A. Preliminaries about subspace distance

Before delving into the proofs, we will prove a few simple preliminaries about subspace angles/distances.

Definition (Distance, Principle angle)**.** Denote the principle angle of $\mathbf{Y}$, $\mathbf{V} \in \mathbb{R}^{n\times k}$ as $\theta(\mathbf{Y},\mathbf{V})$. Then for orthogonal matrix $\mathbf{Y}$ (i.e., $\mathbf{Y}^\top\mathbf{Y} = \mathbf{I}$),

$$\tan\theta(\mathbf{Y},\mathbf{V}) = \|\mathbf{Y}_\perp^\top\mathbf{V}(\mathbf{Y}^\top\mathbf{V})^{-1}\|_2.$$

For orthogonal matrices $\mathbf{Y}$, $\mathbf{V}$,

$$\begin{aligned}
&\cos\theta(\mathbf{Y},\mathbf{V}) = \sigma_{\min}(\mathbf{Y}^\top\mathbf{V}),\\
&\sin\theta(\mathbf{Y},\mathbf{V}) = \|(\mathbf{I}-\mathbf{Y}\mathbf{Y}^\top)\mathbf{V}\|_2 = \|\mathbf{Y}_\perp\mathbf{Y}_\perp^\top\mathbf{V}\|_2 = \|\mathbf{Y}_\perp^\top\mathbf{V}\|_2,\\
&\mathsf{dist}_c(\mathbf{Y},\mathbf{V}) = \min_{\mathbf{Q}\in\mathsf{O}_{k\times k}}\|\mathbf{Y}\mathbf{Q}-\mathbf{V}\|_2
\end{aligned}$$

where $\mathsf{O}_{k\times k}$ is the set of $k\times k$ orthogonal matrices.

Lemma 4 (Equivalence of distance)**.** *Let $\mathbf{Y}$, $\mathbf{V} \in \mathbb{R}^{n\times k}$ be two orthogonal matrices, then we have:*

$$\sin\theta(\mathbf{Y},\mathbf{V}) \le \mathit{dist}_c(\mathbf{Y},\mathbf{V}) \le \sin\theta(\mathbf{Y},\mathbf{V}) + \frac{1-\cos\theta(\mathbf{Y},\mathbf{V})}{\cos\theta(\mathbf{Y},\mathbf{V})} \le 2\tan\theta(\mathbf{Y},\mathbf{V}).$$

Proof of Lemma 4. Suppose

$$\mathbf{Q}^* = \mathrm{argmin}_{\mathbf{Q}\in\mathsf{O}_{k\times k}}\|\mathbf{Y}\mathbf{Q}-\mathbf{V}\|_2.$$

Let's write $\mathbf{V} = \mathbf{Y}\mathbf{Q}^* + \mathbf{R}$, then $\mathsf{dist}_c(\mathbf{Y},\mathbf{V}) = \|\mathbf{R}\|_2$. We have

$$\sin\theta(\mathbf{Y},\mathbf{V}) = \|(\mathbf{I}-\mathbf{Y}\mathbf{Y}^\top)\mathbf{V}\|_2 = \|\mathbf{Y}_\perp\mathbf{Y}_\perp^\top\mathbf{R}\|_2 \le \|\mathbf{R}\|_2$$

On the other hand, suppose $\mathbf{A}\mathbf{D}\mathbf{B}^\top = \mathrm{SVD}(\mathbf{Y}^\top\mathbf{V})$, we know that $\sigma_{\min}(\mathbf{D}) = \sigma_{\min}(\mathbf{Y}^\top\mathbf{V}) = \cos\theta(\mathbf{Y},\mathbf{V})$. Therefore, by $\mathbf{A} = \mathbf{Y}^\top\mathbf{V}\mathbf{B}\mathbf{D}^{-1}$, $\mathbf{A}\mathbf{B}^\top \in \mathsf{O}_{k\times k}$ we have:

$$\begin{aligned}
\mathsf{dist}_c(\mathbf{Y},\mathbf{V}) &\le \|\mathbf{Y}\mathbf{A}\mathbf{B}^\top - \mathbf{V}\|_2 = \|\mathbf{Y}\mathbf{Y}^\top\mathbf{V}\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top - \mathbf{V}\|_2\\
&\le \|\mathbf{Y}\mathbf{Y}^\top\mathbf{V}\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top - \mathbf{Y}\mathbf{Y}^\top\mathbf{V}\|_2 + \|\mathbf{Y}\mathbf{Y}^\top\mathbf{V} - \mathbf{V}\|_2\\
&\le \|\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top - \mathbf{I}\|_2 + \sin\theta(\mathbf{Y},\mathbf{V}) = \|\mathbf{D}^{-1}-\mathbf{I}\|_2 + \sin\theta(\mathbf{Y},\mathbf{V})\\
&= \sin\theta(\mathbf{Y},\mathbf{V}) + \frac{1-\cos\theta(\mathbf{Y},\mathbf{V})}{\cos\theta(\mathbf{Y},\mathbf{V})}.
\end{aligned}$$

Finally, $\sin\theta(\mathbf{Y},\mathbf{V}) \le \tan\theta(\mathbf{Y},\mathbf{V})$ and $\frac{1-\cos\theta(\mathbf{Y},\mathbf{V})}{\cos\theta(\mathbf{Y},\mathbf{V})} \le \tan\theta(\mathbf{Y},\mathbf{V})$ can be verified by definition, so the last inequality follows. □

For convenience in our proofs we will also use the following generalization of incoherence:

Definition (Generalized incoherence)**.** For a matrix $\mathbf{A} \in \mathbb{R}^{n\times k}$, the *generalized incoherence* $\rho(\mathbf{A})$ is defined as:

$$\rho(\mathbf{A}) = \max_{i\in[n]}\left\{\frac{n}{k}\|\mathbf{A}^i\|_2^2\right\}$$

We call it generalized incoherence for obvious reasons: when $\mathbf{A}$ is an orthogonal matrix, then $\rho(\mathbf{A}) = \mu(\mathbf{A})$.

B. Proofs for alternating minimization with clipping

We will show in this section the results for our algorithm based on alternating minimization with a clipping step. The organization is as follows. In Section B.1 we will present the necessary lemmas for the initialization, in Section B.3 we show the decrease of the potential function after one update step, and in Section B.4 we will put everything together, and prove our main theorem.

Before starting with the proofs, we will make a remark which will simplify the exposition.

Without loss of generality, we may assume that

$$\delta = \|\mathbf{W} \odot \mathbf{N}\|_2 \leq \frac{\lambda \sigma_{\min}(\mathbf{M}^*)}{200k} \tag{B.1}$$

Otherwise, we can output the 0 matrix, and the guarantee of all our theorems would be satisfied vacuously.

B.1. SVD-based initialization

We want to show that after initialization, the matrices $\mathbf{X}, \mathbf{Y}$ are close to the ground truth matrix $\mathbf{U}, \mathbf{V}$. Observe that $[\mathbf{X}, \mathbf{\Sigma}, \mathbf{Y}] = \mathrm{SVD}(\mathbf{W} \odot \mathbf{M}) = \mathrm{SVD}(\mathbf{W} \odot (\mathbf{M}^* + \mathbf{N})) = \mathrm{SVD}(\mathbf{W} \odot \mathbf{M}^* + \mathbf{W} \odot \mathbf{N})$. By our assumptions we know that $||\mathbf{W} \odot \mathbf{N}||_2 \leq \delta$ which we are thinking of as small, so the idea is to show that $\mathbf{W} \odot \mathbf{M}^*$ is close to $\mathbf{M}^*$ in spectral norm, then by Wedin's theorem (Wedin, 1972) we will have $\mathbf{X}, \mathbf{Y}$ are close to $\mathbf{U}, \mathbf{V}$. We show that $\mathbf{W} \odot \mathbf{M}^*$ is close to $\mathbf{M}^*$ by the spectral gap property of $\mathbf{W}$ and the incoherence property of $\mathbf{U}, \mathbf{V}$.

Lemma 5 (Spectral lemma)**.** *Let $\mathbf{W}$ be an (entry wise non-negative) matrix in $\mathbb{R}^{n \times n}$ with a spectral gap, i.e. $\mathbf{W} = \mathbf{E} + \gamma n \mathbf{J} \mathbf{\Sigma}_{\mathbf{W}} \mathbf{K}^\top$, where $\mathbf{J}, \mathbf{K}$ are $n \times n$ (column) orthogonal matrices, with $||\mathbf{\Sigma}_{\mathbf{W}}||_2 = 1, \gamma < 1$. Furthermore, for every matrix $\mathbf{H} \in \mathbb{R}^{n \times n}$ such that $\mathbf{H} = \mathbf{A} \mathbf{\Sigma} \mathbf{B}^\top$ ($\mathbf{A}, \mathbf{B}$ not necessarily orthogonal, $\mathbf{\Sigma} \in \mathbb{R}^{k \times k}$ is diagonal) we have*

$$\|(\mathbf{W} - \mathbf{E}) \odot \mathbf{H}\|_2 \leq \gamma k \sigma_{\max}(\mathbf{\Sigma}) \sqrt{\rho(\mathbf{A}) \rho(\mathbf{B})}$$

where $\mathbf{E}$ is the all one matrix.

