A. Appendix

In this section, we provide the proofs of several technical results that are claimed or used in our main paper.

A.1. Proof of Proposition 4.1

The proof follows via a reduction from the so-called SubsetSum problem, which is known to be NP-hard (Garey & Johnson, 1979). Recall that the SubsetSum decision problem is as follows: given n numbers, $a_1, \dots, a_n$ in $\mathbb{R}$, decide if there exists a partition $S \subseteq [n]$ such that

$$\sum_{i \in S} a_i = \sum_{j \in S^c} a_j.$$

We show that if we can solve the mixed linear equations problem in polynomial time, then we can solve the SubsetSum problem, which would thus imply that $P = NP$.

Given $\{a_1, \dots, a_n\}$, we must design a matrix X, and output variable $\mathbf{y}$, such that if we could solve the mixed linear equation problem specified by $(\mathbf{y}, X)$, then we could decide the subset sum problem on $\{a_1, \dots, a_n\}$. To this end, we define:

$$X = \begin{bmatrix} & I_n & \\ & I_n & \\ a_1 & \cdots & a_n \end{bmatrix}, \qquad \mathbf{y} = \begin{pmatrix} \mathbf{1}_{n \times 1} \\ \mathbf{0}_{n \times 1} \\ \sum_i a_i/2 \end{pmatrix}.$$

Here, I_n denotes the $n \times n$ identity matrix, $\mathbf{1}_{n\times 1}$ the $n \times 1$ vector of 1's, and similarly, $\mathbf{0}_{n \times 1}$ the $n \times 1$ vector of 0's. Finding a solution to the mixed linear equations problem amounts to finding a subset $S \subseteq [2n+1]$ of the $2n+1$ constraints, and vectors $\beta^{(1)}, \beta^{(2)} \in \mathbb{R}^n$, so that $\beta^{(1)}$ satisfies the equalities $X_S \beta^{(1)} = \mathbf{y}_S$, and $\beta^{(2)}$ the equalities $X_{S^c}\beta_2 = \mathbf{y}_{S^c}$. Note that S cannot contain i and $n+i$, since these equalities are mutually exclusive. The consequence is that we have $\beta_i^{(1)} \in \{0, 1\}$, with $\beta_i^{(1)} = 1 - \beta_i^{(2)}$. Thus if the first $2n$ constraints are satisfied, the final constraint, therefore, can only be satisfied if we have

$$\sum_{i \in S} a_i = \sum_i a_i \beta_i^{(1)} = \sum_j a_j \beta_j^{(2)} = \sum_{j \in S^c} a_j,$$

thus proving the result.

A.2. Proof of Proposition 4.2

To show that our SVD initialization produces a good initial solution, requires two steps. Recall that Algorithm 5 finds the two dimensional subspace spanned by the top two eigenvectors of the matrix $M = \frac{1}{|\mathcal{S}_*|}\sum_{i \in \mathcal{S}_*} y_i^2 \mathbf{x}_i \otimes \mathbf{x}_i$, and then searches on a discretization of the circle in that subspace for two vectors that minimize the loss function, $\mathcal{L}_+$ evaluated on the samples in $\mathcal{S}_+$.

We first show that the top eigenspace of M is indeed close to the top eigenspace of its expectation, $p_1 \beta_1^* \otimes \beta_1^* + p_2 \beta_2^* \otimes \beta_2^* + I$, i.e., it is close to $\text{span}\{\beta_1^*, \beta_2^*\}$, and that some pair of elements of the discretization are close to (β_1^*, β_2^*). This is the content of lemma A.1. We then show that our loss function $\mathcal{L}_+$ is able to select good points from the discretization.

Our algorithm then uses the loss function $\mathcal{L}_+$ (evaluated on new samples in $\mathcal{S}_+$) to select good points from the grid G. Lemma A.2 shows that as long as the number $\mathcal{S}_+$ of these new samples is large enough, we can upper *and* lower bound, with high probability, the empirically evaluated loss $\mathcal{L}_+(\hat\beta_1, \hat\beta_2)$ of any candidate pair $\hat\beta_1, \hat\beta_2$ by the true error err of that candidate pair. This provides the critical result allowing us to do the correct selection in the 1-d search phase.

