Learning Transition Operators From Sparse Space-Time Samples

Christian Kümmerle, Mauro Maggioni, and Sui Tang

***Abstract*—We consider the nonlinear inverse problem of learning a transition operator $\mathbf{A}$ from partial observations at different times, in particular from sparse observations of entries of its powers $\mathbf{A}, \mathbf{A}^2, \cdots, \mathbf{A}^T$. This *Spatio-Temporal Transition Operator Recovery* problem is motivated by the recent interest in learning time-varying graph signals that are driven by graph operators depending on the underlying graph topology. We address the nonlinearity of the problem by embedding it into a higher-dimensional space of suitable block-Hankel matrices, where it becomes a low-rank matrix completion problem, even if $\mathbf{A}$ is of full rank. For both a uniform and an adaptive random space-time sampling model, we quantify the recoverability of the transition operator via suitable measures of incoherence of these block-Hankel embedding matrices. For graph transition operators these measures of incoherence depend on the interplay between the dynamics and the graph topology. We develop a suitable non-convex iterative reweighted least squares (IRLS) algorithm, establish its quadratic local convergence, and show that, in optimal scenarios, no more than $\mathcal{O}(rn\log(nT))$ space-time samples are sufficient to ensure accurate recovery of a rank-r operator $\mathbf{A}$ of size $n \times n$. This establishes that spatial samples can be substituted by a comparable number of space-time samples. We provide an efficient implementation of the proposed IRLS algorithm with space complexity of order $O(rnT)$ and per-iteration time complexity linear in n. Numerical experiments for transition operators based on several graph models confirm that the theoretical findings accurately track empirical phase transitions, and illustrate the applicability and scalability of the proposed algorithm.**

***Index Terms*—Operator learning; block Hankel matrix completion; iterative reweighted least squares; nonlinear inverse problem; graph signal processing.**

I. Introduction

Signals that arise from social, biological or transport networks are typically interconnected and structured, and can be modeled as residing on graphs. In many modern applications, the graph signals are time-varying and driven by graph operators that are dependent on the underlying graph topology. For example, the traffic flow on the road network is changing during a day; spatial temperature profiles measured by a sensor network vary at different time instances. Estimating such graph signals and dynamical processes from sparse observations is a research topic of wide interest (see for example [ISG18, PGM+16, IRG18, DTFV16, TDKF17, SMMR17, MSMR19, PGM+17] and references therein).

We consider dynamical processes on graphs in the form of (discrete) linear dynamical systems:

$$x_{t+1} = \mathbf{A}x_t : t = 1, \ldots, \tag{1}$$

where x_t is the graph signal and the transition operator $\mathbf{A}$ is typically a function of an algebraic descriptor of the graph structure (e.g. the adjacency matrix of the graph). Examples for such transition operators include the random walk over a graph and its variations, heat operators, and other averaging processes. Powers of $\mathbf{A}$ lead to both multiscale analyses on graphs [CM06], and eigenvectors of $\mathbf{A}$ often play an important role in many machine learning applications such as dimension reduction and clustering [CLL+05]. These models are in part motivated by applications that include modeling traffic in transportation networks [DM16], spatially-distributed atmospheric variables (e.g. temperature, pressure) measured by sensor networks [TDKF17], and neural activity in different regions of the brain [Spo10].

In this paper, we are interested in learning $\mathbf{A}$ from partial space-time observations of a temporal evolution. Let $\mathcal{X}_0 \in \mathbb{R}^{n\times m}$ be a set of m initial states, one per column. We observe a discrete time series of length T satisfying

$$\mathcal{X}_{t+1} = \mathbf{A}\mathcal{X}_t\,, \qquad t \in [T] := 1, \ldots, T \tag{2}$$

via spatio-temporal samples

$$\mathcal{Y}_t = S_t(\mathcal{X}_t)\,, \tag{3}$$

for $t \in [T]$ and $S_t : \mathbb{R}^{n\times m} \to \mathbb{R}^{m_t}$ representing a linear subsampling operator, typically returning a subset of the entries of its input.

If we set aside the time evolution in (2), the task of recovering $\mathcal{X}_t$ at a fixed time t from the observations $\mathcal{Y}_t$ may be considered as a *completion problem*, and may be tackled by *low-rank matrix completion* techniques [CR09, CP11a, DR16, CLC19], even when the number of observations m_t is much less than the number of entries in $\mathcal{X}_t$, if $\mathcal{X}_t$ has low rank. When $S_t \equiv \mathrm{Id}$, the identity map, related problems are also pursued in the model reduction community, to extract dominant eigenvectors of $\mathbf{A}$, for example via the dynamic mode decomposition (DMD) [Sch10, KBBP16].

In this paper, however, we are interested in situations where $T > 1$ and both $\mathcal{X}_t$ and $\mathbf{A}$ are *not* necessarily low rank. At any single time t, we are not able to recover $\mathbf{A}$ from $\mathcal{Y}_t$ and $\mathcal{Y}_{t+1}$, as in practice we often have $m_t, m_{t+1} \lesssim n$, due to application-specific constraints. We seek to compensate the insufficient spatial samples at a single given time t by leveraging the temporal dependent observations across time,

C. Kümmerle is with the Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, 28223, USA, e-mail: kuemmerle@uncc.edu.

M. Maggioni is with the Department of Mathematics and the Department of Applied Mathematics & Statistics, Johns Hopkins University, Baltimore, MD, 21218, e-mail: mauromaggionijhu@icloud.com.

S. Tang is with the Department of Mathematics, University of California, Santa Barbara, Isla Vista, CA 93117, USA, e-mail: suitang@ucsb.edu.

and tackle the fundamental problem of recovering $\mathbf{A}$ from *space-time* samples $\{\mathcal{Y}_t : t \in [T]\}$. We restrict our attention, for simplicity, to recovering $\mathbf{A}$ from *partial observations* of $\mathbf{A}, \mathbf{A}^2, \cdots, \mathbf{A}^T$, which, in the notation above, corresponds to $\mathcal{X}_0 = \mathrm{Id} \in \mathbb{R}^{n \times n}$ in (2). Already in this setting, learning $\mathbf{A}$ is a nonlinear inverse problem (as long as $T > 1$), that is beyond the scope of regular matrix completion, dynamic mode decomposition, or subspace-based techniques [CIML20].

A. Spatio-Temporal Transition Operator Recovery

In the context of dynamical systems, it is expected that the recoverability of a transition operator $\mathbf{A} \in \mathbb{R}^{n \times n}$ depends significantly on the structure of the space-time sampling as well as on spectral properties of $\mathbf{A}$. Denote the sampling locations at $t \in [T]$ by $\Omega_t \subseteq [n] \times [n]$, and let the corresponding subsampling operator $S_t : \mathbb{R}^{n \times n} \to \mathbb{R}^{\Omega_t}$ be

$$S_t(\mathbf{M}) := (\langle E_{i,j}, \mathbf{M} \rangle)_{(i,j) \in \Omega_t} = (\mathbf{M}_{i,j})_{(i,j) \in \Omega_t}\,, \tag{4}$$

where $E_{i,j}$ is the matrix with 1 in its (i,j)-th entry and 0 elsewhere, for $i, j \in [n]$. In applications, Ω_t may correspond to an observation model with mobile sensors that are moved to different locations at different times.

Denoting the set of all possible space-time sampling locations by $I := [n] \times [n] \times [T]$, we define the *sampling set*

$$\Omega := (\Omega_1 \times \{1\}) \cup \cdots \cup (\Omega_T \times \{T\}) \subset I\,. \tag{5}$$

Let $\mathcal{M}_{n_1,n_2}$ denote the set of real $n_1 \times n_2$ matrices, abbreviated as $\mathcal{M}_n$ if $n = n_1 = n_2$. We define the nonlinear monomial operator $\mathcal{Q}_T : \mathcal{M}_n \to \mathcal{M}_n^{\oplus T}$ as

$$\mathcal{Q}_T(\mathbf{A}) := \mathbf{A} \oplus \mathbf{A}^2 \oplus \mathbf{A}^3 \oplus \ldots \oplus \mathbf{A}^T \in \mathcal{M}_n^{\oplus T}\,, \tag{6}$$

and the sampling operator $P_\Omega : \mathcal{M}_n^{\oplus T} \to \mathbb{R}^{|\Omega|}$ as

$$\begin{aligned} \widetilde{\mathbf{X}} :=& \mathbf{X}_1 \oplus \mathbf{X}_2 \oplus \ldots \oplus \mathbf{X}_T \mapsto \\ & P_\Omega(\widetilde{\mathbf{X}}) = \begin{bmatrix} S_1(\mathbf{X}_1), & S_2(\mathbf{X}_2), & \ldots, & S_T(\mathbf{X}_T) \end{bmatrix} \end{aligned} \tag{7}$$

where S_t is as in (4), for $t \in [T]$. We consider the following:

Problem I.1 (Spatio-Temporal Transition Operator Recovery)**.** *Given a space-time sampling set $\Omega \subset I$, recover $\mathbf{A} \in \mathcal{M}_n$ from the space-time samples*

$$\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A})) = \begin{bmatrix} S_1(\mathbf{A}^1), S_2(\mathbf{A}^2), \ldots, S_T(\mathbf{A}^T) \end{bmatrix}\,, \tag{8}$$

or from noisy space-time samples $\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A})) + \eta$, where η is an (unknown) additive noise vector.

We focus on two questions arising naturally in Problem I.1:

- Under which conditions on $\mathbf{A}$ and on the distribution and size of the sampling set Ω can we guarantee to accurately estimate $\mathbf{A}$ in Problem I.1?
- Is there a computationally efficient recovery method to estimate $\mathbf{A}$ in Problem I.1 in these cases?

It is well-known from the literature on other structured inverse problems such as sparse vector recovery and matrix completion that deterministic sampling sets may not enable recovery from a minimal size of the sampling set $|\Omega|$ [FR13,DR16]. For this reason, we focus on random space-time sampling schemes, and consider two types of random models:

1) **Uniform sampling**: for $m \leq n^2 T$, a sampling set Ω consists of m spatio-temporal samples in $[n] \times [n] \times [T]$ sampled uniformly at random without replacement;
2) **Adaptive sampling**: for each space-time index $(i,j,t) \in [n] \times [n] \times [T]$, let $p_{i,j,t} \in [0,1]$. An adaptive sampling set $\Omega \subseteq [n] \times [n] \times [T]$ consists of triplets (i,j,t) drawn from i.i.d. Bernoulli trials with success probabilities $p_{i,j,t}$. The expected total number of samples is therefore $m_{\text{exp}} := \mathbb{E}[|\Omega|] = \sum_{i,j=1}^{n} \sum_{t=1}^{T} p_{i,j,t}$.

While uniform sampling is conceptually simple as it only has one free parameter, $m = |\Omega|$, adaptive sampling is more flexible, in particular because its sampling probabilities $\{p_{i,j,t}\}_{(i,j,t) \in I}$ can be tuned to include prior information about a specific instance of Problem I.1.

B. Our Contribution

We tackle the spatio-temporal transition operator recovery problem by first applying an embedding into a structured block Hankel matrix space, under which the nonlinear relationship between different powers $\mathbf{A}, \mathbf{A}^2, \ldots, \mathbf{A}^T$ of a matrix $\mathbf{A}$ is mapped to a low-rank property of the block Hankel matrix $\mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$, essentially in a one-to-one manner (Theorem II.1). The block Hankel operator $\mathcal{H} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_{d_1 n, d_2 n}$, with parameters $d_1, d_2 \in \mathbb{N}$ s.t. $T = d_1 + d_2 - 1$, is defined as:

$$\mathcal{H}\left(\mathbf{X}_1 \oplus \ldots \oplus \mathbf{X}_T\right) := \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 & \iddots & \mathbf{X}_{d_2} \\ \mathbf{X}_2 & \mathbf{X}_3 & \iddots & \iddots & \iddots \\ \mathbf{X}_3 & \iddots & \iddots & \iddots & \iddots \\ \iddots & \iddots & \iddots & \iddots & \mathbf{X}_{T-1} \\ \mathbf{X}_{d_1} & \iddots & \iddots & \mathbf{X}_{T-1} & \mathbf{X}_T \end{bmatrix} \tag{9}$$

Related embeddings have been used to design computational methods for the solution of classical problems in signal processing and system identification, see Section I-C3 for details.

We then deploy an efficient low-rank optimization algorithm based on Iteratively Reweighted Least Squares (IRLS), which combines an iterative minimization of quadratic majorizing functions with an appropriate smoothing strategy for a log-determinant objective [DDFG10, MF12, KMV21] and, at the same time, respects the block Hankel structure. We address the connection between the choice of a space-time sampling set Ω and the identifiability of $\mathbf{A}$ by proving a local convergence result for the proposed algorithm, called Transition Operator IRLS (`TOIRLS`), which shows that the operator $\mathbf{A}$ can be efficiently computed from a number of spatio-temporal samples that is comparable to the sample complexity of using only spatial samples at time $T = 1$, for random sampling models based on either uniform or adaptive sampling. In particular, we show in Theorem IV.1 that in the noiseless case, with high probability, `TOIRLS` exhibits locally quadratic convergence if initialized close enough to the ground truth block matrix $\mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$, as soon as only $\Omega(\mu_0 r n \log(nT))$ uniform or adaptive samples are provided. An informal version of this result may be stated as follows:

Theorem I.1 (Local Convergence of `TOIRLS`, informal version)**.** *Let $\mathbf{A} \in \mathcal{M}_n$ be a transition operator of rank r, and let $\mathbf{H}_{\mathbf{A}} := \mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$, with $\mathcal{Q}_T(\mathbf{A})$ and $\mathcal{H}$ as in (6) and (9), respectively. Assume that either*

(i) Ω is a space-time sampling set drawn by uniform sampling of cardinality

$$m \gtrsim \mu_0 rn \log(nT), \tag{10}$$

where μ_0 is the incoherence factor (see Definition IV.1) of $\mathbf{H}_{\mathbf{A}}$, or

(ii) Ω is obtained by adaptive sampling with Bernoulli parameters

$$p_{i,j,t} \gtrsim \min\left(\mu_{i,j,t}\frac{r}{nT}\log(nT), 1\right),$$

where $\mu_{i,j,t}$ is a local incoherence (see Definition IV.1) of $\mathbf{H}_{\mathbf{A}}$.

If, additionally, an iterate of `TOIRLS` (Algorithm 1), with observations $\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A}))$, is close enough to $\mathcal{Q}_T(\mathbf{A})$, then, with high probability, the subsequent iterates converge to $\mathcal{Q}_T(\mathbf{A})$ with a quadratic convergence rate.

In Section VI, we provide numerical experiments that illustrate that the order of convergence in Theorem I.1 captures the empirical behavior of `TOIRLS` for several transition operators on random graphs, and that its behavior appears robust to additive noise in the observations.

This result implies that despite the fact that samples are taken across T different powers of $\mathbf{A}$, not only from $\mathbf{A}$ itself, `TOIRLS` recovers $\mathbf{A}$ from essentially as few uniformly random samples as in the classical low-rank matrix completion setting [CR09, KBV09, CT10, Che15, CLC19], where $O(\nu_0 rn \log(n))$ samples are necessary in a uniform sampling model for the unique recovery of $\mathbf{A}$ by any algorithm, where ν_0 is the standard incoherence of $\mathbf{A}$ [Che15].

We also analyze the incoherence of $\mathbf{H}_{\mathbf{A}}$ by relating it to that of $\mathbf{A}$ for several families of transition operators, see Section IV-C. In particular, we show that if $\mathbf{A}$ is an orthogonal matrix or a projection, the incoherence μ_0 of $\mathbf{H}_{\mathbf{A}}$ coincides with the incoherence ν_0 of $\mathbf{A}$, implying that the same order $\Omega(\nu_0 rn \log(n))$ of samples as in conventional matrix completion is sufficient in our setting, at least when $T \lesssim n$.

Unlike in conventional matrix completion, our results are nontrivial also when $\mathbf{A}$ is of full rank, i.e. $r = \operatorname{rank}(\mathbf{A}) = n$: in this case, Theorem I.1 implies that $O(\mu_0 n^2 \log(nT))$ samples, scattered over the T observation times, are sufficient to ensure local convergence of `TOIRLS`, i.e. we pay a multiplicative oversampling factor of $O(\mu_0 \log(nT))$ over the n^2 degrees of freedom of $\mathbf{A}$. In particular, recovery is possible with a budget of only $O(n \log(n))$ sensors and $T = n$ observation times.

Finally, our results, such as Theorem I.1.2, on local incoherence-based sampling of specific space-time locations, inform adaptive sampling schemes that can be more data-efficient than uniform sampling.

Our results are presented in Section IV-B in the context of the spatio-temporal transition operator recovery problem. However, we point out that they are valid more generally for the problem of recovering a low-rank block Hankel matrix via Algorithm 1 if its output is chosen to be the entire matrix $\widetilde{\mathbf{X}}^{(K)} \in \mathcal{M}_n^{\oplus T}$ instead of its restriction to its first block $\mathbf{A}^{(K)}$.

Remark I.1. *In this work, we focus on the algorithmic scheme `TOIRLS` optimizing non-convex surrogates in order to explore the fundamental information-theoretic properties of the underlying problem instead of a more traditional convex approach used for other low-rank optimization problems* [CR09, Che15, DC20, CC14, YKJL17] *(see also Sections I-C1 and I-C4 below) for two reasons:*

- *we are able to ensure fast, albeit local, convergence, with high probability, under minimal assumptions on the sample complexity (10). Using nuclear norm minimization on block Hankel matrices, we surmise that it is possible to also obtain an exact recovery result, albeit with possibly worse dependence on the sample complexity, with additional logarithmic factors in r, μ_0 and potentially T, when using techniques such as a dual certificates or a leave-one-out analysis* [Che15, DC20]*;*
- *using a nuclear norm approach does not by itself lead to a scalable algorithm, as nuclear norm minimization is equivalent to a semidefinite program (SDP) with matrix variables of size $O(nT) \times O(nT)$. While some recent approximate solvers for large-scale SDPs require space of order only $O(rnT + m)$* [YTF+21, DYC+21]*, these methods do not find high-accuracy solutions. On the other hand, methods which provably solve the original SDP (e.g., interior-point methods* [AHO98] *or augmented Lagrangian methods* [STYZ20]*) have storage requirements of $O(n^2T^2)$ or larger.*
  We show in Theorem V.1 that `TOIRLS` is a scalable algorithm with space complexity of $O(rnT + m)$, a per-iteration time complexity linear in n, and quickly leads to high-accuracy solutions thanks to the guaranteed local quadratic convergence rate.

C. Related Work

The transition operator recovery problem and the proposed low-rank modeling have connections to several different fields, which we briefly discuss.

1) Low-Rank Matrix Completion: pioneered by [Faz02, CR09, CT10, Gro11, Che15] and popularized by applications in recommender systems [ZWSP08, KBV09], the problem of recovering a low-rank matrix from a subset of its entries or from underdetermined linear observations has been analyzed using both convex [RFP10, CT10, Che15] and non-convex formulations [KMO10, SL16, CLC19]. The a minimal sufficient condition for global convergence in the case of uniform samples is due to [DC20], where it was shown that $\Omega(\nu_0 rn \log(n) \log(\nu_0 r))$ uniform samples are sufficient for the convex nuclear norm minimization approach to succeed with high probability if ν_0 is the incoherence factor of [Che15], n the dimensionality and r the rank of the matrix to be recovered. Local quadratic convergence in the presence of only $\Omega(\nu_0 rn \log(n))$ random observations was established for low-rank completion for a method similar to `TOIRLS` in [KMV21] and in [ZN22] for a Gauss-Newton method, improving previous works on related algorithms [MF12, FRW11, KS18] and [BNZ21], respectively. Low-rank matrix completion is a special case of the transition operator recovery problem Problem I.1 corresponding to $T = 1$; for $T > 1$, however, Problem I.1 is nonlinear in the transition operator.

A nonlinear generalization of the matrix completion problem that is different from ours was considered in [OWNB17, OPAB$^+$21], where the low-rank properties of *tensorized* data matrices are leveraged. While these problems also involve polynomial dependencies on a ground truth matrix, these dependencies are *columnwise* and do not comprise the rich algebraic structure of matrix polynomials present in Problem I.1. The adaptive sampling model of Section I-A had been considered for $T = 1$ in the works [CBSW15, EWW18].

2) Dynamical Sampling: here the aim is to recover a linear dynamical system from the union of coarse spatial samples at multiple time instances. A mathematical theoretical framework was proposed in [ADK13, ACMT17] for linear systems of the form (1), motivated by the pioneering work of [LV09] that considered the space-time sampling of bandlimited diffusion fields over the real line. Several works [Tan17b, LT19, AHP19, ACC$^+$17, UZ21] focus on the case where the transition operator $\mathbf{A}$ in (1) is known, and the goal is to obtain sampling theorems ensuring exact recovery of the initial state. For the case where $\mathbf{A}$ is unknown, it has been shown that the eigenvalues of the matrix $\mathbf{A}$ can be recovered from the space-time samples of a single trajectory, see [AK14, Tan17a, CT21]. It is typically assumed that the observation operator $\mathcal{S}_t$ is deterministic and independent of t. Our paper is the first one, to our knowledge, to provide results for estimating $\mathbf{A}$ from random space-time samples, i.e., for random subsampling operators $\mathcal{S}_t$ varying over the time t.

3) System Identification: consider a linear time-invariant dynamical system

$$\begin{aligned} x_{t+1} &= \mathbf{A}x_t + \mathbf{B}u_t \\ y_t &= \mathbf{C}x_t + \mathbf{D}u_t \end{aligned} \tag{11}$$

where u_t is the input vector and y_t is the output vector. The parameter estimation problem considered in control theory aims to recover the system matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ from the input-output pairs (u_t, y_t). Classical results show that a necessary condition to ensure identifiability is that $\mathbf{C}$ is full rank [BÅ70]. In general, this problem is ill-posed and the focus is to learn system matrices up to similarity transformations (see subspace identification methods [Lju98, Qin06]) or the impulse response function (also called the Markov parameters) that determines the input-output map, both from a single trajectory [Fat21, OO19, SR19] and from multiple trajectories in [ZL20, SOF20, TBPR17]. In the case of $\mathbf{B} = \mathbf{D} = 0$ and $\mathbf{C} = \mathbf{I}$, a sufficient and necessary condition for the identifiability of $\mathbf{A}$ from a single trajectory with a fixed initial condition is that $\mathbf{A}$ has only one Jordan block for each of its eigenvalues, together with certain constraints on the initial condition [SRS14, DRS20]. The low-rankness of a block-Hankel embedding of suitable powers and products of the matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and $\mathbf{D}$ similar to (9) is known to underlie the Kalman-Ho [HK66, OO19] method for finding a *realization* of the system, and has been explicitly used as optimization objective in [FPST13, MU13, Mar19, GRG18]. However, the observation matrix $\mathbf{C}$ is fixed in all the works the authors are aware of within this line of research, whereas in our setting, $\mathbf{C}$ is random and varies over time t.

4) Recovery of Structured Signals: many structured signal recovery problems can be represented in the following abstract form: for a normed vector space V over $\mathbb{C}$ and a *known* linear operator $\mathbf{A} : V \to V$, one is interested in recovering a signal $f \in V$ that is M-sparse in terms of eigenfunctions $\{v_j\}$ of $\mathbf{A}$, i.e., $f = \sum_{j\in J} c_j v_j$, where $\{c_j\}_j$ is a set of coefficients in $\mathbb{C}$ and $|J| = M$. The goal is to recover $\{c_j\}_{j\in J}$ and $\{v_j\}_{j\in J}$ from observations $\mathcal{F}(\mathbf{A}^\ell f)$ for $\ell = 0, 1, \dots, L$ where $\mathcal{F} : V \to \mathbb{C}$ is a linear functional. For example, let V be a vector space consisting of continuous functions on the real line and $\mathbf{A}$ be a shift operator $(\mathbf{A}f)(x) = f(x+1)$. If one takes $v_j = \mathrm{e}^{a_j x}$ for $j \in [M]$ where $\{a_j\}_{j=1}^M$ are distinct complex numbers and $\mathrm{Im}(a_j) \in [-\pi, \pi)$, $\mathcal{F}(f) = f(x_0)$ for some $x_0 \in \mathbb{R}$ and $L = 2M - 1$, then the recovery problem corresponds to the classical estimation of a sparse exponential sum, called harmonic retrieval. The other instances of this problem include super-resolution, blind deconvolution, recovery of signals with finite rate innovation; we refer to [PP13, HS17] for more details. Many well-known algorithmic approaches for these estimation problems are related to the Hankel matrix formed from the samples $\{\mathcal{F}(\mathbf{A}^\ell f)\}_{\ell=1}^L$, including Prony's method [PT10], matrix pencil methods [HS90], and the algorithms MUSIC [LF16] and ESPRIT [RK89]. A generalization to irregular sets of samples has been considered in [CC14, JLY16, CWW19, KV19], by formulating the signal recovery problem as a low-rank Hankel matrix completion, relating the problem to techniques discussed in Section I-C1 above. Multidimensional versions of these setting have also been considered in these works, leveraging low-rank properties of suitable multilevel Hankel or Toeplitz matrices, as well as in [YXS16]. Specific to the $2D$ case, block Hankel matrices with Hankel blocks arise in these applications. In contrast, here the blocks of the block Hankel matrices we leverage are in general not Hankel, but a general matrix related to a linear dynamical system. Finally, we remark that while the above works focus on recovering *scalar signals*, with the shift operator $\mathbf{A}$ being known, here we address the problem of estimating the unknown $\mathbf{A}$, from partial observations along different trajectories.

5) Graph Learning in Signal Processing: there have been significant research efforts for inferring graph topology from observations of graph signals. This includes [PIM10], where the graph topology is estimated from full observations of a solution of a system of linear SDEs on the graph, via a regularized least squares approach, in particular focusing on the length of time the system needs to be observed in order to estimate the graph topology, as well as on the role of sparsity of the graph topology. Other existing approaches leverage a model based on graph filters [SMMR16, SMMR17], or enforce sparsity [MTF17] or smoothness [Kal16, DTFV16, TDKF17, EPO18] of signals using a penalized likelihood approach. Only a few works consider the graph signals as states of an underlying dynamical system, evolving according to the topology of the graph, e.g., [CIML20], in which case the single-trajectory states are observed via a fixed observation matrix that is static over time. In all cases, the identifiability of graph topology remains to be a challenging problem, as do the recovery algorithms, which lack theoretical guarantees.

D. *Outline*

The paper is organized as follows. In Section II, we present in which sense the transition operator recovery problem is equivalent to a rank minimization problem over block Hankel matrices constrained to an affine space. In Section III, we introduce `TOIRLS`, or Transition Operator Iteratively Reweighted Least Squares, to solve the resulting structured rank minimization problem. We introduce incoherence notions of block Hankel matrices in Section IV, and present Theorem IV.1, our main result that establishes local quadratic convergence of `TOIRLS` for respective sample complexities under both uniform and adaptive space-time sampling models. In Section V, we elaborate on computational considerations for `TOIRLS`, before presenting extensive numerical explorations in Section VI. In Section VII, we provide the proof of the main theorem of Section II and in Section VIII a proof outline of Theorem IV.1. We conclude the main part of the paper in Section IX. Finally, we present useful incoherence estimates in Appendix A, present a complete proof of Theorem IV.1 in Appendix B and detail a practical implementation of `TOIRLS` in Appendix H.

E. *Notation*

We briefly summarize some notational conventions we use in this paper. The set of *orthogonal matrices* of dimension d is denoted by $\mathbb{O}^d = \{\mathbf{M} \in \mathcal{M}_n : \mathbf{M}^*\mathbf{M} = \mathrm{Id}\}$, while Id is the identity matrix (omitting its dimension whenever suitable). If $\mathcal{M}_{d_1 \times d_2}$ and $\mathbf{v} \in \mathbb{R}^d$ is an arbitrary vector of dimension $d := \min(d_1, d_2)$, the operator $\mathrm{dg} : \mathbb{R}^d \to \mathcal{M}_{d_1 \times d_2}$ maps $\mathbf{v}$ to the (generalized) diagonal matrix $\mathrm{dg}(\mathbf{v}) \in \mathcal{M}_{d_1 \times d_2}$ with $\mathrm{dg}(\mathbf{v})_{ij} = \mathbf{v}_i$ if $i = j$ and $\mathrm{dg}(\mathbf{v})_{ij} = 0$ otherwise. For any matrix $\mathbf{H}$, we denote its spectral norm by $\|\mathbf{H}\| := \sigma_1(\mathbf{H})$ and define the spectral norm ball with radius $\xi > 0$ around $\mathbf{H}$ as $\mathcal{B}_{\mathbf{H}}(\xi) := \{\mathbf{M} \in \mathcal{M}_{d_1 n, d_2 n} : \|\mathbf{M} - \mathbf{H}\| \leq \xi\}$.

II. Recovering Transition Operators from Space-Time Samples using Low-Rank Optimization

In this section, we detail an approach to solve the transition operator recovery problem introduced in Section I-A. A fundamental issue is the *nonlinearity* of the operator $\mathcal{Q}_T$: $\mathbf{A} \mapsto \mathcal{Q}_T(\mathbf{A}) := \mathbf{A} \oplus \mathbf{A}^2 \oplus \mathbf{A}^3 \oplus \ldots \oplus \mathbf{A}^T$. We *linearize* this nonlinearity by the transformation into the *structured subspace* $\mathrm{Im}\,\mathcal{H} \subset \mathcal{M}_{d_1 n, d_2 n}$, where $\mathcal{H} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_{d_1 n, d_2 n}$ is the block Hankel operator of (9) with parameters $d_1, d_2 \in \mathbb{N}$. $\mathcal{H}$ maps a direct sums of $(n \times n)$ matrices to a *block Hankel matrix* with d_1 block rows and d_2 block columns. In the remainder of the paper, we call d_1 and d_2 satisfying $T = d_1 + d_2 - 1$ the (first and second) *pencil parameter* of $\mathcal{H}$, in accordance with [HS90, CC14].

The block Hankel operator $\mathcal{H}$ enables us to recover the operator $\mathbf{A}$ and its powers $\mathbf{A}^2, \mathbf{A}^3, \ldots \mathbf{A}^T$ from a block Hankel matrix that is *low rank* (Theorem II.1), an observation which lies at the core of our approach. We use a dedicated low-rank optimization to recover a block Hankel-structured low-rank matrix $\mathcal{H}(\widetilde{\mathbf{X}}^*)$ compatible with the samples $P_\Omega(\mathcal{Q}_T(\mathbf{A}))$ taken at space-time locations Ω. If a sufficient number of random samples Ω from a sampling model in Section I-A are provided, the hope is there is a unique generator $\widetilde{\mathbf{X}}^* = \mathcal{Q}_T(\mathbf{A})$ for the Hankel matrix, from which the transition operator $\mathbf{A}$ can then be directly inferred.

A. *Rank Minimization over Block Hankel Matrices*

As a justification for our search for low-rank matrices in the subspace of block Hankel structured matrices, we establish in Theorem II.1 the strong relationship between the rank of a block Hankel matrix $\mathbf{H} \in \mathcal{M}_{d_1 n, d_2 n}$ and the rank of an underlying "generator" matrix $\mathbf{A}$. We say that $\overset{\square}{\mathbf{H}} \in \mathbb{R}^{Dn}$ is a *square extension* of $\mathbf{H}$ if it is a block Hankel matrix with pencil parameters D and D whose first T block anti-diagonals coincide with the T anti-diagonals of $\mathbf{H}$, and can otherwise have arbitrary entries in the last $2D - d_1 - d_2$ blocks.

Theorem II.1. *Recall the monomial operator* $\mathcal{Q}_T : \mathcal{M}_n \to \mathcal{M}_n^{\oplus T}, \mathbf{A} \mapsto \mathbf{A} \oplus \mathbf{A}^2 \oplus \mathbf{A}^3 \oplus \ldots \oplus \mathbf{A}^T$ *from (6), and the block Hankel operator* $\mathcal{H} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_{d_1 n, d_2 n}$ *from (9), with pencil parameters* d_1, d_2. *Then:*

1) *for any* $\mathbf{A} \in \mathcal{M}_n$,
$$\mathrm{rank}\left(\mathcal{H}(\mathcal{Q}_T(\mathbf{A}))\right) = \mathrm{rank}(\mathbf{A})\,;$$
2) *for any block Hankel matrix* $\mathbf{H} \in \mathcal{M}_{d_1 n, d_2 n}$ *with* $(n \times n)$ *blocks, that has a positive semidefinite square extension* $\overset{\square}{\mathbf{H}} \in \mathcal{M}_{Dn, Dn}$, $D = \max(d_1, d_2)$, *which has its first block* $\mathbf{H}_1 \in \mathcal{M}_n$ *of rank* r, *and at least one other block* $\mathbf{H}_j \in \mathcal{M}_n$, $j > 1$, *of rank* r, *there exists a pair of matrices* $(\mathbf{Y}, \mathbf{M})$, *where* $\mathbf{Y} \in \mathcal{M}_{n,r}$ *and* $\mathbf{M} \in \mathcal{M}_r$ *is symmetric, with* $\mathrm{rank}(\mathbf{H}) = \mathrm{rank}(\mathbf{Y}) = \mathrm{rank}(\mathbf{M}) = r$ *such that*
$$\mathbf{H} = \mathcal{H}\big(\mathbf{Y}\mathbf{Y}^* \oplus \mathbf{Y}\mathbf{M}\mathbf{Y}^* \oplus \ldots \oplus \mathbf{Y}\mathbf{M}^{T-1}\mathbf{Y}^*\big). \quad (12)$$

We refer to Section VII for a proof of Theorem II.1. Related results have appeared in [FH96, YXS16, AC17]; however, Theorem II.1 does not follow from these results.

Theorem II.1 implies a close relationship between a *low-rank property* of block Hankel matrices $\mathcal{H}(\widetilde{\mathbf{X}})$ as in (9) and the existence of an operator $\mathbf{A}$ such that $\mathcal{Q}_T(\mathbf{A}) = \widetilde{\mathbf{X}}$. While Theorem II.1.1 implies that the rank of generator matrix $\mathbf{A}$ is inherited by its block Hankel image, Theorem II.1.2 is a statement in the other direction, i.e. about the existence of an underlying rank-r generator matrix $\mathbf{M}$ of a matrix semigroup associated to a rank-r block Hankel matrix. We show the latter statement only if a positive semidefinite extension exists, noting that similar statements apply if additional, typically weak algebraic constraints are imposed on a general low-rank block Hankel matrix $\mathbf{H} \in \mathcal{H}_{d_1 n, d_2 n}$, see [Tis92, FH96]. In particular, note that is $\mathbf{A} \in \mathcal{M}_n$ if of full rank n, Theorem II.1 implies that $\mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$ is also of rank n. This is, however, *low-rank* if we choose $\min(d_1, d_2) > 1$, as the maximal rank of a $(d_1 n \times d_2 n)$ matrix is $\min(d_1, d_2)n$, and *not* n.

At a high level, Theorem II.1 illustrates that the nonlinear relationship between the matrix powers $\mathbf{A}, \mathbf{A}^2, \ldots, \mathbf{A}^T$ is *translated* to *linear dependences* of the blocks in an associated block Hankel matrix in $\operatorname{Im}\mathcal{H}$, motivating the pursuit of an optimization approach that aims to find a completion of a block Hankel matrix $\mathcal{H}(\widetilde{X})$ that is *both low-rank* and *compatible* with the spatio-temporal measurements parametrized by the sampling operator P_Ω from (7). This suggest a block Hankel structured *rank minimization problem*

$$\min_{\widetilde{\mathbf{X}}\in\mathcal{M}_n^{\oplus T}} \operatorname{rank}\left(\mathcal{H}(\widetilde{\mathbf{X}})\right) \text{ s.t. } P_\Omega(\widetilde{\mathbf{X}}) = \mathbf{y} \tag{13}$$

where $\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A}))$ the subset of observed entries of $\mathcal{Q}_T(\mathbf{A}) = \mathbf{A} \oplus \mathbf{A}^2 \oplus \mathbf{A}^3 \oplus \ldots \oplus \mathbf{A}^T$, indexed by Ω. More generally, for a given regularization parameter $\lambda \geq 0$, we define the *data fitting function* $G^\lambda_{\Omega,\mathbf{y}} : \mathcal{M}_n^{\oplus T} \to \mathbb{R}$

$$G^\lambda_{\Omega,\mathbf{y}}(\widetilde{\mathbf{X}}) = \begin{cases} \iota_{P_\Omega^{-1}(\mathbf{y})}(\widetilde{\mathbf{X}}), & \text{if } \lambda = 0, \\ \frac{1}{\lambda}\left\|P_\Omega(\widetilde{\mathbf{X}}) - \mathbf{y}\right\|_2^2, & \text{if } \lambda > 0, \end{cases} \tag{14}$$

where $\iota_{P_\Omega^{-1}(\mathbf{y})} : \mathcal{M}_n^{\oplus T} \to \mathbb{R} \cup \{\infty\}$ is 0 if $P_\Omega(\widetilde{\mathbf{X}}) = \mathbf{y}$, and ∞ otherwise. We then formulate the rank minimization problem

$$\min_{\widetilde{\mathbf{X}}\in\mathcal{M}_n^{\oplus T}} \operatorname{rank}\left(\mathcal{H}(\widetilde{\mathbf{X}})\right) + G^\lambda_{\Omega,\mathbf{y}}(\widetilde{\mathbf{X}}), \tag{15}$$

which reduces to (13) for $\lambda = 0$. In the presence of inexact measurements with additive noise such that $\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A})) + \eta$ for some $\eta \in \mathbb{R}^m$, it can be beneficial to choose a positive regularization parameter $\lambda > 0$ [BSW11, Klo11].