Proof of Lemma 5. We know that for any unit vectors $x, y \in \mathbb{R}^n$,

$$\begin{aligned}
x^\top \left((\mathbf{W} - \mathbf{E}) \odot \mathbf{H} \right) y &= \sum_{r=1}^{k} \sigma_r x^T \left((\mathbf{W} - \mathbf{E}) \odot \mathbf{A}_r \mathbf{B}_r^\top \right) y \\
&= \gamma n \sum_{r=1}^{k} \sigma_r (\mathbf{A}_r \odot x)^\top \mathbf{J} \mathbf{\Sigma}_{\mathbf{W}} \mathbf{K}^\top (\mathbf{B}_r \odot y) \\
&\leq \gamma n \sum_{r=1}^{k} \sigma_r ||\mathbf{A}_r \odot x||_2 ||\mathbf{J} \mathbf{\Sigma}_{\mathbf{W}} \mathbf{K}^\top||_2 ||\mathbf{B}_r \odot y||_2 \\
&\leq \gamma n \sum_{r=1}^{k} \sigma_r ||\mathbf{A}_r \odot x||_2 ||\mathbf{B}_r \odot y||_2 \\
&\leq \gamma n \sigma_{\max}(\mathbf{\Sigma}) \sqrt{\sum_{r=1}^{k} ||\mathbf{A}_r \odot x||_2^2} \sqrt{\sum_{r=1}^{k} ||\mathbf{B}_r \odot y||_2^2} \\
&\leq \gamma n \sigma_{\max}(\mathbf{\Sigma}) \sqrt{\sum_{i=1}^{n} x_i^2 ||\mathbf{A}^i||_2^2} \sqrt{\sum_{i=1}^{n} y_i^2 ||\mathbf{B}^i||_2^2} \\
&\leq \gamma n \sigma_{\max}(\mathbf{\Sigma}) \sqrt{\frac{k}{n} \rho(\mathbf{A}) \left(\sum_{i=1}^{n} x_i^2 \right)} \sqrt{\frac{k}{n} \rho(\mathbf{B}) \left(\sum_{i=1}^{n} y_i^2 \right)} \\
&\leq \gamma \sigma_{\max}(\mathbf{\Sigma}) k \sqrt{\rho(\mathbf{A}) \rho(\mathbf{B})}.
\end{aligned}$$

The lemma follows from the definition of the operator norm. □

The spectral lemma can be used to prove the initialization condition, when combined with Wedin's theorem.

Lemma 6 (Wedin's Theorem (Wedin, 1972))**.** *Let $\mathbf{M}^*, \widetilde{\mathbf{M}}$ be two matrices whose singular values are $\sigma_1, ..., \sigma_n$ and $\tilde{\sigma}_1, ..., \tilde{\sigma}_n$, let $\mathbf{U}, \mathbf{V}$ and $\mathbf{X}, \mathbf{Y}$ be the first k singular vectors (left and right) of $\mathbf{M}^*, \widetilde{\mathbf{M}}$ respectively. If $\exists \alpha > 0$ such that*

$\max_{r=k+1}^{n} \tilde{\sigma}_r \leq \min_{i=1}^{k} \sigma_i - \alpha$, *then*

$$\max\{\sin\theta(\mathbf{U},\mathbf{X}), \sin\theta(\mathbf{V},\mathbf{Y})\} \leq \frac{||\widetilde{\mathbf{M}} - \mathbf{M}^*||_2}{\alpha}.$$

Lemma 7. *Suppose* $\mathbf{M}^*, \mathbf{W}$ *satisfy all the assumptions, then for* $(\mathbf{X}, \mathbf{\Sigma}, \mathbf{Y}) =$ *rank-*k $\mathrm{SVD}(\mathbf{W} \odot \mathbf{M})$*, we have*

$$\max\{\tan\theta(\mathbf{X},\mathbf{U}), \tan\theta(\mathbf{Y},\mathbf{V})\} \leq \frac{4(\gamma\mu k + \delta)}{\sigma_{\min}(\mathbf{M}^*)}$$

Proof of Lemma 7. We know that

$$\|\mathbf{W} \odot \mathbf{M} - \mathbf{M}^*\|_2 \leq ||\mathbf{W} \odot \mathbf{M}^* - \mathbf{M}^*||_2 + ||\mathbf{W} \odot \mathbf{N}||_2 \leq \gamma\mu k \sigma_{\max}(\mathbf{M}^*) + \delta.$$

Therefore, by Weyl's theorem,

$$\max\{\sigma_r(\mathbf{W} \odot \mathbf{M}) : k+1 \leq r \leq n\} \leq \gamma\mu k + \delta \leq \frac{1}{2}\sigma_{\min}(\mathbf{M}^*).$$

where the last inequality holds because of B.1 and the assumption on γ in the theorem statement.

Now, by Wedin's theorem with $\alpha = \frac{1}{2}\sigma_{\min}(\mathbf{M}^*)$, for $(\mathbf{X}, \mathbf{\Sigma}, \mathbf{Y}) =$ rank-k $\mathrm{SVD}(\mathbf{W} \odot \mathbf{M})$,

$$\max\{\sin\theta(\mathbf{U},\mathbf{X}), \sin\theta(\mathbf{V},\mathbf{Y})\} \leq \frac{2(\gamma\mu k + \delta)}{\sigma_{\min}(\mathbf{M}^*)}$$

Since γ and δ are small enough, so $\sin\theta \leq 1/2$. In this case, we have $\tan\theta \leq 2\sin\theta$, then the lemma follows. □

Finally, this gives us the following guarantee on the initialization:

Lemma 8 (SVD initialization)**.** *Suppose* $\mathbf{M}^*, \mathbf{W}$ *satisfy all the assumptions.*

$$\mathit{dist}_c(\mathbf{V}, \mathbf{Y}_1) \leq 8k\Delta_1, \;\; \rho(\mathbf{Y}_1) \leq \frac{2\mu}{1 - k\Delta_1}$$

where $\Delta_1 = \frac{8(\gamma\mu k + \delta)}{\sigma_{\min}(\mathbf{M}^*)}$.

Proof of Lemma 8. First, consider $\widetilde{\mathbf{Y}}_1$. By Lemma 7 and 4, we get that

$$\mathsf{dist}_c(\widetilde{\mathbf{Y}}_1, \mathbf{V}) \leq \Delta_1$$

which means that $\exists \mathbf{Q} \in \mathsf{O}_{k\times k}$, s.t.

$$\|\widetilde{\mathbf{Y}}_1\mathbf{Q} - \mathbf{V}\|_2 \leq \Delta_1$$

hence

$$\|\widetilde{\mathbf{Y}}_1\mathbf{Q} - \mathbf{V}\|_F \leq k\Delta_1 \leq \frac{1}{4}$$

where the last inequality follows since γ and δ are small enough.

Next, consider $\overline{\mathbf{Y}}_1$. In the clipping step, if $\|\widetilde{\mathbf{Y}}_1^i\| \geq \xi = \frac{2\mu k}{n}$, then $\|\widetilde{\mathbf{Y}}_1^i - \mathbf{V}^i\| \geq \frac{\mu k}{n}$, and $\|\overline{\mathbf{Y}}_1^i - \mathbf{V}^i\| = \|\mathbf{V}^i\| = \frac{\mu k}{n}$. Otherwise, $\overline{\mathbf{Y}}^i = \widetilde{\mathbf{Y}}^i$. So

$$\|\overline{\mathbf{Y}}_1\mathbf{Q} - \mathbf{V}\|_F \leq \|\widetilde{\mathbf{Y}}_1\mathbf{Q} - \mathbf{V}\|_F \leq \frac{1}{4}.$$

Finally, we can argue that $\mathbf{Y}_1$ is close to $\mathbf{V}$. Let's assume that $\mathbf{Y}_1 = \overline{\mathbf{Y}}_1\mathbf{R}^{-1}$, for an upper-triangular $\mathbf{R}$.

