Supplementary material for *Recovering Simultaneously Structured Data via Non-Convex Iteratively Reweighted Least Squares*

This supplement is structured as follows.

- Appendix A presents some details about the experimental setup as well as additional numerical experiments.
- Appendix B.1 presents the proof of Theorem 2.5.
- Appendix B.2 presents the proof of Theorem 2.6.
- Appendix C details some technical results that are used in Appendices B.1 and B.2.

A Experimental Setup and Supplementary Experiments

In this section, we elaborate on the detailed experimental setup that was used in Section 4 of the main paper. Furthermore, we provide additional experiments comparing the behavior of the three methods studied in Section 4 for linear measurement operators $\mathcal{A}$ that are closer to operators that can be encountered in applications of simultaneous low-rank and group-sparse recovery. Finally, we shed light on the evolution of the objective function (4) of `IRLS` (Algorithm 1), including in situations where the algorithm does not manage to recover the ground truth.

A.1 Experimental Setup

The experiments of Section 4 were conducted using MATLAB implementations of the three algorithms on different Linux machines using MATLAB versions R2019b or R2022b. In total, the preparation and execution of the experiments used approximately 1200 CPU hours. The CPU models used in the simulations are Dual 18-Core Intel Xeon Gold 6154, Dual 24-Core Intel Xeon Gold 6248R, Dual 8-Core Intel Xeon E5-2667, 28-Core Intel Xeon E5-2690 v3, 64-Core Intel Xeon Phi KNL 7210-F. For `Sparse Power Factorization (SPF)` [60], we used our custom implementation of [60, Algorithm 4 "rSPF_HTP"] and for `Riemannian adaptive iterative hard thresholding (RiemAdaIHT)` [29], we used an implementation provided to us by Max Pfeffer in private communications. We refer to Section 3 for implementation details for the `IRLS` method Algorithm 1.

In all phase transition experiments, we define *successful recovery* such that the relative Frobenius error $\frac{\|\mathbf{X}^{(K)}-\mathbf{X}_\star\|_F}{\|\mathbf{X}_\star\|_F}$ of the iterate $\mathbf{X}^{(K)}$ returned by the algorithm relative to the simultaneously low-rank and row-sparse ground truth matrix $\mathbf{X}_\star$ is smaller than the threshold 10^{-4}. As stopping criteria, we used the criterion that the relative change of Frobenius norm satisfies $\frac{\|\mathbf{X}^{(k)}-\mathbf{X}^{(k-1)}\|_F}{\|\mathbf{X}^{(k)}\|_F} < \mathrm{tol}$ for `IRLS`, the change in the matrix factors norms satisfy $\|\mathbf{U}_k - \mathbf{U}_{k-1}\| < \mathrm{tol}$ and $\|\mathbf{V}_k - \mathbf{V}_{k-1}\| < \mathrm{tol}$ for `SPF`, and the norm of the Riemannian gradient in `RiemAdaIHT` being smaller than tol for $\mathrm{tol} = 10^{-10}$, or if a maximal number of iterations is reached. This iteration threshold was chosen as $\mathrm{max\_iter} = 250$ for `IRLS` and `SPF` and as $\mathrm{max\_iter} = 2000$ for `RiemAdaIHT`, reflecting the fact that `RiemAdaIHT` is a gradient-type method which might need many iterations to reach a high-accuracy solution. The parameters were chosen so that the stopping criteria do not prevent a method's iterates reaching the recovery threshold if they were to reach $\mathbf{X}_\star$ eventually.

In the experiments, we chose random ground truths $\mathbf{X}_\star \in \mathbb{R}^{n_1 \times n_2}$ of rank r and row-sparsity s such that $\mathbf{X}_\star = \tilde{\mathbf{X}}_\star / \left\|\tilde{\mathbf{X}}_\star\right\|_F$, where $\tilde{\mathbf{X}}_\star = \mathbf{U}_\star \mathrm{diag}(\mathbf{d}_\star)\mathbf{V}_\star^*$, and where $\mathbf{U}_\star \in \mathbb{R}^{n_1 \times r}$ is a matrix with s non-zero rows whose location is chosen uniformly at random and whose entries are drawn from i.i.d. standard Gaussian random variables, $\mathbf{d}$ has i.i.d. standard Gaussian entries and $\mathbf{V}_\star \in \mathbb{R}^{n_2 \times r}$ has likewise i.i.d. standard Gaussian entries.

A.2 Random Rank-One Measurements

In Section 4, we considered only measurement operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ whose matrix representation consists of i.i.d. Gaussian entries, i.e., operators such that there are independent matrices

$\mathbf{A}_1, \dots \mathbf{A}_m$ with i.i.d. standard Gaussian entries such that

$$\mathcal{A}(\mathbf{X})_j = \langle \mathbf{A}_j, \mathbf{X} \rangle_F$$

for any $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$. While it is known that such Gaussian measurement operators satisfy the (r, s)-RIP of Section 4, which is the basis of our convergence theorem Theorem 2.5, in a regime of a near-optimal number of measurements with high probability, practically relevant measurement operators are often more structured; another downside of dense Gaussian measurements is that it is computationally expensive to implement their action on matrices.

In relevant applications of our setup, however, e.g., in sparse phase retrieval [46, 11, 47] or blind deconvolution [59, 83], the measurement operator consists of rank-one measurements. For this reason, we now conduct experiments in settings related to the ones depicted in Figure 1 and Figure 2 Section 4, but for *random rank-one measurements* where the action of $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ on $\mathbf{X}$ can be written as

$$\mathcal{A}(\mathbf{X})_j = \langle \mathbf{a}_j \mathbf{b}_j^*, \mathbf{X} \rangle_F$$

for each $j = 1, \dots, m$, where $\mathbf{a}_j, \mathbf{b}_j$ are independent random standard Gaussian vectors. In Figure 4, we report the phase transition performance of `RiemAdaIHT`, `SPF` and `IRLS` for (256×40)-dimensional ground truths of different row-sparsities and different ranks if we are given such random rank-one measurements.

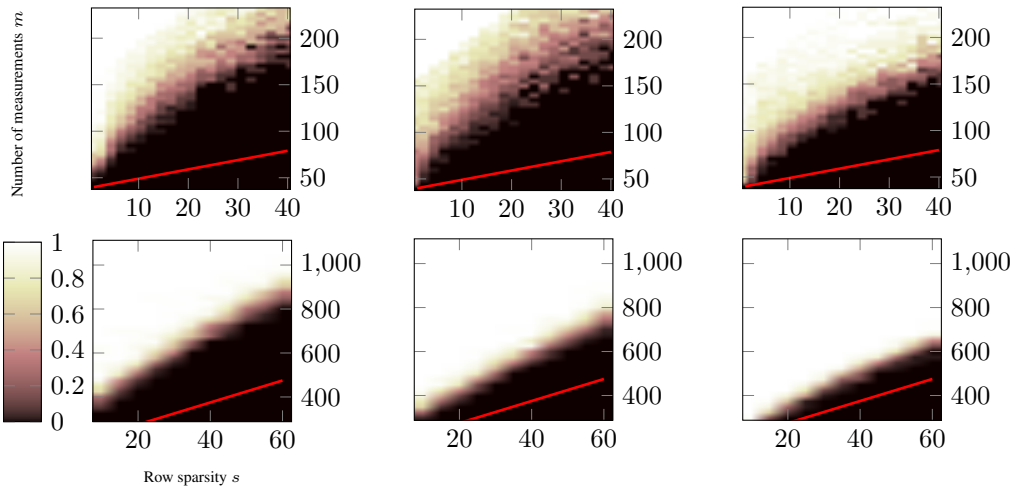

Figure 4: Left column: `RiemAdaIHT`, center: `SPF`, right: `IRLS`. Success rates for the recovery of low-rank and row-sparse matrices from random rank-one measurements. First row: Rank-1 ground truth $\mathbf{X}_\star$ (cf. Figure 1. Second row: Rank-5 ground truth $\mathbf{X}_\star$ (cf. Figure 2).

We observe in Figure 4 that compared to the setting of dense Gaussian measurements, the phase transitions of all three algorithms deteriorate slightly; especially for $r = 1$, one can observe that the transition between no success and high empirical success rate is extends across a larger area. `IRLS` performs clearly best for both $r = 1$ and $r = 5$, whereas `SPF` has the second best performance for $r = 5$. For $r = 1$, it is somewhat unclear whether `RiemAdaIHT` or `SPF` performs better.

A.3 Discrete Fourier Rank-One Measurements

We now revisit the experiments of Appendix A.2 for a third measurement setup motivated from blind deconvolution problems [3, 59, 64, 66, 83, 29], which are prevalent in astronomy, medical imaging and communications engineering [49, 13]. In particular, in these settings, if $\mathbf{z} \in \mathbb{R}^m$ is an (unknown) signal and $\mathbf{w} \in \mathbb{R}^m$ is an (unknown) convolution kernel, assume we are given the entries of their convolution $\widetilde{y} = \mathbf{z} * \mathbf{w}$. If we know that $\mathbf{z} = \mathbf{A}\mathbf{u}$ for some known matrix $\mathbf{A} \in \mathbb{R}^{m \times n_1}$ and an s-sparse vector $\mathbf{u} \in \mathbb{R}^{n_1}$ and $\mathbf{w} = \mathbf{B}\mathbf{v}$ for some known matrix $\mathbf{B} \in \mathbb{R}^{m \times n_2}$ and arbitrary vector $\mathbf{v} \in \mathbb{R}^{n_2}$, applying the discrete Fourier transform (represented via the DFT matrix $\mathbf{F} \in \mathbb{C}^{m \times m}$), we can write the coordinates of

$$\mathbf{y} = \mathbf{F}\widetilde{y} = \operatorname{diag}(\mathbf{F}\mathbf{z})\mathbf{F}\mathbf{w} = \operatorname{diag}(\mathbf{F}\mathbf{A}\mathbf{u})\mathbf{F}\mathbf{B}\mathbf{v}$$

as

$$\mathbf{y}_j = \mathcal{A}(\mathbf{u}\mathbf{v}^*)_j = \langle (\mathbf{F}\mathbf{A})^*_{j,:} \overline{\mathbf{F}\mathbf{B}}_{j,:}, \mathbf{u}\mathbf{v}^* \rangle_F$$

for each $j = 1, \dots, m$, which allows us to write the problem as a simultaneously rank-1 and s-row sparse recovery problem from Fourier-type measurements.

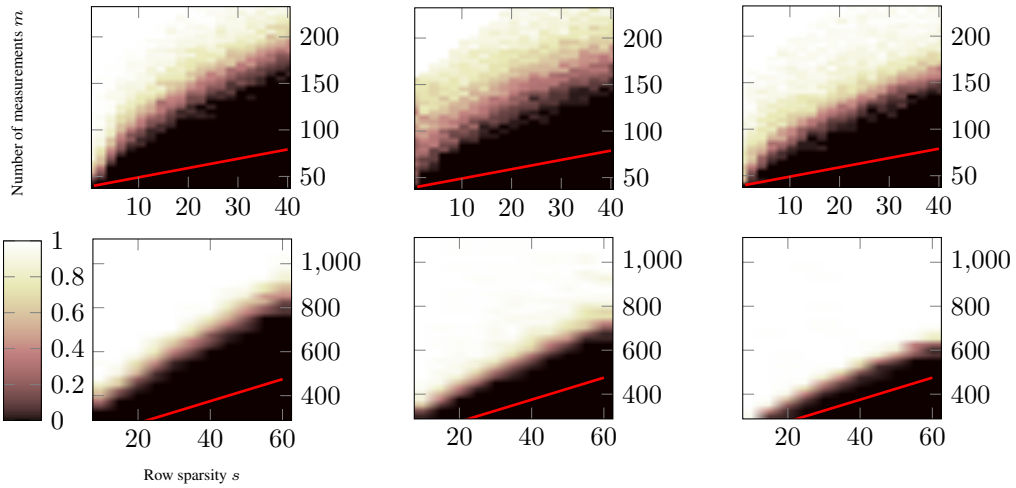

Figure 5: Left column: `RiemAdaIHT`, center: `SPF`, right: `IRLS`. Success rates for the recovery of low-rank and row-sparse matrices from Fourier rank-one measurements. First row: Rank-1 ground truth $\mathbf{X}_\star$. Second row: Rank-5 ground truth $\mathbf{X}_\star$ (cf. Figure 2).

In Figure 5, we report the results of simulations with $\mathbf{A}$ and $\mathbf{B}$ chosen generically as standard real Gaussians for these Fourier-based rank-1 measurements (including for rank-5 ground truths, which goes beyond a blind deconvolution setting). We observe that the transition from no recovery to exact recovery for an increasing number of measurement (with fixed dimension parameters s, n_1 and n_2) happens earlier than for the random Gaussian rank-one measurements of Appendix A.2, but slightly later than for dense Gaussian measurements. Again, `IRLS` exhibits the best empirical data-efficiency with sharpest phase transition curves.

As a summary, we observe that `IRLS` is able to recovery simultaneously low-rank and row-sparse matrices empirically from fewer measurements than state-of-the-art methods for a variety of linear measurement operators, including in cases where the RIP assumption of Definition 2.4 is not satisfied and in cases that are relevant for applications.

A.4 Evolution of Objective Values

While Theorem 2.5 guarantees local convergence if the measurement operator $\mathcal{A}$ is generic enough and contains enough measurements (RIP-assumption), it is instructive to study the behavior of Algorithm 1 in situations where there are *not* enough measurements available to identify a specific low-rank and row-sparse ground truth $\mathbf{X}_\star$ which respect to which the measurements have been taken.

In this setting, Theorem 2.6 guarantees that the behavior of the `IRLS` methods is still benign as the sequence of ε- and δ-smoothed log-objectives $\left(\mathcal{F}_{\varepsilon_k,\delta_k}(\mathbf{X}^{(k)})\right)_{k\geq 1}$ from (4) is non-increasing. In Figure 6, we illustrate the evolution of the relative Frobenius error of an iterate to the ground truth $\mathbf{X}_\star$, the $(\varepsilon_k, \delta_k)$-smoothed logarithmic surrogates $\mathcal{F}_{\varepsilon_k,\delta_k}(\mathbf{X}^{(k)})$ as well as of the rank and row-sparsity parts $\mathcal{F}_{lr,\varepsilon_k}(\mathbf{X}^{(k)})$ and $\mathcal{F}_{sp,\delta_k}(\mathbf{X}^{(k)})$ of the objective, respectively, in two typical situations.

In particular, we can see the evolution of these four quantities in the setting of data of dimensionality $n_1 = 128$, $n_2 \in \{20, 40\}$, $s = 20$ and $r = 5$ created as in the other experiments, where a number of $m = 875$ and $m = 175$ (corresponding to an oversampling factor of 3.0 and 1.0, respectively) dense Gaussian measurements are provided to Algorithm 1.

In the left plot of Figure 6, which corresponds to setting of abundant measurements, we observe that the four quantities all track each other relatively well on a semilogarithmic scale (note that we plot the square roots of the objective values to match the order of the (unsquared) relative Frobenius error),

converging to values between 10^{-13} and 10^{-11} (at which point the stopping criterion of the method applies) within 12 iterations.

In the second plot of Figure 6, the number of measurements exactly matches the number of degrees of freedom of the ground truth, in which case the $\mathbf{X}^{(k)}$ does *not* converge to $\mathbf{X}_\star$. However, it can be seen that Algorithm 1 still finds very meaningful solutions: It can be seen that within 86 iterations, $\mathcal{F}_{\varepsilon_k,\delta_k}(\mathbf{X}^{(k)})$ converges to $\approx 10^{-12}$ (since $\sqrt{\mathcal{F}_{\varepsilon_k,\delta_k}(\mathbf{X}^{(k)})} \approx 10^{-6}$) in a manner that is partially "staircase-like": After 20 initial iterations where $\mathcal{F}_{\varepsilon_k,\delta_k}(\mathbf{X}^{(k)})$ decreases significantly at each iteration, its decrease is dominated by relatively sudden, alternating drops of the (blue) sparsity objective $\mathcal{F}_{sp,\delta_k}(\mathbf{X}^{(k)})$ and the (red) rank objective $\mathcal{F}_{lr,\varepsilon_k}(\mathbf{X}^{(k)})$, which typically do not occur simultaneously.