Rank minimization problems such as (13) and (15) are well-known to be NP-hard in general [RFP10], and different convex and non-convex reformulations of such problems have been studied for *unstructured* problems, i.e., for the case that $\widetilde{\mathbf{X}}$ itself is low-rank [CT10, KMO10, Rec11, Van13, PKCS18, MWCC20]; see [DR16, CLC19] for recent surveys.

While (13) enables us to formulate or problem in the language of optimization and to relate it to a common algorithmic paradigm in machine learning and signal processing, it poses several challenges from an optimization perspective. First, the rank objective is a non-convex and and non-smooth function, so that it is non-trivial to use derivative-based algorithms. Furthermore, unlike most low-rank optimization problems, the search space of (13) is the strict subspace $\left\{\mathcal{H}(\widetilde{\mathbf{X}}) : \widetilde{\mathbf{X}} \in \mathcal{M}_n^{\oplus T}\right\}$ of $\mathcal{M}_{d_1 n, d_2 n}$, making the problem a *structured low-rank optimization problem* [FPST13, Mar19, CC14, CWW18]. Lastly, due to the large dimensionality of the ambient space $\mathcal{M}_{d_1 n, d_2 n}$ even for moderate n and T, only computationally efficient methods can be used for transfer operators of non-trivial size.

III. Our Approach: Iteratively Reweighted Least Squares

In several works in the literature, rank minimization problems have been tackled by designing optimization algorithms that optimize non-convex, *smoothed* objective functions whose minimizers are designed to coincide with those of the rank objective in many cases. It was observed in [FHB03, CESV13, KMV21] that optimizing a log-determinant objective often leads to solutions of underdetermined linear systems of very low-rank even in the presence of relatively few samples. Similarly, objective functions based on the Schatten-p quasi-norm [MF12, LTYL15, OJ17, KS18, GVRH20] and the smoothed clipped absolute deviation (SCAD) of the singular values [MSW20] have been used to derive competitive algorithms for a variety of low-rank matrix recovery problems in signal processing and statistics.

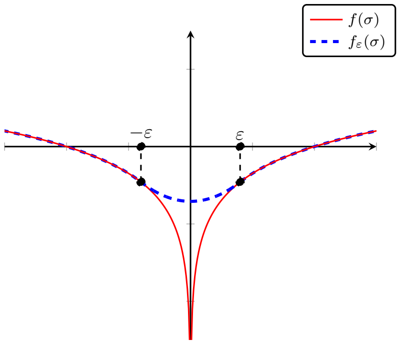


Fig. 1: Illustration of the smoothing $f_\varepsilon(\sigma)$ of $f(\sigma) = \log|\sigma|$.

We propose an algorithm that adapts these ideas to the block Hankel rank minimization problem (15) as presented in Section II-A, which can be interpreted as an *Iteratively Reweighted Least Squares (IRLS)* [HW77, DDFG10, MF12, FRW11, OJ17, KS18] strategy. Instead of optimizing the rank objective (15) directly, let $\varepsilon > 0$ be a smoothing parameter and define the smoothed log-detetermінant objective $F_\varepsilon : \mathcal{M}_{d_1 n, d_2 n} \to \mathbb{R}$ as

$$F_\varepsilon(\mathbf{M}) := \sum_{i=1}^{dn} f_\varepsilon(\sigma_i(\mathbf{M})), \tag{16}$$

where $d = \min(d_1, d_2)$ and

$$f_\varepsilon(\sigma) = \begin{cases} \log|\sigma|, & \text{if } \sigma \geq \varepsilon, \\ \log(\varepsilon) + \frac{1}{2}\left(\frac{\sigma^2}{\varepsilon^2} - 1\right), & \text{if } \sigma < \varepsilon, \end{cases} \tag{17}$$

which is continuously differentiable. If $J_\varepsilon : \mathcal{M}_n^{\oplus T} \to \mathbb{R} \cup \{\infty\}$ is the *ε-smoothed surrogate objective* defined as

$$J_\varepsilon(\widetilde{\mathbf{X}}) = F_\varepsilon(\mathcal{H}(\widetilde{\mathbf{X}})) + G^\lambda_{\Omega,\mathbf{y}}(\widetilde{\mathbf{X}}), \tag{18}$$

for a matrix $\widetilde{\mathbf{X}} \in \mathcal{M}_n^{\oplus T}$, the steps of an iteration of IRLS can be understood as, first, the minimization of a quadratic model function $Q_\varepsilon(\cdot|\mathbf{M}) : \mathcal{M}_{d_1 n, d_2 n} \to \mathcal{M}_{d_1 n, d_2 n}$ that is an appropriate, global upper bound of $J_\varepsilon(\cdot)$, leading to a weighted least squares problems and, second, as an update of the smoothing parameter ε and refinement of the quadratic model function Q_ε using the solution of the last weighted least squares problem. The quadratic model functions $Q_\varepsilon(\cdot|\mathbf{M})$ can be defined implicitly using weight operators, with which we are then able to formulate `TOIRLS`, an IRLS algorithm for transition operator learning (Algorithm 1).

Definition III.1 (see also [Küm19, KKMV22])**.** *Let $\mathbf{M} \in \mathcal{M}_{d_1 n, d_2 n}$ be a matrix with singular value decomposition*

$\mathbf{M} = \mathbf{U}\,\mathrm{dg}(\sigma)\mathbf{V}^*$, *where* $\mathbf{U} \in \mathbb{O}^{d_1 n}$, $\mathbf{V} \in \mathbb{O}^{d_2 n}$, *and* $\varepsilon > 0$.

1) *The* optimal weight operator $W_{\mathbf{M}} : \mathcal{M}_{d_1 n, d_2 n} \to \mathcal{M}_{d_1 n, d_2 n}$ *associated to* $\mathbf{M}$ *and* ε *is the linear operator*

$$W_{\mathbf{M}}(\mathbf{Z}) = \mathbf{U}\Sigma_{\varepsilon,d_1}^{-1}\mathbf{U}^*\mathbf{Z}\mathbf{V}\Sigma_{\varepsilon,d_2}^{-1}\mathbf{V}^*, \tag{19}$$

where $\Sigma_{\varepsilon,d_1} \in \mathcal{M}_{d_1 n}$ *and* $\Sigma_{\varepsilon,d_2} \in \mathcal{M}_{d_2 n}$ *are diagonal with* $(\Sigma_{\varepsilon,d_1})_{ii} = \max(\sigma_i, \varepsilon)$ *for* $i \in [d_1 n]$ *and* $(\Sigma_{\varepsilon,d_2})_{jj} = \max(\sigma_j, \varepsilon)$ *for* $j \in [d_2 n]$.[1]

2) *Let* $\mathcal{H} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_{d_1 n, d_2 n}$ *be the block Hankel operator of (9). We define the* effective weight operator $\widetilde{W}_{\mathbf{M}} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ *as the linear operator*

$$\widetilde{W}_{\mathbf{M}}(\widetilde{\mathbf{Z}}) := \mathcal{H}^* W_{\mathbf{M}} \mathcal{H}(\widetilde{\mathbf{Z}}).$$

The choice of $W_{\mathbf{M}}$ in Definition III.1 in the weighted least squares problem (20) of Algorithm 1 can be regarded as the *geometric operator mean* of the one-sided weight operator notions of the first IRLS papers considering rank optimization [MF12, FRW11]. While a detailed discussion is beyond the scope of this paper, we note that in [Küm19, KKMV22] it is shown that the associated quadratic model function $Q_\varepsilon(\cdot|\mathbf{M})$ not only majorizes the ε-smoothed surrogate objective $J_\varepsilon(\cdot)$ of (18) pointwise, but also is *optimal* in the sense that any smaller weight operator does not lead to majorizing quadratic model functions. Using the pointwise majorization, it is possible to show that the iterates $(\widetilde{\mathbf{X}}^{(k)})_{k\geq 1}$ of Algorithm 1 lead to a monotonically decreasing sequence $\big(J_{\varepsilon_k}(\widetilde{\mathbf{X}}^{(k)})\big)_{k\geq 1}$, and that each accumulation point of $(\widetilde{\mathbf{X}}^{(k)})_{k\geq 1}$ is a stationary point of $J_{\overline{\varepsilon}}(\cdot)$, where $\overline{\varepsilon} := \lim_{k\to\infty} \varepsilon_k$ [KMV21].

While the domain of the weighted least squares step (20) of Algorithm 1 is $\mathcal{M}_n^{\oplus T}$, by the definition of the effective weight operator $\widetilde{W}_{\mathbf{M}}$, a spectral reweighting *in the subspace of block Hankel matrices* is applied *implicitly*. As initialization for $k = 1$, the weight operator $W_{\mathbf{H}_0}$ (19) is chosen to be the identity operator, implying that the *effective* weight operator $\widetilde{W}_{\mathbf{H}_0} = \mathcal{H}^*\mathcal{H} = \mathcal{D}^2$ is a diagonal operator that is constant for each summand of $\mathcal{M}_n^{\oplus T}$, and which amounts to the multiplicity of each block in the block Hankel image (9) defined via the operator $\mathcal{H}$; cf. (44) in Appendix A.

a) Choice of regularization parameter λ*:* the parameter $\lambda \geq 0$ in Algorithm 1 determines which surrogate objective $J_\varepsilon(\cdot)$ is optimized by TOIRLS and which underlying rank objective (15) is chosen. As described in Section II-A, the choice of $\lambda = 0$, which imposes an affine constraint defined by the sampling operator P_Ω and the observation vector $\mathbf{y} \in \mathbb{R}^m$, is appropriate if exact space-time samples are provided to the algorithm. While an optimal choice might correspond to some $\lambda > 0$ in the presence of *inexact* space-time samples that depends on the order of magnitude of the noise, it turns out that $\lambda = 0$ is surprisingly robust to noise in practice, as explored in Section VI-C. Theoretically, this observation is related to the so-called *quotient property* of the measurement operator, which has been used to establish robust guarantees for equality-constrained low-rank and sparse recovery methods [Woj10, CP11b, Liu11, KKM22].

Algorithm 1 TOIRLS Transition Operator Iteratively Reweighted Least Squares

```
**Input:** Indices $\Omega \subset I$, observations $\mathbf{y} \in \mathbb{R}^m$, rank estimate $\widetilde{r} \leq n$, regularization parameter $\lambda \geq 0$, first pencil parameter $1 \leq d_1 \leq n$.
Set $\varepsilon^{(0)} = \infty$ and $\widetilde{W}_{\mathbf{H}_0} = \mathcal{H}^*\mathcal{H}$ with $\mathcal{H} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_{d_1 n, d_2 n}$ as in (9) where $d_2 = T - d_1 + 1$.
**for** $k = 1$ to $K$ **do**
    **Solve weighted least squares problem**
        $\widetilde{\mathbf{X}}^{(k)} := \arg\min_{\widetilde{\mathbf{X}} \in \mathcal{M}_n^{\oplus T}} \left\{ \langle \widetilde{\mathbf{X}}, \widetilde{W}_{\mathbf{H}_{k-1}}(\widetilde{\mathbf{X}}) \rangle + G_\Omega^\lambda(\widetilde{\mathbf{X}}) \right\}$ ,   (20)
    where $G_\Omega^\lambda : \mathcal{M}_n^{\oplus T} \to \mathbb{R}$ is the data fitting function of (14) and $\widetilde{W}_{\widetilde{\mathbf{X}}^{(k-1)}}$ is the effective weight operator of Definition III.1.
    **Update smoothing:** Compute $(\widetilde{r}+1)$-st singular value of $\mathbf{H}_k = \mathcal{H}(\widetilde{\mathbf{X}}^{(k)})$ to update
        $\varepsilon_k := \min\left(\varepsilon_{k-1}, \sigma_{\widetilde{r}+1}\left(\mathbf{H}_k\right)\right)$.   (21)
    **Update weight operator:** For $r_k := \left|\left\{i \in [dn] : \sigma_i\left(\mathbf{H}_k\right) > \varepsilon_k\right\}\right|$, compute reduced rank-$r_k$ singular value decomposition of of $\mathbf{H}_k$ to obtain leading $r_k$ singular values $\sigma_i\left(\mathbf{H}_k\right)$, $i = 1, \ldots, r_k$ and matrices $\mathbf{U}^{(k)} \in \mathbb{R}^{nd_1 \times r_k}$ and $\mathbf{V}^{(k)} \in \mathbb{R}^{nd_2 \times r_k}$, use this to update $\widetilde{W}_{\mathbf{H}_k}$ as defined in Definition III.1.
    $k = k + 1$.
**end for**
Extract the first block $\mathbf{A}^{(K)} := \left[\widetilde{\mathbf{X}}^{(K)}\right]_{1:n,1:n}$ of $\widetilde{\mathbf{X}}^{(K)}$.
**Output:** $\mathbf{A}^{(K)}$.
```

b) Choice of rank estimate $\widetilde{r}$*:* if the rank $r = \mathrm{rank}(\mathbf{A})$ of the transition operator $\mathbf{A}$ to be recovered is known, one should choose $\widetilde{r} = r$. If r is unknown, or if only a vague estimate is available, it is advisable to *overestimate* the true rank, i.e. to choose $\widetilde{r} \geq r$. While exact recovery of the transition operator might need more samples in that case, Algorithm 1 seems to be often able to good estimates for $\mathbf{A}$ in that case.

c) Update rule for smoothing parameter ε_k*:* after each weighted least squares step, the smoothing parameter ε_k is updated, cf. (21), in a non-increasing manner. This distinguishes IRLS from a conventional majorize-minimize (MM) method [Lan16] for the smoothed surrogate objective $J_\varepsilon(\cdot)$ for a fixed ε. Similarly to related IRLS methods [DDFG10, FRW11, VD17, KS18, KMV21], the choice of the update rule quantifies the distance to a matrix of target rank $\widetilde{r}$ that is compatible with the observations $\mathbf{y}$, playing a crucial role in the design of the algorithm due to the non-convexity of $F_{\varepsilon_k}(\cdot)$. If ε_k is large, $J_{\varepsilon_k}(\cdot)$ will possess much fewer non-global minima than if ε_k is small, in which case, however, $F_{\varepsilon_k}(\cdot)$ resembles much more the concave log-determinant objective that is known to constitute a powerful surrogate for the rank function [Fou18].

For a complexity analysis and implementation details, we refer to Section V.

[1] with the convention that $\sigma_i = 0$ for $\min(d_1, d_2)n < i \leq \max(d_1, d_2)n$.

IV. Main Results

In this section, we present a convergence theory for Algorithm 1 for the problem of recovering transition operators from sparse time-space samples.

It has been an open problem to establish global convergence of similar IRLS methods to minimizers of non-smooth, non-convex surrogate objectives such as (18) underlying the respective problems [DDFG10, MF12, KS18, KMV21], despite it being observed numerically in simulations. For this reason, we restrict the convergence analysis for `TOIRLS` to a *local* one, which is based on the assumption we are given an iterate $\widetilde{\mathbf{X}}^{(k)} \in \mathcal{M}_n^{\oplus T}$ that is close to a ground truth which is an image $\mathcal{Q}_T(\mathbf{A})$ of a transition operator $\mathbf{A}$. We quantify this using the set

$$\mathcal{B}_{\mathbf{H_A}}(\xi) := \{\mathbf{H} \in \mathcal{M}_{d_1 n, d_2 n} : \|\mathbf{H} - \mathbf{H_A}\| \le \xi\} \tag{22}$$

that contains matrices close to the block Hankel matrix $\mathbf{H_A} := \mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$.

With Theorem IV.1 in Section IV-B below, we show sufficient conditions on the number of space-time samples, under either the uniform and adaptive sampling model, that, with high probability, guarantee the local convergence of `TOIRLS` to the ground truth, and therefore the recovery of $\mathbf{A}$.

A. Incoherence for Block Hankel Matrices

Due to the coordinate-wise nature of either of our sampling models, even for a fixed dimensionality n and fixed rank r, it cannot be expected that each transition operator $\mathbf{A}$ will require a similar number of samples for successful recovery. In particular, a more *localized* transition operator with a non-zero pattern that is not very distributed will not benefit from space samples at locations associated to its zero coordinates.

In order to quantify which transition operators can be recovered by either of our sampling models, we therefore introduce a notion of *incoherence* for the block Hankel embedding matrix $\mathbf{H_A}$ of a transition operator $\mathbf{A}$. This extends the fundamental ideas in low-rank matrix completion [CR09, Rec11, Che15], where the difficulty of a completion problem is measured by the *incoherence* of a low-rank matrix with respect to the standard basis. We also introduce local incoherence quantities, to be used to guide the adaptive sampling scheme.

Let $\mathbf{T_Z}$ be the tangent space to the manifold of rank-r matrices $\mathcal{M}_r = \{\mathbf{X} \in \mathcal{M}_{d_1 n, d_2 n} : \operatorname{rank}(\mathbf{X}) = r\}$ at $\mathbf{Z} \in \mathcal{M}_{d_1 n, d_2 n}$, where $r \in \mathbb{N}$ and $\mathbf{Z} \in \mathcal{M}_{d_1 n, d_2 n}$ is a rank-r matrix with compact singular value decomposition $\mathbf{Z} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*$ with $\mathbf{U} \in \mathbb{R}^{nd_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{nd_2 \times r}$ with orthonormal columns, and $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$ the diagonal matrix of non-increasing singular values of $\mathbf{Z}$. By [Van13],

$$\mathbf{T_Z} := \{\mathbf{U}\mathbf{M}_1^* + \mathbf{M}_2\mathbf{V}^* : \mathbf{M}_1 \in \mathbb{R}^{nd_2 \times r}, \mathbf{M}_2 \in \mathbb{R}^{nd_1 \times r}\}. \tag{23}$$

Definition IV.1. *Let $\mathbf{Z} \in \mathcal{M}_{d_1 n, d_2 n}$ be a rank-r matrix. Let $\{\mathbf{B}_{i,j,t} : (i,j,t) \in I\}$ be the standard basis of the space of block Hankel matrices.*[2] *Let d_1, d_2 be the pencil parameters of the block Hankel operator $\mathcal{H}$, and $c_s := \frac{T(T+1)}{d_1 d_2}$.*

1) *For $1 \le i,j \le n$ and $1 \le t \le T$, we define the* local incoherence at space-time index (i,j,t) *of $\mathbf{Z}$ as*

$$\mu_{i,j,t} := \frac{nT}{c_s r} \|\mathcal{P}_{\mathbf{T_Z}}(\mathbf{B}_{i,j,t})\|_F^2 \,. \tag{24}$$

2) *We say that $\mathbf{Z}$ is* μ_0-incoherent *if there exists a constant $\mu_0 \ge 1$ such that*

$$\max_{1 \le i,j \le n, 1 \le t \le T} \|\mathcal{P}_{\mathbf{T_Z}}(\mathbf{B}_{i,j,t})\|_F \le \sqrt{\mu_0 c_s \frac{r}{nT}}\,, \tag{25}$$

i.e., if $\max_{1 \le i,j \le n, 1 \le t \le T} \mu_{i,j,t} \le \mu_0$. We call the smallest μ_0 satisfying (25) the incoherence parameter *of $\mathbf{Z}$.*

Intuitively, a rank-r matrix $\mathbf{Z}$ is μ_0-incoherent with small μ_0 if the projections of all elements of the standard basis of the space of block Hankel matrices $\{\mathbf{B}_{i,j,t}\}$ onto the tangent space $T_{\mathbf{Z}}$ associated to $\mathbf{Z}$ are small. In order to use an incoherence notion that is adequate for our purposes of understanding the fundamental difficulty of an instance of Problem I.1, we follow the notion of [Küm19, Definition 3.3.1] in (25), which is a slightly weaker notion than the notions used in the context of structured low-rank matrices [CC14, (27)] and [CWW19]. In fact, μ_0 in (25) can be upper bounded by the incoherence parameter of [Rec11, CC14] (see also [KMV21, Remark B.1.], the discussion of Section IV-C and Lemma A.2).

B. Local Quadratic Convergence of `TOIRLS`

We are now ready to state local convergence guarantees of `TOIRLS` (Algorithm 1) for the recovery of transition operators from space-time samples.

Theorem IV.1 (Local Quadratic Convergence of `TOIRLS`). *There exist absolute constants $\widetilde{c}_0, C$ such that the following holds. Let $\mathbf{A} \in \mathcal{M}_n$ be a rank-r transition operator, let $\mathbf{H_A} = \mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$ be the block Hankel matrix associated to the first T time scales of $\mathbf{A}$, where $\mathcal{H} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_{d_1 n, d_2 n}$ is the block Hankel embedding map (9) with pencil parameters d_1, d_2. Let $\widetilde{\mathbf{X}}^{(k)}$ be the k-th iterate of Algorithm 1 with inputs Ω, $\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A}))$ and $\widetilde{r} = r$, assume that the smoothing parameter (21) satisfies $\varepsilon_k = \sigma_r(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}))$. Let $\kappa := \sigma_1(\mathbf{H_A})/\sigma_r(\mathbf{H_A})$ denote the condition number of $\mathbf{H_A}$.*

Suppose that one of the following statements holds:

1) [Uniform sampling] $\mathbf{H_A}$ *is μ_0-incoherent and that Ω is a random subset of cardinality m uniformly drawn without replacement in the set of space-time samples $I = [n] \times [n] \times [T]$, with*

$$m = \Omega(c_s \mu_0 r n \log(nT)), \tag{26}$$

and, furthermore, $\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) \in \mathcal{B}_{\mathbf{H_A}}(R_0)$ with[3]

$$R_0 := \widetilde{c}_0 \left(\frac{\mu_0}{nT}\right)^{3/2} \frac{r^{1/2}}{\kappa(dn-r)^{1/2}} \sigma_r(\mathbf{H_A})\,. \tag{27}$$

2) [Adaptive sampling] *With $(\mu_{i,j,t})_{(i,j,t) \in I}$ being the local incoherences (24) of $\mathbf{H_A}$, Ω consists of random index triplets $(i,j,t) \in I$ that are independently observed*

[2] See Lemma A.1 in Appendix A for an explicit representation.

[3] recall that $d := \min(d_1, d_2)$

according to Bernoulli distributions with probabilities $p_{i,j,t}$ each satisfying

$$p_{i,j,t} \geq \min\left(Cc_s \frac{\mu_{i,j,t}\log(nT)}{nT} r, 1 \right), \qquad (28)$$

and, furthermore, $\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) \in \mathcal{B}_{\mathbf{H_A}}(R_0)$ with

$$R_0 := \widetilde{c}_0 \min_{(i,j,t)\in I} \left(\frac{\mu_{i,j,t}\log(nT)}{nT} \right)^{3/2} \frac{r^{1/2}\sigma_r(\mathbf{H_A})}{\kappa(dn-r)^{1/2}}. \qquad (29)$$

Then, with probability of of at least $1-2n^{-2}$, the subsequent iterates of `TOIRLS` *(Algorithm 1) converge to $\mathbf{H_A}$, i.e. $\mathcal{H}(\widetilde{\mathbf{X}}^{(k+\ell)}) \xrightarrow{\ell\to\infty} \mathbf{H_A}$, with quadratic convergence rate: for a dimension-dependent constant ν.*[4]

$$\begin{aligned} &\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k+\ell+1)}) - \mathbf{H_A}\| \\ &\leq \min(\nu\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k+\ell)}) - \mathbf{H_A}\|^2, \|\mathcal{H}(\widetilde{\mathbf{X}}^{(k+\ell)}) - \mathbf{H_A}\|). \end{aligned}$$

Theorem IV.1 justifies that the spatio-temporal transition operator recovery problem can be solved efficiently using `TOIRLS` given a number of random samples that is, up to constants, only logarithmically larger than the $r(2n-r) = O(rn)$ free parameters that are required to describe a rank-r transition operator $\mathbf{A} \in \mathcal{M}_n$. In the case of adaptive sampling, the condition (28) can be translated into a bound on the number of expected samples m_{exp} since $m_{\text{exp}} = \mathbb{E}[|\Omega|] = \sum_{(i,j,t)\in I} p_{i,j,t} \geq Cc_s\frac{r}{nT}\log(nT)\sum_{(i,j,t)\in I}\mu_{i,j,t}$ (if the constants in (28) are small enough to attain the minimum in the first argument). See Section IV-C for further discussion.

The proximity assumptions (27) and (29), which ensure that the spectral norm error of subsequent iterates of `TOIRLS` decreases with a quadratic convergence rate, are comparably restrictive due to their dependence on the n, d and T, which makes it hard to find an initialization that satisfy the conditions for large-scale problems. However, extending the convergence radius of IRLS methods remains an open problem even for simpler problems such as sparse vector and unstructured low-rank matrix recovery if a non-convex objective such as (16) is used [DDFG10, KS18, KMV21]. In Section VI-A, we provide numerical experiments illustrating that in practice, exact recovery of transition operators is observed empirically with an empirical probability of essentially 1 once enough samples are provided, even if `TOIRLS` is initialized with the natural, data-agnostic weights of $\widetilde{W}_{\mathbf{H}_0} = \mathcal{H}^*\mathcal{H}$ as in Algorithm 1.

The statements of Theorem IV.1 address the case of exact observations $\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A}))$ that are not perturbed by any noise. In Section VI-C, we provide numerical experiments suggesting that `TOIRLS` is in practice robust in the presence of noisy observations (including with a choice of the regularization parameter $\lambda = 0$).

The proof strategy for Theorem IV.1 is outlined in Section VIII and detailed in Appendix C.

We note that while Section IV-B focuses on the the behavior of Algorithm 1 in the context of the recovery of a transition operator $\mathbf{A}$ from space-time samples, it *is* possible to extend the applicability of Theorem IV.1 to Algorithm 1 recovering—more generally—rank-r block Hankel matrices by choosing as output the entire matrix $\widetilde{\mathbf{X}}^{(K)} \in \mathcal{M}_n^{\oplus T}$ instead of its restriction to its first block $\mathbf{A}^{(K)}$. In particular, in this setting, the ground truth $\mathbf{H_A}$ can be substituted by any ground truth $\mathcal{H}(\widetilde{\mathbf{X}}_0)$ with $\operatorname{rank}(\mathcal{H}(\widetilde{\mathbf{X}}_0)) = r$, using the same notions of (local) incoherence as in the presented results.

Remark IV.1. *We recall that the pencil parameter d_1 is a free parameter in Algorithm 1. Theorem IV.1 also has implications for the choice of d_1: for uniform sampling, the sampling complexity (26) is minimized if we choose d_1 such that $c_s \cdot \mu_0$ (both c_s and μ_0 depend on d_1) is minimized. The factor $c_s = \frac{T(T+1)}{d_1 d_2} = \frac{T(T+1)}{d_1(T-d_1+1)}$ is minimized for $d_1 = \lfloor (T+1)/2 \rfloor$, yielding a roughly square block Hankel matrix. Such a choice is observed to be favorable also for other problems using structured low-rank optimization* [CC14, CCY22].

A priori, the dependence of μ_0 on d_1 is unclear; however, numerical experiments conducted in Section VI-A2 suggest that this choice of d_1 also minimizes the product $c_s\mu_0$ at least in some of situations we consider.

C. Examples and Discussion of Sample Complexity

We now attempt to better understand the implications of Section IV-B and, in particular, the sample complexity conditions (26) and (28) for uniform and adaptive sampling schemes. We provide sufficient conditions on the sample complexity by providing bounds on the incoherence parameter μ_0 and local incoherences $\mu_{i,j,t}$, respectively, in various examples.

It is instructive to relate μ_0, the incoherence of the block Hankel matrix $\mathbf{H_A} = \mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$, with the now-classical incoherence notion [CT10, Che15] of the transition operator $\mathbf{A}$ (which coincides with $\mathbf{H_A}$ in the static case of $T = 1$). While in general there is no direct relationship between these two notions, as the singular vectors of $\mathbf{H_A}$ may not be always expressed in terms of the singular vectors of $\mathbf{A}$, In two particular cases, when $\mathbf{A}$ is an orthogonal matrix or a positive semi-definite matrix, it is possible to establish a simple relationship between these incoherence notions.

a) Orthogonal matrices: If $\mathbf{A} \in \mathbb{O}^n := \{\mathbf{X} \in \mathcal{M}_n : \mathbf{X}^*\mathbf{X} = \mathrm{Id}\}$, it holds that $\operatorname{rank}(\mathbf{A}) = n$. In this case, the incoherence parameter μ_0 of $\mathbf{H_A}$ satisfies

$$\mu_0 \leq 1 =: \widetilde{\mu}_0$$

and, furthermore, the local incoherences $\mu_{i,j,t}$ of $\mathbf{H_A}$ satisfy

$$\sum_{(i,j,t)\in I} \mu_{i,j,t} \leq Tn^2,$$

see Appendix A1 for details. The two parts of Theorem IV.1 therefore imply that for both uniform and adaptive sampling, $\Theta(c_s n^2\log(Tn))$ space-time samples are sufficient to establish local convergence of IRLS with high probability. These results are consistent with the intuition that a dynamical system driven by an orthogonal transition operator is energy-preserving, and from the bound $\Theta(c_s n^2\log(Tn))$, we see that up to a logarithmic factor of $\log(Tn)$, space-time samples contain a comparable amount of information to that of static samples. As the resulting sample complexity bound is of the same order in both cases, we expect adaptive sampling and uniform sampling

[4] See Appendix C for a possible choices for ν.

to exhibit similar behavior for orthogonal transition operators. We refer to Section VI-A1 for numerical experiments.

b) Positive semi-definite matrices: Let $d_1 \leq d_2$ without loss of generality. If the transition operator is a positive semidefinite matrix $\mathbf{A} = \sum_{i=1}^r \lambda_i \mathbf{u}_i \mathbf{u}_i^*$ with (positive) eigenvalues λ_i and corresponding eigenvectors $\mathbf{u}_i$, we show that the incoherence parameter μ_0 of $\mathcal{H}\big(\mathcal{Q}_T(\mathbf{A})\big)$ satisfies

$$\mu_0 \leq \max_{1\leq i\leq n} \sum_{\ell=1}^{r} \frac{n d_2 (\mathbf{u}_\ell)_i^2}{r(\sum_{s=0}^{d_1-1} \lambda_\ell^{2s})} =: \widetilde{\mu}_0. \qquad (30)$$

In particular, if $\mathbf{A}$ is a rank-r projection and if $d_1 = d_2$, this bound becomes (with e_i denoting the i-th canonical basis vector)

$$\widetilde{\mu}_0 = \max_{1\leq i\leq n} \frac{n}{r} \|\mathbf{U}^* e_i\|_2^2 =: \nu_0\,,$$

which coincides with the incoherence constant of $\mathbf{A}$ as defined in the low-rank matrix completion literature [CT10, Che15]. In this case, we obtain a space-time sampling bound $\Theta(c_s \nu_0 r n \log(nT))$, which is just slightly more than the necessary condition of $\Theta(c_s \nu_0 r n \log(n))$ for exact recovery by any method under a uniform sampling model [CT10, Che15]. For adaptive sampling, we show an upper bound for $\sum_{i,j,t} p_{i,j,t}$ as $\Theta(rn\log(nT)\log(T))$ ($T \geq 3$), and this bound can be improved to $\Theta(rn\log(nT))$ if $\mathbf{A}$ is a rank r projection. Our bound is comparable with the one obtained in [CBSW15, Theorem 2] for Bernoulli sampling for low-rank matrix completion, namely, $\Theta(rn\log^2(n))$ for the case $T = 1$. We refer to Appendix A2 for proofs of the presented estimates.

In general, $\widetilde{\mu}_0$ could be larger or smaller than ν_0, depending on the interplay of the spectrum of $\mathbf{A}$ with the coherence of the eigenvectors. For very *spiky* operators $\mathbf{A}$ with large incoherence ν_0 and quickly decaying spectrum, however, the best estimate we obtain from (30) is $\widetilde{\mu}_0 \leq d_2 \nu_0$. This implies that in such a setting, our estimates lead to a sufficient condition of $\Omega(c_s \nu_0 r n T \log(nT))$ required samples, which is rather pessimistic.

V. Computational Considerations

If $\mathbf{H}_{k-1} = \mathcal{H}(\widetilde{\mathbf{X}}^{(k-1)})$ is the block Hankel matrix at iteration $k-1$, the solution $\widetilde{\mathbf{X}}^{(k)}$ of the weighted least squares (20) can be written as (see Lemma A.8 in Appendix H)

$$\widetilde{\mathbf{X}}^{(k)} = \widetilde{W}_{\mathbf{H}_{k-1}}^{-1} P_\Omega^* \left(\lambda \,\mathrm{Id} + P_\Omega \widetilde{W}_{\mathbf{H}_{k-1}}^{-1} P_\Omega^*\right)^{-1} (\mathbf{y})\,.$$

However, using this formula directly can be impractical as we have no explicit representation of the inverse $\widetilde{W}_{\mathbf{H}_{k-1}}^{-1} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ of the effective weight operator $\widetilde{W}_{\mathbf{H}_{k-1}} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$, unlike in the case of *unstructured* low-rank optimization, where the optimization domain is not restricted to a strict linear subspace such as $\mathcal{H}(\mathcal{M}_n^{\oplus T}) \subset \mathcal{M}_{d_1 n, d_2 n}$ and for which a related IRLS method was studied in [KMV21]. A space and memory-efficient implementation of the weighted least squares step leveraging an underlying "low-rank plus diagonal" structure of $\widetilde{W}_{\mathbf{H}_{k-1}}$ can still be achieved, as can be seen in Theorem V.1.

Theorem V.1. *Let $\widetilde{\mathbf{X}}^{(k-1)} \in \mathcal{M}_{d_1,d_2}$ be the $(k-1)$-st iterate of* `TOIRLS` *(Algorithm 1) for an observation vector $\mathbf{y} \in \mathbb{R}^m$ with $m = |\Omega|$, $\widetilde{r} = r$, and $\lambda \geq 0$. Assume that $r_{k-1} = r$. Then an approximation of the k-th iterate $\widetilde{\mathbf{X}}^{(k)}$ of* `TOIRLS` *can be computed within N_{CG_inner} steps of a conjugate method solving a $O(rnT) \times O(rnT)$ linear system with space complexity of $O(rnT+m)$ and in $O(N_{CG\_inner} rT(m+n\log T+nrT))$ time.*

Theorem V.1 follows directly from Lemma A.9 in Appendix H, using the implementation outlined in Algorithm 2. The linear systems solved within Lemma A.9 can be shown to be well-conditioned under reasonable assumptions, in which case a constant number $N_{\text{CG\_inner}}$ of CG iterations is sufficient to obtain an *accurate* approximation of $\widetilde{\mathbf{X}}^{(k)}$.

As stated in the weight operator update step of Algorithm 1, the action of the effective weight operator $\widetilde{W}_{\mathbf{H}_{k-1}}^{-1}$ only uses information about the r_{k-1} leading singular vector pairs and singular values of $\mathcal{H}(\widetilde{\mathbf{X}}^{(k-1)})$. In particular, if for all iterations where the smoothing update (21) is such that $\varepsilon_k = \sigma_{\widetilde{r}+1}(\mathbf{H}_k)$, it holds that $r_k = \widetilde{r}$. This means that, in this case, only $\widetilde{r}$ singular values and singular vector pairs of $\mathbf{H}_k$ need to be computed in the weight update step of Algorithm 1, and these can be computed up to high precision using matrix-matrix multiplications with a randomized block Krylov method in [MM15, YGL18] in $O(mTr + rT(\log T + rT)n + Tnr^2)$ time (using fast multiplication with block circulant matrices, see also proof of Lemma A.9).