$$\sin\theta(\mathbf{V},\mathbf{Y}_1) = \|\mathbf{V}_\perp^\top\mathbf{Y}_1\|_2 = \|\mathbf{V}_\perp^\top(\overline{\mathbf{Y}}_1 - \mathbf{V}\mathbf{Q}^{-1})\mathbf{R}^{-1}\|_2 \leq \|\overline{\mathbf{Y}}_1\mathbf{Q} - \mathbf{V}\|_2\|\mathbf{R}^{-1}\|_2 \leq \frac{1}{\sigma_{\min}(\overline{\mathbf{Y}}_1)}\|\overline{\mathbf{Y}}_1\mathbf{Q} - \mathbf{V}\|_F$$

where the second inequality follows because the singular values of $\mathbf{R}$ and $\overline{\mathbf{Y}}_1$ are the same. Note that

$$\sigma_{\min}(\overline{\mathbf{Y}}_1) \geq \sigma_{\min}(\mathbf{V}) - \|\overline{\mathbf{Y}}_1 - \mathbf{V}\|_F \geq \sigma_{\min}(\mathbf{V}) - k\Delta_1 = 1 - k\Delta_1 \geq \frac{1}{2}$$

So

$$\sin\theta(\mathbf{V}, \mathbf{Y}_1) \leq 2\|\overline{\mathbf{Y}}_1\mathbf{Q} - \mathbf{V}\|_F \leq \frac{1}{2}.$$

In this case, we have $\tan\theta(\mathbf{V}, \mathbf{Y}_1) \leq 2\sin\theta(\mathbf{V}, \mathbf{Y}_1)$ and thus

$$\mathsf{dist}_c(\mathbf{V}, \mathbf{Y}_1) \leq 2\tan\theta(\mathbf{V}, \mathbf{Y}_1) \leq 4\sin\theta(\mathbf{V}, \mathbf{Y}_1) \leq 8\|\overline{\mathbf{Y}}_1\mathbf{Q} - \mathbf{V}\|_2 \leq 8\|\overline{\mathbf{Y}}_1\mathbf{Q} - \mathbf{V}\|_F \leq 8k\Delta_1.$$

For $\rho(\mathbf{Y}_1)$, observe that $\mathbf{Y}_1^i = \overline{\mathbf{Y}}^i\mathbf{R}^{-1}$, so

$$\|\mathbf{Y}_1^i\| \leq \|\overline{\mathbf{Y}}_1^i\|\|\mathbf{R}^{-1}\|_2 \leq \frac{\xi}{\sigma_{\min}(\overline{\mathbf{Y}}_1)} \leq \frac{\xi}{1 - k\Delta_1}$$

which leads to the bound. □

B.2. Random initialization

With respect to the random initialization, the lemma we will need is the following one:

Lemma 9 (Random initialization)**.** *Let $\mathbf{Y}$ be a random matrix in $\mathbb{R}^{n\times k}$ generated as $\mathbf{Y}_{i,j} = b_{i,j}\frac{1}{\sqrt{n}}$, where $b_{i,j}$ are independent, uniform $\{-1, 1\}$ variables. Furthermore, let $\|\mathbf{W}\|_\infty \leq \frac{\lambda n}{k^2\mu\log^2 n}$. Then, with probability at least $1 - \frac{1}{n^2}$ over the draw of $\mathbf{Y}$,*

$$\forall i, \sigma_{\min}\left(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}\right) \geq \frac{1}{4}\frac{\lambda}{k\mu}.$$

Proof of Lemma 9. Notice that $\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y} = \sum_j(\mathbf{Y}^j)^\top(\mathbf{D}_i)_j\mathbf{Y}^j$, and each of the terms $(\mathbf{Y}^j)^\top(\mathbf{D}_i)_j\mathbf{Y}^j$ is independent. Furthermore, it's easy to see that $\mathbb{E}[(\mathbf{Y}^j)^\top(\mathbf{D}_i)_j(\mathbf{Y}^j)] = \frac{1}{n}(\mathbf{D}_i)_j$, $\forall j$. By linearity of expectation it follows that $\mathbb{E}[\sum_j(\mathbf{Y}^j)^\top(\mathbf{D}_i)_j\mathbf{Y}^j] = \frac{1}{n}\sum_j(\mathbf{D}_i)_j$.

Now, we claim $\sum_j(\mathbf{D}_i)_j \geq \frac{\overline{\lambda}n}{k\mu}$. Indeed, by Assumption **(A3)** we have for any vector $a \in \mathbb{R}^n$

$$a^\top\mathbf{V}^\top\mathbf{D}_i\mathbf{V}a = \sum_j(\mathbf{D}_i)_j\langle\mathbf{V}^j, a\rangle^2 \geq \underline{\lambda}.$$

On the other hand, however, by incoherence of $\mathbf{V}$, $\sum_j(\mathbf{D}_i)_j\langle\mathbf{V}^j, a\rangle^2 \leq \sum_j(\mathbf{D}_i)_j\frac{\mu k}{n}$. Hence, $\sum_j(\mathbf{D}_i)_j \geq \overline{\lambda}\frac{n}{k\mu}$. Putting things together, we get

$$\mathbb{E}[\sum_j(\mathbf{Y}^j)^\top(\mathbf{D}_i)_j\mathbf{Y}^j] \geq \frac{\lambda}{k\mu}$$

Denote

$$B := \|(\mathbf{Y}^j)^\top(\mathbf{D}_i)_j\mathbf{Y}^j\|_2 \leq \frac{k}{n}(\mathbf{D}_i)_j \leq \frac{\underline{\lambda}}{k\mu\log^2 n}$$

where the first inequality follows from our sampling procedure, and the last inequality by the assumption that $\|\mathbf{W}\|_\infty \leq \frac{\underline{\lambda}n}{k^2\mu\log^2 n}$.

Since all the random variables $(\mathbf{Y}^j)^\top(\mathbf{D}_i)_j\mathbf{Y}^j$ are independent, applying Matrix Chernoff we get that

$$\Pr\left[\sum_j(\mathbf{Y}^j)^\top(\mathbf{D}_i)_j(\mathbf{Y}^j) \leq (1-\delta)\frac{\underline{\lambda}}{k\mu}\right] \leq n\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\frac{\underline{\lambda}}{k\mu B}} \leq n\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\log^2 n}$$

Picking $\delta = \frac{3}{4}$, and union bounding over all i, with probability at least $1 - \frac{1}{n^2}$, for all i,

$$\sigma_{\min}\left(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}\right) \geq \frac{1}{4}\frac{\lambda}{k\mu}$$

as needed. □

B.3. Update

We now prove the two key technical lemmas (Lemma 10 and Lemma 11) and then use them to prove that the updates make progress towards the ground truth. We prove them for $\mathbf{Y}_t$ and use them to show $\mathbf{X}_t$ improves, while completely analogous arguments also hold when switching the role of the two iterates. Note that we measure the distance between $\mathbf{Y}_t$ and $\mathbf{V}$ by $\mathsf{dist}_c(\mathbf{Y}_t, \mathbf{V}) = \min_{\mathbf{Q}\in\mathsf{O}_{k\times k}} \|\mathbf{Y}_t\mathbf{Q} - \mathbf{V}\|$ where $\mathsf{O}_{k\times k}$ is the set of $k \times k$ orthogonal matrices. For simplicity of notations, in these two lemmas, we let $\mathbf{Y}_o = \mathbf{Y}_t\mathbf{Q}^*$ where $\mathbf{Q}^* = \mathrm{argmin}_{\mathbf{Q}\in\mathsf{O}_{k\times k}}\|\mathbf{Y}_t\mathbf{Q} - \mathbf{V}\|$.

We first show that there can only be a few i's such that the spectral property of $\mathbf{Y}_o^\top\mathbf{D}_i\mathbf{Y}_o$ can be bad, when $\mathbf{Y}_o$ is close to $\mathbf{V}$. Let $(\mathbf{D}_i)_j$ be the j-th diagonal entry in $\mathbf{D}_i$, that is, $(\mathbf{D}_i)_j = \mathbf{W}_{i,j}$.

Lemma 10. *Let $\mathbf{Y}_o$ be a (column) orthogonal matrix in $\mathbb{R}^{n\times k}$, and $\epsilon \in (0,1)$. If $\|\mathbf{Y}_o - \mathbf{V}\|_F^2 \le \frac{\epsilon^3\underline{\lambda}^2 n}{128\mu k D_1}$ for $D_1 = \max_{i\in[n]}\sum_j(\mathbf{D}_i)_j$, then*

$$\left|\left\{i \in [n] \left| \sigma_{\min}(\mathbf{Y}_o^\top\mathbf{D}_i\mathbf{Y}_o) \le (1-\epsilon)\underline{\lambda}\right.\right\}\right| \le \frac{1024\mu^2k^2\gamma^2 D_1}{\epsilon^4\underline{\lambda}^3}\|\mathbf{V} - \mathbf{Y}_o\|_F^2.$$

Proof of Lemma 10. For a value $g > 0$ which we will specify shortly, we call $j \in [n]$ "good" if $\|\mathbf{Y}_o^j - \mathbf{V}^j\|^2 \le g^2$. Denote the set of "good" j's as $\mathcal{S}_g$.