Now we are ready to prove the result. Suppose the conditions of lemma A.1 hold. Then we are guaranteed the existence of $(\bar\beta_1, \bar\beta_2)$ in the grid G with δ-resolution, such that $\max_i \|\bar\beta_i - \beta_i^*\| < \delta$. Next, let $(\beta_1^{(0)}, \beta_2^{(0)})$ be the output of our SVD initialization, and let err denote their distance from (β_1^*, β_2^*). By definition, the vectors $(\beta_1^{(0)}, \beta_2^{(0)})$ minimize the loss function $\mathcal{L}_+$ taken on inputs $\mathcal{S}_+$, and hence $\mathcal{L}_+(\beta_1^{(0)}, \beta_2^{(0)}) \le \mathcal{L}_+(\bar\beta_1, \bar\beta_2)$. Using the lower bound from lemma A.2, applied to $(\beta_1^{(0)}, \beta_2^{(0)})$ we have:

$$\frac{1}{5}\sqrt{\min\{p_1,p_2\}}\mathrm{err} \le \sqrt{\frac{\mathcal{L}_+(\beta_1^{(0)}, \beta_2^{(0)})}{|\mathcal{S}_+|}}.$$

From the upper bound applied to $(\bar{\beta}_1, \bar{\beta}_2)$, we have

$$\sqrt{\frac{\mathcal{L}_+(\bar{\beta}_1, \bar{\beta}_2)}{|\mathcal{S}_+|}} \le 1.1\delta.$$

Recalling that $\mathcal{L}_+(\beta_1^{(0)}, \beta_2^{(0)}) \le \mathcal{L}_+(\bar{\beta}_1, \bar{\beta}_2)$, and taking

$$\delta \le \frac{2}{11}\widehat{c}\|\beta_1^* - \beta_2^*\|_2 \sqrt{\min\{p_1,p_2\}}^3,$$

we combine to finally obtain:

$$\begin{aligned}\mathrm{err} &\le \frac{11}{2}\frac{\delta}{\sqrt{\min\{p_1,p_2\}}} \\ &\le \widehat{c}\min\{p_1,p_2\}\|\beta_1^* - \beta_2^*\|_2.\end{aligned}$$

where $\widehat{c}$ is as in the statement of proposition 4.2.

A.3. Proof of Proposition 4.3

Using standard concentration results, in lemma A.1, we have shown if

$$|\mathcal{S}_*| > c(1/\widetilde{\delta})^2 k \log^2 k,$$

with probability at least $1 - \frac{1}{k^2}$,

$$\|M - \mathbb{E}(M)\| < 3\widetilde{\delta}$$

Hence, we have

$$\big||\lambda_1^* - \lambda_2^*| - |\lambda_1 - \lambda_2|\big| \le 6\widetilde{\delta}.$$

The approximate error of Δ_b^* can be bounded as:

$$\begin{aligned}2p_b|\Delta_b^* - \Delta_b| &\le 6\widetilde{\delta} + (p_b^2 - p_{-b}^2)[\frac{1}{\lambda_{-b}^* - \lambda_b^*} - \frac{1}{\lambda_{-b} - \lambda_b}] \\ &\le 6\widetilde{\delta} + |p_b^2 - p_{-b}^2|\frac{6\widetilde{\delta}}{(\lambda_{-b}^* - \lambda_b^*)(\lambda_{-b} - \lambda_b)} \\ &\le 6\widetilde{\delta} + |p_b^2 - p_{-b}^2|\frac{6\widetilde{\delta}}{|\lambda_{-b}^* - \lambda_b^*|(|\lambda_{-b}^* - \lambda_b^*| - 6\widetilde{\delta})} \\ &\le 6\widetilde{\delta} + |p_b^2 - p_{-b}^2|\frac{12\widetilde{\delta}}{|\lambda_{-b}^* - \lambda_b^*|^2}\end{aligned}$$

In the last inequality we use $\widetilde{\delta} \le \frac{|\lambda_1^* - \lambda_2^*|}{12}$.