We conclude that one iteration of `TOIRLS` can be computed with a time complexity that is *linear* in the dimension n of the transition operator $\mathbf{A}$, at least if it is of rank $r = O(1)$. For example, if $|\Omega| = m = \Theta(rn\log(nT))$ space-time samples of an $O(1)$-incoherent ground truth $\mathbf{A}$ are provided uniformly at random, one full `TOIRLS` iteration using Lemma A.9 takes $O(nT^2\log(nT))$ time.

VI. Numerical Experiments

In this section we explore the numerical performance of `TOIRLS`, Algorithm 1 for estimating transition operators from sparse space-time samples. We consider operators $\mathbf{A} \in \mathbb{R}^{n\times n}$ associated with random graph models, as well as orthogonal matrices $\mathbf{A}$. These experiments are meant to shed light on the sharpness of our sampling complexity results Theorem IV.1, and verify they are consistent with the empirically observed behavior. While there are no dedicated computational approaches to our recovery problem available in the literature, we include also comparisons with the interior-point algorithm [WMNO06] used in the nonlinear optimization wrapper `fmincon` of MATLAB, minimizing the objective $f : \mathcal{M}_n \to \mathbb{R}$

$$f(\mathbf{B}) := \|P_\Omega(\mathcal{Q}_T(\mathbf{B})) - P_\Omega(\mathcal{Q}_T(\mathbf{A}))\|_2^2 \qquad (31)$$

using finite difference gradient approximations.

a) Numerical setup: since the number of degrees of freedom is $r(2n-r)$ for a rank-r matrix $\mathbf{A} \in \mathcal{M}_n$ and $r(n-(r-1)/2)$ for a symmetric $(n\times n)$ rank-r matrix $\mathbf{A}$, we define, for $m_{total} = |\Omega|$ space-time samples, the oversampling factor ρ as, respectively,

$$\rho = \frac{m_{total}}{r(2n-r)} \qquad \text{and} \qquad \rho = \frac{m_{total}}{r(n-(r-1)/2)}\,.$$

The average number of spatial samples taken at each time instance is $m_{single} = m_{total}/T$. We use m_1 to denote the number of samples taken at $T = 1$. We will use both the uniform and adaptive schemes described in Section I-A.

In the numerical experiments, we use TOIRLS as outlined in Algorithm 1 using the implementation described in Appendix H for computing the tangent spaces, and solving the linear systems associated to the weighted least squared problems with a conjugate gradient method.[5] Unless stated otherwise, we use Algorithm 1 with stopping criterion combining a maximal number of iterations $N_0 = 250$, a tolerance $\text{tol} = 10^{-11}$ with respect to the relative change in Frobenius norm, and $\text{tol\_CG} = 10^{-13}$ for the conjugate gradient step. We provide the true $\text{rank}(\mathbf{A})$ of the ground truth as the rank estimate $\widetilde{r} = \text{rank}(\mathbf{A})$ to the algorithm. If not stated otherwise, we provide TOIRLS with the pencil parameter $d_1 = \lfloor (T+1)/2 \rfloor$, leading an (approximately) square dimensionality of the block Hankel embedding space $\text{ran}(\mathcal{H})$.

b) Evaluation metrics: we define the recovery error of an estimator $\hat{\mathbf{A}}$ of $\mathbf{A}$ as

$$\text{Rec}_{\mathbf{A}} := \|\hat{\mathbf{A}} - \mathbf{A}\|_F / \|\mathbf{A}\|_F.$$

For a random model, unless stated otherwise, we run 10 independent trials and report the mean and standard deviation of the recovery errors.

c) Graph Topology-Induced Transition Operators: We consider operators representing dynamics on different graphs and random graph models. Let $\mathcal{G} = (V, E, \mathbf{W})$ be an *undirected weighted graph* with n vertices $V = \{v_1, \cdots, v_n\}$, edges $E \subset \mathcal{V} \times \mathcal{V}$ and adjacency matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$, i.e., $\mathbf{W}_{ij} = 1$ if $(i, j) \in E$ and $\mathbf{W}_{ij} = 0$ otherwise. The *degree* of a vertex $v_i \in \mathcal{V}$ is $\deg(v_i) = \sum_{j=1}^{n} \mathbf{W}_{ij}$. Given a graph $\mathcal{G}$, a variety of associated transition operators can be defined, encoding structural information about the graph [Chu97] and associating to the graph certain dynamical processes on it.

Definition VI.1. *The* normalized diffusion operator *of a graph* $\mathcal{G} = (V, E, \mathbf{W})$ *is* $\mathbf{A} := (\mathbf{D}^{-1})^{\frac{1}{2}} \mathbf{W} (\mathbf{D}^{-1})^{\frac{1}{2}}$, *where* $\mathbf{D} := \text{diag}(\deg(v_i))_{v_i \in V}$ *and* $\mathbf{D}^{-1}$ *denotes its pseudo-inverse. The* normalized graph Laplacian operator *is* $\mathbf{L} = \text{Id} - \mathbf{A}$. *The* random walk matrix $\mathbf{P}$ *is* $\mathbf{D}^{-1}\mathbf{W}$, *and the* heat diffusion operator *for time parameter* $\tau > 0$ *is* $\exp(-\tau \mathbf{L})$.

A. Recoverability in the Noiseless Setting

We first investigate the empirical recoverability of transition operators $\mathbf{A}$ by Algorithm 1 from spatio-temporal samples Ω given different numbers of time steps T, sampling schemes and different sample complexities. Furthermore, we consider different types of transition operators that include both full and low-rank matrices, symmetric and non-symmetric matrices, orthogonal matrices and operators associated to the topology of graphs.

1) Dependence on Number of Time Steps T: For a first experiment, we fix the number $m = |\Omega|$ of uniformly sampled space-time samples from $\mathcal{Q}_T(\mathbf{A}) := \mathbf{A} \oplus \mathbf{A}^2 \oplus \mathbf{A}^3 \oplus \ldots \oplus \mathbf{A}^T$ and consider different choices of T.

a) Random orthogonal matrices: we consider random orthogonal matrices $\mathbf{A} \in \mathbb{R}^{n \times n}$, with $n = 50$, sampled from the Haar measure on the orthogonal group $\mathcal{O}(n) = \{\mathbf{A} \in \mathbb{R}^{n \times n} : \mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}\}$. In this case, $\mathbf{A}$ has $n^2 = 2500$ degrees of freedom,[6] and we first fix the total number of space-time samples (uniform sampling) to $m_{\text{total}} = 7500$, corresponding to an oversampling factor of $\rho = 3$. We investigate the performance of the proposed approach when T is between 10 and 50, i.e., the average spatial samples per time instance ranges from 750 to 150. We report the results in Table I: our approach is able to recover $\mathbf{A}$ accurately after about 30 IRLS iterations, even when T grows larger, increasing the apparent nonlinearity of the problem. This is consistent with our theoretical analysis: in short, in an energy-preserving system, one can trade spatial samples for an equal amount of temporal samples without loss of information.

T	$\text{Rec}_{\mathbf{A}}$	m_{single}	$\text{dof}(\mathbf{A})$	ρ	Iterations
10	$(2.3 \pm 0.2) \cdot 10^{-14}$	750	2500	3	20.9 ± 0.3
20	$(5.7 \pm 0.8) \cdot 10^{-14}$	375	2500	3	25.1 ± 0.6
30	$(1.1 \pm 0.2) \cdot 10^{-13}$	250	2500	3	27.5 ± 0.8
40	$(1.1 \pm 0.1) \cdot 10^{-13}$	187.5	2500	3	30.2 ± 2
50	$(1.8 \pm 0.3) \cdot 10^{-13}$	150	2500	3	32.8 ± 2

TABLE I: The estimation errors for random orthogonal matrices of size 50×50 using uniform sampling with replacement.

In Figure 2, we report on an experiment with the same data and sampling model, but where we vary both the number of time steps $T = 1, \ldots, 40$ and the oversampling factor $\rho = 1, \ldots, 3.5$. For 24 random instances, we visualize the empirical probability of exact recovery (defined as a relative Frobenius error of $\text{Rec}_{\mathbf{A}} < 10^{-4}$). We observe the existence of a sharp phase transition between no recovery and exact recovery for all instances, at an oversampling factor between

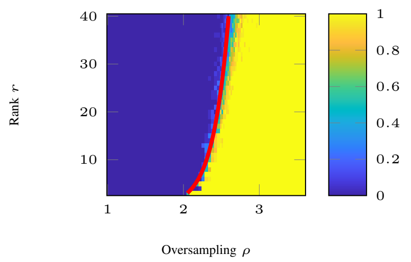


Fig. 2: Phase transition plot for orthogonal matrices, with oversampling factor ρ on the x-axis and time steps T on y-axis. Yellow corresponds to exact recovery for all random instances of the problem, blue corresponds to no recovery. Red line: $1 + 0.21 \log(nT)$.

[5] For problem instances with relatively large ambient dimension n, such as the Minnesota road network graph of Section VI-A4, we use a MATLAB implementation that follows closely the steps outlined in the proof of Lemma A.9.2. For problems with larger number of time steps T, we used matrix-vector multiplications in Algorithm 2 that include antiaveraging of block Hankel matrices instead of block-wise fast Fourier transforms as these turned out to be faster for the problem dimensions we were interested in.

[6] In fact, an orthogonal matrix has only $n(n-1)/2$ degrees of freedom; however, as reconstruction method is oblivious to the orthogonality constraints, we neglect these in our calculation.

$\rho = 2$ and $\rho = 2.7$, depending weakly on T. This is consistent with Theorem I.1, which predicts exact recovery from $\rho n^2 \gtrsim n^2 \log(nT)$ samples, since here $\mu_0 = 1$, cf. Section IV-C. In fact, the phase transition in Figure 2 occurs at around $\rho \approx 1 + 0.21 \log(nT)$ for the tested parameters.

b) Erdős-Rényi Graphs.: In the next experiment, we consider graph matrices $\mathbf{A}$ associated with Erdős-Rényi graphs [ER59, Gil59] with $n = 60$ nodes with connectivity probability of $p = 0.8$. With $\mathcal{T}_r(\cdot)$ the map from a matrix to its best rank-r approximation, for r between 1 and 60, we create rank-truncated continuous-time heat diffusion operators $\mathbf{A} = \mathcal{T}_r\left(\exp(-\tau\mathbf{L})\right) \in \mathbb{R}^{n\times n}$, with $\mathbf{L}$ the normalized graph Laplacian as in Definition VI.1, $\tau = 0.4$, on an instantiation of an Erdős-Rényi graph. In Figure 3, we depict $\mathcal{Q}_T(\mathbf{A})$ for such a transition operator, with $r = 20$ and $T = 7$ time steps.

As they are symmetric, such matrices have $\mathrm{dof}(\mathbf{A}) = r(n-(r-1)/2)$ degrees of freedom. In Figure 4, we visualize the recovery performance of `TOIRLS` for varying numbers of samples $m = |\Omega|$, for three different numbers of time steps T. In the left column of Figure 4, we see that the phase transition for $T = 1$ occurs extremely close to the information theoretical limit—in this case, the setting coincides with low-rank matrix completion via `MatrixIRLS` as described in [KMV21]. For $T = 4$, the transition occurs at around $m = 1.5rn$. Apart from the fact it is expected that generally, the phase transition will occur at larger sample complexities than for $T = 1$ due its the logarithmic dependence on T, it is remarkable that the quadratic dependence of $\mathrm{dof}(\mathbf{A})$ on r is not reflected in the empirical transition curve. However, this is still compatible with Theorem IV.1, as dependence on r in the sufficient condition is linear. As expected due to the logarithmic dependence on T, we observe a similar, but slightly deteriorated transition curve for $T = 7$.

2) Choice of Pencil Parameter d_1.: In the experiments of Section VI-A1, we always chose the first pencil parameter d_1 so that block Hankel matrices $\mathcal{H}(\widetilde{\mathbf{X}})$ are square or as square as possible, i.e., such that $d_1 = \lfloor (T+1)/2 \rfloor$.

Revisiting the experiments of Section VI-A1 for the Erdős-Rényi graph model and $T = 7$ time steps, we now explore the sensitivity of the problem to the choice of d_1. In Figure 5, we observe that for $d_1 = 1/d_2 = 7$, the phase transition occurs only for significantly more samples m than for the square choice of $d_1 = d_2 = 4$; for example, it can be seen that for $r = 20$, the transition is at $m = 3600$ or $\rho \approx 3.56$ for $d_1 = 1$, whereas it is at $m = 2500$ or $\rho \approx 2.47$ for $d_1 = 4$. For $d_1 = 1$, the recovery problem becomes impossible if the rank of $\mathbf{A}$ satisfies $r = 60$ due to a lack of any low-rank property of the embedding matrix $\mathcal{H}\big(\mathcal{Q}_T(\mathbf{A})\big)$, and the experiment indicates that even for lower ranks $r \ll 60$, this choice of d_1 is disadvantageous. For $d_1 = 2$ and $d_1 = 3$, the behavior is quite similar to the square case in this example with a just slightly worse phase transition.

Furthermore, we illustrate in the last column of Figure 5 the values of the d_1-dependent product $c_s\mu_0$, where $c_s = \frac{T(T+1)}{d_1 d_2}$ is the constant of Definition IV.1 and μ_0 is the incoherence parameter (25) of $\mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$ for a given choice of the pencil parameter d_1. The values are illustrated with a one standard deviation confidence interval across 24 realizations of the Erdős-Rényi model. We observe that $c_s\mu_0$ is minimal for $d_4 = 1$ for essentially all ranks r, indicating that our sample complexity bound (26) in Theorem IV.1 indeed justifies a square choice for the pencil parameter such that $d_1 = \lfloor (T+1)/2 \rfloor$.

3) Uniform Sampling vs. Adaptive Sampling: Next, we explore the empirical benefits of *adaptive sampling* compared to uniform sampling for the recovery of transition operators $\mathbf{A}$. In particular, we assume that we have knowledge about the local incoherences $\mu_{i,j,t}$ of $\mathbf{A}$ for all $(i,j,t) \in I$, see (24) in Section IV-A, and design an adaptive sampling scheme with probabilities $p_{i,j,t} = c\mu_{i,j,t}$ for all $(i,j,t) \in I$, where $c > 0$ is chosen such that the expected number of samples $m_{\text{exp}} = \mathbb{E}[|\Omega|] = \sum_{(i,j,t)\in I} p_{i,j,t}$. We vary then m_{exp} in a similar manner as m above for uniform sampling. We note that this is not a very realistic sampling scheme, since local incoherences are not immediately accessible as they require the knowledge of $\mathbf{A}$. An implementable approximation of the ideal adaptive sampling scheme was proposed in [CBSW15] for the related low-rank matrix completion problem; however, an application to our setting is beyond the scope of this paper.

In Figure 6, we illustrate a realization of an expected number of $m_{\text{exp}} = 3000$ adaptive samples in the Erdős-Rényi setting of Section VI-A1, corresponding to an oversampling factor $\rho \approx 2.97$. Applying `TOIRLS` to the recovery of heat diffusion operators associated with Erdős-Rényi graphs from adaptive sampling, we report the results of the experiment of Section VI-A1 in Figure 7. It can be seen that for $T = 1$ the phase transition is very similar to the one corresponding to uniform sampling (see Figure 4), as it was already close to the information theoretic threshold $\rho = 1$ (red curve). For the dynamic cases $T = 4, 7$, we see that we obtain a modest improvement compared to uniform sampling, with the phase transition exceeding the line $m_{\text{exp}} = 1.5rn$ especially for large r and $T = 4$, and achieving recoverability of full rank operators from around $6{,}000$ samples for $T = 7$, unlike

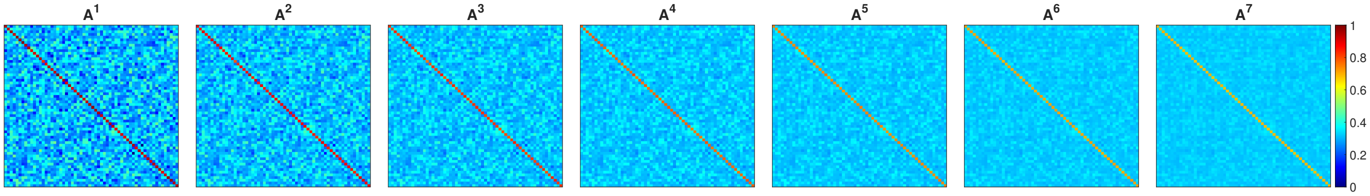


Fig. 3: Left column: Realization of Erdős-Rényi graph with $n = 60$ and $p = 0.8$. Other columns: Aggregation of matrix of powers $\mathcal{Q}_T(\mathbf{A})$ of rank-20 truncation $\mathbf{A}$ of heat diffusion operator (color scheme normalized across powers, log-scale).

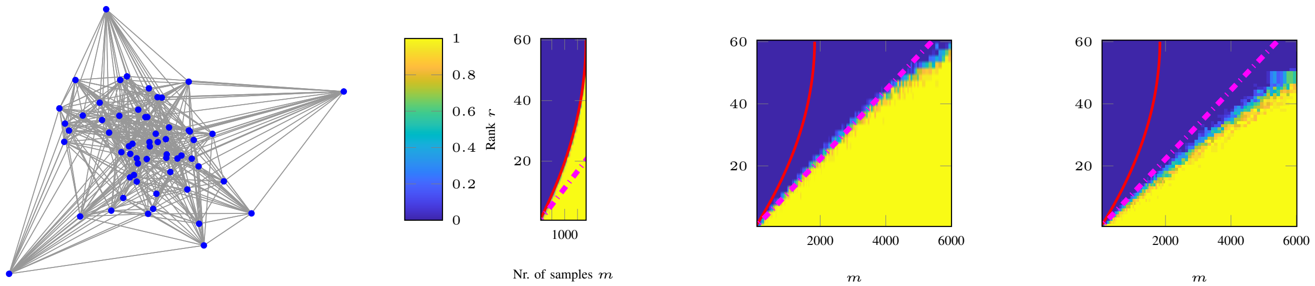


Fig. 4: Left: Erdős-Rényi graph with $n = 60$, $p = 0.8$. Center to right: Phase transition plots for rank r-approximations of heat diffusion operators of Erdős-Rényi graphs, *uniform sampling*. Increasing rank on y-axis, increasing number of samples m on x-axis. Center left: $T = 1$. Center right: $T = 4$. Right: $T = 7$. Red curved line: Number of degrees of freedom $\mathrm{dof}(\mathbf{A}) = r(n - (r - 1)/2)$ of operator $\mathbf{A}$; Pink dotted line: $1.5rn$.

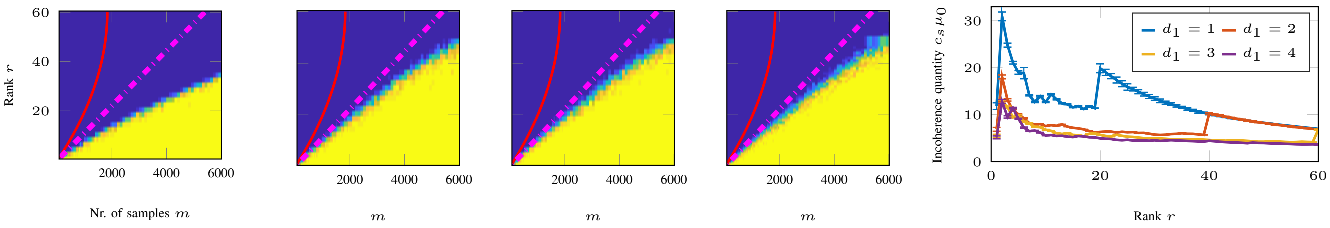


Fig. 5: First columns: Experiment as in Figure 4 for $T = 7$, for different pencil parameters d_1. From left to right: $d_1 = 1$, $d_1 = 2$, $d_1 = 3$ and $d_1 = 4$. Last column: Value of $c_s\mu_0$ of $\mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$ for $d \in \{1, \ldots, 4\}$.

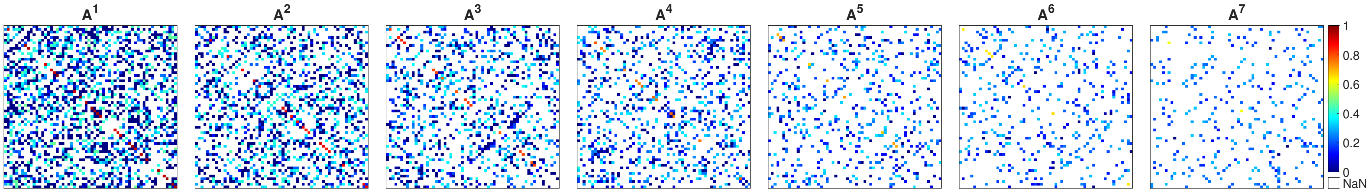


Fig. 6: Adaptive space-time samples $P_\Omega(\mathcal{Q}_T(\mathbf{A}))$ of rank-20 truncation $\mathbf{A}$ of heat diffusion operator (1010 degrees of freedom, log-scale) with $m_{\text{exp}} = 3000$.

in the uniform sampling case.

a) Community Graphs: the improved efficiency of the adaptive sampling scheme has been rather modest for the heat diffusion operator based on a Erdős-Rényi graph in our parameter setting due to the relatively benign spectral decay of $\mathbf{A}$. We now consider a *community graph* with $n = 60$ vertices and 10 communities (eight of size 5, one of size 13 and one of size 7), with dense connections within a community, and independent random inter-community edges with probability $1/10$. We use the Graph Signal Processing (GSP) toolbox [PPS+14] to create such graphs, and define the associated transition matrix as the truncated random walk matrix $\mathbf{A} = \mathcal{T}_r(\mathbf{P}) = \mathcal{T}_r\left(\mathbf{D}^{-1}\mathbf{W}\right)$, cf. Definition VI.1. Note that this matrix is in general asymmetric. For $r = 20$ and $T = 7$, we visualize $\mathcal{Q}_T(\mathbf{A})$, i.e., $\mathbf{A}$ and its powers $\mathbf{A}^2, \ldots, \mathbf{A}^7$ in Figure 8, together with an example of adaptive samples for this setting with $m_{\text{exp}} = 3,000$, computed based on local incoherences. Comparing Figure 8 with the adaptive sampling pattern for the Erdős-Rényi heat diffusion model (Figure 6), we note that the sampling density for larger time steps such as $t = 5, 6, 7$ is smaller for community graphs, indicating that the adaptive sampling focuses now more on smaller time scales than for the Erdős-Rényi model. This is expected, since the spectrum of the (untruncated) transition matrix decays faster than for the Erdős-Rényi heat diffusion operator above, indicating that sampling large time steps is less informative than sampling earlier time steps, see also Figure 8.

Empirically, this is confirmed in the experiment of Figure 9, where we report on the phase transition for both adaptive sampling and uniform sampling, considering $T = 7$ steps of a random walk. Unlike for the Erdős-Rényi transition operators (Figure 4 and Figure 7), we observe a significant difference between adaptive and uniform sampling for this model, as the uniform sampling scheme requires approximately the double amount of samples to obtain exact recovery, with this phase transition being located at around $m = 4.8rn$ (uniform sampling) and $m = 2.4rn$ (adaptive sampling), respectively.

4) Dependence on Graph Topology: We now elucidate how the recovery of transition operators $\mathbf{A}$ by Algorithm 1 depends on the *topology* of an underlying *graph*.

a) Random walk matrix: We recall that a random walk matrix, cf. Definition VI.1, is suitable to reveal structural information of a graph: The multiplicity of the eigenvalue 1 is equal to the number of connected components; the second

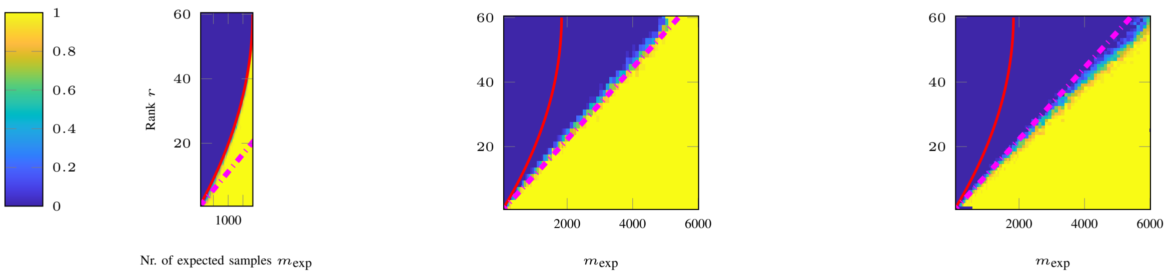


Fig. 7: Experiment as in Figure 4 for *adaptive sampling* with m_{exp} expected space-time samples. Left: $T = 1$. Center: $T = 4$. Right: $T = 7$. Red curved line: Number of degrees of freedom $\text{dof}(\mathbf{A}) = r(n - (r - 1)/2)$ of operator $\mathbf{A}$; Pink dotted line: $1.5rn$.

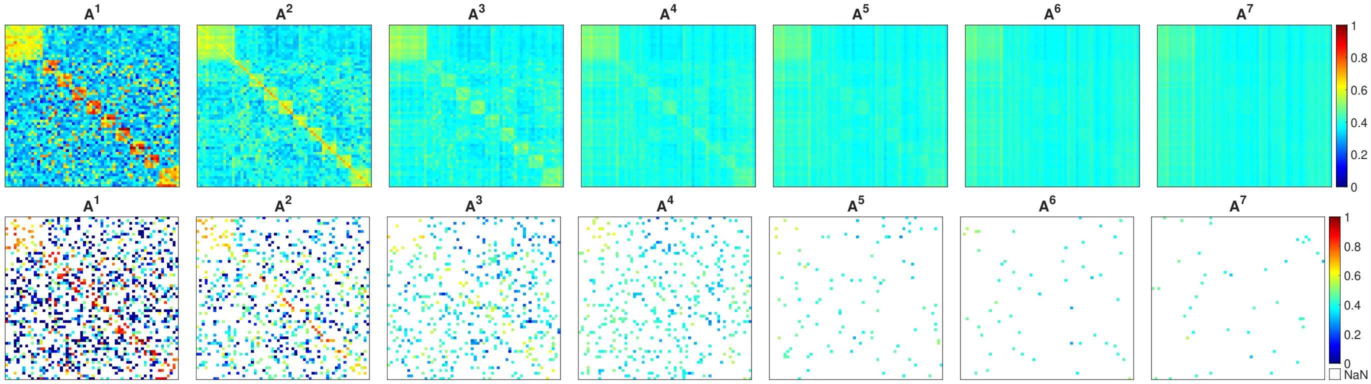


Fig. 8: First row: Rank-r truncation $\mathbf{A}$ of random walk transition operator of random community graph with 60 nodes for $r = 20$, aggregation $\mathcal{Q}_T(\mathbf{A})$ of $T = 7$ matrix powers; Second row: Adaptive space-time samples $P_\Omega(\mathcal{Q}_T(\mathbf{A}))$ of rank-20 truncation $\mathbf{A}$ (log-scale) with $m_{\text{exp}} = 3000$

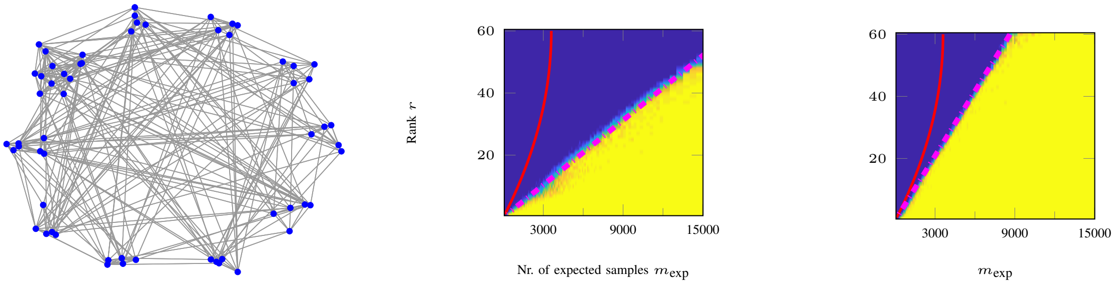


Fig. 9: Left: Community graph with $n = 60$ vertices in 10 communities; Center and right: Phase transition, uniform vs. adaptive sampling for the recovery of truncated random walk matrices of random community graph, $T = 7$ time steps. Center: Uniform sampling, pink dotted line: $4.8rn$. Right: Adaptive sampling, pink dotted line: $2.4rn$. Red curved line: $\text{dof}(\mathbf{A}) = r(2n - r)$ of $(n \times n)$-matrix $\mathbf{A}$ of rank-r.

largest eigenvalue λ_2 that describes the mixing rate of the random walks; the spectral gap $|\lambda_1 - \lambda_2|$ represents how well the graph is connected. We refer the readers to [Chu97] for a detailed discussion.

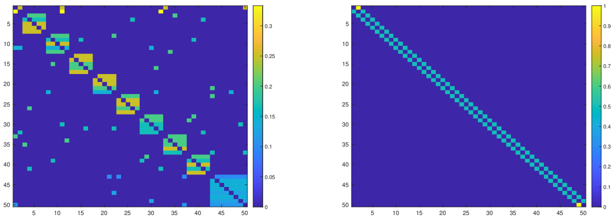
In Table II, we report on experiments on the recovery of (full-rank) random walk matrices $\mathbf{A} = \mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$ associated to two different graphs, both with $n = 50$ nodes: A very regular *path graph* , and a more irregular community graph (one community of size 9, eight of size 5, and a single node) with inter-cluster connection probability of $1/50$, cf. Figure 10. As in Section VI-A3, we use both uniform and adaptive sampling. We denote the number of degrees of freedom by $\text{dof}(\mathbf{A})$ and, in case of adaptive sampling, the number of samples located at time $t = 1$ (averaged across 10 realizations) as $\overline{\mathrm{m}_1}$.

We observe that for the path graph, recovery by Algorithm 1 is possible for uniform sampling at an oversampling factor of $\rho = 3$, as the recovery error $\text{Rec}_{\mathbf{A}}$ is of the order of magnitude of the algorithmic tolerance with $\text{Rec}_{\mathbf{A}} \approx 10^{-10}$, while exact recovery fails for $\rho = 2.8$ even if adaptive sampling is chosen; i.e., the performance is essentially the same for uniform and adaptive sampling. For the community graph, for which we now consider $T = 10$ time steps instead of $T = 5$, an oversampling factor of $\rho = 8$ is not sufficient for uniform sampling to recover the transition operator, however, a much smaller sample complexity corresponding to $\rho = 3.5$ leads

Models	Sampling	T	$\mathrm{m}_{\mathrm{single}}$	$\overline{\mathrm{m}_1}$	$\mathrm{dof}(\mathbf{A})$	ρ	$\mathrm{Rec}_{\mathbf{A}}$
Path graph	uniform	5	1500		2500	3	$9.6 \cdot 10^{-11} \pm 9.2 \cdot 10^{-20}$[7]
Path graph	adaptive	5		2169	2500	2.8	$7.2 \cdot 10^{-4} \pm 2.3 \cdot 10^{-3}$
Community	uniform	10	2000		2500	8	$7.6 \cdot 10^{-5} \pm \cdot 5.8 \cdot 10^{-8}$
Community	adaptive	10		2297	2500	3.5	$4.4 \cdot 10^{-13} \pm \cdot 5.8 \cdot 10^{-13}$

TABLE II: Recovery errors $\mathrm{Rec}_{\mathbf{A}}$ of Algorithm 1 for random walk matrix of path/community graphs for different sampling sets.

already to exact recovery for adaptive sampling. While these graphs are simple examples, they show illustrate that the difficulty of the setup is negatively affected by the irregularity of the underlying graph.



(a) Community graph

(b) Path graph

Fig. 10: The plot of two random walk matrices used in the simulation. Both of matrices are sparse with significant nonzero entries.

b) Heat diffusion operator: We now revisit the heat diffusion operators $\mathbf{A} = \mathcal{T}_r(\exp(-\tau\mathbf{L}))$ from Section VI-A1, focussing on how the number of time steps T, the heat diffusion scale τ and the structure of the underlying graph $\mathcal{G}$ determine the recoverability of $\mathbf{A}$ from time-space samples that are sampled uniformly at random, for settings of slightly larger scale. To that end, we consider a Swiss roll graph with $n = |V| = 200$ nodes and graph representing the roads of the state of Minnesota [DH11, KAB+19] with $n = |V| = 2642$ nodes, using the default settings of the GSP toolbox [PPS+14]; see Figure 11 for a visualization.

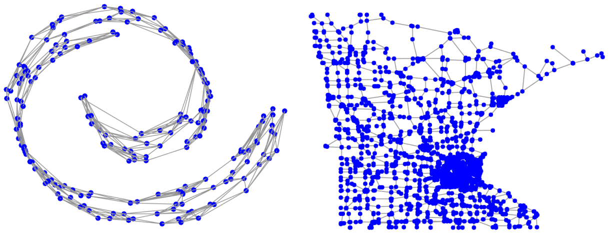


Fig. 11: Left: Swiss roll graph. Right: Minnesota road network graph.

Furthermore, we consider heat diffusion operators corresponding to *slow* and *fast* energy dissipation, corresponding to small and large choices of τ. Since we choose $r = 10$, we have that $\lambda_{min}(\mathbf{A}) = \exp(-\tau\lambda_{10}(\mathbf{L}))$, which is larger for slow energy dissipation and smaller or faster energy dissipation. Here, $\lambda_{10}(\mathbf{L})$ corresponds to the 10th largest eigenvalue of the Laplacian $\mathbf{L}$.

In Table III, we report parameter choices and sample sizes and the smallest oversampling factor ρ (stepsize 0.5) that allows for accurate recovery, i.e., $\mathrm{Rec}_{\mathbf{A}}$ of the order of the stopping condition, of the transition operator via Algorithm 1. We observe the transition operator is recoverable even if the expected number of samples $\mathrm{m}_{\mathrm{single}}$ for a single time step is below the number of degrees of freedom $\mathrm{dof}(\mathbf{A})$, which would not be possible in the static setting of $T = 1$. Furthermore, we see that in the case of faster energy dissipation ($\lambda_{min}(\mathbf{A})$ small), the threshold oversampling factor ρ is larger. This is consistent with the worse bound $\widetilde{\mu}_0$ in (30) and the sufficient condition (26) for the local convergence of Algorithm 1.

B. Comparison with Black-Box Nonlinear Optimization

We now compare the performance of Algorithm 1 for the problem studied in this paper to the one of a black-box nonlinear optimization solver applied to the nonlinear least squares objective (31) [WMNO06], as used by the wrapper function `fmincon` of MATLAB. For `fmincon`, we use the zero-padded observations $P_\Omega^* P_\Omega(\mathcal{Q}_T(\mathbf{A}))$ as initialization.

Unlike Algorithm 1, this method is not able to algorithmically utilize the low-rank structure of the problem in the case of rank-truncated transition operators $\mathbf{A}$, which is why it is not suitable to handle large-scale problem instances with many unknowns such as, for example, those associated to the Minnesota road network graph considered in Table III–in fact, it is infeasible to run it on a personal computer already for transition operator sizes of $n > 200$, unlike Algorithm 1.

Instead, we consider rank-truncated heat diffusion operator associated to the Swiss roll graph used in the first row of Table III, and the random walk matrices associated to a path graph and the community graph of Table II, each with uniform sampling.

Setting the maximal number of iterations equal to 200, we report the observed recovery errors $\mathrm{Rec}_{\mathbf{A}}$ in Table IV. We see that for the oversampling factors for which Algorithm 1 essentially leads to exact recovery, `fmincon` exhibits recovery errors of the order 10^{-1} or 10^{-2} for the Swiss roll and community graph models, indicating that exact recovery does not happen. For the path graph, on the other hand, the recovery error is of order 10^{-7}. Taking the often larger computational cost of the generic nonlinear optimization solver into account, while it cannot be ruled out that an increase in the iteration number will eventually lead to smaller errors, we conclude that Algorithm 1 works significantly better than `fmincon` for irregular graphs.