Then for every unit vector $a \in \mathbb{R}^k$,

$$\begin{aligned}
a^\top\mathbf{Y}_o^\top\mathbf{D}_i\mathbf{Y}_o a &= \sum_{j\in[n]}(\mathbf{D}_i)_j\langle a, \mathbf{Y}_o^j\rangle^2 \\
&\ge \sum_{j\in\mathcal{S}_g}(\mathbf{D}_i)_j\langle a, \mathbf{Y}_o^j\rangle^2 \\
&= \sum_{j\in\mathcal{S}_g}(\mathbf{D}_i)_j\left(\langle a, \mathbf{V}^j\rangle + \langle a, \mathbf{Y}_o^j - \mathbf{V}^j\rangle\right)^2 \\
&\ge (1-\frac{\epsilon}{4})\sum_{j\in\mathcal{S}_g}(\mathbf{D}_i)_j\langle a, \mathbf{V}^j\rangle^2 - \frac{4-\epsilon}{\epsilon}\sum_{j\in S_g}(\mathbf{D}_i)_j\langle a, \mathbf{Y}_o^j - \mathbf{V}^j\rangle^2 \\
&\qquad \text{(Using the fact } \forall x, y \in \mathbb{R} : (x+y)^2 \ge (1-\epsilon_0)x^2 - \frac{1-\epsilon_0}{\epsilon_0}y^2) \\
&\ge (1-\frac{\epsilon}{4})\sum_{j\in\mathcal{S}_g}(\mathbf{D}_i)_j\langle a, \mathbf{V}^j\rangle^2 - \frac{4-\epsilon}{\epsilon}g^2\sum_{j\in[n]}(\mathbf{D}_i)_j \\
&\ge (1-\frac{\epsilon}{4})\sum_{j\in[n]}(\mathbf{D}_i)_j\langle a, \mathbf{V}^j\rangle^2 - \frac{\mu k}{n}\sum_{j\in[n]-\mathcal{S}_g}(\mathbf{D}_i)_j - \frac{4-\epsilon}{\epsilon}g^2\sum_{j\in[n]}(\mathbf{D}_i)_j
\end{aligned}$$

By Assumption **(A3)**, we know that

$$\sum_{j\in[n]}(\mathbf{D}_i)_j\langle a, \mathbf{V}^j\rangle^2 = a^T\mathbf{V}^\top\mathbf{D}_i\mathbf{V}a \ge \sigma_{\min}(\mathbf{V}^\top\mathbf{D}_i\mathbf{V}) \ge \underline{\lambda}$$

Moreover, recall $D_1 = \max_{i\in[n]}\sum_j(\mathbf{D}_i)_j$, so when $g^2 \le \frac{\epsilon^2\underline{\lambda}}{16D_1}$,

$$\frac{4-\epsilon}{\epsilon}g^2\sum_{j\in[n]}(\mathbf{D}_i)_j \le \frac{\epsilon\underline{\lambda}}{4}$$

Let us consider now $\sum_{j\in[n]-S_g}(\mathbf{D}_i)_j$. Define:

$$\mathcal{S} = \left\{ i \in [n] \left| \frac{\mu k}{n}\sum_{j\in[n]-\mathcal{S}_g}(\mathbf{D}_i)_j \ge \frac{\epsilon\underline{\lambda}}{4} \right.\right\}$$

Then it is sufficient to bound $|\mathcal{S}|$.

For $\mathcal{S}_g$, observe that

$$\sum_j \|\mathbf{V}^j - \mathbf{Y}_o^j\|_2^2 = \|\mathbf{V} - \mathbf{Y}_o\|_F^2$$

Which implies that

$$|[n] - \mathcal{S}_g| = \text{size}\left([n] - \mathcal{S}_g\right) \leq \frac{\|\mathbf{V} - \mathbf{Y}_o\|_F^2}{g^2}$$

Let u_S be the indicator vector of $\mathcal{S}$, and u_g be the indicator vector of $[n] - \mathcal{S}_g$, we know that

$$\begin{aligned} u_S^\top \mathbf{W} u_g &= \sum_{i \in \mathcal{S}} \sum_{j \in [n] - \mathcal{S}_g} (\mathbf{D}_i)_j \\ &\geq \frac{\epsilon \underline{\lambda} n}{4\mu k} |\mathcal{S}| \end{aligned}$$

On the other hand,

$$\begin{aligned} u_S^\top \mathbf{W} u_g &= u_S^\top \mathbf{E} u_g + u_S^\top (\mathbf{W} - \mathbf{E}) u_g \\ &\leq |\mathcal{S}||[n] - \mathcal{S}_g| + \gamma n \sqrt{|\mathcal{S}||[n] - \mathcal{S}_g|} \end{aligned}$$

Putting these two inequalities together, we have

$$|[n] - \mathcal{S}_g| + \gamma n \sqrt{\frac{|[n] - \mathcal{S}_g|}{|\mathcal{S}|}} \geq \frac{\epsilon \underline{\lambda} n}{4\mu k}$$

Which implies when $|[n] - \mathcal{S}_g| \leq \frac{\epsilon \underline{\lambda} n}{8\mu k}$, we have:

$$|\mathcal{S}| \leq \frac{64\mu^2 k^2 \gamma^2 |[n] - \mathcal{S}_g|}{\epsilon^2 \underline{\lambda}^2} \leq \frac{64\mu^2 k^2 \gamma^2 \|\mathbf{V} - \mathbf{Y}_o\|_F^2}{\epsilon^2 \underline{\lambda}^2 g^2}$$

Then, setting $g^2 = \frac{\epsilon^2 \underline{\lambda}}{16 D_1}$, we have:

$$\left|\left\{ i \in [n] \,\middle|\, \sigma_{\min}(\mathbf{Y}_o^\top \mathbf{D}_i \mathbf{Y}_o) \leq (1 - \epsilon)\underline{\lambda} \right\}\right| \leq |\mathcal{S}| \leq \frac{1024 \mu^2 k^2 \gamma^2 D_1}{\epsilon^4 \underline{\lambda}^3} \|\mathbf{V} - \mathbf{Y}_o\|_F^2$$

which is what we need.

□

Lemma 11. *Let $\mathbf{Y}_o$ be a (column) orthogonal matrix in $\mathbb{R}^{n \times k}$. Then we have*

$$\sum_{i \in [n]} \|\mathbf{V}^\top \mathbf{Y}_\perp \mathbf{Y}_\perp^\top \mathbf{D}_i \mathbf{Y}_o\|_2^2 \leq \gamma^2 \rho(\mathbf{Y}_o) n k^3 \|\mathbf{Y}_o - \mathbf{V}\|_2^2$$

Proof of Lemma 11. We want to bound the spectral norm of $\mathbf{V}^\top \mathbf{Y}_\perp \mathbf{Y}_\perp^\top \mathbf{D}_i \mathbf{Y}_o$, for a fixed $j \in [k]$, let $\mathbf{Y}_j$ be the j-th column of $\mathbf{Y}_o$ and $\widetilde{\mathbf{V}}_j$ be the j-th column of $\mathbf{Y}_\perp \mathbf{Y}_\perp^\top \mathbf{V}$.

For fixed $j, j' \in [k]$, consider a new vector $x^{j,j'} \in \mathbb{R}^n$ such that $x_i^{j,j'} = (\widetilde{\mathbf{V}}_j)_i (\mathbf{Y}_{j'})_i$.

Note that $\langle \widetilde{\mathbf{V}}_j, \mathbf{Y}_{j'} \rangle = 0$, which implies that $\sum_i x_i^{j,j'} = 0$.

Let us consider $\mathbf{V}_j^\top \mathbf{Y}_\perp \mathbf{Y}_\perp^\top \mathbf{D}_i \mathbf{Y}_{j'}$, we know that

$$\begin{aligned}\mathbf{V}_j^\top \mathbf{Y}_\perp \mathbf{Y}_\perp^\top \mathbf{D}_i \mathbf{Y}_{j'} &= \sum_{s\in[n]} (\mathbf{D}_i)_s (\widetilde{\mathbf{V}}_j)_s (\mathbf{Y}_{j'})_s \\ &= \sum_{s\in[n]} (\mathbf{D}_i)_s x_s^{j,j'}\end{aligned}$$

Which implies that

$$\begin{aligned}\sum_{i\in[n]} \left(\sum_{s\in[n]} (\mathbf{D}_i)_s x_s^{j,j'} \right)^2 &= \|\mathbf{W} x^{j,j'}\|_2^2 \\ &= \|(\mathbf{W}-\mathbf{E}) x^{j,j'}\|_2^2 \qquad (\text{since } \mathbf{E} x^{j,j'} = 0) \\ &\leq \gamma^2 n^2 \|x^{j,j'}\|_2^2\end{aligned}$$