This illustrates the *self-balancing* property of the two objective terms in the IRLS objective $\mathcal{F}_{\varepsilon_k,\delta_k}(\mathbf{X}^{(k)})$: while the final iterate at iteration $k = 86$ is not of the target row-sparsity $s = 20$ and $r = 5$, it is still 20-row sparse and has essentially rank 6. This means that Algorithm 1 has found an alternative parsimonious solution to the simultaneous low-rank and row-sparse recovery problem that is just slightly less parsimonious.

Arguably, this robust performance in the low-data regime of `IRLS` is rather unique, and to the best of our knowledge, not shared by methods such as `SPF` or `RiemAdaIHT`, which typically breakdown in such a regime.

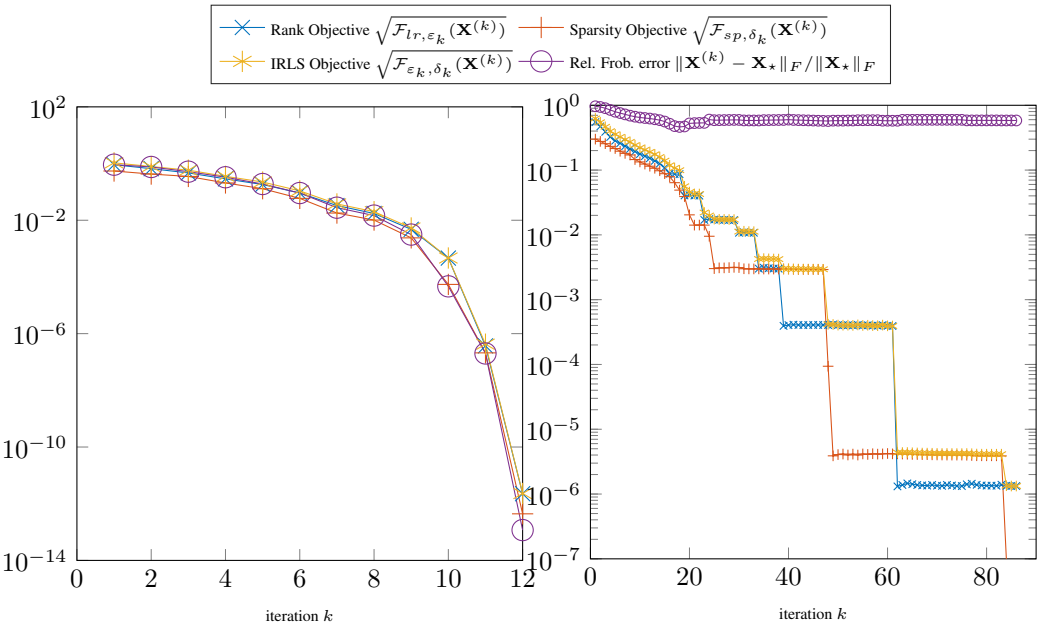

Figure 6: Objective/ error quantities of iterates $\mathbf{X}^{(k)}$ for iterations k. Left: Typical result for $n_1 = 128$, $n = 40$, $m = 875$. Right: Typical result for $n_1 = 128$, $n = 20$, $m = 175$.

A.5 Robustness under Noisy Measurements

The convergence theory for the `IRLS` method Algorithm 1 established in Theorem 2.5 assume that *exact* linear measurements $\mathbf{y} = \mathcal{A}(\mathbf{X}_\star)$ of a row-sparse and low-rank ground truth $\mathbf{X}_\star$ are provided to the algorithm. However, in practice, one would expect that the linear measurement model is only approximately accurate. For IRLS for sparse vector recovery, theoretical guarantees have been established for this case in [26, 57]. We do not extend such results to the simultaneously structured case, but we provide numerical evidence that `IRLS` as defined in Algorithm 1 can be used directly also for noisy measurements.

To this end, we conduct an experiment in the problem setup of Figure 1 in Section 4 for a fixed row-sparsity of $s = 40$, in which the measurements provided to the algorithms `IRLS`, `RiemAdaIHT`

and SPF are such that

$$\mathbf{y} = \mathcal{A}(\mathbf{X}_\star) + \mathbf{w},$$

where $\mathbf{w}$ is a Gaussian vector (i.i.d. entries) with standard deviation of $\sigma = \sqrt{\frac{\|\mathcal{A}(\mathbf{X}_\star)\|_2^2}{m \cdot \mathrm{SNR}}}$ and where SNR is a varying signal-to-noise ratio. We consider SNRs between 10 and 10^{12}, and report the resulting relative Frobenius error statistics in Figure 7.

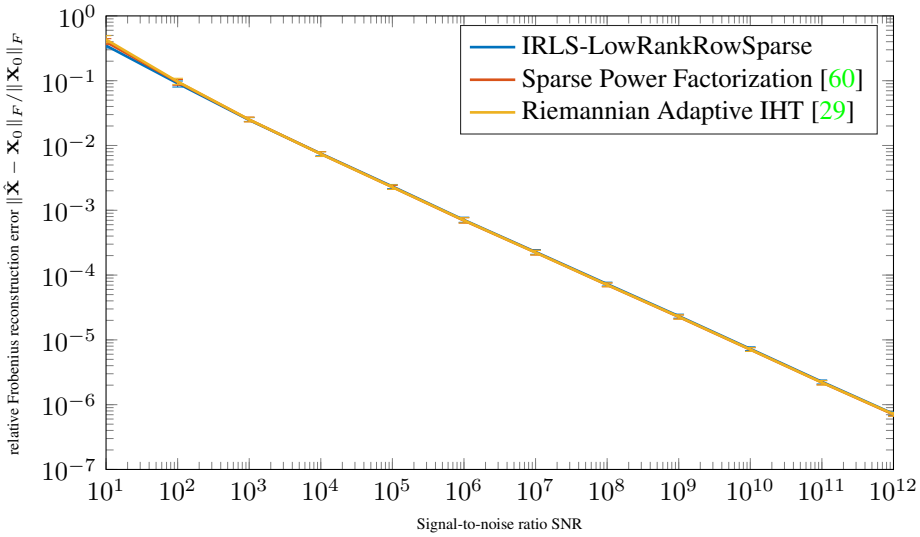

Figure 7: Median relative Frobenius reconstruction errors of different algorithms given noisy Gaussian measurements, $n_1 = 256$, $n_2 = 40$, row-sparsity $s = 40$ and rank $r = 1$, oversampling factor of 3. Error bars correspond to 25% and 75% percentiles.

We observe that the reconstruction error is consistently roughly proportional to the inverse square root of the signal-to-noise ratio, for all three algorithms considered. This suggests that IRLS is as noise robust as comparable algorithms, and expected to be return estimates of the original ground truth that has a reconstruction error that is of the order of the norm of the noise.

B Proofs

The following two sections contain the proofs of our main results. Let us begin with some helpful observations.

First note that the low-rank promoting part $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ of our weight operator can be re-written as

$$W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{Z}) = [\mathbf{U} \quad \mathbf{U}_\perp] \left(\mathbf{H}(\boldsymbol{\sigma}^{(k)}, \varepsilon_k) \circ \left(\begin{bmatrix} \mathbf{U}^* \\ \mathbf{U}_\perp^* \end{bmatrix} \mathbf{Z} \, [\mathbf{V} \quad \mathbf{V}_\perp] \right) \right) \begin{bmatrix} \mathbf{V}^* \\ \mathbf{V}_\perp^* \end{bmatrix}, \tag{18}$$

where

$$\begin{aligned} \mathbf{H}(\boldsymbol{\sigma}^{(k)}, \varepsilon_k) &:= \left[\min\left(\varepsilon_k / \sigma_i^{(k)}, 1\right) \min\left(\varepsilon_k / \sigma_j^{(k)}, 1\right) \right]_{i,j=1}^{n_1,n_2} \\ &= \left[\begin{array}{c|c} \left(\frac{\varepsilon_k^2}{\sigma_i^{(k)} \sigma_j^{(k)}} \right)_{i,j=1}^{r_k} & \left(\frac{\varepsilon_k}{\sigma_i^{(k)}} \right)_{i,j=1}^{r_k,d_2} \\ \hline \left(\frac{\varepsilon_k}{\sigma_j^{(k)}} \right)_{i,j=1}^{d_1,r_k} & \mathbf{1} \end{array} \right] \in \mathbb{R}^{n_1 \times n_2}. \end{aligned}$$

Consequently, all weight operators in Definition 2.1 are self-adjoint and positive. Whereas for $\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}$ this is obvious, for $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}$ it follows from the matrix representation

$$W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k} = \Big([\mathbf{U} \quad \mathbf{U}_\perp] \otimes [\mathbf{V} \quad \mathbf{V}_\perp] \Big) \mathbf{D}_{\mathbf{H}(\boldsymbol{\sigma}^{(k)},\varepsilon_k)} \Big([\mathbf{U} \quad \mathbf{U}_\perp]^* \otimes [\mathbf{V} \quad \mathbf{V}_\perp]^* \Big)$$

where $\mathbf{D}_{\mathbf{H}(\boldsymbol{\sigma}^{(k)},\varepsilon_k)} \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$ is a diagonal matrix with the entries of $\mathbf{H}(\boldsymbol{\sigma}^{(k)}, \varepsilon_k)$, which are all positive, on its diagonal.

B.1 Proof of Theorem 2.5

Before approaching the proof of Theorem 2.5, let us collect various important observations. In order to keep the presentation concise, we defer part of the proofs to Appendix C.

For a rank-r matrix $\mathbf{Z} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*$, we define the tangent space of the manifold of rank-r matrices at $\mathbf{Z}$ as

$$T_{\mathbf{U},\mathbf{V}} := \{\mathbf{U}\mathbf{Z}_1^* + \mathbf{Z}_2\mathbf{V}^* \colon \mathbf{Z}_1 \in \mathbb{R}^{n_2 \times r}, \mathbf{Z}_2 \in \mathbb{R}^{n_1 \times r}\}. \tag{19}$$

In a similar manner, we can define for $\mathbf{Z} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^* \in \mathcal{M}_{r,s}$ and $S = \operatorname{supp}(\mathbf{Z}) = \operatorname{supp}(\mathbf{U}) \subset [n_1]$ the tangent space of $\mathcal{M}_r$ restricted to S as

$$T_{\mathbf{U},\mathbf{V},S} := \{\mathbf{U}\mathbf{Z}_1^* + \mathbf{Z}_2\mathbf{V}^* \colon \mathbf{Z}_1 \in \mathbb{R}^{n_2 \times r}, \mathbf{Z}_2 \in \mathbb{R}^{n_1 \times r} \text{ with } \operatorname{supp}(\mathbf{Z}_2) = S\}. \tag{20}$$

As the following lemma shows, orthogonal projections onto the sets $\mathcal{M}_r^{n_1,n_2}$, $\mathcal{N}_s^{n_1,n_2}$, $T_{\mathbf{U},\mathbf{V}}$, and $T_{\mathbf{U},\mathbf{V},S}$ can be efficiently computed.

Lemma B.1. *We denote the projection operators onto $\mathcal{M}_r^{n_1,n_2}$ and $\mathcal{N}_s^{n_1,n_2}$ by T_r and H_s. T_r truncates a matrix to the r dominant singular values; H_s sets all but the s in ℓ_2-norm largest rows to zero. In case of ambiguities (multiple singular values/rows of same magnitude), by convention we choose the r (respectively s) with smallest index.*

For $\mathbf{U}$ and $\mathbf{V}$ fixed, the orthogonal projection onto $T_{\mathbf{U},\mathbf{V}}$ is given by

$$\mathbb{P}_{\mathbf{U},\mathbf{V}} := \mathbb{P}_{T_{\mathbf{U},\mathbf{V}}}\mathbf{Z} = \mathbf{U}\mathbf{U}^*\mathbf{Z} + \mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^*.$$

For $S \subset [n_1]$ and $\mathbf{U}, \mathbf{V}$ fixed with $\operatorname{supp}(\mathbf{U}) = S$, the orthogonal projection onto $T_{\mathbf{U},\mathbf{V},S}$ is given by

$$\begin{aligned}\mathbb{P}_{\mathbf{U},\mathbf{V},S} := \mathbb{P}_{T_{\mathbf{U},\mathbf{V},S}}\mathbf{Z} &= \mathbb{P}_S(\mathbf{U}\mathbf{U}^*\mathbf{Z} + \mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^*) \\ &= \mathbf{U}\mathbf{U}^*\mathbf{Z} + \mathbb{P}_S\mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^*,\end{aligned}$$

where $\mathbb{P}_S$ projects to the row support S, i.e., it sets all rows to zero which are not indexed by S.

The proof of Lemma B.1 is provided in Appendix C.2. In contrast to the above named projections, the projection onto $\mathcal{M}_{r,s}^{n_1,n_2}$ is not tractable. However, [29, Lemma 2.4] shows that locally $\mathbb{P}_{\mathcal{M}_{r,s}}$ can be replaced by the concatenation of T_r and H_s, i.e., for $\mathbf{Z}_\star \in \mathcal{M}_{r,s}$ and $\mathbf{Z} \approx \mathbf{Z}_\star$, one has that

$$\mathbb{P}_{\mathcal{M}_{r,s}}(\mathbf{Z}) = \mathsf{T}_r(\mathsf{H}_s(\mathbf{Z})).$$

For a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ and $i \in [n_1]$, we set $\rho_i(\mathbf{X}) = \|(\mathbf{X})_{i',:}\|_2$ where i' is a row index corresponding to the i-th largest row of $\mathbf{X}$ in ℓ_2-norm. More precisely, if $\overline{\mathbf{X}}$ is a decreasing rearrangement of $\mathbf{X}$ with rows ordered by magnitude in ℓ_2-norm, then $\rho_i(\mathbf{X}) = \|(\overline{\mathbf{X}})_{i,:}\|_2$. As the following lemma shows, the quantity $\rho_s(\mathbf{X})$ determines a local neighborhood of $\mathbf{X}$ on which H_s preserves the row-support.

Lemma B.2. *Let $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ be a matrix with row-support $S \subset [n_1]$ and $|S| = s$. Then, for any $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}$ with $\|\mathbf{X} - \mathbf{Z}\|_{\infty,2} := \max_{i \in [n_1]} \|\mathbf{X}_{i,:} - \mathbf{Z}_{i,:}\|_2 \leq \frac{1}{2}\rho_s(\mathbf{X})$ the matrix $\mathsf{H}_s(\mathbf{Z})$ has row-support S.*

Proof: Note that

$$\max_{i \in [n_1]} \|(\mathbf{Z})_{i,:} - (\mathbf{X})_{i,:}\|_2 \leq \frac{1}{2}\rho_s(\mathbf{X})$$

implies that any non-zero row of $\mathbf{X}$ corresponds to a non-zero row of $\mathsf{H}_s(\mathbf{Z})$ and hence yields the claim. ∎

A first important observation is that if $\mathcal{A}$ has the (r, s)-RIP, then the norm of kernel elements of $\mathcal{A}$ is bounded in the following way.

Lemma B.3. *If $\mathcal{A}$ has the (r, s)-RIP with $\delta \in (0, 1)$ and $\mathbf{U} \in \mathbb{R}^{n_1 \times r}, \mathbf{V} \in \mathbb{R}^{n_2 \times r}$ with $\operatorname{supp}(\mathbf{U}) = S$, $|S| \leq s$, then*

$$\|\boldsymbol{\Xi}\|_F \leq \sqrt{1 + \frac{\|\mathcal{A}\|_{2\to 2}^2}{(1-\delta)}} \left\|\mathbb{P}_{T_{\mathbf{U},\mathbf{V},S}}^{\perp}(\boldsymbol{\Xi})\right\|_F,$$

for all $\boldsymbol{\Xi} \in \ker(\mathcal{A})$.

The proof of Lemma B.3 is presented in Appendix C.3.

Remark B.4. *If $\mathcal{A}$ is a Gaussian operator with standard deviation $\sqrt{\frac{1}{m}}$, one has with high probability that $\|\mathcal{A}\|_{2\to 2}^2 \approx \frac{n_1 n_2}{m}$.*

We use of the following lemma to characterize the solution of the weighted least squares problem (11). Its proof is analogous to [54, Lemma B.7] and [26, Lemma 5.2].