C. Robustness to Noisy Observations

In all previous experiments in Section VI, we have assumed that the observations $\mathbf{y} \in \mathbb{R}^m$ provided to the recovery method

Models	$\lambda_{min}(\mathbf{A})$	T	$\mathrm{m}_{\text{single}}$	$\mathrm{dof}(\mathbf{A})$	ρ	$\mathrm{Rec}_{\mathbf{A}}$
Swiss roll	0.21	6	1629	1955	5	$1.4 \cdot 10^{-13} \pm 4.2 \cdot 10^{-14}$
Swiss roll	0.89	6	978	1955	3	$1.9 \cdot 10^{-12} \pm 3.4 \cdot 10^{-12}$
Minnesota	0.64	5	32082	26375	6	$8.3 \cdot 10^{-13} \pm 5.2 \cdot 10^{-13}$
Minnesota	0.9	5	21100	26375	4	$3.8 \cdot 10^{-11} \pm 6.6 \cdot 10^{-11}$

TABLE III: Parameter choices for accurate recovery of rank-10 heat diffusion operators using uniform sampling with replacement, sample complexities at phase transition.

Models	Sampling	n	Rank r	T	$\mathrm{m}_{\text{single}}$ or $\overline{\mathrm{m}_1}$	$\mathrm{dof}(\mathbf{A})$	ρ	$\mathrm{Rec}_{\mathbf{A}}$
Swiss roll	uniform	200	10	6	1629	1955	5	$5.5 \cdot 10^{-1} \pm 3.3 \cdot 10^{-2}$
Path graph	adaptive	50	50	5	2169	2500	3	$6.8 \cdot 10^{-8} \pm 2.4 \cdot 10^{-8}$
Community	adaptive	50	50	10	2297	2500	3.5	$1.5 \cdot 10^{-2} \pm 2.3 \cdot 10^{-3}$

TABLE IV: Recovery errors using interior-point solver of nonlinear least squares formulation (31) (`fmincon`).

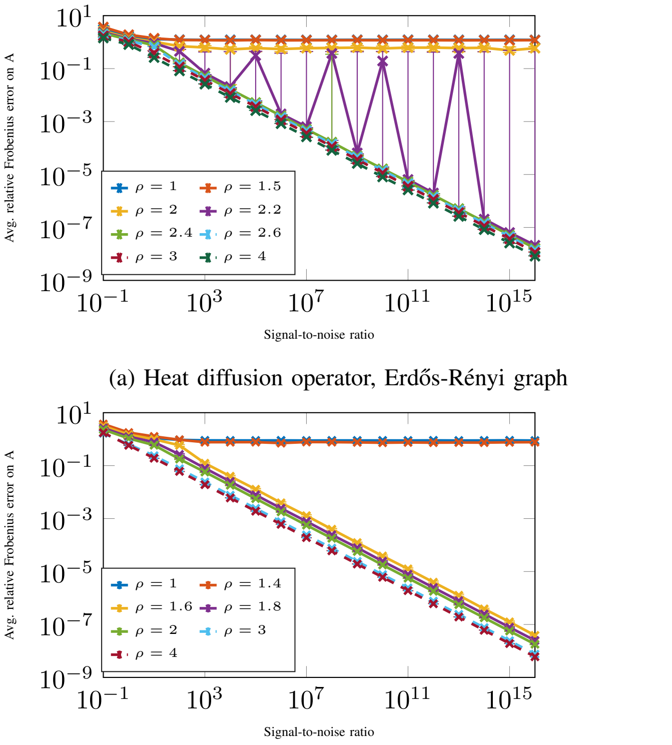
correspond to *exact* space-time samples $\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A}))$. While, taken literally, the local convergence statements of Theorem IV.1 only apply to this setting, for practical applicability it is important that the problem is also solvable in the presence of *additive noise* such that $\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A})) + \eta$, where the noise η is unknown to the algorithm. We investigate the noise robustness of the IRLS approach of Algorithm 1 by reconsidering the Erdős-Rényi and community graph transition operator models of Section VI-A1 and Section VI-A3 for noisy observations with random spherical noise such that $\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A})) + \eta = P_\Omega(\mathcal{Q}_T(\mathbf{A})) + \frac{\|P_\Omega(\mathcal{Q}_T(\mathbf{A}))\|_2}{\sqrt{\mathrm{SNR}}}\mathbf{v}$, where $\mathbf{v}$ is a vector drawn uniformly at random from the unit sphere and SNR correponds to the signal-to-noise ration $\mathrm{SNR} = \|P_\Omega(\mathcal{Q}_T(\mathbf{A}))\|_2^2 / \|\eta\|_2^2$. Despite the presence of noise, we apply Algorithm 1 with regularization parameter $\lambda = 0$.

We observe that for sample complexities below the phase transition thresholds in Section VI-A1 and Section VI-A3, the resulting recovery errors $\mathrm{Rec}_{\mathbf{A}}$ are consistently of the order 10^{-1} to 10^{1}, as even in the noiseless case recovery is not possible. For ρ chosen above the phase transition threshold, we observe a linear decrease in $\mathrm{Rec}_{\mathbf{A}}$ with respect to the SNR in the log-log plots of Figure 12 with an approximate slope of $-1/2$, empirically supporting the relationship

$$\|\hat{\mathbf{A}} - \mathbf{A}\|_F \asymp 1/\sqrt{\mathrm{SNR}} \asymp \|\eta\|_2 \tag{32}$$

whenever exact recovery occur in the noiseless case. For $\rho = 2.2$ in Figure 12(a), we see that the accuracy of the outputs of Section VI has large variance: this is because the sample complexity is set to be right at the phase transition for the uniform sampling (see Figure 4).

We recall that in view of (15) and the majorization-minimization interpretation of IRLS [DDFG10, KMV21, KKMV22], it is possible to choose a regularization parameter λ in Algorithm 1 that is adapted to the noise level via cross validation, leading to a potential improvement in the dependency of $\mathrm{Rec}_{\mathbf{A}}$ with respect to $\|\eta\|_2$. This necessitates determining an additional free parameter in the method. In fact, Figure 12 and (32) suggest that the improvement might be modest, of the order of a constant, and that the choice $\lambda = 0$ may be a valid option even in the case of noisy observations.



(a) Heat diffusion operator, Erdős-Rényi graph

(b) Random walk, Community graph

Fig. 12: Median recovery error $\mathrm{Rec}_{\mathbf{A}}$ with 25% and 75% quantiles, vs. signal-to-noise ratio (SNR), 100 realizations, different oversampling factors ρ. $T = 7$ time steps, $n = 50$ nodes, rank $r = 20$.

VII. Proof of the Low-Rank Property of Block Hankel Matrix

Before proceeding with the proof of Theorem II.1, we show a corollary of Theorem II.1 that generalizes the well-known Vandermonde decomposition for Hankel matrices.

Corollary VII.1 (Generalized Vandermonde Decomposition)**.** *Let* $\mathbf{H} \in \mathcal{M}_{d_1 n, d_2 n}$ *be a block Hankel matrix with* $(n \times n)$ *blocks, that has a positive semidefinite square extension* $\overset{\square}{\mathbf{H}} \in \mathcal{M}_{Dn,Dn}$*,* $D = \max(d_1, d_2)$*, such that*

- *its first block* $\mathbf{H}_1 \in \mathcal{M}_n$ *if of rank* r*, and*
- *at least one other block* $\mathbf{H}_j \in \mathcal{M}_n$*,* $j > 1$*, is of rank* r*.*

Then there exists a triple $(\mathbf{U}, \mathbf{N}, \Sigma)$ *with* $\mathbf{U} \in \mathcal{M}_{n,r}$ *with orthonormal columns,* $\Sigma \in \mathcal{M}_r$ *positive definite diagonal, and* $\mathbf{N} \in \mathcal{M}_r$ *that is* Σ*-self-adjoint*[8]*, such that each block* $\mathbf{H}_k$ *satisfies*

$$\mathbf{H}_k = \mathbf{U}\mathbf{N}^{k-1}\Sigma\mathbf{U}^*\,,$$

and $\mathbf{H}$ *has the generalized Vandermonde decomposition*

$$\mathbf{H} = \mathcal{V}_{d_1}(\mathbf{U}, \mathbf{N})\Sigma\mathcal{V}_{d_2}(\mathbf{U}, \mathbf{N})^*, \tag{33}$$

where

$$\mathcal{V}_m(\mathbf{N}, \Lambda) = \begin{pmatrix}(\mathbf{U}\mathbf{N}^0)^* & (\mathbf{U}\mathbf{N}^1)^* & \dots & (\mathbf{U}\mathbf{N}^{m-1})^*\end{pmatrix}^* \\ \in \mathbb{R}^{nm \times r}$$

is a generalized Vandermonde matrix *with* $m \in \mathbb{N}$*.*

Proof of Corollary VII.1. If $\mathbf{Y} \in \mathcal{M}_{n,r}$ and $\mathbf{M} \in \mathcal{M}_r$ with $\operatorname{rank}(\mathbf{H}) = \operatorname{rank}(\mathbf{Y}) = \operatorname{rank}(\mathbf{M}) = r$ are the matrices from Theorem II.1.2, we can write the square extension $\overset{\square}{\mathbf{H}} \in \mathcal{M}_{Dn,Dn}$ of $\mathbf{H}$ as $\overset{\square}{\mathbf{H}} = \mathbf{L}\mathbf{L}^*$, where $\mathbf{L} = \begin{bmatrix}\mathbf{L}_1^*, & \dots, & \mathbf{L}_D^*\end{bmatrix}^* \in \mathcal{M}_{Dn,r}$ is a block matrix with $\mathbf{L}_1 = \mathbf{Y}$ and

$$\mathbf{L}_{j+1} = \mathbf{Y}\mathbf{M}^j$$

for $1 \leq j \leq D-1$. Let $\mathbf{Y} = \mathbf{U}\widetilde{\Sigma}\mathbf{V}^*$ be a singular value decomposition, where $\widetilde{\Sigma} \in \mathcal{M}_r$ contains the non-zero singular values of $\mathbf{Y}$, $\mathbf{U} \in \mathcal{M}_{n,r}$ the corresponding left singular vectors and $\mathbf{V} \in \mathcal{M}_r$ the right singular vectors in their columns. Defining $\Sigma = \widetilde{\Sigma}^2$ and $\mathbf{N} = \widetilde{\Sigma}\mathbf{V}^*\mathbf{M}\mathbf{V}\widetilde{\Sigma}^{-1}$, we can write each block $\mathbf{H}_k$ of the block Hankel matrix $\mathbf{H}$ as

$$\begin{aligned}\mathbf{H}_k &= \mathbf{Y}\mathbf{M}^{k-1}\mathbf{Y}^* = \mathbf{U}\widetilde{\Sigma}\mathbf{V}^*\mathbf{M}^{k-1}\mathbf{V}\widetilde{\Sigma}^*\mathbf{U}^* \\ &= \mathbf{U}\widetilde{\Sigma}\mathbf{V}^*\mathbf{M}^{k-1}\mathbf{V}\widetilde{\Sigma}^{-1}\widetilde{\Sigma}^2\mathbf{U}^* = \mathbf{U}\mathbf{N}^{k-1}\Sigma\mathbf{U}^*,\end{aligned}$$

since

$$\begin{aligned}\mathbf{N}^k &= (\widetilde{\Sigma}\mathbf{V}^*\mathbf{M}\mathbf{V}\widetilde{\Sigma}^{-1})^k = \widetilde{\Sigma}\mathbf{V}^*\mathbf{M}\mathbf{V}\widetilde{\Sigma}^{-1}(\widetilde{\Sigma}\mathbf{V}^*\mathbf{M}\mathbf{V}\widetilde{\Sigma}^{-1})^{k-1} \\ &= \widetilde{\Sigma}\mathbf{V}^*\mathbf{M}^k\mathbf{V}\widetilde{\Sigma}^{-1}\end{aligned}$$

for each $k \in \mathbb{N}$, using the fact that $\mathbf{V}$ has orthogonal columns. Furthermore, we can verify that $\mathbf{N}$ is Σ-adjoint, since

$$\Sigma\mathbf{N}^* = \widetilde{\Sigma}^2\widetilde{\Sigma}^{-1}\mathbf{V}^*\mathbf{M}\mathbf{V}\widetilde{\Sigma} = \widetilde{\Sigma}\mathbf{V}^*\mathbf{M}\mathbf{V}\widetilde{\Sigma}^{-1}\widetilde{\Sigma}^2 = \mathbf{N}\Sigma,$$

using also that $\mathbf{M}$ is symmetric and that $\Sigma = \widetilde{\Sigma}^2$. The Σ-adjointness of $\mathbf{N}$ allows us to write the occurrence of $\mathbf{H}_k$ in the i-th row block and the j-th column block as

$$\mathbf{H}_k = \mathbf{U}\mathbf{N}^{i-1}\Sigma(\mathbf{N}^*)^{j-1}\mathbf{U}^*$$

for i and j satisfying $k = i + j - 1$. From this, we see that $\mathbf{H}$ attains the generalized Vandermonde decomposition (33) since

$$\mathbf{H}_k = \left(\mathcal{V}_{d_1}(\mathbf{U}, \mathbf{N})\Sigma\mathcal{V}_{d_2}(\mathbf{U}, \mathbf{N})^*\right)_{i,j} = \mathbf{U}\mathbf{N}^{i-1}\Sigma(\mathbf{N}^*)^{j-1}\mathbf{U}^*$$

for i, j with $k = i + j - 1$, using the definition $\mathcal{V}_m(\mathbf{N}, \Lambda) = \begin{pmatrix}(\mathbf{U}\mathbf{N}^0)^* & (\mathbf{U}\mathbf{N}^1)^* & \dots & (\mathbf{U}\mathbf{N}^{m-1})^*\end{pmatrix}^*$ for $m = d_1$ and $m = d_2$, respectively. □

We now continue with the proof of Theorem II.1. The proof has similarities to the proof of a similar result for block Toeplitz matrices [YXS16, Lemma 2].

Proof of Theorem II.1. For the first statement of Theorem II.1, let $\mathbf{A}$ be a rank-r matrix and denote by $\mathbf{A} = \mathbf{U}\mathbf{J}\mathbf{U}^{-1}$ its Jordan decomposition, i.e., $\mathbf{U} \in \mathbb{C}^{n \times n}$ is invertible and $\mathbf{J} \in \mathbb{C}^{n \times n}$ is an upper triangular matrix with $\operatorname{rank}(\mathbf{J}) = r$. This decomposition shows that $\mathcal{H}\big(\mathcal{Q}_T(\mathbf{A})\big) \in \mathcal{M}_{d_1 n, d_2 n}$ is similar to the matrix

$$\begin{aligned}\mathbf{S} :&= \begin{pmatrix}\mathbf{J} & \mathbf{J}^2 & \dots & \mathbf{J}^{d_1} \\ \mathbf{J}^2 & \mathbf{J}^3 & \dots & \mathbf{J}^{d_1+1} \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{J}^{d_2} & \mathbf{J}^{d_2+1} & \dots & \mathbf{J}^T\end{pmatrix} \\ &= \begin{pmatrix}\mathbf{J} \\ \mathbf{J}^2 \\ \vdots \\ \mathbf{J}^{d_2}\end{pmatrix}\begin{pmatrix}\mathrm{Id} & \mathbf{J} & \cdots & \mathbf{J}^{d_1-1}\end{pmatrix} =: \mathbf{S}_1\mathbf{S}_2,\end{aligned}$$

where $\mathrm{Id} \in \mathbb{R}^{n \times n}$ is the identity matrix. We note that $\operatorname{rank}(\mathbf{S}_1) \geq \operatorname{rank}(\mathbf{J}) = r$, as the linear independence of each r of its columns is implied by the linear independence of any r columns of $\mathbf{J}$. To show the reverse inequality, we observe that the rows of the lower blocks of $\mathbf{S}_1$ are all, in fact, in the row space of $\mathbf{J}$ as these blocks are simply subsequent powers of $\mathbf{J}$, which implies that $\operatorname{rank}(\mathbf{S}_1) \leq \operatorname{rank}(\mathbf{J}) = r$. Furthermore, it holds that $\operatorname{rank}(\mathbf{S}_2) = n$ due to the full rank of $\mathrm{Id} \in \mathbb{R}^{n \times n}$. The decomposition above shows that $\operatorname{rank}(\mathbf{S}) \leq \min\{r, n\} = r$. On the other hand, Let $\mathbf{E}_1 = \begin{pmatrix}\mathrm{Id} & 0 & 0 & \cdots & 0\end{pmatrix} \in \mathbb{R}^{n \times nd_2}$. Thus, $\mathbf{S}_1 = \mathbf{S}\mathbf{E}_1^\top$ and $\operatorname{rank}(\mathbf{S}\mathbf{E}_1^\top) = r \leq \min\{\operatorname{rank}(\mathbf{S}), \operatorname{rank}(\mathbf{E}_1)\}$, which finally implies $\operatorname{rank}(\mathbf{S}) = r$.

We continue with the proof of the second statement of Theorem II.1. Since the block Hankel matrix $\mathbf{H} \in \mathcal{M}_{d_1 n, d_2 n}$ has a square extension $\overset{\square}{\mathbf{H}} \in \mathcal{M}_{Dn,Dn}$ that is positive semidefinite and of rank r, there exists a block matrix $\mathbf{L} \in \mathcal{M}_{Dn \times r}$ such that

$$\begin{aligned}\overset{\square}{\mathbf{H}} &= \begin{bmatrix}\mathbf{H}_1 & \mathbf{H}_2 & \iddots & \mathbf{H}_D \\ \mathbf{H}_2 & \iddots & \iddots & \iddots \\ \iddots & \iddots & \iddots & \mathbf{H}_{2D-2} \\ \mathbf{H}_D & \iddots & \mathbf{H}_{2D-2} & \mathbf{H}_{2D-1}\end{bmatrix} \\ &= \mathbf{L}\mathbf{L}^* = \begin{bmatrix}\mathbf{L}_1 \\ \mathbf{L}_2 \\ \vdots \\ \mathbf{L}_D\end{bmatrix}\begin{bmatrix}\mathbf{L}_1^* & \mathbf{L}_2^* \dots & \mathbf{L}_D^*\end{bmatrix},\end{aligned}$$

where the columns of $\mathbf{L}$ are linear independent. In the last equality, we wrote $\mathbf{L} = \begin{bmatrix}\mathbf{L}_1^*, & \dots, & \mathbf{L}_D^*\end{bmatrix}^*$ with block row matrix with blocks $\mathbf{L}_j \in \mathcal{M}_{n,r}$ for each $j = 1, \dots, D$. Now we denote the lower submatrix of $\mathbf{L}$ as $\mathbf{L}_L = \begin{bmatrix}\mathbf{L}_1^*, & \dots, & \mathbf{L}_{D-1}^*\end{bmatrix}^*$ and the upper submatrix $\mathbf{L}$ as $\mathbf{L}_U = \begin{bmatrix}\mathbf{L}_2^*, & \dots, & \mathbf{L}_D^*\end{bmatrix}^*$. Due to the Hankel structure of $\overset{\square}{\mathbf{H}}$, the $D-1$ left lower and $D-1$ right upper blocks coincide, and therefore $\mathbf{L}_U\mathbf{L}_L^* = \mathbf{L}_L\mathbf{L}_U^*$. Since by assumption at least one

[8] which means that $\mathbf{N}\Sigma = \Sigma\mathbf{N}^*$.

other matrix block of $\overset{\square}{\mathbf{H}}$, besides the first one, is of rank $r = \operatorname{rank}(\overset{\square}{\mathbf{H}})$, it follows that both $\mathbf{L}_L$ and $\mathbf{L}_U$ have full column rank. Furthermore, since this matrix is symmetric and its columns are both in the span of the columns of $\mathbf{L}_L$ and $\mathbf{L}_U$, then the column spans of $\mathbf{L}_U$ and $\mathbf{L}_L$ coincide. Thus, there exists a unique invertible matrix $\mathbf{M} \in \mathcal{M}_r$ such that

$$\mathbf{L}_U = \mathbf{L}_L \mathbf{M}. \tag{34}$$

Comparing the $D-1$ blocks of the latter matrix, we note that $\mathbf{L}_{j+1} = \mathbf{L}_j \mathbf{M}$ for each $j = 1, \ldots, D-1$, and therefore

$$\mathbf{L}_{j+1} = \mathbf{L}_1 \mathbf{M}^j \tag{35}$$

for each $j = 1, \ldots, D-1$. Inserting this into $\overset{\square}{\mathbf{H}} = \mathbf{L}\mathbf{L}^*$, we observe that $\mathbf{H}_2 = \mathbf{L}_1 \mathbf{M} \mathbf{L}_1^* = \mathbf{L}_1 \mathbf{M}^* \mathbf{L}_1^*$, implying that $\mathbf{M}$ is symmetric. The representation (12) of $\mathbf{H}$ follows with $\mathbf{Y} = \mathbf{L}_1$ from inserting the definition of the right hand side and multiplying the resulting block matrices. □

VIII. Proof Outline for Theorem IV.1

The proof of Theorem IV.1 consists of multiple steps. First, we formulate a local restricted isometry property, Property VIII.1, and show that it holds with high probability for uniform and adaptive sampling if enough samples are provided or if the local sampling probabilities are large enough, respectively (Lemma VIII.1). Using perturbation arguments, we then show with Lemmas VIII.2 and VIII.3 that this regularity also extends to the neighborhood of the ground truth.

In order to have a chance of establishing recovery guarantees, it is necessary to understand when a coodinatewise sampling operator provided related to the sampling set $\Omega \subset I$ is invertible restricted to a subspace associated to low-rank matrices.

Let $\mathcal{P}_{\mathbf{T_Z}} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ (or, in short, $\mathcal{P}_{\mathbf{T}}$) be the orthogonal projection onto $\mathbf{T_Z}$. If $\mathcal{H}$ is the block Hankel operator (9) and $E_{i,j,t}$ is element (i,j,t) of the standard basis of $\mathcal{M}_n^{\oplus T}$, s.t.

$$\langle E_{i,j,t}, \widetilde{\mathbf{X}} \rangle_F = \langle E_{i,j}, \mathbf{X}_t \rangle_F = (\mathbf{X}_t)_{i,j}$$

for any $\widetilde{\mathbf{X}} = \mathbf{X}_1 \oplus \mathbf{X}_2 \oplus \mathbf{X}_3 \oplus \ldots \oplus \mathbf{X}_T \in \mathcal{M}_n^{\oplus T}$, then we define *normalized block Hankel operator* $\mathcal{G} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_{d_1 n, d_2 n}$ as

$$\mathcal{G}(\widetilde{\mathbf{X}}) := \sum_{t=1}^{T} \sum_{i,j=1}^{n} \langle E_{i,j,t}, \widetilde{\mathbf{X}} \rangle_F \frac{\mathcal{H}(E_{i,j,t})}{\|\mathcal{H}(E_{i,j,t})\|_F}. \tag{36}$$

With this definition, we formulate the following property.

Property VIII.1 (Local restricted isometry property)**.** *Let $\mathbf{Z} \in \mathcal{M}_n$ be of rank r, let $\mathbf{T_Z} \subset \mathcal{M}_{d_1 n, d_2 n}$ be the associated tangent space (23) to the manifold of rank-r matrices, and let $\alpha > 0$. Let $\mathcal{R}_\Omega : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ be a self-adjoint, normalized sampling operator relative to a sampling set $\Omega \subset I$. We say that $\mathcal{R}_\Omega$ satisfies the* local restricted isometry property *with respect to $\mathbf{T_Z}$ and constant α if*

$$\|\mathcal{P}_{\mathbf{T_Z}} \mathcal{G} \mathcal{R}_\Omega \mathcal{G}^* \mathcal{P}_{\mathbf{T_Z}} - \mathcal{P}_{\mathbf{T_Z}} \mathcal{G} \mathcal{G}^* \mathcal{P}_{\mathbf{T_Z}}\| \leq \alpha, \tag{37}$$

where $\mathcal{P}_{\mathbf{T_Z}} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ is the orthogonal projection onto the linear subspace $\mathbf{T_Z}$.

Condition (37) is referred as a *local* restricted isometry property since it is not a restricted isometry property with respect to the entire manifold of low-rank matrices [RFP10, DR16], but rather one that holds with respect to a particular (tangent) subspace associated to the low-rank matrix manifold around a point. Similar conditions have been used for structured low-rank matrix completion [CC14, Lemma 1], [YKJL17, Lemma 20].

With Lemma VIII.1, we establish Property VIII.1 with high probability for uniform and adaptive sampling with respect to the block Hankel matrix $\mathbf{H_A}$ associated to a transition operator $\mathbf{A}$.

Lemma VIII.1 (Local RIP for sampling operators)**.** *Let $\mathbf{A} \in \mathcal{M}_n$ be of rank-r, let $\mathbf{T} := \mathbf{T_{H_A}}$ be the tangent space to the rank-r matrix manifold at the block Hankel matrix $\mathbf{H_A} = \mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$ associated to $\mathcal{Q}_T(\mathbf{A})$. Let $0 < \alpha < 1$ and $\mathcal{G} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_{d_1 n, d_2 n}$ be the normalized block Hankel operator of (36). There exists a constant $C > 0$ such that the following holds:*

1) [Uniform sampling model] *Suppose that Ω is a random subset of cardinality m uniformly drawn without replacement among the set of space-time samples $I = [n] \times [n] \times [T]$. Let $\mathcal{R}_\Omega : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ be the normalized sampling operator*

$$\mathbf{L} \to \mathcal{R}_\Omega(\mathbf{L}) := \sum_{(i,j,t) \in \Omega} \frac{n^2 T}{m} \langle E_{i,j,t}, \mathbf{L} \rangle_F E_{i,j,t}. \tag{38}$$

Then $\mathcal{R}_\Omega$ satisfies the local restricted isometry property with respect to $\mathbf{H_A}$ with constant α (Property VIII.1), with probability at least $1 - n^{-2}$, provided that

$$m \geq \frac{C c_s}{\alpha^2} \mu_0 r n \log(nT), \tag{39}$$

if $\mathbf{H_A}$ is μ_0-incoherent as per Definition IV.1.

2) [Adaptive sampling] *Suppose that Ω consists of random index triplets $(i,j,t) \in I$ that are independently observed according to Bernoulli distributions with probabilities $(p_{i,j,t})_{(i,j,t) \in I}$. Let $\mathcal{R}_\Omega : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ be the normalized sampling operator*

$$\mathbf{L} \to \mathcal{R}_\Omega(\mathbf{L}) := \sum_{(i,j,t) \in \Omega} \frac{1}{p_{i,j,t}} \langle E_{i,j,t}, \mathbf{L} \rangle_F E_{i,j,t}. \tag{40}$$

Then $\mathcal{R}_\Omega$ satisfies the local restricted isometry property with respect to $\mathbf{H_A}$ with constant α (Property VIII.1), with probability at least $1 - n^{-2}$, provided that, for each $(i,j,t) \in I$, we have

$$p_{i,j,t} \geq \min\left(\frac{C c_s}{\alpha^2} \mu_{i,j,t} \frac{r}{nT} \log(nT), 1 \right), \tag{41}$$

if $(\mu_{i,j,t})_{(i,j,t) \in I}$ are the local incoherences (24) of $\mathbf{H_A}$ as in Definition IV.1.

We note that the incoherence parameters of the matrix $\mathbf{H_A}$ play an important role in quantifying the number of space-time samples that are sufficient to establish Property VIII.1 for sampling operators. The proof can be found in Appendix D.

In Lemma VIII.2, we extend Property VIII.1 to a neighborhood of $\mathbf{H_A}$.

Lemma VIII.2. *Assume that the local restricted isometry property Property VIII.1 holds true for a normalized sampling operator $\mathcal{R}_\Omega : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ with respect to $\mathbf{H}_\mathbf{A} = \mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$, where $\mathbf{A}$ has a rank-r, and constant $\alpha > 0$. If $\mathbf{H} \in \mathcal{M}_{d_1 n, d_2 n}$ is of rank r and*

$$\mathbf{H} \in \mathcal{B}_{\mathbf{H}_\mathbf{A}}\left(\frac{\alpha}{8}\left(\sqrt{\|\mathcal{R}_\Omega\|\,(1+\alpha)}+1\right)^{-1}\sigma_r(\mathbf{H}_\mathbf{A})\right), \qquad (42)$$

then

$$\|\mathcal{P}_{\mathrm{T}_\mathbf{H}}\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\mathcal{P}_{\mathrm{T}_\mathbf{H}} - \mathcal{P}_{\mathrm{T}_\mathbf{H}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathrm{T}_\mathbf{H}}\| \leq 2\alpha,$$

where $\mathbf{T}_\mathbf{H}$ is the tangent space to the rank-r manifold at $\mathbf{H}$.

The proof is postponed to Appendix E.

As the next step, we establish a null space-type property that shows that not too much mass can be concentrated on the tangent space $\mathbf{T}_\mathbf{H}$ among block Hankel matrices in the null space of the sampling operator.

Lemma VIII.3. *Let $\mathcal{R}_\Omega : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ be a normalized sampling operator as in (38) or (40). If $\mathbf{H} \in \mathcal{M}_{d_1 n, d_2 n}$ is of rank r and $T_\mathbf{H} \subset \mathcal{M}_{d_1 n, d_2 n}$ is the tangent space (23) to the rank-r manifold at $\mathbf{H}$, then*

$$\|\mathcal{P}_{\mathrm{T}_\mathbf{H}}\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\mathcal{P}_{\mathrm{T}_\mathbf{H}} - \mathcal{P}_{\mathrm{T}_\mathbf{H}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathrm{T}_\mathbf{H}}\| \leq \frac{2}{5}, \qquad (43)$$

implies

$$\|\mathcal{H}(\eta)\|_F^2 \leq \frac{5}{3}\left(\|\mathcal{R}_\Omega\| + 8/5\right)\left\|\mathcal{P}_{\mathbf{T}_\mathbf{H}^\perp}\mathcal{H}(\eta)\right\|_F^2$$

for each $\eta \in \ker \mathcal{R}_\Omega$.

We refer to Appendix F for the proof.

Finally, Proposition VIII.1 relates Property VIII.1 with respect to the block Hankel matrix of a transition operator to the local quadratic convergence of `TOIRLS`.

Proposition VIII.1 (Local Convergence with Quadratic Rate)**.** *There exists an absolute constant c_0 such that the following holds. Assume that $\mathbf{A} \in \mathcal{M}_n$ is of rank r, and that Property VIII.1 holds for the normalized sampling operator $\mathcal{R}_\Omega : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ with respect to $\mathbf{H}_\mathbf{A} := \mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$ and constant $\alpha = 1/5$. Let $\widetilde{\mathbf{X}}^{(k)}$ is the k-th iterate of `TOIRLS` Algorithm 1 with inputs: Ω, $\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A}))$, and $\widetilde{r} = r$. If we assume that the smoothing parameter fulfills $\varepsilon_k = \sigma_{r+1}(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}))$ and if $\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) \in \mathcal{B}_{\mathbf{H}_\mathbf{A}}\left(c_0\|\mathcal{R}_\Omega\|^{-3/2}r^{-1}\kappa^{-1}(dn-r)^{-1/2}\sigma_r(\mathbf{H}_\mathbf{A})\right)$, where $\kappa := \sigma_1(\mathbf{H}_\mathbf{A})/\sigma_r(\mathbf{H}_\mathbf{A})$ is the condition number of $\mathbf{H}_\mathbf{A}$, then there exists ν such that for all $\ell \geq 0$*

$$\begin{aligned}&\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k+\ell+1)}) - \mathbf{H}_\mathbf{A}\| \\ &\quad \leq \min(\nu\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k+\ell)}) - \mathbf{H}_\mathbf{A}\|^2, \|\mathcal{H}(\widetilde{\mathbf{X}}^{(k+\ell)}) - \mathbf{H}_\mathbf{A}\|).\end{aligned}$$

In other words, $\mathcal{H}(\widetilde{\mathbf{X}}^{(k+\ell)}) \xrightarrow{\ell\to\infty} \mathbf{H}_\mathbf{A}$ with quadratic convergence rate.

The proof (see Appendix G) crucially relies on estimates from [KMV21] on the action of the weight operator $W_{\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})}$ of Definition III.1 where $\widetilde{\mathbf{X}}^{(k)}$ is an `TOIRLS` iterate, and combines them with Lemma VIII.3.

Putting the results of this section together amounts finally to the proof of Theorem IV.1, which is detailed in Appendix C.

IX. Conclusion & Outlook

In this paper, we developed a framework for the learning of linear transition operators from random sparse observations of space-time samples, and provided a local convergence analysis for the non-convex optimization approach `TOIRLS` for solving that problem, quantifying the number and distribution of samples sufficient for convergence. The current work could be extended in several directions: the presented convergence analysis for `TOIRLS` is inherently local, i.e., requires an iterate that is already close to a low-rank ground truth matrix. The empirical results suggest that a global convergence of `TOIRLS` might be possible, despite being beyond the scope of this paper. Furthermore, not only entrywise, but general linear sampling operators could be considered, cf. (7), as well as applications to a broader family of dynamical systems such as linear time-invariant systems with input terms (11). Finally, it would be of interest to combine the setup considered in this paper with additional prior knowledge on the transition operator $\mathbf{A}$, such as sparsity, which is common for example in the context of graph transition operators.

Appendix

A. Incoherence Estimates

In this subsection, we provide estimates for the incoherence parameters $\mu_{i,j,t}$ and μ_0 of the block Hankel matrix $\mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$ associated with a linear operator $\mathbf{A} \in \mathcal{M}_n$, which have been defined in Definition IV.1. Also for use in other proofs, we state the following result that elucidates the action of the block Hankel operator $\mathcal{H}$.

Lemma A.1. *Recalling the normalized block Hankel operator $\mathcal{G} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_{d_1 n, d_2 n}$ of (36), let $\{E_{i,j,t}\}_{(i,j,t)\in I}$ be the standard basis of $\mathcal{M}_n^{\oplus T}$ and $\{\mathbf{B}_{i,j,t}\}_{(i,j,t)\in I}$ the standard basis of the space of block Hankel matrices $\mathcal{H}(\mathcal{M}_n^{\oplus T})$. Then we have that the diagonal operator $\mathcal{D} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$*

$$\begin{aligned}\mathcal{D}(E_{i,j,t}) :&= \|\mathcal{H}(E_{i,j,t})\|_F E_{i,j,t} \\ &= \sqrt{\min(t, T+1-t, d_1, d_2)}E_{i,j,t},\end{aligned} \qquad (44)$$

for each $(i,j,t) \in I$, satisfies $\mathcal{H} = \mathcal{G}\mathcal{D}$, which is equivalent to

$$\mathbf{B}_{i,j,t} = \mathcal{H}(\mathcal{D}^{-1}E_{i,j,t}) = \mathcal{G}(E_{i,j,t})\,.$$

Furthermore, it holds that $\mathcal{H}^*\mathcal{H} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ satisfies, for all $(i,j,t) \in I$,

$$\mathcal{H}^*\mathcal{H}(E_{i,j,t}) = \min(t, T+1-t, d_1, d_2)E_{i,j,t}.$$

Proof. The first statement follows by combining the definition (44) with (36), and by counting the number of occurrences of each block in (9). The last statement follows from

$$\begin{aligned}\langle E_{i,j,t}, \mathcal{H}^*\mathcal{H}(E_{i,j,t})\rangle &= \left\|\mathcal{H}(E_{i,j,t})\right\|_F^2 = \left\|\mathcal{H}(E_{i,j,t})\right\|_F^2 \\ &= \left\|\mathcal{D}(E_{i,j,t})\right\|_F^2 = \min(t, T+1-t, d_1, d_2)\end{aligned}$$

for all $(i,j,t) \in I$ due to (44). □

With the following lemma, we bound the local incoherence parameter as defined in Definition IV.1 by the incoherence parameter based on the related incoherence parameter in (45) in the spirit of [CC14, (27)] and [CWW19].

Lemma A.2. *Let $\mathbf{Z} \in \mathcal{M}_{d_1 n, d_2 n}$ be of rank r with leading left and right singular vector matrices $\mathbf{U} \in \mathbb{R}^{nd_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{nd_2 \times r}$, respectively, and let $\mathbf{T_Z}$ be the associated tangent space (23). Suppose there exists a positive constant $\mu_0 > 0$ such that*

$$\max_{1\le i\le j\le n, 1\le t\le T} \|\mathbf{U}^*\mathbf{B}_{i,j,t}\|_F \le \sqrt{\mu_0 \frac{r}{nd_1}},$$
$$\max_{1\le i\le j\le n, 1\le t\le T} \|\mathbf{B}_{i,j,t}\mathbf{V}\|_F \le \sqrt{\mu_0 \frac{r}{nd_2}}, \tag{45}$$

where $\{\mathbf{B}_{i,j,t}\}_{(i,j,t)\in I}$ is the standard basis of the space of block Hankel matrices $\mathcal{H}(\mathcal{M}_n^{\oplus T})$.

