A. Reduction to Linear System

Here we show that solving linear systems in $\mathbf{B}$ is inherent in solving the top-k generalized eigenvector problem in the worst case and we provide evidence a $\sqrt{\kappa(\mathbf{B})}$ factor in the running time is essential for a broad class of iterative methods for the problem.

Let $\mathbf{M}$ be a symmetric positive definite matrix and suppose we wish to solve the linear system $\mathbf{M}\mathbf{x} = \mathbf{m}$, i.e. compute $\mathbf{x}_*$ with $\mathbf{M}\mathbf{x}_* = \mathbf{m}$. If we set $\mathbf{A} = \mathbf{m}\mathbf{m}^\top$ and $\mathbf{B} = \mathbf{M}$ then

$$\operatorname*{argmax}_{\mathbf{x}^\top\mathbf{B}\mathbf{x}=1} \mathbf{x}^\top\mathbf{A}\mathbf{x} = \frac{\mathbf{B}^{-1}\mathbf{m}}{\mathbf{m}^\top\mathbf{B}^{-1}\mathbf{m}}$$

and consequently computing the top-1 generalized eigenvector yields the solution to the linear system. Therefore, the problem of computing top-k generalized eigenvectors is in general harder than the problem of solving symmetric positive definite linear systems.

Moreover, it is well known that any method which starts at $\mathbf{m}$ and iteratively applies $\mathbf{M}$ to linear combinations of the points computed so far must apply $\mathbf{M}$ at least $\Omega(\sqrt{\kappa(\mathbf{B})})$ in order to halve the error in the standard norm for the problem (Shewchuk, 1994). Consequently, methods that solve the top-1 generalized eigenvector problem by simply applying $\mathbf{A}$ and $\mathbf{B}$, which is the same as applying $\mathbf{M}$ and taking linear combinations with $\mathbf{m}$, must apply $\mathbf{M}$ at least $\Omega(\sqrt{\kappa(\mathbf{M})})$ times to achieve small error, unless they exploit more structure of $\mathbf{M}$ or the initialization.

B. Solving Linear System via Accelerated Gradient Descent

Algorithm 4 Nesterov's accelerated gradient descent

```
Input: learning rate $\eta$, factor $Q$, initial point $\mathbf{x}_0$, $T$.
Output: minimizer $x^\star$ of $f$ .
  for $t = 0, \cdots, T-1$ do
    $\mathbf{y}_{t+1} \leftarrow \mathbf{x}_t - (1/\beta) \cdot \nabla f(\mathbf{x}_t)$
    $\mathbf{x}_{t+1} \leftarrow \mathbf{y}_{t+1} + (\sqrt{Q}-1)/(\sqrt{Q}+1) \cdot (\mathbf{y}_{t+1} - \mathbf{y}_t)$
  end for
  Return $\mathbf{y}_T$.
```

Since we use accelerated gradient descent in our main theorems, for completeness, we put the algorithm and cite its result about iteration complexity here without proof.

Theorem 8 ((Nesterov, 1983)). *Let f be α-strongly convex and β-smooth, then accelerated gradient descent with learning rate $\eta = \frac{1}{\beta}$ and $Q = \beta/\alpha$ satisfies:*

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq 2(f(\mathbf{x}_0) - f(\mathbf{x}^\star)) \exp(-\frac{t}{\sqrt{Q}}) \tag{2}$$

C. Proofs of Main Theorem

In this section we will prove Theorems 5, 6 and 7.

C.1. Rank-1 Setting

We first prove our claim that $\mathbf{B}^{-1}\mathbf{A}$ has an eigenbasis.

Lemma 9. *Let $(\mathbf{u}_i, \sigma_i)$ be the eigenpairs of the symmetric matrix $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$. Then $\mathbf{B}^{-1/2}\mathbf{u}_i$ is an eigenvector of $\mathbf{B}^{-1}\mathbf{A}$ with eigenvalue σ_i.*

Proof. The proof is straightforward.

$$\mathbf{B}^{-1}\mathbf{A}\left(\mathbf{B}^{-1/2}\mathbf{u}_i\right) = \mathbf{B}^{-1/2}\left(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{u}_i\right) = \sigma_i\mathbf{B}^{-1/2}\mathbf{u}_i.$$

□

Denote the eigenpairs of $\mathbf{B}^{-1}\mathbf{A}$ by $(\lambda_i, \mathbf{v}_i)$, the above lemma further tells us that $\mathbf{v}_i^\top \mathbf{B}\mathbf{v}_j = \mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$.

Recall that we defined the angle between $\mathbf{w}$ and $\mathbf{v}_1$ in the $\mathbf{B}$-norm: $\theta\left(\mathbf{w}, \mathbf{v}_1\right) = \arccos\left(|\mathbf{v}_1^\top \mathbf{B}\mathbf{w}|\right)$.

