A. Auxiliary Lemmas

In this section, we prove Lemma A.1 and a few auxiliary lemmas that we will need for the proofs of Theorem 2.4 and Theorem 3.6.

Lemma A.1. Let $\mathrm{x} \in \mathbb{R}^{d_2}$ be distributed according to distribution $\mathcal{D}$ with $\mathbb{E}_{\mathrm{x}}[\mathrm{x}\mathrm{x}^\top] = \mathrm{I}$. Then, for $\ell(\mathrm{U},\mathrm{V}) := \mathbb{E}_{\mathrm{x}}[\|\mathrm{y} - \mathrm{U}\mathrm{V}^\top\mathrm{x}\|^2]$ and $f(\mathrm{U},\mathrm{V}) := \mathbb{E}_{b,\mathrm{x}}[\|\mathrm{y} - \frac{1}{\theta}\mathrm{U}\,\mathrm{diag}(\mathrm{b})\mathrm{V}^\top\mathrm{x}\|^2]$, it holds that

$$f(\mathrm{U},\mathrm{V}) = \ell(\mathrm{U},\mathrm{V}) + \lambda\sum_{i=1}^{r}\|\mathrm{u}_i\|^2\|\mathrm{v}_i\|^2. \tag{8}$$

Furthermore, $\ell(\mathrm{U},\mathrm{V}) = \|\mathrm{M} - \mathrm{U}\mathrm{V}^\top\|_F^2$.

Proof of Lemma A.1. The proof closely follows (Cavazza et al., 2018). Recall that $\mathrm{y} = \mathrm{M}\mathrm{x}$, for some unknown $\mathrm{M} \in \mathbb{R}^{d_2\times d_1}$. Observe that

$$\begin{aligned} f(\mathrm{U},\mathrm{V}) &= \mathbb{E}_{\mathrm{x}}[\|\mathrm{y}\|^2] + \frac{1}{\theta^2}\mathbb{E}_{\mathrm{b},\mathrm{x}}[\|\mathrm{U}\,\mathrm{diag}(\mathrm{b})\mathrm{V}^\top\mathrm{x}\|^2] \\ &\quad - \frac{2}{\theta}\mathbb{E}_{\mathrm{x}}[\langle \mathrm{M}\mathrm{x}, \mathbb{E}_{\mathrm{b}}[\mathrm{U}\,\mathrm{diag}(\mathrm{b})\mathrm{V}^\top]\mathrm{x}\rangle] \end{aligned} \tag{9}$$

where we used the fact that $\mathrm{y} = \mathrm{M}\mathrm{x}$. We have the following set of equalities for the second term on the right hand side of Equation (9):

$$\begin{aligned} \mathbb{E}_{\mathrm{b},\mathrm{x}}[\|\mathrm{U}\,\mathrm{diag}(\mathrm{b})\mathrm{V}^\top\mathrm{x}\|^2] &= \mathbb{E}_{\mathrm{x}}\sum_{i=1}^{d_2}\mathbb{E}_{\mathrm{b}}\left(\sum_{j=1}^{r}u_{ij}b_j\mathrm{v}_j^\top\mathrm{x}\right)^2 \\ &= \mathbb{E}_{\mathrm{x}}\sum_{i=1}^{d_2}\mathbb{E}_{\mathrm{b}}[\sum_{j,k=1}^{r}u_{ij}u_{ik}b_jb_k(\mathrm{v}_j^\top\mathrm{x})(\mathrm{v}_k^\top\mathrm{x})] \\ &= \mathbb{E}_{\mathrm{x}}\sum_{i=1}^{d_2}\sum_{j,k=1}^{r}u_{ij}u_{ik}(\theta^2\mathbb{1}_{j\neq k} + \theta\mathbb{1}_{j=k})(\mathrm{v}_j^\top\mathrm{x})(\mathrm{v}_k^\top\mathrm{x}) \\ &= \theta^2\mathbb{E}_{\mathrm{x}}[\|\mathrm{U}\mathrm{V}^\top\mathrm{x}\|^2] + (\theta-\theta^2)\mathbb{E}_{\mathrm{x}}\sum_{i=1}^{d_2}\sum_{j=1}^{r}u_{ij}^2(\mathrm{v}_j^\top\mathrm{x})^2 \\ &= \theta^2\mathbb{E}_{\mathrm{x}}[\|\mathrm{U}\mathrm{V}^\top\mathrm{x}\|^2] + (\theta-\theta^2)\sum_{j=1}^{r}\|\mathrm{v}_j\|^2\sum_{i=1}^{d_2}u_{ij}^2 \\ &= \theta^2\mathbb{E}_{\mathrm{x}}[\|\mathrm{U}\mathrm{V}^\top\mathrm{x}\|^2] + (\theta-\theta^2)\sum_{j=1}^{r}\|\mathrm{v}_j\|^2\|\mathrm{u}_j\|^2, \end{aligned} \tag{10}$$

where the second to last equality follows because $\mathbb{E}_{\mathrm{x}}[(\mathrm{v}_j^\top\mathrm{x})^2] = \mathrm{v}_j^\top\mathbb{E}_{\mathrm{x}}[\mathrm{x}\mathrm{x}^\top]\mathrm{v}_j = \|\mathrm{v}_j\|^2$. For the third term in Equation (9) we have:

$$\langle \mathrm{M}\mathrm{x}, \mathbb{E}_{\mathrm{b}}[\mathrm{U}\,\mathrm{diag}(\mathrm{b})\mathrm{V}^\top]\mathrm{x}\rangle = \theta\langle \mathrm{M}\mathrm{x}, \mathrm{U}\mathrm{V}^\top\mathrm{x}\rangle \tag{11}$$

Plugging Equations (10) and (11) into (9), we get

$$\begin{aligned} f(\mathrm{U},\mathrm{V}) &= \mathbb{E}_{\mathrm{x}}[\|\mathrm{y}\|^2] + \mathbb{E}_{\mathrm{x}}[\|\mathrm{U}\mathrm{V}^\top\mathrm{x}\|^2] - 2\mathbb{E}_{\mathrm{x}}\langle \mathrm{M}\mathrm{x}, \mathrm{U}\mathrm{V}^\top\mathrm{x}\rangle \\ &\quad + \frac{1-\theta}{\theta}\sum_{i=1}^{r}\|\mathrm{u}_i\|^2\|\mathrm{v}_i\|^2 \end{aligned} \tag{12}$$

It is easy to check that the first three terms in Equation (12) sum to $\ell(\mathrm{U},\mathrm{V})$. Furthermore, since for any $\mathrm{A} \in \mathbb{R}^{d_2\times d_1}$ it holds that $\|\mathrm{A}\mathrm{x}\|^2 = \|\mathrm{A}\|_F^2$, we should have $\ell(\mathrm{U},\mathrm{V}) = \|\mathrm{M} - \mathrm{U}\mathrm{V}^\top\|_F^2$. □

Lemma A.2. For any pair of integers ρ and r, and for any $\lambda \in \mathbb{R}_+$, it holds that

$$(\mathrm{I}_\rho + \frac{\lambda}{r}\mathbb{1}\mathbb{1}^\top)^{-1} = \mathrm{I}_\rho - \frac{\lambda}{r+\lambda\rho}\mathbb{1}\mathbb{1}^\top.$$

Lemma A.2 is an instance of the Woodbury's matrix identity. Here, we include a proof for completeness.