Next, we calculate approximation error of eigenvectors. Note that $\mathbb{E}(\frac{M-I}{2}) = p_1\beta_1^* \otimes \beta_1^* + p_2\beta_2^* \otimes \beta_2^*$, we have

$$\{\lambda_1^*, \lambda_2^*\} = \{\frac{1+\kappa}{2}, \frac{1-\kappa}{2}\}.$$

Using lemma A.3, we have,

$$\|\mathbf{v}_b - \mathbf{v}_b^*\|_2^2 \le \frac{6\widetilde{\delta}}{\kappa} + \frac{24\widetilde{\delta}}{1-\kappa} \le \frac{24\widetilde{\delta}}{\kappa(1-\kappa)},\ b = 1, 2.$$

Then

$$\|\beta_b^* - \beta_b\|_2 \le \left|\sqrt{\frac{1-\Delta_b^*}{2}}\mathbf{v}_b^* - \sqrt{\frac{1-\Delta_b}{2}}\mathbf{v}_b\right| + \left|\sqrt{\frac{1+\Delta_b^*}{2}}\mathbf{v}_{-b}^* - \sqrt{\frac{1+\Delta_b}{2}}\mathbf{v}_{-b}\right|. \tag{14}$$

Note that

$$\begin{aligned}
\left|\sqrt{\frac{1-\Delta_b^*}{2}}\mathbf{v}_b^* - \sqrt{\frac{1-\Delta_b}{2}}\mathbf{v}_b\right| &= \sqrt{\frac{1-\Delta_b^*}{2}}\mathbf{v}_b^* - \sqrt{\frac{1-\Delta_b^*}{2}}\mathbf{v}_b + \sqrt{\frac{1-\Delta_b^*}{2}}\mathbf{v}_b - \sqrt{\frac{1-\Delta_b}{2}}\mathbf{v}_b\Bigg| \\
&\le \sqrt{\frac{1-\Delta_b^*}{2}}\|\mathbf{v}_b - \mathbf{v}_b^*\|_2 + \left|\sqrt{\frac{1-\Delta_b^*}{2}} - \sqrt{\frac{1-\Delta_b}{2}}\right|\|\mathbf{v}_b\|_2 \\
&\le \|\mathbf{v}_b - \mathbf{v}_b^*\|_2 + \left|\sqrt{\frac{1-\Delta_b^*}{2}} - \sqrt{\frac{1-\Delta_b}{2}}\right| \\
&\le \|\mathbf{v}_b - \mathbf{v}_b^*\|_2 + \sqrt{\frac{1}{2}\left|\Delta_b - \Delta_b^*\right|}.
\end{aligned}$$

Plug the above result back to (14), we obtain

$$\begin{aligned}
\|\beta_b^* - \beta_b\|_2 &\lesssim \sqrt{\left|\Delta_b - \Delta_b^*\right|} + \sum_b \|\mathbf{v}_b - \mathbf{v}_b^*\|_2 \\
&\lesssim \sqrt{\frac{\widetilde{\delta}}{\kappa(1-\kappa)}} + \frac{1}{\sqrt{\min\{p_1, p_2\}}}\sqrt{\widetilde{\delta} + \frac{\widetilde{\delta}}{\kappa^2}} \\
&\lesssim \sqrt{\frac{\widetilde{\delta}}{\min\{p_1, p_2\}}} \times \sqrt{\frac{1}{\kappa(1-\kappa)} + \frac{1}{\kappa^2}} \\
&= \sqrt{\frac{\widetilde{\delta}}{\min\{p_1, p_2\}}}\frac{1}{\kappa\sqrt{1-\kappa}}.
\end{aligned}$$

By setting the above upper bound to be less than $\widehat{c}\min\{p_1, p_2\}\|\beta_1^* - \beta_2^*\|_2$, we complete the proof.

A.4. Proof of Proposition 4.5

It's equivalent to show that $J_b = J_b^*, b = 1, 2$. Let's consider $b = 1$, that is for all $p_1 * |\mathcal{S}_t|$ samples that are generated by $y = \mathbf{x}^T\beta_1^*$. For simplicity, let β_1, β_2 denote $\beta_1^{(t-1)}, \beta_2^{(t-1)}$, we need

$$\left(\mathbf{x}^T(\beta_1^* - \beta_1)\right)^2 < \left(\mathbf{x}^T(\beta_1^* - \beta_2)\right)^2.$$