Lemma B.5. *Let $\mathcal{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^m$. Let $W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ be the weight operator (8) defined based on the information of $\mathbf{X}^{(k)} \in \mathbb{R}^{n_1\times n_2}$. Then the solution of the weighted least squares step (11) of Algorithm 1*

$$\mathbf{X}^{(k+1)} = \underset{\mathcal{A}(\mathbf{X})=\mathbf{y}}{\arg\min} \langle \mathbf{X}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\mathbf{X})\rangle, \tag{21}$$

is unique and solves (21) if and only if

$$\mathcal{A}(\mathbf{X}^{(k+1)}) = \mathbf{y} \qquad \text{and} \qquad \langle W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\mathbf{X}^{(k+1)}), \mathbf{\Xi}\rangle = 0 \ \text{ for all } \ \mathbf{\Xi} \in \ker \mathcal{A}. \tag{22}$$

For any iterate $\mathbf{X}^{(k)}$ of Algorithm 1, we furthermore abbreviate the tangent space (20) of the fixed rank-r manifold $\mathcal{M}_r$ restricted to S at $\mathsf{H}_s(\mathbf{X}^{(k)})$ by

$$T_k = T_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}},S}, \tag{23}$$

where $\widetilde{\mathbf{U}} \in \mathbb{R}^{n_1\times r}$ and $\widetilde{\mathbf{V}} \in \mathbb{R}^{n_2\times r}$ are matrices with leading[4] r singular vectors of $\mathsf{H}_s(\mathbf{X}^{(k)})$ as columns, and $S \in [n_1]$ is the support set of the s rows of $\mathbf{X}^{(k)}$ with largest ℓ_2-norm.

The following lemma is the first crucial tool for showing local quadratic convergence of Algorithm 1.

Lemma B.6. *Let $\mathbf{X}_\star \in \mathcal{M}_{r,s}$ and let $\mathbf{X}^{(k)}$ be the k-th iterate of Algorithm 1 with rank and sparsity parameters $\widetilde{r} = r$ and $\widetilde{s} = s$, let δ_k, ε_k be such that s_k and r_k from Definition 2.1 satisfy $s_k \geq s$ and $r_k \geq r$. Assume that there exists a constant $c > 1$ such that*

$$\|\mathbf{\Xi}\|_F \leq c \left\| \mathbb{P}_{T_k^\perp}(\mathbf{\Xi}) \right\|_F \qquad \text{for all} \qquad \mathbf{\Xi} \in \ker(\mathcal{A}), \tag{24}$$

where $T_k = T_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}},S}$ is as defined in (23) for matrices $\widetilde{\mathbf{U}} \in \mathbb{R}^{n_1\times r}$ and $\widetilde{\mathbf{V}} \in \mathbb{R}^{n_2\times r}$ of leading r left and right singular vectors of $\mathsf{H}_s(\mathbf{X}^{(k)})$ and $S \subset [n_1]$ is the support set of $\mathsf{H}_s(\mathbf{X}^{(k)})$. Assume furthermore that

$$\|\mathbf{X}^{(k)} - \mathbf{X}_\star\| \leq \min\left\{ \frac{1}{2}\rho_s(\mathbf{X}_\star), \min\left\{ \frac{1}{48}, \frac{1}{19c} \right\} \sigma_r(\mathbf{X}_\star) \right\}. \tag{25}$$

Then,

$$\begin{aligned} \left\| \mathbf{X}^{(k+1)} - \mathbf{X}_\star \right\| \leq 4c^2 \min\left\{ \frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k} \right\}^2 \\ \cdot \left(\left\| W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{X}_\star) \right\|_* + \left\| \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k} \cdot \mathbf{X}_\star \right\|_{1,2} \right), \end{aligned}$$

where $\|\mathbf{M}\|_{1,2} = \sum_i \|\mathbf{M}_{i,:}\|_2$ denotes the row-sum norm of a matrix $\mathbf{M}$, and $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}$ and $\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}$ are the weight operators (9) and (10) from Definition 2.1.

The proof of Lemma B.6 is presented in Appendix C.4.

Remark B.7. *By revisiting the proof of Lemma B.6 (omit the bound in (48) and keep the term $\langle \mathbf{\Xi}, \bar{W}\mathbf{\Xi}\rangle$ until the end), one can show under the same assumptions as in Lemma B.6 that*

$$\|\mathbf{\Xi}\|_F^2 \leq 4c^2 \min\left\{ \frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k} \right\}^2 \left\langle \mathbf{\Xi}, \left(\mathbb{P}^\perp_{\mathbf{U},\mathbf{V}} W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k} \mathbb{P}^\perp_{\mathbf{U},\mathbf{V}} + \mathbb{P}_{S^c} \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k} \mathbb{P}_{S^c} \right) \mathbf{\Xi} \right\rangle,$$

where $\mathbf{U}$ and $\mathbf{V}$ are containing the left and right singular vectors of $\mathbf{X}^{(k)}$, see Definition 2.1.

[4] As $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ might not be unique, any set of r leading singular vectors can be chosen in this definition.

The contribution of the norms of the weighted $\mathbf{X}_\star$ terms in Lemma B.6 can be controlled by Lemmas B.8 and B.9 below.

Lemma B.8. *Let $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}: \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ be the rank-based weight operator (9) that uses the spectral information of $\mathbf{X}^{(k)}$ and let $\mathbf{X}_\star \in \mathbb{R}^{n_1\times n_2}$ be a rank-r matrix. Assume that there exists $0<\zeta<\frac{1}{2}$ such that*

$$\max\{\varepsilon_k, \|\mathbf{X}^{(k)}-\mathbf{X}_\star\|\} \leq \zeta\sigma_r(\mathbf{X}_\star). \tag{26}$$

Then for each $1\leq q\leq\infty$,

$$\left\|W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{X}_\star)\right\|_{S_q} \leq \frac{r^{1/q}}{(1-\zeta)\sigma_r(\mathbf{X}_\star)}\left(\frac{1}{1-\zeta}\varepsilon_k^2+\varepsilon_k K_q\|\mathbf{X}^{(k)}-\mathbf{X}_\star\|+2\|\mathbf{X}^{(k)}-\mathbf{X}_\star\|^2\right)$$

and

$$\left\|W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{X}_\star)\right\|_{S_q} \leq \frac{1}{(1-\zeta)\sigma_r(\mathbf{X}_\star)}\left(\frac{r^{1/q}}{1-\zeta}\varepsilon_k^2+2\left\|\mathbf{X}^{(k)}-\mathbf{X}_\star\right\|_{S_q}\left(\varepsilon_k+\|\mathbf{X}^{(k)}-\mathbf{X}_\star\|\right)\right)$$

where K_q is such that $K_q=2^{1/q}$ for $1\leq q\leq 2$ and $4\leq q$, $K_q=\sqrt{2}$ for $2<q\leq 4$ and $K_q=1$ for $q=\infty$.

Lemma B.9. *Let $\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\in\mathbb{R}^{n_1\times n_1}$ be the row-sparsity-based weight operator (10) that uses the current iterate $\mathbf{X}^{(k)}$ with $\delta_k=\min\left(\delta_{k-1},\rho_{s+1}(\mathbf{X}^{(k)})\right)$ and let $\mathbf{X}_\star\in\mathbb{R}^{n_1\times n_2}$ be an s-row-sparse matrix. Assume that there exists $0<\zeta<\frac{1}{2}$ such that*

$$\|\mathbf{X}^{(k)}-\mathbf{X}_\star\|_{\infty,2}=\max_{i\in[n_1]}\|(\mathbf{X}^{(k)})_{i,:}-(\mathbf{X}_\star)_{i,:}\|_2\leq\zeta\rho_s(\mathbf{X}_\star), \tag{27}$$

where $\rho_s(\mathbf{M})$ denotes the ℓ_2-norm of the in ℓ_2-norm s-largest row of $\mathbf{M}$. Then

$$\|\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\cdot\mathbf{X}_\star\|_{1,2}\leq\frac{s\delta_k^2}{(1-\zeta)^2\rho_s(\mathbf{X}_\star)}$$

Lemma B.8 is a refined version of [54, Lemma B.9] the proof of which we omit here.[5] The proof of Lemma B.9 is provided in Appendix C.5. Finally, the following lemma will allow us to control the decay of the IRLS parameters δ_k and ε_k.

Lemma B.10 ([54, Lemma B.5])**.** *Let $\mathbf{X}_\star\in\mathcal{M}_{r,s}$, assume that $\mathcal{A}$ has the (r,s)-RIP with $\delta\in(0,1)$, and let us abbreviate $n=\min\{n_1,n_2\}$.*

Assume that the k-th iterate $\mathbf{X}^{(k)}$ of Algorithm 1 with $\widetilde{r}=r$ and $\widetilde{s}=r$ updates the smoothing parameters in (12) such that one of the statements $\varepsilon_k=\sigma_{r+1}(\mathbf{X}^{(k)})$ or $\delta_k=\rho_{s+1}(\mathbf{X}^{(k)})$ is true, and that $r_k\geq r$ and $s_k\geq s$. Furthermore, let

$$\varepsilon_k\leq\frac{1}{48}\sigma_r(\mathbf{X}_\star)$$

with $c_{\|\mathcal{A}\|_{2\to2}}=\sqrt{1+\frac{\|\mathcal{A}\|^2_{2\to2}}{(1-\delta)}}$, let $\boldsymbol{\Xi}^{(k)}:=\mathbf{X}^{(k)}-\mathbf{X}_\star$ satisfy

$$\|\boldsymbol{\Xi}^{(k)}\|\leq\min\left\{\frac{1}{2}\rho_s(\mathbf{X}_\star),\min\left\{\frac{1}{48},\frac{1}{21c_{\|\mathcal{A}\|_{2\to2}}}\right\}\sigma_r(\mathbf{X}_\star)\right\}. \tag{28}$$

Then

$$\|\boldsymbol{\Xi}^{(k)}\|_F\leq 2\sqrt{2}\sqrt{n}c_{\|\mathcal{A}\|_{2\to2}}\sqrt{4\varepsilon_k^2+\delta_k^2}.$$

The proof of Lemma B.10 is provided in Appendix C.6. We finally have all the tools to prove Theorem 2.5. Note that (14) implies

$$\|\mathbf{X}^{(k)}-\mathbf{X}_\star\|\leq\min\left\{\frac{1}{48c^2_{\|\mathcal{A}\|_{2\to2}}}\min\left\{\frac{\sigma_r(\mathbf{X}_\star)}{r},\frac{\rho_s(\mathbf{X}_\star)}{s}\right\},\frac{1}{4\mu\sqrt{5n}c_{\|\mathcal{A}\|_{2\to2}}}\right\}, \tag{29}$$

[5] This result is a technical result of an unpublished paper. In this paper, we only use that result as a tool. If the reviewers think that adding the proof is relevant here, we are happy to provide it.

and

$$\varepsilon_k \leq \frac{1}{48}\sigma_r(\mathbf{X}_\star) \tag{30}$$

which we will use in the proof below. The latter follows from the fact that for $\tilde{r} = r$

$$\varepsilon_k = \min\left(\varepsilon_{k-1}, \sigma_{r+1}(\mathbf{X}^{(k)})\right) \leq \sigma_{r+1}(\mathbf{X}^{(k)}) \leq ||\mathbf{X}^{(k)} - \mathbf{X}_*|| \leq \sigma_r(\mathbf{X}_*)/48.$$

Proof of Theorem 2.5 : First note, that by assumption $\tilde{r} = r$ and $\tilde{s} = s$. Furthermore, since $\frac{1}{48c^2_{\|\mathcal{A}\|_{2\to 2}}} \leq \frac{1}{2}$, the closeness assumption (29) implies that $\mathsf{H}_s(\mathbf{X}^{(k)})$ and $\mathbf{X}_\star$ share the same support due to Lemma B.2.

Let $\mathbf{X}^{(k)}$ be the k-th iterate of Algorithm 1. Since the operator $\mathcal{A}\colon \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ has the (r,s)-RIP with $\delta \in (0,1)$, Lemma B.3 yields for all $\mathbf{U} \in \mathbb{R}^{n_1\times r}$,$\mathbf{V} \in \mathbb{R}^{n_2\times r}$ with $\operatorname{supp}(\mathbf{U}) = S$, $|S| \leq s$, that

$$\|\mathbf{\Xi}\|_F \leq c_{\|\mathcal{A}\|_{2\to 2}} \left\|\mathbb{P}^{\perp}_{T_{\mathbf{U},\mathbf{V},S}}(\mathbf{\Xi})\right\|_F,$$

for any $\mathbf{\Xi} \in \ker(\mathcal{A})$. Furthermore, due to our assumption that $\tilde{s} = s$ and $\tilde{r} = r$, the smoothing parameter update rules in (12), i.e., $\delta_k = \min\left(\delta_{k-1}, \rho_{s+1}(\mathbf{X}^{(k)})\right)$ and $\varepsilon_k = \min\left(\varepsilon_{k-1}, \sigma_{r+1}(\mathbf{X}^{(k)})\right)$, imply that $r_k \geq r$ and $s_k \geq s$ for all k. We can thus apply Lemma B.6 for $\mathbf{\Xi}^{(k)} := \mathbf{X}^{(k)} - \mathbf{X}_\star$ (note at this point that (29) implies the closeness assumption (25) of Lemma B.6) and obtain

$$\begin{aligned}\left\|\mathbf{\Xi}^{(k+1)}\right\| &= \left\|\mathbf{X}^{(k+1)} - \mathbf{X}_\star\right\| \\ &\leq 4c^2_{\|\mathcal{A}\|_{2\to 2}} \min\left\{\frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k}\right\}^2 \left(\left\|W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{X}_\star)\right\|_* + \left\|\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\cdot \mathbf{X}_\star\right\|_{1,2}\right),\end{aligned} \tag{31}$$

where $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}\colon \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ is the low-rank promoting part (9) of the weight operator associated to $\mathbf{X}^{(k)}$ and $\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k} \in \mathbb{R}^{n_1\times n_1}$ the sparsity promoting part (10). Since by assumption

$$\max(\varepsilon_k, \|\mathbf{\Xi}^{(k)}\|) \leq \frac{1}{48}\sigma_r(\mathbf{X}_\star),$$

Lemma B.8 yields

$$\left\|W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{X}_\star)\right\|_* \leq 0.995\frac{42}{40\sigma_r(\mathbf{X}_\star)}\left(\varepsilon_k^2 r + 2\varepsilon_k\|\mathbf{X}^{(k)} - \mathbf{X}_\star\| + 2\|\mathbf{X}^{(k)} - \mathbf{X}_\star\|^2\right). \tag{32}$$

Similarly, by assumption

$$\|\mathbf{\Xi}^{(k)}\|_{\infty,2} \leq \|\mathbf{\Xi}^{(k)}\| \leq \frac{1}{48s}\rho_s(\mathbf{X}_\star) \leq \frac{1}{48}\rho_s(\mathbf{X}_\star),$$

such that Lemma B.9 yields

$$\|\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\cdot \mathbf{X}_\star\|_{1,2} \leq 0.995\frac{21s\delta_k^2}{20\rho_s(\mathbf{X}_\star)}. \tag{33}$$

Inserting (32) and (33) into (31) we obtain that

$$\begin{aligned}\left\|\mathbf{\Xi}^{(k+1)}\right\| \leq\ & 0.995\cdot 4.2c^2_{\|\mathcal{A}\|_{2\to 2}} \min\left\{\frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k}\right\}^2 \\ &\cdot \left(\frac{r}{\sigma_r(\mathbf{X}_\star)}\left(\varepsilon_k^2 + 2\varepsilon_k\|\mathbf{\Xi}^{(k)}\| + 2\|\mathbf{\Xi}^{(k)}\|^2\right) + \frac{2s}{\rho_s(\mathbf{X}_\star)}\delta_k^2\right).\end{aligned} \tag{34}$$

Due to the assertion that $r_k \geq r$, it holds that $\varepsilon_k \leq \sigma_{r+1}(\mathbf{X}^{(k)})$. Therefore, Lemma C.3 yields that

$$\left(\varepsilon_k^2 + 2\varepsilon_k\|\mathbf{\Xi}^{(k)}\| + 2\|\mathbf{\Xi}^{(k)}\|^2\right) \leq 5\|\mathbf{\Xi}^{(k)}\|^2.$$

and, since $s_k \geq s$, also that

$$\delta_k^2 \leq \|\mathbf{\Xi}^{(k)}\|^2_{\infty,2} \leq \|\mathbf{\Xi}^{(k)}\|^2,$$

since $\delta_k \leq \rho_{s+1}(\mathbf{X}^{(k)})$ in this case.

