candidate step sizes generated until one is accepted. For each additional candidate, stage (iii) mentioned in subsection 3.2 must be performed, followed by application of SeMP. Assuming that at least one of the first L generated candidates satisfies its upper bound (which was always the case in our simulations), the extra cost is thus given in the worst case by $L-1$ times the cost of these two operations.

A competitive alternative to ISS is based on the step size selection rule (17) proposed in [38]. This expression is the higher-order analogue of that used in the matrix NIHT algorithm of [43]. Here, the gradient undergoes a multilinear transformation so that each mode is projected onto the corresponding modal subspace of $\boldsymbol{\mathcal{X}}_{k-1}$. In the case of SeMPIHT, note that each $\mathbf{U}^{(p)}$ in (17) must be replaced by the matrix $\bar{\mathbf{U}}^{(p)}$ computed by SeMP (see Algorithm 1) at iteration $k-1$. As in [43], this is motivated by the expectation that little change occurs from one iterate to another in terms of those subspaces, in which case (17) is approximately optimal. The cost implied by its use is of

$$\mathcal{O}\left(\sum_{p=1}^{P} R_1 \dots R_p N_p \dots N_P\right) + \mathcal{O}\left(\sum_{p=1}^{P} N_1 \dots N_p R_p \dots R_P\right) + c_{\mathscr{A}} + \mathcal{O}(M) + \mathcal{O}(\bar{N})$$

operations, where the first two terms are associated with the multilinear transformation in (17), $c_{\mathscr{A}}$ denotes the cost of applying the MO $\mathscr{A}$ (as discussed in subsection 3.2), and the last two terms are related to the calculation of the norms.

For convenience, we give a concrete description of SeMPIHT with adaptive step size in Algorithm 1. Two stopping criteria are used. At each iteration k we check whether the condition

$$\|\boldsymbol{\mathcal{X}}_k - \boldsymbol{\mathcal{X}}_{k-1}\|_F \leq \epsilon \|\boldsymbol{\mathcal{X}}_{k-1}\|_F, \tag{32}$$

with $\epsilon > 0$, is satisfied for two consecutive estimates. If so, convergence is declared, and the algorithm stops. Otherwise, it keeps running until a maximum number of iterations $K_{\max}$ is met.

4. Performance improvement with gradual rank increase. More often than not, tensors measured in applications possess modal singular spectra which decay steadily, instead of having an exactly low mrank. In that case, gradually increasing the mrank of the estimated model along iterations can improve recovery [31]. We pursue this idea here, proposing a continuation technique, called gradual rank increase (GRI), which starts off with a small mrank and conducts the algorithm through increasingly complex estimates.

There are several ways in which one can implement a GRI scheme. A fairly simple one starts with a chosen mrank $\mathbf{r}_1$ and then runs Algorithm 1 for a maximum of $K'_{\max} < K_{\max}$ iterations or until (32) is satisfied. The outcome $\hat{\boldsymbol{\mathcal{X}}}^{\star}_{\mathbf{r}_1}$ is then used to initialize a subsequent run in which the mrank components are set as $[\mathbf{r}_2]_p = \min\{[\mathbf{r}_{\max}]_p, [\mathbf{r}_1 + \mathbf{i}]_p\}$ for all p, where $\mathbf{i} \in \mathbb{N}^P$ is a prescribed increment and $\mathbf{r}_{\max}$ is

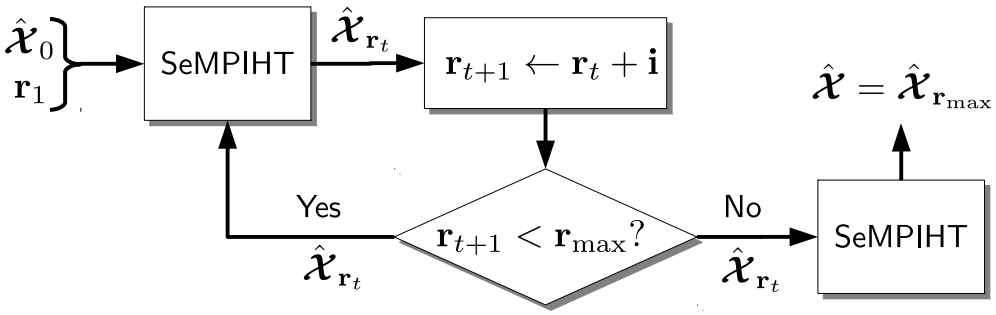

FIG. 1. *Diagram of SeMPIHT algorithm with GRI heuristic.*

the (final) target mrank. This process is repeated until $\mathbf{r}_{\max}$ is reached, at which point a final run is performed, as depicted in Figure 1. Note that a sequence of increasingly complex estimates $\hat{\boldsymbol{\mathcal{X}}}^{\star}_{\mathbf{r}_t}$, $t = 1, 2, \ldots$, is produced before outputting $\hat{\boldsymbol{\mathcal{X}}}^{\star} = \hat{\boldsymbol{\mathcal{X}}}^{\star}_{\mathbf{r}_{\max}}$.

A disadvantage of the above scheme is that one cannot separately control the iteration at which each mrank component is incremented. If, e.g., $\mathbf{i} = \mathbf{1}$ and $[\mathbf{r}_{\max}]_p \ll [\mathbf{r}_{\max}]_q$, then the algorithm reaches $[\mathbf{r}_{\max}]_p$ many iterations before reaching $[\mathbf{r}_{\max}]_q$. But we would rather assign to each component a growth rate proportional to its magnitude. To this end, we can check the convergence of each modal subspace basis matrix $\mathbf{U}^{(p)}$ separately. An even simpler strategy is to predefine modal rank profiles specifying values for the mrank components at each iteration, until attaining the target mrank at iteration $\bar{K}_{\max}$. From that point, normal operation is resumed. For instance, if $\mathbf{r}_{\max} = (R, 2R, 10R)$, then one can increment $[\mathbf{r}]_p$ by one unit at every $10R/[\mathbf{r}_{\max}]_p$ iterations, so that $\mathbf{r}_{\max}$ is attained at iteration $\bar{K}_{\max} = 10R$.

5. Simulation results. In the following, we thoroughly evaluate SeMPIHT and compare it with other LRTR/TC algorithms by means of computer simulations. For simplicity, our simulations concern only the recovery of third-order tensors (i.e., $P = 3$). We note also that the ISS heuristic is always employed with parameters $L = 3$, $\alpha = 0.5$, $\beta = 0.7$ (see subsection 3.5). All reported experiments were performed in MATLAB R2013a running on an Intel Xeon ES-2630v2 2.60 GHz with 32 GB of 1866 MHz RAM memory.

The main performance criterion used in our experiments is the normalized squared error (NSE). Given a tensor of interest $\boldsymbol{\mathcal{X}}^{\star}$ and an estimate $\hat{\boldsymbol{\mathcal{X}}}^{\star}$ obtained by applying a recovery algorithm to some measurement vector $\mathbf{y} = \mathscr{A}(\boldsymbol{\mathcal{X}}^{\star})$, we define

$$\text{NSE}(\hat{\boldsymbol{\mathcal{X}}}^{\star}; \boldsymbol{\mathcal{X}}^{\star}) \triangleq \frac{\|\boldsymbol{\mathcal{X}}^{\star} - \hat{\boldsymbol{\mathcal{X}}}^{\star}\|_F^2}{\|\boldsymbol{\mathcal{X}}^{\star}\|_F^2}. \tag{33}$$

When recovery is performed for N_r realizations, providing N_r pairs $(\hat{\boldsymbol{\mathcal{X}}}^{\star}_l, \boldsymbol{\mathcal{X}}^{\star}_l)$, $l \in [\![N_r]\!]$, we often employ the normalized (sample) mean square error (NMSE)

$$\text{NMSE}(\hat{\boldsymbol{\mathcal{X}}}^{\star}; \boldsymbol{\mathcal{X}}^{\star}) = \frac{1}{N_r} \sum_{l=1}^{N_r} \text{NSE}(\hat{\boldsymbol{\mathcal{X}}}^{\star}_l; \boldsymbol{\mathcal{X}}^{\star}_l), \tag{34}$$

whose arguments may be omitted for simplicity whenever they are clear from the context.