From lemma 5.1,

$$\begin{aligned}
\mathbb{P}\left[\left(\mathbf{x}^T(\beta_1^* - \beta_1)\right)^2 < \left(\mathbf{x}^T(\beta_1^* - \beta_2)\right)^2\right] &\ge 1 - \frac{\|\beta_1^* - \beta_1\|_2}{\|\beta_1^* - \beta_2\|_2} \qquad (15)\\
&\ge 1 - 2\frac{\|\beta_1^* - \beta_1\|_2}{\|\beta_1^* - \beta_2^*\|_2} \qquad (16)\\
&\ge 1 - \frac{2c_1}{k^2}. \qquad (17)
\end{aligned}$$

Then we use union bound for $p_1 * |\mathcal{S}_t|$ samples in J_1^*,

$$\mathbb{P}\left[\left(\mathbf{x}_i^T(\beta_1^* - \beta_1)\right)^2 < \left(\mathbf{x}_i^T(\beta_1^* - \beta_2)\right)^2,\ for\ all\ i \in J_1^*\right] \ge 1 - p_1 c_2 k \times \frac{2c_1}{k^2} \ge 1 - \frac{c'}{k}.$$

So all samples are correctly clustered with high probability.

As $\frac{1}{\min(p_1, p_2)}k < |\mathcal{S}_t|$, number of samples in J_1 and J_2 are both greater than k. Therefore, least square solution reveals the ground truth. In other words, $err^{(t)} = 0$.

A.5. Proof of Lemma 5.1

(1)
Without loss of generality, we assume $T\{u,v\} = T\{\mathbf{e}_1, \mathbf{e}_2\}$. Let x_1, x_2 denote $\mathbf{x}^T\mathbf{e}_1, \mathbf{x}^T\mathbf{e}_2$. As x_1, x_2 are independent Gaussian random variables, we have $x_1 = A\cos\theta, x_2 = A\sin\theta$, where A is Rayleigh random variable, and θ is uniformly distributed over $[0, 2\pi)$. Conditioning on $(\mathbf{x}^T u)^2 > (\mathbf{x}^T v)^2$, the range of θ is truncated to be $[\theta_0, \theta_0 + \alpha_{(u,v)}] \cup [\theta_0 + \pi, \theta_0 + \pi + \alpha_{(u,v)}]$ for some θ_0. It is not hard to see the eigenvalues of covariance matrix of (x_1, x_2) are $1 + \frac{\sin\alpha_{(u,v)}}{\alpha_{(u,v)}}, 1 - \frac{\sin\alpha_{(u,v)}}{\alpha_{(u,v)}}$. As the rest if the eigenvalues of Σ are 1, this completes the proof.
(2)
Note that

$$\mathbb{P}\left[(\mathbf{x}^T u)^2 > (\mathbf{x}^T v)^2\right] = \frac{\alpha_{(u,v)}}{\pi}.$$

If $\|u\|_2 > \|v\|_2$, $\alpha_{(u,v)} > \frac{\pi}{2}$, when $\|u\|_2 < \|v\|_2$,

$$\cos\alpha_{(u,v)} \geq \frac{\|v\|_2^2 - \|u\|_2^2}{\|u\|_2^2 + \|v\|_2^2}.$$

Note that for any $\alpha \in [0, \pi/2]$, $\alpha \leq \frac{\pi}{2}\sin\alpha$. We have

$$\mathbb{P}\left[(\mathbf{x}^T u)^2 > (\mathbf{x}^T v)^2\right] \leq \frac{1}{2}\sin\alpha_{(u,v)} \leq \frac{\|u\|_2\|v\|_2}{\|u\|_2^2 + \|v\|_2^2} \leq \frac{\|u\|_2}{\|v\|_2}.$$

A.6. Supporting Lemmas

Lemma A.1. *For any given $\delta > 0$, let G denote the grid points, at resolution δ, of the unit circle on the subspace spanned by the top two eigenvectors of M, formed with $|S_*|$ samples. Then, there exists an absolute constant c such that if*

$$|S_*| \geq c(1/\tilde{\delta})^2 k \log^2 k,$$

where

$$\tilde{\delta} = \frac{\delta^2}{384}(1 - \sqrt{1 - 4(1 - \langle\beta_1^*, \beta_2^*\rangle^2)p_1 p_2}),$$

then

$$\min_{\mathbf{a}\in G} \|\beta_i^* - \mathbf{a}\| \leq \delta, i = 1, 2,$$

with probability at least $1 - O\left(\frac{1}{k^2}\right)$.