Thus, using the assertion that one of the statements $\varepsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)})$ or $\delta_k = \rho_{s+1}(\mathbf{X}^{(k)})$ is true, we obtain from (34) that

$$\left\|\mathbf{\Xi}^{(k+1)}\right\| \leq 0.995\cdot 4.2c^2_{\|\mathcal{A}\|_{2\to 2}}\cdot\left(\frac{5r}{\sigma_r(\mathbf{X}_\star)} + \frac{2s}{\rho_s(\mathbf{X}_\star)}\right)\left\|\mathbf{\Xi}^{(k)}\right\|^2. \tag{35}$$

For $\left\|\boldsymbol{\Xi}^{(k)}\right\| < \frac{1}{48} c_{\|\mathcal{A}\|_{2\to 2}}^{-2} \min\{\frac{\sigma_r(\mathbf{X}_\star)}{r}, \frac{\rho_s(\mathbf{X}_\star)}{s}\}$ (as implied by (29)), this yields

$$\|\boldsymbol{\Xi}^{(k+1)}\| < 0.9\|\boldsymbol{\Xi}^{(k)}\| \tag{36}$$

and the quadratic error decay

$$\|\boldsymbol{\Xi}^{(k+1)}\| \leq \mu\|\boldsymbol{\Xi}^{(k)}\|^2$$

if we define $\mu = 4.179 c_{\|\mathcal{A}\|_{2\to 2}}^2 \left(\frac{5r}{\sigma_r(\mathbf{X}_\star)} + \frac{2s}{\rho_s(\mathbf{X}_\star)}\right)$.

To show the remaining statement, we need to argue that the assertions of Theorem 2.5 are satisfied not only for k, but for any $k+\ell$ with $\ell \geq 1$. For this, it is sufficient to show that

1. $r_{k+1} \geq r$,
2. $s_{k+1} \geq s$,
3. $\varepsilon_{k+1} \leq \frac{1}{48}\sigma_r(\mathbf{X}_\star)$,
4. (29) holds for $\mathbf{X}^{(k+1)}$, and that
5. one of the statements $\varepsilon_{k+1} = \sigma_{r+1}(\mathbf{X}^{(k+1)})$ or $\delta_{k+1} = \rho_{s+1}(\mathbf{X}^{(k+1)})$ is true,

as in this case, $\mathbf{X}^{(k+\ell)} \overset{\ell\to\infty}{\to} \mathbf{X}_\star$ follows by induction due to successive application of (36).

For 1. and 2., we see that this follows from the smoothing parameter update rules (12) which imply that $\varepsilon_{k+1} \leq \sigma_{r+1}(\mathbf{X}^{(k+1)})$ and $\delta_{k+1} \leq \rho_{s+1}(\mathbf{X}^{(k+1)})$.

3. follows from (30) and the fact that due to (12), $(\varepsilon_k)_{k\geq 1}$ is non-increasing. 4. is satisfied due to (36) and (29).

To show 5., we note that due to (29), the assertion (28) is satisfied, and therefore it follows from (35) and Lemma B.10 that

$$\left\|\boldsymbol{\Xi}^{(k+1)}\right\| \leq 4.179 c_{\|\mathcal{A}\|_{2\to 2}}^2 \left(\frac{5r}{\sigma_r(\mathbf{X}_\star)} + \frac{2s}{\rho_s(\mathbf{X}_\star)}\right) \left\|\boldsymbol{\Xi}^{(k)}\right\| \cdot 2\sqrt{2}\sqrt{n} c_{\|\mathcal{A}\|_{2\to 2}} \sqrt{4\varepsilon_k^2 + \delta_k^2}.$$

We now distinguish the case (i) $\delta_k < \varepsilon_k$ and the case (ii) $\delta_k \geq \varepsilon_k$.

In case (i), it holds that

$$\begin{aligned}
\sigma_{r+1}(\mathbf{X}^{(k+1)}) \leq \left\|\boldsymbol{\Xi}^{(k+1)}\right\| &\leq 4.179 c_{\|\mathcal{A}\|_{2\to 2}}^2 \left(\frac{5r}{\sigma_r(\mathbf{X}_\star)} + \frac{2s}{\rho_s(\mathbf{X}_\star)}\right) 2\sqrt{10n} c_{\|\mathcal{A}\|_{2\to 2}} \left\|\boldsymbol{\Xi}^{(k)}\right\| \varepsilon_k \\
&= \mu 2\sqrt{10n} c_{\|\mathcal{A}\|_{2\to 2}} \left\|\boldsymbol{\Xi}^{(k)}\right\| \varepsilon_k \\
&< \varepsilon_k,
\end{aligned}$$

where the last inequality holds since by (29) the k-th iterate $\mathbf{X}^{(k)}$ additionally satisfies

$$\|\mathbf{X}^{(k)} - \mathbf{X}_\star\| < \frac{1}{2\mu\sqrt{10} c_{\|\mathcal{A}\|_{2\to 2}}}. \tag{37}$$

In this case, due to the smoothing parameter update rule (12), we have that $\varepsilon_{k+1} = \sigma_{r+1}(\mathbf{X}^{(k+1)})$.

In case (ii), we have likewise that

$$\rho_{s+1}(\mathbf{X}^{(k+1)}) \leq \left\|\boldsymbol{\Xi}^{(k+1)}\right\| \leq \mu 2\sqrt{10n} c_{\|\mathcal{A}\|_{2\to 2}} \left\|\boldsymbol{\Xi}^{(k)}\right\| \delta_k < \delta_k,$$

due to (35), Lemma B.10, and (37). Hence, $\delta_{k+1} = \rho_{s+1}(\mathbf{X}^{(k+1)})$ which shows the remaining statement 5. and concludes the proof of Theorem 2.5. ∎

B.2 Proof of Theorem 2.6

1.) Let $\varepsilon, \delta > 0$ be arbitrary. Due to the additive structure of $\mathcal{F}_{\varepsilon,\delta}(\cdot)$, cf. (4), it is sufficient to establish that

$$\mathcal{F}_{sp,\delta}(\mathbf{Z}) \leq \mathcal{Q}_{sp,\delta}(\mathbf{Z}|\mathbf{X}) = \mathcal{F}_{sp,\delta}(\mathbf{X}) + \langle \nabla \mathcal{F}_{sp,\delta}(\mathbf{X}), \mathbf{Z} - \mathbf{X}\rangle + \frac{1}{2}\langle \mathbf{Z} - \mathbf{X}, \mathbf{W}_{\mathbf{X},\delta}^{sp}(\mathbf{Z} - \mathbf{X})\rangle \tag{38}$$

for any $\mathbf{Z}, \mathbf{X} \in \mathbb{R}^{n_1\times n_2}$, where $\mathbf{W}_{\mathbf{X},\delta}^{sp} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ is defined analogously to (10) and

$$\mathcal{F}_{lr,\varepsilon}(\mathbf{Z}) \leq \mathcal{Q}_{lr,\varepsilon}(\mathbf{Z}|\mathbf{X}) = \mathcal{F}_{lr,\varepsilon}(\mathbf{X}) + \langle \nabla \mathcal{F}_{lr,\varepsilon}(\mathbf{X}), \mathbf{Z} - \mathbf{X}\rangle + \frac{1}{2}\langle \mathbf{Z} - \mathbf{X}, W_{\mathbf{X},\varepsilon}^{lr}(\mathbf{Z} - \mathbf{X})\rangle, \tag{39}$$

for any $\mathbf{Z}, \mathbf{X} \in \mathbb{R}^{n_1\times n_2}$, where $W_{\mathbf{X},\varepsilon}^{lr} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ is defined analogously to (9).

The argument for (38) is standard in the IRLS literature [2, 73, 76] and is based on the facts that both $\mathcal{Q}_{sp,\delta}(\mathbf{Z}|\mathbf{X})$ and $\mathcal{F}_{sp,\delta}(\mathbf{Z})$ are row-wise separable, and that $t \mapsto f_{\sqrt{\tau}}(\sqrt{t})$ is concave and therefore majorized by its linearization: indeed, let $g_\tau : \mathbb{R} \to \mathbb{R}$ be such that

$$g_\tau(t) := \begin{cases} \frac{1}{2}\tau \log(e|t|/\tau), & \text{if } |t| > \tau, \\ \frac{1}{2}|t|, & \text{if } |t| \leq \tau. \end{cases}$$

The function $g_\tau(\cdot)$ is continuously differentiable with derivative $g'_\tau(t) = \frac{\tau}{2\max(|t|,\tau)}\operatorname{sign}(t)$ and furthermore, concave restricted to the non-negative domain $\mathbb{R}_{\geq 0}$.

Therefore, it holds for any $t, t' \in \mathbb{R}_{\geq 0}$ that

$$g_\tau(t) \leq g_\tau(t') + g'_\tau(t')(t - t').$$

We recall the definition $f_\tau(t) = \frac{1}{2}\tau^2 \log(et^2/\tau^2)$ for $|t| > \tau$ and $f_\tau(t) = \frac{1}{2}t^2$ for $|t| \leq \tau$ from (2) with derivative $f'_\tau(t) = \frac{\max(t^2,\tau^2)t}{t^2} = \frac{\tau^2 t}{\max(t^2,\tau^2)}$. Thus, for any $x, z \in \mathbb{R}$, it follows that

$$\begin{aligned} f_\tau(z) = g_{\tau^2}(z^2) &\leq g_{\tau^2}(x^2) + g'_{\tau^2}(x^2)(z^2 - x^2) \\ &= f_\tau(x) + \frac{\tau^2}{2\max(x^2,\tau^2)}(z^2 - x^2), \end{aligned}$$

and inserting $\tau = \delta$, $z = \|\mathbf{Z}_{i,:}\|_2$, $x = \|\mathbf{X}_{i,:}\|_2$ and summing over $i = 1, \dots n_1$ implies that

$$\begin{aligned} \mathcal{F}_{sp,\delta}(\mathbf{Z}) &= \sum_{i=1}^{n_1} f_\delta(\|\mathbf{Z}_{i,:}\|_2) \leq \mathcal{F}_{sp,\delta}(\mathbf{X}) + \sum_{i=1}^{n_1} \frac{\delta^2}{2\max(\|\mathbf{X}_{i,:}\|_2^2, \delta^2)}(\|\mathbf{Z}_{i,:}\|_2^2 - \|\mathbf{X}_{i,:}\|_2^2) \\ &= \mathcal{F}_{sp,\delta}(\mathbf{X}) + \sum_{i=1}^{n_1} \frac{\delta^2}{\max(\|\mathbf{X}_{i,:}\|_2^2, \delta^2)}\langle \mathbf{X}_{i,:}, \mathbf{Z}_{i,:} - \mathbf{X}_{i,:}\rangle + \frac{1}{2}\sum_{i=1}^{n_1} \frac{\|\mathbf{Z}_{i,:} - \mathbf{X}_{i,:}\|_2^2}{\max(\|\mathbf{X}_{i,:}\|_2^2/\delta^2, 1)} \end{aligned}$$

From the chain rule, it follows that for all $i = 1, \dots, n_1$ for which $\mathbf{X}_{i,:} \neq 0$,

$$\frac{d}{d\mathbf{X}_{i,:}} f_\delta(\|\mathbf{X}_{i,:}\|_2) = f'_\delta(\|\mathbf{X}_{i,:}\|_2)\frac{d\|\mathbf{X}_{i,:}\|_2}{d\mathbf{X}_{i,:}} = \frac{\delta^2\|\mathbf{X}_{i,}\|_2}{\max(\|\mathbf{X}_{i,:}\|_2^2, \delta^2)}\frac{\mathbf{X}_{i,:}}{\|\mathbf{X}_{i,:}\|_2}, = \frac{\delta^2 \mathbf{X}_{i,:}}{\max(\|\mathbf{X}_{i,:}\|_2^2, \delta^2)} \tag{40}$$

and therefore

$$\mathcal{F}_{sp,\delta}(\mathbf{Z}) \leq \mathcal{F}_{sp,\delta}(\mathbf{X}) + \langle \nabla\mathcal{F}_{sp,\delta}(\mathbf{X}), \mathbf{Z} - \mathbf{X}\rangle + \frac{1}{2}\langle \mathbf{Z} - \mathbf{X}, \mathbf{W}^{sp}_{\mathbf{X},\delta}(\mathbf{Z} - \mathbf{X})\rangle$$

which shows the majorization of (38), recalling the definition $\mathbf{W}^{sp}_{\mathbf{X},\delta} = \operatorname{diag}\left(\max\left(\|\mathbf{X}_{i,:}\|_2^2/\delta^2, 1\right)_{i=1}^{d_1}\right)^{-1}\Big)$ of (10).

The majorization of (39) is non-trivial but follows in a straightforward way from [53, Theorem 2.4] as the objective $\mathcal{F}_{lr,\varepsilon}(\mathbf{Z})$ corresponds to the one of [53, Theorem 2.4] up to a multiplicative factor of ε^2 and constant additive factors, and since the weight operator $W^{lr}_{\mathbf{X},\varepsilon}$ corresponds to the weight operator used in [53, Chapter 2] for $p = 0$.

2.) Due to the definition (3) of $\mathcal{F}_{sp,\delta_k}(\cdot)$ and the derivative computation of (40), we observe that

$$\nabla\mathcal{F}_{sp,\delta_k}(\mathbf{X}^{(k)}) = \operatorname{diag}\left(\left(\max\left(\left\|(\mathbf{X}^{(k)})_{i,:}\right\|_2^2/\delta_k^2, 1\right)^{-1}\right)_{i=1}^{d_1}\right)\mathbf{X}^{(k)} = \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k} \cdot \mathbf{X}^{(k)},$$

comparing the resulting term with the definition of (10) of $\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}$. Furthermore, an analogue equality follows from the the formula

$$\nabla\mathcal{F}_{lr,\varepsilon_k}(\mathbf{X}^{(k)}) = [\mathbf{U} \quad \mathbf{U}_\perp]\operatorname{diag}\left(\left(\sigma_i^{(k)}\max\left((\sigma_i^{(k)})^2/\varepsilon_k^2, 1\right)^{-1}\right)_{i=1}^{d}\right)\begin{bmatrix}\mathbf{V}^* \\ \mathbf{V}_\perp^*\end{bmatrix}$$

with $\sigma_i^{(k)} = \sigma_i(\mathbf{X}^{(k)})$ for any $i \leq d$, which is a direct consequence from the calculus of spectral functions Lemma C.1, and inserting into the low-rank promoting weight operator formula (9)

$$W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{X}^{(k)}) = [\mathbf{U} \quad \mathbf{U}_\perp]\,\boldsymbol{\Sigma}^{-1}_{\varepsilon_k}\operatorname{diag}\left(\left(\sigma_i^{(k)}\right)_{i=1}^{d}\right)\boldsymbol{\Sigma}^{-1}_{\varepsilon_k}\begin{bmatrix}\mathbf{V}^* \\ \mathbf{V}_\perp^*\end{bmatrix} = \nabla\mathcal{F}_{lr,\varepsilon_k}(\mathbf{X}^{(k)})$$

Inserting $\nabla \mathcal{F}_{sp,\delta_k}(\mathbf{X}^{(k)}) = \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k} \cdot \mathbf{X}^{(k)}$ and $\nabla \mathcal{F}_{lr,\varepsilon_k}(\mathbf{X}^{(k)}) = W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{X}^{(k)})$ into the definitions of $\mathcal{Q}_{lr,\varepsilon_k}(\mathbf{Z}|\mathbf{X}^{(k)})$ and $\mathcal{Q}_{sp,\delta_k}(\mathbf{Z}|\mathbf{X}^{(k)})$, we see that it holds that

$$\mathcal{Q}_{lr,\varepsilon_k}(\mathbf{Z}|\mathbf{X}^{(k)}) = \mathcal{F}_{lr,\varepsilon_k}(\mathbf{X}^{(k)}) + \frac{1}{2}\left(\langle \mathbf{Z}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{Z})\rangle - \langle \mathbf{X}^{(k)}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{X}^{(k)})\rangle\right)$$

and

$$\mathcal{Q}_{sp,\delta_k}(\mathbf{Z}|\mathbf{X}^{(k)}) = \mathcal{F}_{sp,\delta_k}(\mathbf{X}^{(k)}) + \frac{1}{2}\left(\langle \mathbf{Z}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbf{Z}\rangle - \langle \mathbf{X}^{(k)}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbf{X}^{(k)}\rangle\right).$$