5.1. Tensor models. Two types of synthetic tensors are considered in our experiments:

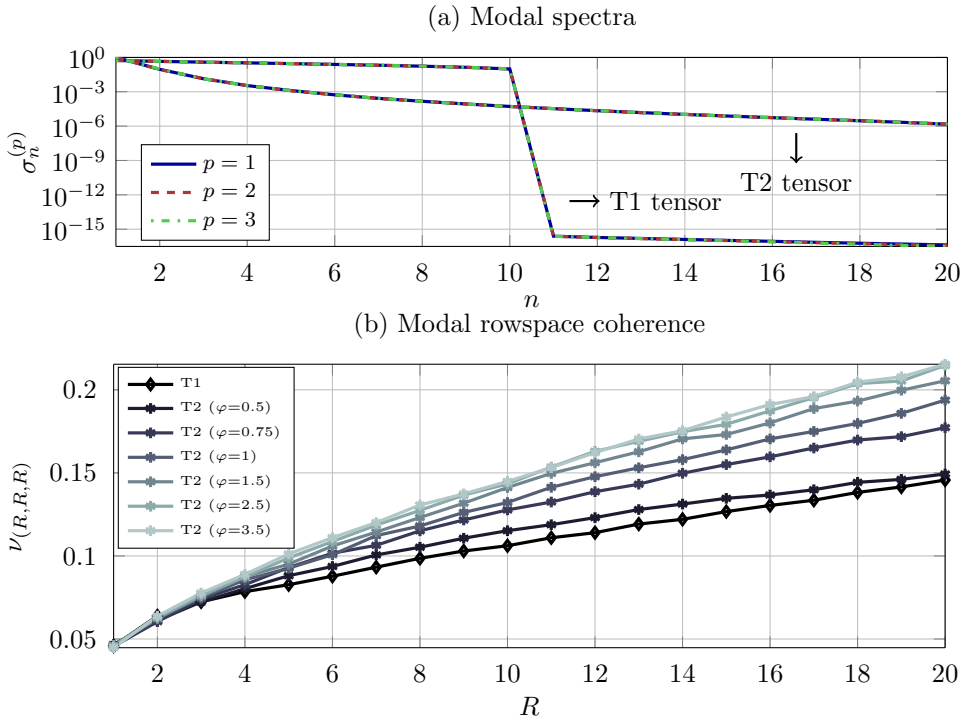

FIG. 2. *Typical behavior of the considered random tensor models:* (a) *modal singular spectra;* (b) *row space coherence (see* (35)*).*

- A *type*-1 (T1) tensor consists of a Tucker model having an $R_1 \times \cdots \times R_P$ core and $N_p \times R_p$ factors. Thus, by construction, it belongs to $\mathcal{L}_{\mathbf{r}}$. All factors and the core have standard Gaussian entries. As a consequence, T1 tensors have highly concentrated nonzero modal singular values.
- A *type*-2 (T2) tensor generally has full mrank but exhibits exponentially decaying modal singular spectra. To impose this property, we adopt the Tucker model used in [49, sec. 2.3], which has an $N_1 \times \cdots \times N_P$ Gaussian core and matrix factors $\mathbf{A}_p = \mathbf{Q}_p\mathbf{S}_p \in \mathbb{R}^{N_p \times N_p}$, where $\mathbf{Q}_p$ is orthogonal and $\mathbf{S}_p = \operatorname{diag}(1, 2^{-\varphi}, \ldots, N_p^{-\varphi})$, with $\varphi > 0$.

The typical spectral characteristics of T1 and T2 tensors are illustrated in Figure 2(a). Specifically, it shows the average modal singular spectra of 500 realizations of $20 \times 20 \times 20$ T1 ($\mathbf{r} = (10, 10, 10)$) and T2 ($\varphi = 3$) tensors, which are normalized to have unit Frobenius norm. The average nth singular value of the mode-p unfolding of the generated tensors is denoted by $\sigma_n^{(p)}$. Numerically, the T2 tensors have full mrank. Consistently with the shown behavior, the mean squared error of the best mrank-(R, R, R) approximation as a function of R displays an abrupt variation for the T1 tensors; that of the T2 tensors decays steadily and smoothly.

Another relevant property of these tensor models is highlighted in Figure 2(b). Namely, we generated 500 realizations of T1 tensors of varying mrank $\mathbf{r} = (R, R, R)$ and T2 tensors with varying φ, all having dimensions $20 \times 20 \times 20$, and then plotted the average modal row space coherence (see [5, 26]) of their approximate projections

onto $\mathcal{L}_{\mathbf{r}}$, i.e.,

(35)
$$\nu_{\mathbf{r}}(\boldsymbol{\mathcal{X}}) = \min_{p} \max_{n\in[\![N]\!]} \left\| \mathcal{P}_{\mathcal{W}_p}\left(\mathbf{e}_n^{(\bar{N}_p)}\right)\right\|_2^2, \qquad \text{where } \begin{cases} \mathcal{W}_p = \text{rowspace}\left((\boldsymbol{\mathcal{X}}_{\mathbf{r}})_{\langle p\rangle}\right), \\ \boldsymbol{\mathcal{X}}_{\mathbf{r}} = \mathcal{S}_{\mathbf{r}}(\boldsymbol{\mathcal{X}}), \end{cases}$$

where $\mathbf{e}_n^{(\bar{N}_p)}$ is the nth canonical basis vector of $\mathbb{R}^{\bar{N}_p}$. Note that $\boldsymbol{\mathcal{X}}_{\mathbf{r}} = \boldsymbol{\mathcal{X}}$ for T1 tensors of mrank $\mathbf{r}$, while for T2 tensors $\boldsymbol{\mathcal{X}}_{\mathbf{r}} \neq \boldsymbol{\mathcal{X}}$. The (approximate) projection is performed because we are ultimately interested in the properties of the best mrank-$\mathbf{r}$ approximation of $\boldsymbol{\mathcal{X}}$, since it is this approximation which is sought by SeMPIHT, the difference $\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_{\mathbf{r}}$ being regarded as a modeling error (cf. Theorem 5). Figure 2(b) indicates that the modal row space coherence of the (approximately) projected T2 tensors grows with φ. Also, the gap among the curves grows with R. As we shall see in what follows, this has important negative implications when trying to complete T2 tensors sampled uniformly at random.

We would like to draw attention to the fact that, although the modal spectra of T2 tensors are more akin to those of most real-world tensors, to date most published works have exclusively considered T1 (or similar) tensors in computer experiments with synthetic data.

5.2. Effect of gradual rank increase. In this section, we discuss the effects of the GRI heuristic by drawing upon experimental results. This allows us to show its motivation and better understand how it works, based on empirical grounds. To this end, we resort to Monte Carlo simulations involving the recovery of $20 \times 20 \times 20$ tensors by employing Algorithm 1 with ISS.

We first employ Gaussian MOs. For each value of $\rho = M/20^3 \in \{0.10, 0.25, 0.40\}$, $N_r = 100$ realizations of an MO $\mathscr{A}$ are generated by drawing the entries of its associated matrix $\mathbf{A} \in \mathbb{R}^{M\times 20^3}$ (such that $\mathscr{A}(\boldsymbol{\mathcal{X}}) = \mathbf{A}\,\text{vec}(\boldsymbol{\mathcal{X}})$) from a zero-mean Gaussian distribution of variance $1/M$. Each MO is then used to sense T1 tensors having mrank (R, R, R), with $R \in [\![15]\!]$, and T2 tensors with spectral decay factors $\varphi \in \{\frac{3}{2}, \frac{7}{2}\}$. When recovering T1 tensors, the target mrank always matches mrank$(\boldsymbol{\mathcal{X}}^\star)$, and we set $K_{\max} = 1000$ and $\epsilon = 10^{-10}$ for the stopping criterion (32). The algorithm is run once initialized with the null tensor (initialization I) and then three more times with random initializations (initialization II).

In the recovery of T2 tensors, we vary the target mrank (R, R, R) and run the algorithm twice for each R: once initialized with the null tensor (initialization I) and once using the solution obtained with mrank $(R-1, R-1, R-1)$ to initialize the run in which $\mathbf{r} = (R, R, R)$ (initialization II). Note that the latter initialization strategy is closely related to our GRI heuristic. Again, $K_{\max} = 1000$, but a specific ϵ was chosen for each combination of φ and ρ by a trial and error procedure.

The NMSE of the estimates provided by SeMPIHT is shown in Figure 3(a),(c),(e). In the case of T1 tensors, only the best outcome among the runs with initialization II is kept for computing (34). For T2 tensors, we also plot NMSE$(\mathcal{S}_{\mathbf{r}}(\boldsymbol{\mathcal{X}}^\star); \boldsymbol{\mathcal{X}}^\star)$, which gives an approximate lower bound. Figure 3(a) displays a sharp transition from success to failure in the recovery of T1 tensors, which is a typical behavior in parsimonious signal recovery problems. Concerning T2 tensors, Figure 3(c),(e) shows that the NMSE gets quite close to the lower bound when inside the region of successful recovery of T1 tensors (cf. Figure 3(a)), regardless of the initialization. Beyond that region, a gap appears: results obtained with initialization I rapidly degrade, while those for initialization II degrade (or even improve) only slightly before stabilizing. The rate of deviation from the lower bound depends on ρ and φ, in conformity with