To measure the distance from optimality, we use the following potential function for normalized vector $\mathbf{w}$ ($\|\mathbf{w}\|_{\mathbf{B}} = 1$):

$$\tan\theta(\mathbf{w}, \mathbf{v}_1) = \frac{\sqrt{1 - |\mathbf{v}_1^\top \mathbf{B}\mathbf{w}|^2}}{|\mathbf{v}_1^\top \mathbf{B}\mathbf{w}|}. \tag{3}$$

Lemma 10. *Consider any* $\mathbf{w}$ *such that* $\|\mathbf{w}\|_{\mathbf{B}} = 1$ *and* $\tan\theta(\mathbf{w}, \mathbf{v}_1) \leq \epsilon$. *Then, we have:*

$$\cos^2\theta(\mathbf{w}, \mathbf{v}_1) = (\mathbf{v}_1^\top \mathbf{B}\mathbf{w})^2 \geq 1 - \epsilon^2 \textit{ and } \mathbf{w}^\top \mathbf{A}\mathbf{w} \geq \lambda_1(1 - \epsilon^2).$$

Proof. Clearly,

$$(\mathbf{v}_1^\top \mathbf{B}\mathbf{w})^2 = \cos^2\theta(\mathbf{w}, \mathbf{v}_1) = \frac{1}{1 + \tan^2\theta(\mathbf{w}, \mathbf{v}_1)} \geq \frac{1}{1 + \epsilon^2} \geq 1 - \epsilon^2,$$

proving the first part. For the second part, we have the following:

$$\begin{aligned} \mathbf{w}^\top \mathbf{A}\mathbf{w} &= \sum_{i,j} (\mathbf{v}_i^\top \mathbf{B}\mathbf{w})(\mathbf{v}_j^\top \mathbf{B}\mathbf{w})\mathbf{v}_i^\top \mathbf{A}\mathbf{v}_j = \sum_{i,j} \lambda_j (\mathbf{v}_i^\top \mathbf{B}\mathbf{w})(\mathbf{v}_j^\top \mathbf{B}\mathbf{w})\mathbf{v}_i^\top \mathbf{B}\mathbf{v}_j \\ &= \sum_i \lambda_i (\mathbf{v}_i^\top \mathbf{B}\mathbf{w})^2 \geq \lambda_1 (\mathbf{v}_1^\top \mathbf{B}\mathbf{w})^2 \geq (1 - \epsilon^2)\lambda_1, \end{aligned}$$

proving the lemma. □

Proof of Theorem 5. We will show that the potential function $\tan\theta(\mathbf{w}_t, \mathbf{v}_1)$ decreases geometrically with t. This will directly provides an upper bound for $\sin\theta(\mathbf{w}_t, \mathbf{v}_1)$. For simplicity, through out the proof we will simply denote $\theta(\mathbf{w}_t, \mathbf{v}_i)$ as θ_t.

Recall the updates in Algorithm 1, suppose at time t, we have $\mathbf{w}_t$ such that $\|\mathbf{w}_t\|_{\mathbf{B}} = 1$. Let us say

$$\mathbf{w}_{t+1} = \frac{1}{Z}(\mathbf{B}^{-1}\mathbf{A}\mathbf{w}_t + \xi) \tag{4}$$

where Z is some normalization factor, and ξ is the error in solving the least squares. We will first prove the geometric convergence claim assuming

$$\|\xi\|_{\mathbf{B}} \leq \frac{|\lambda_1| - |\lambda_2|}{4} \min\{\cos\theta_t, \sin\theta_t\}, \tag{5}$$

and then bound the time taken by black-box linear system solver to provide such an accuracy. Since $\mathbf{w}_t$ can be written as $\mathbf{w}_t = \sum_i \left(\mathbf{w}_t^\top \mathbf{B}\mathbf{v}_i\right)\mathbf{v}_i$, we know $\mathbf{B}^{-1}\mathbf{A}\mathbf{w}_t = \sum_{i=1}^d \lambda_i \left(\mathbf{w}_t^\top \mathbf{B}\mathbf{v}_i\right)\mathbf{v}_i$. Since $\|\mathbf{w}_{t+1}\|_{\mathbf{B}} = 1$ and $\mathbf{v}_i^\top \mathbf{B}\mathbf{v}_j = \delta_{ij}$, we have

$$\begin{aligned} \tan\theta_{t+1} &= \frac{\sqrt{Z^2 - |\mathbf{v}_1^\top \mathbf{B} Z\mathbf{w}_{t+1}|^2}}{|\mathbf{v}_1^\top \mathbf{B} Z\mathbf{w}_{t+1}|} \leq \frac{\sqrt{\sum_{i=2}^d \left(\mathbf{w}_t^\top \mathbf{B}\mathbf{v}_i\right)^2 \lambda_i^2} + \|\xi\|_{\mathbf{B}}}{|\left(\mathbf{w}_t^\top \mathbf{B}\mathbf{v}_1\right)\lambda_1| - \|\xi\|_{\mathbf{B}}} \\ &\leq \frac{\sqrt{1 - \left(\mathbf{w}_t^\top \mathbf{B}\mathbf{v}_1\right)^2}}{|\mathbf{w}_t^\top \mathbf{B}\mathbf{v}_1|} \times \frac{|\lambda_2| + \frac{\|\xi\|_{\mathbf{B}}}{\sqrt{1 - \left(\mathbf{w}_t^\top \mathbf{B}\mathbf{v}_1\right)^2}}}{|\lambda_1| - \frac{\|\xi\|_{\mathbf{B}}}{|\mathbf{w}_t^\top \mathbf{B}\mathbf{v}_1|}} = \tan\theta_t \times \frac{|\lambda_2| + \frac{\|\xi\|_{\mathbf{B}}}{\sqrt{1 - \left(\mathbf{w}_t^\top \mathbf{B}\mathbf{v}_1\right)^2}}}{|\lambda_1| - \frac{\|\xi\|_{\mathbf{B}}}{|\mathbf{w}_t^\top \mathbf{B}\mathbf{v}_1|}} \end{aligned}$$