Proof of Lemma A.2. The proof simply follows from the following set of equations.

$$\begin{aligned} &(\mathrm{I}_\rho + \frac{\lambda}{r}\mathbb{1}\mathbb{1}^\top)(\mathrm{I}_\rho - \frac{\lambda}{r+\lambda\rho}\mathbb{1}\mathbb{1}^\top) \\ &= \mathrm{I}_\rho + \frac{\lambda}{r}\mathbb{1}\mathbb{1}^\top - \frac{\lambda}{r+\lambda\rho}\mathbb{1}\mathbb{1}^\top - \frac{\lambda^2}{r(r+\lambda\rho)}\mathbb{1}\mathbb{1}^\top\mathbb{1}\mathbb{1}^\top \\ &= \mathrm{I}_\rho + \left(\frac{\lambda}{r} - \frac{\lambda}{r+\lambda\rho} - \frac{\rho\lambda^2}{r(r+\lambda\rho)}\right)\mathbb{1}\mathbb{1}^\top = \mathrm{I}_\rho \end{aligned}$$

□

Lemma A.3. Let $\lambda > 0$ be a constant. Let $\mathrm{a} \in \mathbb{R}_+^d$ such that $a_i \geq a_{i+1}$ for all $i \in [d-1]$. For $\mathrm{r} \leq d$, let the function $g : [r] \to \mathbb{R}$ be defined as

$$\begin{aligned} g(\rho) := &\sum_{i=1}^{\rho}\left(\frac{\lambda\sum_{k=1}^{\rho}a_k}{r+\lambda\rho}\right)^2 + \sum_{i=\rho+1}^{d}a_i^2 \\ &+ \frac{\lambda}{r}\left(\sum_{i=1}^{\rho}\left(a_i - \frac{\lambda\sum_{k=1}^{\rho}a_k}{r+\lambda\rho}\right)\right)^2. \end{aligned}$$

Then $g(\rho)$ is monotonically non-increasing in ρ.

Proof of Lemma A.3. Let denote the sum of the top τ elements of a by $\mathrm{h}_\tau = \sum_{i=1}^{\tau}a_i$. Furthermore, let the sum of squared of τ bottom elements of a be denoted by $t_\tau = \sum_{i=\tau+1}^{d}a_i^2$. We can simplify $g(\rho)$ and give it in terms of h_ρ and t_ρ as follows:

$$\begin{aligned} g(\rho) =& \rho\left(\frac{\lambda h_\rho}{r+\lambda\rho}\right)^2 + t_\rho + \frac{\lambda}{r}\left(\left(1 - \frac{\lambda\rho}{r+\lambda\rho}\right)h_\rho\right)^2 \\ =& \frac{\rho\lambda^2 + \lambda r}{(r+\lambda\rho)^2}(h_\rho)^2 + t_\rho \\ =& \frac{\lambda h_\rho^2}{r+\lambda\rho} + t_\rho \end{aligned}$$

It suffices to show that $g(\rho+1) \le g(\rho)$ for all $\rho \in [r-1]$.

$$
\begin{aligned}
g(\rho+1) &= \frac{\lambda h_{\rho+1}^2}{r+\lambda\rho+\lambda} + t_{\rho+1} \\
&= \frac{\lambda}{r+\lambda\rho+\lambda}\left(h_\rho^2 + \lambda_{\rho+1}^2(\mathrm{M}) + 2\lambda_{\rho+1}(\mathrm{M})h_\rho\right) \\
&\quad - \lambda_{\rho+1}^2(\mathrm{M}) + t_\rho \\
&= g(\rho) - \frac{\lambda^2 h_\rho^2}{(r+\lambda\rho)(r+\lambda\rho+\lambda)} - \lambda_{\rho+1}^2(\mathrm{M}) \\
&\quad + \frac{\lambda}{r+\lambda\rho+\lambda}\left(\lambda_{\rho+1}^2(\mathrm{M}) + 2\lambda_{\rho+1}(\mathrm{M})h_\rho\right) \\
&= g(\rho) - \frac{\lambda^2 h_\rho^2}{(r+\lambda\rho)(r+\lambda\rho+\lambda)} - \frac{(r+\lambda\rho)\lambda_{\rho+1}^2(\mathrm{M})}{r+\lambda\rho+\lambda} \\
&\quad + \frac{\lambda}{r+\lambda\rho+\lambda}\left(2\lambda_{\rho+1}(\mathrm{M})h_\rho\right) \\
&= g(\rho) - \frac{\left(\lambda h_\rho - (r+\lambda\rho)\lambda_{\rho+1}^2(\mathrm{M})\right)^2}{(r+\lambda\rho)(r+\lambda\rho+\lambda)} \le g(\rho).
\end{aligned}
$$

Hence $g(\rho)$ is monotonically non-increasing in ρ. □

B. Proofs of Theorems in Section 2

Proof of Theorem 2.2. Consider the matrix $\mathrm{G}_1 := \mathrm{G}_\mathrm{U} - \frac{\operatorname{Tr}\mathrm{G}_\mathrm{U}}{r}\mathrm{I}_r$. We exhibit an orthogonal transformation Q, such that $\mathrm{Q}^\top \mathrm{G}_1 \mathrm{Q}$ is zero on its diagonal. Observe that

$$\mathrm{Q}^\top \mathrm{G}_\mathrm{U} \mathrm{Q} = \mathrm{Q}^\top \mathrm{G}_1 \mathrm{Q} + \frac{\operatorname{Tr}\mathrm{G}_\mathrm{U}}{r}\mathrm{I}_r,$$

so that all diagonal elements of G_U are equal to $\frac{\operatorname{Tr}\mathrm{G}_\mathrm{U}}{r}$, i.e. G_U is equalized.

Our construction closely follows the proof of a classical theorem in matrix analysis, which states that any trace zero matrix is a commutator (Albert and Muckenhoupt, 1957; Kahan, 1999). For the zero trace matrix G_1, we first show that there exists a unit vector w_{11} such that $\mathrm{w}_{11}^\top \mathrm{G}_1 \mathrm{w}_{11} = 0$.

Claim 1. Assume G is a zero trace matrix and let $\mathrm{G} = \sum_{i=1}^r \lambda_i \mathrm{u}_i \mathrm{u}_i^\top$ be an eigendecomposition of G. Then $\mathrm{w} = \frac{1}{\sqrt{r}}\sum_{i=1}^r \mathrm{u}_i$ has a vanishing Rayleigh quotient, that is, $\mathrm{w}^\top \mathrm{G} \mathrm{w} = 0$, and $\|\mathrm{w}\| = 1$.

Proof of Claim 1. First, we notice that w has unit norm

$$\|\mathrm{w}\|^2 = \|\frac{1}{\sqrt{r}}\sum_{i=1}^r \mathrm{u}_i\|^2 = \frac{1}{r}\|\sum_{i=1}^r \mathrm{u}_i\|^2 = \frac{1}{r}\sum_{i=1}^r \|\mathrm{u}_i\|^2 = 1.$$

It is easy to see that w has a zero Rayleigh quotient

$$
\begin{aligned}
\mathrm{w}^\top \mathrm{G} \mathrm{w} &= (\frac{1}{\sqrt{r}}\sum_{i=1}^r \mathrm{u}_i)^\top \mathrm{G} (\frac{1}{\sqrt{r}}\sum_{i=1}^r \mathrm{u}_i) \\
&= \frac{1}{r}\sum_{i,j=1}^r \mathrm{u}_i \mathrm{G} \mathrm{u}_j = \frac{1}{r}\sum_{i=1}^r \lambda_j \mathrm{u}_i^\top \mathrm{u}_j = \frac{1}{r}\sum_{i=1}^r \lambda_i = 0.
\end{aligned}
$$

Let $\mathrm{W}_1 := [\mathrm{w}_{11}, \mathrm{w}_{12}, \cdots, \mathrm{w}_{1d}]$ be such that $\mathrm{W}_1^\top \mathrm{W}_1 = \mathrm{W}_1 \mathrm{W}_1^\top = \mathrm{I}_d$. Observe that $\mathrm{W}_1^\top \mathrm{G}_1 \mathrm{W}_1$ has zero on its first diagonal elements

$$\mathrm{W}_1^\top \mathrm{G}_1 \mathrm{W}_1 = \begin{bmatrix} 0 & \mathrm{b}_1^\top \\ \mathrm{b}_1 & \mathrm{G}_2 \end{bmatrix}$$

The principal submatrix G_2 also has a zero trace. With a similar argument, let $\mathrm{w}_{22} \in \mathbb{R}^{d-1}$ be such that $\|\mathrm{w}_{22}\| = 1$ and $\mathrm{w}_{22}^\top \mathrm{G}_2 \mathrm{w}_{22} = 0$ and define $\mathrm{W}_2 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & \mathrm{w}_{22} & \mathrm{w}_{23} & \cdots & \mathrm{w}_{2d} \end{bmatrix} \in \mathbb{R}^{d\times d}$ such that $\mathrm{W}_2^\top \mathrm{W}_2 = \mathrm{W}_2 \mathrm{W}_2^\top = \mathrm{I}_d$, and observe that

$$(\mathrm{W}_1\mathrm{W}_2)^\top \mathrm{G}_1 (\mathrm{W}_1\mathrm{W}_2) = \begin{bmatrix} 0 & \cdot & \cdots \\ \cdot & 0 & \cdots \\ \vdots & \vdots & \mathrm{G}_2 \end{bmatrix}.$$