Then, for each $(i,j,t) \in [n] \times [n] \times [T]$,

$$\|P_{\mathbf{T_Z}}(\mathbf{B}_{i,j,t})\|_F^2 \le \mu_0 c_s \frac{r}{nT},$$

where $c_s = \frac{T(T+1)}{d_1 d_2}$. In particular, $\mathbf{Z}$ is μ_0-incoherent in the sense of Definition IV.1.

Proof. It is well-known [Rec11, Eq. (3)] that the action of the projection operator $P_{\mathbf{T_Z}}$ can be written such that

$$\begin{aligned} P_{\mathbf{T}}(\mathbf{M}) &= \mathbf{UU}^*\mathbf{M} + \mathbf{MVV}^* - \mathbf{UU}^*\mathbf{MVV}^* \\ &= \mathbf{UU}^*\mathbf{M}(\mathbf{I} - \mathbf{VV}^*) + \mathbf{MVV}^* \end{aligned}$$

for any matrix $\mathbf{M}$. Therefore, we estimate that

$$\begin{aligned} \|P_{\mathbf{T}}(\mathbf{M})\|_F^2 &= \|\mathbf{UU}^*\mathbf{M}(\mathbf{I} - \mathbf{VV}^*)\|_F^2 + \|\mathbf{MVV}^*\|_F^2 \\ &\le \|\mathbf{UU}^*\mathbf{M}\|_F^2 \|(\mathbf{I} - \mathbf{VV}^*)\|^2 + \|\mathbf{MVV}^*\|_F^2 \\ &\le \|\mathbf{UU}^*\mathbf{M}\|_F^2 + \|\mathbf{MVV}^*\|_F^2 \\ &\le \|\mathbf{U}^*\mathbf{M}\|_F^2 + \|\mathbf{MV}\|_F^2. \end{aligned} \tag{46}$$

Thus, (45) implies that

$$\|P_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F^2 \le \mu_0 \frac{r}{nd_1} + \mu_0 \frac{r}{nd_2} = \frac{\mu_0 r}{n} \frac{d_1 + d_2}{d_1 d_2} = \mu_0 c_s \frac{r}{nT},$$

setting $\mathbf{M} = \mathbf{B}_{i,j,t}$ for any $(i,j,t) \in I$. □

Before providing the proof of the incoherence estimates for the examples of Section IV-C, we note that it follows from (28) in Theorem IV.1 that for adaptive sampling with probabilities satisfying $p_{i,j,t} \ge C c_s \mu_{i,j,t} \frac{r}{nT} \log(nT)$ for all $(i,j,t) \in I$, where C is the constant from (28).2 and $c_s = \frac{T(T+1)}{d_1 d_2}$, a number of

$$\begin{aligned} m_{\text{exp}} := \mathbb{E}[|\Omega|] &= \sum_{(i,j,t)\in I} p_{i,j,t} \\ &\ge C c_s \sum_{(i,j,t)\in I} \mu_{i,j,t} \frac{r}{nT} \log(nT) \end{aligned} \tag{47}$$

expected samples will enable local convergence of `TOIRLS` in the adaptive model for Ω.

1) Orthogonal Matrices : To understand the incoherences of block Hankel matrices $\mathbf{H_A} = \mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$ associated to orthonormal matrices $\mathbf{A} \in \mathbb{O}^n$, we observe that a compact singular value decomposition of $\mathbf{H_A}$ can be given by $\mathbf{H_A} = \mathbf{U_{H_A}} \Sigma \mathbf{V}_{\mathbf{H_A}}^*$ with

$$\mathbf{U_{H_A}} = \frac{1}{\sqrt{d_1}} \begin{pmatrix} \mathbf{A} \\ \mathbf{A}^2 \\ \vdots \\ \mathbf{A}^{d_1} \end{pmatrix}, \quad \mathbf{V_{H_A}} = \frac{1}{\sqrt{d_2}} \begin{pmatrix} \mathbf{I} \\ \mathbf{A} \\ \cdots \\ \mathbf{A}^{d_2 - 1} \end{pmatrix},$$

$\mathbf{U_{H_A}} \in \mathcal{M}_{nd_1, n}$, $\mathbf{V_{H_A}} \in \mathcal{M}_{nd_2, n}$ and $\Sigma = \sqrt{d_1 d_2}\,\mathrm{Id} \in \mathcal{M}_n$. Using this, we obtain the following proposition.

Proposition A.1. *If the transition operator $\mathbf{A} \in \mathbb{O}^n$ is an orthogonal matrix, then*

1) *$\mathbf{H_A} = \mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$ is μ_0-incoherent with $\mu_0 \le 1$.*
2) *The local incoherences of $\mathbf{H_A}$ satisfy $\sum_{(i,j,t)\in I} \mu_{i,j,t} \le nT^2$.*

As a consequence, for both uniform and adaptive sampling, a sample complexity of order $\Theta(n^2 \log(nT))$ is sufficient to satisfy the assumption of Theorem IV.1.

Proof. First, it is straightforward to verify from a block-wise computation that $\|\mathbf{U}_{\mathbf{H_A}}^* \mathbf{B}_{i,j,t}\|_F = \frac{1}{\sqrt{d_1}}$ for each $(i,j,t) \in I$, since $\|(\mathbf{A}^j)^* \mathbf{M}\|_F = \|\mathbf{M}\|_F$ for each j, for each block $\mathbf{M}$ of $\mathbf{B}_{i,j,t}$, due to the preservation of norms through multiplication with orthogonal matrices. Similarly, $\|\mathbf{B}_{i,j,t}\mathbf{V_{H_A}}\|_F = \frac{1}{\sqrt{d_2}}$ for each $(i,j,t) \in I$.

This implies that (45) is satisfied with μ_0 as $r = \operatorname{rank}(\mathbf{H_A}) = \operatorname{rank}(\mathbf{A}) = n$. In view of Lemma A.2, it follows that $\mathbf{H_A}$ is μ_0-incoherent in the sense of Definition IV.1 with $\mu_0 \le 1$ since

$$\|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F^2 \le c_s \frac{1}{T}$$

for each $(i,j,t) \in I$, where $\mathcal{P}_{\mathbf{T}}$ is the projection operator onto the subspace $\mathbf{T} = \mathbf{T_{H_A}}$. This shows the first statement of Proposition A.1

Estimating the sum of local incoherences $\mu_{i,j,t}$ of $\mathbf{H_A}$, we obtain

$$\begin{aligned} \sum_{(i,j,t)\in I} \mu_{i,j,t} &= \sum_{1\le i\le j\le n, t=1,\dots,T} \frac{T}{c_s} \|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F^2 \\ &= \sum_{1\le i\le j\le n, t=1,\dots,T} \frac{d_1 d_2}{(d_1 + d_2)} \|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F^2 \\ &\le \sum_{\substack{1\le i\le j\le n \\ t=1,\dots,T}} \frac{d_1 d_2}{(d_1 + d_2)} \left(\|\mathbf{U}_{\mathbf{H_A}}^* \mathbf{B}_{i,j,t}\|_F^2 + \|\mathbf{B}_{i,j,t}\mathbf{V_{H_A}}\|_F^2\right) \\ &\le \sum_{1\le i\le j\le n, t=1,\dots,T} \frac{d_1 d_2}{(d_1 + d_2)} \left(\frac{1}{d_1} + \frac{1}{d_2}\right) \\ &= \sum_{1\le i\le j\le n, t=1,\dots,T} 1 = Tn^2, \end{aligned}$$

using (46) in the inequality.

Therefore, in view of (47), a sufficient number of expected space-time samples m_{exp} to enable the local convergence guarantee of Proposition A.1 is $m_{\text{exp}} = \Theta(n^2 \log(nT))$. □

2) Positive Semi-Definite Matrices: We now justify the bounds of Section IV-C for positive semidefinite transition operators $\mathbf{A} \in \mathcal{M}_n$. To this end, we provide a closed formula for a singular value decomposition for the associated block Hankel matrix $\mathbf{H}_\mathbf{A} = \mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$ in Theorem A.1.

Theorem A.1. *Suppose $\mathbf{A} = \sum_{\ell=1}^r \lambda_\ell \mathbf{u}_i \mathbf{u}_i^* = \mathbf{U}\Lambda\mathbf{U}^*$ is a positive semidefinite matrix with r positive eigenvalues $\lambda_1, \ldots, \lambda_r$ and $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$ are the corresponding eigenvectors, so that $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ and $\mathbf{U} = \begin{pmatrix}\mathbf{u}_1 & \ldots & \mathbf{u}_r\end{pmatrix}$. For a vector $\mathbf{u} \in \mathbb{R}^n$, an integer m and a scalar λ, we define*

$$\mathcal{V}_{m,\lambda}(\mathbf{u}) = \begin{pmatrix} \mathbf{u} \\ \lambda\mathbf{u} \\ \vdots \\ \lambda^{m-1}\mathbf{u} \end{pmatrix} \in \mathbb{R}^{nm}$$

and the generalized Vandermonde matrix, *in $\mathcal{M}_{nm,r}$,*

$$\mathcal{V}_m(\mathbf{U},\Lambda) = \begin{pmatrix} | & & | \\ \mathcal{V}_{m,\lambda_1}(\mathbf{u}_1) & \ldots & \mathcal{V}_{m,\lambda_r}(\mathbf{u}_r) \\ | & & | \end{pmatrix} = \begin{pmatrix} \mathbf{U}\Lambda^0 \\ \mathbf{U}\Lambda^1 \\ \vdots \\ \mathbf{U}\Lambda^{m-1} \end{pmatrix}.$$

Let $\overrightarrow{\Lambda^m} := \operatorname{diag}(\overrightarrow{\lambda_\ell^m})_{\ell=1}^r$ with $\overrightarrow{\lambda_\ell^m} = \sqrt{\sum_{s=0}^{m-1} \lambda_\ell^{2s}}$ for all $\ell \in [r]$. Then a compact singular value decomposition of $\mathbf{H}_\mathbf{A}$ can be written such that

$$\mathbf{H}_\mathbf{A} = \mathbf{U}_{\mathbf{H}_\mathbf{A}} \mathbf{D} \mathbf{V}_{\mathbf{H}_\mathbf{A}}^* \tag{48}$$

where

$$\begin{aligned} \mathbf{U}_{\mathbf{H}_\mathbf{A}} &= \mathcal{V}_{d_1}(\mathbf{U},\Lambda)(\overrightarrow{\Lambda^{d_1}})^{-1} \\ \mathbf{V}_{\mathbf{H}_\mathbf{A}} &= \mathcal{V}_{d_1}(\mathbf{U},\Lambda)(\overrightarrow{\Lambda^{d_2}})^{-1} \\ \mathbf{D} &= (\overrightarrow{\Lambda^{d_1}})\Lambda(\overrightarrow{\Lambda^{d_2}}). \end{aligned}$$

In particular, the nonzero singular values of $\mathbf{H}_\mathbf{A}$ are

$$\lambda_\ell \sqrt{\sum_{t=0}^{d_1-1} \lambda_\ell^{2t}} \sqrt{\sum_{t=0}^{d_2-1} \lambda_\ell^{2t}},$$

for $\ell = 1, \ldots, r$, and the first r right and left singular vectors are $\{(\overrightarrow{\lambda_\ell^{d_2}})^{-1}\mathcal{V}_{d_2,\lambda_1}(\mathbf{u}_i)\}_{\ell=1}^r$ and $\{(\overrightarrow{\lambda_\ell^{d_1}})^{-1}\mathcal{V}_{d_1,\lambda_1}(\mathbf{u}_i)\}_{\ell=1}^r$, respectively.

Proof. To prove the statements, the equality (48) can be verified expanding the right hand side, and furthermore, since $\mathbf{U}^*\mathbf{U} = \mathrm{Id}_r$, the orthogonality of the columns of the singular vector matrices can be verified, i.e. $\mathbf{U}_{\mathbf{H}_\mathbf{A}}^* \mathbf{U}_{\mathbf{H}_\mathbf{A}} = \mathrm{Id}_r$ and $\mathbf{V}_{\mathbf{H}_\mathbf{A}}^* \mathbf{V}_{\mathbf{H}_\mathbf{A}} = \mathrm{Id}_r$. □

Following the notation in (48), we compute estimates of $\|\mathbf{U}_{\mathbf{H}_\mathbf{A}}(\mathbf{B}_{i,j,t})\|_F$ and $\|(\mathbf{B}_{i,j,t})\mathbf{V}_{\mathbf{H}_\mathbf{A}}\|_F$. As a preparation of what follows, we recall from (44) that

$$\|\mathcal{H}(E_{i,j,t})\|_F = \begin{cases} \sqrt{t} & \text{if } t \le \min\{d_1,d_2\}, \\ \sqrt{\min\{d_1,d_2\}} & \text{if } \begin{array}{l} \min\{d_1,d_2\} < t \\ \le \max\{d_1,d_2\} \end{array}, \\ \sqrt{T+1-t} & \text{if } t > \max\{d_1,d_2\}, \end{cases}$$

if $E_{i,j,t}$ is the standard basis matrix of index $(i,j,t) \in I$ of $\mathcal{M}_n^{\oplus T}$. The above observation yields the following lemma.

Lemma A.3. *Let $\mathbf{B}_{i,j,t} = \mathcal{H}(E_{i,j,t})/\|\mathcal{H}(E_{i,j,t})\|_F \in \mathcal{M}_{d_1 n, d_2 n}$ be the standard basis matrix of index $(i,j,t) \in I$ of $\mathcal{H}(\mathcal{M}_n^{\oplus T})$. For $2 \le d_1 \le d_2$, we have the following identities if $\mathbf{U}_{\mathbf{H}_\mathbf{A}}$ and $\mathbf{V}_{\mathbf{H}_\mathbf{A}}$ are as in Theorem A.1:*

$$\|\mathbf{U}_{\mathbf{H}_\mathbf{A}}^* \mathbf{B}_{i,j,t}\|_F =$$

$$= \begin{cases} \sqrt{\sum_{\ell=1}^r \frac{\sum_{s=0}^{t-1} \lambda_\ell^{2s}}{\sum_{s=0}^{d_1-1} \lambda_\ell^{2s}} \frac{\|\mathbf{u}_\ell^* E_{ij}\|^2}{t}}, & \text{if } t < d_1, \\ \sqrt{\sum_{\ell=1}^r \frac{\|\mathbf{u}_\ell^* E_{ij}\|^2}{d_1}}, & \text{if } d_1 \le t \le d_2, \\ \sqrt{\sum_{\ell=1}^r \frac{\sum_{s=t-d_2}^{d_1-1} \lambda_\ell^{2s}}{\sum_{s=0}^{d_1-1} \lambda_\ell^{2s}} \frac{\|\mathbf{u}_\ell^* E_{ij}\|^2}{T+1-t}}, & \text{if } t > d_2, \end{cases}$$

and

$$\|\mathbf{B}_{i,j,t} \mathbf{V}_{\mathbf{H}_\mathbf{A}}\|_F =$$

$$= \begin{cases} \sqrt{\sum_{\ell=1}^r \frac{\sum_{s=0}^{t-1} \lambda_\ell^{2s}}{\sum_{s=0}^{d_2-1} \lambda_\ell^{2s}} \frac{\|E_{ij}\mathbf{u}_\ell\|^2}{t}}, & \text{if } t < d_1, \\ \sqrt{\sum_{\ell=1}^r \frac{\sum_{s=t-d_1}^{t-1} \lambda_\ell^{2s}}{\sum_{s=0}^{d_2-1} \lambda_\ell^{2s}} \frac{\|E_{ij}\mathbf{u}_\ell\|^2}{d_1}}, & \text{if } d_1 \le t \le d_2, \\ \sqrt{\sum_{\ell=1}^r \frac{\sum_{s=t-d_1}^{d_2-1} \lambda_\ell^{2s}}{\sum_{s=0}^{d_2-1} \lambda_\ell^{2s}} \frac{\|E_{ij}\mathbf{u}_\ell\|^2}{T+1-t}}, & \text{if } t > d_2. \end{cases}$$

We restrict our attention to *normalized* positive semidefinite transition operators $\mathbf{A}$ whose eigenvalues are within the interval $[0,1]$. To simplify the analysis and avoid unnecessary technicalities, we restrict ourselves to the case of $d_1 = d_2$.

Proposition A.2. *Let $\mathbf{A}$ be a positive semidefinite transition operator as in A.1 and assume that $0 < \lambda_r \le \ldots \le \lambda_1 \le 1$. Assume also that $d_1 = d_2$ and $T = d_1 + d_2 - 1 \ge 3$. Then:*

1) *$\mathbf{H}_\mathbf{A} = \mathcal{H}(\mathcal{Q}_T(\mathbf{A}))$ is μ_0-incoherent with*

$$\mu_0 \le \max_{1 \le i \le n} \sum_{\ell=1}^r \frac{n d_2 (\mathbf{u}_\ell)_i^2}{r(1 + \lambda_\ell^2 + \cdots + \lambda_\ell^{2(d_2-1)})}.$$

2) *The local incoherences of $\mathbf{H}_\mathbf{A}$ satisfy*

$$\sum_{(i,j,t) \in I} \mu_{i,j,t}$$

$$\le \frac{n^2 d_2}{r} \left(\sum_{t=1}^{d_2-1} \sum_{\ell=1}^r \frac{(1 + \lambda_\ell^{2(d_2-t)}) \sum_{s=0}^{t-1} \lambda_\ell^{2s}}{t \sum_{s=0}^{d_2-1} \lambda_\ell^{2s}} + \frac{1}{d_2} \right) \tag{49}$$

$$\le 4.4 n^2 T \log(T). \tag{50}$$

Consequently, for the adaptive sampling model, it is possible to satisfy the assumption of Theorem IV.1 with

$$m_{exp} = \mathbb{E}[|\Omega|] = \Theta(rn \log(nT) \log(T))$$

expected samples.

Proof. 1. Define the function $g : [0,1] \times \mathbb{N} \times \mathbb{N} \mapsto \mathbb{R}$ such that

$$g(\lambda, d, t) = \frac{\sum_{s=0}^{t-1} \lambda^s}{t \sum_{s=0}^{d-1} \lambda^s}. \tag{51}$$

We observe that g has following properties:

- For a fixed $d \in \mathbb{N}$ and $\lambda \in [0,1]$, $g(\lambda, d, t)$ is a decreasing function with respect to $t \in \mathbb{N}$.
- For fixed $d \in \mathbb{N}$ and $t \in \mathbb{N}$, $g(\lambda, d, t)$ is a decreasing function with respect to λ on $[0, \infty)$.

Using Lemma A.3 and the properties of the function g, we obtain that

$$\max_{1\leq i\leq j\leq n, 1\leq t\leq T} \|\mathbf{U}_{\mathbf{H_A}}^*(\mathbf{B}_{i,j,t})\|_F = \max_{1\leq i\leq j\leq n} \|\mathbf{U}_{\mathbf{H_A}}^*(\mathbf{B}_{i,j,1})\|_F = \max_{1\leq i\leq n} \sqrt{\sum_{\ell=1}^{r} \frac{(\mathbf{u}_\ell)_i^2}{\sum_{s=0}^{d_1-1}\lambda_\ell^{2s}}}.$$

By symmetry, we have

$$\max_{1\leq i\leq j\leq n, 1\leq t\leq T} \|\mathbf{B}_{i,j,t}\mathbf{V}_{\mathbf{H_A}}\|_F = \max_{1\leq i\leq n} \sqrt{\sum_{\ell=1}^{r} \frac{(\mathbf{u}_\ell)_i^2}{\sum_{s=0}^{d_2-1}\lambda_\ell^{2s}}}.$$

From this and from the fact that $d_1 \leq d_2$, we see that we can choose

$$\mu_0 = \max_{1\leq i\leq n} \sum_{\ell=1}^{r} \frac{nd_2(\mathbf{u}_\ell)_i^2}{r(\sum_{s=0}^{d_1-1}\lambda_\ell^{2s})}$$

to satisfy the inequalities of (45), from which it follows that $\mathbf{H_A}$ is μ_0-incoherent by Lemma A.2.

2. To obtain an upper bound for $\sum_{i,j,t}\mu_{i,j,t}$, we use that

$$\sum_{1\leq i,j\leq n} \left(\|\mathbf{U}_{\mathbf{H_A}}^*\mathbf{B}_{i,j,t}\|_F^2 + \|\mathbf{B}_{i,j,t}\mathbf{V}_{\mathbf{H_A}}\|_F^2\right) = \begin{cases} \sum_{\ell=1}^{r} 2g(\lambda_\ell^2, d_2, t)n, & \text{for } t < d_2, \\ \frac{2n}{d_2}, & \text{for } t = d_2, \\ \sum_{\ell=1}^{r} 2\lambda_\ell^{2(t-d_2)} g(\lambda_\ell^2, d_2, T-t+1)n, & \text{for } t > d_2. \end{cases}$$

Therefore

$$\sum_{1\leq i,j\leq n}\sum_{t=1}^{T} \left(\|\mathbf{U}_{\mathbf{H_A}}^*(\mathbf{B}_{i,j,t})\|_F^2 + \|\mathbf{B}_{i,j,t}\mathbf{V}_{\mathbf{H_A}}\|_F^2\right) = \sum_{t=1}^{d_2-1}\sum_{\ell=1}^{r} 2(1+\lambda_\ell^{2(d_2-t)})g(\lambda_\ell^2, d_2, t)n + \frac{2n}{d_2} = 2G(\Lambda, T)n,$$

where $G(\Lambda, T) := \sum_{t=1}^{d_2-1}\sum_{\ell=1}^{r}(1+\lambda_\ell^{2(d_2-t)})g(\lambda_\ell^2, d_2, t) + \frac{1}{d_2}$ is a constant that only depends on the eigenvalues $(\lambda_\ell)_\ell$, d_2 and T. Finally, it follows then from (46) that

$$\sum_{(i,j,t)\in I} \mu_{i,j,t} \leq \sum_{(i,j,t)\in I} \frac{nd_2^2}{2rd_2}\left(\|\mathbf{U}_{\mathbf{H_A}}^*\mathbf{B}_{i,j,t}\|_F^2 + \|\mathbf{B}_{i,j,t}\mathbf{V}_{\mathbf{H_A}}\|_F^2\right) \leq \frac{n^2 d_2}{r} G(\Lambda, T),$$

which amounts to the first desired bound (49). Moreover, using the properties of g described previously, one can show that

$$g(\lambda_\ell^2, d_2, t) \leq g(0, d_2, t) = \frac{1}{t}.$$

Since furthermore $\sum_{t=1}^{d_2-1}\frac{1}{t} < c_\gamma + \log(d_2)$, where $c_\gamma < 0.58$ is the Euler-Mascheroni constant, we have

$$G(\Lambda, L) \leq \sum_{t=1}^{d_2-1}\sum_{\ell=1}^{r} \frac{1+\lambda_\ell^{2(d_2-t)}}{t} + \frac{1}{d_2} \leq 2r(c_\gamma + \log(d_2)) + \frac{1}{d_2} \leq (1.16 + 2\log(d_2) + 1/(rd_2))r \leq 4.4r\log(T),$$

using that $\lambda_\ell \leq 1$ for all $\ell \in [r]$ in the second inequality and $2 \leq d_2 \leq T$ in the last inequality. This yields the second desired bound (50) and concludes the proof.

Therefore, analgously as to the argument in the proof of Proposition A.1, it follows that a sufficient number of expected space-time samples m_{exp} to enable the local convergence guarantee of Proposition A.1 is $m_{\text{exp}} = \Theta(rn\log(nT)\log(T))$. □

We conclude this section by noting that if $\mathbf{A}$ is a rank-r projection, this amounts to a positive semidefinite transition operator with $\lambda_1 = \lambda_2 = \ldots = \lambda_r = 1$. In this case, the function g of (51) can be simplified to $g(1, d, t) = \frac{1}{d}$, which simplifies the expression for $G(\Lambda, L)$ to $G(\Lambda, L) = (2d_2 - 1)/d_2 = 2 - \frac{1}{d_2}$. This means that in fact, for rank-r projection matrices, $m_{\text{exp}} = \Theta(rn\log(nT))$ expected samples in the adaptive regime are sufficient.

B. Proofs

In the next sections, we provide the proofs of the main local convergence result for `TOIRLS`, Theorem IV.1, as well as the proofs of Lemmas VIII.1 to VIII.3 and proposition VIII.1 which are auxiliary results for proving Theorem IV.1.

C. Proof of Theorem IV.1

In this section, we provide the proof of Theorem IV.1, which is based on combining Proposition VIII.1 and Lemma VIII.1. As an additional ingredient, we bound the spectral norm $\|\mathcal{R}_\Omega\|$ of the normalized sampling operators $\mathcal{R}_\Omega$ of (38) and (40).

Lemma A.4. *Let Ω be a random subset of the index set $I = [n] \times [n] \times [T]$ of size m that is sampled uniformly i.i.d. with replacement, where $m < n^2T$. Let $\beta > 1$. Then with probability at least $1 - (n^2T)^{1-\beta}$, the maximal number of repetitions of any entry in Ω is less than $\frac{8}{3}\beta\log(nT)$ for $n\sqrt{T} \geq 9$ and $\beta > 1$.*

Consequently, we have that with probability of at least $1 - (n^2T)^{1-\beta}$, the operator $\mathcal{R}_\Omega : \mathcal{M}_{d_1,d_2} \to \mathcal{M}_{d_1,d_2}$ of (38) fulfills

$$\|\mathcal{R}_\Omega\| \leq \frac{8}{3}\beta\frac{n^2T}{m}\log(nT),$$

where $\|\mathcal{R}_\Omega\|$ is the spectral norm of $\mathcal{R}_\Omega$.

The proof of Lemma A.4 is a simple adaptation of [Rec11, Proposition 5]. We proceed to the proof of our main result, Theorem IV.1.

Proof of Theorem IV.1.1. By choosing $\beta = 2$ in Lemma A.4, it follows that with probability of at least $1 - (n^2T)^{-1}$,

$$\|\mathcal{R}_\Omega\|_2 \leq \frac{16}{3}\frac{n^2T}{m}\log(nT) \tag{52}$$

Recall that $d = \min(d_1, d_2)$ was chosen to be the minimum of the pencil parameters d_1 and d_2, which satisfy $d_1 + d_2 - 1 = T$. Let c_0 be the constant of Proposition VIII.1 and C the constant of Lemma VIII.1.1.

Fix now $\alpha = 1/5$. From the statement of Lemma VIII.1.1, if follows that if

$$m \geq 25Cc_s\mu_0 rn\log(nT), \tag{53}$$

with a probability at least $1 - n^{-2}$ the normalized sampling operator $\mathcal{R}_\Omega : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ of (38) satisfies

$$\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| \leq \frac{1}{5},$$

i.e., Property VIII.1 is satisfied with respect to $\mathbf{T} = \mathbf{T}_{\mathbf{H_A}}$ and constant $\alpha = 1/5$.

Let now $\widetilde{\mathbf{X}}^{(k)}$ be such that $\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})$ satisfies assumption (27) of Theorem IV.1.1. It follows from Proposition VIII.1 and (52) that on an event E of probability of at least $1-(n^2T)^{-1}-n^{-2} \geq 1 - 2n^{-2}$, if c_0 is the constant of Proposition VIII.1, C the constant of (39) and $\widetilde{c}_0 := c_0(75Cc_s/16)^{3/2}$, it holds that

$$\begin{aligned}
&\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) - \mathbf{H_A}\| \\
&\leq \widetilde{c}_0\mu_0^{3/2}(nT)^{-3/2}r^{1/2}\kappa^{-1}(dn-r)^{-1/2}\sigma_r(\mathbf{H_A}) \\
&= c_0\frac{(25Cc_s)^{3/2}\mu_0^{3/2}r^{1/2}}{(16/3)^{3/2}n^{3/2}T^{3/2}\kappa(dn-r)^{1/2}}\sigma_r(\mathbf{H_A}) \\
&= c_0\frac{(25Cc_s)^{3/2}\mu_0^{3/2}r^{3/2}n^{3/2}\log^{3/2}(nT)\sigma_r(\mathbf{H_A})}{(16/3)^{3/2}n^3T^{3/2}\log^{3/2}(nT)r\kappa(dn-r)^{1/2}} \\
&\leq c_0\frac{m^{3/2}\sigma_r(\mathbf{H_A})}{(16/3)^{3/2}n^3T^{3/2}\log^{3/2}(nT)r\kappa(dn-r)^{1/2}} \\
&\leq c_0\frac{\sigma_r(\mathbf{H_A})}{\|\mathcal{R}_\Omega\|^{3/2}r\kappa(dn-r)^{1/2}},
\end{aligned}$$

using also (53) in the second inequality.

Therefore, the conclusion of Proposition VIII.1 holds with constant (see the proof of Proposition VIII.1 in Appendix G)

$$\begin{aligned}
\nu &= \frac{20}{3\sigma_r(\mathbf{H_A})}(1+6\kappa)\left(\|\mathcal{R}_\Omega\| + 8/5\right)r \\
&\leq \frac{20}{3\sigma_r(\mathbf{H_A})}(1+6\kappa)\left(\frac{16}{3}\frac{n^2T}{m}\log(nT) + 8/5\right)r,
\end{aligned}$$

which means that $\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k+1)}) - \mathbf{H_A}\| \leq \min(\nu\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) - \mathbf{H_A}\|^2, \|\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})-\mathbf{H_A}\|)$ and furthermore, $\mathcal{H}(\widetilde{\mathbf{X}}^{(k+\ell)}) \xrightarrow{\ell\to\infty} \mathbf{H_A}$, on the event E from above.

This finishes the proof of Theorem IV.1.1. □

Proof of Theorem IV.1.2. Let $C > 0$ be the constant of (28). To show Theorem IV.1 in the case of adaptive sampling, we recall the definition

$$\mathbf{L} \to \mathcal{R}_\Omega(\mathbf{L}) = \sum_{(i,j,t)\in\Omega}\frac{1}{p_{i,j,t}}\langle E_{i,j,t}, \mathbf{L}\rangle_F E_{i,j,t}.$$

of the normalized sampling operator $\mathcal{R}_\Omega : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ in this case, cf. (40).

Fix $\alpha = 1/5$. It follows from the definition of $\mathcal{R}_\Omega$ and the Bernoulli sampling model that

$$\|\mathcal{R}_\Omega\| \leq \min_{(i,j,t)\in I}\frac{1}{p_{i,j,t}} \leq \frac{nT}{25Cc_sr\log(nT)\min_{(i,j,t)\in I}\mu_{i,j,t}}, \tag{54}$$

using assumption (28) in the last inequality. Under the same assumption, it follows from Lemma VIII.1.2 that with probability at least $1 - n^{-2}$, the local isometry property on $\mathbf{T} = \mathbf{T}_{\mathbf{H_A}}$ with constant $1/5$ holds, i.e.,

$$\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| \leq \frac{1}{5},$$

which entails that Property VIII.1 is satisfied for $\alpha = 1/5$. As above, it follows from Proposition VIII.1 and (54) that on an event of probability at least $1 - n^{-2} \geq 1 - 2n^{-2}$, if c_0 is the constant of Proposition VIII.1, C the constant of (28) and $\widetilde{c}_0 = c_0(25Cc_s)^{3/2}$,

$$\begin{aligned}
&\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) - \mathbf{H_A}\| \\
&\leq \widetilde{c}_0\frac{\min_{(i,j,t)\in I}\mu_{i,j,t}^{3/2}r^{1/2}\log^{3/2}(nT)}{(nT)^{3/2}\kappa(dn-r)^{1/2}}\sigma_r(\mathbf{H_A}) \\
&\leq c_0\frac{(25Cc_s)^{3/2}r^{1/2}\log^{3/2}(nT)\min_{(i,j,t)\in I}\mu_{i,j,t}^{3/2}}{(nT)^{3/2}\kappa(dn-r)^{1/2}}\sigma_r(\mathbf{H_A}) \\
&\leq c_0\frac{1}{\|\mathcal{R}_\Omega\|^{3/2}r\kappa(dn-r)^{1/2}}\sigma_r(\mathbf{H_A}),
\end{aligned}$$

and therefore, the conclusion of Proposition VIII.1 holds with constant

$$\begin{aligned}
\nu &= \frac{20}{3\sigma_r(\mathbf{H_A})}(1+6\kappa)\left(\|\mathcal{R}_\Omega\| + 8/5\right)r \\
&\leq \frac{20}{3\sigma_r(\mathbf{H_A})}(1+6\kappa)\left(\frac{nT(\min_{(i,j,t)\in I}\mu_{i,j,t})^{-1}}{25Cc_sr\log(nT)} + 8/5\right)r,
\end{aligned}$$

which means that $\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k+1)}) - \mathbf{H_A}\| \leq \min(\nu\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) - \mathbf{H_A}\|^2, \|\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})-\mathbf{H_A}\|)$ and furthermore, $\mathcal{H}(\widetilde{\mathbf{X}}^{(k+\ell)}) \xrightarrow{\ell\to\infty} \mathbf{H_A}$, and therefore concludes the proof of Theorem IV.1. □

D. Proof of Lemma VIII.1

In this section, we prove Lemma VIII.1, our main result about the regularity of the normalized sampling operators $\mathcal{R}_\Omega : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ for the uniform and adaptive sampling models, see (38) and (40), respectively. The proof uses a non-commutative Bernstein inequality:

Lemma A.5 (Noncommutative Bernstein inequality, cf. [Rec11, Theorem 4] or [Ver18, Theorem 5.4.1])**.** *Let $\mathcal{Z}_1, \ldots, \mathcal{Z}_m$ be independent, Hermitian zero-mean random operators of dimension $n^2d_1d_2 \times n^2d_1d_2$. Suppose that $\rho^2 = \|\mathbb{E}\sum_{\ell=1}\mathcal{Z}_\ell\mathcal{Z}_\ell\|$ and $\|\mathcal{Z}_\ell\| \leq M$ almost surely for all $\ell \in [m]$. Then for any $\alpha > 0$,*

$$\mathbb{P}\left(\left\|\sum_{\ell=1}^m \mathcal{Z}_\ell\right\|\right) \leq 2n^2d_1d_2\exp\left(\frac{-\alpha^2/2}{\rho^2 + M\alpha/3}\right).$$

The proof of Lemma VIII.1.1 follows the proof idea of [CC14, Lemma 3] and [LLJY18, Lemma 23].

Proof of Lemma VIII.1.1. If $\{E_{i,j,t}\}_{(i,j,t)\in I}$ is the standard basis of $\mathcal{M}_n^{\oplus T}$ and $\{\mathbf{B}_{i,j,t}\}_{(i,j,t)\in I}$ the standard basis of the space of block Hankel matrices $\mathcal{H}(\mathcal{M}_n^{\oplus T})$, we recall from Lemma A.1 that $\mathcal{G}(E_{i,j,t}) = \mathbf{B}_{i,j,t}$ for each $(i,j,t) \in I$.