By definition of θ_t, we know $\cos\theta_t = |\mathbf{w}_t^\top \mathbf{B}\mathbf{v}_1|$ and $\sin\theta_t = \sqrt{1 - \left(\mathbf{w}_t^\top \mathbf{B}\mathbf{v}_1\right)^2}$ giving us

$$\tan\theta_{t+1} \leq \tan\theta_t \times \frac{|\lambda_2| + \frac{\|\xi\|_{\mathbf{B}}}{\sin\theta_t}}{|\lambda_1| - \frac{\|\xi\|_{\mathbf{B}}}{\cos\theta_t}}.$$

Since $\|\xi\|_{\mathbf{B}} \leq \frac{|\lambda_1|-|\lambda_2|}{4}\min\{\cos\theta_t, \sin\theta_t\}$, we have that

$$\tan\theta_{t+1} \leq \frac{|\lambda_1|+3|\lambda_2|}{3|\lambda_1|+|\lambda_2|} \times \tan\theta_t.$$

Letting $\gamma = \frac{3|\lambda_1|+|\lambda_2|}{|\lambda_1|+3|\lambda_2|}$, this shows that $G(\mathbf{w}_t) \leq \gamma^t G(\mathbf{w}_0)$. Recalling the definition of eigengap $\rho = 1 - \frac{|\lambda_2|}{|\lambda_1|}$, choosing t to be

$$t \geq \frac{2}{\rho}\log\left(\frac{1}{\epsilon\cos\theta_0}\right) \geq \frac{\log\left(\frac{\tan\theta_0}{\epsilon}\right)}{\left(\frac{1}{\gamma}-1\right)} \geq \frac{\log\left(\frac{\tan\theta_0}{\epsilon}\right)}{\log\left(\frac{1}{\gamma}\right)}, \tag{6}$$

we are guaranteed that $\sin\theta_t \leq \tan\theta_t \leq \epsilon$. This number of iterations $\frac{2}{\rho}\log\left(\frac{1}{\epsilon\cos\theta_0}\right)$ could be further decompose into two phase: 1) initial phase $\frac{2}{\rho}\log\frac{1}{\cos\theta_0}$ which mainly caused by large initial angle, 2) convergence phase $\frac{2}{\rho}\log\frac{1}{\epsilon}$ which is mainly due to the high accuracy ϵ we need.

We now focus on how to obtain the iterate $\mathbf{w}_{t+1}$ using accelerated gradient descent such that the error ξ has norm bounded as in (5).

Let $f(\mathbf{w}) \stackrel{\text{def}}{=} \frac{1}{2}\mathbf{w}^\top\mathbf{B}\mathbf{w} - \mathbf{w}^\top\mathbf{A}\mathbf{w}_t$ and recall that in each iteration, we use linear system solver to solve the following optimization problem:

$$\min_{\mathbf{w}} f(\mathbf{w}). \tag{7}$$

The minimizer of (7) is $\mathbf{B}^{-1}\mathbf{A}\mathbf{w}_t$. Define ϵ_{init} and ϵ_{des} as initial error and required destination error of linear system solver $\|\mathbf{w} - \mathbf{B}^{-1}\mathbf{A}\mathbf{w}_t\|_{\mathbf{B}}^2$. Observe that for any $\mathbf{w}$ we have equality,

$$\|\mathbf{w} - \mathbf{B}^{-1}\mathbf{A}\mathbf{w}_t\|_{\mathbf{B}}^2 = 2(f(\mathbf{w}) - f(\mathbf{B}^{-1}\mathbf{A}\mathbf{w}_t)) \tag{8}$$

Eq.(5) directly poses a condition on ϵ_{des}:

$$\epsilon_{\text{des}} \leq \frac{(|\lambda_1|-|\lambda_2|)^2}{16}\min\{\cos^2\theta_t, \sin^2\theta_t\}$$

Since we initialize Algorithm 4 with $\beta_t\mathbf{w}_t$, where $\beta_t \stackrel{\text{def}}{=} \frac{\mathbf{w}_t^\top\mathbf{A}\mathbf{w}_t}{\mathbf{w}_t^\top\mathbf{B}\mathbf{w}_t}$, the initial error can be bounded as follows:

$$\begin{aligned}
\epsilon_{\text{init}} &= 2(f(\beta_t\mathbf{w}_t) - f(\mathbf{B}^{-1}\mathbf{A}\mathbf{w}_t)) \\
&= 2(\min_{\beta} f(\beta\mathbf{w}_t) - f(\mathbf{B}^{-1}\mathbf{A}\mathbf{w}_t)) \leq 2(f(\lambda_1\mathbf{w}_t) - f(\mathbf{B}^{-1}\mathbf{A}\mathbf{w}_t)) \\
&= \|\lambda_1\mathbf{w}_t - \mathbf{B}^{-1}\mathbf{A}\mathbf{w}_t\|_{\mathbf{B}}^2 \\
&= \sum_{i\geq 2}(\lambda_1-\lambda_i)^2\left(\mathbf{w}_t^\top\mathbf{B}\mathbf{v}_i\right)^2 \leq \lambda_1^2(1-\left(\mathbf{w}_t^\top\mathbf{B}\mathbf{v}_1\right)^2) = \lambda_1^2\sin^2\theta_t.
\end{aligned}$$

This means that we wish to decrease the ratio of final to initial error smaller than

$$\frac{\epsilon_{\text{des}}}{\epsilon_{\text{init}}} \leq \frac{(|\lambda_1|-|\lambda_2|)^2}{16}\min\{\cos^2\theta_t, \sin^2\theta_t\} \times \frac{1}{\lambda_1^2\sin^2\theta_t} = \frac{\rho^2}{16}\min\left\{\frac{1}{\tan^2\theta_t}, 1\right\}. \tag{9}$$

Recall we defined $\mathcal{T}(\delta)$ as the time for linear system solver to reduce the error by a factor δ. Therefore, in the initial phase where θ_t is large, it would be suffice to solve linear system up to factor $\delta = \frac{\rho^2\cos^2\theta_0}{16} \leq \frac{\rho^2}{16\tan\theta_t}$. In convergence phase, where θ_t is small, choose $\delta = \frac{\rho^2}{16}$ would be sufficient.