This procedure can be applied recursively so that for the *equalizer* $\mathrm{Q} = \mathrm{W}_1 \mathrm{W}_2 \cdots \mathrm{W}_d$ we have

$$\mathrm{Q}^\top \mathrm{G}_1 \mathrm{Q} = \begin{bmatrix} 0 & \cdot & \cdots & \cdot \\ \cdot & 0 & \cdots & \cdot \\ \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdot & 0 \end{bmatrix}.$$

□

Proof of Theorem 2.3. Let us denote the squared column norms of U by $\mathrm{n}_\mathrm{u} = (\|\mathrm{u}_1\|^2, \ldots, \|\mathrm{u}_r\|^2)$. Observe that for any weight matrix U:

$$
\begin{aligned}
R(\mathrm{U},\mathrm{U}) &= \lambda \sum_{i=1}^r \|\mathrm{u}_i\|^4 = \frac{\lambda}{r}\|1_r\|^2 \|\mathrm{n}_\mathrm{u}\|^2 \\
&\ge \frac{\lambda}{r}\langle 1_r, \mathrm{n}_\mathrm{u}\rangle^2 = \frac{\lambda}{r}\left(\sum_{i=1}^r \|\mathrm{u}_i\|^2\right)^2 = \frac{\lambda}{r}\|\mathrm{U}\|_F^4,
\end{aligned}
$$

where $1_r \in \mathbb{R}^r$ is the vector of all ones and the inequality is due to Cauchy-Schwartz. Hence, the regularizer is lower bounded by $\frac{\lambda}{r}\|\mathrm{U}\|_F^4$, with equality if and only if n_u is parallel to 1_r, i.e. when U is equalized. Now, if U is not equalized, by Theorem 2.2 there exist a rotation matrix Q such that UQ is equalized, which implies $R(\mathrm{UQ},\mathrm{UQ}) < R(\mathrm{U},\mathrm{U})$. Together with rotational invariance of the loss function, this gives a contradiction with global optimality U. Hence, if U is a global optimum then it is equalized and we have $R(\mathrm{U},\mathrm{U}) = \lambda\sum_{i=1}^r \|\mathrm{u}_i\|^4 = \frac{\lambda}{r}\|\mathrm{U}\|_F^4$. □

Proof of Theorem 2.4. By Theorem 2.3, if W is an optimum of Problem 4, then it holds that $\lambda\sum_{i=1}^r \|\mathrm{w}_i\|^4 = \frac{\lambda}{r}\|\mathrm{W}\|_F^4$. Also, by Theorem 2.2, it is always possible to equalize

any given weight matrix. Hence, Problem 4 reduces to the following problem:

$$\min_{\mathrm{W}\in\mathbb{R}^{d\times r}} \|\mathrm{M}-\mathrm{W}\mathrm{W}^\top\|_F^2+\frac{\lambda}{r}\|\mathrm{W}\|_F^4 \tag{13}$$

Let $\mathrm{M}=\mathrm{U}_\mathrm{M}\Lambda_\mathrm{M}\mathrm{U}_\mathrm{M}^\top$ and $\mathrm{W}=\mathrm{U}_\mathrm{W}\Sigma_\mathrm{W}\mathrm{V}_\mathrm{W}^\top$ be an eigen-decomposition of M and a full SVD of W respectively, such that $\lambda_i(\mathrm{M})\geq\lambda_{i+1}(\mathrm{M})$ and $\sigma_i(\mathrm{W})\geq\sigma_{i+1}(\mathrm{W})$ for all $i\in[d-1]$. Rewriting objective of Problem 13 in terms of these decompositions gives:

$$\begin{aligned}
&\|\mathrm{M}-\mathrm{W}\mathrm{W}^\top\|_F^2+\frac{\lambda}{r}\|\mathrm{W}\|_F^4\\
&=\|\mathrm{U}_\mathrm{M}\Lambda_\mathrm{M}\mathrm{U}_\mathrm{M}^\top-\mathrm{U}_\mathrm{W}\Sigma_\mathrm{W}\Sigma_\mathrm{W}^\top\mathrm{U}_\mathrm{W}^\top\|_F^2+\frac{\lambda}{r}\|\mathrm{U}_\mathrm{W}\Sigma_\mathrm{W}\mathrm{V}_\mathrm{W}^\top\|_F^4\\
&=\|\Lambda_\mathrm{M}-\mathrm{U}'\Sigma_\mathrm{W}\Sigma_\mathrm{W}^\top\mathrm{U}'^\top\|_F^2+\frac{\lambda}{r}\|\Sigma_\mathrm{W}\|_F^4\\
&=\|\Lambda_\mathrm{M}\|_F^2+\|\Lambda_\mathrm{W}\|_F^2-2\langle\Lambda_\mathrm{M},\mathrm{U}'\Lambda_\mathrm{W}\mathrm{U}'^\top\rangle+\frac{\lambda}{r}\left(\mathrm{Tr}(\Lambda_\mathrm{W})\right)^2
\end{aligned}$$

where $\Lambda_\mathrm{W}:=\Sigma_\mathrm{W}\Sigma_\mathrm{W}^\top$ and $\mathrm{U}'=\mathrm{U}_\mathrm{M}^\top\mathrm{U}_\mathrm{W}$. By Von Neumann's trace inequality, for a fixed Σ_W we have that

$$\langle\Lambda_\mathrm{M},\mathrm{U}'\Lambda_W\mathrm{U}'^\top\rangle\leq\sum_{i=1}^d\lambda_i(\mathrm{M})\lambda_i(\mathrm{W}),$$

where the equality is achieved when $\Lambda_i(\mathrm{W})$ have the same ordering as $\Lambda_i(\mathrm{M})$ and $\mathrm{U}'=\mathrm{I}$, i.e. $\mathrm{U}_\mathrm{M}=\mathrm{U}_\mathrm{W}$. Now, Problem 13 is reduced to

$$\begin{aligned}
&\min_{\substack{\|\Lambda_\mathrm{W}\|_0\leq r,\\ \Lambda_\mathrm{W}\geq 0}}\|\Lambda_\mathrm{M}-\Lambda_\mathrm{W}\|_F^2+\frac{\lambda}{r}\left(\mathrm{Tr}(\Lambda_\mathrm{W})\right)^2\\
&=\min_{\bar\lambda\in\mathbb{R}_+^r}\sum_{i=1}^r\left(\lambda_i(\mathrm{M})-\bar\lambda_i\right)^2+\sum_{i=r+1}^d\lambda_i^2(\mathrm{M})+\frac{\lambda}{r}\left(\sum_{i=1}^r\bar\lambda_i\right)^2
\end{aligned}$$

The Lagrangian is given by

$$\begin{aligned}
L(\bar\lambda,\alpha)=&\sum_{i=1}^r\left(\lambda_i(\mathrm{M})-\bar\lambda_i\right)^2+\sum_{i=r+1}^d\lambda_i^2(\mathrm{M})\\
&+\frac{\lambda}{r}\left(\sum_{i=1}^r\bar\lambda_i\right)^2-\sum_{i=1}^r\alpha_i\bar\lambda_i
\end{aligned}$$

The KKT conditions ensures that at the optima it holds for all $i\in[r]$ that

$$\begin{aligned}
&\bar\lambda_i\geq 0,\ \alpha_i\geq 0,\ \bar\lambda_i\alpha_i=0\\
&2(\bar\lambda_i-\lambda_i(\mathrm{M}))+\frac{2\lambda}{r}\left(\sum_{i=1}^r\bar\lambda_i\right)-\alpha_i=0
\end{aligned}$$