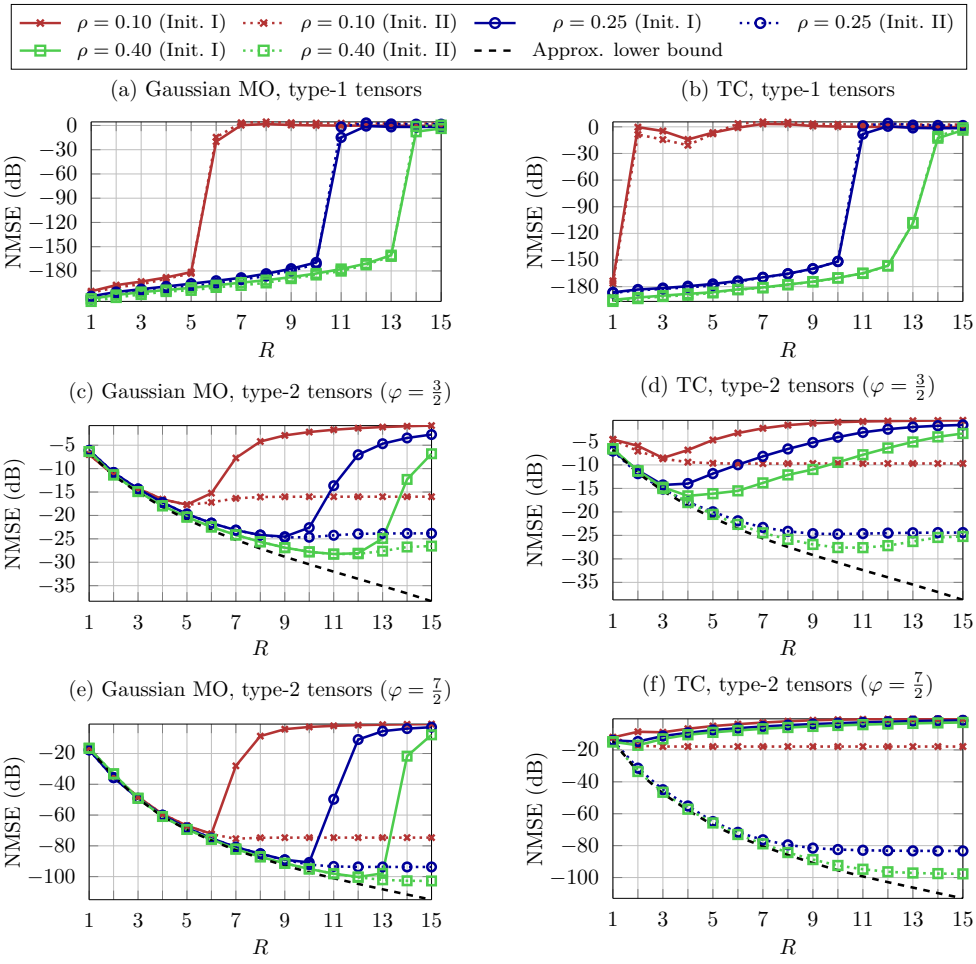

FIG. 3. *Effect of GRI on recovery performance of SeMPIHT with Gaussian (*(a), (c), *and* (e)*) sensing and in TC (*(b), (d), *and* (f)*). The approximate lower bound in* (c)–(f) *is computed as* $NMSE(\mathcal{S}_{\mathbf{r}}(\boldsymbol{\mathcal{X}}^{\star}); \boldsymbol{\mathcal{X}}^{\star})$.

(29).

The results of a similar experiment performed with (uniformly) random SOs are shown in Figure 3(b),(d),(f). As we can see, transition into failure happens in Figure 3(b) for $\rho = 0.10$ as soon as $R = 2$, against $R = 6$ in the Gaussian case. Also, the results are very poor for T2 tensors with initialization I, even in a favorable regime (i.e., where recovery of T1 tensors succeeds). Moreover, the performance worsens as the singular values decay rate φ grows, which is explained by the behavior shown in Figure 2(b), as the recoverability of $\boldsymbol{\mathcal{X}}^{\star}_{\mathbf{r}}$ depends on $\mathbf{r}$ and on some measure of coherence. The use of initialization II does a remarkable job in avoiding such a degradation. Indeed, the results for $\rho = 0.25$ and $\rho = 0.40$ are similar to those obtained with Gaussian sensing. For $\rho = 0.10$, not enough measurements seem to be available for achieving comparable results.

Let us now interpret these outcomes in light of the results of subsection 3.4. Since $\boldsymbol{\mathcal{X}}^{\star} \in \mathcal{L}_{\mathbf{r}}$ in our experiment with T1 tensors, Corollary 6 guarantees convergence to

the global minimizer $\boldsymbol{\mathcal{X}}^\star$ whenever $\mathscr{A}$ satisfies the stated RIC condition, regardless of the initialization (and despite the nonconvexity of (6)). Hence, when using Gaussian sensing, initialization plays no role in the recovery regime (with high probability), which is corroborated by Figure 3(a). Our results suggest that, in the phase transition region, the influence of initialization comes into play, as the insufficiency of measurements vis-à-vis the number of DOFs causes convergence to local minima (or inability to converge), with rapidly increasing probability as R grows. Similar remarks hold for T2 tensors, in that the iterates approach a ball centered at a best approximation $\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} \in \mathcal{L}_{\mathbf{r}}$ of $\boldsymbol{\mathcal{X}}^\star$ regardless of the initialization for appropriate $\mathscr{A}$ (cf. Theorem 5), which explains Figure 3(c),(e).

Now, when a too high mrank (with respect to ρ) is chosen to model a T2 tensor, gradually increasing the mrank stabilizes the approximation error, or at least mitigates its degradation. Apparently, this happens because, once the phase transition region is reached, the lack of sufficient information causes convergence to a local minimum not far from the initial point. In particular, when completing T2 tensors, this continuation strategy delivers good results despite their nonideal coherence properties. It also brings computational advantages: ξ is smaller, leading to a faster convergence, and the cost of $\mathcal{S}_{\mathbf{r}}$ is reduced when $\mathbf{r}$ has small components.

5.3. Empirical sampling bounds. In this section, we numerically estimate how many measurements are necessary for recovering a model with a given complexity. More precisely, the idea is to find, for several values of ρ, the maximum normalized number of DOFs $\bar{\Phi}(\mathbf{r}) = \Phi(\mathbf{r})/\bar{N}$ up to which recovery is highly likely. For simplicity, we take $N_1 = N_2 = N_3 = N$ and sort all possible values of $\bar{\Phi}(\mathbf{r})$ by considering every mrank $\mathbf{r} = (R_1, R_2, R_3)$ such that (i) $R_1 \leq R_2 \leq R_3$ and (ii) $R_3 \leq R_1 R_2$. This entails no loss of generality, as constraint (i) avoids redundant tuples, while constraint (ii) eliminates those which are not feasible.[10] Then, for each $\rho \in \{0.05, 0.10, \ldots, 1\}$, we start from the simplest model, $\mathbf{r} = (1, 1, 1)$, and generate 15 joint realizations of an MO $\mathscr{A}$ and a T1 tensor $\boldsymbol{\mathcal{X}}^\star \in \mathcal{L}_{\mathbf{r}}$. Recovery of $\boldsymbol{\mathcal{X}}^\star$ from $\mathbf{y} = \mathscr{A}(\boldsymbol{\mathcal{X}}^\star)$ is declared successful when $\mathrm{NSE}(\hat{\boldsymbol{\mathcal{X}}}^\star; \boldsymbol{\mathcal{X}}^\star) \leq -90$ dB. If all 15 runs are successful, then the process is repeated with the next model of higher complexity (in terms of $\bar{\Phi}(\mathbf{r})$). When failure occurs for some $\mathbf{r}'$, then the value $\bar{\Phi}(\mathbf{r})$ of the immediately less complex model is declared to be the frontier of the recovery region. To reduce computing time, instead of starting from $\mathbf{r} = (1, 1, 1)$ for every level of ρ, we start from the mrank tuple associated with the frontier obtained for the immediately preceding undersampling rate (i.e., for $\rho - 0.05$). The stopping criteria parameters are set as $\epsilon = 10^{-8}$ and $K_{\max} = 1500$. Gaussian MOs and SOs are generated as described in subsection 5.2.

The results obtained for $N \in \{10, 15, 20\}$ are shown in Figure 4. When using Gaussian operators (GOs), the maximum $\bar{\Phi}(\mathbf{r})$ clearly grows approximately linearly with ρ for all N. Moreover, the improvement due to ISS is remarkable, as the slope becomes much higher (about 0.9) than with fixed step size (about 0.17). Hence, $M \geq M_{\min} = \mathcal{O}(\Phi(\mathbf{r}))$ Gaussian measurements (are highly likely to) suffice for recovery, with $M_{\min} \approx \frac{1}{0.9}\Phi(\mathbf{r}) = 1.11\,\Phi(\mathbf{r})$ when using ISS and $M_{\min} \approx \frac{1}{0.17}\Phi(\mathbf{r}) = 5.88\,\Phi(\mathbf{r})$ when $\mu_k = 1$. So, despite the quite loose sampling bounds implied by Theorem 5, in practice SeMPIHT with ISS succeeds for a quasi-optimal number of Gaussian measurements. On the other hand, the relation between $\bar{\Phi}(\mathbf{r})$ and ρ is no longer linear in TC.

For the sake of comparison, the same procedure is applied with $N = 20$ to ISS-

[10] Note that $\mathrm{mrank}(\boldsymbol{\mathcal{X}}) = \mathbf{r}$ is equivalent to the existence of a Tucker model constrained as discussed in section 1, whose core can only have a mode-3 unfolding with orthogonal rows if $R_3 \leq R_1 R_2$.

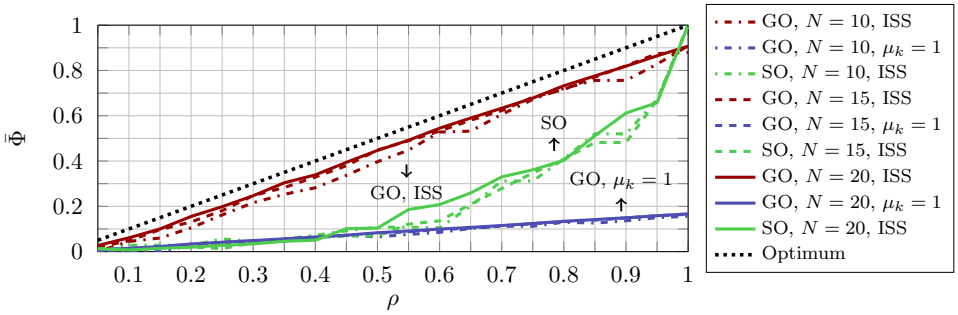

FIG. 4. *Estimated (normalized) number of DOFs which can be recovered by SeMPIHT for each level of ρ, using Gaussian MOs (GO) and sampling MOs (SO). Recovery was successful in all* 15 *realizations for values of $\bar{\Phi}$ below or over the curve.*

TIHT [19], MnRA [52], geomCG [31], TMac [49], and an ADMM scheme based on SNN minimization (SNNM) [44]. In the latter, the penalty parameter η is adapted along the iterations to accelerate convergence, as discussed in [3, sec. 3.4.1], and observations are taken as constraints ($\lambda \to 0$). Having been devised specifically for TC, the performances of geomCG and TMac are only evaluated[11] with SOs. For a fair comparison, a variant of MnRA using ISS is also included. All methods are initialized with the null tensor and cannot exceed $K_{\max} = 1500$ iterations.