Therefore, adding the computational cost of Algorithm 1 other than by linear system solver, it's not hard to get the total running time will be bounded by

$$\frac{2}{\rho}\left(\log\frac{1}{\cos\theta_0}\cdot\mathcal{T}(\frac{\rho^2\cos^2\theta_0}{16}) + \log\frac{1}{\epsilon}\cdot\mathcal{T}(\frac{\rho^2}{16})\right) + \frac{2}{\rho}(\text{nnz}\,(\mathbf{A}) + \text{nnz}\,(\mathbf{B}) + d)\log\frac{1}{\epsilon\cos\theta_0}.$$

Furthermore, if we run Nesterov's accelerated gradient descent (Algorithm 4) on function $f(\mathbf{w})$ to solve the linear systems. Since the condition number of the optimization problem (7) is $\kappa(\mathbf{B})$, by Theorem 8, we know $\mathcal{T}(\delta) = O(\mathrm{nnz}\,(\mathbf{B})\sqrt{\kappa(\mathbf{B})}\log\frac{1}{\delta})$. Substituting this gives runtime:

$$O\left(\frac{\mathrm{nnz}\,(\mathbf{B})\sqrt{\kappa(\mathbf{B})}}{\rho}\left(\log\frac{1}{\cos\theta_0}\log\frac{1}{\rho\cos\theta_0}+\log\frac{1}{\epsilon}\log\frac{1}{\rho}\right)+\frac{1}{\rho}\mathrm{nnz}\,(\mathbf{A})\log\frac{1}{\epsilon\cos\theta_0}\right).$$

which finishes the proof. □

C.2. Top-k Setting

To prove the convergence of subspace, we need a notion of angle between subspaces. The standard definition the is principal angles.

Definition 11 (Principal angles)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be subspaces of $\mathbb{R}^d$ of dimension at least k. The principal angles $0 \leq \theta^{(1)} \leq \cdots \leq \theta^{(k)}$ between $\mathcal{X}$ and $\mathcal{Y}$ with respect to $\mathbf{B}$-based scalar product are defined recursively via:*

$$\theta^{(i)}(\mathcal{X},\mathcal{Y}) = \min\{\arccos(\frac{\langle\mathbf{x},\mathbf{y}\rangle_{\mathbf{B}}}{\|\mathbf{x}\|_{\mathbf{B}}\|\mathbf{y}\|_{\mathbf{B}}}) : \mathbf{x}\in\mathcal{X}, \mathbf{y}\in\mathcal{Y}, \mathbf{x}\perp_{\mathbf{B}}\mathbf{x}_j, \mathbf{y}\perp_{\mathbf{B}}\mathbf{y}_j \textit{ for all } j<i\}$$
$$(\mathbf{x}_i,\mathbf{y}_i) \in \mathrm{argmin}\{\arccos(\frac{\langle\mathbf{x},\mathbf{y}\rangle_{\mathbf{B}}}{\|\mathbf{x}\|_{\mathbf{B}}\|\mathbf{y}\|_{\mathbf{B}}}) : \mathbf{x}\in\mathcal{X}, \mathbf{y}\in\mathcal{Y}, \mathbf{x}\perp_{\mathbf{B}}\mathbf{x}_j, \mathbf{y}\perp_{\mathbf{B}}\mathbf{y}_j \textit{ for all } j<i\}$$

For matrices $\mathbf{X}$ and $\mathbf{Y}$, we use $\theta_j(\mathbf{X},\mathbf{Y})$ to denote the j-th principal angle between their range.

Since for our interest, we only care the largest principal angle, thus, in the following proof, without ambiguity, for $\mathbf{X},\mathbf{Y}\in\mathbb{R}^{d\times k}$, we use $\theta(\mathbf{X},\mathbf{Y})$ to indicate $\theta^{(k)}(\mathbf{X},\mathbf{Y})$. Next lemma will tells us this definition of $\theta(\mathbf{X},\mathbf{Y})$ to be the largest principal angle is same as what we defined in the main paper Definition 4.

Lemma 12. *Let $\mathbf{X},\mathbf{Y}\in\mathbb{R}^{d\times k}$ be orthonormal bases (w.r.t $\mathbf{B}$) for subspace $\mathcal{X},\mathcal{Y}$ respectively. Let $\mathbf{X}_\perp$ be an orthonormal basis for orthogonal complement of $\mathcal{X}$ (w.r.t $\mathbf{B}$). Then we have*

$$\cos\theta(\mathcal{X},\mathcal{Y}) = \sigma_k(\mathbf{X}^\top\mathbf{B}\mathbf{Y}), \quad \sin\theta(\mathcal{X},\mathcal{Y}) = \|\mathbf{X}_\perp^\top\mathbf{B}\mathbf{Y}\| \tag{10}$$

and assuming $\mathbf{X}^\top\mathbf{B}\mathbf{Y}$ is invertible ($\theta(\mathcal{X},\mathcal{Y}) < \frac{\pi}{2}$), we have:

$$\tan\theta(\mathcal{X},\mathcal{Y}) = \|\mathbf{X}_\perp^\top\mathbf{B}\mathbf{Y}(\mathbf{X}^\top\mathbf{B}\mathbf{Y})^{-1}\| \tag{11}$$