Let $\rho=|i:\bar\lambda_i>0|\leq r$ be the number of nonzero $\bar\lambda_i$. For $i=1,\dots,\rho$ we have $\alpha_i=0$, hence

$$\begin{aligned}
&\bar\lambda_i+\frac{\lambda}{r}\left(\sum_{i=1}^\rho\bar\lambda_i\right)=\lambda_i(\mathrm{M})\\
&\implies(\mathrm{I}_\rho+\frac{\lambda}{r}11^\top)\bar\lambda_{1:\rho}=\lambda_{1:\rho}(\mathrm{M})\\
&\implies\bar\lambda_{1:\rho}=(\mathrm{I}_\rho-\frac{\lambda}{r+\lambda\rho}11^\top)\lambda_{1:\rho}(\mathrm{M})\\
&\implies\bar\lambda_{1:\rho}=\lambda_{1:\rho}(\mathrm{M})-\frac{\lambda\rho\kappa_\rho}{r+\lambda\rho}1_\rho\\
&\implies\Lambda_\mathrm{W}=(\Lambda_\mathrm{M}-\frac{\lambda\rho\kappa_\rho}{r+\lambda\rho}\mathrm{I}_d)_+
\end{aligned}$$

where $\kappa_\rho:=\frac{1}{\rho}\sum_{i=1}^\rho\lambda_i(\mathrm{M})$ and the second implication is due to Lemma A.2. It only remains to find the optimal ρ. Let's define the function

$$\begin{aligned}
g(\rho):=&\sum_{i=1}^\rho\left(\lambda_i(\mathrm{M})-\bar\lambda_i\right)^2+\sum_{i=\rho+1}^d\lambda_i^2(\mathrm{M})+\frac{\lambda}{r}\left(\sum_{i=1}^\rho\bar\lambda_i\right)^2\\
=&\sum_{i=1}^\rho\left(\frac{\lambda\sum_{k=1}^\rho\lambda_k(\mathrm{M})}{r+\lambda\rho}\right)^2+\sum_{i=\rho+1}^d\lambda_i(\mathrm{M})^2\\
&+\frac{\lambda}{r}\left(\sum_{i=1}^\rho\left(\lambda_i(\mathrm{M})-\frac{\lambda\sum_{k=1}^\rho\lambda_k(\mathrm{M})}{r+\lambda\rho}\right)\right)^2.
\end{aligned}$$

By Lemma A.3, $g(\rho)$ is monotonically non-increasing in ρ, hence ρ should be the largest *feasible* integer, i.e.

$$\rho=\max\{j:\ \lambda_j>\frac{\lambda j\kappa}{r+\lambda j}\}.$$

□

Proof of Remark 5.2. For $\tilde{\mathrm{U}}$ to have equal column norms, it suffices to show that $\tilde{\mathrm{U}}^\top\tilde{\mathrm{U}}$ is constant on its diagonal. Next, we note that

$$\begin{aligned}
\tilde{\mathrm{U}}^\top\tilde{\mathrm{U}}&=\mathrm{Q}^\top\mathrm{U}^\top\mathrm{U}\mathrm{Q}\\
&=(\mathrm{V}\mathrm{Z}_k)^\top(\mathrm{W}\Sigma\mathrm{V}^\top)^\top(\mathrm{W}\Sigma\mathrm{V}^\top)(\mathrm{V}\mathrm{Z}_k)\\
&=\mathrm{Z}_k^\top\mathrm{V}^\top\mathrm{V}\Sigma\mathrm{W}^\top\mathrm{W}\Sigma\mathrm{V}^\top\mathrm{V}\mathrm{Z}_k\\
&=\mathrm{Z}_k^\top\Sigma^2\mathrm{Z}_k
\end{aligned}$$

It remains to show that for any diagonal matrix D, $\mathrm{Z}_k^\top\mathrm{D}\mathrm{Z}_k$ is diagonalized. First note that

$$\mathrm{Z}_2\mathrm{Z}_2^\top=\frac{1}{2}\begin{bmatrix}1&1\\-1&1\end{bmatrix}\begin{bmatrix}1&-1\\1&1\end{bmatrix}=\mathrm{I}_2$$

so that Z_2 is indeed a rotation. By induction, it is easy to see that Z_k is a rotation for all k. Now, we show that Z_k equalizes any diagonal matrix D. Observe that

$$[\mathrm{Z}_k^\top\mathrm{D}\mathrm{Z}_k]_{ii}=\sum_{i=1}^{2^{k-1}}D_{ii}z_{ji}^2=\sum_{i=1}^{2^{k-1}}D_{ii}2^{-k+1}=2^{1-k}\mathrm{Tr}\,\mathrm{D}$$

so that all the diagonal elements are identically equal to the average of the diagonal elements of D. □

C. Proofs of Theorems in Section 3

Proof of Theorem 3.3. Let $\mathrm{U}\mathrm{V}^\top = \mathrm{W}\Sigma\mathrm{Y}^\top$ be a compact SVD of $\mathrm{U}\mathrm{V}^\top$. Define $\tilde{\mathrm{U}} := \mathrm{W}\Sigma^{1/2}$ and $\tilde{\mathrm{V}} := \mathrm{Y}\Sigma^{1/2}$ and observe that $\tilde{\mathrm{U}}\tilde{\mathrm{V}}^\top = \mathrm{U}\mathrm{V}^\top$. Furthermore, let $\mathrm{G}_{\tilde{\mathrm{U}}} = \tilde{\mathrm{U}}^\top\tilde{\mathrm{U}}$ and $\mathrm{G}_{\tilde{\mathrm{V}}} = \tilde{\mathrm{V}}^\top\tilde{\mathrm{V}}$ be their Gram matrices. Observe that $\mathrm{G}_{\tilde{\mathrm{U}}} = \mathrm{G}_{\tilde{\mathrm{V}}} = \Sigma$. Hence, by Theorem 2.2, there exists a rotation Q such that $\bar{\mathrm{V}} := \tilde{\mathrm{V}}\mathrm{Q}$ and $\bar{\mathrm{U}} := \tilde{\mathrm{U}}\mathrm{Q}$ are equalized, with $\|\bar{\mathrm{u}}_i\|^2 = \|\bar{\mathrm{v}}_i\|^2 = \frac{1}{r}\operatorname{Tr}\Sigma$. □

Proof of Theorem 3.4. Define

$$\mathrm{n}_{\mathrm{u},\mathrm{v}} = (\|\mathrm{u}_1\|\|\mathrm{v}_1\|, \ldots, \|\mathrm{u}_r\|\|\mathrm{v}_r\|)$$

and observe that

$$\begin{aligned} R(\mathrm{U},\mathrm{V}) &= \lambda\sum_{i=1}^r \|\mathrm{u}_i\|^2\|\mathrm{v}_i\|^2 \\ &= \frac{\lambda}{r}\|\mathrm{n}_{\mathrm{u},\mathrm{v}}\|^2\|1_r\|^2 \geq \frac{\lambda}{r}\langle \mathrm{n}_{\mathrm{u},\mathrm{v}}, 1_r\rangle^2 \\ &= \frac{\lambda}{r}\left(\sum_{i=1}^r \|\mathrm{u}_i\|\|\mathrm{v}_i\|\right)^2 \end{aligned}$$

where the inequality is due to Cauchy-Schwartz, and it holds with equality if and only if $\mathrm{n}_{\mathrm{u},\mathrm{v}}$ is parallel to 1_r. Let $(\bar{\mathrm{U}}, \bar{\mathrm{V}})$ be a global optima of Problem 6. The inequality above together with Theorem 3.3 imply that $\bar{\mathrm{U}}$ and $\bar{\mathrm{V}}$ should be jointly equalized up to dilation transformations, hence the first equality claimed by the theorem.