This comparison is shown in Figure 5. In the Gaussian sensing setting of Figure 5(a), the sampling requirements of SeMPIHT and ISS-TIHT are almost identical, while those of ISS-MnRA are a little stricter. Though MnRA with fixed step size $\mu_k = 1$ displays quite a poor performance, $\bar{\Phi}$ still grows roughly linearly with ρ. In its turn, the behavior of the SNNM approach is markedly different, abruptly improving in the region $\rho > 0.8$. Such a nonlinear relation is expected, as discussed in subsection 1.1. In the TC scenario of Figure 5(b), SeMPIHT and ISS-TIHT have generally the least strict sampling requirements, with geomCG competing closely for $\rho \geq 0.7$. TMac's performance is less satisfying but slightly better than that of MnRA for $0.7 \leq \rho \leq 0.95$. Here, ISS does not improve MnRA's sampling requirements. Finally, the SNNM approach displays an overwhelmingly poor performance in comparison with the others.

5.4. Convergence and computational cost. In order to evaluate the studied algorithms with respect to their convergence speed and computational cost, they are applied to recover 60 realizations of $N \times N \times N$ T1 and T2 tensors sensed by GOs and SOs. At each iteration, we measure the NSE of the current solution with respect to $\boldsymbol{\mathcal{X}}^\star$ and also the time spent. Results concerning T2 tensors are displayed along with an average (approximate) lower bound calculated as in subsection 5.2. SeMPIHT is run both with the ISS heuristic and with the NTIHT step size selection rule (17). When (and only when) T2 tensors are recovered, SeMPIHT is also run with GRI (in which case the ISS heuristic is used). The tolerance parameter used in geomCG's rank increase condition (cf. [31, eq. 4.2]) is set as $\delta = 0.1$. TMac's adaptive weight heuristic is used, starting with weights $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$ [49]. The ADMM scheme

[11] We employ the implementations provided by their authors, obtained from http://anchp.epfl.ch/geomCG and http://www.math.ucla.edu/~wotaoyin/papers/tmac.html. Yet, we have replaced geomCG's MEX routines by MATLAB code, which turns out to be much faster in our setting (as suggested by [10]).

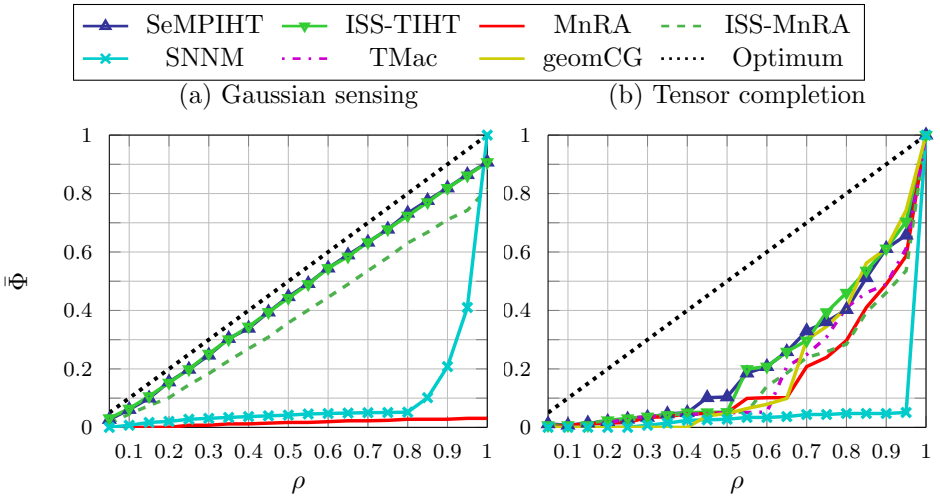

FIG. 5. *Estimated (normalized) number of DOFs which can be recovered by several algorithms for each level of ρ, using Gaussian MOs* (a) *and sampling MOs* (b). *The measured tensors have dimensions $N_p = 20$ for all p. Recovery was successful in all* 15 *realizations for values of $\bar{\Phi}$ below or over the curve.*

for SNNM is again run with $\lambda \to 0$, and the penalty parameter is adapted as described by [3, sec. 3.4.1].

We start by considering GOs. In this case, $N = 20$, $\rho = 0.25$, T1 tensors have mrank $\mathbf{r} = (3, 3, 3)$, and T2 tensors have decay parameter $\varphi = 2.5$. The results for T1 tensors in terms of the NMSE achieved at each iteration are shown in Figure 6(a). The average elapsed time until completion of each iteration is shown on the abscissa. In this scenario, both SeMPIHT (with ISS) and ISS-TIHT outperform the other algorithms, having practically indistinguishable performances. This happens because the cost of applying the Gaussian MO dominates that of the projection. Figure 6(b) displays the results obtained for T2 tensors modeled with the mrank $\mathbf{r} = (9, 9, 9)$. The GRI used in SeMPIHT follows the first procedure described in section 4, with $K'_{\max} = 1$, $\mathbf{i} = (1, 1, 1)$, and $\mathbf{r}_1 = (1, 1, 1)$. One can see that all algorithms reach reasonably close to the bound except for SNNM. Among them, SeMPIHT with GRI is clearly the fastest to converge. Now, in Figure 6(c), the model mrank is set as $\mathbf{r} = (13, 13, 13)$, which yields too high a value of $\Phi(\mathbf{r})$ for $\rho = 0.25$. In this case, we have set $K'_{\max} = 2$. Note that the GRI technique prevents the degradation brought by mrank overestimation, while the performances of the other IHT algorithms are severely deteriorated. This robustness with respect to mrank overestimation is valuable, since in practice one generally does not know which mrank values fall inside the recovery region for a given M.

Figure 7 displays the results obtained for TC, with $N = 300$ and $\rho = 0.2$. The T1 tensors and T2 tensors are generated with, respectively, $\mathbf{r} = (30, 30, 30)$ and $\varphi = 2$. Upon inspection of Figure 7(a), it is clear that the SeMPIHT variants (with ISS and with NTIHT step size selection) are the most efficient in recovering T1 tensors. The gap between SeMPIHT with ISS and ISS-TIHT is due to the reduced cost of the thresholding operator. The NTIHT variant is even faster in this scenario. For the recovery of T2 tensors, the mrank is set as $\mathbf{r} = (90, 90, 90)$, and we choose $K'_{\max} = 1$. Both geomCG and TMac are run with their mrank increase heuristics [31, 49], with

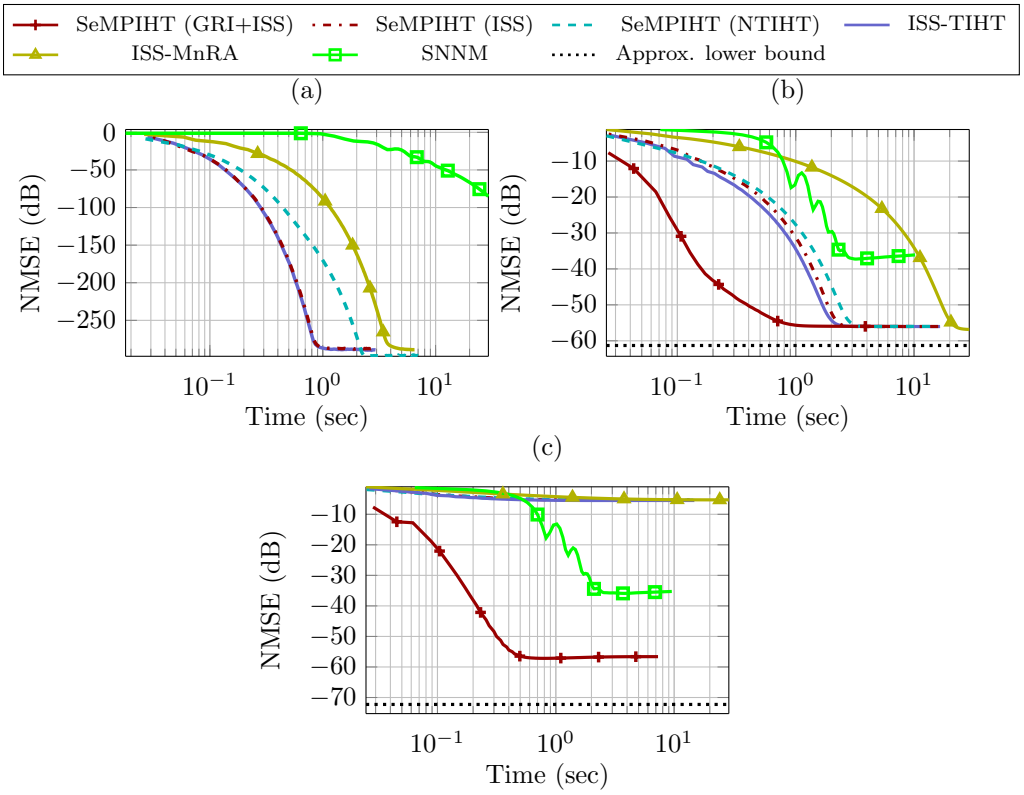

FIG. 6. *Convergence of several algorithms in a Gaussian sensing scenario where* $\rho = 0.25$ *and* $N_1 = N_2 = N_3 = 20$*:* (a) *T*1 *tensors of mrank* $\mathbf{r} = (3, 3, 3)$*;* (b) *T*2 *tensors* ($\varphi = 2.5$) *modeled with mrank* $\mathbf{r} = (9, 9, 9)$*;* (c) *T*2 *tensors* ($\varphi = 2.5$) *modeled with mrank* $\mathbf{r} = (13, 13, 13)$.

initial mrank $\mathbf{r}_1 = (1, 1, 1)$ and unit increments. SeMPIHT uses the same settings. Figure 7(b) shows that the IHT algorithms without GRI clearly fail, which is due to the nonideal coherence properties of the T2 tensors. Among the others, SeMPIHT with GRI provides the best performance, followed by TMac. Unlike the other methods, geomCG's results have large variance due to the occurrence of two realizations with outstandingly poor results. So, we also plot in Figure 7(b) the median of its NSE per iteration, which yields a reasonable behavior in terms of final error, but at a large computing cost.