Proof. In order to prove the result, we make use of standard concentration results.

Let $\Sigma = \mathbb{E}[M]$. We observe that $\mathbb{P}\left[|y| > \sqrt{2\alpha\log k}\right] \leq n^{-\alpha}$, $\mathbb{P}\left[\|\mathbf{x}\|_2^2 \geq 3k\right] \leq e^{-k/3}$. Suppose N is much less than $O\left(k^{10}\right)$, where the constant is arbitrarily chosen here. Set $\alpha = 12$. Then with probability at least $1 - O\left(\frac{1}{k^2}\right)$, The vectors $y_i\mathbf{x}_i$ are all supported in a ball with radius $\sqrt{72k\log k}$. Directly following theorem 5.44 in (Vershynin, 2010), we claim that when $N > C(1/\tilde{\delta})^2 k\log^2 k$,

$$\|M - \Sigma\| \leq \tilde{\delta}\|\Sigma\| \leq 3\tilde{\delta}.$$

We use $\sigma_i(A)$ to denote the i'th biggest eigenvalue of the positive semidefinite matrix A. By simple algebraic calculation we get $\sigma_1(\Sigma) = 2 + \kappa$, $\sigma_2(\Sigma) = 2 - \kappa$, where $\kappa = \sqrt{1 - 4(1 - \langle\beta_1^*, \beta_2^*\rangle^2)p_1p_2}$. The top two eigenvectors of Σ are denoted as $\mathbf{v}_1^*$, $\mathbf{v}_2^*$. We use $\mathbf{v}_1$, $\mathbf{v}_2$ to denote the top two eigenvectors of M. Lemma A.3 yields that

$$\begin{aligned}
\|\mathbf{v}_i^* - \mathcal{P}_{T(\mathbf{v}_1,\mathbf{v}_2)}\mathbf{v}_i^*\|_2^2 &\leq \frac{12\tilde{\delta}}{\sigma_2(M) - \sigma_3(M)} \\
&\leq \frac{12\tilde{\delta}}{\sigma_2(\Sigma) - \sigma_3(\Sigma) - 6\tilde{\delta}} \\
&= \frac{12\tilde{\delta}}{1 - \kappa - 6\tilde{\delta}} \\
&= \frac{24\tilde{\delta}}{1 - \kappa}, i = 1, 2.
\end{aligned}$$

The last inequality holds when $\tilde{\delta} \leq \frac{1-\kappa}{12}$. Using the fact that for any two vectors $\mathbf{a}, \mathbf{b}$,$\|\mathbf{a}+\mathbf{b}\|_2^2 \leq 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$, we conclude that

$$\|\beta_i^* - \mathcal{P}_{T(\mathbf{v}_1,\mathbf{v}_2)}\beta_i^*\|_2^2 \leq \frac{48\tilde{\delta}}{1-\kappa}, i = 1, 2.$$

Let $w = \|\beta_i^* - \mathcal{P}_{T(\mathbf{u},\mathbf{v})}\beta_i^*\|_2$. Then, by simple geometric relation,

$$\begin{aligned}\min_{\mathbf{a}\in\mathbb{S}^{k-1}\cap T_{(\mathbf{u},\mathbf{v})}} \|\mathbf{a} - \beta_i^*\|_2^2 &\leq 2 - 2\sqrt{1-w^2} \\ &\leq 2w^2 \\ &\leq (\frac{\epsilon}{2})^2, i = 1, 2.\end{aligned}$$

Consider the δ-resolution grid G. We observe that for any point in $\mathbb{S}^{k-1} \cap T_{(\mathbf{u},\mathbf{v})}$, there exists a point in G that is within $\delta/2$ away from it. By triangle inequality, we end up with

$$\min_{\mathbf{a}\in W} \|\mathbf{a} - \beta_i^*\|_2 \leq \delta. \tag{18}$$

□

Lemma A.2. *Let $\hat{\beta}_1, \hat{\beta}_2$ be any two given vectors with error defined by* $\mathrm{err} := \max_{i=1,2} \|\hat{\beta}_i - \beta_i^*\|$. *There exist constants $c_1, c_2 > 0$ such that as long as we have enough testing samples,*

$$|\mathcal{S}_+| \geq c_1 k / \min\{p_1, p_2\},$$

then with probability at least $1 - O\left(e^{-c_2 k}\right)$

$$\sqrt{\frac{\mathcal{L}_+(\hat{\beta}_1, \hat{\beta}_2)}{|\mathcal{S}_+|}} \leq 1.1\,\mathrm{err}$$

and

$$\sqrt{\frac{\mathcal{L}_+(\hat{\beta}_1, \hat{\beta}_2)}{|\mathcal{S}_+|}} \geq \frac{1}{5}\sqrt{\min\{p_1, p_2\}}\min\left\{\mathrm{err}, \frac{1}{2}\|\beta_1^* - \beta_2^*\|_2\right\}.$$