Observe that

$$\begin{aligned}\|x^{j,j'}\|_2^2 &= \sum_{i\in[n]} (x_i^{j,j'})^2 \\ &= \sum_{i\in[n]} (\widetilde{\mathbf{V}}_j)_i^2 (\mathbf{Y}_{j'})_i^2 \\ &\leq \frac{\rho(\mathbf{Y}_o)k}{n} \sum_{i\in[n]} (\widetilde{\mathbf{V}}_j)_i^2 \\ &= \frac{\rho(\mathbf{Y}_o)k}{n} \|\widetilde{\mathbf{V}}_j\|_2^2 \\ &\leq \frac{\rho(\mathbf{Y}_o)k}{n} \|\mathbf{Y}_\perp \mathbf{Y}_\perp^\top \mathbf{V}\|_2^2 \\ &= \frac{\rho(\mathbf{Y}_o)k}{n} \|\mathbf{Y}_\perp \mathbf{Y}_\perp^\top (\mathbf{Y}_o - \mathbf{V})\|_2^2 \\ &\leq \frac{\rho(\mathbf{Y}_o)k}{n} \|\mathbf{Y}_o - \mathbf{V}\|_2^2.\end{aligned}$$

Which implies

$$\sum_{i\in[n]} \left(\sum_{s\in[n]} (\mathbf{D}_i)_s x_s^{j,j'} \right)^2 \leq \gamma^2 \rho(\mathbf{Y}_o) nk \|\mathbf{Y}_o - \mathbf{V}\|_2^2$$

Now we are ready to bound $\mathbf{V}^\top \mathbf{Y}_\perp \mathbf{Y}_\perp^\top \mathbf{D}_i \mathbf{Y}_o$. Note that

$$\begin{aligned}\|\mathbf{V}^\top \mathbf{Y}_\perp \mathbf{Y}_\perp^\top \mathbf{D}_i \mathbf{Y}_o\|_2^2 &\leq \|\mathbf{V}^\top \mathbf{Y}_\perp \mathbf{Y}_\perp^\top \mathbf{D}_i \mathbf{Y}_o\|_F^2 \\ &\leq \sum_{j,j'\in[k]} \left(\mathbf{V}_j^\top \mathbf{Y}_\perp \mathbf{Y}_\perp^\top \mathbf{D}_i \mathbf{Y}_{j'}\right)^2 \\ &= \sum_{j,j'\in[k]} \left(\sum_{s\in[n]} (\mathbf{D}_i)_s x_s^{j,j'} \right)^2.\end{aligned}$$

This implies that

$$\sum_{i\in[n]} \|\mathbf{V}^\top \mathbf{Y}_\perp \mathbf{Y}_\perp^\top \mathbf{D}_i \mathbf{Y}_o\|_2^2 \leq \sum_{i\in[n]} \sum_{j,j'\in[k]} \left(\sum_{s\in[n]} (\mathbf{D}_i)_s x_s^{j,j'} \right)^2 \leq \gamma^2 \rho(\mathbf{Y}_o) nk^3 \|\mathbf{Y}_o - \mathbf{V}\|_2^2.$$

as needed.

□

We now use the two technical lemmas to prove the guarantees for the iterate after one update step.

Lemma 12 (Update, main)**.** *Let $\mathbf{Y}$ be a (column) orthogonal matrix in $\mathbb{R}^{n\times k}$, and $\mathsf{dist}_c^2(\mathbf{Y},\mathbf{V}) \leq \min\{\frac{1}{2}, \frac{\underline{\lambda}^2 n}{384\mu k^2 D_1}\}$ for $D_1 = \max_{i\in[n]}\sum_j (\mathbf{D}_i)_j$.*

Define $\widetilde{\mathbf{X}} \leftarrow \mathrm{argmin}_{\mathbf{X}\in\mathbb{R}^{n\times k}} \left\|\mathbf{M}-\mathbf{X}\mathbf{Y}^\top\right\|_{\mathbf{W}}$. Let $\overline{\mathbf{X}}$ a $n\times k$ matrix such that for each row:

$$\overline{\mathbf{X}}^i = \begin{cases} \widetilde{\mathbf{X}}^i & \text{if } \|\widetilde{\mathbf{X}}^i\|_2^2 \leq \xi = \frac{2\mu k}{n} \\ 0 & \text{otherwise.} \end{cases}$$

Suppose $\overline{\mathbf{X}}$ has QR decomposition $\overline{\mathbf{X}} = \mathbf{X}\mathbf{R}$. Then
(1) $\|\overline{\mathbf{X}} - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\|_F^2 \leq \Delta_u^2 := \left(\frac{108\xi\mu^2k^3\gamma^2 D_1}{\underline{\lambda}^2} + \frac{160\gamma^2\mu\rho(\mathbf{Y})k^4}{\underline{\lambda}^2}\right)\mathsf{dist}_c(\mathbf{Y},\mathbf{V})^2 + \frac{160k}{\underline{\lambda}^2}\|\mathbf{W}\odot\mathbf{N}\|_2^2$.
(2) If $\Delta_u \leq \frac{1}{8}\sigma_{\min}(\mathbf{M}^*)$, then

$$\mathsf{dist}_c(\mathbf{U},\mathbf{X}) \leq \frac{8}{\sigma_{\min}(\mathbf{M}^*) - 2\Delta_u}\Delta_u \;\; \text{and} \;\; \rho(\mathbf{X}) \leq \frac{4\mu}{\sigma_{\min}(\mathbf{M}^*) - 2\Delta_u}.$$

Proof of Lemma 12. (1) By KKT condition, we know that for orthogonal $\mathbf{Y}$, the optimal $\widetilde{\mathbf{X}}$ satisfies

$$\left(\mathbf{W}\odot\left[\mathbf{M} - \widetilde{\mathbf{X}}\mathbf{Y}^\top\right]\right)\mathbf{Y} = 0$$

which implies that the i-th row $\widetilde{\mathbf{X}}^i$ of $\widetilde{\mathbf{X}}$ is given by

$$\widetilde{\mathbf{X}}^i = \mathbf{M}^i\mathbf{D}_i\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}\right)^{-1} = (\mathbf{M}^*)^i\mathbf{D}_i\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}\right)^{-1} + \mathbf{N}^i\mathbf{D}_i\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}\right)^{-1}.$$

Let us consider the first term, by $\mathbf{M}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, we know that

$$\begin{aligned}(\mathbf{M}^*)^i\mathbf{D}_i\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}\right)^{-1} &= \mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{D}_i\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}\right)^{-1} \\ &= \mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{Y}\mathbf{Y}^\top + \mathbf{Y}_\perp\mathbf{Y}_\perp^\top)\mathbf{D}_i\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}\right)^{-1} \\ &= \mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y} + \mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}_\perp\mathbf{Y}_\perp^\top\mathbf{D}_i\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}\right)^{-1}\end{aligned}$$

which implies that

$$\widetilde{\mathbf{X}}^i - \mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y} = \mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}_\perp\mathbf{Y}_\perp^\top\mathbf{D}_i\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}\right)^{-1} + \mathbf{N}^i\mathbf{D}_i\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}\right)^{-1}$$

Let us consider set

$$\mathcal{S}_1 = \left\{ i\in[n] \,\middle|\, \sigma_{\min}(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}) \leq \frac{\underline{\lambda}}{4}\right\}$$

Now we have:

$$\begin{aligned}\sum_{i\notin\mathcal{S}_1}\left\|\widetilde{\mathbf{X}}^i - \mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\right\|_2^2 &\leq \frac{16}{\underline{\lambda}^2}\sum_{i\notin\mathcal{S}_1}\left(2\|\mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}_\perp\mathbf{Y}_\perp^\top\mathbf{D}_i\mathbf{Y}\|_2^2 + 2\|\mathbf{N}^i\mathbf{D}_i\mathbf{Y}\|_2^2\right) \\ &\leq \frac{32\mu k\|\mathbf{\Sigma}\|_2^2}{n\underline{\lambda}^2}\sum_{i\notin\mathcal{S}_1}\|\mathbf{V}^\top\mathbf{Y}_\perp\mathbf{Y}_\perp^\top\mathbf{D}_i\mathbf{Y}\|_2^2 + \frac{32}{\underline{\lambda}^2}\sum_{i\in[n]}\|\mathbf{N}^i\mathbf{D}_i\mathbf{Y}\|_2^2 \\ &\leq \frac{32\mu k\|\mathbf{\Sigma}\|_2^2}{n\underline{\lambda}^2}\sum_{i\in[n]}\|\mathbf{V}^\top\mathbf{Y}_\perp\mathbf{Y}_\perp^\top\mathbf{D}_i\mathbf{Y}\|_2^2 + \frac{32}{\underline{\lambda}^2}\|(\mathbf{W}\odot\mathbf{N})\mathbf{Y}\|_F^2 \\ &\leq \Delta_g := \frac{32\gamma^2\mu\rho(\mathbf{Y})k^4}{\underline{\lambda}^2}\mathsf{dist}_c(\mathbf{Y},\mathbf{V})^2 + \frac{32k}{\underline{\lambda}^2}\|(\mathbf{W}\odot\mathbf{N})\|_2^2.\end{aligned}$$

where the last inequality is due to Lemma 11. Note that since $\xi = \frac{2\mu k}{n} \geq 2\|\mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\|_2^2$, this implies

$$\left|\left\{i \in [n] - \mathcal{S}_1 \left| \|\widetilde{\mathbf{X}}^i\|_2^2 \geq \xi \right.\right\}\right| \quad \leq \quad \left|\left\{i \in [n] - \mathcal{S}_1 \left| \|\widetilde{\mathbf{X}}^i - \mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\|_2^2 \geq \frac{\xi}{2} \right.\right\}\right| \leq \frac{2\Delta_g}{\xi}.$$