5.5. Completion of real-world data. Finally, aiming to assess the performance of SeMPIHT in a scenario involving real-world data, we have performed the reconstruction of the hyperspectral image corresponding to the Gualtar scene described in [14], which is shown in Figure 8 for two different wavelengths. (Only the image taken at 11:44am has been used.) This data tensor has dimensions $1024 \times 1344 \times 33$, where the first two modes correspond to the spatial dimensions of the image and the third refers to the number of acquired wavelengths (from 400 to 720 nm at 10-nm intervals). The applied MOs again correspond to a uniformly distributed random sampling of the tensor components, with $\rho \in \{0.15, 0.30, 0.45\}$. We set the model mrank to $\mathbf{r} = (300, 350, 15)$, which gives an approximate lower bound of $\mathrm{NSE}(\mathcal{S}_{\mathbf{r}}(\boldsymbol{\mathcal{X}}^\star); \boldsymbol{\mathcal{X}}^\star) = -40.6$ dB for the methods that explicitly impose low-mrank

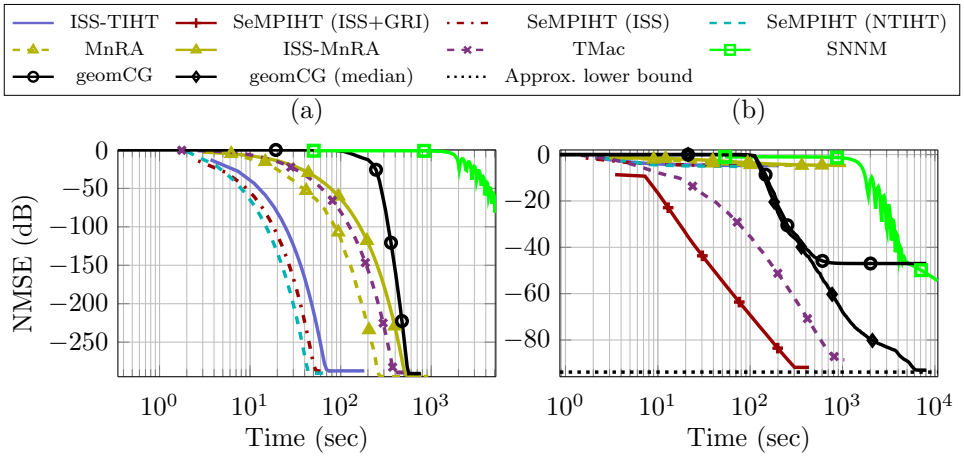

FIG. 7. *Convergence of several algorithms in a TC scenario where* $\rho = 0.2$ *and* $N_1 = N_2 = N_3 = 300$*:* (a) *T*1 *tensors of mrank* $\mathbf{r} = (30, 30, 30)$*;* (b) *T*2 *tensors* ($\varphi = 2$) *modeled with mrank* $\mathbf{r} = (90, 90, 90)$.

FIG. 8. *Hyperspectral image Gualtar scene* [14] *at two wavelengths.*

constraints. Here, the only evaluated IHT method is SeMPIHT (with and without GRI), due to its observed superiority in Figure 7. Among the other methods, geomCG is not included, because it takes too much time when a model having large mrank components is used. The second GRI technique of section 4 is employed for SeMPIHT, so that each mrank component is increased in a quasi-linear fashion with rate proportional to its magnitude until iteration $\bar{K}_{\max} = 150$, from which 50 more iterations are performed with the mrank set at its target. TMac is also run with its rank increasing heuristic, for a maximum of 500 iterations, and the ADMM algorithm for SNN again uses the adaptive penalty parameter for convergence acceleration.

The obtained results are shown in Figure 9. They are displayed in terms of the NSE per iteration, as only a single realization is performed per value of ρ. For $\rho = 0.15$, SeMPIHT with GRI clearly outperforms all other algorithms, converging faster and attaining a smaller NSE. For $\rho = 0.30$ and $\rho = 0.45$, its performance is close to that of TMac. The importance of GRI is very well highlighted, as it significantly accelerates the convergence of SeMPIHT and, furthermore, allows approaching the NSE lower bound for $\rho = 0.15$.

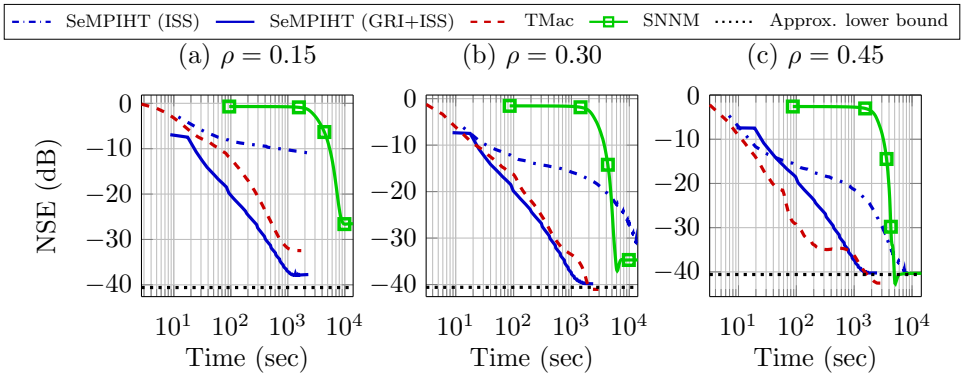

FIG. 9. *Performance of TC algorithms in the reconstruction of the* $1024 \times 1344 \times 33$ *hyperspectral image Gualtar scene* [14]*, which is uniformly sampled at random with subsampling rate* ρ*.*

6. Conclusions. We have proposed an iterative hard thresholding algorithm, called SeMPIHT, to address a constrained least-squares formulation of low-rank tensor recovery in which the solution must have bounded multilinear rank. The employed hard thresholding operator, SeMP, consists of a chain of sequentially optimal modal projections. This yields an approximate projector which enjoys the same quasi-optimality property of the truncated HOSVD while requiring less computing effort. Moreover, the sequential optimality of the modal projections has allowed us to derive a performance bound for SeMPIHT based solely on RIC conditions, which is still an open problem for TIHT. However, the order of the exploited RIC only takes into account the low rank of a single mode, thus leading to loose sampling bounds for certain random (e.g., Gaussian) measurement ensembles. Nonetheless, our systematic empirical evaluation shows that perfect recovery is achieved by SeMPIHT (and also by ISS-TIHT) with a number of Gaussian measurements which scales linearly with the intrinsic complexity of the model, as measured by its number of degrees of freedom. Moreover, the constant governing this linear relation is close to 1, meaning a quasi-optimal recovery performance is observed.

Our numerical studies have validated the theoretical results and have also shown that a gradual rank increase heuristic plays a significant role in achieving good results when the tensor data possess fast decaying modal spectra, stabilizing the estimation error when the model mrank is overestimated and accelerating the algorithm. It is especially important in TC, where it can avoid degradation due to nonideal coherence properties of measured tensors. A simulation scenario involving the completion of a hyperspectral imaging data tensor has further substantiated these observations, corroborating the usefulness of SeMPIHT.

Deriving recovery guarantees based on stricter RICs, in order to theoretically establish the quasi-optimality of SeMPIHT observed for Gaussian sensing, remains an open problem. Another important aspect which necessitates further development is the derivation of the minimum number of measurements needed for tensor completion, which is the most practically relevant scenario.

Appendix A. Proof of our main result. We first state two necessary lemmas and then proceed to the demonstration of Theorem 5.

LEMMA 7. *Let* $\mathbf{U}^{(p)} \in \mathcal{V}_{R_p}(\mathbb{R}^{N_p})$, $p \in [\![P]\!]$, *and define*

$$\mathcal{U} = \left\{ \boldsymbol{\mathcal{X}} : \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{G}} \times_1 \mathbf{U}^{(1)} \times_2 \cdots \times_P \mathbf{U}^{(P)} \text{ for some } \boldsymbol{\mathcal{G}} \in \mathbb{R}^{R_1 \times \cdots \times R_P} \right\} \subset \mathcal{L}_{\mathbf{r}},$$

with $\mathbf{r} = (R_1, \ldots, R_P)$. *Denote* $\mathscr{A}_{\mathcal{U}} = \mathscr{A}\mathcal{P}_{\mathcal{U}}$, *where* $\mathcal{P}_{\mathcal{U}}$ *is the orthogonal projector onto* $\mathcal{U}$, *and assume* $\mathscr{A}$ *has an RIC* $\delta_{\mathbf{r}} < 1$. *Then,* $\|\mathscr{A}_{\mathcal{U}}^{\dagger}\mathscr{A}_{\mathcal{U}} - \mathcal{I}\| \leq \delta_{\mathbf{r}}$, *where* $\mathcal{I}$ *is the identity over* $\boldsymbol{\mathcal{T}}$.