To see the second equality, note that if U and V are jointly equalized, then

$$\|\mathrm{u}_i\|^2 = \|\mathrm{v}_i\|^2 = \frac{1}{r}\operatorname{Tr}\Sigma,$$

where Σ is the matrix of singular values of $\mathrm{U}\mathrm{V}^\top$. Hence,

$$\begin{aligned} R(\mathrm{U},\mathrm{V}) = \frac{\lambda}{r}\left(\sum_{i=1}^r \|\mathrm{u}_i\|\|\mathrm{v}_i\|\right)^2 &= \frac{\lambda}{r}\left(\frac{1}{r}\sum_{i=1}^r \operatorname{Tr}\Sigma\right)^2 \\ &= \frac{\lambda}{r}\left(\operatorname{Tr}\Sigma\right)^2 \end{aligned}$$

which is equal to $\frac{\lambda}{r}\|\bar{\mathrm{U}}\bar{\mathrm{V}}^\top\|_*^2$ as claimed. □

Proof of Theorem 3.6. By Theorem 3.4, if (X,Y) is an optimum of Problem 6, then it holds that

$$\lambda\sum_{i=1}^r \|\mathrm{x}_i\|^2\|\mathrm{y}_i\|^2 = \frac{\lambda}{r}\|\mathrm{X}\mathrm{Y}^\top\|_*^2.$$

Hence, Problem 6 reduces to the following problem:

$$\min_{\mathrm{X}\in\mathbb{R}^{d_1\times r},\mathrm{Y}\in\mathbb{R}^{d_2\times r}} \|\mathrm{M} - \mathrm{X}\mathrm{Y}^\top\|_F^2 + \frac{\lambda}{r}\|\mathrm{X}\mathrm{Y}^\top\|_*^2 \tag{14}$$

Let $\mathrm{M} = \mathrm{U}_\mathrm{M}\Sigma_\mathrm{M}\mathrm{V}_\mathrm{M}^\top$ and $\mathrm{W} := \mathrm{X}\mathrm{Y}^\top = \mathrm{U}_\mathrm{W}\Sigma_\mathrm{W}\mathrm{V}_\mathrm{W}^\top$ be full SVDs of M and W respectively, such that $\sigma_i(\mathrm{M}) \geq \sigma_{i+1}(\mathrm{M})$ and $\sigma_i(\mathrm{W}) \geq \sigma_{i+1}(\mathrm{W})$ for all $i \in [d-1]$ where $d = \min\{d_1, d_2\}$. Rewriting objective of Problem 14 in terms of these decompositions,

$$\begin{aligned} &\|\mathrm{M} - \mathrm{X}\mathrm{Y}^\top\|_F^2 + \frac{\lambda}{r}\|\mathrm{X}\mathrm{Y}^\top\|_*^2 \\ &= \|\mathrm{U}_\mathrm{M}\Sigma_\mathrm{M}\mathrm{V}_\mathrm{M}^\top - \mathrm{U}_\mathrm{W}\Sigma_\mathrm{W}\mathrm{V}_\mathrm{W}^\top\|_F^2 + \frac{\lambda}{r}\|\mathrm{U}_\mathrm{W}\Sigma_\mathrm{W}\mathrm{V}_\mathrm{W}^\top\|_*^2 \\ &= \|\Sigma_\mathrm{M} - \mathrm{U}'\Sigma_\mathrm{W}\mathrm{V}'^\top\|_F^2 + \frac{\lambda}{r}\|\Sigma_\mathrm{W}\|_*^2 \\ &= \|\Sigma_\mathrm{M}\|_F^2 + \|\Sigma_\mathrm{W}\|_F^2 - 2\langle\Sigma_\mathrm{M}, \mathrm{U}'\Sigma_\mathrm{W}\mathrm{U}'^\top\rangle + \frac{\lambda}{r}\|\Sigma_\mathrm{W}\|_*^2 \end{aligned}$$

where $\mathrm{U}' = \mathrm{U}_\mathrm{M}^\top\mathrm{U}_\mathrm{W}$. By Von Neumann's trace inequality, for a fixed Σ_W we have that $\langle\Sigma_\mathrm{M}, \mathrm{U}'\Sigma_\mathrm{W}\mathrm{U}'^\top\rangle \leq \sum_{i=1}^d \sigma_i(\mathrm{M})\sigma_i(\mathrm{W})$, where the equality is achieved when $\Sigma_i(\mathrm{W})$ have the same ordering as $\Sigma_i(\mathrm{M})$ and $\mathrm{U}' = \mathrm{I}$, i.e. $\mathrm{U}_\mathrm{M} = \mathrm{U}_\mathrm{W}$. Now, Problem 14 is reduced to

$$\begin{aligned} &\min_{\substack{\|\Sigma_\mathrm{W}\|_0\leq r,\\ \Sigma_\mathrm{W}\geq 0}} \|\Sigma_\mathrm{M} - \Sigma_\mathrm{W}\|_F^2 + \frac{\lambda}{r}\|\Sigma_\mathrm{W}\|_*^2 \\ &= \min_{\bar{\sigma}\in\mathbb{R}_+^r} \sum_{i=1}^r (\sigma_i(\mathrm{M}) - \bar{\sigma}_i)^2 + \sum_{i=r+1}^d \sigma_i^2(\mathrm{M}) + \frac{\lambda}{r}\left(\sum_{i=1}^r \bar{\sigma}_i\right)^2 \end{aligned}$$

The Lagrangian is given by

$$\begin{aligned} L(\bar{\lambda}, \alpha) = &\sum_{i=1}^r (\sigma_i(\mathrm{M}) - \bar{\sigma}_i)^2 + \sum_{i=r+1}^d \sigma_i^2(\mathrm{M}) \\ &+ \frac{\lambda}{r}\left(\sum_{i=1}^r \bar{\sigma}_i\right)^2 - \sum_{i=1}^r \alpha_i\bar{\sigma}_i \end{aligned}$$

The KKT conditions ensures that $\forall i = 1, \ldots, r$,

$$\begin{aligned} &\bar{\sigma}_i \geq 0,\ \alpha_i \geq 0,\ \bar{\sigma}_i\alpha_i = 0 \\ &2(\bar{\sigma}_i - \sigma_i(\mathrm{M})) + \frac{2\lambda}{r}\left(\sum_{i=1}^r \bar{\sigma}_i\right) - \alpha_i = 0 \end{aligned}$$

Let $\rho = |i : \bar{\sigma}_i > 0| \leq r$ be the number of nonzero $\bar{\sigma}_i$. For

$i = 1, \dots, \rho$ we have $\alpha_i = 0$, hence

$$\begin{aligned}
&\bar{\sigma}_i + \frac{\lambda}{r}\left(\sum_{i=1}^{\rho}\bar{\sigma}_i\right) = \sigma_i(\mathrm{M}) \\
&\implies (\mathrm{I}_\rho + \frac{\lambda}{r}11^\top)\bar{\sigma}_{1:\rho} = \sigma_{1:\rho}(\mathrm{M}) \\
&\implies \bar{\sigma}_{1:\rho} = (\mathrm{I}_\rho - \frac{\lambda}{r+\lambda\rho}11^\top)\sigma_{1:\rho}(\mathrm{M}) \\
&\implies \bar{\sigma}_{1:\rho} = \sigma_{1:\rho}(\mathrm{M}) - \frac{\lambda\rho\kappa_\rho}{r+\lambda\rho}1_\rho \\
&\implies \Sigma_{\mathrm{W}} = (\Sigma_{\mathrm{M}} - \frac{\lambda\rho\kappa_\rho}{r+\lambda\rho}\mathrm{I}_d)_+
\end{aligned}$$

where $\kappa_\rho = \frac{1}{\rho}\sum_{i=1}^{\rho}\sigma_i(\mathrm{M})$ and the second implication holds since $(\mathrm{I}_\rho + \frac{\lambda}{r}11^\top)^{-1} = \mathrm{I}_\rho - \frac{\lambda}{r+\lambda\rho}11^\top$. It only remains to find the optimal ρ. Let's define the function

$$\begin{aligned}
g(\rho) :=& \sum_{i=1}^{\rho}(\sigma_i(\mathrm{M}) - \bar{\sigma}_i)^2 + \sum_{i=\rho+1}^{d}\sigma_i^2(\mathrm{M}) + \frac{\lambda}{r}\left(\sum_{i=1}^{\rho}\bar{\sigma}_i\right)^2 \\
=& \sum_{i=1}^{\rho}\left(\frac{\lambda\sum_{k=1}^{\rho}\sigma_k(\mathrm{M})}{r+\lambda\rho}\right)^2 + \sum_{i=\rho+1}^{d}\sigma_i(\mathrm{M})^2 \\
&+ \frac{\lambda}{r}\left(\sum_{i=1}^{\rho}\left(\sigma_i(\mathrm{M}) - \frac{\lambda\sum_{k=1}^{\rho}\sigma_k(\mathrm{M})}{r+\lambda\rho}\right)\right)^2.
\end{aligned}$$

By Lemma A.3, $g(\rho)$ is monotonically non-increasing in ρ, hence ρ should be the largest *feasible* integer, i.e.

$$\rho = \max\{j : \sigma_j > \frac{\lambda j\kappa_j}{r+\lambda j}\}.$$

□

D. Proofs of Theorems in Sections 4

In this section for ease of notation we let λ_i denote $\lambda_i(\mathrm{M})$. Furthermore, with slight abuse of notation we let $f(\mathrm{U})$, $\ell(\mathrm{U})$ and $R(\mathrm{U})$ denote the objective, the loss function and the regularizer, respectively.