Therefore, we see that the weighted least squares solution $\mathbf{X}^{(k+1)}$ of (11) for $k+1$ coincides with the minimizer of

$$\begin{aligned}
&\min_{\mathbf{Z}:\mathcal{A}(\mathbf{Z})=\mathbf{y}} \left[\mathcal{Q}_{lr,\varepsilon_k}(\mathbf{Z}|\mathbf{X}^{(k)}) + \mathcal{Q}_{sp,\delta_k}(\mathbf{Z}|\mathbf{X}^{(k)})\right] \\
&= \min_{\mathbf{Z}:\mathcal{A}(\mathbf{Z})=\mathbf{y}} \Bigg[\mathcal{F}_{lr,\varepsilon_k}(\mathbf{X}^{(k)}) + \mathcal{F}_{sp,\delta_k}(\mathbf{X}^{(k)}) \\
&\qquad + \frac{1}{2}\left(\langle \mathbf{Z}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\mathbf{Z})\rangle - \langle \mathbf{X}^{(k)}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\mathbf{X}^{(k)})\rangle\right)\Bigg]
\end{aligned} \tag{41}$$

with the weight operator $W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}$ of (8), which implies that

$$\mathcal{Q}_{lr,\varepsilon_k}(\mathbf{X}^{(k+1)}|\mathbf{X}^{(k)}) + \mathcal{Q}_{sp,\delta_k}(\mathbf{X}^{(k+1)}|\mathbf{X}^{(k)}) \leq \mathcal{Q}_{lr,\varepsilon_k}(\mathbf{X}^{(k)}|\mathbf{X}^{(k)}) + \mathcal{Q}_{sp,\delta_k}(\mathbf{X}^{(k)}|\mathbf{X}^{(k)}). \tag{42}$$

Using the majorization (16) established in Statement 1 of Theorem 2.6 and (42), it follows that

$$\begin{aligned}
\mathcal{F}_{\varepsilon_k,\delta_k}(\mathbf{X}^{(k+1)}) &\leq \mathcal{Q}_{lr,\varepsilon_k}(\mathbf{X}^{(k+1)}|\mathbf{X}^{(k)}) + \mathcal{Q}_{sp,\delta_k}(\mathbf{X}^{(k+1)}|\mathbf{X}^{(k)}) \\
&\leq \mathcal{Q}_{lr,\varepsilon_k}(\mathbf{X}^{(k)}|\mathbf{X}^{(k)}) + \mathcal{Q}_{sp,\delta_k}(\mathbf{X}^{(k)}|\mathbf{X}^{(k)}) \\
&= \mathcal{F}_{lr,\varepsilon_k}(\mathbf{X}^{(k)}) + \mathcal{F}_{sp,\delta_k}(\mathbf{X}^{(k)}) = \mathcal{F}_{\varepsilon_k,\delta_k}(\mathbf{X}^{(k)}),
\end{aligned} \tag{43}$$

using in the third line that $\mathcal{Q}_{lr,\varepsilon_k}(\mathbf{X}^{(k)}|\mathbf{X}^{(k)}) = \mathcal{F}_{lr,\varepsilon_k}(\mathbf{X}^{(k)})$ and $\mathcal{Q}_{sp,\delta_k}(\mathbf{X}^{(k)}|\mathbf{X}^{(k)}) = \mathcal{F}_{sp,\delta_k}(\mathbf{X}^{(k)})$.

To conclude, it suffices to show that $\varepsilon \mapsto \mathcal{F}_{\varepsilon,\delta_k}(\mathbf{X}^{(k+1)})$ and $\delta \mapsto \mathcal{F}_{\varepsilon_k,\delta}(\mathbf{X}^{(k+1)})$ are non-decreasing functions, since (43) then extends to

$$\mathcal{F}_{\varepsilon_{k+1},\delta_{k+1}}(\mathbf{X}^{(k+1)}) \leq \mathcal{F}_{\varepsilon_k,\delta_{k+1}}(\mathbf{X}^{(k+1)}) \leq \mathcal{F}_{\varepsilon_k,\delta_k}(\mathbf{X}^{(k+1)}) \leq \mathcal{F}_{\varepsilon_k,\delta_k}(\mathbf{X}^{(k)}),$$

where we used that the sequences ε_k and δ_k defined in Algorithm 1 are decreasing. So let us prove this last claim. We define for $t \in \mathbb{R}$ the function $h_t : \mathbb{R}_{>0} \to \mathbb{R}$ such that $h_t(\tau) = f_\tau(t)$, i.e.,

$$h_t(\tau) = \begin{cases} \frac{1}{2}t^2, & \text{if } \tau \geq |t|, \\ \frac{1}{2}\tau^2 \log(et^2/\tau^2), & \text{if } \tau < |t|. \end{cases}$$

This function is continuously differentiable with $h_t'(\tau) = 0$ for all $\tau > |t|$ and

$$h_t'(\tau) = \tau\left(\log(et^2/\tau^2) - 1\right)$$

for $\tau < |t|$, which implies that $h_t'(\tau) \geq 0$ for all $\tau \geq 0$ and thus shows that $\varepsilon \mapsto \mathcal{F}_{\varepsilon,\delta_k}(\mathbf{X}^{(k+1)})$ and $\delta \mapsto \mathcal{F}_{\varepsilon_k,\delta}(\mathbf{X}^{(k+1)})$ are non-decreasing functions due to the additive structure of $\mathcal{F}_{\varepsilon,\delta}(\mathbf{X}^{(k+1)})$ and (4).

3.) First, we argue that $(\mathbf{X}^{(k)})_{k\geq 1}$ is a bounded sequence: Indeed, if $\overline{\varepsilon} := \lim_{k\to\infty} \varepsilon_k > 0$ and $\overline{\delta} := \lim_{k\to\infty} \delta_k > 0$, we note that

$$\begin{aligned}
&\frac{1}{2}\overline{\varepsilon}^2 \log(e\|\mathbf{X}^{(k)}\|^2/\overline{\varepsilon}^2) + \frac{1}{2}\overline{\delta}^2 \log(e \max_i \|\mathbf{X}^{(k)}\|_{\infty,2}/\overline{\delta}^2) \\
&= \frac{1}{2}\overline{\varepsilon}^2 \log(e\sigma_1^2(\mathbf{X}^{(k)})/\overline{\varepsilon}^2) + \frac{1}{2}\overline{\delta}^2 \log(e \max_i \|\mathbf{X}^{(k)}_{i,:}\|_2/\overline{\delta}^2) \\
&\leq \frac{1}{2}\varepsilon_k^2 \log(e\sigma_1^2(\mathbf{X}^{(k)})/\varepsilon_k^2) + \frac{1}{2}\delta_k^2 \log(e \max_i \|\mathbf{X}^{(k)}_{i,:}\|_2/\delta_k^2) \\
&\leq \mathcal{F}_{lr,\varepsilon_k}(\mathbf{X}^{(k)}) + \mathcal{F}_{sp,\delta_k}(\mathbf{X}^{(k)}) = \mathcal{F}_{\varepsilon_k,\delta_k}(\mathbf{X}^{(k)}) \leq \mathcal{F}_{\varepsilon_1,\delta_1}(\mathbf{X}^{(1)}) \\
&\leq \frac{1}{2}\min(d_1,d_2)\sigma_1^2(\mathbf{X}^{(1)}) + \frac{1}{2}d_1 \max_i \|\mathbf{X}^{(1)}_{i,:}\|_2^2 =: C_{\mathbf{X}^{(1)}},
\end{aligned}$$

which implies that $\{\|\mathbf{X}^{(k)}\|\}_{k\geq 1}$ is bounded by a constant that depends on $C_{\mathbf{X}^{(1)}}$.

Furthermore, we note that the optimality condition of (11) (see Lemma B.5) implies that $\mathbf{X}^{(k+1)}$ satisfies

$$\langle W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\mathbf{X}^{(k+1)}), \mathbf{\Xi}\rangle = 0 \text{ for all } \mathbf{\Xi} \in \ker \mathcal{A} \text{ and } \mathcal{A}(\mathbf{X}^{(k+1)}) = \mathbf{y}.$$

Choosing $\mathbf{\Xi} = \mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}$ and using the notation $W^{(k)} = W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}$ we see that

$$\begin{aligned}
&\langle \mathbf{X}^{(k+1)}, W^{(k)}(\mathbf{X}^{(k+1)})\rangle - \langle \mathbf{X}^{(k)}, W^{(k)}(\mathbf{X}^{(k)})\rangle \\
&= \langle \mathbf{X}^{(k+1)}, W^{(k)}(\mathbf{X}^{(k+1)})\rangle - \langle \mathbf{X}^{(k)}, W^{(k)}(\mathbf{X}^{(k)})\rangle - 2\langle W^{(k)}(\mathbf{X}^{(k+1)}), \mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\rangle \\
&= -\Big(\langle \mathbf{X}^{(k+1)}, W^{(k)}(\mathbf{X}^{(k+1)})\rangle - 2\langle W^{(k)}(\mathbf{X}^{(k)}), \mathbf{X}^{(k+1)}\rangle + \langle \mathbf{X}^{(k)}, W^{(k)}(\mathbf{X}^{(k)})\rangle\Big) \\
&= -\big\langle (\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}), W^{(k)}(\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)})\big\rangle.
\end{aligned} \tag{44}$$

Due to the definition of $W^{(k)}$, we note that its smallest singular value (interpreted as matrix operator) can be lower bounded by

$$\begin{aligned}
\sigma_{\min}(W^{(k)}) &\geq \sigma_{\min}\big(W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}\big) + \sigma_{\min}\big(\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\big) \geq \delta_k^2 / \max_i \big\|\mathbf{X}^{(k)}_{i,:}\big\|_2^2 + \varepsilon_k^2/\sigma_1^2(\mathbf{X}^{(k)}) \\
&\geq \bar{\delta}^2/c_{\text{sp},\mathbf{X}^{(1)}} + \bar{\varepsilon}^2/c_{\text{lr},\mathbf{X}^{(1)}},
\end{aligned}$$

where $c_{\text{sp},\mathbf{X}^{(1)}}$ and $c_{\text{lr},\mathbf{X}^{(1)}}$ are constants that satisfy $c_{\text{sp},\mathbf{X}^{(1)}} \leq \bar{\delta}^2 \exp(C_{\mathbf{X}^{(1)}}/\bar{\delta}^2 - 1)$ and $c_{\text{lr},\mathbf{X}^{(1)}} \leq \bar{\varepsilon}^2 \exp(C_{\mathbf{X}^{(1)}}/\bar{\varepsilon}^2 - 1)$.

Combining this with (44), the monotonicity according to Statement 2 of Theorem 2.6, and (41), it follows that

$$\begin{aligned}
\mathcal{F}_{\varepsilon_k,\delta_k}(\mathbf{X}^{(k)}) - \mathcal{F}_{\varepsilon_{k+1},\delta_{k+1}}(\mathbf{X}^{(k+1)}) &\geq \frac{1}{2}\big\langle (\mathbf{X}^{(k)} - \mathbf{X}^{(k+1)}), W^{(k)}(\mathbf{X}^{(k)} - \mathbf{X}^{(k+1)})\big\rangle \\
&\geq \frac{1}{2}\left(\bar{\delta}^2/c_{\text{sp},\mathbf{X}^{(1)}} + \bar{\varepsilon}^2/c_{\text{lr},\mathbf{X}^{(1)}}\right)\|\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\|_F^2.
\end{aligned}$$

Summing over all k, this implies that $\lim_{k\to\infty}\|\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\|_F = 0$.

Since $(\mathbf{X}^{(k)})_{k\geq 1}$ is bounded, each subsequence of $(\mathbf{X}^{(k)})_{k\geq 1}$ has a convergent subsequence. Let $(\mathbf{X}^{(k_\ell)})_{\ell\geq 1}$ be such a sequence with $\lim_{\ell\to\infty}\mathbf{X}^{(k_\ell)} = \bar{\mathbf{X}}$, i.e., $\bar{\mathbf{X}}$ is an accumulation point of the sequence. As the weight operator $W^{(k_\ell)}$ depends continuously on $\mathbf{X}^{(k_\ell)}$, there exists a weight operator $\bar{W}: \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ such that $\bar{W} = \lim_{\ell\to\infty} W^{(k_\ell)}$.

Since $\lim_{k\to\infty}\|\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\|_F = 0$, it also holds that $\mathbf{X}^{(k_\ell+1)} \to \bar{\mathbf{X}}$ and therefore

$$\langle \nabla \mathcal{F}_{\bar{\varepsilon},\bar{\delta}}(\bar{\mathbf{X}}), \mathbf{\Xi}\rangle = \langle \bar{W}(\bar{\mathbf{X}}), \mathbf{\Xi}\rangle = \lim_{\ell\to\infty}\langle W^{(k_\ell)}(\mathbf{X}^{(k_\ell+1)}), \mathbf{\Xi}\rangle = 0$$

for all $\mathbf{\Xi} \in \ker\mathcal{A}$. The statement is shown as this is equivalent to $\bar{\mathbf{X}}$ being a stationary point of $\mathcal{F}_{\bar{\varepsilon},\bar{\delta}}(\cdot)$ subject to the linear constraint $\{\mathbf{Z} \in \mathbb{R}^{n_1\times n_2} : \mathcal{A}(\mathbf{Z}) = \mathbf{y}\}$.

C Technical addendum

C.1 Auxiliary Results

In the proof of Theorem 2.6, we use the following result about the calculus of spectral functions.

Lemma C.1 ([63],[35, Proposition 7.4]). *Let $F: \mathbb{R}^{d_1\times d_2} \to \mathbb{R}$ be a spectral function $F = f\circ\sigma$ with an associated function $f: \mathbb{R}^d \to \mathbb{R}$ that is absolutely permutation symmetric. Then, F is differentiable at $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$ if and only if f is differentiable at $\sigma(\mathbf{X}) \in \mathbb{R}^d$.*

In this case, the gradient ∇F of F at $\mathbf{X}$ is given by

$$\nabla F(\mathbf{X}) = \mathbf{U}\operatorname{diag}\big(\nabla f(\sigma(\mathbf{X}))\big)\mathbf{V}^*$$

if $\mathbf{X} = \mathbf{U}\operatorname{diag}\big(\sigma(\mathbf{X})\big)\mathbf{V}^*$ for unitary matrices $\mathbf{U} \in \mathbb{R}^{d_1\times d_1}$ and $\mathbf{V} \in \mathbb{R}^{d_2\times d_2}$.[6]

[6]Here, for $\mathbf{v} \in \mathbb{R}^{\min(d_1,d_2)}$, $\operatorname{diag}(\mathbf{v}) \in \mathbb{R}^{d_1\times d_2}$ refers to the matrix with diagonal elements $\mathbf{v}_i$ on its main diagonal and zeros elsewhere.

C.2 Proof of Lemma B.1

The projection operators for $\mathcal{M}_r^{n_1,n_2}$, $\mathcal{N}_s^{n_1,n_2}$, and $T_{\mathbf{U},\mathbf{V}}$ are well-known, see e.g. [8]. To see the final statement assume that $\mathbf{U}$ has row-support S and note that $\mathbb{P}_{\mathbf{U},\mathbf{V},S}$ is idempotent, i.e.,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{U},\mathbf{V},S}\mathbb{P}_{\mathbf{U},\mathbf{V},S}\mathbf{Z}\\
&= \mathbf{U}\mathbf{U}^*(\mathbf{U}\mathbf{U}^*\mathbf{Z} + \mathbb{P}_S\mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^*) + \mathbb{P}_S(\mathbf{U}\mathbf{U}^*\mathbf{Z} + \mathbb{P}_S\mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^*)\mathbf{V}\mathbf{V}^*\\
&- \mathbf{U}\mathbf{U}^*(\mathbf{U}\mathbf{U}^*\mathbf{Z} + \mathbb{P}_S\mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^*)\mathbf{V}\mathbf{V}^*\\
&= \mathbf{U}\mathbf{U}^*\mathbf{Z} + \mathbf{U}\mathbf{U}^*\mathbb{P}_S\mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^* + \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^* + \mathbb{P}_S\mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^*\\
&- \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbb{P}_S\mathbf{Z}\mathbf{V}\mathbf{V}^* + \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^*\\
&= \mathbf{U}\mathbf{U}^*\mathbf{Z} + \mathbb{P}_S\mathbf{Z}\mathbf{V}\mathbf{V}^* - \mathbf{U}\mathbf{U}^*\mathbf{Z}\mathbf{V}\mathbf{V}^* = \mathbb{P}_{\mathbf{U},\mathbf{V},S}\mathbf{Z}.
\end{aligned}$$

One can easily check that $\mathbb{P}_{\mathbf{U},\mathbf{V},S}$ acts as identity when applied to matrices in $T_{\mathbf{U},\mathbf{V},S}$ and that $\mathbb{P}_{\mathbf{U},\mathbf{V},S} = \mathbb{P}^*_{\mathbf{U},\mathbf{V},S}$ since $\langle \mathbf{Z}', \mathbb{P}_{\mathbf{U},\mathbf{V},S}\mathbf{Z}\rangle_F = \langle \mathbb{P}_{\mathbf{U},\mathbf{V},S}\mathbf{Z}', \mathbf{Z}\rangle_F$, for any $\mathbf{Z}, \mathbf{Z}'$. This proves the claim.