Proof. Our proof is an extension of the argument supporting [15, eq. (6.2)] (given in the context of CS). Consider $\boldsymbol{\mathcal{X}} \in \mathcal{U}$, for which $\mathscr{A}_{\mathcal{U}}(\boldsymbol{\mathcal{X}}) = \mathscr{A}(\boldsymbol{\mathcal{X}})$. By the definition of $\delta_{\mathbf{r}}$, we deduce $\|\mathscr{A}_{\mathcal{U}}(\boldsymbol{\mathcal{X}})\|_F^2 - \|\boldsymbol{\mathcal{X}}\|_F^2 \leq \delta_{\mathbf{r}}\|\boldsymbol{\mathcal{X}}\|_F^2$. Rewriting the left-hand side of this inequality, we obtain

$$\langle \mathscr{A}_{\mathcal{U}}(\boldsymbol{\mathcal{X}}), \mathscr{A}_{\mathcal{U}}(\boldsymbol{\mathcal{X}}) \rangle - \langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \rangle = \langle (\mathscr{A}_{\mathcal{U}}^{\dagger}\mathscr{A}_{\mathcal{U}} - \mathcal{I})(\boldsymbol{\mathcal{X}}), \boldsymbol{\mathcal{X}} \rangle \leq \delta_{\mathbf{r}}\|\boldsymbol{\mathcal{X}}\|_F^2.$$

Assuming that $\|\boldsymbol{\mathcal{X}}\|_F \neq 0$, dividing by $\|\boldsymbol{\mathcal{X}}\|_F^2$, and taking the maximum with respect to $\boldsymbol{\mathcal{X}} \in \mathcal{U} \setminus \{\mathcal{O}\}$ yields

$$\max_{\boldsymbol{\mathcal{X}} \in \mathcal{U} \setminus \{\mathcal{O}\}} \frac{\|\mathscr{A}_{\mathcal{U}}(\boldsymbol{\mathcal{X}})\|_F^2}{\|\boldsymbol{\mathcal{X}}\|_F^2} - 1 = \max_{\boldsymbol{\mathcal{X}} \in \mathcal{U} \setminus \{\mathcal{O}\}} \frac{\langle (\mathscr{A}_{\mathcal{U}}^{\dagger}\mathscr{A}_{\mathcal{U}} - \mathcal{I})(\boldsymbol{\mathcal{X}}), \boldsymbol{\mathcal{X}} \rangle}{\|\boldsymbol{\mathcal{X}}\|_F^2} \leq \delta_{\mathbf{r}}. \tag{36}$$

Now, note that for any $\boldsymbol{\mathcal{Z}} \in \boldsymbol{\mathcal{T}}$, $\|\mathscr{A}_{\mathcal{U}}(\boldsymbol{\mathcal{Z}})\|_F^2 = \|\mathscr{A}_{\mathcal{U}}(\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{Z}}))\|_F^2$ and $\|\boldsymbol{\mathcal{Z}}\|_F^2 \geq \|\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{Z}})\|_F^2$. Consequently, the maximum must be the same over the whole space, because

$$\max_{\boldsymbol{\mathcal{Z}} \neq \mathcal{O}} \frac{\|\mathscr{A}_{\mathcal{U}}(\boldsymbol{\mathcal{Z}})\|_F^2}{\|\boldsymbol{\mathcal{Z}}\|_F^2} \leq \max_{\boldsymbol{\mathcal{Z}} \neq \mathcal{O}} \frac{\|\mathscr{A}_{\mathcal{U}}(\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{Z}}))\|_F^2}{\|\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{Z}})\|_F^2} = \max_{\boldsymbol{\mathcal{X}} \in \mathcal{U} \setminus \{\mathcal{O}\}} \frac{\|\mathscr{A}_{\mathcal{U}}(\boldsymbol{\mathcal{X}})\|_F^2}{\|\boldsymbol{\mathcal{X}}\|_F^2},$$

and therefore (36) implies

$$\max_{\boldsymbol{\mathcal{Z}} \neq \mathcal{O}} \frac{\langle (\mathscr{A}_{\mathcal{U}}^{\dagger}\mathscr{A}_{\mathcal{U}} - \mathcal{I})(\boldsymbol{\mathcal{Z}}), \boldsymbol{\mathcal{Z}} \rangle}{\|\boldsymbol{\mathcal{Z}}\|_F^2} \leq \delta_{\mathbf{r}}.$$

Finally, since $\mathscr{A}_{\mathcal{U}}^{\dagger}\mathscr{A}_{\mathcal{U}} - \mathcal{I}$ is self-adjoint, the left-hand side of the above expression is precisely the definition of its operator norm, and thus the proof is complete. ∎

The next lemma is an extension of [15, Lemma 6.20] (which also applies to CS).

LEMMA 8. *If* $\mathcal{U} \subseteq \mathcal{L}_{\mathbf{r}}$ *and* $\mathscr{A}$ *has an RIC* $\delta_{\mathbf{r}} < 1$, *then for all* $\mathbf{e} \in \mathbb{R}^M$ *we have*

$$\|\mathcal{P}_{\mathcal{U}}\mathscr{A}^{\dagger}(\mathbf{e})\|_F \leq \sqrt{1 + \delta_{\mathbf{r}}}\|\mathbf{e}\|_2.$$

Proof. We assume $\|\mathcal{P}_{\mathcal{U}}\mathscr{A}^{\dagger}(\mathbf{e})\|_F \neq 0$ (otherwise the result is trivial) and start by deriving

$$\|\mathcal{P}_{\mathcal{U}}\mathscr{A}^{\dagger}(\mathbf{e})\|_F^2 = \langle \mathcal{P}_{\mathcal{U}}\mathscr{A}^{\dagger}(\mathbf{e}), \mathcal{P}_{\mathcal{U}}\mathscr{A}^{\dagger}(\mathbf{e}) \rangle = \langle \mathbf{e}, \mathscr{A}\mathcal{P}_{\mathcal{U}}\mathscr{A}^{\dagger}(\mathbf{e}) \rangle \leq \|\mathbf{e}\|_2\|\mathscr{A}\mathcal{P}_{\mathcal{U}}\mathscr{A}^{\dagger}(\mathbf{e})\|_2. \tag{37}$$

Now, by definition of $\delta_{\mathbf{r}}$ (see (11)), $\|\mathscr{A}\mathcal{P}_{\mathcal{U}}\mathscr{A}^{\dagger}(\mathbf{e})\|_F \leq \sqrt{1 + \delta_{\mathbf{r}}}\,\|\mathcal{P}_{\mathcal{U}}\mathscr{A}^{\dagger}(\mathbf{e})\|_F$. Combining this inequality with (37) and dividing both sides by $\|\mathcal{P}_{\mathcal{U}}\mathscr{A}^{\dagger}(\mathbf{e})\|_F$ yields the desired result. ∎

Proof of Theorem 5. For simplicity, we assume, without loss of generality, that $\pi = (p, p_2, \ldots, p_P) = (1, 2, \ldots, P)$. To describe the computation of $\mathcal{S}_{\mathbf{r}}$ at each iteration, we use the notation

$$\begin{aligned} \boldsymbol{\mathcal{V}}_0 &= \boldsymbol{\mathcal{X}}_{k-1} + \mu_k \mathscr{A}^{\dagger}\left(\mathbf{y} - \mathscr{A}\left(\boldsymbol{\mathcal{X}}_{k-1}\right)\right) \\ &= \boldsymbol{\mathcal{X}}_{k-1} + \mu_k \mathscr{A}^{\dagger}\mathscr{A}\left(\boldsymbol{\mathcal{X}}_{\mathbf{r}}^{\star} - \boldsymbol{\mathcal{X}}_{k-1}\right) + \mu_k \mathscr{A}^{\dagger}(\mathscr{A}(\boldsymbol{\mathcal{X}}^{\star} - \boldsymbol{\mathcal{X}}_{\mathbf{r}}^{\star}) + \mathbf{e}), \end{aligned} \tag{38}$$

$(\boldsymbol{\mathcal{V}}_p)_{\langle p\rangle} = \mathscr{H}_{R_p}((\boldsymbol{\mathcal{V}}_{p-1})_{\langle p\rangle})$, where $\mathscr{H}_{R_p}$ is the same as in (13), with $R = R_p$, and $\boldsymbol{\mathcal{X}}_k = \mathcal{S}_{\mathbf{r}}(\boldsymbol{\mathcal{V}}_0) = \boldsymbol{\mathcal{V}}_P$. The result is then obtained by bounding the errors of the approximations $\boldsymbol{\mathcal{V}}_1, \ldots, \boldsymbol{\mathcal{V}}_P$. First, note that

$$(\boldsymbol{\mathcal{V}}_p)_{\langle p\rangle} \in \underset{\operatorname{rank}(\mathbf{Z})\leq R_p}{\arg\min} \left\|\mathbf{Z} - (\boldsymbol{\mathcal{V}}_{p-1})_{\langle p\rangle}\right\|_F \implies$$
$$\forall \boldsymbol{\mathcal{Z}} \in \mathcal{L}_{\mathbf{r}}, \|\boldsymbol{\mathcal{V}}_p - \boldsymbol{\mathcal{V}}_{p-1}\|_F \leq \|\boldsymbol{\mathcal{Z}} - \boldsymbol{\mathcal{V}}_{p-1}\|_F,$$

which, together with $\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} \in \mathcal{L}_{\mathbf{r}}$, implies

$$\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{V}}_p\|_F \leq \|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{V}}_{p-1}\|_F + \|\boldsymbol{\mathcal{V}}_p - \boldsymbol{\mathcal{V}}_{p-1}\|_F \leq 2\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{V}}_{p-1}\|_F.$$