Proof. Our notation here, namely, J_1, J_2, J_1^*, J_2^*, is consistent with proof of Theorem 4.4. Note that we have:

$$\mathcal{L}(\beta_1, \beta_2) = \sum_i \min_{z_i} z_i(y_i - \mathbf{x}_i^T\beta_1)^2 + (1 - z_i)(y_i - \mathbf{x}_i^T\beta_2)^2.$$

For the upper bound, we assign label z_i as the true label. Then,

$$\mathcal{L} \leq \sum_{i\in J_1^*}(\mathbf{x}_i^T(\beta_1^* - \beta_1))^2 + \sum_{i\in J_2^*}(\mathbf{x}_i^T(\beta_2^* - \beta_2))^2.$$

When $|\mathcal{S}_+| \geq C\frac{k}{\min\{p_1,p_2\}}$, then the number of samples in set J_1^*,J_2^* is also greater than Ck. Following standard concentration results, there exist constants C, c_1, such that with probability greater than $1 - e^{-c_1 k}$, we have

$$\|\frac{1}{p_j|\mathcal{S}_+|}\sum_{i\in J_j^*}(\mathbf{x}_i\mathbf{x}_i^T) - I\| \leq 0.21, j = 1, 2.$$

We have

$$\begin{aligned}\mathcal{L} &\leq 1.21 p_1|\mathcal{S}_+|\|\beta_1 - \beta_1^*\|_2^2 + 1.21 p_2|\mathcal{S}_+|\|\beta_2 - \beta_2^*\|_2^2 \\ &\leq 1.21|\mathcal{S}_+|\mathrm{err}^2.\end{aligned}$$

For the lower bound, we observe that

$$\mathcal{L} = \underbrace{\sum_{i\in J_1\cap J_1^*}(\mathbf{x}_i^T(\beta_1-\beta_1^*))^2 + \sum_{i\in J_2\cap J_1^*}(\mathbf{x}_i^T(\beta_2-\beta_1^*))^2}_{A1} + \underbrace{\sum_{i\in J_1\cap J_2^*}(\mathbf{x}_i^T(\beta_1-\beta_2^*))^2 + \sum_{i\in J_2\cap J_2^*}(\mathbf{x}_i^T(\beta_2-\beta_2^*))^2}_{A2}.$$

First we consider the first term, A1. Note a simple fact that $\|\beta_1-\beta_1^*\|_2 < \|\beta_2-\beta_1^*\|_2$ or $\|\beta_1-\beta_1^*\|_2 > \|\beta_2-\beta_1^*\|_2$. In the first case, from Lemma 5.1, $\mathbb{E}\left[|J_1\cap J_1^*|\right] \geq \frac{1}{2}p_1|\mathcal{S}_+|$. From Hoeffding's inequality and concentration result (see proof of Lemma 5.1 for similar techniques), for any $\delta \in (0, 1-\frac{2}{\pi})$, there exist constants C', c_1', such that when $N \geq C'k/p_1$, with probability at least $1-e^{-c_1'k}$,

$$\sum_{i\in J_1\cap J_1^*}(\mathbf{x}_i^T(\beta_1-\beta_1^*))^2 \geq \frac{1}{4}p_1|\mathcal{S}_+|(1-\frac{1}{\pi}-\delta)\|\beta_1-\beta_1^*\|_2^2.$$

In the second case, we have a similar result:

$$\sum_{i\in J_2\cap J_1^*}(\mathbf{x}_i^T(\beta_2-\beta_1^*))^2 \geq \frac{1}{4}p_1|\mathcal{S}_+|(1-\frac{1}{\pi}-\delta)\|\beta_2-\beta_1^*\|_2^2.$$