We first assume a slightly different sampling model than that considered in the statement of Lemma VIII.1: let $\Omega = \{(i_\ell, j_\ell, t_\ell)\}_{\ell=1}^m \subset I$ be a set of m indices sampled uniformly i.i.d. *with* replacement. For $\ell \in [m]$, define the operators $\mathcal{Z}_\ell$ and $\widetilde{\mathcal{Z}}_\ell$ such that

$$\begin{aligned}
\mathcal{Z}_\ell &:= \frac{n^2T}{m}\widetilde{\mathcal{Z}}_\ell - \frac{1}{m}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}} \\
&:= \frac{n^2T}{m}\mathcal{P}_{\mathbf{T}}\mathcal{G}E_{i_\ell,j_\ell,t_\ell}E_{i_\ell,j_\ell,t_\ell}^*\mathcal{G}^*\mathcal{P}_{\mathbf{T}} - \frac{1}{m}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}.
\end{aligned}$$

Then the expectation $\mathbb{E}[\widetilde{\mathcal{Z}}_\ell]$ of $\widetilde{\mathcal{Z}}_\ell$ satisfies

$$\begin{aligned}\mathbb{E}[\widetilde{\mathcal{Z}}_\ell] &= \mathbb{E}\left[\mathcal{P}_{\mathbf{T}}\mathcal{G}E_{i_\ell,j_\ell,t_\ell}E^*_{i_\ell,j_\ell,t_\ell}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\right]\\ &= \frac{1}{n^2T}\sum_{i,j=1,i\leq j}^{n}\sum_{t=1}^{T}\mathcal{P}_{\mathbf{T}}\mathcal{G}E_{i,j,t}E^*_{i,j,t}\mathcal{G}^*\mathcal{P}_{\mathbf{T}} \\ &= \frac{1}{n^2T}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\end{aligned} \tag{55}$$

and furthermore,

$$\mathbb{E}[\mathcal{Z}_\ell] = \frac{n^2T}{m}\mathbb{E}[\widetilde{\mathcal{Z}}_\ell] - \frac{1}{m}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}} = 0.$$

Since for any $\mathbf{M} \in \mathbb{R}^{d_1 n \times d_2 n}$,

$$\widetilde{\mathcal{Z}}_\ell(\mathbf{M}) = \langle\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell}),\mathbf{M}\rangle_F\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell}),$$

we obtain

$$\begin{aligned}\|\widetilde{\mathcal{Z}}_\ell(\mathbf{M})\|_F &\leq |\langle\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell}),\mathbf{M}\rangle_F|\,\|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell})\|_F\\ &\leq \|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell})\|_F^2\|\mathbf{M}\|_F\end{aligned}$$

by Cauchy-Schwarz, and thus obtain

$$\begin{aligned}\left\|\widetilde{\mathcal{Z}}_\ell\right\| &\leq \|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell})\|_F^2\\ &\leq \max_{1\leq i\leq j\in[n],t\in[T]}\|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F^2 \leq \frac{\mu_0 c_s r}{nT},\end{aligned} \tag{56}$$

using the incoherence assumption on $\mathbf{H}_{\mathbf{A}}$ in the last inequality, as well as $d_1+d_2-1=T$ and the definition of $c_s = T(T+1)/(d_1d_2)$. Analogously, we estimate that

$$\begin{aligned}\left\|\frac{1}{m}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\right\| &\leq \frac{1}{m}\sum_{i,j=1,i\leq j}^{n}\left\|\sum_{t=1}^{T}\mathcal{P}_{\mathbf{T}}\mathcal{G}E_{i,j,t}E^*_{i,j,t}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\right\|\\ &\leq \frac{n^2T}{m}\frac{\mu_0c_sr}{nT} = \frac{\mu_0c_srn}{m}.\end{aligned} \tag{57}$$

We observe that if $\mathcal{A}$ and $\mathcal{B}$ are positive semidefinite operators, it holds that $\|\mathcal{A}-\mathcal{B}\| \leq \max(\|\mathcal{A}\|,\|\mathcal{B}\|)$. Therefore, it follows from (56) and (57) that

$$\|\mathcal{Z}_\ell\| \leq \max\left(\frac{n^2T}{m}\frac{\mu_0c_sr}{nT},\frac{\mu_0c_srn}{m}\right) = \frac{\mu_0c_srn}{m} \tag{58}$$

almost surely for all $\ell \in [m]$, as the operators involved are positive semidefinite. Further we compute that

$$\begin{aligned}&\mathbb{E}\left[\mathcal{Z}_\ell\mathcal{Z}_\ell\right]\\ &= \frac{(n^2T)^2}{m^2}\mathbb{E}\left[\widetilde{\mathcal{Z}}_\ell\widetilde{\mathcal{Z}}_\ell\right] - \frac{n^2T}{m^2}\mathbb{E}\left[\widetilde{\mathcal{Z}}_\ell\right]\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\\ &- \frac{n^2T}{m^2}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\mathbb{E}\left[\widetilde{\mathcal{Z}}_\ell\right] + \frac{1}{m^2}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\\ &= \frac{(n^2T)^2}{m^2}\mathbb{E}\left[\widetilde{\mathcal{Z}}_\ell\widetilde{\mathcal{Z}}_\ell\right] - \frac{1}{m^2}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}},\end{aligned}$$

using that $\mathbb{E}\left[\widetilde{\mathcal{Z}}_\ell\right] = \frac{1}{n^2T}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}$, cf. (55). In order to estimate the latter terms, we observe that for any $\mathbf{M} \in \mathbb{R}^{d_1n\times d_2n}$,

$$\begin{aligned}\widetilde{\mathcal{Z}}_\ell\widetilde{\mathcal{Z}}_\ell(\mathbf{M}) &= \langle\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell}),\widetilde{\mathcal{Z}}_\ell(\mathbf{M})\rangle\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell})\\ &= \|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell})\|_F^2\,\langle\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell}),\mathbf{M}\rangle\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell})\\ &= \|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell})\|_F^2\,\widetilde{\mathcal{Z}}_l(\mathbf{M})\end{aligned}$$

and therefore

$$\begin{aligned}\left\|\mathbb{E}\widetilde{\mathcal{Z}}_\ell\widetilde{\mathcal{Z}}_\ell\right\| &= \max_{\|M\|_F=1}\left\|\mathbb{E}\,\|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i_\ell,j_\ell,t_\ell})\|_F^2\,\widetilde{\mathcal{Z}}_\ell(\mathbf{M})\right\|_F\\ &\leq \max_{i\leq j\in[n]t\in[T]}\|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F^2\left\|\mathbb{E}\widetilde{\mathcal{Z}}_\ell\right\|\\ &\leq \frac{\mu_0c_sr}{nT}\frac{1}{n^2T}\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| \leq \frac{\mu_0c_sr}{n^3T^2},\end{aligned} \tag{59}$$

using (56) in the second inequality and the fact $\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| \leq 1$ in the third in equality. For the expectation of the squares of $\mathcal{Z}_\ell$, we obtain

$$\begin{aligned}&\sum_{\ell=1}^{m}\|\mathbb{E}\mathcal{Z}_\ell\mathcal{Z}_\ell\|\\ &\leq \frac{1}{m^2}\sum_{\ell=1}^{m}\max\Big(n^2T^2\left\|\mathbb{E}\widetilde{\mathcal{Z}}_\ell\widetilde{\mathcal{Z}}_\ell\right\|,\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\|\Big)\\ &\leq \sum_{\ell=1}^{m}\max\left(\frac{(n^2T)^2}{(m)^2}\frac{\mu_0c_sr}{n^3T^2},\frac{1}{m^2}\right)\\ &= \sum_{\ell=1}^{m}\max\left(\frac{n}{m^2}\frac{\mu_0c_sr}{1},\frac{1}{m^2}\right)\\ &\leq \frac{\mu_0c_srn}{m},\end{aligned} \tag{60}$$

using again that $\|\mathcal{A}-\mathcal{B}\| \leq \max(\|\mathcal{A}\|,\|\mathcal{B}\|)$ for positive semidefinite operators in the second inequality, (59) and the fact that $\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| \leq 1$ in the third inequality.

Next, recalling the definition (38) of the normalized sampling operator $\mathcal{R}_\Omega : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ of the statement of Lemma VIII.1, we observe that $\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\mathcal{P}_{\mathbf{T}} = \frac{n^2T}{m}\sum_{\ell=1}^{m}\widetilde{\mathcal{Z}}_\ell$.

Since the $\mathcal{Z}_\ell$'s are Hermitian, we can now apply the matrix Bernstein inequality [Rec11, Theorem 4] in form of Lemma A.5 above to obtain, for $0 < \alpha < 1$, the estimate

$$\begin{aligned}&\mathbb{P}\left(\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| \geq \alpha\right)\\ &\leq 2n^2d_1d_2\exp\left(-\frac{m\alpha^2}{2\mu_0c_srn(1+\alpha/3)}\right)\\ &\leq 2\frac{(T+1)^2}{4}n^2\exp\left(-\frac{m\alpha^2}{2\mu_0c_srn(4/3)}\right)\\ &= \frac{(T+1)^2n^2}{2}\exp\left(-\frac{3m\alpha^2}{8\mu_0c_srn}\right),\end{aligned}$$

using the norm estimates of (58) and (60) to estimate the respective quantities in Lemma A.5. From this, we see that

$$\mathbb{P}\left(\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| \geq \alpha\right) \leq n^{-2}$$

if $\frac{1}{2}(T+1)^2n^4 \leq \exp\left(\frac{3m\alpha^2}{8\mu_0c_srn}\right)$, which is further implied by the condition

$$m \geq \frac{16c_s}{3\alpha^2}\mu_0rn\log(nT).$$

This shows that for the constant $C := \frac{16}{3}$, if (39) is fulfilled, then with probability at least $1-n^{-2}$,

$$\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| < \alpha$$

if m i.i.d. samples are uniformly sampled with replacement. With the argument of [Rec11, Proposition 3], we conclude that the statement with the same probability bound holds true for the sampling model where Ω is a random subset of cardinality m, uniformly drawn without replacement, if m satisfies (39), which finishes the proof.

□

Proof of Lemma VIII.1.2. To show the second part of Lemma VIII.1, we consider for each $(i,j,t) \in I$ a random variable $\delta_{i,j,t}$ that is 1 if $(i,j,t) \in \Omega$ and 0 otherwise. With that notation, $\mathcal{R}_\Omega$ of (38) can be written as

$$\begin{aligned}\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\mathcal{P}_{\mathbf{T}} &= \sum_{(i,j,t)\in I} \frac{\delta_{i,j,t}}{p_{i,j,t}} \mathcal{P}_{\mathbf{T}}\mathcal{G}E_{i,j,t}E^*_{i,j,t}\mathcal{G}^*\mathcal{P}_{\mathbf{T}} \\ &=: \sum_{(i,j,t)\in I} \frac{\delta_{i,j,t}}{p_{i,j,t}} \widetilde{\mathcal{Z}}_{i,j,t},\end{aligned}$$

defining operators $\widetilde{\mathcal{Z}}_{i,j,t} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ for each $(i,j,t) \in I$. With this, we obtain

$$\begin{aligned}&\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}} \\ &= \sum_{(i,j,t)\in I} \frac{\delta_{i,j,t}}{p_{i,j,t}} \widetilde{\mathcal{Z}}_{i,j,t} - \sum_{(i,j,t)\in I} \widetilde{\mathcal{Z}}_{i,j,t} \\ &= \sum_{(i,j,t)\in I} \left(\frac{\delta_{i,j,t}}{p_{i,j,t}} - 1\right) \widetilde{\mathcal{Z}}_{i,j,t} =: \sum_{(i,j,t)\in I} \mathcal{Z}_{i,j,t},\end{aligned}$$

defining the random operators $\mathcal{Z}_{i,j,t}$. Based on the assumption on the sampling model, the $\mathcal{Z}_{i,j,t}$ are independent and as the $\delta_{i,j,t}$ are Bernoulli variables with success probabilities $p_{i,j,t}$, it follows that

$$\mathbb{E}[\mathcal{Z}_{i,j,t}] = \left(\frac{\mathbb{E}[\delta_{i,j,t}]}{p_{i,j,t}} - 1\right) \mathcal{P}_{\mathbf{T}}\mathcal{G}E_{i,j,t}E^*_{i,j,t}\mathcal{G}^*\mathcal{P}_{\mathbf{T}} = 0.$$

Let $\mathbf{M} \in \mathbb{R}^{d_1 n \times d_2 n}$ be arbitrary. Since

$$\begin{aligned}\widetilde{\mathcal{Z}}_{i,j,t}(\mathbf{M}) &= \langle \mathbf{B}_{i,j,t}, \mathcal{P}_{\mathbf{T}}(\mathbf{M})\rangle_F \mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t}) \\ &= \langle \mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t}), \mathbf{M}\rangle_F \mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\end{aligned} \tag{61}$$

we obtain

$$\begin{aligned}\|\widetilde{\mathcal{Z}}_{i,j,t}(\mathbf{M})\|_F &\le |\langle \mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t}), \mathbf{M}\rangle_F| \, \|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F \\ &\le \|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F^2 \|\mathbf{M}\|_F \\ &\le \frac{r(d_1+d_2)}{nd_1d_2}\mu_{i,j,t}\|\mathbf{M}\|_F \le \frac{c_s r}{nT}\mu_{i,j,t}\|\mathbf{M}\|_F,\end{aligned}$$

using the definition of the local incoherence factor $\mu_{i,j,t}$, cf. Definition IV.1. This implies that

$$\|\widetilde{\mathcal{Z}}_{i,j,t}\| \le \frac{c_s r}{nT}\mu_{i,j,t}$$

almost surely for each $i \le j \le n$ and each $t \le T$ and, since $\delta_{i,j,t}/p_{i,j,t}\widetilde{\mathcal{Z}}_{i,j,t}$ and $\widetilde{\mathcal{Z}}_{i,j,t}$ are both positive semidefinite operators, that

$$\|\mathcal{Z}_{i,j,t}\| \le \max\left(\frac{\delta_{i,j,t}}{p_{i,j,t}}\|\widetilde{\mathcal{Z}}_{i,j,t}\|, \|\widetilde{\mathcal{Z}}_{i,j,t}\|\right) \tag{62}$$

$$\le \frac{1}{p_{i,j,t}}\|\widetilde{\mathcal{Z}}_{i,j,t}\| \le \frac{c_s r}{p_{i,j,t} nT}\mu_{i,j,t} \tag{63}$$

almost surely as well. Furthermore, for the expectation of the squares of $\mathcal{Z}_{i,j,t}$ we obtain

$$\begin{aligned}&\mathbb{E}\mathcal{Z}_{i,j,t}\mathcal{Z}_{i,j,t} \\ &= \mathbb{E}\left[\frac{\delta_{i,j,t}^2}{p_{i,j,t}^2}\widetilde{\mathcal{Z}}_{i,j,t}\widetilde{\mathcal{Z}}_{i,j,t}\right] - 2\mathbb{E}\left[\frac{\delta_{i,j,t}}{p_{i,j,t}}\widetilde{\mathcal{Z}}_{i,j,t}\widetilde{\mathcal{Z}}_{i,j,t}\right] + \widetilde{\mathcal{Z}}_{i,j,t}\widetilde{\mathcal{Z}}_{i,j,t} \\ &= \frac{p_{i,j,t}}{p_{i,j,t}^2}\widetilde{\mathcal{Z}}_{i,j,t}\widetilde{\mathcal{Z}}_{i,j,t} - \widetilde{\mathcal{Z}}_{i,j,t}\widetilde{\mathcal{Z}}_{i,j,t} \\ &= \left(\frac{1}{p_{i,j,t}} - 1\right)\widetilde{\mathcal{Z}}_{i,j,t}\widetilde{\mathcal{Z}}_{i,j,t}.\end{aligned}$$

Now, using (61) and observing that for any $\mathbf{M} \in \mathbb{R}^{d_1 n \times d_2 n}$

$$\begin{aligned}\widetilde{\mathcal{Z}}_{i,j,t}\widetilde{\mathcal{Z}}_{i,j,t}(\mathbf{M}) &= \|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F^2 \langle \mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t}), \mathbf{M}\rangle_F \mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t}) \\ &= \|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F^2 \mathcal{P}_{\mathbf{T}}\mathbf{B}_{i,j,t}\mathbf{B}^*_{i,j,t}\mathcal{P}_{\mathbf{T}}(\mathbf{M}),\end{aligned}$$

we obtain the spectral norm bound

$$\begin{aligned}\left\|\widetilde{\mathcal{Z}}_{i,j,t}\widetilde{\mathcal{Z}}_{i,j,t}\right\| &\le \|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F^2 \left\|\mathcal{P}_{\mathbf{T}}\mathbf{B}_{i,j,t}\mathbf{B}^*_{i,j,t}\mathcal{P}_{\mathbf{T}}\right\| \\ &\le \frac{r(d_1+d_2)}{nd_1d_2}\mu_{i,j,t} \le \frac{c_s r}{nT}\mu_{i,j,t},\end{aligned}$$

using the definition of $\mu_{i,j,t}$ and the fact that $\left\|\mathcal{P}_{\mathbf{T}}\mathbf{B}_{i,j,t}\mathbf{B}^*_{i,j,t}\mathcal{P}_{\mathbf{T}}\right\| \le 1$, as well as $d_1 + d_2 - 1 = T$ and the definition of $c_s = T(T+1)/(d_1d_2)$ in the last inequality. Due to a similar argument as made in (62), we obtain that

$$\begin{aligned}&\left\|\sum_{(i,j,t)\in I}\mathbb{E}\mathcal{Z}_{i,j,t}\mathcal{Z}_{i,j,t}\right\| = \left\|\sum_{(i,j,t)\in I}\left(\frac{1}{p_{i,j,t}} - 1\right)\widetilde{\mathcal{Z}}_{i,j,t}\widetilde{\mathcal{Z}}_{i,j,t}\right\| \\ &= \left\|\sum_{(i,j,t)\in I}\left(\frac{1}{p_{i,j,t}} - 1\right)\|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F^2\mathcal{P}_{\mathbf{T}}\mathbf{B}_{i,j,t}\mathbf{B}^*_{i,j,t}\mathcal{P}_{\mathbf{T}}\right\| \\ &\le \max_{(i',j',t')\in I}\left(\frac{1}{p_{i,j,t}} - 1\right)\|\mathcal{P}_{\mathbf{T}}(\mathbf{B}_{i,j,t})\|_F^2 \\ &\quad \cdot \left\|\sum_{(i,j,t)\in I}\mathcal{P}_{\mathbf{T}}\mathbf{B}_{i,j,t}\mathbf{B}^*_{i,j,t}\mathcal{P}_{\mathbf{T}}\right\| \\ &\le \max_{(i',j',t')\in I}\frac{c_s r}{nT}\frac{\mu_{i',j',t'}}{p_{i',j,'t'}}\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| \\ &\le \max_{(i',j',t')\in I}\frac{c_s r}{nT}\frac{\mu_{i',j',t'}}{p_{i',j,'t'}} =: \widetilde{c}\end{aligned} \tag{64}$$

using the formulas for $\mathbb{E}\mathcal{Z}_{i,j,t}\mathcal{Z}_{i,j,t}$ and $\widetilde{\mathcal{Z}}_{i,j,t}\widetilde{\mathcal{Z}}_{i,j,t}$ from above, the fact that the $\mathcal{P}_{\mathbf{T}}\mathbf{B}_{i,j,t}\mathbf{B}^*_{i,j,t}\mathcal{P}_{\mathbf{T}}$ are all positive semidefinite and the assumption the $p_{i,j,t} \le 1$ for all $(i,j,t) \in I$. Furthermore, we used the definition of the local coherences $\mu_{i',j',t'}$ from Definition IV.1 in the first inequality, and the fact that $\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| \le 1$ in the last inequality.

As the $\mathcal{Z}_{i,j,t}$ are Hermitian, we can now use (64) and (62) to apply the matrix Bernstein inequality Lemma A.5 to estimate that

$$\begin{aligned}
&\mathbb{P}\left(\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| \geq \alpha\right) \\
&\leq 2n^2 d_1 d_2 \exp\left(-\frac{\alpha^2/2}{\widetilde{c} + \widetilde{c}\alpha/3}\right) \\
&\leq 2\frac{(T+1)^2}{4} n^2 \exp\left(-\frac{\alpha^2/2}{\widetilde{c} + \widetilde{c}\alpha/3}\right) \\
&\leq \frac{1}{2}(T+1)^2 n^2 \exp\left(-\frac{3\alpha^2}{8\widetilde{c}}\right) \leq n^{-2},
\end{aligned}$$

where the last inequality holds if $\widetilde{c}^{-1} \geq \frac{8}{3\alpha^2}\left(4\log(n) + \log(1/2) + 2\log(T+1)\right)$, which, in view of the definition of $\widetilde{c}$ from (64) , is implied by the condition

$$p_{i,j,t} \geq \frac{32}{3\alpha^2}\mu_{i,j,t} c_s \frac{r}{nT} \log((T+1)n)$$

for all $1 \leq i \leq j \leq n, 1 \leq t \leq T$.

This shows that there exists an absolute constant $C > 1$ such that if (28) is fulfilled for each $(i,j,t) \in I$, with probability at least $1 - n^2$, it holds that

$$\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| < \alpha,$$

which finishes the proof of Lemma VIII.1. □

E. *Proof of Lemma VIII.2*

To show the perturbation result of Lemma VIII.2, we use ideas from the proof of [CWW19, Lemma 8]. As an auxiliary result, we also use the following lemma.

Lemma A.6 ([WCCL20, Lemma 4.2], [KMV21, Eq. (30)])**.** *If* $\mathbf{T} := \mathbf{T}_{\mathbf{H_A}}$ *and* $\mathbf{T}_{\mathbf{H}}$ *are the tangent spaces of the rank-r matrix manifold at* $\mathbf{H_A}$ *and* $\mathbf{H}$*, respectively, then*

$$\|\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\| \leq \frac{4\|\mathbf{H_A} - \mathbf{H}\|}{\sigma_r(\mathbf{H_A})}.$$

Proof of Lemma VIII.2. Recall that $\mathbf{T} = T_{\mathbf{H_A}} \subset \mathcal{M}_{d_1 n, d_2 n}$ is the tangent space onto $\mathcal{M}_r$ at $\mathbf{H_A}$. For any $\mathbf{Z} \in \mathcal{M}_{d_1 n, d_2 n}$, we have

$$\begin{aligned}
&\|\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}(\mathbf{Z})\|_F^2 \\
&= \langle \mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}(\mathbf{Z}), \mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}(\mathbf{Z})\rangle = \left\langle \mathcal{G}^*\mathcal{P}_{\mathbf{T}}(\mathbf{Z}), \mathcal{R}_{\Omega}^2\mathcal{G}^*\mathcal{P}_{\mathbf{T}}(\mathbf{Z})\right\rangle \\
&\leq \|\mathcal{R}_{\Omega}\| \left\langle \mathcal{G}^*\mathcal{P}_{\mathbf{T}}(\mathbf{Z}), \mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}(\mathbf{Z})\right\rangle \\
&= \|\mathcal{R}_{\Omega}\| \left\langle \mathbf{Z}, \mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}(\mathbf{Z})\right\rangle \\
&= \|\mathcal{R}_{\Omega}\| \big(\left\langle \mathbf{Z}, \left(\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\right)\mathbf{Z}\right\rangle \\
&\qquad\qquad + \left\langle \mathbf{Z}, \mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\mathbf{Z}\right\rangle \big) \\
&\leq \|\mathcal{R}_{\Omega}\| \left(\alpha\|\mathbf{Z}\|_F^2 + \|\mathbf{Z}\|_F^2\right) = \|\mathcal{R}_{\Omega}\|\,(1+\alpha)\,\|\mathbf{Z}\|_F^2
\end{aligned}$$

using the fact that $\mathcal{R}_{\Omega}$ is self-adjoint in the second inequality and Property VIII.1 in the last inequality. From this, it follows that

$$\|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_{\Omega}\| = \|\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| \leq \sqrt{\|\mathcal{R}_{\Omega}\|\,(1+\alpha)}. \tag{65}$$

With this preparation, we can now apply the triangle inequality multiple times to estimate that

$$\begin{aligned}
&\|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathrm{T}_{\mathbf{H}}} - \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\| \\
&\leq \|(\mathcal{P}_{\mathrm{T}_{\mathbf{H}}} - \mathcal{P}_{\mathbf{T}})\,\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\| + \|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^*\left(\mathcal{P}_{\mathrm{T}_{\mathbf{H}}} - \mathcal{P}_{\mathbf{T}}\right)\| \\
&+ \|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathbf{T}}\| \\
&+ \|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{G}^*\left(\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\right)\| + \|(\mathcal{P}_{\mathrm{T}_{\mathbf{H}}} - \mathcal{P}_{\mathbf{T}})\,\mathcal{G}\mathcal{G}^*\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\| \\
&\leq \|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}} - \mathcal{P}_{\mathbf{T}}\|\,\|\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\| + \|\mathcal{P}_{\mathbf{T}}\mathcal{G}\mathcal{R}_{\Omega}\|\,\|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}} - \mathcal{P}_{\mathbf{T}}\| + \alpha \\
&+ \|\mathcal{P}_{\mathbf{T}}\|\,\|\mathcal{G}\mathcal{G}^*\|\,\|\mathcal{P}_{\mathbf{T}} - \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\| + \|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}} - \mathcal{P}_{\mathbf{T}}\|\,\|\mathcal{G}\mathcal{G}^*\|\,\|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\| \\
&\leq \|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}} - \mathcal{P}_{\mathbf{T}}\|\left(2\,\|\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\|\right) + \alpha + 2\,\|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}} - \mathcal{P}_{\mathbf{T}}\| \\
&\leq \frac{8\,\|\mathbf{H} - \mathbf{H_A}\|}{\sigma_r(\mathbf{H_A})}\left(\sqrt{\|\mathcal{R}_{\Omega}\|\,(1+\alpha)} + 1\right) + \alpha \\
&\leq \alpha + \alpha = 2\alpha,
\end{aligned}$$

using the sub-multiplicativity of the spectral norm multiple times, Property VIII.1, in the second inequality, and (65) and Lemma A.6 in the penultimate inequality. Finally, we conclude the proof by using the closeness assumption (42) in the last inequality. □

F. *Proof of Lemma VIII.3*

We present the proof of Lemma VIII.3, which is inspired by the proofs of [YKJL17, Lemma 20] and [CC14, Lemma 1], but refines the respective arguments.

Proof of Lemma VIII.3. Let $\eta \in \ker \mathcal{R}_{\Omega}$. Due to the entrywise nature of the normalized sampling operator $\mathcal{R}_{\Omega}$, it holds that $\eta \in \ker \mathcal{R}_{\Omega}$ if and only if $\mathcal{D}\eta \in \ker \mathcal{R}_{\Omega}$ due to the diagonality of the diagonal operator $\mathcal{D} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ from (44). Therefore, it holds that $\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{H}(\eta) = \mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{G}\mathcal{D}(\eta) = \mathcal{R}_{\Omega}\mathcal{D}(\eta) = 0$, as $\mathcal{G}^*\mathcal{G} = \mathrm{Id}$ is the identity operator and as $\mathcal{H} = \mathcal{G}\mathcal{D}$ due to Lemma A.1, which implies further that $\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^*\mathcal{H}(\eta) = 0$

Furthermore, this also implies that

$$\left(\mathrm{Id} - \mathcal{G}\mathcal{G}^*\right)\mathcal{H}(\eta) = \left(\mathcal{G}\mathcal{D} - \mathcal{G}\mathcal{G}^*\mathcal{G}\mathcal{D}\right)\eta = \left(\mathcal{G}\mathcal{D} - \mathcal{G}\mathcal{D}\right)\eta = 0.$$

Therefore, taking the scalar product with $\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\mathcal{H}(\eta)$, we note that

$$\begin{aligned}
0 &= \langle \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\mathcal{H}(\eta), \left(\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*\right)\mathcal{H}(\eta)\rangle \\
&= \langle \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\mathcal{H}(\eta), \left(\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*\right)\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\mathcal{H}(\eta)\rangle \\
&+ \langle \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\mathcal{H}(\eta), \left(\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*\right)\mathcal{P}_{T_{\mathbf{H}}^{\perp}}\mathcal{H}(\eta)\rangle,
\end{aligned} \tag{66}$$

and furthermore, taking the scalar product with $\mathcal{P}_{T_{\mathbf{H}}^{\perp}}\mathcal{H}(\eta)$, we also observe that

$$\begin{aligned}
0 &= \langle \mathcal{P}_{T_{\mathbf{H}}^{\perp}}\mathcal{H}(\eta), \left(\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*\right)\mathcal{H}(\eta)\rangle \\
&= \langle \mathcal{P}_{T_{\mathbf{H}}^{\perp}}\mathcal{H}(\eta), \left(\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*\right)\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\mathcal{H}(\eta)\rangle \\
&+ \langle \mathcal{P}_{T_{\mathbf{H}}^{\perp}}\mathcal{H}(\eta), \left(\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*\right)\mathcal{P}_{T_{\mathbf{H}}^{\perp}}\mathcal{H}(\eta)\rangle,
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\langle \mathcal{P}_{T_{\mathbf{H}}^{\perp}}\mathcal{H}(\eta), \left(\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*\right)\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\mathcal{H}(\eta)\rangle \\
&= \langle \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\mathcal{H}(\eta), \left(\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*\right)\mathcal{P}_{T_{\mathbf{H}}^{\perp}}\mathcal{H}(\eta)\rangle \\
&= -\langle \mathcal{P}_{T_{\mathbf{H}}^{\perp}}\mathcal{H}(\eta), \left(\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*\right)\mathcal{P}_{T_{\mathbf{H}}^{\perp}}\mathcal{H}(\eta)\rangle,
\end{aligned}$$

where we used in the first equality the fact that $\mathcal{G}\mathcal{R}_{\Omega}\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*$ is self-adjoint as a sum of self-adjoint operators.

Inserting this into (66), we obtain

$$\begin{aligned}&\langle \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta), (\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*)\,\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\rangle \\ &= \langle \mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta), (\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*)\,\mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta)\rangle.\end{aligned} \tag{67}$$

We now bound the left and right hand side of the latter equality separately. On the one hand, we obtain a lower bound

$$\begin{aligned}&|\langle \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta), (\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*)\,\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\rangle| \\ &\geq |\langle \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta), \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\rangle| \\ &- |\langle \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta), (\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^* - \mathcal{G}\mathcal{G}^*)\,\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\rangle| \\ &= \|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\|_F^2 - \\ &|\langle \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta), (\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\,\mathcal{P}_{\mathrm{T}_{\mathbf{H}}} - \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{G}\mathcal{G}^*\,\mathcal{P}_{\mathrm{T}_{\mathbf{H}}})\,\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\rangle| \\ &\geq \|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\|_F^2 - \\ &\|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\,\mathcal{P}_{\mathrm{T}_{\mathbf{H}}} - \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{G}\mathcal{G}^*\,\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\|\,\|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\|_F^2 \\ &\geq \|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\|_F^2 - \frac{2}{5}\,\|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\|_F^2\,,\end{aligned}$$

using the projection property ${\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}}^2 = \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}$ in the equality and (43) in the last inequality, which implies that

$$\begin{aligned}&\|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\|_F^2 \\ &\leq \frac{5}{3}\,|\langle \mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta), (\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*)\,\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\rangle|\end{aligned} \tag{68}$$

On the other hand, we have the upper bounds

$$\begin{aligned}&\left|\langle \mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta), (\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^* + \mathrm{Id} - \mathcal{G}\mathcal{G}^*)\,\mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta)\rangle\right| \\ &\leq \left\|\mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta)\right\|_F \|\mathcal{G}\mathcal{R}_\Omega\mathcal{G}^*\| \left\|\mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta)\right\|_F \\ &+ \left\|\mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta)\right\|_F \|\mathrm{Id} - \mathcal{G}\mathcal{G}^*\| \left\|\mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta)\right\|_F \\ &\leq \|\mathcal{G}\|\,\|\mathcal{R}_\Omega\|\,\|\mathcal{G}^*\| \left\|\mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta)\right\|_F^2 + \left\|\mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta)\right\|_F^2 \\ &\leq (\|\mathcal{R}_\Omega\| + 1) \left\|\mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta)\right\|_F^2\,,\end{aligned}$$

using the sub-mulitiplicativity of the spectral norm and the fact that $\mathrm{Id} - \mathcal{G}\mathcal{G}^*$ is a projection in the third inequality, and observing that $\|\mathcal{G}\| \leq 1$ and$\|\mathcal{G}^*\| \leq 1$ in the last inequality. Combining this with (67) and (68), this implies

$$\|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\|_F^2 \leq \frac{5}{3}\,(\|\mathcal{R}_\Omega\| + 1) \left\|\mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta)\right\|_F^2$$

and therefore

$$\begin{aligned}\|\mathcal{H}(\eta)\|_F^2 &= \|\mathcal{P}_{\mathrm{T}_{\mathbf{H}}}\,\mathcal{H}(\eta)\|_F^2 + \left\|\mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta)\right\|_F^2 \\ &\leq \frac{5}{3}\,(\|\mathcal{R}_\Omega\| + 8/5) \left\|\mathcal{P}_{T_{\mathbf{H}}^\perp}\mathcal{H}(\eta)\right\|_F^2\,,\end{aligned}$$

which finishes the proof. □

Next, we provide an auxiliary result of similar flavor as Lemma VIII.3 to be used in the convergence analysis of `TOIRLS`.

Lemma A.7. *Assume that Property VIII.1 holds true for a normalized sampling operator $\mathcal{R}_\Omega : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ with respect to a rank-r matrix $\mathbf{H}_{\mathbf{A}} \in \mathcal{M}_{d_1n,d_2n}$ and constant $\alpha = 1/5$. If $\widetilde{\mathbf{X}}^{(k)} \in \mathcal{M}_n^{\oplus T}$ is such that the best rank-r approximation of a matrix $\mathbf{H}_k := \mathcal{H}(\widetilde{\mathbf{X}}^{(k)})$, i.e.,*

$$\mathcal{T}_r(\mathbf{H}_k) = \underset{\mathbf{Z}\in\mathcal{M}_{d_1n,d_2n}:\mathrm{rank}(\mathbf{Z})\leq r}{\arg\min} \|\mathbf{Z} - \mathbf{H}_k\| \tag{69}$$

satisfies

$$\mathcal{T}_r(\mathbf{H}_k) \in \mathcal{B}_{\mathbf{H}_{\mathbf{A}}}\left(\frac{\sigma_r(\mathbf{H}_{\mathbf{A}})}{32\sqrt{r}\kappa\left(\sqrt{6\|\mathcal{R}_\Omega\|/5} + 1\right)}\right),$$

then it holds that

$$\|\mathcal{H}(\eta^{(k)})\| < \sqrt{\frac{20}{3}\,(\|\mathcal{R}_\Omega\| + 8/5)\sqrt{dn - r}\sigma_{r+1}(\mathbf{H}_k)}$$

where $\eta^{(k)} = \widetilde{\mathbf{X}}^{(k)} - \mathcal{Q}_T(\mathbf{A})$.

Proof. If $\mathbf{T}_k := T_{\mathcal{T}_r(\mathbf{H}_k)}$ is tangent space onto the manifold of rank-r matrices at $\mathcal{T}_r(\mathbf{H}_k)$ and if $\mathbf{U}_\perp^{(k)} \in \mathbb{R}^{d_1n\times(d_1n-r)}$ and $\mathbf{V}_\perp^{(k)} \in \mathbb{R}^{d_2n\times(d_2n-r)}$ are the matrices with the last d_1n-r and last d_2n-r left and right singular vectors of $\mathbf{H}_k$, respectively, we can write the action of the projection $\mathcal{P}_{\mathbf{T}_k^\perp} : \mathcal{M}_{d_1n,d_2n} \to \mathcal{M}_{d_1n,d_2n}$ onto the orthogonal complement $\mathbf{T}_k^\perp$ of $\mathbf{T}_k$ as

$$\mathcal{P}_{\mathbf{T}_k^\perp}(\mathbf{Z}) = \mathbf{U}_\perp^{(k)}\mathbf{U}_\perp^{(k)*}\mathbf{Z}\mathbf{V}_\perp^{(k)}\mathbf{V}_\perp^{(k)*},$$

cf., e.g., [Rec11]. Let $d = \min(d_1, d_2)$. If $\Sigma_k^\perp \in \mathbb{R}^{(d_1n-r)\times(d_2n-r)}$ is diagonal with the last $dn - r$ singular values of $\mathbf{H}_k$ ordered in an non-increasing way, we observe that

$$\mathcal{P}_{\mathbf{T}_k^\perp}(\mathbf{H}_k) = \mathbf{U}_\perp^{(k)}\Sigma_k^\perp\mathbf{V}_\perp^{(k)*}.$$

Now, if $\mathbf{H}_{\mathbf{A}} = \mathbf{U}_0\Sigma_0\mathbf{V}_0^T$ is a compact singular value decomposition of $\mathbf{H}_{\mathbf{A}}$, we estimate

$$\begin{aligned}&\|\mathcal{P}_{\mathbf{T}_k^\perp}(\mathcal{H}(\eta^{(k)}))\|_F \leq \|\mathcal{P}_{\mathbf{T}_k^\perp}(\mathbf{H}_k)\|_F + \|\mathcal{P}_{\mathbf{T}_k^\perp}(\mathbf{H}_{\mathbf{A}})\|_F \\ &\leq \sqrt{\sum_{i=r+1}^{dn}\sigma_i^2(\mathbf{H}_k)} + \left\|\mathbf{U}_\perp^{(k)}\mathbf{U}_\perp^{(k)*}\mathbf{H}_{\mathbf{A}}\mathbf{V}_\perp^{(k)}\mathbf{V}_\perp^{(k)*}\right\|_F \\ &\leq \sqrt{\sum_{i=r+1}^{dn}\sigma_i^2(\mathbf{H}_k)} + \|\mathbf{U}_\perp^{(k)}\| \left\|\mathbf{U}_\perp^{(k)*}\mathbf{H}_{\mathbf{A}}\mathbf{V}_\perp^{(k)}\right\|_F \|\mathbf{V}_\perp^{(k)*}\| \\ &\leq \sqrt{dn - r}\sigma_{r+1}(\mathbf{H}_k) + \|\mathbf{U}_\perp^{(k)*}\mathbf{U}_0\|\|\mathbf{\Sigma}_0\|_F\|\mathbf{V}_0^*\mathbf{V}_\perp^{(k)}\|\end{aligned}$$

using the definition of $\eta^{(k)}$, the triangle inequality, the fact that $\|\mathbf{A}\mathbf{B}\|_F \leq \|\mathbf{A}\|\|\mathbf{B}\|_F$ for all matrices $\mathbf{A}$ and $\mathbf{B}$, and that $\|\mathbf{V}_\perp^{(k)*}\| = \|\mathbf{U}_\perp^{(k)}\| = 1$.