Let $\mathcal{S}_2 = \left\{i \in [n] - \mathcal{S}_1 \left| \|\widetilde{\mathbf{X}}^i\|_2^2 \geq \xi \right.\right\}$, we have:

$$\begin{aligned}
\left\|\overline{\mathbf{X}} - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\right\|_F^2 &= \sum_{i=1}^n \left\|\overline{\mathbf{X}}^i - \mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\right\|_2^2 \qquad \text{(because } \|\overline{\mathbf{X}}^i\|_2^2 \leq \xi \quad \text{and } \|\mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\|_2^2 \leq \xi) \\
&\leq \sum_{i \in \mathcal{S}_1 \cup \mathcal{S}_2} 2\xi + \sum_{i \notin \mathcal{S}_1 \cup \mathcal{S}_2} \left\|\widetilde{\mathbf{X}}^i - \mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\right\|_2^2 \\
&\leq 2\xi(|\mathcal{S}_1| + |\mathcal{S}_2|) + \sum_{i \notin \mathcal{S}_1 \cup \mathcal{S}_2} \left\|\widetilde{\mathbf{X}}^i - \mathbf{U}^i\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\right\|_2^2 \\
&\leq 2\xi|\mathcal{S}_1| + 4\Delta_g + \Delta_g.
\end{aligned}$$

By Lemma 10, we know that $|\mathcal{S}_1| \leq \frac{54\mu^2k^3\gamma^2D_1}{\underline{\lambda}^2}\|\mathbf{V} - \mathbf{Y}\|_2^2$. Further plugging in Δ_g, we have

$$\begin{aligned}
&\left\|\overline{\mathbf{X}} - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\right\|_F^2 \\
\leq\ & 2\xi\frac{54\mu^2k^3\gamma^2D_1}{\underline{\lambda}^2}\|\mathbf{V} - \mathbf{Y}\|_2^2 + \frac{160\gamma^2\mu\rho(\mathbf{Y})k^4}{\underline{\lambda}^2}\|\mathbf{Y} - \mathbf{V}\|_2^2 + \frac{160k}{\underline{\lambda}^2}\|(\mathbf{W} \odot \mathbf{N})\|_2^2 \\
=\ & \left(\frac{108\xi\mu^2k^3\gamma^2D_1}{\underline{\lambda}^2} + \frac{160\gamma^2\mu\rho(\mathbf{Y})k^4}{\underline{\lambda}^2}\right)\|\mathbf{Y} - \mathbf{V}\|_2^2 + \frac{160k}{\underline{\lambda}^2}\|(\mathbf{W} \odot \mathbf{N})\|_2^2.
\end{aligned}$$

(2) Denote $\mathbf{B} = \mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}$. Then,

$$\sin\theta(\mathbf{U}, \mathbf{X}) = \|\mathbf{U}_\perp^\top\mathbf{X}\|_2 = \|\mathbf{U}_\perp^\top(\overline{\mathbf{X}} - \mathbf{U}\mathbf{B})\mathbf{R}^{-1}\|_2 \leq \|\overline{\mathbf{X}} - \mathbf{U}\mathbf{B}\|_2\|\mathbf{R}^{-1}\|_2 = \frac{1}{\sigma_{\min}(\overline{\mathbf{X}})}\|\overline{\mathbf{X}} - \mathbf{U}\mathbf{B}\|_2$$

Since $\|\overline{\mathbf{X}} - \mathbf{U}\mathbf{B}\|_2 \leq \Delta_u$, we have

$$\sigma_{\min}(\overline{\mathbf{X}}) \geq \sigma_{\min}(\mathbf{U}\mathbf{B}) - \Delta_u = \sigma_{\min}(\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}) - \Delta_u \geq \sigma_{\min}(\mathbf{M}^*)\cos\theta(\mathbf{Y}, \mathbf{V}) - \Delta_u.$$

By the assumption $\cos\theta(\mathbf{Y}, \mathbf{V}) \geq 1/2$, so

$$\sin\theta(\mathbf{U}, \mathbf{X}) \leq \frac{2}{\sigma_{\min}(\mathbf{M}^*) - 2\Delta_u}\Delta_u.$$

When $\Delta_u \leq \frac{1}{8}\sigma_{\min}(\mathbf{M}^*)$, the right hand side is smaller than $1/3$, so $\cos\theta(\mathbf{U}, \mathbf{X}) \geq 1/2$, and thus $\tan\theta(\mathbf{U}, \mathbf{X}) \leq 2\sin\theta(\mathbf{U}, \mathbf{X})$. Then the statement on $\mathsf{dist}_c(\mathbf{U}, \mathbf{X})$ follows from $\mathsf{dist}_c(\mathbf{U}, \mathbf{X}) \leq 2\tan\theta(\mathbf{U}, \mathbf{X}) \leq 4\sin\theta(\mathbf{U}, \mathbf{X})$.

Finally, observe that $\mathbf{X}^i = \overline{\mathbf{X}}^i\mathbf{R}^{-1}$, so

$$\|\mathbf{X}^i\|_2 \leq \|\overline{\mathbf{X}}^i\|_2\|\mathbf{R}^{-1}\|_2 \leq \frac{\xi}{\sigma_{\min}(\overline{\mathbf{X}})}$$

which leads to the bound. □

B.4. Putting everything together: proofs of the main theorems

Finally, in this section we put things together and prove the main theorems.

We first proceed to the SVD-initialization based algorithm:

Theorem 1. *If* $\mathbf{M}^*, \mathbf{W}$ *satisfy assumptions* ***(A1)-(A3)****, and*

$$\gamma = O\left(\min\left\{\sqrt{\frac{n}{D_1}}\frac{\underline{\lambda}}{\tau\mu^{3/2}k^2}, \frac{\underline{\lambda}}{\tau^{3/2}\mu k^2}\right\}\right),$$

then after $O(\log(1/\epsilon))$ *rounds Algorithm 1 with initialization from Algorithm 3 outputs a matrix* $\widetilde{\mathbf{M}}$ *that satisfies*

$$||\widetilde{\mathbf{M}} - \mathbf{M}^*||_2 \leq O\left(\frac{k\tau}{\underline{\lambda}}\right)||\mathbf{W}\odot\mathbf{N}||_2 + \epsilon.$$

The running time is polynomial in n *and* $\log(1/\epsilon)$.

Proof of Theorem 1. We first show by induction $\mathsf{dist}_c(\mathbf{X}_t, \mathbf{U}) \leq \frac{1}{2^t} + 70\frac{k}{\underline{\lambda}\sigma_{\min}(\mathbf{M}^*)}\delta$ for $t > 1$, and $\mathsf{dist}_c(\mathbf{Y}_t, \mathbf{U}) \leq \frac{1}{2^t} + 70\frac{k}{\underline{\lambda}\sigma_{\min}(\mathbf{M}^*)}\delta$ for $t \geq 1$.