Proof. By definition of principal angle, it's easy to show $\cos\theta(\mathcal{X},\mathcal{Y}) = \sigma_k(\mathbf{X}^\top\mathbf{B}\mathbf{Y})$. The projection operator onto subspace $\mathcal{X}$ is $\mathbf{X}\mathbf{X}^\top\mathbf{B}$. It's also easy to show $\mathbf{X}\mathbf{X}^\top\mathbf{B}+\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{B}=\mathbf{I}$ Then, we have:

$$\begin{aligned}&(\mathbf{X}_\perp^\top\mathbf{B}\mathbf{Y})^\top\mathbf{X}_\perp^\top\mathbf{B}\mathbf{Y} = \mathbf{Y}^\top\mathbf{B}\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{B}\mathbf{Y}\\ &= \mathbf{Y}^\top\mathbf{B}(\mathbf{I}-\mathbf{X}\mathbf{X}^\top\mathbf{B})\mathbf{Y} = \mathbf{Y}^\top\mathbf{B}\mathbf{Y} - (\mathbf{X}^\top\mathbf{B}\mathbf{Y})^\top(\mathbf{X}^\top\mathbf{B}\mathbf{Y}) = \mathbf{I} - (\mathbf{X}^\top\mathbf{B}\mathbf{Y})^\top(\mathbf{X}^\top\mathbf{B}\mathbf{Y})\end{aligned} \tag{12}$$

Therefore:

$$\|\mathbf{X}_\perp^\top\mathbf{B}\mathbf{Y}\|^2 = 1-\sigma_k^2(\mathbf{X}^\top\mathbf{B}\mathbf{Y}) = 1-\cos^2\theta(\mathcal{X},\mathcal{Y}) = \sin^2\theta(\mathcal{X},\mathcal{Y}) \tag{13}$$

Similarily:

$$\begin{aligned}&[\mathbf{X}_\perp^\top\mathbf{B}\mathbf{Y}(\mathbf{X}^\top\mathbf{B}\mathbf{Y})^{-1}]^\top\mathbf{X}_\perp^\top\mathbf{B}\mathbf{Y}(\mathbf{X}^\top\mathbf{B}\mathbf{Y})^{-1}\\ =&[(\mathbf{X}^\top\mathbf{B}\mathbf{Y})^{-1}]^\top[\mathbf{I}-(\mathbf{X}^\top\mathbf{B}\mathbf{Y})^\top(\mathbf{X}^\top\mathbf{B}\mathbf{Y})](\mathbf{X}^\top\mathbf{B}\mathbf{Y})^{-1}\\ =&[(\mathbf{X}^\top\mathbf{B}\mathbf{Y})^{-1}]^\top(\mathbf{X}^\top\mathbf{B}\mathbf{Y})^{-1}-\mathbf{I}\end{aligned} \tag{14}$$

Therefore:

$$\|\mathbf{X}_\perp^\top\mathbf{B}\mathbf{Y}(\mathbf{X}^\top\mathbf{B}\mathbf{Y})^{-1}\|^2 = \frac{1}{\sigma_k^2(\mathbf{X}^\top\mathbf{B}\mathbf{Y})}-1 = \frac{1}{\cos^2\theta(\mathcal{X},\mathcal{Y})}-1 = \tan^2\theta(\mathcal{X},\mathcal{Y}) \tag{15}$$

Obviously, $\theta(\mathcal{X},\mathcal{Y})$ is acute, thus $\sin\theta(\mathcal{X},\mathcal{Y}) > 0$ and $\tan\theta(\mathcal{X},\mathcal{Y}) > 0$, which finishes the proof. □

Similar to the top one case, for simplicity, we denote $\theta_t \overset{\text{def}}{=} \theta(\mathbf{W}_t, \mathbf{V})$, where $\mathbf{V} \in \mathbb{R}^{d\times k}$ is top k eigen-vector of generalized eigenvalue problem. Now we are ready to prove the theorem. We also denote $\mathbf{V}_\perp \in \mathbb{R}^{d\times(d-k)}$. Also throughout the proof, for any matrix $\mathbf{X}$, we use notation $\|\mathbf{X}\|_\mathbf{B} \equiv \|\mathbf{B}^{\frac{1}{2}}\mathbf{X}\| \equiv \sqrt{\|\mathbf{X}^\top\mathbf{B}\mathbf{X}\|}$ and $\|\mathbf{X}\|_{\mathbf{B},F} = \|\mathbf{B}^{\frac{1}{2}}\mathbf{X}\|_F = \sqrt{\text{tr}(\mathbf{X}^\top\mathbf{B}\mathbf{X})}$.