Let $1-\frac{2}{\pi}-\delta = 0.3$ and choose C', c_1' to let the above results also hold for A2. We then conclude that when $N > C'\frac{k}{\min\{p_1,p_2\}}$,

$$\mathcal{L} \geq \frac{0.3}{4}p_1|\mathcal{S}_+|\min\{\|\beta_1-\beta_1^*\|_2^2, \|\beta_2-\beta_1^*\|_2^2\} + \frac{0.3}{4}p_2|\mathcal{S}_+|\min\{\|\beta_1-\beta_2^*\|_2^2, \|\beta_2-\beta_2^*\|_2^2\}. \tag{19}$$

When $\|\beta_1-\beta_1^*\|_2 < \|\beta_2-\beta_1^*\|_2$ and $\|\beta_2-\beta_2^*\|_2 < \|\beta_1-\beta_2^*\|_2$, (19) implies

$$\mathcal{L} \geq \frac{1}{25}\min\{p_1,p_2\}|\mathcal{S}_+|\mathrm{err}^2. \tag{20}$$

When $\|\beta_1-\beta_1^*\|_2 > \|\beta_2-\beta_1^*\|_2$ and $\|\beta_2-\beta_2^*\|_2 < \|\beta_1-\beta_2^*\|_2$, we have

$$\mathcal{L} \geq \frac{1}{25}\min\{p_1,p_2\}|\mathcal{S}_+|(\|\beta_2-\beta_1^*\|_2^2 + \|\beta_2-\beta_2^*\|_2^2) \tag{21}$$

$$\geq \frac{1}{25}\min\{p_1,p_2\}|\mathcal{S}_+|\frac{1}{4}\|\beta_1^*-\beta_2^*\|_2^2. \tag{22}$$

Note that it is impossible for $\|\beta_1-\beta_1^*\|_2 > \|\beta_2-\beta_1^*\|_2$ and $\|\beta_2-\beta_2^*\|_2 > \|\beta_1-\beta_2^*\|_2$ both to be true. Otherwise, we could switch the subscripts of the two β's. Putting (20) and (22) together, we complete the proof. □

Lemma A.3. *Suppose symmetric matrix $\Sigma \in \mathbb{R}^{n\times n}$ has eigenvalues $\lambda_1 \geq \lambda_2 > \lambda_3 \ldots$ with corresponding normalized eigenvectors denoted as $u_1, u_2, u_3, \ldots$. Let M be another symmetric matrix with eigenvalues: $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 > \tilde{\lambda}_3 \ldots$ and eigenvectors $\tilde{u}_1, \tilde{u}_2, \tilde{u}_3, \ldots$. (a) Let span$\{u_1, u_2\}$ denote the hyperplane spanned by u_1 u_2. If $\|M-\Sigma\|_2 \leq \varepsilon$, for $\varepsilon < \frac{\lambda_2-\lambda_3}{2}$ we have*

$$\|\tilde{u}_i - \mathcal{P}_{T(u_1,u_2)}\tilde{u}_i\|_2^2 \leq \frac{4\varepsilon}{\lambda_2-\lambda_3}, i = 1, 2. \tag{23}$$

Moreover, if $\lambda_1 \neq \lambda_2$,

$$\|u_1 - \tilde{u}_1\|_2^2 \leq \frac{4\epsilon}{\lambda_1-\lambda_2} \tag{24}$$

$$\|u_2 - \tilde{u}_2\|_2^2 \leq \frac{4\epsilon}{\lambda_1-\lambda_2} + \frac{8\epsilon}{\lambda_2-\lambda_3} \tag{25}$$

Proof. Suppose $\tilde{u}_1 = \alpha_1 u_1 + \beta_1 u_2 + \gamma_1 w$, $\tilde{u}_2 = \alpha_2 u_1 + \beta_2 u_2 + \gamma_2 v$, where w, v are vector orthogonal to $span\{u_1, u_2\}$. We have $\alpha_1^2 + \beta_1^2 + \gamma_1^2 = \alpha_2^2 + \beta_2^2 + \gamma_2^2 = 1$. Since $\|M - \Sigma\|_2 \le \varepsilon$,

$$\tilde{u}_1^T M \tilde{u}_1 \ge \lambda_1 - \varepsilon \tag{26}$$

$$\begin{aligned} \tilde{u}_1^T M \tilde{u}_1 &\le \tilde{u}_1^T (M - \Sigma)\tilde{u}_1 + \tilde{u}_1^T \Sigma \tilde{u}_1 && (27)\\ &\le \varepsilon + \tilde{u}_1^T \Sigma \tilde{u}_1. && (28) \end{aligned}$$