By the classical perturbation bound due to Wedin [Wed72, Ste06], cf. also [KMV21, Lemma B.6],

$$\max\{\|\mathbf{U}_\perp^{(k)*}\mathbf{U}_0\|, \|\mathbf{V}_0^*\mathbf{V}_\perp^{(k)}\|\} \leq \sqrt{2}\frac{1}{\zeta}\|\mathcal{H}(\eta^{(k)})\|\,,$$

if $\mathbf{H}_k \in \mathcal{B}_{\mathbf{H}_{\mathbf{A}}}(\zeta)$ with $0 < \zeta < 1$. By assumption $\zeta < 1/2$, so

$$\begin{aligned}&\|\mathcal{P}_{\mathbf{T}_k^\perp}(\mathcal{H}(\eta^{(k)}))\|_F \\ &\leq \sqrt{dn - r}\sigma_{r+1}(\mathbf{H}_k) + 8\|\mathcal{H}(\eta^{(k)})\|^2\sqrt{r}\sigma_1(\mathbf{H}_{\mathbf{A}}) \\ &\leq \sqrt{dn - r}\sigma_{r+1}(\mathbf{H}_k) + 8\|\mathcal{H}(\eta^{(k)})\|^2\sqrt{r}\kappa\sigma_r(\mathbf{H}_{\mathbf{A}}).\end{aligned}$$

Due to our assumptions, we can apply Lemma VIII.2 for $\alpha = 1/5$ and further Lemma VIII.3 for $\eta = \eta^{(k)}$ to estimate that

$$\begin{aligned}&\|\mathcal{H}(\eta^{(k)})\| \leq \|\mathcal{H}(\eta^{(k)})\|_F\\ &\leq \sqrt{\frac{5}{3}\left(\|\mathcal{R}_\Omega\|+8/5\right)}\|\mathcal{P}_{\mathbf{T}_k^\perp}(\mathcal{H}(\eta^{(k)}))\|_F\\ &\leq \sqrt{\frac{5}{3}\left(\|\mathcal{R}_\Omega\|+8/5\right)}\\ &\left(\sqrt{dn-r}\cdot\sigma_{r+1}(\mathbf{H}_k)+8\|\mathcal{H}(\eta^{(k)})\|^2\sqrt{r}\kappa\sigma_r(\mathbf{H}_\mathbf{A})\right)\\ &\leq \sqrt{\frac{5}{3}\left(\|\mathcal{R}_\Omega\|+8/5\right)}\sqrt{dn-r}\cdot\sigma_{r+1}(\mathbf{H}_k)\\ &\qquad\qquad+\frac{8\sqrt{r}\kappa\sqrt{\frac{5}{3}\left(\|\mathcal{R}_\Omega\|+8/5\right)}}{(32\sqrt{r}\kappa)\left(\sqrt{6\|\mathcal{R}_\Omega\|/5}+1\right)}\|\mathcal{H}(\eta^{(k)})\|\\ &< \sqrt{\frac{5}{3}\left(\|\mathcal{R}_\Omega\|+8/5\right)}\sqrt{dn-r}\cdot\sigma_{r+1}(\mathbf{H}_k)+\frac{1}{2}\|\mathcal{H}(\eta^{(k)})\|.\end{aligned}$$

Rearranging this estimate, we obtain

$$\|\mathcal{H}(\eta^{(k)})\| < \sqrt{\frac{20}{3}\left(\|\mathcal{R}_\Omega\|+8/5\right)}\sqrt{dn-r}\sigma_{r+1}(\mathbf{H}_k).$$

□

G. *Proof of Proposition VIII.1*

In this section, we provide the proof of Proposition VIII.1. For this purpose, we use key results of [KMV21], adapted to our notation.

Proposition A.3 ([KMV21, Lemma B.8 and Lemma B.9])**.** *Let $\mathbf{H}_\mathbf{A} = \mathcal{H}(\mathcal{Q}_T(\mathbf{A})) \in \operatorname{Ran}\mathcal{H}$ be a matrix of rank r, let $\widetilde{\mathbf{X}}^{(k)}$ be the k-th iterate of Algorithm 1 for input parameters Ω, $\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A}))$, $\lambda = 0$ and $\widetilde{r} = r$. For $\mathbf{H}_k := \mathcal{H}(\widetilde{\mathbf{X}}^{(k)})$, assume that $\varepsilon_k = \sigma_{r+1}(\mathbf{H}_k)$ and that*

$$\|\mathcal{H}(\eta)\|_F \leq C\|\mathcal{P}_{\mathbf{T}_k^\perp}\mathcal{H}(\eta)\|_F \qquad \textit{for all } \eta \in \ker\mathcal{R}_\Omega$$

for some constant C, where $\mathbf{T}_k = \mathbf{T}_{\mathcal{T}_r(\mathbf{H}_k)}$ is the tangent space onto the manifold of rank-r matrices at $\mathcal{T}_r(\mathbf{H}_k)$. Then

$$\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k+1)})-\mathbf{H}_\mathbf{A}\| \leq C^2\varepsilon_k^2\|W_{\mathbf{H}_k}(\mathbf{H}_\mathbf{A})\|_{S_1},$$

where $W_{\mathbf{H}_k}: \mathcal{M}_{d_1n,d_2n} \to \mathcal{M}_{d_1n,d_2n}$ is the optimal weight operator of $\mathbf{H}_k$ as in (19), and $\|\cdot\|_{S_1}$ describes the Schatten-1 norm.

Furthermore, if additionally $\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) \in \mathcal{B}_{\mathbf{H}_\mathbf{A}}(\zeta)$ for some $0 < \zeta < 1$, then

$$\begin{aligned}&\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k+1)})-\mathbf{H}_\mathbf{A}\| \leq C^2r(1-\zeta)^{-2}\sigma_r(\mathbf{H}_\mathbf{A})^{-1}\\ &\left(\varepsilon_k^2+4\varepsilon_k\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})-\mathbf{H}_\mathbf{A}\|\kappa+2\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k+1)})-\mathbf{H}_\mathbf{A}\|^2\kappa\right),\end{aligned}$$

where $\kappa = \sigma_1(\mathbf{H}_\mathbf{A})/\sigma_r(\mathbf{H}_\mathbf{A})$ is the condition number of $\mathbf{H}_\mathbf{A}$.

We can now put Lemma VIII.3, Lemma VIII.2 and Proposition A.3 together to prove Proposition VIII.1, showing that we attain locally quadratic convergence under the stated assumptions.

Proof of Proposition VIII.1. Let $k \in \mathbb{N}$ and $\widetilde{\mathbf{X}}^{(k)}$ be the k-th iterate of Algorithm 1 with inputs Ω, $\mathbf{y} = P_\Omega(\mathcal{Q}_T(\mathbf{A}))$, $\lambda = 0$ and $\widetilde{r} = r$. First, we observe that

$$\begin{aligned}\zeta_3 = \min\left(\zeta_1,\zeta_2,\zeta_3,\frac{1}{2}\right) := \min\Bigg(&\frac{(32\sqrt{r}\kappa)^{-1}}{\sqrt{6\|\mathcal{R}_\Omega\|/5}+1},\\ &\frac{3}{20r}\frac{(1+6\kappa)^{-1}\sigma_r(\mathbf{H}_\mathbf{A})}{\|\mathcal{R}_\Omega\|+8/5},\\ &\frac{3^{\frac{3}{2}}\left(1+6\kappa\right)^{-1}\sigma_r(\mathbf{H}_\mathbf{A})}{20^{\frac{3}{2}}r\left(\|\mathcal{R}_\Omega\|+8/5\right)^{\frac{3}{2}}(dn-r)^{\frac{1}{2}}},\frac{1}{2}\Bigg).\end{aligned}$$

We note that Property VIII.1 is satisfied with respect to $\mathbf{H}_\mathbf{A}$ and constant $\alpha = 1/5$. Since $\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) \in \mathcal{B}_{\mathbf{H}_\mathbf{A}}(\zeta_3)$ (as $\zeta_3 < \zeta_2$), we also have that a best rank-r approximation $\mathcal{T}_r(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}))$ of $\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})$ satisfies

$$\begin{aligned}&\left\|\mathcal{T}_r(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}))-\mathbf{H}_\mathbf{A}\right\|\\ &\leq \left\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})-\mathbf{H}_\mathbf{A}\right\|+\left\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})-\mathcal{T}_r(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}))\right\|\\ &\leq 2\left\|\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})-\mathbf{H}_\mathbf{A}\right\| \leq 2\zeta_3 \leq \frac{1}{40}\left(\sqrt{6\|\mathcal{R}_\Omega\|/5}+1\right)^{-1},\end{aligned}$$

from which it follows due to Lemma VIII.2 that (43) holds true for $\mathbf{H} := \mathcal{T}_r(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}))$. Thus, Lemma VIII.3 implies that

$$\|\mathcal{H}(\eta^{(k)})\|_F^2 \leq \frac{5}{3}\left(\|\mathcal{R}_\Omega\|+8/5\right)\left\|\mathcal{P}_{\mathbf{T}_k^\perp}\mathcal{H}(\eta^{(k)})\right\|_F^2,$$

where $\eta^{(k)} := \widetilde{\mathbf{X}}^{(k)} - \mathcal{Q}_T(\mathbf{A})$ and $\mathbf{T}_k = \mathbf{T}_\mathbf{H} = \mathbf{T}_{\mathcal{T}_r(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}))}$ is tangent space onto the manifold of rank-r matrices at $\mathbf{H}$.

Next, since $\varepsilon_k = \sigma_{r+1}(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}))$, it follows from the Eckardt-Young-Mirsky theorem [Mir60] that

$$\begin{aligned}\varepsilon_k &= \sigma_{r+1}(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}))\\ &= \|\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})-\mathbf{H}\| \leq \|\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})-\mathbf{H}_\mathbf{A}\| = \|\mathcal{H}(\eta^{(k)})\|,\end{aligned}$$

where we used that $\mathbf{H}_\mathbf{A}$ is of rank r in the inequality. Since also $\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) \in \mathcal{B}_{\mathbf{H}_\mathbf{A}}(1/2)$, it follows therefore from Proposition A.3 that

$$\begin{aligned}&\|\mathcal{H}(\eta^{(k+1)})\|\\ &\leq \frac{20}{3}\left(\|\mathcal{R}_\Omega\|+8/5\right)r\sigma_r(\mathbf{H}_\mathbf{A})^{-1}\\ &\qquad\left(\varepsilon_k^2+4\varepsilon_k\|\mathcal{H}(\eta^{(k)})\|\kappa+2\|\mathcal{H}(\eta^{(k)})\|^2\kappa\right)\\ &\leq \frac{20}{3}\left(\|\mathcal{R}_\Omega\|+8/5\right)r\sigma_r(\mathbf{H}_\mathbf{A})^{-1}\left(1+6\kappa\right)\|\mathcal{H}(\eta^{(k)})\|^2,\end{aligned} \tag{70}$$

where $\kappa = \sigma_1(\mathbf{H}_\mathbf{A})/\sigma_r(\mathbf{H}_\mathbf{A})$ is the condition number of $\mathbf{H}_\mathbf{A}$ and $\eta^{(k+1)} = \widetilde{\mathbf{X}}^{(k+1)} - \mathcal{Q}_T(\mathbf{A})$.

If, additionally, we assume that $\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) \in \mathcal{B}_{\mathbf{H}_\mathbf{A}}(\zeta_2)$, then we can further bound the right hand side of (70) to obtain

$$\|\mathcal{H}(\eta^{(k+1)})\| < \|\mathcal{H}(\eta^{(k)})\|, \tag{71}$$

and also obtain a quadratic decay in the spectral error

$$\|\mathcal{H}(\eta^{(k+1)})\| \leq \nu\|\mathcal{H}(\eta^{(k)})\|^2,$$

with $\nu = \frac{20}{3\sigma_r(\mathbf{H}_\mathbf{A})}(1+6\kappa)\left(\|\mathcal{R}_\Omega\|+8/5\right)r$.

What remains to be shown is that the $(r+1)$-st singular value $\mathcal{H}\left(\widetilde{\mathbf{X}}^{(k)}\right)$ is strictly decreasing from one iterate to the next. If $\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) \in \mathcal{B}_{\mathbf{H_A}}(\zeta_1)$, it follows that

$$\begin{aligned}
&\sigma_{r+1}(\mathcal{H}(\widetilde{\mathbf{X}}^{(k+1)})) \leq \|\mathcal{H}(\eta^{(k+1)})\| \\
&\leq \frac{20}{3}\left(\|\mathcal{R}_\Omega\| + 8/5\right) r \sigma_r(\mathbf{H_A})^{-1} (1+6\kappa) \|\mathcal{H}(\eta^{(k)})\|^2 \\
&< \left(\frac{20}{3}\left(\|\mathcal{R}_\Omega\| + 8/5\right)\right)^{3/2} r \qquad (72)\\
&\qquad \cdot (1+6\kappa)\sqrt{dn-r}\,\sigma_{r+1}(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}))\frac{\|\mathcal{H}(\eta^{(k)})\|}{\sigma_r(\mathbf{H_A})} \\
&\leq \sigma_{r+1}(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})),
\end{aligned}$$

using (70) in the second inequality, Lemma A.7 in the third inequality, and $\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}) \in \mathcal{B}_{\mathbf{H_A}}(\zeta_3)$ in the last inequality.

Finally, we recall the update rule (21), which gives that (72) implies that $\varepsilon_{k+1} = \sigma_{r+1}(\mathcal{H}(\widetilde{\mathbf{X}}^{(k+1)}))$, so that (71) ensures that the assumptions of Proposition VIII.1 are not only fulfilled for iteration k, but also for iteration $k+1$. By induction, this implies that $\mathcal{H}(\widetilde{\mathbf{X}}^{(k+\ell)}) \xrightarrow{\ell \to \infty} \mathbf{H_A}$, concluding the proof of Proposition VIII.1. □

H. Computational Details

In this section, we detail some aspects of anx efficient implementations of `TOIRLS`, cf. Algorithm 1.

1) Explicit Expression for Weighted Least Squares Solution: First, we justify the explicit formula of Section V provided for the $k+1$-st iterate $\widetilde{\mathbf{X}}^{(k+1)}$ of `TOIRLS`.

Lemma A.8. *For any $\lambda \geq 0$, it holds that the solution $\widetilde{\mathbf{X}}^{(k+1)}$ of the weighted least squares problem (20) satisfies*

$$\widetilde{\mathbf{X}}^{(k+1)} = \widetilde{W}_{\mathbf{H}_k}^{-1} P_\Omega^* \left(\lambda \operatorname{Id} + P_\Omega \widetilde{W}_{\mathbf{H}_k}^{-1} P_\Omega^*\right)^{-1}(\mathbf{y}), \qquad (73)$$

where $\widetilde{W}_{\mathbf{H}_k}^{-1} : \mathcal{M}_{d_1 n, d_2 n} \to \mathcal{M}_{d_1 n, d_2 n}$ is the inverse of the effective weight operator $\widetilde{W}_{\mathbf{H}_k} : \mathcal{M}_{d_1 n, d_2 n} \to \mathcal{M}_{d_1 n, d_2 n}$ of Definition III.1.

Proof. Using the substitution $\widetilde{\mathbf{X}'} = \widetilde{W}_{\mathbf{H}_k}^{1/2}(\widetilde{\mathbf{X}})$ in (20), we obtain that

$$\widetilde{\mathbf{X}}^{(k+1)} = \widetilde{W}_{\mathbf{H}_k}^{-1/2}\left(\widetilde{\mathbf{X}'}^{(k)}\right)$$

where

$$\begin{aligned}
\widetilde{\mathbf{X}'}^{(k+1)} &= \operatorname*{arg\,min}_{\widetilde{\mathbf{X}'} \in \mathcal{M}_n^{\oplus T}} \left\{ \langle \widetilde{\mathbf{X}}', \widetilde{\mathbf{X}}' \rangle + \frac{1}{\lambda}\left\| P_\Omega(\widetilde{W}_{\mathbf{H}_k}^{-1/2}(\widetilde{\mathbf{X}}')) - \mathbf{y} \right\|_2^2) \right\} \\
&= \operatorname*{arg\,min}_{\widetilde{\mathbf{X}'} \in \mathcal{M}_n^{\oplus T}} \left\{ \lambda \|\widetilde{\mathbf{X}}'\|_F^2 + \left\| \mathcal{F}(\widetilde{\mathbf{X}}') - \mathbf{y} \right\|_2^2 \right\} \\
&= \mathcal{F}^* \left(\lambda \operatorname{Id} + \mathcal{F}\mathcal{F}^*\right)^{-1} \mathbf{y} \\
&= \widetilde{W}_{\mathbf{H}_k}^{-1/2} P_\Omega^* \left(\lambda \operatorname{Id} + P_\Omega \widetilde{W}_{\mathbf{H}_k}^{-1} P_\Omega^*\right)^{-1}(\mathbf{y}),
\end{aligned}$$

defining $\mathcal{F} := P_\Omega \circ \widetilde{W}_{\mathbf{H}_k}^{-1/2} : \mathcal{M}_{d_1 n, d_2 n} \to \mathbb{R}^m$ to interpret the problem as a ridge regression/ℓ_2-penalized least squares problem in the second equality, and using the inner product implementation of ridge regression in the third equality (see, e.g., [FLZZ20, Theorem 2.4]). This shows (73).

For $\lambda = 0$, we note that

$$\widetilde{\mathbf{X}}^{(k+1)} = \operatorname*{arg\,min}_{\widetilde{\mathbf{X}} \in \mathcal{M}_n^{\oplus T} : P_\Omega(\widetilde{\mathbf{X}}) = \mathbf{y}} \langle \widetilde{\mathbf{X}}, \widetilde{W}_{\mathbf{H}_k}(\widetilde{\mathbf{X}}) \rangle = \widetilde{W}_{\mathbf{H}_k}^{-1/2}\left(\widetilde{\mathbf{X}'}^{(k)}\right)$$

with

$$\begin{aligned}
\widetilde{\mathbf{X}'}^{(k+1)} &= \operatorname*{arg\,min}_{\widetilde{\mathbf{X}'} \in \mathcal{M}_n^{\oplus T} : \mathcal{F}(\widetilde{\mathbf{X}'}) = \mathbf{y}} \|\widetilde{\mathbf{X}}'\|_F^2 = \mathcal{F}^* \left(\mathcal{F}\mathcal{F}^*\right)^{-1} \mathbf{y} \\
&= \widetilde{W}_{\mathbf{H}_k}^{-1/2} P_\Omega^* \left(P_\Omega \widetilde{W}_{\mathbf{H}_k}^{-1} P_\Omega^*\right)^{-1}(\mathbf{y}),
\end{aligned}$$

using an analogous substitution as above and the fact that $\mathcal{F}^\dagger = \mathcal{F}^* \left(\mathcal{F}\mathcal{F}^*\right)^{-1}$ is the Moore-Penrose inverse of $\mathcal{F}$ as defined above. □

2) Efficient Implementation of `TOIRLS`: In this section we outline the main computational steps of an efficient implementation of `TOIRLS`. In particular, we provide an algorithm, Algorithm 2, for computing the weighted least squares solution (73) essentially via a conjugate gradient method applied to a $(r_k(nd_1 + nd_2 + r_k) \times r_k(nd_1 + nd_2 + r_k)) = O(rnT) \times O(rnT)$ linear system.

In the following, we let $S_k := \mathbb{R}^{r_k(nd_1 + nd_2 + r_k)}$, and we recall from the proof of Lemma A.7 that if $\widetilde{\mathbf{X}}^{(k)} \in \mathcal{M}_{d_1 n, d_2 n}$ is the k-th iterate of `TOIRLS` and $\mathbf{H}_k := \mathcal{H}(\widetilde{\mathbf{X}}^{(k)})$, $\mathbf{T}_k := \mathbf{T}_{\mathcal{T}_{r_k}(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)}))}$ denotes tangent space onto the manifold of rank-r_k matrices at $\mathcal{T}_{r_k}\left(\mathcal{H}(\widetilde{\mathbf{X}}^{(k)})\right)$ (here, with r_k instead of r), where $\mathcal{T}_{r_k}(\mathbf{H}_k)$ is the best rank-r r_k approximation of $\mathbf{H}_k$, cf. (69). Given the subspace $\mathbf{T}_k \subset \mathcal{M}_{d_1 n, d_2 n}$, we let $P_{\mathbf{T}_k} : S_k \to \mathbf{T}_k$ be the parametrization operator defined, for $\gamma \in S_k$, as

$$P_{\mathbf{T}_k}(\gamma) := \mathbf{U}^{(k)} \Gamma_1 \mathbf{V}^{(k)*} + \mathbf{U}^{(k)} \Gamma_2 + \Gamma_3 \mathbf{V}^{(k)*},$$

where $\Gamma_1 \in \mathbb{R}^{r_k \times r_k}, \Gamma_2 \in \mathbb{R}^{r_k \times nd_2}$ and $\Gamma_d \in \mathbb{R}^{nd_1 \times r_k}$ are matricizations of the first r_k^2, central $r_k nd_2$ and final $r_k nd_2$ coordinates of $\gamma \in S_k$, respectively. We note that the projection operator $\mathcal{P}_{\mathbf{T_k}} : \mathcal{M}_{d_1 n, d_2 n} \to \mathcal{M}_{d_1 n, d_2 n}$ can be implemented via $\mathcal{P}_{\mathbf{T}_k} = P_{\mathbf{T}_k} P_{\mathbf{T}_k}^* \cdot$ Recall that $\mathcal{G} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_{d_1 n, d_2 n}$ is the normalized block Hankel operator (36) and $\mathcal{D} : \mathcal{M}_n^{\oplus T} \to \mathcal{M}_n^{\oplus T}$ the diagonal operator of (44). Finally, $\mathbf{I}_m$ and $\mathbf{I}_{S_k}$ are identity matrices on $\mathbb{R}^m$ and S_k, respectively, and $\mathbf{D}_{S_k} : S_k \to S_k$ is a diagonal matrix that contains coordinates $\left(\Sigma_{\varepsilon_k, d_1 n}\right)_{ii}^{-1} \left(\Sigma_{\varepsilon_k, d_2 n}\right)_{jj} = \max\left(\sigma_i^{(k)}, \varepsilon_k\right)^{-1} \max\left(\sigma_j^{(k)}, \varepsilon_k\right)^{-1}$ for $(i,j) \in \{(i,j) \in [nd_1] \times [nd_2] : i \leq r_k \text{ or } j \leq r_k\}$ on its diagonal, which is related to the weight operator $W_{\mathbf{H}_k} : \mathcal{M}_{d_1 n, d_2 n} \to \mathcal{M}_{d_1 n, d_2 n}$ of (19) by

$$W_{\mathbf{H}_k}(\mathbf{Z}) = \left(P_{T_k} \mathbf{D}_{S_k} P_{T_k}^* + \epsilon_k^{-2}\left(\operatorname{Id} - P_{T_k} P_{T_k}^*\right)\right) \mathbf{Z}, \qquad (74)$$

cf. [KMV21, Appendix A, Eq. (12)]. With these notational conventions, we can formulate Algorithm 2.

Lemma A.9 shows that Algorithm 2 indeed computes the weighted least squares solution $\widetilde{\mathbf{X}}^{(k)}$.

Lemma A.9. *Let $\widetilde{\mathbf{X}}^{(k)} \in \mathcal{M}_{d_1 n, d_2 n}$ be the k-th iterate of `TOIRLS` (Algorithm 1) for an observation vector $\mathbf{y} \in \mathbb{R}^m$ with $m = |\Omega|, \lambda \geq 0$ and smoothing parameter $\varepsilon_k > 0$, let $\mathbf{H}_k = \mathcal{H}\left(\widetilde{\mathbf{X}}^{(k)}\right)$ and $r_k = |\{i \in [dn] : \sigma_i(\mathbf{H}_k) > \varepsilon_k\}|$. Then the following statements hold.*

1) The $(k+1)$-st iterate $\widetilde{\mathbf{X}}^{(k+1)}$ of Algorithm 1 satisfies

$$\widetilde{\mathbf{X}}^{(k+1)} = \mathcal{D}^{-1} \widetilde{P}_\Omega^*(\mathbf{p}_{k+1}) + \mathcal{D}^{-1} \mathcal{G}^* P_{\mathbf{T}_k}(\gamma_{k+1})$$

Algorithm 2 Implementation of $k+1$-st weighted least squares step of `TOIRLS`

```
**Input:** Set $\Omega$, observations $\mathbf{y} \in \mathbb{R}^m$, $\lambda \geq 0$, matrices $\mathbf{U}^{(k)} \in \mathbb{R}^{d_1 \times r_k}$, $\mathbf{V}^{(k)} \in \mathbb{R}^{d_2 \times r_k}$ of singular vectors and $r_k$ leading singular values $\sigma_1^{(k)}, \dots, \sigma_{r_k}^{(k)}$ of $\mathbf{H}_k$ , smoothing parameter $\varepsilon_k$, initialization $\gamma_{k+1}^{(0)} = P_{T_k}^* P_{T_{k-1}}(\gamma_k) \in \mathbb{R}^{r_k(nd_1+nd_2+r_k)}$ where $\gamma_k \in \mathbb{R}^{r_{k-1}(nd_1+nd_2+r_{k-1})}$ is respective parameter (76) of the $k$-th iteration.
Let $\mathbf{K} := \lambda\varepsilon_k^{-2}\mathbf{I} + P_\Omega \mathcal{D}^{-2} P_\Omega^*$ and
$$\mathbf{M} := P_{\mathbf{T}_k}^* \mathcal{G}\mathcal{D}^{-1} P_\Omega^* \mathbf{K}^{-1} P_\Omega \mathcal{D}^{-1} \mathcal{G}^* P_{\mathbf{T}_k} + \frac{\mathbf{D}_{S_k}^{-1}}{\mathbf{D}_{S_k}^{-1} - \varepsilon_k^2 \mathbf{I}_{S_k}} - P_{\mathbf{T}_k}^* \mathcal{G}\mathcal{G}^* P_{\mathbf{T}_k}. \quad (75)$$
  Compute $\mathbf{h}_k^0 := P_{\mathbf{T}_k}^* \mathcal{G}\mathcal{D}^{-1} P_\Omega^* \mathbf{K}^{-1} \mathbf{y} - \mathbf{M}\gamma_{k+1}^{(0)} \in \mathbb{R}^{r_k(nd_1+nd_2+r_k)}$.
  Solve linear system
      $$\mathbf{M}\Delta\gamma_{k+1} = \mathbf{h}_k^0 \quad (76)$$
    for $\Delta\gamma_{k+1} \in S_k$ by the *conjugate gradient* method [HS52, Meu06].
  Compute $\gamma_{k+1} = \gamma_{k+1}^{(0)} + \Delta\gamma_{k+1}$.
  Compute residual $\mathbf{p}_{k+1} := \mathbf{K}^{-1}(\mathbf{y} - P_\Omega \mathcal{D}^{-1} \mathcal{G}^* P_{\mathbf{T}_k}(\gamma_{k+1})) \in \mathbb{R}^m$ where
      $$\mathbf{K} := \lambda\varepsilon_k^{-2}\mathbf{I} + P_\Omega \mathcal{D}^{-2} P_\Omega^*. \quad (77)$$
**Output:** $\mathbf{p}_{k+1} \in \mathbb{R}^m$ and $\gamma_{k+1} \in \mathbb{R}^{r_k(nd_1+nd_2+r_k)}$.
```

where $\mathbf{p}_{k+1} \in \mathbb{R}^m$ and $\gamma_{k+1} \in \mathbb{R}^{r_k(nd_1+nd_2+r_k)}$ is the output of Algorithm 2 if the linear system of (76) is solved exactly.

2) *The vector $\gamma_{k+1} \in \mathbb{R}^{r_k(nd_1+nd_2+r_k)}$ corresponding to an iterative solution of (76) using N_{CG_inner} iterations of a conjugate gradient method*[9] *can be computed in $O(N_{CG\_inner} r_k T(m + n\log T + nr_kT))$ time. Thus, a parametrization of an approximation of the $(k+1)$-st iterate $\widetilde{\mathbf{X}}^{(k+1)}$ of Algorithm 1 can be computed in $O(N_{CG\_inner} r_k T(m + n\log T + nr_kT))$ time.*

The statement of Lemma A.9.2 enables Algorithm 2 to compute an *accurate* approximation of $\widetilde{\mathbf{X}}^{(k+1)}$ in $O(r_kT(m+n\log T + nr_kT))$ time in many situations, in particular, if $\widetilde{\mathbf{X}}^{(k)}$ is close to an image $\mathcal{Q}_T(\mathbf{A})$ of a transition operator $\mathbf{A}$ satisfying $P_\Omega(\mathcal{Q}_T(\mathbf{A})) = \mathbf{y}$ and if the normalized sampling operator $\mathcal{R}_\Omega$ associated to the sampling set Ω satisfies a local restricted isometry property (37), as in this case, it can be shown that the condition number of the matrix $\mathbf{M}$ of linear system (76) is a *small constant*. We do not provide the full proof for that statement as it amounts to a variant of [KMV21, Theorem 4.2] and its proof.

Proof of Lemma A.9.1. We recall from Lemma A.8 that

$$\begin{aligned}\widetilde{\mathbf{X}}^{(k+1)} &\\ &= \widetilde{W}_{\mathbf{H}_k}^{-1} P_\Omega^* \left(\lambda\,\mathrm{Id} + P_\Omega \widetilde{W}_{\mathbf{H}_k}^{-1} P_\Omega^*\right)^{-1}(\mathbf{y}) \\ &= \mathcal{D}^{-1}\left(\widetilde{W}_{\mathcal{G}}^{(k)}\right)^{-1} \widetilde{P}_\Omega^* \left(\lambda\,\mathrm{Id} + \widetilde{P}_\Omega \left(\widetilde{W}_{\mathcal{G}}^{(k)}\right)^{-1} \widetilde{P}_\Omega^*\right)^{-1}(\mathbf{y})\end{aligned} \tag{78}$$

using the notation $\widetilde{W}_{\mathcal{G}}^{(k)} := \mathcal{G}^* W_{\mathrm{H}_k} \mathcal{G}$ and $\widetilde{P}_\Omega := P_\Omega \mathcal{D}^{-1}$. The last inequality holds due to

$$\begin{aligned}\widetilde{W}_{\mathbf{H}_k}^{-1} &= (\mathcal{H}^* W_{\mathbf{H}_k} \mathcal{H})^{-1} = (\mathcal{D}\mathcal{G}^* W_{\mathbf{H}_k} \mathcal{G}\mathcal{D})^{-1} \\ &= \mathcal{D}^{-1}\left(\widetilde{W}_{\mathcal{G}}^{(k)}\right)^{-1} \mathcal{D}^{-1}\end{aligned}$$

Using (74) and $\mathcal{G}^*\mathcal{G} = \mathrm{Id}$, we can write

$$\begin{aligned}\left(\widetilde{W}_{\mathcal{G}}^{(k)}\right)^{-1} &= (\mathcal{G}^* W_{\mathbf{H}_k} \mathcal{G})^{-1} \\ &= \left(\mathcal{G}^*\left(P_{\mathbf{T}_k}\left(\mathbf{D}_{S_k} - \varepsilon_k^{-2}\mathbf{I}_{S_k}\right) P_{\mathbf{T}_k}^* + \varepsilon_k^{-2}\mathrm{Id}\right)\mathcal{G}\right)^{-1} \\ &= \left(\mathcal{G}^* P_{\mathbf{T}_k}\left(\mathbf{D}_{S_k} - \varepsilon_k^{-2}\mathbf{I}_{S_k}\right) P_{\mathbf{T}_k}^* \mathcal{G} + \varepsilon_k^{-2}\mathrm{Id}\right)^{-1}.\end{aligned}$$

For the next step, we recall the Sherman-Morrison-Woodbury formula [Woo50, FRW11], [HJ12, (0.7.4.1)], which states that for any invertible matrices $\mathbf{B}, \mathbf{C}$ and matrices $\mathbf{E}, \mathbf{F}$ of appropriate dimensions,

$$(\mathbf{B} + \mathbf{E}\mathbf{C}\mathbf{F}^*)^{-1} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{E}\left(\mathbf{C}^{-1} + \mathbf{F}^*\mathbf{B}^{-1}\mathbf{E}\right)^{-1}\mathbf{F}^*\mathbf{B}^{-1}. \tag{79}$$

Applying (79) for $\mathbf{B} = \varepsilon_k^{-2}\mathrm{Id}, \mathbf{E} = \mathcal{G}^* P_{\mathbf{T}_k}, \mathbf{F}^* = P_{\mathbf{T}_k}^* \mathcal{G}$ and $\mathbf{C} = \mathbf{D}_{S_k} - \varepsilon_k^{-2}\mathbf{I}_{S_k}$ yields then

$$\begin{aligned}\left(\widetilde{W}_{\mathcal{G}}^{(k)}\right)^{-1} &= \varepsilon_k^2\Big[\mathrm{Id} \\ &- \mathcal{G}^* P_{\mathbf{T}_k}\left(\varepsilon_k^{-2}\left(\mathbf{D}_{S_k} - \varepsilon_k^{-2}\mathbf{I}_{S_k}\right)^{-1} + P_{\mathbf{T}_k}^* \mathcal{G}\mathcal{G}^* P_{\mathbf{T}_k}\right)^{-1} P_{\mathbf{T}_k}^* \mathcal{G}\Big].\end{aligned} \tag{80}$$

Thus,

$$\left(\lambda\,\mathrm{Id} + \widetilde{P}_\Omega\left(\widetilde{W}_{\mathcal{G}}^{(k)}\right)^{-1} \widetilde{P}_\Omega^*\right)^{-1} = \varepsilon_k^{-2}\left(\lambda\varepsilon_k^{-2}\mathbf{I} + \widetilde{P}_\Omega \widetilde{P}_\Omega^* - \Xi\right)^{-1}$$

where

$$\Xi := \widetilde{P}_\Omega \mathcal{G}^* P_{\mathbf{T}_k}\left(\frac{\varepsilon_k^{-2}\mathbf{I}_{S_k}}{\mathbf{D}_{S_k} - \varepsilon_k^{-2}\mathbf{I}_{S_k}} + P_{\mathbf{T}_k}^* \mathcal{G}\mathcal{G}^* P_{\mathbf{T}_k}\right)^{-1} P_{\mathbf{T}_k}^* \mathcal{G} \widetilde{P}_\Omega^*.$$

Here, we can again use Sherman-Morrison-Woodbury (79) with $\mathbf{C} = \mathbf{N}^{-1}, \mathbf{E} = \mathbf{F} = \widetilde{\mathbf{E}}$ and $\mathbf{B} = \mathbf{K}$ with $\mathbf{K}$ from (77), $\widetilde{\mathbf{E}} = \widetilde{P}_\Omega \mathcal{G}^* P_{\mathbf{T}_k}$ and

$$\begin{aligned}\mathbf{N} &:= \frac{\mathbf{D}_{S_k}^{-1}}{\mathbf{D}_{S_k}^{-1} - \varepsilon_k^2\mathbf{I}_{S_k}} - P_{\mathbf{T}_k}^* \mathcal{G}\mathcal{G}^* P_{\mathbf{T}_k} \\ &= -\left(\frac{\varepsilon_k^{-2}\mathbf{I}_{S_k}}{\mathbf{D}_{S_k} - \varepsilon_k^{-2}\mathbf{I}_{S_k}} + P_{\mathbf{T}_k}^* \mathcal{G}\mathcal{G}^* P_{\mathbf{T}_k}\right)\end{aligned} \tag{81}$$