Therefore, as $\boldsymbol{\mathcal{X}}_k = \boldsymbol{\mathcal{V}}_P$, iterating over this inequality for $p = 2, \ldots, P$, we deduce

$$\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_k\|_F \leq 2^{P-1}\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{V}}_1\|_F. \tag{39}$$

Now, to bound $\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{V}}_1\|_F$, we employ the same reasoning as in [18, Lemma 4.1]. Let

$$\mathcal{U} = \left\{\boldsymbol{\mathcal{Z}} : \operatorname{colspace}(\mathbf{Z}_{\langle 1\rangle}) \subset \operatorname{colspace}\left((\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}})_{\langle 1\rangle}\right) + \operatorname{colspace}\left((\boldsymbol{\mathcal{V}}_1)_{\langle 1\rangle}\right) + \operatorname{colspace}\left((\boldsymbol{\mathcal{X}}_{k-1})_{\langle 1\rangle}\right)\right\},$$

so that $\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}}, \boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{X}}_{k-1} \in \mathcal{U} \subset \mathcal{L}_{\bar{\mathbf{r}}_1}$. We thus have

$$\begin{aligned}\|\boldsymbol{\mathcal{V}}_1 - \boldsymbol{\mathcal{V}}_0\|_F^2 &= \|\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{V}}_1 - \boldsymbol{\mathcal{V}}_0)\|_F^2 + \|\mathcal{P}_{\mathcal{U}^\perp}(\boldsymbol{\mathcal{V}}_1 - \boldsymbol{\mathcal{V}}_0)\|_F^2 \\ &= \|\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{V}}_1 - \boldsymbol{\mathcal{V}}_0)\|_F^2 + \|\mathcal{P}_{\mathcal{U}^\perp}(\boldsymbol{\mathcal{V}}_0)\|_F^2 \end{aligned} \tag{40}$$

and also

$$\|\boldsymbol{\mathcal{V}}_1 - \boldsymbol{\mathcal{V}}_0\|_F^2 \leq \|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{V}}_0\|_F^2 = \|\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{V}}_0)\|_F^2 + \|\mathcal{P}_{\mathcal{U}^\perp}(\boldsymbol{\mathcal{V}}_0)\|_F^2, \tag{41}$$

which follows from $(\boldsymbol{\mathcal{V}}_1)_{\langle 1\rangle} = \mathscr{H}_{R_1}((\boldsymbol{\mathcal{V}}_0)_{\langle 1\rangle})$ and $\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} \in \mathcal{L}_{\mathbf{r}} \cap \mathcal{U}$. Combining (40) and (41), we obtain

$$\|\boldsymbol{\mathcal{V}}_1 - \mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{V}}_0)\|_F = \|\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{V}}_1 - \boldsymbol{\mathcal{V}}_0)\|_F \leq \|\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{V}}_0)\|_F = \|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{V}}_0)\|_F.$$

Hence, using the above equation and (38), we have

$$\begin{aligned}\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{V}}_1\|_F \leq\ & \|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{V}}_0)\|_F + \|\boldsymbol{\mathcal{V}}_1 - \mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{V}}_0)\|_F \\ &\leq 2\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{V}}_0)\|_F = 2\|\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{V}}_0)\|_F \\ &= 2\left\|\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1}) - \mu_k\mathcal{P}_{\mathcal{U}}\mathscr{A}^\dagger\mathscr{A}(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1}) - \mu_k\mathcal{P}_{\mathcal{U}}\mathscr{A}^\dagger(\mathscr{A}(\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}^\star_{\mathbf{r}}) + \mathbf{e})\right\|_F \\ &= 2\|(1-\mu_k)\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1}) - \mu_k\mathcal{P}_{\mathcal{U}}\left(\mathscr{A}^\dagger\mathscr{A} - \mathcal{J}\right)(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1}) \\ &\qquad - \mu_k\mathcal{P}_{\mathcal{U}}\mathscr{A}^\dagger(\mathscr{A}(\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}^\star_{\mathbf{r}}) + \mathbf{e})\|_F \\ &\leq 2|1-\mu_k|\,\|\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1})\|_F + 2\mu_k\left\|\mathcal{P}_{\mathcal{U}}(\mathscr{A}^\dagger\mathscr{A} - \mathcal{J})(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1})\right\|_F \\ &\quad + 2\mu_k\left\|\mathcal{P}_{\mathcal{U}}\mathscr{A}^\dagger(\mathscr{A}(\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}^\star_{\mathbf{r}}) + \mathbf{e})\right\|_F. \end{aligned} \tag{42}$$

It follows from the nonexpansiveness of $\mathcal{P}_{\mathcal{U}}$ that $\|\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1})\|_F \leq \|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1}\|_F$. By noting that $\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}}, \boldsymbol{\mathcal{X}}_{k-1} \in \mathcal{U}$ and using the notation $\mathscr{A}_{\mathcal{U}} = \mathscr{A}\mathcal{P}_{\mathcal{U}}$, we have also

$$\begin{aligned}\mathcal{P}_{\mathcal{U}}(\mathscr{A}^\dagger\mathscr{A} - \mathcal{J})(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1}) &= \mathcal{P}_{\mathcal{U}}(\mathscr{A}^\dagger\mathscr{A} - \mathcal{J})\mathcal{P}_{\mathcal{U}}(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1}) \\ &= (\mathscr{A}_{\mathcal{U}}^\dagger\mathscr{A}_{\mathcal{U}} - \mathcal{J})(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1}).\end{aligned}$$

Thus, from Lemma 7 and the fact that $\mathcal{U} \subset \mathcal{L}_{\bar{\mathbf{r}}_1}$ we derive the bound

$$\left\|\mathcal{P}_{\mathcal{U}}(\mathscr{A}^\dagger\mathscr{A} - \mathcal{J})\left(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1}\right)\right\|_F = \left\|(\mathscr{A}^\dagger_{\mathcal{U}}\mathscr{A}_{\mathcal{U}} - \mathcal{J})\left(\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1}\right)\right\|_F \leq \delta_{\bar{\mathbf{r}}_1}\left\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1}\right\|_F.$$

Finally, resorting to Lemma 8, the last term of (42) can be bounded as

$$\left\|\mathcal{P}_{\mathcal{U}}\mathscr{A}^\dagger\left(\mathscr{A}\left(\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}^\star_{\mathbf{r}}\right) + \mathbf{e}\right)\right\|_F \leq \sqrt{1+\delta_{\bar{\mathbf{r}}_1}}\left\|\mathscr{A}(\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}^\star_{\mathbf{r}}) + \mathbf{e}\right\|_2.$$

The above inequalities, combined with (39), yield

$$\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_k\|_F \leq \xi_k\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_{k-1}\|_F + 2^P\mu_k\sqrt{1+\delta_{\bar{\mathbf{r}}_1}}\|\mathscr{A}\left(\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}^\star_{\mathbf{r}}\right) + \mathbf{e}\|_2,$$

where $\xi_k \triangleq 2^P(|1-\mu_k| + \mu_k\delta_{\bar{\mathbf{r}}_1})$. We consider two choices of step size:

- For $\mu_k = 1$, the assumption $\delta_{\bar{\mathbf{r}}_1} < 2^{-P}$ implies $\xi_k = 2^P\delta_{\bar{\mathbf{r}}_1} < 1$.
- If (17) is employed, it follows from the definition of the RIC that $(1+\delta_{\bar{\mathbf{r}}_1})^{-1} \leq \mu_k \leq (1-\delta_{\bar{\mathbf{r}}_1})^{-1}$. We then have two cases: (i) if $\mu_k > 1$, then $|1-\mu_k| + \mu_k\delta_{\bar{\mathbf{r}}_1} = \mu_k(1+\delta_{\bar{\mathbf{r}}_1}) - 1 \leq 2\delta_{\bar{\mathbf{r}}_1}(1-\delta_{\bar{\mathbf{r}}_1})^{-1}$; (ii) similarly, if $\mu_k \leq 1$, then $|1-\mu_k|+\mu_k\delta_{\bar{\mathbf{r}}_1} = \mu_k(\delta_{\bar{\mathbf{r}}_1}-1)+1 \leq 2\delta_{\bar{\mathbf{r}}_1}(1+\delta_{\bar{\mathbf{r}}_1})^{-1} \leq 2\delta_{\bar{\mathbf{r}}_1}(1-\delta_{\bar{\mathbf{r}}_1})^{-1}$. It can be checked that the condition $\delta_{\bar{\mathbf{r}}_1} < 1/(2^{P+1}+1)$ implies $2\delta_{\bar{\mathbf{r}}_1}(1-\delta_{\bar{\mathbf{r}}_1})^{-1} < 2^{-P}$, thus yielding $\xi_k < 1$ in both cases.