It is easy to see that the gradient of the objective of Problem 4 is given by

$$\nabla f(\mathrm{U}) = 4(\mathrm{U}\mathrm{U}^\top - \mathrm{M})\mathrm{U} + 4\lambda\mathrm{U}\,\mathrm{diag}(\mathrm{U}^\top\mathrm{U}).$$

We first make the following important observation about the critical points of Problem 4.

Lemma D.1. If U is a critical point of Problem 4, then it holds that $\mathrm{U}\mathrm{U}^\top \preceq \mathrm{M}$.

Proof of Lemma D.1. Since $\nabla f(\mathrm{U}) = 0$, we have that

$$(\mathrm{M} - \mathrm{U}\mathrm{U}^\top)\mathrm{U} = \lambda\mathrm{U}\,\mathrm{diag}(\mathrm{U}^\top\mathrm{U})$$

multiply both sides from right by $\mathrm{U}^\top$ and rearrange to get

$$\mathrm{M}\mathrm{U}\mathrm{U}^\top = \mathrm{U}\mathrm{U}^\top\mathrm{U}\mathrm{U}^\top + \lambda\mathrm{U}\,\mathrm{diag}(\mathrm{U}^\top\mathrm{U})\mathrm{U}^\top \tag{15}$$

Note that the right hand side is symmetric, which implies that the left hand side must be symmetric as well, i.e.

$$\mathrm{M}\mathrm{U}\mathrm{U}^\top = (\mathrm{M}\mathrm{U}\mathrm{U}^\top)^\top = \mathrm{U}\mathrm{U}^\top\mathrm{M},$$

so that M and $\mathrm{U}\mathrm{U}^\top$ commute. Note that in Equation (15), $\mathrm{U}\,\mathrm{diag}(\mathrm{U}^\top\mathrm{U})\mathrm{U}^\top \succeq 0$. Thus, $\mathrm{M}\mathrm{U}\mathrm{U}^\top \succeq \mathrm{U}\mathrm{U}^\top\mathrm{U}\mathrm{U}^\top$. Let $\mathrm{U}\mathrm{U}^\top = \mathrm{W}\Gamma\mathrm{W}^\top$ be a compact eigendecomposition of $\mathrm{U}\mathrm{U}^\top$. We get

$$\mathrm{M}\mathrm{U}\mathrm{U}^\top = \mathrm{M}\mathrm{W}\Gamma\mathrm{W}^\top \succeq \mathrm{U}\mathrm{U}^\top\mathrm{U}\mathrm{U}^\top = \mathrm{W}\Gamma^2\mathrm{W}^\top.$$

Multiplying from right and left by $\mathrm{W}\Gamma^{-1}$ and $\mathrm{W}^\top$ respectively, we have that

$$\mathrm{W}^\top\mathrm{M}\mathrm{W} \succeq \Gamma$$

which completes the proof. □

Lemma D.1 allows us to bound different norms of the critical points, as will be seen later in the proofs.

To explore the landscape properties of Problem 4, we first focus on the non-equalized critical points in Lemma D.2. We show that the set of non-equalized critical points does not include any local optima. Furthermore, all such points are strict saddles. Therefore, we turn our focus to the equalized critical points in Lemma D.3. We show all such points inherit the eigenspace of the input matrix M. This allows us to give a closed-form characterization of all the equalized critical points in terms of the eigendecompostion of M. We then show that if λ is chosen appropriately, all such critical points that are not global optima, are strict saddle points.

Lemma D.2. All local minima of Problem 4 are equalized. Moreover, all critical points that are not equalized, are strict saddle points.

Proof of Lemma D.2. We show that if U is not equalized, then any ϵ-neighborhood of U contains a point with objective strictly smaller than $f(\mathrm{U})$. More formally, for any $\epsilon > 0$, we exhibit a rotation Q_ϵ such that $\|\mathrm{U} - \mathrm{U}\mathrm{Q}_\epsilon\|_F \le \epsilon$ and $f(\mathrm{U}\mathrm{Q}_\epsilon) < f(\mathrm{U})$. Let U be a critical point of Problem 4 that is not equalized, i.e. there exists two columns of U with different norms. Without loss of generality, let $\|\mathrm{u}_1\| > \|\mathrm{u}_2\|$. We design a rotation matrix Q such that it is almost an isometry, but it moves mass from u_1 to u_2. Consequently, the new factor becomes "less un-equalized" and achieves a smaller regularizer, while preserving the value of the loss. To that end, define

$$\mathrm{Q}_\delta := \begin{bmatrix} \sqrt{1-\delta^2} & -\delta & 0 \\ \delta & \sqrt{1-\delta^2} & 0 \\ 0 & 0 & \mathrm{I}_{r-2} \end{bmatrix}$$

and let $\hat{\mathrm{U}} := \mathrm{U}\mathrm{Q}_\delta$. It is easy to verify that Q_ϵ is indeed a rotation. First, we show that for any ϵ, as long as $\delta^2 \leq \frac{\epsilon^2}{2\,\mathrm{Tr}(\mathrm{M})}$, we have $\hat{\mathrm{U}} \in \mathcal{B}_\epsilon(\mathrm{U})$:

$$\begin{aligned}\|\mathrm{U}-\hat{\mathrm{U}}\|_F^2 &= \sum_{i=1}^r \|\mathrm{u}_i - \hat{\mathrm{u}}_i\|^2 \\ &= \|\mathrm{u}_1 - \sqrt{1-\delta^2}\mathrm{u}_1 - \delta \mathrm{u}_2\,\|^2 \\ &+ \|\mathrm{u}_2 - \sqrt{1-\delta^2}\mathrm{u}_2 + \delta \mathrm{u}_1\,\|^2 \\ &= 2(1-\sqrt{1-\delta^2})(\|\mathrm{u}_1\|^2 + \|\mathrm{u}_2\|^2) \\ &\leq 2\delta^2\,\mathrm{Tr}(\mathrm{M}) \leq \epsilon^2\end{aligned}$$

where the second to last inequality follows from Lemma D.1, because $\|\mathrm{u}_1\|^2 + \|\mathrm{u}_2\|^2 \leq \|\mathrm{U}\|_F^2 = \mathrm{Tr}(\mathrm{U}\mathrm{U}^\top) \leq \mathrm{Tr}(\mathrm{M})$, and also the fact that $1-\sqrt{1-\delta^2} = \frac{1-1+\delta^2}{1+\sqrt{1-\delta^2}} \leq \delta^2$.

Next, we show that for small enough δ, the value of the function at $\hat{\mathrm{U}}$ is strictly smaller than that of U. Observe that

$$\begin{aligned}\|\hat{\mathrm{u}}_1\|^2 &= (1-\delta^2)\|\mathrm{u}_1\|^2 + \delta^2\|\mathrm{u}_2\|^2 + 2\delta\sqrt{1-\delta^2}\mathrm{u}_1^\top\mathrm{u}_2 \\ \|\hat{\mathrm{u}}_2\|^2 &= (1-\delta^2)\|\mathrm{u}_2\|^2 + \delta^2\|\mathrm{u}_1\|^2 - 2\delta\sqrt{1-\delta^2}\mathrm{u}_1^\top\mathrm{u}_2\end{aligned}$$

and the remaining columns will not change, i.e. for $i = 3, \cdots, r$, $\hat{\mathrm{u}}_i = \mathrm{u}_i$. Together with the fact that Q_δ preserves the norms, i.e. $\|\mathrm{U}\|_F = \|\mathrm{U}\mathrm{Q}_\delta\|_F$, we get

$$\|\hat{\mathrm{u}}_1\|^2 + \|\hat{\mathrm{u}}_2\|^2 = \|\mathrm{u}_1\|^2 + \|\mathrm{u}_2\|^2. \tag{16}$$

Let $\delta = -c\cdot\mathrm{sgn}(\mathrm{u}_1^\top\mathrm{u}_2)$ for a small enough $c > 0$ such that $\|\mathrm{u}_2\| < \|\hat{\mathrm{u}}_2\| \leq \|\hat{\mathrm{u}}_1\| < \|\mathrm{u}_1\|$. Using Equation (16), This implies that $\|\hat{\mathrm{u}}_1\|^4 + \|\hat{\mathrm{u}}_2\|^4 < \|\mathrm{u}_1\|^4 + \|\mathrm{u}_2\|^4$, which in turn gives us $R(\hat{\mathrm{U}}) < R(\mathrm{U})$ and hence $f(\hat{\mathrm{U}}) < f(\mathrm{U})$. Therefore, a non-equalized critical point cannot be local minimum, hence the first claim of the lemma.