Proof of Theorem 5. Let $\mathbf{V} \in \mathbb{R}^{d\times k}, \Lambda_\mathbf{V} \in \mathbb{R}^{k\times k}$ be the top k generalized eigen-pairs; and $\mathbf{V}_\perp \in \mathbb{R}^{d\times(d-k)}, \Lambda_{\mathbf{V}_\perp} \in \mathbb{R}^{(d-k)\times(d-k)}$ be the remaining $(d-k)$ generalized eigen-pairs (assume all eigen-vectors normalized w.r.t. $\mathbf{B}$). Then, we have:

$$\begin{aligned}\mathbf{A} &= \mathbf{B}(\mathbf{V}\Lambda_\mathbf{V}\mathbf{V}^\top + \mathbf{V}_\perp\Lambda_{\mathbf{V}_\perp}\mathbf{V}_\perp^\top)\mathbf{B}\\ \mathbf{B} &= \mathbf{B}(\mathbf{V}\mathbf{V}^\top + \mathbf{V}_\perp\mathbf{V}_\perp^\top)\mathbf{B}\end{aligned}$$

By approximately solving $\text{argmin}_{\mathbf{W}\in\mathbb{R}^{d\times k}} \text{tr}(\frac{1}{2}\mathbf{W}^\top\mathbf{B}\mathbf{W} - \mathbf{W}^\top\mathbf{A}\mathbf{W}_t)$ and Gram-Schmidt process, we have:

$$\mathbf{W}_{t+1} = (\mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t + \xi)\mathbf{R} \tag{16}$$

where $\mathbf{R} \in \mathbb{R}^{k\times k}$ is an invertable matrix generated by Gram-Schmidt process.

We will follow the same strategy as in top 1 case, which will first prove the geometric convergence of $\tan\theta_t$ assuming

$$\|\xi\|_\mathbf{B} \le \frac{|\lambda_k| - |\lambda_{k+1}|}{4}\min\{\sin\theta_t, \cos\theta_t\} \tag{17}$$

Note here ξ is a matrix, and $\|\xi\|_\mathbf{B} = \|\mathbf{B}^{\frac{1}{2}}\xi\| = \sqrt{\|\xi^\top\mathbf{B}\xi\|}$. Then we will bound the time taken by black-box linear system solver to provide such an accuracy.

By definition of $\tan\theta_t$ and linear algebra calculation, we have

$$\begin{aligned}\tan\theta_{t+1} &= \|\mathbf{V}_\perp^\top\mathbf{B}\mathbf{W}_{t+1}(\mathbf{V}^\top\mathbf{B}\mathbf{W}_{t+1})^{-1}\|\\ &= \|\mathbf{V}_\perp^\top\mathbf{B}\tilde{\mathbf{W}}_{t+1}(\mathbf{V}^\top\mathbf{B}\tilde{\mathbf{W}}_{t+1})^{-1}\|\\ &= \|(\Lambda_{\mathbf{V}_\perp}\mathbf{V}_\perp^\top\mathbf{B}\mathbf{W}_t + \mathbf{V}_\perp^\top\mathbf{B}\xi)(\Lambda_\mathbf{V}\mathbf{V}^\top\mathbf{B}\mathbf{W}_t + \mathbf{V}^\top\mathbf{B}\xi)^{-1}\|\\ &\le \frac{\|(\Lambda_{\mathbf{V}_\perp}\mathbf{V}_\perp^\top\mathbf{B}\mathbf{W}_t + \mathbf{V}_\perp^\top\mathbf{B}\xi)(\mathbf{V}^\top\mathbf{B}\mathbf{W}_t)^{-1}\|}{\sigma_k(\Lambda_\mathbf{V} + \mathbf{V}^\top\mathbf{B}\xi(\mathbf{V}^\top\mathbf{B}\mathbf{W}_t)^{-1})}\\ &\le \frac{\|\Lambda_{\mathbf{V}_\perp}\|\tan\theta_t + \|\mathbf{V}_\perp^\top\mathbf{B}\xi(\mathbf{V}^\top\mathbf{B}\mathbf{W}_t)^{-1}\|}{\sigma_k(\Lambda_\mathbf{V}) - \|\mathbf{V}^\top\mathbf{B}\xi(\mathbf{V}^\top\mathbf{B}\mathbf{W}_t)^{-1}\|}\\ &\le \frac{\|\Lambda_{\mathbf{V}_\perp}\|\tan\theta_t + \|\mathbf{V}_\perp^\top\mathbf{B}\xi\|\|\mathbf{V}^\top\mathbf{B}\mathbf{W}_t)^{-1}\|}{\sigma_k(\Lambda_\mathbf{V}) - \|\mathbf{V}^\top\mathbf{B}\xi\|\|(\mathbf{V}^\top\mathbf{B}\mathbf{W}_t)^{-1}\|}\\ &= \frac{\|\Lambda_{\mathbf{V}_\perp}\|\tan\theta_t + \frac{\|\mathbf{V}_\perp^\top\mathbf{B}\xi\|}{\cos\theta_t}}{\sigma_k(\Lambda_\mathbf{V}) - \frac{\|\mathbf{V}^\top\mathbf{B}\xi\|}{\cos\theta_t}}\\ &\le \tan\theta_t\frac{|\lambda_{k+1}| + \frac{\|\xi\|_\mathbf{B}}{\sin\theta_t}}{|\lambda_k| - \frac{\|\xi\|_\mathbf{B}}{\cos\theta_t}} \qquad (18)\end{aligned}$$

Since $\|\xi\|_\mathbf{B} \le \frac{|\lambda_k|-|\lambda_{k+1}|}{4}\min\{\sin\theta_t, \cos\theta_t\}$, we have that:

$$\begin{aligned}\tan\theta_{t+1} &\le \frac{|\lambda_k| + 3|\lambda_{k+1}|}{3|\lambda_k| + |\lambda_{k+1}|}\tan\theta_t \qquad (19)\\ &= (1 - \frac{2(|\lambda_k| - |\lambda_{k+1}|)}{3|\lambda_k| + |\lambda_{k+1}|})\tan\theta_t \le \exp(-\frac{|\lambda_k| - |\lambda_{k+1}|}{2|\lambda_k|})\tan\theta_t \qquad (20)\end{aligned}$$