C.3 Proof of Lemma B.3

Let $\boldsymbol{\Xi} \in \ker(\mathcal{A})$. Note that

$$0 = \|\mathcal{A}(\boldsymbol{\Xi})\|_2 = \left\|\mathcal{A}(\mathbb{P}_{T_{\mathbf{U},\mathbf{V},S}}(\boldsymbol{\Xi}) + \mathbb{P}^{\perp}_{T_{\mathbf{U},\mathbf{V},S}}(\boldsymbol{\Xi}))\right\|_2 \geq \left\|\mathcal{A}(\mathbb{P}_{T_{\mathbf{U},\mathbf{V},S}}(\boldsymbol{\Xi}))\right\|_2 - \left\|\mathcal{A}(\mathbb{P}^{\perp}_{T_{\mathbf{U},\mathbf{V},S}}(\boldsymbol{\Xi}))\right\|_2$$

By the RIP we hence get that

$$\begin{aligned}
\left\|\mathbb{P}_{T_{\mathbf{U},\mathbf{V},S}}(\boldsymbol{\Xi})\right\|_F^2 &\leq \frac{1}{1-\delta}\left\|\mathcal{A}(\mathbb{P}_{T_{\mathbf{U},\mathbf{V},S}}(\boldsymbol{\Xi}))\right\|_2^2 \leq \frac{1}{1-\delta}\left\|\mathcal{A}(\mathbb{P}^{\perp}_{T_{\mathbf{U},\mathbf{V},S}}(\boldsymbol{\Xi}))\right\|_2^2\\
&\leq \frac{\|\mathcal{A}\|_{2\to 2}^2}{(1-\delta)}\left\|\mathbb{P}^{\perp}_{T_{\mathbf{U},\mathbf{V},S}}(\boldsymbol{\Xi})\right\|_F^2.
\end{aligned}$$

Consequently,

$$\|\boldsymbol{\Xi}\|_F^2 = \left\|\mathbb{P}_{T_{\mathbf{U},\mathbf{V},S}}(\boldsymbol{\Xi})\right\|_F^2 + \left\|\mathbb{P}^{\perp}_{T_{\mathbf{U},\mathbf{V},S}}(\boldsymbol{\Xi})\right\|_F^2 \leq \left(1 + \frac{\|\mathcal{A}\|_{2\to 2}^2}{(1-\delta)}\right)\left\|\mathbb{P}^{\perp}_{T_{\mathbf{U},\mathbf{V},S}}(\boldsymbol{\Xi})\right\|_F^2.$$

C.4 Proof of Lemma B.6

In the proof of Lemma B.6 we will use the following fact.

Lemma C.2. *Let $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}$ be the weight operator defined in (9), which is based on the matrices $\mathbf{U} \in \mathbb{R}^{n_1\times r_k}$ and $\mathbf{V} \in \mathbb{R}^{n_2\times r_k}$ of leading r_k left and right singular vectors of $\mathbf{X}^{(k)}$. If $\mathbf{M} \in T_{\mathbf{U},\mathbf{V}}$, then $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{M}) \in T_{\mathbf{U},\mathbf{V}}$. If $\mathbf{M} \in T^{\perp}_{\mathbf{U},\mathbf{V}}$, then $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{M}) \in T^{\perp}_{\mathbf{U},\mathbf{V}}$.*

Proof: If $\mathbf{M} \in T_{\mathbf{U},\mathbf{V}}$, there exist $\mathbf{M}_1 \in \mathbb{R}^{r_k\times r_k}$, $\mathbf{M}_2 \in \mathbb{R}^{r_k\times(n_2-r_k)}$, $\mathbf{M}_3 \in \mathbb{R}^{(n_1-r_k)\times r_k}$ such that

$$\mathbf{M} = \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{M}_1 & \mathbf{M}_2\\ \mathbf{M}_3 & 0\end{bmatrix}\begin{bmatrix}\mathbf{V}^*\\ \mathbf{V}_\perp^*\end{bmatrix},$$

e.g., see [90, Proposition 2.1]. We thus observe that the weight operator $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ from (18) satisfies

$$\begin{aligned}
&W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{M})\\
&= \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}\left(\mathbf{H}(\boldsymbol{\sigma}^{(k)},\varepsilon_k) \circ \left(\begin{bmatrix}\mathbf{U}^*\\ \mathbf{U}_\perp^*\end{bmatrix}\begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{M}_1 & \mathbf{M}_2\\ \mathbf{M}_3 & 0\end{bmatrix}\begin{bmatrix}\mathbf{V}^*\\ \mathbf{V}_\perp^*\end{bmatrix}\begin{bmatrix}\mathbf{V} & \mathbf{V}_\perp\end{bmatrix}\right)\right)\begin{bmatrix}\mathbf{V}^*\\ \mathbf{V}_\perp^*\end{bmatrix}\\
&= \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}\left(\mathbf{H}(\boldsymbol{\sigma}^{(k)},\varepsilon_k) \circ \begin{bmatrix}\mathbf{M}_1 & \mathbf{M}_2\\ \mathbf{M}_3 & 0\end{bmatrix}\right)\begin{bmatrix}\mathbf{V}^*\\ \mathbf{V}_\perp^*\end{bmatrix}\\
&= \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}\mathbf{H}_1^{(k)} \circ \mathbf{M}_1 & \mathbf{H}_2^{(k)} \circ \mathbf{M}_2\\ \mathbf{H}_3^{(k)} \circ \mathbf{M}_3 & 0\end{bmatrix}\begin{bmatrix}\mathbf{V}^*\\ \mathbf{V}_\perp^*\end{bmatrix} \in T_{\mathbf{U},\mathbf{V}}.
\end{aligned}$$

Similarly, if $\mathbf{M} \in T^{\perp}_{\mathbf{U},\mathbf{V}}$, there exists $\mathbf{M}_4 \in \mathbb{R}^{(n_1-r_k)\times(n_2-r_k)}$ such that $\mathbf{M} = \mathbf{U}_\perp\mathbf{M}_4(\mathbf{V}_\perp)^*$ and

$$\begin{aligned}
W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{M}) &= \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}\left(\mathbf{H}(\boldsymbol{\sigma}^{(k)},\varepsilon_k) \circ \begin{bmatrix}0 & 0\\ 0 & \mathbf{M}_4\end{bmatrix}\right)\begin{bmatrix}\mathbf{V}^*\\ \mathbf{V}_\perp^*\end{bmatrix}\\
&= \begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}\begin{bmatrix}0 & 0\\ 0 & \mathbf{M}_4/\varepsilon_k^2\end{bmatrix}\begin{bmatrix}\mathbf{V}^*\\ \mathbf{V}_\perp^*\end{bmatrix} \in T^{\perp}_{\mathbf{U},\mathbf{V}}.
\end{aligned}$$

■

Proof of Lemma B.6: Let $\mathbf{\Xi} \in \mathbb{R}^{n_1 \times n_2}$ be arbitrary. We start with some simple but technical observations. First note that by Lemma B.1, if $T_k = T_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}},S}$,

$$\begin{aligned}\mathbb{P}_{T_k^\perp}\mathbf{\Xi} &= (\mathbf{Id} - \mathbb{P}_{T_k})(\mathbf{\Xi}) \\ &= \mathbf{\Xi} - \left(\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^*\mathbf{\Xi} + \mathbb{P}_S\mathbf{\Xi}\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^*\mathbf{\Xi}\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^*\right) \\ &= (\mathbf{Id} - \mathbb{P}_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}})\mathbf{\Xi} + (\mathbf{Id} - \mathbb{P}_S)\mathbf{\Xi}\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* \\ &= \mathbb{P}^\perp_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}}\mathbf{\Xi} + \mathbb{P}_{S^c}\mathbf{\Xi}\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^*,\end{aligned} \tag{45}$$

with

$$\left\langle \mathbb{P}^\perp_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}}\mathbf{\Xi}, \mathbb{P}_{S^c}\mathbf{\Xi}\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* \right\rangle = \left\langle \widetilde{\mathbf{U}}_\perp\widetilde{\mathbf{U}}_\perp^*\mathbf{\Xi}\widetilde{\mathbf{V}}_\perp\widetilde{\mathbf{V}}_\perp^*, \mathbb{P}_{S^c}\mathbf{\Xi}\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* \right\rangle = 0, \tag{46}$$

where we used that $\mathbb{P}^\perp_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}}\mathbf{\Xi} = (\mathbf{Id} - \mathbb{P}_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}})\mathbf{\Xi} = \widetilde{\mathbf{U}}_\perp\widetilde{\mathbf{U}}_\perp^*\mathbf{\Xi}\widetilde{\mathbf{V}}_\perp\widetilde{\mathbf{V}}_\perp^*$, for $\widetilde{\mathbf{U}}_\perp \in \mathbb{R}^{n_1 \times (n_1 - r)}$ and $\widetilde{\mathbf{V}}_\perp \in \mathbb{R}^{n_2 \times (n_2 - r)}$ being the complementary orthonormal bases of $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$, and that $\widetilde{\mathbf{V}}_\perp^*\widetilde{\mathbf{V}} = \mathbf{0}$.

Second, let now $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times r}$ be matrices with r leading left and right singular vectors of $\mathbf{X}^{(k)}$ in their columns which coincide with the matrices $\mathbf{U}$ and $\mathbf{V}$ from Definition 2.1 in their first r columns. Then it follows from (45) and (46) that

$$\begin{aligned}&\left\|\mathbb{P}_{T_k^\perp}(\mathbf{\Xi})\right\|_F^2 \\ &= \left\|\mathbb{P}^\perp_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}}\mathbf{\Xi}\right\|_F^2 + \left\|\mathbb{P}_{S^c}\mathbf{\Xi}\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^*\right\|_F^2 \\ &= \left\|\mathbb{P}^\perp_{\mathbf{U},\mathbf{V}}\mathbf{\Xi} + (\mathbb{P}^\perp_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}} - \mathbb{P}^\perp_{\mathbf{U},\mathbf{V}})\mathbf{\Xi}\right\|_F^2 + \left\|\mathbb{P}_{S^c}\mathbf{\Xi}(\mathbf{V}\mathbf{V}^* + (\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* - \mathbf{V}\mathbf{V}^*))\right\|_F^2 \\ &\leq 2\left(\left\|\mathbb{P}^\perp_{\mathbf{U},\mathbf{V}}\mathbf{\Xi}\right\|_F^2 + \left\|(\mathbb{P}^\perp_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}} - \mathbb{P}^\perp_{\mathbf{U},\mathbf{V}})\mathbf{\Xi}\right\|_F^2 + \|\mathbb{P}_{S^c}\mathbf{\Xi}\mathbf{V}\mathbf{V}^*\|_F^2 + \left\|\mathbb{P}_{S^c}\mathbf{\Xi}(\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* - \mathbf{V}\mathbf{V}^*)\right\|_F^2\right).\end{aligned} \tag{47}$$

By an argument analogous to (46), we observe that

$$\left\|\mathbb{P}^\perp_{\mathbf{U},\mathbf{V}}\mathbf{\Xi}\right\|_F^2 + \|\mathbb{P}_{S^c}\mathbf{\Xi}\mathbf{V}\mathbf{V}^*\|_F^2 = \langle \mathbb{P}^\perp_{\mathbf{U},\mathbf{V}}\mathbf{\Xi} + \mathbb{P}_{S^c}\mathbf{\Xi}\mathbf{V}\mathbf{V}^*, \mathbb{P}^\perp_{\mathbf{U},\mathbf{V}}\mathbf{\Xi} + \mathbb{P}_{S^c}\mathbf{\Xi}\mathbf{V}\mathbf{V}^* \rangle = \langle \widetilde{\mathbf{\Xi}}, \widetilde{\mathbf{\Xi}} \rangle,$$

where $\widetilde{\mathbf{\Xi}} = \mathbb{P}_{\mathcal{M}}(\mathbf{\Xi})$ is an element of the subspace $\mathcal{M} = \mathcal{M}_1 \oplus \mathcal{M}_2 \subset \mathbb{R}^{n_1 \times n_2}$ that is the direct sum of the subspaces $\mathcal{M}_1 := \{\mathbb{P}^\perp_{\mathbf{U},\mathbf{V}}\mathbf{Z} : \mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}\}$ and $\mathcal{M}_2 := \{\mathbb{P}_{S^c}\mathbf{Z}\mathbf{V}\mathbf{V}^* : \mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}\}$.

Let now $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}$ be the rank promoting part of the weight operator from (9) and $\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}$ be the row-sparsity promoting part from (10). Note that the restriction of

$$\bar{W} := \mathbb{P}^\perp_{\mathbf{U},\mathbf{V}}W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}\mathbb{P}^\perp_{\mathbf{U},\mathbf{V}} + \mathbb{P}_{S^c}\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_{S^c}$$

to $\mathcal{M}$ is invertible as its first summand is invertible on $\mathcal{M}_1 = T^\perp_{\mathbf{U},\mathbf{V}}$, its second summand is invertible on $\mathcal{M}_2$ (recall that the weight operators are positive definite), and $\mathcal{M}_1 \perp \mathcal{M}_2$. Therefore it holds that

$$\begin{aligned}&\left\|\mathbb{P}^\perp_{\mathbf{U},\mathbf{V}}\mathbf{\Xi}\right\|_F^2 + \|\mathbb{P}_{S^c}\mathbf{\Xi}\mathbf{V}\mathbf{V}^*\|_F^2 \\ &= \langle \widetilde{\mathbf{\Xi}}, \widetilde{\mathbf{\Xi}} \rangle = \left\langle \bar{W}^{1/2}_{|\mathcal{M}}\widetilde{\mathbf{\Xi}}, \bar{W}^{-1}_{|\mathcal{M}}\bar{W}^{1/2}_{|\mathcal{M}}\widetilde{\mathbf{\Xi}} \right\rangle \\ &\leq \sigma_1\left(\bar{W}^{-1}_{|\mathcal{M}}\right)\left\langle \widetilde{\mathbf{\Xi}}, \bar{W}_{|\mathcal{M}}\widetilde{\mathbf{\Xi}} \right\rangle = \frac{1}{\sigma_{\min}\left(\bar{W}_{|\mathcal{M}}\right)}\left\langle \widetilde{\mathbf{\Xi}}, \bar{W}\widetilde{\mathbf{\Xi}} \right\rangle \\ &\leq \frac{1}{\sigma_{\min}\left(\left(\mathbb{P}^\perp_{\mathbf{U},\mathbf{V}}W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}\mathbb{P}^\perp_{\mathbf{U},\mathbf{V}}\right)_{|\mathcal{M}}\right) + \sigma_{\min}\left(\left(\mathbb{P}_{S^c}\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_{S^c}\right)_{|\mathcal{M}}\right)}\left\langle \widetilde{\mathbf{\Xi}}, \bar{W}\widetilde{\mathbf{\Xi}} \right\rangle \\ &\leq \frac{1}{\frac{\varepsilon_k^2}{\sigma^2_{r+1}(\mathbf{X}^{(k)})} + \frac{\delta_k^2}{\rho^2_{s+1}(\mathbf{X}^{(k)})}}\left\langle \mathbf{\Xi}, \bar{W}\mathbf{\Xi} \right\rangle.\end{aligned}$$