First, by Lemma 8, $\mathbf{Y}_1$ satisfies

$$\mathsf{dist}_c(\mathbf{V}, \mathbf{Y}_1) \leq 8k\Delta_1 = \frac{64k(\gamma\mu k + \delta)}{\sigma_{\min}(\mathbf{M}^*)}.$$

Since $\gamma = O\left(\frac{1}{\tau k^2\mu}\right)$, the base case follows. Now proceed to the inductive step and prove the statement for $t+1$ assuming it is true for t. Now we can apply Lemma 12. By taking the constants within the $O(\cdot)$ notation for γ sufficiently small and by the inductive hypothesis, we have

$$\left(\frac{108\xi\mu^2k^3\gamma^2D_1}{\underline{\lambda}^2} + \frac{160\gamma^2\mu\rho(\mathbf{Y}_1)k^4}{\underline{\lambda}^2}\right) \leq \frac{1}{100}\sigma^2_{\min}(\mathbf{M}^*)$$

and

$$\Delta_u \leq \frac{1}{8}\sigma_{\min}(\mathbf{M}^*).$$

By Lemma 12, we get

$$\begin{aligned}
&\mathsf{dist}_c(\mathbf{U}, \mathbf{X}_{t+1}) \leq \frac{2}{\sigma_{\min}(\mathbf{M}^*) - 2\Delta_u}\Delta_u \leq \frac{8}{3\sigma_{\min}(\mathbf{M}^*)}\Delta_u \\
&= \frac{8}{3\sigma_{\min}(\mathbf{M}^*)}\sqrt{\left(\frac{108\xi\mu^2k^3\gamma^2D_1}{\underline{\lambda}^2} + \frac{160\gamma^2\mu\rho(\mathbf{Y}_1)k^4}{\underline{\lambda}^2}\right)\mathsf{dist}_c^2(\mathbf{U}, \mathbf{X}_t) + \frac{160k}{\underline{\lambda}^2}\delta^2} \\
&\leq \frac{8}{3\sigma_{\min}(\mathbf{M}^*)}\left(\sqrt{\left(\frac{108\xi\mu^2k^3\gamma^2D_1}{\underline{\lambda}^2} + \frac{160\gamma^2\mu\rho(\mathbf{Y}_1)k^4}{\underline{\lambda}^2}\right)\mathsf{dist}_c^2(\mathbf{Y}_t, \mathbf{V})} + \sqrt{\frac{160k}{\underline{\lambda}^2}\delta^2}\right) \qquad \text{(using } \sqrt{a+b} \leq \sqrt{a} + \sqrt{b}\text{)} \\
&\leq \frac{1}{2}\mathsf{dist}_c(\mathbf{Y}_t, \mathbf{V}) + \frac{35\sqrt{k}}{\underline{\lambda}\sigma_{\min}(\mathbf{M}^*)}\delta
\end{aligned}$$

so the statement also holds for $t+1$. This completes the proof for bounding $\mathsf{dist}_c(\mathbf{X}_t, \mathbf{U})$ and $\mathsf{dist}_c(\mathbf{Y}_t, \mathbf{V})$.

Given the bounds on $\mathsf{dist}_c(\mathbf{X}_t, \mathbf{U})$ and $\mathsf{dist}_c(\mathbf{Y}_t, \mathbf{V})$, we are now ready to prove the theorem statement. For simplicity, let $\overline{\mathbf{X}}$ denote $\overline{\mathbf{X}}_{T+1}$ and $\mathbf{Y}$ denote $\mathbf{Y}_T$, so the algorithm outputs $\widetilde{\mathbf{M}} = \overline{\mathbf{X}}\mathbf{Y}$.

By Lemma 12,

$$\|\overline{\mathbf{X}} - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\|_F^2 \leq \Delta_u^2 := \left(\frac{108\xi\mu^2k^3\gamma^2D_1}{\underline{\lambda}^2} + \frac{160\gamma^2\mu\rho(\mathbf{Y})k^4}{\underline{\lambda}^2}\right)\mathsf{dist}_c(\mathbf{Y}, \mathbf{V})^2 + \frac{160k}{\underline{\lambda}^2}\|\mathbf{W}\odot\mathbf{N}\|_2^2.$$

Plugging the choice of γ and noting $\xi = \frac{2\mu k}{n}$ and $\rho(\mathbf{Y}) = O(\mu/\sigma_{\min}(\mathbf{M}^*))$, we have

$$\|\overline{\mathbf{X}} - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\|_F^2 \leq \Delta_u^2 = O\left(\mathsf{dist}_c(\mathbf{Y}, \mathbf{V})^2\right) + \frac{160k}{\underline{\lambda}^2}\|\mathbf{W}\odot\mathbf{N}\|_2^2$$

which leads to

$$\|\overline{\mathbf{X}} - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\|_F \le \Delta_u \le O\left(\mathsf{dist}_c(\mathbf{Y},\mathbf{V})\right) + \frac{16\sqrt{k}}{\underline{\lambda}}\|\mathbf{W}\odot\mathbf{N}\|_2.$$

Now consider $\|\mathbf{M}^* - \widetilde{\mathbf{M}}\|_2 = \|\mathbf{M}^* - \overline{\mathbf{X}}\mathbf{Y}^\top\|_2$. By definition, we know that there exists $\mathbf{Q}$ such that $\mathbf{Y} = \mathbf{V}\mathbf{Q} + \Delta_y$ where $\|\Delta_y\|_2 = O(\mathsf{dist}_c(\mathbf{Y},\mathbf{V}))$. Also, let $\mathbf{R} = \overline{\mathbf{X}} - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}$.

$$\begin{aligned}\widetilde{\mathbf{M}} - \mathbf{M}^* &= \left[\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{V}\mathbf{Q}+\Delta_y) + \mathbf{R}\right](\mathbf{V}\mathbf{Q}+\Delta_y)^\top - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{Q}\Delta_y^\top + \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\Delta_y(\mathbf{V}\mathbf{Q}+\Delta_y)^\top + \mathbf{R}(\mathbf{V}\mathbf{Q}+\Delta_y)^\top \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{Q}\Delta_y^\top + \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\Delta_y\mathbf{Y}^\top + \mathbf{R}\mathbf{Y}^\top.\end{aligned}$$

Therefore,

$$\begin{aligned}\|\widetilde{\mathbf{M}} - \mathbf{M}^*\|_2 &\le \|\mathbf{U}\mathbf{\Sigma}\|_2\|\mathbf{Q}\|_2\|\Delta_y\|_2 + \|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\|_2\|\Delta_y\|_2\|\mathbf{Y}\|_2 + \|\mathbf{R}\|_2\|\mathbf{Y}\|_2 \\ &\le 2\|\Delta_y\|_2 + \|\mathbf{R}\|_2 \\ &\le O\left(\mathsf{dist}_c(\mathbf{Y},\mathbf{V})\right) + \frac{16\sqrt{k}}{\underline{\lambda}}\|\mathbf{W}\odot\mathbf{N}\|_2.\end{aligned}$$

Combining this with the bound on $\mathsf{dist}_c(\mathbf{Y}_T,\mathbf{V})$, the theorem then follows. □

Next, we show the main theorem for random initialization:

Theorem 3 (Main, random initialization)**.** *Suppose* $\mathbf{M}^*, \mathbf{W}$ *satisfy assumptions* ***(A1)-(A3)*** *with*

$$\begin{aligned}\gamma &= O\left(\min\left\{\sqrt{\frac{n}{D_1}}\frac{\underline{\lambda}}{\tau\mu^2k^{5/2}}, \frac{\underline{\lambda}}{\tau^{3/2}\mu^{3/2}k^{5/2}}\right\}\right), \\ \|\mathbf{W}\|_\infty &= O\left(\frac{\underline{\lambda}n}{k^2\mu\log^2 n}\right),\end{aligned}$$

where $D_1 = \max_{i\in[n]}\|\mathbf{W}^i\|_1$*. Then after* $O(\log(1/\epsilon))$ *rounds Algorithm 1 using initialization from Algorithm 4 outputs a matrix* $\widetilde{\mathbf{M}}$ *that with probability at least* $1 - 1/n^2$ *satisfies*

$$\|\widetilde{\mathbf{M}} - \mathbf{M}^*\|_2 \le O\left(\frac{k\tau}{\underline{\lambda}}\right)\|\mathbf{W}\odot\mathbf{N}\|_2 + \epsilon.$$

The running time is polynomial in n *and* $\log(1/\epsilon)$*.*

Proof of Theorem 3. Let $\mathbf{Y}$ be initialized using the random initialization algorithm 4. Consider applying the proof in Lemma 12, with $\mathcal{S}_1$ being modified to be

$$\mathcal{S}_1 = \left\{i\in[n]\,\middle|\,\sigma_{\min}(\mathbf{Y}^\top\mathbf{D}_i\mathbf{Y}) \le \frac{\underline{\lambda}}{4\mu k}\right\}$$

But with this modification, $\mathcal{S}_1 = \emptyset$, with high probability. Then the same calculation from Lemma 12 (which now doesn't need to use Lemma 10 at all since $\mathcal{S}_1 = \emptyset$) gives

$$\left\|\overline{\mathbf{X}} - \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{Y}\right\|_F^2 \le \Delta_g\mu k$$

But following part (2) of the same Lemma, we get that if $\Delta_g\mu k < \frac{1}{8}\sigma_{\min}(M^*)$,

$$\mathsf{dist}_c(\mathbf{U},\mathbf{X}) \le \frac{2}{\sigma_{\min}(\mathbf{M}^*) - 2\Delta_g\mu k}\Delta_g\mu k$$

So, in order to argue by induction in 1 exactly as before, we only need to check that after the update step for $\mathbf{X}$, $\mathsf{dist}_c(\mathbf{U}, \mathbf{X})$ is small enough to apply Lemma 12 for later steps. Indeed, we have:

$$\mathsf{dist}_c(\mathbf{U}, \mathbf{X}) \leq \frac{2}{\sigma_{\min}(\mathbf{M}^*) - 2\Delta_g \mu k} \Delta_g \mu k \leq \sqrt{\min\left\{\frac{1}{2}, \frac{\underline{\lambda}^2 n}{384 \mu k^2 D_1}\right\}}$$

Noticing that Δ_g has a quadratic dependency on γ, we see that if

$$\gamma = O\left(\min\left\{\sqrt{\frac{n}{D_1}}\frac{\underline{\lambda}\sigma_{\min}(\mathbf{M}^*)}{\mu^2 k^{5/2}}, \frac{\underline{\lambda}\sigma_{\min}^{3/2}(\mathbf{M}^*)}{\mu^{3/2} k^{5/2}}\right\}\right),$$

the inequality is indeed satisfied.