Recall in this problem $\rho = 1 - \frac{|\lambda_{k+1}|}{|\lambda_k|}$, therefore, we know:

$$\sin\theta_t \le \tan\theta_t \le \exp(-\frac{\rho}{2}\cdot t)\tan\theta_0 \le \exp(-\frac{\rho}{2}\cdot t)\frac{1}{\cos\theta_0} \tag{21}$$

If we want $\sin\theta_t \le \epsilon$, which gives iterations:

$$t \ge \frac{2}{\rho}\log\frac{1}{\epsilon\cos\theta_0} \tag{22}$$

Let $f(\mathbf{W}) = \text{tr}(\frac{1}{2}\mathbf{W}^\top\mathbf{B}\mathbf{W} - \mathbf{W}^\top\mathbf{A}\mathbf{W}_t)$. For this problem, we can view $\mathbf{W}$ as a dk dimensional vector, and use linear system to solve this d, k dimensional problem. Therefore, if we represent $\mathbf{W}$ in terms of matrix, the corresponding linear system error is $\|\mathbf{W} - \mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t\|_{\mathbf{B},F}$, recall $\|\mathbf{W}\|_{\mathbf{B},F} = \|\mathbf{B}^{\frac{1}{2}}\mathbf{W}\|_F = \sqrt{\text{tr}(\mathbf{W}^\top\mathbf{B}\mathbf{W})}$. To satisfy the accuracy requirement, we only need

$$\epsilon_{\text{des}} = \|\xi\|^2_{\mathbf{B},F} \le \frac{(|\lambda_k| - |\lambda_{k+1}|)^2}{16}\min\{\sin^2\theta_t, \cos^2\theta_t\} \tag{23}$$

Recall we initialize the linear system solver with $\mathbf{W}_t\Gamma_t$ with $\Gamma_t = (\mathbf{W}_t^\top\mathbf{B}\mathbf{W}_t)^{-1}(\mathbf{W}_t^\top\mathbf{A}\mathbf{W}_t)$, we then have

$$\begin{aligned}
\epsilon_{\text{init}} = &\|\mathbf{W}_t\Gamma_t - \mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t\|^2_{\mathbf{B},F} = \text{tr}[(\mathbf{W}_t\Gamma_t - \mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t)^\top\mathbf{B}(\mathbf{W}_t\Gamma_t - \mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t)]\\
=&2[f(\mathbf{W}_t\Gamma_t) - f(\mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t)] = 2[\operatorname*{argmin}_{\Gamma\in\mathbb{R}^{k\times k}} f(\mathbf{W}_t\Gamma) - f(\mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t)]
\end{aligned} \tag{24}$$

Let $\hat{\Gamma}_t = (\mathbf{V}^\top\mathbf{B}\mathbf{W}_t)^{-1}\Lambda_{\mathbf{V}}(\mathbf{V}^\top\mathbf{B}\mathbf{W}_t)$, and observe $\|\xi\|^2_{\mathbf{B},F} = \|\mathbf{B}^{\frac{1}{2}}\xi\|^2_F = \|\mathbf{V}^\top\mathbf{B}\xi\|^2_F + \|\mathbf{V}_\perp^\top\mathbf{B}\xi\|^2_F$ (Pythagorean theorem under $\mathbf{B}$ norm), then we have:

$$\begin{aligned}
\epsilon_{\text{init}} =&\|\mathbf{W}_t\Gamma_t - \mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t\|^2_{\mathbf{B},F} = 2[\operatorname*{argmin}_{\Gamma\in\mathbb{R}^{k\times k}} f(\mathbf{W}_t\Gamma) - f(\mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t)]\\
\le&2[f(\mathbf{W}_t\hat{\Gamma}_t) - f(\mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t)] = \|\mathbf{W}_t\hat{\Gamma}_t - \mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t\|^2_{\mathbf{B},F}\\
=&\|\mathbf{V}^\top\mathbf{B}(\mathbf{W}_t\hat{\Gamma}_t - \mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t)\|^2_F + \|\mathbf{V}_\perp^\top\mathbf{B}(\mathbf{W}_t\hat{\Gamma}_t - \mathbf{B}^{-1}\mathbf{A}\mathbf{W}_t)\|^2_F\\
=&\|\mathbf{V}^\top\mathbf{B}\mathbf{W}_t\hat{\Gamma}_t - \Lambda_{\mathbf{V}}\mathbf{V}^\top\mathbf{B}\mathbf{W}_t\|^2_F + \|\mathbf{V}_\perp^\top\mathbf{B}\mathbf{W}_t\hat{\Gamma}_t - \Lambda_{\mathbf{V}_\perp}\mathbf{V}_\perp^\top\mathbf{B}\mathbf{W}_t\|^2_F\\
=&0 + \|\mathbf{V}_\perp^\top\mathbf{B}\mathbf{W}_t\hat{\Gamma}_t - \Lambda_{\mathbf{V}_\perp}\mathbf{V}_\perp^\top\mathbf{B}\mathbf{W}_t\|^2_F\\
\le&k\|\mathbf{V}_\perp^\top\mathbf{B}\mathbf{W}_t\hat{\Gamma}_t - \Lambda_{\mathbf{V}_\perp}\mathbf{V}_\perp^\top\mathbf{B}\mathbf{W}_t\|^2\\
\le&2k\sin^2\theta_t(\|\hat{\Gamma}_t\|^2 + \|\Lambda_{\mathbf{V}_\perp}\|^2) \le 4k|\lambda_1|^2\tan^2\theta_t
\end{aligned} \tag{25}$$