In the first inequality, we used that $\bar{W}_{|\mathcal{M}}$ is positive definite. In the second inequality, we used that $\sigma_{\min}(A + B) \geq \sigma_{\min}(A) + \sigma_{\min}(B)$, for any positive semidefinite operators A and B, and and in the third inequality that $\langle \widetilde{\mathbf{\Xi}}, \bar{W}\widetilde{\mathbf{\Xi}} \rangle \leq \langle \mathbf{\Xi}, \bar{W}\mathbf{\Xi} \rangle$. The latter observation can be deduced as follows: Note that, by the self-adjointness of $\bar{W}$,

$$\begin{aligned}\langle \mathbf{\Xi}, \bar{W}\mathbf{\Xi} \rangle &= \langle \mathbb{P}_{\mathcal{M}}(\mathbf{\Xi}), \bar{W}\mathbb{P}_{\mathcal{M}}(\mathbf{\Xi}) \rangle + \langle \mathbb{P}_{\mathcal{M}}(\mathbf{\Xi}), \bar{W}\mathbb{P}^\perp_{\mathcal{M}}(\mathbf{\Xi}) \rangle + \langle \mathbb{P}^\perp_{\mathcal{M}}(\mathbf{\Xi}), \bar{W}\mathbb{P}_{\mathcal{M}}(\mathbf{\Xi}) \rangle + \langle \mathbb{P}^\perp_{\mathcal{M}}(\mathbf{\Xi}), \bar{W}\mathbb{P}^\perp_{\mathcal{M}}(\mathbf{\Xi}) \rangle \\ &= \langle \widetilde{\mathbf{\Xi}}, \bar{W}\widetilde{\mathbf{\Xi}} \rangle + 2\langle \mathbb{P}^\perp_{\mathcal{M}}(\mathbf{\Xi}), \bar{W}\mathbb{P}_{\mathcal{M}}(\mathbf{\Xi}) \rangle + \langle \mathbb{P}^\perp_{\mathcal{M}}(\mathbf{\Xi}), \bar{W}\mathbb{P}^\perp_{\mathcal{M}}(\mathbf{\Xi}) \rangle.\end{aligned}$$

Since $\bar{W}$ is positive semi-definite (due to the fact that both $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}$ and $\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}$ are positive definite), all that remains is to argue that the mixed term on the right-hand side vanishes. To this end, note that $\mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}\mathbf{Z} = \mathbb{P}_{S^c}\mathbf{Z}\mathbf{V}\mathbf{V}^* = 0$, for any $\mathbf{Z} \in \mathcal{M}_{\perp}$ and compute

$$\begin{aligned}\langle \mathbb{P}^{\perp}_{\mathcal{M}}(\boldsymbol{\Xi}), \bar{W}\mathbb{P}_{\mathcal{M}}(\boldsymbol{\Xi})\rangle &= \langle \mathbb{P}^{\perp}_{\mathcal{M}}(\boldsymbol{\Xi}), \mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}(W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}(\mathbb{P}_{\mathcal{M}_1}(\boldsymbol{\Xi}))))\rangle + \langle \mathbb{P}^{\perp}_{\mathcal{M}}(\boldsymbol{\Xi}), \mathbb{P}_{S^c}\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_{S^c}\mathbb{P}_{\mathcal{M}_2}(\boldsymbol{\Xi})\rangle \\ &= \langle \mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}(\mathbb{P}^{\perp}_{\mathcal{M}}(\boldsymbol{\Xi})), W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}(\mathbb{P}_{\mathcal{M}_1}(\boldsymbol{\Xi})))\rangle + \langle \mathbb{P}_{S^c}\mathbb{P}^{\perp}_{\mathcal{M}}(\boldsymbol{\Xi})\mathbf{V}\mathbf{V}^*, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_{S^c}\boldsymbol{\Xi}\rangle \\ &= 0.\end{aligned}$$

We can now continue by estimating

$$\begin{aligned}\left\langle \boldsymbol{\Xi}, \bar{W}\boldsymbol{\Xi}\right\rangle &= \left\langle \boldsymbol{\Xi}, \mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}\mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi}\right\rangle + \left\langle \boldsymbol{\Xi}, \mathbb{P}_{S^c}\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_{S^c}\boldsymbol{\Xi}\right\rangle \\ &= \left\langle \mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}(\boldsymbol{\Xi}), W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}\mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}(\boldsymbol{\Xi})\right\rangle + \left\langle \mathbb{P}_{S^c}\boldsymbol{\Xi}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_{S^c}\boldsymbol{\Xi}\right\rangle \\ &\leq \left\langle \boldsymbol{\Xi}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}\boldsymbol{\Xi}\right\rangle + \left\langle \boldsymbol{\Xi}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\boldsymbol{\Xi}\right\rangle \\ &= \left\langle \boldsymbol{\Xi}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}\boldsymbol{\Xi}\right\rangle,\end{aligned} \tag{48}$$

using the positive semidefiniteness of $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}$ and $\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}$ in the last inequality. To be precise, the last inequality can be argued as follows: Due to complimentary supports S and S^c, we see that

$$\begin{aligned}\langle \boldsymbol{\Xi}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\boldsymbol{\Xi}\rangle &= \langle \mathbb{P}_S\boldsymbol{\Xi}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_S\boldsymbol{\Xi}\rangle + \langle \mathbb{P}_{S^c}\boldsymbol{\Xi}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_{S^c}\boldsymbol{\Xi}\rangle \\ &\quad + \underbrace{\langle \mathbb{P}_S\boldsymbol{\Xi}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_{S^c}\boldsymbol{\Xi}\rangle}_{=0} + \underbrace{\langle \mathbb{P}_{S^c}\boldsymbol{\Xi}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_S\boldsymbol{\Xi}\rangle}_{=0} \\ &\geq \langle \mathbb{P}_{S^c}\boldsymbol{\Xi}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_{S^c}\boldsymbol{\Xi}\rangle.\end{aligned} \tag{49}$$

Similarly, we note that $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}$ acts diagonally on $T_{\mathbf{U},\mathbf{V}}$ and $T^{\perp}_{\mathbf{U},\mathbf{V}}$. Indeed, we have by Lemma C.2 that if $\mathbf{M} \in T_{\mathbf{U},\mathbf{V}}$, then $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{M}) \in T_{\mathbf{U},\mathbf{V}}$ and if $\mathbf{M} \in T^{\perp}_{\mathbf{U},\mathbf{V}}$, then $W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{M}) \in T^{\perp}_{\mathbf{U},\mathbf{V}}$, which implies

$$\langle \mathbb{P}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi})\rangle = 0 \quad \text{and} \quad \langle \mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbb{P}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi})\rangle = 0$$

due to the orthogonality of elements in $T^{\perp}_{\mathbf{U},\mathbf{V}}$ and $T_{\mathbf{U},\mathbf{V}}$, respectively, and therefore it follows from $\boldsymbol{\Xi} = T_{\mathbf{U},\mathbf{V}}(\boldsymbol{\Xi}) + T^{\perp}_{\mathbf{U},\mathbf{V}}(\boldsymbol{\Xi})$ that

$$\begin{aligned}\langle \boldsymbol{\Xi}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\boldsymbol{\Xi})\rangle &= \langle \mathbb{P}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbb{P}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi})\rangle + \langle \mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi})\rangle \\ &\quad + \underbrace{\langle \mathbb{P}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi})\rangle}_{=0} + \underbrace{\langle \mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbb{P}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi})\rangle}_{=0} \\ &\geq \langle \mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}}\boldsymbol{\Xi})\rangle.\end{aligned} \tag{50}$$

Combining the previous estimates with (47) and noticing that

$$\frac{1}{\frac{\varepsilon_k^2}{\sigma^2_{r+1}(\mathbf{X}^{(k)})} + \frac{\delta_k^2}{\rho^2_{s+1}(\mathbf{X}^{(k)})}} \leq \min\left\{\frac{\sigma^2_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k^2}, \frac{\rho^2_{s+1}(\mathbf{X}^{(k)})}{\delta_k^2}\right\},$$

we obtain

$$\begin{aligned}\left\|\mathbb{P}_{T_k^{\perp}}(\boldsymbol{\Xi})\right\|_F^2 &\leq 2\left(\min\left\{\frac{\sigma^2_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k^2}, \frac{\rho^2_{s+1}(\mathbf{X}^{(k)})}{\delta_k^2}\right\}\left\langle \boldsymbol{\Xi}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}\boldsymbol{\Xi}\right\rangle\right) \\ &\quad + 2\left\|(\mathbb{P}^{\perp}_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}} - \mathbb{P}^{\perp}_{\mathbf{U},\mathbf{V}})\boldsymbol{\Xi}\right\|_F^2 + 2\left\|\mathbb{P}_{S^c}\boldsymbol{\Xi}(\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* - \mathbf{V}\mathbf{V}^*)\right\|_F^2.\end{aligned} \tag{51}$$

Next, we control the last two summands in (51) using matrix perturbation results. Recall that $\mathbf{U}_\star \in \mathbb{R}^{n_1\times r}$ and $\mathbf{V}_\star \in \mathbb{R}^{n_2\times r}$ are the singular vector matrices of the reduced singular value decomposition of $\mathbf{X}_\star$. First observe that

$$\begin{aligned}&\left\|(\mathbb{P}_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}} - \mathbb{P}_{\mathbf{U}_\star,\mathbf{V}_\star})\boldsymbol{\Xi}\right\|_F \\ &= \left\|(\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^* - \mathbf{U}_\star\mathbf{U}_\star^*)\boldsymbol{\Xi}(\mathbf{Id} - \mathbf{V}_\star\mathbf{V}_\star^*) + (\mathbf{Id} - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^*)\boldsymbol{\Xi}(\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* - \mathbf{V}_\star\mathbf{V}_\star^*)\right\|_F \\ &\leq \left\|\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^* - \mathbf{U}_\star\mathbf{U}_\star^*\right\| \|\boldsymbol{\Xi}\|_F \|\mathbf{Id} - \mathbf{V}_\star\mathbf{V}_\star^*\| + \left\|\mathbf{Id} - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^*\right\| \|\boldsymbol{\Xi}\|_F \left\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* - \mathbf{V}_\star\mathbf{V}_\star^*\right\| \\ &\leq \left(\left\|\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^* - \mathbf{U}_\star\mathbf{U}_\star^*\right\| + \left\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* - \mathbf{V}_\star\mathbf{V}_\star^*\right\|\right)\|\boldsymbol{\Xi}\|_F.\end{aligned}$$

Now note that, by [12, Lemma 1] and [65, Theorem 3.5], we obtain

$$\begin{aligned}\left\|\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^* - \mathbf{U}_\star\mathbf{U}_\star^*\right\| + \left\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* - \mathbf{V}_\star\mathbf{V}_\star^*\right\| &\leq 2\left(\|\mathbf{U}_\star\widetilde{\mathbf{U}}_\perp^*\| + \|\mathbf{V}_\star\widetilde{\mathbf{V}}_\perp^*\|\right) \leq 4\frac{\|\mathsf{H}_s(\mathbf{X}^{(k)}) - \mathbf{X}_\star\|}{\sigma_r(\mathbf{X}_\star) - \sigma_{r+1}(\mathsf{H}_s(\mathbf{X}^{(k)}))} \\ &\leq 4\frac{\|\mathbf{X}^{(k)} - \mathbf{X}_\star\|}{\sigma_r(\mathbf{X}_\star) - \sigma_{r+1}(\mathbf{X}^{(k)})} \leq 4\frac{\|\mathbf{X}^{(k)} - \mathbf{X}_\star\|}{(1-1/48)\sigma_r(\mathbf{X}_\star)},\end{aligned} \tag{52}$$

where we used some small observations in the last two inequalities: First, $\sigma_{r+1}(\mathsf{H}_s(\mathbf{X}^{(k)})) \leq \sigma_{r+1}(\mathbf{X}^{(k)})$, which follows from the rectangular Cauchy interlacing theorem [27, Theorem 23]. Second, according to Lemma B.2 and (25), the row-support S of $\mathsf{H}_s(\mathbf{X}^{(k)})$ coincides with the row-support $S_\star = \{i \in [n_1] : \|(\mathbf{X}_\star)_{i,:}\|_2 \neq 0\}$ of $\mathbf{X}_\star$ and hence $\mathsf{H}_s(\mathbf{X}^{(k)}) - \mathbf{X}_\star$ is a submatrix of $\mathbf{X}^{(k)} - \mathbf{X}_\star$. Finally, $\sigma_{r+1}(\mathbf{X}^{(k)}) = \|\mathsf{T}_r(\mathbf{X}^{(k)}) - \mathbf{X}^{(k)}\| \leq \|\mathbf{X}_\star - \mathbf{X}^{(k)}\| \leq \frac{1}{48}\sigma_r(\mathbf{X}_\star)$ due to (25). Consequently,

$$\left\|(\mathbb{P}_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}} - \mathbb{P}_{\mathbf{U}_\star,\mathbf{V}_\star})\boldsymbol{\Xi}\right\|_F \leq 4\frac{\|\mathbf{X}^{(k)} - \mathbf{X}_\star\|}{(1-1/48)\sigma_r(\mathbf{X}_\star)}\|\boldsymbol{\Xi}\|_F$$

and, by a similar argument,

$$\|(\mathbb{P}_{\mathbf{U}_\star,\mathbf{V}_\star} - \mathbb{P}_{\mathbf{U},\mathbf{V}})\boldsymbol{\Xi}\|_F \leq 4\frac{\|\mathbf{X}^{(k)} - \mathbf{X}_\star\|}{(1-1/48)\sigma_r(\mathbf{X}_\star)}\|\boldsymbol{\Xi}\|_F,$$

such that it follows from $\mathbb{P}^\perp_{\mathbf{U},\mathbf{V}} = \mathbf{Id} - \mathbb{P}_{\mathbf{U},\mathbf{V}}$ and $\mathbb{P}^\perp_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}} = \mathbf{Id} - \mathbb{P}_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}}$ that

$$\begin{aligned}\left\|(\mathbb{P}^\perp_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}} - \mathbb{P}^\perp_{\widetilde{\mathbf{U}},\mathbf{V}})\boldsymbol{\Xi}\right\|_F &= \left\|(\mathbb{P}_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}} - \mathbb{P}_{\mathbf{U},\mathbf{V}})\boldsymbol{\Xi}\right\|_F \leq \left\|(\mathbb{P}_{\widetilde{\mathbf{U}},\widetilde{\mathbf{V}}} - \mathbb{P}_{\mathbf{U}_\star,\mathbf{V}_\star})\boldsymbol{\Xi}\right\|_F + \left\|(\mathbb{P}_{\mathbf{U}_\star,\mathbf{V}_\star} - \mathbb{P}_{\mathbf{U},\mathbf{V}})\boldsymbol{\Xi}\right\|_F \\ &\leq 8\frac{\|\mathbf{X}^{(k)} - \mathbf{X}_\star\|}{(1-1/48)\sigma_r(\mathbf{X}_\star)}\|\boldsymbol{\Xi}\|_F.\end{aligned} \tag{53}$$

To estimate the fourth summand in (51), we argue analogously that

$$\begin{aligned}\left\|\mathbb{P}_{S^c}\boldsymbol{\Xi}(\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* - \mathbf{V}\mathbf{V}^*)\right\|_F &\leq \left\|\boldsymbol{\Xi}(\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* - \mathbf{V}\mathbf{V}^*)\right\|_F \leq \left(\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^* - \mathbf{V}_\star\mathbf{V}_\star^*\| + \|\mathbf{V}_\star\mathbf{V}_\star^* - \mathbf{V}\mathbf{V}^*\|\right)\|\boldsymbol{\Xi}\|_F \\ &\leq 2\left(\|\mathbf{V}_\star\widetilde{\mathbf{V}}_\perp^*\| + \|\mathbf{V}_\star\mathbf{V}_\perp^*\|\right)\|\boldsymbol{\Xi}\|_F \\ &\leq 2\left(\frac{\|\mathsf{H}_s(\mathbf{X}^{(k)}) - \mathbf{X}_\star\|}{\sigma_r(\mathbf{X}_\star) - \sigma_{r+1}(\mathsf{H}_s(\mathbf{X}^{(k)}))} + \frac{\|\mathbf{X}^{(k)} - \mathbf{X}_\star\|}{\sigma_r(\mathbf{X}_\star) - \sigma_{r+1}(\mathbf{X}^{(k)})}\right)\|\boldsymbol{\Xi}\|_F \\ &\leq 4\frac{\|\mathbf{X}^{(k)} - \mathbf{X}_\star\|}{\sigma_r(\mathbf{X}_\star) - \sigma_{r+1}(\mathbf{X}^{(k)})}\|\boldsymbol{\Xi}\|_F \leq 4\frac{\|\mathbf{X}^{(k)} - \mathbf{X}_\star\|}{(1-1/48)\sigma_r(\mathbf{X}_\star)}\|\boldsymbol{\Xi}\|_F,\end{aligned} \tag{54}$$

using again that $\sigma_{r+1}(\mathbf{X}^{(k)}) \leq \|\mathbf{X}_\star - \mathbf{X}^{(k)}\| \leq \frac{1}{48}\sigma_r(\mathbf{X}_\star)$ due to (25).