Combining (26) and (27), using $\tilde{u}_1^T \Sigma \tilde{u}_1 = \alpha_1^2 \lambda_1 + \beta_1^2 \lambda_2 + \gamma_1^2 \lambda_3$, we get

$$\alpha_1^2 \lambda_1 + \beta_1^2 \lambda_2 + \gamma_1^2 \lambda_3 \ge \lambda_1 - 2\varepsilon \tag{29}$$

Since $\alpha_1^2 \lambda_1 + \beta_1^2 \lambda_2 + \gamma_1^2 \lambda_3 \le (1 - \gamma_1^2)\lambda_1 + \gamma_1^2 \lambda_3$, it implies that

$$\gamma_1^2 \le \frac{2\varepsilon}{\lambda_1 - \lambda_3} \le \frac{2\varepsilon}{\lambda_2 - \lambda_3}. \tag{30}$$

We assume $\lambda_1 \neq \lambda_2$. Otherwise , the above inequality also holds for $\tilde{u}_2$, then the proof of (23) is completed. By using another upper bound $\alpha_1^2 \lambda_1 + \beta_1^2 \lambda_2 + \gamma_1^2 \lambda_3 \le \alpha_1^2 \lambda_1 + (1 - \alpha_1^2)\lambda_2$, the following inequality α_1^2 holds

$$\alpha_1^2 \ge 1 - \frac{2\varepsilon}{\lambda_1 - \lambda_2}. \tag{31}$$

Note $\|\tilde{u}_2 - \mathcal{P}_{T(u_1, u_2)} \tilde{u}_2\|_2^2 = \gamma_1^2$, we get the distance bound of u_1. Next, we show the bound for $\tilde{u}_2$. Similar to (29),

$$\alpha_2^2 \lambda_1 + \beta_2^2 \lambda_2 + \gamma_2^2 \lambda_3 \ge \lambda_2 - 2\varepsilon. \tag{32}$$

Again, by using $\alpha_2^2 \lambda_1 + \beta_2^2 \lambda_2 + \gamma_2^2 \lambda_3 \le \alpha_2^2 \lambda_1 + (1 - \alpha_2^2)\lambda_2$, we get

$$\gamma_2^2 \le \frac{2\varepsilon + \alpha_2^2(\lambda_1 - \lambda_2)}{\lambda_2 - \lambda_3}. \tag{33}$$

We use the condition that $\tilde{u}_1$ $\tilde{u}_2$ are orthogonal. Hence, $\alpha_1^2 \alpha_2^2 \le (1 - \alpha_1^2)(1 - \alpha_2^2)$. It is easy to see $\alpha_1^2 + \alpha_2^2 \le 1$. Plugging it into (33) and using (31) result in

$$\gamma_2^2 \le \frac{4\varepsilon}{\lambda_2 - \lambda_3}. \tag{34}$$

Through (30) and (34), we complete the proof of (23).

Using some intermediate results, we derive the bounds for eigenvectors in the case $\lambda_1 \neq \lambda_2$.

$$\begin{aligned} \|u_1 - \tilde{u}_1\|_2^2 &= (1 - \alpha_1)^2 + \beta_1^2 + \gamma_1^2 \\ &= (1 - \alpha_1)^2 + 1 - \alpha_1^2 \\ &\le 2(1 - \alpha_1^2) \\ &\le \frac{4\epsilon}{\lambda_1 - \lambda_2}. \end{aligned}$$

The last inequality follows from (31).

Similarly,

$$\begin{aligned} \|u_2 - \tilde{u}_2\|_2^2 &\le 2(1 - \beta_2^2) \\ &= 2(\alpha_2^2 + \gamma_2^2) \\ &\le 2(1 - \alpha_1^2 + \gamma_2^2) \\ &\le \frac{4\epsilon}{\lambda_1 - \lambda_2} + \frac{8\epsilon}{\lambda_2 - \lambda_3}. \end{aligned}$$

We obtain the last inequality from (31) and (34). □