[9] or of related iterative solvers based on matrix-vector multiplication

to obtain

$$\begin{aligned}\widetilde{\mathbf{y}} &:= \left(\lambda \operatorname{Id} + \widetilde{P}_\Omega \left(\widetilde{W}_{\mathcal{G}}^{(k)}\right)^{-1} \widetilde{P}_\Omega^*\right)^{-1} (\mathbf{y}) \\ &= \varepsilon_k^{-2}\mathbf{K}^{-1}(\mathbf{y}) - \varepsilon_k^{-2}\mathbf{K}^{-1}\widetilde{\mathbf{E}}\left(\widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\widetilde{\mathbf{E}} + \mathbf{N}\right)^{-1} \widetilde{\mathbf{E}}^*\mathbf{K}^{-1}(\mathbf{y}) \\ &= \varepsilon_k^{-2}\mathbf{K}^{-1}\left(\mathbf{y} - \widetilde{\mathbf{E}}\gamma_{k+1}\right),\end{aligned} \tag{82}$$

using the notation that $\gamma_{k+1} \in S_k$ is solution to the invertible linear system

$$\left(\widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\widetilde{\mathbf{E}} + \mathbf{N}\right)\gamma_{k+1} = \widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\mathbf{y}$$

Furthermore, $\mathbf{y}, \widetilde{\mathbf{y}}$ and γ_{k+1} are related by

$$\begin{aligned}&\varepsilon_k^2\widetilde{\mathbf{E}}^*(\widetilde{\mathbf{y}}) \\ &= \widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\left(\mathbf{y} - \widetilde{\mathbf{E}}\gamma_{k+1}\right) \\ &= \widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\left(\mathbf{y} - \widetilde{\mathbf{E}}\left(\widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\widetilde{\mathbf{E}} + \mathbf{N}\right)^{-1}\widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\mathbf{y}\right) = \widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\mathbf{y} \\ &\quad - \left(\widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\widetilde{\mathbf{E}} + \mathbf{N} - \mathbf{N}\right)\left(\widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\widetilde{\mathbf{E}} + \mathbf{N}\right)^{-1}\widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\mathbf{y} \\ &= \mathbf{N}\left(\widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\widetilde{\mathbf{E}} + \mathbf{N}\right)^{-1}\widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\mathbf{y}.\end{aligned} \tag{83}$$

Inserting these equalities back into the expression (78) for $\widetilde{\mathbf{X}}^{(k+1)}$, we observe that

$$\begin{aligned}\widetilde{\mathbf{X}}^{(k+1)} &= \mathcal{D}^{-1}\big(\widetilde{W}_{\mathcal{G}}^{(k)}\big)^{-1}\widetilde{P}_\Omega^*\big(\lambda\operatorname{Id} + \widetilde{P}_\Omega\big(\widetilde{W}_{\mathcal{G}}^{(k)}\big)^{-1}\widetilde{P}_\Omega^*\big)^{-1}(\mathbf{y}) \\ &= \mathcal{D}^{-1}\big(\widetilde{W}_{\mathcal{G}}^{(k)}\big)^{-1}\widetilde{P}_\Omega^*(\widetilde{\mathbf{y}}) \\ &= \varepsilon_k^2\mathcal{D}^{-1}\big[\operatorname{Id} + \mathcal{G}^* P_{\mathbf{T}_k}\mathbf{N}^{-1}P_{\mathbf{T}_k}^*\mathcal{G}\big]\widetilde{P}_\Omega^*(\widetilde{\mathbf{y}}) \\ &= \varepsilon_k^2\mathcal{D}^{-1}\widetilde{P}_\Omega^*(\widetilde{\mathbf{y}}) + \varepsilon_k^2\mathcal{D}^{-1}\mathcal{G}^*P_{\mathbf{T}_k}\mathbf{N}^{-1}\widetilde{\mathbf{E}}^*(\widetilde{\mathbf{y}}) \\ &= \varepsilon_k^2\mathcal{D}^{-1}\widetilde{P}_\Omega^*(\widetilde{\mathbf{y}}) \\ &\quad + \varepsilon_k^2\mathcal{D}^{-1}\mathcal{G}^*P_{\mathbf{T}_k}\left(\widetilde{\mathbf{E}}^*\mathbf{K}^{-1}\widetilde{\mathbf{E}} + \mathbf{N}\right)^{-1}\widetilde{\mathbf{E}}^*\mathbf{K}^{-1}(\widetilde{\mathbf{y}}) \\ &= \varepsilon_k^2\mathcal{D}^{-1}\widetilde{P}_\Omega^*(\widetilde{\mathbf{y}}) + \varepsilon_k^2\mathcal{D}^{-1}\mathcal{G}^*P_{\mathbf{T}_k}(\gamma_{k+1}) \\ &= \mathcal{D}^{-1}\widetilde{P}_\Omega^*\mathbf{K}^{-1}\left(\mathbf{y} - \widetilde{E}\gamma_{k+1}\right) + \mathcal{D}^{-1}\mathcal{G}^*P_{\mathbf{T}_k}(\gamma_{k+1}) \\ &= \mathcal{D}^{-1}\widetilde{P}_\Omega^*(\mathbf{p}_{k+1}) + \mathcal{D}^{-1}\mathcal{G}^*P_{\mathbf{T}_k}(\gamma_{k+1}).\end{aligned}$$

In the second equality, we used the definition of $\widetilde{\mathbf{y}}$; in the third equality, we used that $\left(\widetilde{W}_{\mathcal{G}}^{(k)}\right)^{-1} = \varepsilon_k^2\left[\operatorname{Id} + \mathcal{G}^*P_{\mathbf{T}_k}\mathbf{N}^{-1}P_{\mathbf{T}_k}^*\mathcal{G}\right]$, which follows from (80) and (81). In the fifth equality, we used (83); in the sixth equality the definition of γ_{k+1}. In the seventh equality, we used (82) and in the last equality, we use the definition $\mathbf{p}_{k+1} = \mathbf{K}^{-1}\left(\mathbf{y} - P_\Omega\mathcal{D}^{-1}\mathcal{G}^*P_{\mathbf{T}_k}(\gamma_{k+1})\right) = \mathbf{K}^{-1}\left(\mathbf{y} - \widetilde{\mathbf{E}}(\gamma_{k+1})\right)$ of the residual $\mathbf{p}_{k+1}$. This concludes the proof of Lemma A.9. □

Proof of Lemma A.9.2. In line 3 of Algorithm 2, γ_{k+1} is computed by adding $\gamma_{k+1}^{(0)}$ (which is an input) and $\Delta\gamma_{k+1}$, the solution of the positive definite linear system of (76), which requires $O(r_k(nd_1 + nd_2 + r_k))$ computations.

To solve (76), a conjugate gradient method can be applied, whose cost crucially depends on the cost of executing matrix-vector multiplications with the matrix $\mathbf{M}$ from (75). We address the matrix-vector multiplication cost of the three summands of $\mathbf{M}$ separately, as we can obtain $\mathbf{M}\gamma$ for $\gamma \in \mathbb{R}^{r_k(nd_1+nd_2+r_k)}$ by adding/substracting the three resulting vectors in additional $3r_k(nd_1 + nd_2 + r_k)$ time.

- *Matrix-vector multiplication* with $\frac{\mathbf{D}_{S_k}^{-1}}{\mathbf{D}_{S_k}^{-1} - \varepsilon_k^2\mathbf{I}_{S_k}}$: this matrix is diagonaly, since $\mathbf{D}_{S_k}$ is, resulting in a time complexity of $r_k(nd_1 + nd_2 + r_k)$.
- *Matrix-vector* multiplication with $P_{\mathbf{T}_k}^*\mathcal{G}\mathcal{D}^{-1}P_\Omega^*\mathbf{K}^{-1}P_\Omega\mathcal{D}^{-1}\mathcal{G}^*P_{\mathbf{T}_k}$. This can me implemented by the successive application of the three operators $P_{\mathbf{T}_k}^*\mathcal{G}\mathcal{D}^{-1}P_\Omega^*$, $\mathbf{K}^{-1}$ and $P_\Omega\mathcal{D}^{-1}\mathcal{G}^*P_{\mathbf{T}_k}$. Applying $P_\Omega\mathcal{D}^{-1}\mathcal{G}^*P_{\mathbf{T}_k}$ can be done in $O(mTr_k + r_k^2nT) + mT = O(mTr_k + r_k^2nT)$ time by evaluating the tangent space matrix returned by $P_{\mathbf{T}_k}$ at mT locations (which correspond to the support set of $\mathcal{H}(P_\Omega^*(\mathbf{y}))$) via Algorithm 4 of the paper [KMV21] and averaging the entries across the Hankel blocks. $\mathbf{K}^{-1}$ can be applied by in $r_k(nd_1 + nd_2 + r_k)$ time as $P_\Omega\mathcal{D}^{-2}P_\Omega^*$ is a diagonal $(m \times m)$ matrix whose i-th diagonal entry is the inverse number of occurrences of the Hankel block which corresponds to the i-th observation in Ω. Finally, the application of $P_{\mathbf{T}_k}^*\mathcal{G}\mathcal{D}^{-1}P_\Omega^*$ can be implemented via Algorithm 3 of [KMV21] as the image of $\mathcal{G}\mathcal{D}^{-1}P_\Omega^*$ is a sparse $(nd_1 \times nd_2)$ matrix with a support set of size mT, giving a time complexity of $O(mTr_k + r_k^2nT)$. Thus, the total time complexity of the entire matrix-vector multiplication is $O(mTr_k + r_k^2nT)$.
- *Matrix-vector multiplication* with $P_{\mathbf{T}_k}^*\mathcal{G}\mathcal{G}^*P_{\mathbf{T}_k}$. We observe that block Hankel matrices of size $(nd_1 \times nd_2)$ with $(n \times n)$ blocks can be embedded into a $(nT \times nT)$ block circulant matrix (up to reordering of columns), cf., e.g., [KS99, Section 8.3.1], and such block circulant matrices can be diagonalized by a "block" discrete Fourier transform. Using the fast Fourier transform across blocks, it is possible to compute the image of $P_{\mathbf{T}_k}^*\mathcal{G}\mathcal{G}^*P_{\mathbf{T}_k}$ in $O(r_knT\log T + r_k^2T^2n)$ time, see also [Küm19, Section 3.4] for a related algorithm for Hankel matrices.

We refer to our MATLAB implementation for further details on the above. Overall, the matrix-vector multiplication of M with a vector $\gamma \in \mathbb{R}^{r_k(nd_1+nd_2+r_k)}$ can be performed in $O(r_kT(m + n\log T + nr_kT))$ time. Since the computations necessary to obtain $\mathbf{h}_k^0$ and $\mathbf{p}_{k+1}$ in line 1 and 4, respectively, involve only operations whose order we quantified above, this concludes the proof of Lemma A.9.2. □

References

[AC17] F. Andersson and M. Carlsson, *On the structure of positive semi-definite finite rank general domain Hankel and Toeplitz operators in several variables*, Complex Anal. Oper. Theory **11** (2017), no. 4, 755–784.

[ACC+17] A. Aldroubi, C. Cabrelli, A. F Cakmak, U. Molter, and A Petrosyan, *Iterative actions of normal operators*, J. Funct. Anal. **272** (2017), no. 3, 1121–1146.

[ACMT17] A. Aldroubi, C. Cabrelli, U. Molter, and S. Tang, *Dynamical sampling*, Appl. Comput. Harmon. Anal. **42** (2017), no. 3, 378–401.

[ADK13] A. Aldroubi, J. Davis, and I. Krishtal, *Dynamical sampling: Time–space trade-off*, Appl. Comput. Harmon. Anal. **34** (2013), no. 3, 495–503.

[AHO98] F. Alizadeh, J.-P. A Haeberly, and M. L Overton, *Primal-dual interior-point methods for semidefinite programming: convergence rates, stability and numerical results*, SIAM J. Optim. **8** (1998), no. 3, 746–768.

[AHP19] A. Aldroubi, L. Huang, and A. Petrosyan, *Frames induced by the action of continuous powers of an operator*, J. Math. Anal. Appl. **478** (2019), no. 2, 1059–1084.

[AK14] A. Aldroubi and I. Krishtal, *Krylov subspace methods in dynamical sampling*, arXiv preprint arXiv:1412.1538 (2014).

[BÅ70] R. Bellman and K. J. Åström, *On structural identifiability*, Math. Biosci. **7** (1970), no. 3-4, 329–339.

[BNZ21] J. Bauch, B. Nadler, and P. Zilber, *Rank 2r iterative least squares: efficient recovery of ill-conditioned low rank matrices from few entries*, SIAM J. Math. Data Sci. **3** (2021), no. 1, 439–465.

[BSW11] F. Bunea, Y. She, and M. H Wegkamp, *Optimal selection of reduced rank estimators of high-dimensional matrices*, Ann. Stat **39** (2011), no. 2, 1282–1309.

[CBSW15] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward, *Completing any low-rank matrix, provably*, J. Mach. Learn. Res. (JMLR) **16** (2015), no. 1, 2999–3034.

[CC14] Y. Chen and Y. Chi, *Robust Spectral Compressed Sensing via Structured Matrix Completion*, IEEE Trans. Inf. Theory **60** (2014), no. 10, 6576–6601.

[CCY22] H. Cai, J.-F. Cai, and J. You, *Structured gradient descent for fast robust low-rank hankel matrix completion*, arXiv preprint arXiv:2204.03316 (2022).

[CESV13] E. J. Candès, Y. Eldar, T. Strohmer, and V. Voroninski, *Phase Retrieval via Matrix Completion*, SIAM J. Imag. Sci. **6** (2013), no. 1, 199–225.

[Che15] Y. Chen, *Incoherence-Optimal Matrix Completion*, IEEE Trans. Inf. Theory **61** (2015), no. 5, 2909–2923.

[Chu97] F. R. K. Chung, *Spectral graph theory*, Vol. 92 of the CBMS Regional Conference Series in Mathematics, American Mathematical Society, 1997.

[CIML20] M. Coutino, E. Isufi, T. Maehara, and G. Leus, *State-space network topology identification from partial observations*, IEEE Trans. Signal Inf. Process. Netw. **6** (2020), 211–225.

[CLC19] Y. Chi, Y. M. Lu, and Y. Chen, *Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview*, IEEE Trans. Signal Process. **67** (2019), no. 20, 5239–5269.

[CLL$^+$05] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps*, Proc. Natl. Acad. Sci. U.S.A. **102** (2005), no. 21, 7426–7431.

[CM06] R. R. Coifman and M. Maggioni, *Diffusion wavelets*, Appl. Comp. Harm. Anal. **21** (2006), no. 1, 53–94.

[CP11a] E. J. Candès and Y. Plan, *Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements*, IEEE Trans. Inf. Theory **57** (2011), no. 4, 2342–2359.

[CP11b] E. J. Candès and Y. Plan, *Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements*, IEEE Trans. Inf. Theory **57** (2011), no. 4, 2342–2359.

[CR09] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math. **9** (2009), no. 6, 717–772.

[CT10] E. J. Candès and T. Tao, *The Power of Convex Relaxation: Near-Optimal Matrix Completion*, IEEE Trans. Inf. Theory **56** (2010), no. 5, 2053–2080.

[CT21] J. Cheng and S. Tang, *Estimate the spectrum of affine dynamical systems from partial observations of a single trajectory data*, Inverse Probl. **38** (2021), no. 1, 015004.

[CWW18] J.-F. Cai, T. Wang, and K. Wei, *Spectral compressed sensing via projected gradient descent*, SIAM J. Optim. **28** (2018), no. 3, 2625–2653.

[CWW19] ———, *Fast and provable algorithms for spectrally sparse signal reconstruction via low-rank Hankel matrix completion*, Appl. Comput. Harmon. Anal. **46** (2019), no. 1, 94–121.

[DC20] L. Ding and Y. Chen, *Leave-one-out approach for matrix completion: Primal and dual analysis*, IEEE Trans. Inf. Theory. **66** (2020), no. 11, 7274–7301.

[DDFG10] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, *Iteratively reweighted least squares minimization for sparse recovery*, Commun. Pure Appl. Math. **63** (2010), 1–38.

[DH11] T. A. Davis and Y. Hu, *The university of Florida sparse matrix collection*, ACM Trans. Math. Softw. **38** (2011), no. 1.

[DM16] J. A. Deri and J. M. F. Moura, *New York City taxi analysis with graph signal processing*, 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2016, pp. 1275–1279.

[DR16] M. A. Davenport and J. Romberg, *An Overview of Low-Rank Matrix Recovery From Incomplete Observations*, IEEE J. Sel. Topics Signal Process. **10** (2016), 608–622.

[DRS20] X Duan, J. Rubin, and D Swigon, *Identification of affine dynamical systems from a single trajectory*, Inverse Probl. **36** (2020), no. 8, 085004.

[DTFV16] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, *Learning laplacian matrix in smooth graph signal representations*, IEEE Trans. Signal Process. **64** (2016), no. 23, 6160–6173.

[DYC$^+$21] L. Ding, A. Yurtsever, V. Cevher, J. A Tropp, and M. Udell, *An optimal-storage approach to semidefinite programming using approximate complementarity*, SIAM J. Optim. **31** (2021), no. 4, 2695–2725.

[EPO18] H. E Egilmez, E. Pavez, and A. Ortega, *Graph learning from filtered signals: Graph system and diffusion kernel identification*, IEEE Trans. Signal Inf. Process. Netw. **5** (2018), no. 2, 360–374.

[ER59] P. Erdös and A. Rényi, *On random graphs*, Publicationes Mathematicae Debrecen **6** (1959), 290.

[EWW18] A. Eftekhari, M. B Wakin, and R. A Ward, *MC2: a two-phase algorithm for leveraged matrix completion*, Inf. Inference **7** (2018), no. 3, 581–604.

[Fat21] S. Fattahi, *Learning partially observed linear dynamical systems from logarithmic number of samples*, Proceedings of the 3rd Conference on Learning for Dynamics and Control, 2021, pp. 60–72.

[Faz02] M. Fazel, *Matrix rank minimization with applications*, Ph.D. Thesis, Electrical Engineering Department, Stanford University, 2002.

[FH96] S. Feldmann and G. Heinig, *Vandermonde factorization and canonical representations of block Hankel matrices*, Linear Algebra Appl. **241** (1996), 247–278.

[FHB03] M. Fazel, H. Hindi, and S. P. Boyd, *Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices*, Proc. Am. Control Conf., 2003, pp. 2156–2162.

[FLZZ20] J. Fan, R. Li, C.-H. Zhang, and H. Zou, *Statistical foundations of data science*, Chapman and Hall/CRC, 2020.

[Fou18] S. Foucart, *Concave Mirsky Inequality and Low-Rank Recovery*, SIAM J. Matrix Anal. Appl. **39** (2018), no. 1, 99–103.

[FPST13] M. Fazel, T. K. Pong, D. Sun, and P. Tseng, *Hankel matrix rank minimization with applications to system identification and realization*, SIAM J. Matrix Anal. Appl. **34** (2013), no. 3, 946–977.

[FR13] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Applied and Numerical Harmonic Analysis, Springer New York, 2013.

[FRW11] M. Fornasier, H. Rauhut, and R. Ward, *Low-rank matrix recovery via iteratively reweighted least squares minimization*, SIAM J. Optim. **21** (2011), no. 4, 1614–1640.

[Gil59] E. N. Gilbert, *Random Graphs*, Ann. Math. Stat. **30** (1959), no. 4, 1141 –1144.

[GRG18] C. Grussler, A. Rantzer, and P. Giselsson, *Low-rank optimization with convex constraints*, IEEE Trans. Automat. Contr. **63** (2018), no. 11, 4000–4007.

[Gro11] D. Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Trans. Inf. Theory. **57** (2011), no. 3, 1548–1566.

[GVRH20] P. Giampouras, R. Vidal, A. Rontogiannis, and B. Haeffele, *A novel variational form of the Schatten-p quasi-norm*, Advances in Neural Information Processing Systems 33 (NeurIPS), 2020.

[HJ12] R. A. Horn and C. R. Johnson, *Matrix analysis*, 2nd ed., Cambridge University Press, 2012.

[HK66] B. Ho and R. E Kálmán, *Effective construction of linear state-variable models from input/output functions*, at-Automatisierungstechnik **14** (1966), no. 1-12, 545–548.

[HS17] R. Heckel and M. Soltanolkotabi, *Generalized line spectral estimation via convex optimization*, IEEE Trans. Inf. Theory. **64** (2017), no. 6, 4001–4023.

[HS52] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Natl. Bur. Stand. (U. S.) **49** (1952), no. 1.

[HS90] Y. Hua and T. K. Sarkar, *Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise*, IEEE Trans. Signal Process. **38** (1990), no. 5, 814–824.

[HW77] P. W. Holland and R. E. Welsch, *Robust regression using iteratively reweighted least-squares*, Commun. Stat. Theory Methods **6** (1977), 813–827.

[IRG18] V. N Ioannidis, D. Romero, and G. B Giannakis, *Inference of spatio-temporal functions over graphs via multikernel kriged kalman filtering*, IEEE Trans. Signal Process. **66** (2018), no. 12, 3228–3239.

[ISG18] V. N. Ioannidis, Y. Shen, and G. B. Giannakis, *Semi-blind inference of topologies and signals over graphs*, 2018 IEEE Data Science Workshop (DSW), 2018, pp. 165–169.

[JLY16] K. H. Jin, D. Lee, and J. C. Ye, *A general framework for compressed sensing and parallel MRI using annihilating filter based low-rank Hankel matrix*, IEEE Trans. Comput. Imag. **2** (2016), no. 4, 480–495.

[KAB+19] S. P Kolodziej, M. Aznaveh, M. Bullock, J. David, T. A Davis, M. Henderson, Y. Hu, and R. Sandstrom, *The SuiteSparse matrix collection website interface*, J. Open Source Softw. **4** (2019), no. 35, 1244.

[Kal16] V. Kalofolias, *How to learn a graph from smooth signals*, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS), 2016, pp. 920–929.

[KBBP16] J N. Kutz, S. L Brunton, B. W Brunton, and J. L Proctor, *Dynamic mode decomposition: data-driven modeling of complex systems*, SIAM, 2016.

[KBV09] Y. Koren, R. Bell, and C. Volinsky, *Matrix Factorization Techniques for Recommender Systems*, Computer **42** (2009), no. 8, 30–37.

[KKM22] F. Krahmer, C. Kümmerle, and O. Melnyk, *On the robustness of noise-blind low-rank recovery from rank-one measurements*, Linear Algebra Appl. **652** (2022), 37–81.

[KKMV22] C. Kümmerle, F. Krahmer, T. Masák, and C. M. Verdun, *Optimal Weight Operators of Spectral Functions for Fast Low-Rank Optimization*, 2022. in preparation.

[Klo11] O. Klopp, *Rank penalized estimators for high-dimensional matrices*, Electron. J. Stat. **5** (2011), 1161–1183.

[KMO10] R. H. Keshavan, A. Montanari, and S. Oh, *Matrix Completion From a Few Entries*, IEEE Trans. Inf. Theory **56** (2010), no. 6, 2980–2998.

[KMV21] C. Kümmerle and C. Mayrink Verdun, *A Scalable Second Order Method for Ill-Conditioned Matrix Completion from Few Samples*, Proceedings of the International Conference on Machine Learning (ICML), 2021.

[KS18] C. Kümmerle and J. Sigl, *Harmonic Mean Iteratively Reweighted Least Squares for Low-Rank Matrix Recovery*, J. Mach. Learn. Res. (JMLR) **19** (2018), no. 47, 1–49.

[KS99] T. Kailath and A. H Sayed, *Fast reliable algorithms for matrices with structure*, SIAM, 1999.

[Küm19] C. Kümmerle, *Understanding and enhancing data recovery algorithms: From noise-blind sparse recovery to reweighted methods for low-rank matrix optimization*, Ph.D. Thesis, Department of Mathematics, Technical University of Munich, 2019.

[KV19] C. Kümmerle and C. M Verdun, *Completion of structured low-rank matrices via iteratively reweighted least squares*, 2019 13th international conference on sampling theory and applications (sampta), 2019, pp. 1–5.

[Lan16] K. Lange, *MM Optimization Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016.

[LF16] W. Liao and A. Fannjiang, *Music for single-snapshot spectral estimation: Stability and super-resolution*, Appl. Comput. Harmon. Anal. **40** (2016), no. 1, 33–67.

[Liu11] Y.-K. Liu, *Universal low-rank matrix recovery from Pauli measurements*, Advances in Neural Information Processing Systems 24 (NIPS), 2011, pp. 1638–1646.

[Lju98] L. Ljung, *System Identification*, Signal Analysis and Prediction, 1998, pp. 163–173.

[LLJY18] K. Lee, Y. Li, K. H. Jin, and J. C. Ye, *Unified Theory for Recovery of Sparse Signals in a General Transform Domain*, IEEE Trans. Inf. Theory **64** (2018), no. 8, 5457–5477.

[LT19] C.-K. Lai and S. Tang, *Undersampled windowed exponentials and their applications*, Acta Appl. Math. **164** (2019), no. 1, 65–81.

[LTYL15] C. Lu, J. Tang, S. Yan, and Z. Lin, *Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm*, IEEE Trans. Image Process. **25** (2015), no. 2, 829–839.

[LV09] Y. M Lu and M. Vetterli, *Spatial super-resolution of a diffusion field by temporal oversampling in sensor networks*, 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2009, pp. 2249–2252.

[Mar19] I. Markovsky, *Structured low-rank approximation: Algorithms, implementation, applications*, 2nd ed., Springer International Publishing, 2019.

[Meu06] G. Meurant, *The Lanczos and Conjugate Gradient Algorithms: From Theory to Finite Precision Computations*, Society for Industrial and Applied Mathematics (SIAM), 2006.

[MF12] K. Mohan and M. Fazel, *Iterative reweighted algorithms for matrix rank minimization*, J. Mach. Learn. Res. (JMLR) **13** (2012), no. 1, 3441–3473.

[Mir60] L. Mirsky, *Symmetric Gauge Functions And Unitarily Invariant Norms*, Q. J. Math. **11** (1960), no. 1, 50–59.

[MM15] C. Musco and C. Musco, *Randomized Block Krylov Methods for Stronger and Faster Approximate Singular Value Decomposition*, Advances in Neural Information Processing Systems 28 (NIPS), 2015, pp. 1396–1404.

[MSMR19] G. Mateos, S. Segarra, A. G Marques, and A. Ribeiro, *Connecting the dots: Identifying network structure via graph signal processing*, IEEE Signal Process. Mag. **36** (2019), no. 3, 16–43.

[MSW20] R. Mazumder, D. Saldana, and H. Weng, *Matrix completion with nonconvex regularization: Spectral operators and scalable algorithms*, Stat. Comput. (2020), 1–26.

[MTF17] H. P. Maretic, D. Thanou, and P. Frossard, *Graph learning under sparsity priors*, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, 2017, pp. 6523–6527.

[MU13] I. Markovsky and K. Usevich, *Structured low-rank approximation with missing data*, SIAM J. Matrix Anal. Appl. **34** (2013), no. 2, 814–830.

[MWCC20] C. Ma, K. Wang, Y. Chi, and Y. Chen, *Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution*, Found. Comput. Math **20** (2020), 451–632.

[OJ17] G. Ongie and M. Jacob, *A Fast Algorithm for Convolutional Structured Low-rank Matrix Recovery*, IEEE Trans. Comput. Imaging **3** (2017), no. 4, 535–550.

[OO19] S. Oymak and N. Ozay, *Non-asymptotic Identification of LTI Systems from a Single Trajectory*, 2019 American Control Conference (ACC), 2019, pp. 5655–5661.

[OPAB+21] G. Ongie, D. Pimentel-Alarcón, L. Balzano, R. Willett, and R. D Nowak, *Tensor methods for nonlinear matrix completion*, SIAM J. Math. Data Sci. **3** (2021), no. 1, 253–279.

[OWNB17] G. Ongie, R. Willett, R. D Nowak, and L. Balzano, *Algebraic variety models for high-rank matrix completion*, Proceedings of the International Conference on Machine Learning (ICML), 2017, pp. 2691–2700.

[PGM+16] B. Pasdeloup, V. Gripon, G. Mercier, D. Pastor, and M. G Rabbat, *Characterization and inference of weighted graph topologies from observations of diffused signals*, arXiv preprint arXiv:1605.02569 (2016).

[PGM+17] ______, *Characterization and inference of graph diffusion processes from observations of stationary signals*, IEEE Trans. Signal Inf. Process. Netw. **4** (2017), no. 3, 481–496.

[PIM10] J. Pereira, M. Ibrahimi, and A. Montanari, *Learning Networks of Stochastic Differential Equations*, Advances in Neural Information Processing Systems 23 (NIPS), 2010.

[PKCS18] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, *Finding Low-Rank Solutions via Nonconvex Matrix Factorization, Efficiently and Provably*, SIAM J. Imaging Sci. **11** (2018), no. 4, 2165–2204.

[PP13] T. Peter and G. Plonka, *A generalized Prony method for reconstruction of sparse sums of eigenfunctions of linear operators*, Inverse Probl. **29** (2013), no. 2, 025001.

[PPS+14] N. Perraudin, J. Paratte, D. Shuman, L. Martin, V. Kalofolias, P. Vandergheynst, and D. K Hammond, *Gspbox: A toolbox for signal processing on graphs*, arXiv preprint arXiv:1408.5781 (2014).

[PT10] D. Potts and M. Tasche, *Parameter estimation for exponential sums by approximate prony method*, Signal Process. **90** (2010), no. 5, 1631–1642.

[Qin06] S J. Qin, *An overview of subspace identification*, Computers & chemical engineering **30** (2006), no. 10-12, 1502–1513.

[Rec11] B. Recht, *A Simpler Approach to Matrix Completion*, J. Mach. Learn. Res. (JMLR) **12** (2011), 3413–3430.

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-*

rank solutions of linear matrix equations via nuclear norm minimization, SIAM Rev. **52** (2010), no. 3, 471–501.

[RK89] R. Roy and T. Kailath, *Esprit-estimation of signal parameters via rotational invariance techniques*, IEEE. Trans. Acoust. **37** (1989), no. 7, 984–995.

[Sch10] P. J Schmid, *Dynamic mode decomposition of numerical and experimental data*, J. Fluid Mech. **656** (2010), 5–28.

[SL16] R. Sun and Z. Q. Luo, *Guaranteed Matrix Completion via Non-Convex Factorization*, IEEE Trans. Inf. Theory **62** (2016), no. 11, 6535–6579.

[SMMR16] S. Segarra, G. Mateos, A. G Marques, and A. Ribeiro, *Blind identification of graph filters*, IEEE Trans. Signal Process. **65** (2016), no. 5, 1146–1159.

[SMMR17] S. Segarra, A. G Marques, G. Mateos, and A. Ribeiro, *Network topology inference from spectral templates*, IEEE Trans. Signal Inf. Process. Netw. **3** (2017), no. 3, 467–483.

[SOF20] Y. Sun, S. Oymak, and M. Fazel, *Finite sample system identification: improved rates and the role of regularization*, Proceedings of the 2nd Conference on Learning for Dynamics and Control, 2020, pp. 16–25.

[Spo10] O. Sporns, *Networks of the brain*, MIT press, 2010.

[SR19] T. Sarkar and A. Rakhlin, *Near optimal finite time identification of arbitrary linear dynamical systems*, Proceedings of the International Conference on Machine Learning (ICML), 2019, pp. 5610–5618.

[SRS14] S Stanhope, J. E Rubin, and D. Swigon, *Identifiability of linear and linear-in-parameters dynamical systems from a single trajectory*, SIAM J. Appl. Dyn. **13** (2014), no. 4, 1792–1815.

[Ste06] M. Stewart, *Perturbation of the SVD in the presence of small singular values*, Linear Algebra Appl. **419** (2006), no. 1, 53–77.

[STYZ20] D. Sun, K.-C. Toh, Y. Yuan, and X.-Y. Zhao, *Sdpnal+: A matlab software for semidefinite programming with bound constraints (version 1.0)*, Optim. Methods Softw. **35** (2020), no. 1, 87–115.

[Tan17a] S. Tang, *System identification in dynamical sampling*, Adv. Comput. Math. **43** (2017), no. 3, 555–580.

[Tan17b] ______, *Universal spatiotemporal sampling sets for discrete spatially invariant evolution processes*, IEEE Trans. Inf. Theory . **63** (2017), no. 9, 5518–5528.

[TBPR17] S. Tu, R. Boczar, A. Packard, and B. Recht, *Non-asymptotic analysis of robust control from coarse-grained identification*, arXiv preprint arXiv:1707.04791 (2017).

[TDKF17] D. Thanou, X. Dong, D. Kressner, and P. Frossard, *Learning heat diffusion graphs*, IEEE Trans. Signal Inf. Process. Netw. **3** (2017), no. 3, 484–499.

[Tis92] M. Tismenetsky, *Factorizations of hermitian block hankel matrices*, Linear Algebra Appl. **166** (1992), 45–63.

[UZ21] A. Ulanovskii and I. Zlotnikov, *Reconstruction of bandlimited functions from space–time samples*, J. Funct. Anal. **280** (2021), no. 9, 108962.

[Van13] B. Vandereycken, *Low-Rank Matrix Completion by Riemannian Optimization*, SIAM J. Optim. **23** (2013), no. 2, 1214–1236.

[VD17] S. Voronin and I. Daubechies, *An iteratively reweighted least squares algorithm for sparse regularization*, Functional Analysis, Harmonic Analysis, and Image Processing, 2017, pp. 391–411.

[Ver18] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2018.

[WCCL20] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung, *Guarantees of Riemannian optimization for low rank matrix completion*, Inverse Probl. Imaging **14** (2020), no. 2, 233–265.

[Wed72] P.-Å. Wedin, *Perturbation bounds in connection with singular value decomposition*, BIT **12** (1972), no. 1, 99–111.

[WMNO06] R. A Waltz, J. L. Morales, J. Nocedal, and D. Orban, *An interior algorithm for nonlinear optimization that combines line search and trust region steps*, Math. Program. **107** (2006), no. 3, 391–408.

[Woj10] P. Wojtaszczyk, *Stability and instance optimality for Gaussian measurements in compressed sensing*, Found. Comput. Math. **10** (2010), no. 1, 1–13.

[Woo50] M. A. Woodbury, *Inverting modified matrices*, Memorandum Rept. 42, Statistical Research Group (1950).

[YGL18] Q. Yuan, M. Gu, and B. Li, *Superlinear convergence of randomized block Lanczos algorithm*, 2018 IEEE International Conference on Data Mining (ICDM), 2018, pp. 1404–1409.

[YKJL17] J. C. Ye, J. M. Kim, K. H. Jin, and K. Lee, *Compressive Sampling Using Annihilating Filter-Based Low-Rank Interpolation*, IEEE Trans. Inf. Theory **63** (2017), no. 2, 777–801.

[YTF$^+$21] A. Yurtsever, J. A Tropp, O. Fercoq, M. Udell, and V. Cevher, *Scalable semidefinite programming*, SIAM J. Math. Data Sci. **3** (2021), no. 1, 171–200.

[YXS16] Z. Yang, L. Xie, and P. Stoica, *Vandermonde decomposition of multilevel toeplitz matrices with application to multidimensional super-resolution*, IEEE Trans. Inf. Theory. **62** (2016), no. 6, 3685–3701.

[ZL20] Y. Zheng and N. Li, *Non-asymptotic identification of linear dynamical systems using multiple trajectories*, IEEE Contr. Syst. Lett. **5** (2020), no. 5, 1693–1698.

[ZN22] P. Zilber and B. Nadler, *Gnmr: A provable one-line algorithm for low rank matrix recovery*, SIAM J. Math. Data Sci. **4** (2022), no. 2, 909–934.

[ZWSP08] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, *Large-Scale Parallel Collaborative Filtering for the Netflix Prize*, Algorithmic Aspects in Information and Management, 2008, pp. 337–348.