Let now $\boldsymbol{\Xi}^{(k+1)} = \mathbf{X}^{(k+1)} - \mathbf{X}_\star$. Combining (24) and (51)-(52) we can proceed to estimate that

$$\begin{aligned}&\|\boldsymbol{\Xi}^{(k+1)}\|_F^2 \\ &\leq c^2\left\|\mathbb{P}_{T_k^\perp}(\boldsymbol{\Xi}^{(k+1)})\right\|_F^2 \\ &\leq 2c^2\min\left\{\frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k}\right\}^2\left\langle\boldsymbol{\Xi}^{(k+1)}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}\boldsymbol{\Xi}^{(k+1)}\right\rangle \\ &+ 2\left[\left(\frac{8}{47/48}\right)^2 + \left(\frac{4}{47/48}\right)^2\right]c^2\|\boldsymbol{\Xi}^{(k+1)}\|_F^2\frac{\|\mathbf{X}^{(k)} - \mathbf{X}_\star\|^2}{\sigma_r^2(\mathbf{X}_\star)} \\ &\leq 2c^2\min\left\{\frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k}\right\}^2\left\langle\boldsymbol{\Xi}^{(k+1)}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\boldsymbol{\Xi}^{(k+1)})\right\rangle + 167c^2\|\boldsymbol{\Xi}^{(k+1)}\|_F^2\frac{1}{(19c)^2} \\ &\leq 2c^2\min\left\{\frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k}\right\}^2\left\langle\boldsymbol{\Xi}^{(k+1)}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\boldsymbol{\Xi}^{(k+1)})\right\rangle + \frac{1}{2}\|\boldsymbol{\Xi}^{(k+1)}\|_F^2,\end{aligned} \tag{55}$$

where the third inequality follows from (8) and (25). Hence, rearranging (55) yields

$$\|\boldsymbol{\Xi}^{(k+1)}\|_F^2 \leq 4c^2\min\left\{\frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k}\right\}^2\left\langle\boldsymbol{\Xi}^{(k+1)}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\boldsymbol{\Xi}^{(k+1)})\right\rangle.$$

By Lemma B.5, we know that $\mathbf{X}^{(k+1)}$ fulfills

$$0 = \langle \Xi^{(k+1)}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\mathbf{X}^{(k+1)}) \rangle = \langle \Xi^{(k+1)}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\Xi^{(k+1)}) \rangle + \langle \Xi^{(k+1)}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\mathbf{X}_\star) \rangle$$

such that we conclude that

$$\begin{aligned}
&\left\| \Xi^{(k+1)} \right\|^2 \leq \left\| \Xi^{(k+1)} \right\|_F^2 \\
&\leq 4c^2 \min\left\{ \frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k} \right\}^2 \langle \Xi^{(k+1)}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\Xi^{(k+1)}) \rangle \\
&= -4c^2 \min\left\{ \frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k} \right\}^2 \langle \Xi^{(k+1)}, W_{\mathbf{X}^{(k)},\varepsilon_k,\delta_k}(\mathbf{X}_\star) \rangle \\
&= -4c^2 \min\left\{ \frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k} \right\}^2 \left(\langle \Xi^{(k+1)}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{X}_\star) \rangle + \langle \Xi^{(k+1)}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k} \cdot \mathbf{X}_\star \rangle \right) \\
&\leq 4c^2 \min\left\{ \frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k} \right\}^2 \left(\left\| W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{X}_\star) \right\|_* \left\| \Xi^{(k+1)} \right\| + \left\| \Xi^{(k+1)} \right\| \left\| \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k} \cdot \mathbf{X}_\star \right\|_{1,2} \right) \\
&= 4c^2 \min\left\{ \frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k} \right\}^2 \left(\left\| W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}(\mathbf{X}_\star) \right\|_* + \left\| \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k} \cdot \mathbf{X}_\star \right\|_{1,2} \right) \left\| \Xi^{(k+1)} \right\|,
\end{aligned}$$

which completes the proof. We used in the penultimate line Hölder's inequality and that

$$|\langle \mathbf{A}, \mathbf{B} \rangle_F| = \left| \sum_{i,j} A_{i,j} B_{i,j} \right| \leq \sum_i \|\mathbf{A}_{i,:}\|_2 \|\mathbf{B}_{i,:}\|_2 \leq \left(\max_i \|\mathbf{A}_{i,:}\|_2 \right) \cdot \sum_i \|\mathbf{B}_{i,:}\|_2 \leq \|\mathbf{A}\| \, \|\mathbf{B}\|_{1,2},$$

for all matrices $\mathbf{A}, \mathbf{B}$. ∎

C.5 Proof of Lemma B.9

Note that by Lemma B.2 and (27), $S := \operatorname{supp}(\mathsf{H}_s(\mathbf{X}^{(k)})) = \operatorname{supp}(\mathbf{X}_\star)$. Since by assumption $\delta_k \leq \rho_s(\mathbf{X}^{(k)})$ we have by definition of $\mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}$ that $\mathbf{Z} := \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k} \cdot \mathbf{X}_\star$ is a matrix with row-support S and rows

$$\mathbf{Z}_{i,:} = \min\left\{ \frac{\delta_k^2}{\left\| (\mathbf{X}^{(k)})_{i,:} \right\|_2^2}, 1 \right\} (\mathbf{X}_\star)_{i,:}, = \frac{\delta_k^2}{\left\| (\mathbf{X}^{(k)})_{i,:} \right\|_2^2} (\mathbf{X}_\star)_{i,:}$$

for $i \in S$. Now note that if (27) holds, then

$$\|\mathbf{Z}_{i,:}\|_2 = \frac{\delta_k^2 \|(\mathbf{X}_\star)_{i,:}\|_2}{\|(\mathbf{X}^{(k)})_{i,:}\|_2^2} \leq \frac{\delta_k^2}{(1-\zeta)^2 \rho_s(\mathbf{X}_\star)},$$

where we used in the last estimate that with (27) and $\|(\mathbf{X}_\star)_{i,:}\|_2 \geq \rho_s(\mathbf{X}_\star)$, for $i \in S$, we have

$$\begin{aligned}
\|(\mathbf{X}^{(k)})_{i,:}\|_2^2 &\geq (\|(\mathbf{X}_\star)_{i,:}\|_2 - \|(\mathbf{X}^{(k)})_{i,:} - (\mathbf{X}_\star)_{i,:}\|_2)^2 \geq (\|(\mathbf{X}_\star)_{i,:}\|_2 - \zeta \rho_s(\mathbf{X}_\star))^2 \\
&\geq \Big((1-\zeta)\rho_s(\mathbf{X}_\star) \Big) \Big((1-\zeta) \|(\mathbf{X}_\star)_{i,:}\|_2 \Big),
\end{aligned}$$

for all $i \in S$. The claim easily follows since $\mathbf{Z}$ has only s non-zero rows.

C.6 Proof of Lemma B.10

In the proof of Lemma B.10, we use a simple technical observation.

Lemma C.3. *Let $\mathbf{X}_\star \in \mathcal{M}_{r,s}$, let $\mathbf{X}^{(k)}$ be the k-th iterate of Algorithm 1, and abbreviate $\Xi^{(k)} = \mathbf{X}^{(k)} - \mathbf{X}_\star$. Then the following two statements hold true:*

1. *If $\varepsilon_k \leq \sigma_{r+1}(\mathbf{X}^{(k)})$, then $\varepsilon_k \leq \|\Xi^{(k)}\|$.*
2. *If $\delta_k \leq \rho_{s+1}(\mathbf{X}^{(k)})$, then $\delta_k \leq \|\Xi^{(k)}\|_{\infty,2}$.*

Proof: By defining $[\mathbf{X}^{(k)}]_r$ to be the best rank-r approximation of $\mathbf{X}^{(k)}$ in any unitarily invariant norm, we bound

$$\varepsilon_k \leq \sigma_{r+1}(\mathbf{X}^{(k)}) = \|\mathbf{X}^{(k)} - [\mathbf{X}^{(k)}]_r\| \leq \|\mathbf{X}^{(k)} - \mathbf{X}_\star\| = \|\mathbf{\Xi}^{(k)}\|,$$

where the inequality follows the fact that $\mathbf{X}_\star$ is a rank-r matrix.
Similarly, for the second statement, we have that

$$\delta_k \leq \rho_{s+1}(\mathbf{X}^{(k)}) = \|\mathbf{X}^{(k)} - \mathsf{H}_s(\mathbf{X}^{(k)})\|_{\infty,2} \leq \|\mathbf{X}^{(k)} - \mathbf{X}_\star\|_{\infty,2} = \|\mathbf{\Xi}^{(k)}\|_{\infty,2},$$

using that $\mathbf{X}_\star$ is s-row sparse. ∎

Proof of Lemma B.10: First, we note that, using Lemma B.3, the observation in Remark B.7 yields

$$\begin{aligned}\|\mathbf{\Xi}\|_F^2 \leq 4c_{\|\mathcal{A}\|_{2\to 2}}^2 \min\left\{\frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k}\right\}^2 \\ \cdot \left\langle \mathbf{\Xi}, \left(\mathbb{P}_{\mathbf{U},\mathbf{V}}^{\perp} W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k} \mathbb{P}_{\mathbf{U},\mathbf{V}}^{\perp} + \mathbb{P}_{S^c} \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k} \mathbb{P}_{S^c}\right)\mathbf{\Xi}\right\rangle\end{aligned} \tag{56}$$

for all $\mathbf{\Xi} \in \mathbb{R}^{n_1 \times n_2}$ as the assumption (28) implies $\|\mathbf{X}^{(k)} - \mathbf{X}_\star\| \leq \min\left\{\frac{1}{48}, \frac{1}{19c}\right\}\sigma_r(\mathbf{X}_\star)$. Thus, this holds in particular also for $\mathbf{\Xi}^{(k)} = \mathbf{X}^{(k)} - \mathbf{X}_\star$. (Recall that $\mathbf{U}$ and $\mathbf{V}$ contain the leading singular vectors of $\mathbf{X}^{(k)}$, see Definition 2.1.) We estimate that

$$\begin{aligned}&\sqrt{\left\langle \mathbb{P}_{\mathbf{U},\mathbf{V}}^{\perp}\mathbf{\Xi}^{(k)}, W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k}\mathbb{P}_{\mathbf{U},\mathbf{V}}^{\perp}\mathbf{\Xi}^{(k)}\right\rangle} = \left\|(W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k})^{1/2}(\mathbb{P}_{\mathbf{U},\mathbf{V}}^{\perp}(\mathbf{\Xi}^{(k)}))\right\|_F \\ &\leq \left\|(W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k})^{1/2}(\mathbb{P}_{\mathbf{U},\mathbf{V}}^{\perp}(\mathbf{X}_\star))\right\|_F + \left\|(W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k})^{1/2}(\mathbb{P}_{\mathbf{U},\mathbf{V}}^{\perp}(\mathbf{X}^{(k)}))\right\|_F \\ &\leq \left\|(W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k})^{1/2}(\mathbf{X}_\star)\right\|_F + \sqrt{\sum_{i=r+1}^{\min(n_1,n_2)} \frac{\sigma_i^2}{\max\left(\frac{\sigma_i^2}{\varepsilon_k^2}, 1\right)}} \leq \left\|(W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k})^{1/2}(\mathbf{X}_\star)\right\|_F + \sqrt{n-r}\varepsilon_k\end{aligned}$$

Furthermore, since $\max(\varepsilon_k, \|\mathbf{\Xi}^{(k)}\|) \leq \frac{1}{48}\sigma_r(\mathbf{X}_\star)$ by assumption and $\varepsilon_k \leq \|\mathbf{\Xi}^{(k)}\|$ by Lemma C.3, we can use a variant of Lemma B.8 to obtain

$$\begin{aligned}\left\|(W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k})^{1/2}(\mathbf{X}_\star)\right\|_F &\leq \frac{48}{47}\left(\sqrt{r}\varepsilon_k + 2\varepsilon_k\frac{\|\mathbf{\Xi}^{(k)}\|_F}{\sigma_r(\mathbf{X}_\star)} + 2\frac{\|\mathbf{\Xi}^{(k)}\|\|\mathbf{\Xi}^{(k)}\|_F}{\sigma_r(\mathbf{X}_\star)}\right) \\ &\leq 1.04\sqrt{r}\varepsilon_k + \frac{4.16\|\mathbf{\Xi}^{(k)}\|}{\sigma_r(\mathbf{X}_\star)}\|\mathbf{\Xi}^{(k)}\|_F.\end{aligned}$$

On the other hand, we note that $\mathbf{\Xi}^{(k)}$ restricted to S^c coincides with the restriction of $\mathbf{X}^{(k)}$ to S^c under assumption (28), cf. Lemma B.2, and therefore

$$\begin{aligned}\left\langle \mathbb{P}_{S^c}\mathbf{\Xi}^{(k)}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_{S^c}\mathbf{\Xi}^{(k)}\right\rangle &= \left\langle \mathbb{P}_{S^c}\mathbf{X}^{(k)}, \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k}\mathbb{P}_{S^c}\mathbf{X}^{(k)}\right\rangle \\ &= \sum_{i=s+1}^{n_1} \frac{\|(\mathbf{X}^{(k)})_{i,:}\|_2^2}{\max\{\|(\mathbf{X}^{(k)})_{i,:}\|_2^2/\delta_k^2, 1\}} \leq (n_1 - s)\delta_k^2.\end{aligned}$$

With the estimate of above, this implies that

$$\begin{aligned}&\left\langle \mathbf{\Xi}, \left(\mathbb{P}_{\mathbf{U},\mathbf{V}}^{\perp} W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k} \mathbb{P}_{\mathbf{U},\mathbf{V}}^{\perp} + \mathbb{P}_{S^c} \mathbf{W}^{sp}_{\mathbf{X}^{(k)},\delta_k} \mathbb{P}_{S^c}\right)\mathbf{\Xi}\right\rangle \\ &\leq \left(\left\|(W^{lr}_{\mathbf{X}^{(k)},\varepsilon_k})^{1/2}(\mathbf{X}_\star)\right\|_F + \sqrt{n-r}\varepsilon_k\right)^2 + (n_1 - s)\delta_k^2 \\ &\leq \frac{13}{4}r\varepsilon_k^2 + \frac{52\|\mathbf{\Xi}^{(k)}\|^2}{\sigma_r^2(\mathbf{X}_\star)}\|\mathbf{\Xi}^{(k)}\|_F^2 + 3(n-r)\varepsilon_k^2 + (n_1 - s)\delta_k^2.\end{aligned}$$

Inserting these estimates into (56), we obtain

$$\|\mathbf{\Xi}^{(k)}\|_F^2 \leq 4c_{\|\mathcal{A}\|_{2\to 2}}^2 \min\left\{\frac{\sigma_{r+1}(\mathbf{X}^{(k)})}{\varepsilon_k}, \frac{\rho_{s+1}(\mathbf{X}^{(k)})}{\delta_k}\right\}^2 \left(\frac{13}{4}n\varepsilon_k^2 + n_1\delta_k^2 + \frac{52\|\mathbf{\Xi}^{(k)}\|^2}{\sigma_r^2(\mathbf{X}_\star)}\|\mathbf{\Xi}^{(k)}\|_F^2\right).$$

If now either one of the two equations $\varepsilon_k = \sigma_{r+1}(\mathbf{X}^{(k)})$ or $\delta_k = \rho_{s+1}(\mathbf{X}^{(k)})$ is true, it follows that

$$\begin{aligned}\|\mathbf{\Xi}^{(k)}\|_F^2 &\leq c_{\|\mathcal{A}\|_{2\to 2}}^2\left(13n\varepsilon_k^2 + 4n_1\delta_k^2\right) + c_{\|\mathcal{A}\|_{2\to 2}}^2\frac{208\|\mathbf{\Xi}^{(k)}\|^2}{\sigma_r^2(\mathbf{X}_\star)}\|\mathbf{\Xi}^{(k)}\|_F^2 \\ &\leq c_{\|\mathcal{A}\|_{2\to 2}}^2\left(13n\varepsilon_k^2 + 4n_1\delta_k^2\right) + \frac{1}{2}\|\mathbf{\Xi}^{(k)}\|_F^2\end{aligned}$$

if the proximity condition $\|\mathbf{\Xi}^{(k)}\| = \|\mathbf{X}^{(k)} - \mathbf{X}_\star\| \leq \frac{1}{21c_{\|\mathcal{A}\|_{2\to 2}}}\sigma_r(\mathbf{X}_\star)$ is satisfied. Rearranging the latter inequality yields the conclusion of Lemma B.10. ∎