We now prove the second part of the lemma. Let U be a critical point that is not equalized. To show that U is a strict saddle point, it suffices to show that the Hessian has a negative eigenvalue. In here, we exhibit a curve along which the second directional derivative is negative. Assume, without loss of generality that $\|\mathrm{u}_1\| > \|\mathrm{u}_2\|$ and consider the curve

$$\Delta(t) := [(\sqrt{1-t^2}-1)\mathrm{u}_1 + t\mathrm{u}_2, (\sqrt{1-t^2}-1)\mathrm{u}_2 - t\mathrm{u}_1, 0_{d,r-2}]$$

It is easy to check that for any $t \in \mathbb{R}$, $\ell(\mathrm{U}+\Delta(t)) = \ell(\mathrm{U})$ since $\mathrm{U}+\Delta(t)$ is essentially a rotation on U and ℓ is invariant under rotations. Observe that

$$\begin{aligned}&g(t) := f(\mathrm{U}+\Delta(t)) \\ &= f(\mathrm{U}) + \|\sqrt{1-t^2}\mathrm{u}_1 + t\mathrm{u}_2\|^4 - \|\mathrm{u}_1\|^4 \\ &+ \|\sqrt{1-t^2}\mathrm{u}_2 - t\mathrm{u}_1\|^4 - \|\mathrm{u}_2\|^4 \\ &= f(\mathrm{U}) - 2t^2(\|\mathrm{u}_1\|^4 + \|\mathrm{u}_2\|^4) + 8t^2(\mathrm{u}_1\mathrm{u}_2)^2 \\ &+4t^2\|\mathrm{u}_1\|^2\|\mathrm{u}_2\|^2 + 4t\sqrt{1-t^2}\mathrm{u}_1^\top\mathrm{u}_2(\|\mathrm{u}_1\|^2 - \|\mathrm{u}_2\|^2) + O(t^3).\end{aligned}$$

The derivative of g then is given as

$$\begin{aligned}g'(t) &= -4t(\|\mathrm{u}_1\|^4 + \|\mathrm{u}_2\|^4) + 16t(\mathrm{u}_1\mathrm{u}_2)^2 + 8t\|\mathrm{u}_1\|^2\|\mathrm{u}_2\|^2 \\ &+4(\sqrt{1-t^2} - \frac{t^2}{\sqrt{1-t^2}})(\mathrm{u}_1^\top\mathrm{u}_2)(\|\mathrm{u}_1\|^2 - \|\mathrm{u}_2\|^2) + O(t^2).\end{aligned}$$

Since U is a critical point and f is continuously differentiable, it should hold that $g'(0) = 4(\mathrm{u}_1^\top\mathrm{u}_2)(\|\mathrm{u}_1\|^2 - \|\mathrm{u}_2\|^2) = 0$. Since by assumption $\|\mathrm{u}_1\|^2 - \|\mathrm{u}_2\|^2 > 0$, it should be the case that $\mathrm{u}_1^\top\mathrm{u}_2 = 0$. We now consider the second order directional derivative:

$$\begin{aligned}g''(0) &= -4(\|\mathrm{u}_1\|^4 + \|\mathrm{u}_2\|^4) + 16(\mathrm{u}_1\mathrm{u}_2)^2 + 8\|\mathrm{u}_1\|^2\|\mathrm{u}_2\|^2 \\ &= -4(\|\mathrm{u}_1\|^2 - \|\mathrm{u}_2\|^2)^2 < 0\end{aligned}$$

which completes the proof. □

We now focus on the critical points that are equalized, i.e. points U such that $\nabla f(\mathrm{U}) = 0$ and $\mathrm{diag}(\mathrm{U}^\top\mathrm{U}) = \frac{\|\mathrm{U}\|_F^2}{r}\mathrm{I}$.

Lemma D.3. Assume that $\lambda < \frac{r\lambda_r}{\sum_{i=1}^r \lambda_i - r\lambda_r}$. Then all equalized local minima are global. All other equalized critical points are strict saddle points.

Proof of Lemma D.3. Let $\mathrm{U} = \mathrm{W}\Sigma\mathrm{V}^\top$ be a compact SVD of the rank-r' weight matrix U. We have:

$$\begin{aligned}&\nabla f(\mathrm{U}) = 4(\mathrm{U}\mathrm{U}^\top - \mathrm{M})\mathrm{U} + 4\lambda\mathrm{U}\,\mathrm{diag}(\mathrm{U}^\top\mathrm{U}) = 0 \\ \implies &\mathrm{U}\mathrm{U}^\top\mathrm{U} + \frac{\lambda\|\mathrm{U}\|_F^2}{r}\mathrm{U} = \mathrm{M}\mathrm{U} \\ \implies &\mathrm{W}\Sigma^3\mathrm{V}^\top + \frac{\lambda\|\Sigma\|_F^2}{r}\mathrm{W}\Sigma\mathrm{V}^\top = \mathrm{M}\mathrm{W}\Sigma\mathrm{V}^\top \\ \implies &\Sigma^2 + \frac{\lambda\|\Sigma\|_F^2}{r}\mathrm{I} = \mathrm{W}^\top\mathrm{M}\mathrm{W}\end{aligned}$$

Since the left hand side of the above equality is diagonal, it implies that $\mathrm{W} \in \mathbb{R}^{d\times r'}$ corresponds to some r' eigenvectors of M. Let $\mathcal{E} \subseteq [d]$, $|\mathcal{E}| = r'$ denote the set of eigenvectors of M that are present in W. Note that the above is equivalent of the following system of linear equations:

$$(\mathrm{I} + \frac{\lambda}{r}1\mathrm{1}^\top)\sigma^2 = \vec{\lambda},$$

where $\sigma^2 := \mathrm{diag}(\Sigma^2)$ and $\vec{\lambda} = \mathrm{diag}(\mathrm{W}^\top\mathrm{M}\mathrm{W})$. By Lemma A.2, the solution to this linear system is given by

$$\sigma^2 = (\mathrm{I} - \frac{\lambda}{r+\lambda r'})\vec{\lambda}. \tag{17}$$

The set $\mathcal{E}$ belongs to one of the following categories:

1. $\mathcal{E} = [r']$, $r' = \rho$
2. $\mathcal{E} = [r']$, $r' < \rho$
3. $\mathcal{E} \neq [r']$

The case $\mathcal{E} = [r'],\ r' > \rho$ is excluded from the above partition, since whenever $\mathcal{E} = [r']$, it should hold that $r' \leq \rho$. To see this, note that due to $\mathrm{U} = \mathrm{W}\Sigma\mathrm{V}^\top$ being a compact SVD of M, it holds that $\sigma_j > 0$ for all $j \in [r']$. Specifically for $j = r'$, plugging $\sigma_{r'} > 0$ back to Equation (17), we get

$$\lambda_{r'} > \frac{\lambda \sum_{i=1}^{r'} \lambda_i}{r + \lambda r'} = \frac{\lambda r' \kappa_{r'}}{r + \lambda r'}.$$

Then it follows from definition of ρ in Theorem 2.4 that $r' \leq \rho$. We provide a case by case analysis for the above partition here.

Case 1. **[$\mathcal{E} = [r'],\ r' = \rho$]** When W corresponds to the top-ρ eigenvectors of M, we retrieve the global optimal solution described by Theorem 2.4. Therefore, all such critical points are global minima.