With that, the theorem statement follows.

□

B.5. Estimating $\sigma_{\max}(\mathbf{M}^*)$

Finally, we show that we can estimate $\sigma_{\max}(\mathbf{M}^*)$ up to a very good accuracy, so that we can apply our main theorems to matrices with arbitrary $\sigma_{\max}(\mathbf{M}^*)$. This is quite easy: the estimate of it is just $\|\mathbf{W} \odot \mathbf{M}\|_2$. Then, the following lemma holds:

Lemma 13. *It $\gamma = o(\frac{1}{k\mu})$ and $\delta = \|\mathbf{W} \odot \mathbf{N}\|_2 = o(\sigma_{\max}(\mathbf{M}^*))$ then $\|\mathbf{W} \odot \mathbf{M}\|_2 = (1 \pm o(1))(\sigma_{\max}(\mathbf{M}^*))$*

Proof. We proceed separately for the upper and lower bound.

For the upper bound, we have

$$\begin{aligned}\|\mathbf{W} \odot \mathbf{M}\|_2 &= \|\mathbf{W} \odot \mathbf{M}^* + \mathbf{W} \odot \mathbf{N}\|_2 \leq \|\mathbf{W} \odot \mathbf{M}^*\|_2 + \|\mathbf{W} \odot \mathbf{N}\|_2 \\ &\leq \|(\mathbf{W} - \mathbf{E}) \odot \mathbf{M}^*\|_2 + \|\mathbf{E} \odot \mathbf{M}^*\|_2 + \|\mathbf{W} \odot \mathbf{N}\|_2 \\ &\leq \gamma k \mu \sigma_{\max}(\mathbf{M}^*) + \sigma_{\max}(\mathbf{M}^*) + \delta \leq (1 + o(1))\sigma_{\max}(\mathbf{M}^*). \qquad \text{(by Lemma 5)}\end{aligned}$$

For the lower bound, completely analogously we have

$$\begin{aligned}\|\mathbf{W} \odot \mathbf{M}\|_2 &= \|\mathbf{W} \odot \mathbf{M}^* + \mathbf{W} \odot \mathbf{N}\|_2 \geq \|\mathbf{W} \odot \mathbf{M}^*\|_2 - \|\mathbf{W} \odot \mathbf{N}\|_2 \\ &\geq \|\mathbf{E} \odot \mathbf{M}^*\|_2 - \|(\mathbf{W} - \mathbf{E}) \odot \mathbf{M}^*\|_2 - \|\mathbf{W} \odot \mathbf{N}\|_2 \\ &\geq \sigma_{\max}(\mathbf{M}^*) - \gamma k \mu \sigma_{\max}(\mathbf{M}^*) - \delta \geq (1 - o(1))\sigma_{\max}(\mathbf{M}^*) \qquad \text{(by Lemma 5)}\end{aligned}$$

which finishes the proof. □

Given this, the reduction to the case $\sigma_{\max}(\mathbf{M}^*) \leq 1$ is obvious: first, we scale the matrix $\mathbf{M}$ down by our estimate of $\sigma_{\max}(\mathbf{M}^*)$ and run our algorithm with, say, four times as many rounds. After this, we rescale the resulting matrix $\widetilde{\mathbf{M}}$ by our estimate of $\sigma_{\max}(\mathbf{M}^*)$, after which the claim of Theorems 1 and 3 follows.

C. Empirical verification of the spectral gap property

Experiments on the performance of the alternating minimization can be found in related work (e.g., (Lu et al., 1997; Srebro & Jaakkola, 2003)). Therefore, we focus on verifying the key assumption, i.e., the spectral gap property of the weight matrix (Assumption $(\mathbf{A2})$).

Here we consider the application of computing word embeddings by factorizing the co-occurrence matrix between the words, which is one of the state-of-the-art techniques for mapping words to low-dimensional vectors (about 300 dimension) in natural language processing. There are many variants (e.g., (Levy & Goldberg, 2014; Pennington et al., 2014; Arora et al., 2016)); we consider the following simple approach. Let X be the co-occurrence matrix, where $\mathbf{X}_{i,j}$ is the number

of times that word i and word j appear together within a window of small size (we use size 10 here) in the given corpus. Then the word embedding by weighted low rank problem is

$$\min_{\mathbf{V}} \sum_{i,j} f(\mathbf{X}_{i,j}) \left(\log \left(\frac{\mathbf{X}_{i,j}}{X} \right) - \langle \mathbf{V}_i, \mathbf{V}_j \rangle \right)^2$$

where $X = \sum_{i,j} \mathbf{X}_{i,j}$, $\mathbf{V}_i$'s are the vectors for the words, and $f(x) = \max\{\mathbf{X}_{i,j}, 100\}$ for a large corpus and $f(x) = \max\{\mathbf{X}_{i,j}, 10\}$ for a small corpus.

We focus on the weight matrix $\mathbf{W}_{i,j} = f(\mathbf{X}_{i,j})$. It has been observed that using $\mathbf{X}_{i,j}$ as weights is roughly the maximum likelihood estimator under certain probabilistic model and is better than using uniform weights. It has also been verified that using the truncated weight $f(\mathbf{X}_{i,j})$ is better than using $\mathbf{X}_{i,j}$. Our experiments suggest that $f(\mathbf{X}_{i,j})$ is better partially due to the requirement that the weight matrix should have the spectral gap property for the algorithm to succeed.

We consider two large corpus (Wikipedia corpus (Wikimedia, 2012), about 3G tokens; a subset of Commoncrawl corpus (Buck et al., 2014), about 20G tokens). For each corpus, we pick the top n words ($n = 500, 1000, \ldots, 5000$) and compute the spectral gap $\|\mathbf{W} - \mathbf{E}\|_2$ where $\mathbf{W}$ is the weight matrix corresponding to the words, and $\mathbf{E}$ is the all-one matrix. Note that a scaling of $\mathbf{W}$ does not affect the problem, so we enumerate different scaling of $\mathbf{W}$ (from 2^{-20} to 2^{10}) and plot the best spectral gap. We compare the two variants: with threshold ($\mathbf{W}_{i,j} = f(\mathbf{X}_{i,j})$), and without threshold ($\mathbf{W}_{i,j} = \mathbf{X}_{i,j}$).

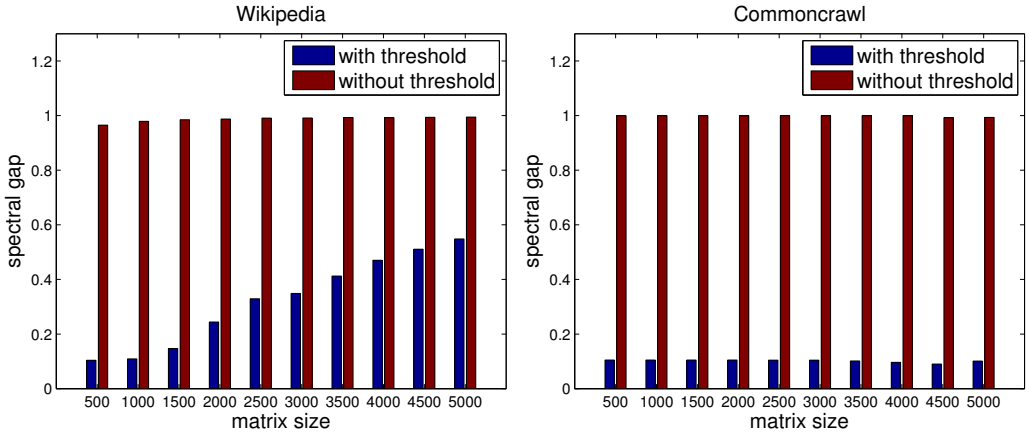

Figure 1. Spectral gap of the weight matrix for word embeddings on Wikipedia corpus. x-axis: number of words (size of the matrix); y-axis: the spectral gap $\|\mathbf{W} - \mathbf{E}\|_2$ where $\mathbf{E}$ is the all-one matrix.

The results are shown in Figure 1. Without threshold, there is almost no spectral gap. With threshold, there is a decent gap, though with the increase of the matrix size, the gap become smaller because larger vocabulary includes more uneven co-occurrence entries and thus more noise. This suggests that thresholding can make the weight matrix nicer for the algorithm, and thus leads to better performance.