Defining $\xi \triangleq \sup_k \xi_k < 1$, it follows that

$$\begin{aligned}\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_k\|_F &\leq \xi^k\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_0\|_F + \left(\sum_{l=0}^{k-1}\xi^l\right)2^P\sqrt{1+\delta_{\bar{\mathbf{r}}_1}}\|\mathscr{A}(\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}^\star_{\mathbf{r}}) + \mathbf{e}\|_2 \\ &\leq \xi^k\|\boldsymbol{\mathcal{X}}^\star_{\mathbf{r}} - \boldsymbol{\mathcal{X}}_0\|_F + 2^P\frac{\sqrt{1+\delta_{\bar{\mathbf{r}}_1}}}{1-\xi}\left\|\mathscr{A}(\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}^\star_{\mathbf{r}}) + \mathbf{e}\right\|_2,\end{aligned}$$

as claimed. To conclude, note that the same reasoning holds for any other MPO $\pi = (p, p_2, \ldots, p_P)$, in which case the role of $\delta_{\bar{\mathbf{r}}_1}$ is played more generally by $\delta_{\bar{\mathbf{r}}_p}$. ∎

REFERENCES

[1] T. Blumensath and M. E. Davies, *Iterative thresholding for sparse approximations*, J. Fourier Anal. Appl., 14 (2008), pp. 629–654.

[2] T. Blumensath and M. E. Davies, *Normalized iterative hard thresholding: Guaranteed stability and performance*, IEEE J. Sel. Topics Signal Process., 4 (2010), pp. 298–309.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Machine Learning, 3 (2011), pp. 1–122.

[4] E. J. Candès and Y. Plan, *Matrix completion with noise*, Proc. IEEE, 98 (2010), pp. 925–936.

[5] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math., 9 (2009), pp. 717–772.

[6] E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203–4215.

[7] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, *The convex geometry of linear inverse problems*, Found. Comput. Math., 12 (2012), pp. 805–849.

[8] P. L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, New York, 2011, pp. 185–212.

[9] P. Comon, J. M. F. ten Berge, L. De Lathauwer, and J. Castaing, *Generic and typical ranks of multi-way arrays*, Linear Algebra Appl., 430 (2009), pp. 2997–3007.

[10] C. Da Silva and F. J. Herrmann, *Optimization on the hierarchical Tucker manifold—applications to tensor completion*, Linear Algebra Appl., 481 (2015), pp. 131–173.

[11] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278, https://doi.org/10.1137/S0895479896305696.

[12] V. de Silva and L.-H. Lim, *Tensor rank and the ill-posedness of the best low-rank approximation problem*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1084–1127, https://doi.org/10.1137/06066518X.

[13] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

[14] D. H. Foster, K. Amano, and S. M. Nascimento, *Time-lapse ratios of cone excitations in natural scenes*, Vision Res., 120 (2016), pp. 45–60.

[15] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Appl. Numer. Harmon. Anal. 2013, Birkhäuser/Springer, New York, 2013.

[16] S. Friedland and L.-H. Lim, *Nuclear Norm of Higher-Order Tensors*, preprint, https://arxiv.org/abs/1410.6072, 2014.

[17] S. Gandy, B. Recht, and I. Yamada, *Tensor completion and low-n-rank tensor recovery via convex optimization*, Inverse Problems, 27 (2011), 025010.

[18] D. Goldfarb and S. Ma, *Convergence of fixed-point continuation algorithms for matrix rank minimization*, Found. Comput. Math., 11 (2011), pp. 183–210.

[19] J. H. de. M. Goulart and G. Favier, *An iterative hard thresholding algorithm with improved convergence for low-rank tensor recovery*, in European Signal Processing Conference (EUSIPCO), Nice, France, 2015.

[20] L. Grasedyck, *Hierarchical singular value decomposition of tensors*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2029–2054, https://doi.org/10.1137/090764189.

[21] L. Grasedyck, M. Kluge, and S. Krämer, *Variants of alternating least squares tensor completion in the tensor train format*, SIAM J. Sci. Comput., 37 (2015), pp. A2424–A2450, https://doi.org/10.1137/130942401.

[22] L. Grasedyck, D. Kressner, and C. Tobler, *A literature survey of low-rank tensor approximation techniques*, GAMM-Mitt., 36 (2013), pp. 53–78.

[23] N. Halko, P.-G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288, https://doi.org/10.1137/090771806.

[24] C. J. Hillar and L.-H. Lim, *Most tensor problems are NP-hard*, J. ACM, 60 (2013), article 45.

[25] F. L. Hitchkock, *The expression of a tensor or a polyadic as a sum of products*, J. Math. Phys., 6 (1927), pp. 164–189.

[26] B. Huang, C. Mu, D. Goldfarb, and J. Wright, *Provable low-rank tensor recovery*, Pacific J. Optim., 11 (2015), pp. 339–364.

[27] P. Jain, R. Meka, and I. S. Dhillon, *Guaranteed rank minimization via singular value projection*, in Proceedings of the 23rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 2010, pp. 937–945.

[28] H. Kasai and B. Mishra, *Riemannian Preconditioning for Tensor Completion*, preprint, https://arxiv.org/abs/1506.02159, 2015.

[29] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500, https://doi.org/10.1137/07070111X.

[30] N. Kreimer, A. Stanton, and M. D. Sacchi, *Tensor completion based on nuclear norm minimization for 5D seismic data reconstruction*, Geophysics, 78 (2013), pp. V273–V284.

[31] D. Kressner, M. Steinlechner, and B. Vandereycken, *Low-rank tensor completion by Riemannian optimization*, BIT, 54 (2014), pp. 447–468.

[32] A. Kyrillidis and V. Cevher, *Matrix recipes for hard thresholding methods*, J. Math. Imaging Vision, 48 (2014), pp. 235–265.

[33] J. Liu, P. Musialski, P. Wonka, and J. Ye, *Tensor completion for estimating missing values in visual data*, in Proceedings of the 12th IEEE International Conference on Computer Vision, Kyoto, Japan, 2009, pp. 2114–2121.

[34] J. Liu, P. Musialski, P. Wonka, and J. Ye, *Tensor completion for estimating missing values in visual data*, IEEE Trans. Pattern Anal. Mach. Intell., 35 (2013), pp. 208–220.

[35] C. Mu, B. Huang, J. Wright, and D. Goldfarb, *Square deal: Lower bounds and improved relaxations for tensor recovery*, in Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014, pp. 73–81.

[36] H. Rauhut, R. Schneider, and Z. Stojanac, *Low rank tensor recovery via iterative hard thresholding*, in Proceedings of the 10th International Conference on Sampling Theory and Applications, Bremen, Germany, 2013, pp. 21–24.

[37] H. Rauhut, R. Schneider, and Z. Stojanac, *Tensor completion in hierarchical tensor representations*, in Compressed Sensing and Its Applications, H. Boche, R. Calderbank, G. Kutyniok, and J. Vybíral, eds., Appl. Numer. Harmon. Anal., Springer, New York, 2015, pp. 419–450.

[38] H. Rauhut, R. Schneider, and Z. Stojanac, *Low Rank Tensor Recovery via Iterative Hard Thresholding*, preprint, https://arxiv.org/abs/1602.05217, 2016.

[39] H. Rauhut and Ž. Stojanac, *Tensor Theta Norms and Low Rank Recovery*, preprint, https://arxiv.org/abs/1505.05175, 2015.

[40] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Rev., 52 (2010), pp. 471–501, https://doi.org/10.1137/070697835.

[41] M. Rudelson and R. Vershynin, *On sparse reconstruction from Fourier and Gaussian measurements*, Comm. Pure Appl. Math., 61 (2008), pp. 1025–1045.

[42] M. Signoretto, R. Van de Plas, B. De Moor, and J. A. K. Suykens, *Tensor versus matrix completion: A comparison with application to spectral data*, IEEE Signal Process. Lett., 18 (2011), pp. 403–406.

[43] J. Tanner and K. Wei, *Normalized iterative hard thresholding for matrix completion*, SIAM J. Sci. Comput., 35 (2013), pp. S104–S125, https://doi.org/10.1137/120876459.

[44] R. Tomioka, K. Hayashi, and H. Kashima, *Estimation of Low-Rank Tensors via Convex Optimization*, preprint, https://arxiv.org/abs/1010.0789, 2010.

[45] R. Tomioka, K. Hayashi, and H. Kashima, *On the extension of trace norm to tensors*, in NIPS Workshop on Tensors, Kernels, and Machine Learning, Whistler, Canada, 2010.

[46] R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima, *Statistical performance of convex tensor decomposition*, in Proceedings of the 24th International Conference on Neural Information Processing Systems, Granada, Spain, 2011, pp. 972–980.

[47] L. R. Tucker, *Some mathematical notes on three-mode factor analysis*, Psychometrika, 31 (1966), pp. 279–311.

[48] N. Vannieuwenhoven, R. Vandebril, and K. Meerbergen, *A new truncation strategy for the higher-order singular value decomposition*, SIAM J. Sci. Comput., 34 (2012), pp. A1027–A1052, https://doi.org/10.1137/110836067.

[49] Y. Xu, R. Hao, W. Yin, and Z. Su, *Parallel matrix factorization for low-rank tensor completion*, Inverse Probl. Imaging, 9 (2015), pp. 601–624.

[50] Y. Yang, Y. Feng, and J. Suykens, *A rank-one tensor updating algorithm for tensor completion*, IEEE Signal Process. Lett., 22 (2015), pp. 1633–1637.

[51] M. Yuan and C.-H. Zhang, *On Tensor Completion via Nuclear Norm Minimization*, preprint, https://arxiv.org/abs/1405.1773, 2014.

[52] M. Zhang, L. Yang, and Z.-H. Huang, *Minimum n-rank approximation via iterative hard thresholding*, Appl. Math. Comput., 256 (2015), pp. 860–875.