Case 2. **[$\mathcal{E} = [r'],\ r' < \rho$]** Let $\mathrm{W}_r := [\mathrm{W}, \mathrm{W}_\perp]$ be the top-r eigenvectors of M and $\mathrm{V}_\perp$ span the orthogonal subspace of V, i.e. $\mathrm{V}_r := [\mathrm{V}, \mathrm{V}_\perp]$ be an orthonormal basis for $\mathbb{R}^r$. Define $\mathrm{U}(t) = \mathrm{W}_r \Sigma' \mathrm{V}_r^\top$ where $\sigma_i' = \sqrt{\sigma_i^2 + t^2}$ for $i \leq r$. Observe that

$$\mathrm{U}(t)^\top \mathrm{U}(t) = \mathrm{V}\Sigma\mathrm{V}^\top + t^2 \mathrm{V}_r^\top \mathrm{V}_r = \mathrm{U}^\top \mathrm{U} + t^2 \mathrm{I}_r$$

so that for all t, the parametric curve $\mathrm{U}(t)$ is equalized. The value of the loss function at $\mathrm{U}(t)$ is given by:

$$\begin{aligned}\ell(\mathrm{U}(t)) &= \sum_{i=1}^{r} (\lambda_i - \sigma_i^2 - t^2)^2 + \sum_{i=r+1}^{d} (\lambda_i)^2 \\ &= \ell(\mathrm{U}) + rt^4 - 2t^2 \sum_{i=1}^{r} (\lambda_i - \sigma_i^2).\end{aligned}$$

Furthermore, since $\mathrm{U}(t)$ is equalized, we obtain the following form for the regularizer:

$$\begin{aligned}R(\mathrm{U}(t)) &= \frac{\lambda}{r} \|\mathrm{U}(t)\|_F^4 = \frac{\lambda}{r} \left(\|\mathrm{U}\|_F^2 + rt^2\right)^2 \\ &= \ell(\mathrm{U}) + \lambda r t^4 + 2\lambda t^2 \|\mathrm{U}\|_F^2.\end{aligned}$$

Now define $g(t) := \ell(\mathrm{U}(t)) + R(\mathrm{U}(t))$ and observe

$$\begin{aligned}g(t) = \ell(\mathrm{U}) + R(\mathrm{U}) + rt^4 - 2t^2 \sum_{i=1}^{r} (\lambda_i - \sigma_i^2) \\ + \lambda r t^4 + 2\lambda t^2 \|\mathrm{U}\|_F^2.\end{aligned}$$

It is easy to verify that $g'(0) = 0$. Moreover, the second derivative of g at the origin is given as:

$$\begin{aligned}g''(0) &= -4 \sum_{i=1}^{r} (\lambda_i - \sigma_i^2) + 4\lambda \|\mathrm{U}\|_F^2 \\ &= -4 \sum_{i=1}^{r} \lambda_i + 4(1 + \lambda) \|\mathrm{U}\|_F^2 \\ &= -4 \sum_{i=1}^{r} \lambda_i + 4 \frac{r + r\lambda}{r + \lambda r'} \sum_{i=1}^{r'} \lambda_i\end{aligned}$$

where the last equality follows from the fact Equation (17) and the fact that $\|\mathrm{U}\|_F^2 = \sum_{i=1}^{r'} \sigma_i^2$. To get a sufficient condition for U to be a strict saddle point, we set $g''(0) < 0$:

$$\begin{aligned}&-4 \sum_{i=r'+1}^{r} \lambda_i + 4 \frac{(r - r')\lambda}{r + \lambda r'} \sum_{i=1}^{r'} \lambda_i < 0 \\ \implies &\frac{(r - r')\lambda}{r + \lambda r'} \sum_{i=1}^{r'} \lambda_i < \sum_{i=r'+1}^{r} \lambda_i \\ \implies &\lambda < \frac{(r + \lambda r') \sum_{i=r'+1}^{r} \lambda_i}{(r - r') \sum_{i=1}^{r'} \lambda_i} \\ \implies &\lambda (1 - \frac{r' \sum_{i=r'+1}^{r} \lambda_i}{(r - r') \sum_{i=1}^{r'} \lambda_i}) < \frac{r \sum_{i=r'+1}^{r} \lambda_i}{(r - r') \sum_{i=1}^{r'} \lambda_i} \\ \implies &\lambda < \frac{r \sum_{i=r'+1}^{r} \lambda_i}{(r - r') \sum_{i=1}^{r'} \lambda_i - r' \sum_{i=r'+1}^{r} \lambda_i} \\ \implies &\lambda < \frac{r h(r')}{\sum_{i=1}^{r'} (\lambda_i - h(r'))}\end{aligned}$$

where $h(r') := \frac{\sum_{i=r'+1}^{r} \lambda_i}{r - r'}$ is the average of the eigenvalues $\lambda_{r'+1}, \cdots, \lambda_r$. It is easy to see that the right hand side is monotonically decreasing with r', since $h(r')$ monotonically decrease with r'. Hence, it suffices to make sure that λ is smaller than the right hand side for the choice of $r' = r - 1$, i.e. $\lambda < \frac{r\lambda_r}{\sum_{i=1}^{r} (\lambda_i - \lambda_r)}$.

Case 3. **[$\mathcal{E} \neq [r']$]** We show that all such critical points are strict saddle points. Let w' be one of the top r' eigenvectors that are missing in W. Let $j \in \mathcal{E}$ be such that w_j is not among the top r' eigenvectors of M. For any $t \in [0, 1]$, let $\mathrm{W}(t)$ be identical to W in all the columns but the j^{th} one, where $\mathrm{w}_j(t) = \sqrt{1 - t^2}\mathrm{w}_j + t\mathrm{w}'$. Note that $\mathrm{W}(t)$ is still an orthogonal matrix for all values of t. Define the parametrized curve $\mathrm{U}(t) := \mathrm{W}(t)\Sigma\mathrm{V}^\top$ for $t \in [0, 1]$ and observe that:

$$\begin{aligned}\|\mathrm{U} - \mathrm{U}(t)\|_F^2 &= \sigma_j^2 \|\mathrm{w}_j - \mathrm{w}_j(t)\|^2 \\ &= 2\sigma_j^2 (1 - \sqrt{1 - t^2}) \leq t^2 \operatorname{Tr} \mathrm{M}\end{aligned}$$

That is, for any $\epsilon > 0$, there exist a $t > 0$ such that $\mathrm{U}(t)$ belongs to the ϵ-ball around U. We show that $f(\mathrm{U}(t))$ is strictly smaller than $f(\mathrm{U})$, which means U cannot be a local minimum. Note that this construction of $\mathrm{U}(t)$ guarantees that $R(\mathrm{U}') = R(\mathrm{U})$. In particular, it is easy to see that $\mathrm{U}(t)^\top \mathrm{U}(t) = \mathrm{U}^\top \mathrm{U}$, so that $\mathrm{U}(t)$ remains equalized for all values of t. Moreover, we have that

$$\begin{aligned}&f(\mathrm{U}(t)) - f(\mathrm{U}) = \|\mathrm{M} - \mathrm{U}(t)\mathrm{U}(t)^\top\|_F^2 - \|\mathrm{M} - \mathrm{U}\mathrm{U}^\top\|_F^2 \\ &= -2 \operatorname{Tr}(\Sigma^2 \mathrm{W}(t)^\top \mathrm{M}\mathrm{W}(t)) + 2 \operatorname{Tr}(\Sigma^2 \mathrm{W}^\top \mathrm{M}\mathrm{W}) \\ &= -2\sigma_j^2 t^2 (\mathrm{w}_j(t)^\top \mathrm{M} \mathrm{w}_j(t) - \mathrm{w}_j^\top \mathrm{M} \mathrm{w}_j) < 0,\end{aligned}$$

where the last inequality follows because by construction $\mathrm{w}_j(t)^\top \mathrm{M}\mathrm{w}_j(t) > \mathrm{w}_j^\top \mathrm{M}\mathrm{w}_j$. Define $g(t) := f(\mathrm{U}(t)) = \ell(\mathrm{U}(t)) + R(\mathrm{U}(t))$. To see that such saddle points are non-degenerate, it suffices to show $g''(0) < 0$. It is easy to check that the second directional derivative at the origin is given by

$$g''(0) = -4\sigma_j^2(\mathrm{w}_j(t)^\top \mathrm{M}\mathrm{w}_j(t) - \mathrm{w}_j^\top \mathrm{M}\mathrm{w}_j) < 0,$$

which completes the proof. □

Proof of Lemma 4.1. Follows from Lemma D.2 □

Proof of Theorem 4.3. Follows from Lemma D.2 and Lemma D.3. □