The last step is correct since $\|\Lambda_{\mathbf{V}_\perp}\| \le |\lambda_1|$ and $\|\hat{\Gamma}_t\| \le \|(\mathbf{V}^\top\mathbf{B}\mathbf{W}_t)^{-1}\|\|\Lambda_{\mathbf{V}}\|\|\mathbf{V}^\top\mathbf{B}^{\frac{1}{2}}\|\|\mathbf{B}^{\frac{1}{2}}\mathbf{W}_t\| \le \frac{1}{\cos\theta_t}|\lambda_1|$

This means we wish to decrease the ratio of final to initial error smaller than:

$$\frac{\epsilon_{\text{des}}}{\epsilon_{\text{init}}} \le \frac{\rho^2}{64k\gamma^2}\min\{\frac{1}{\cos^2\theta_t}, \frac{\sin^2\theta_t}{\cos^4\theta_t}\} \tag{26}$$

where $\gamma = \frac{|\lambda_1|}{|\lambda_k|}$. Therefore, a two phase analysis of running time depending on θ_t is large or small similar to top 1 case would gives the total runtime:

$$\frac{2}{\rho}\left(\log\frac{1}{\cos\theta_0}\cdot\mathcal{T}(\frac{\rho^2\cos^4\theta_0}{64k\gamma^2}) + \log\frac{1}{\epsilon}\cdot\mathcal{T}(\frac{\rho^2}{64k\gamma^2})\right) + \frac{2}{\rho}\left(\text{nnz}\left(\mathbf{A}\right)k + \text{nnz}\left(\mathbf{B}\right)k + dk^2\right)\log\frac{1}{\epsilon\cos\theta_0},$$

if we are using the accelerated gradient descent to solve the linear system, we are essentially solve k disjoint optimization problem, with each problem dimension d and condition number $\kappa(\mathbf{B})$. Directly apply Theorem 8 gives runtime

$$O\left(\frac{\text{nnz}\left(\mathbf{B}\right)k\sqrt{\kappa(\mathbf{B})}}{\rho}\left(\log\frac{1}{\cos\theta_0}\log\frac{k\gamma}{\rho\cos\theta_0} + \log\frac{1}{\epsilon}\log\frac{k\gamma}{\rho}\right) + \frac{\left(\text{nnz}\left(\mathbf{A}\right)k + dk^2\right)}{\rho}\log\frac{1}{\epsilon\cos\theta_0}\right).$$

□

Finally, since both results Theorem 5 and Theorem 7 are stated in terms of initialization θ_0, here we will give probablistic guarantee for random initialization.

Lemma 13 (Random Initialization). *Let top k eigen-vector be $\mathbf{V} \in \mathbb{R}^{d\times k}$, and the remaining eigen-vector be $\mathbf{V}_\perp \in \mathbb{R}^{d\times(d-k)}$. If we initialize $\mathbf{W}_0$ as in Algorithm 2, then With at least probability $1-\eta$, we have:*

$$\tan\theta_0 = \|\mathbf{V}_\perp^\top\mathbf{B}\mathbf{W}_0(\mathbf{V}^\top\mathbf{B}\mathbf{W}_0)^{-1}\| \le O(\frac{\sqrt{\kappa(\mathbf{B})dk}}{\eta}) \tag{27}$$

Proof. Recall $\tilde{\mathbf{W}}$ is entry-wise sampled from standard Gaussian, and

$$\begin{aligned}\tan\theta_0 =& \|\mathbf{V}_\perp^\top\mathbf{B}\mathbf{W}_0(\mathbf{V}^\top\mathbf{B}\mathbf{W}_0)^{-1}\| = \|\mathbf{V}_\perp^\top\mathbf{B}\tilde{\mathbf{W}}_0(\mathbf{V}^\top\mathbf{B}\tilde{\mathbf{W}}_0)^{-1}\| \le \frac{\|\mathbf{V}_\perp^\top\mathbf{B}\tilde{\mathbf{W}}_0\|}{\sigma_k(\mathbf{V}^\top\mathbf{B}\tilde{\mathbf{W}}_0)}\\ \le& \frac{\|\mathbf{V}_\perp^\top\mathbf{B}\tilde{\mathbf{V}_\perp}\|}{\sigma_k(\mathbf{V}^\top\mathbf{B}\tilde{\mathbf{V}})}\frac{\|\tilde{\mathbf{V}_\perp}^\top\tilde{\mathbf{W}}_0\|}{\sigma_k(\tilde{\mathbf{V}}^\top\tilde{\mathbf{W}}_0)}\end{aligned} \tag{28}$$

Where $\tilde{\mathbf{V}_\perp}, \tilde{\mathbf{V}}$ are the right singular vectors of $\mathbf{V}_\perp^\top\mathbf{B}, \mathbf{V}^\top\mathbf{B}$ respectively. Then, we have first term:

$$\frac{\|\mathbf{V}_\perp^\top\mathbf{B}\tilde{\mathbf{V}_\perp}\|}{\sigma_k(\mathbf{V}^\top\mathbf{B}\tilde{\mathbf{V}})} = \frac{\|\mathbf{V}_\perp^\top\mathbf{B}\|}{\sigma_k(\mathbf{V}^\top\mathbf{B})} \le \frac{\|\mathbf{V}_\perp^\top\mathbf{B}^{\frac{1}{2}}\|\|\mathbf{B}^{\frac{1}{2}}\|}{\sigma_k(\mathbf{V}^\top\mathbf{B}^{\frac{1}{2}})\sigma_{\min}(\mathbf{B}^{\frac{1}{2}})} = \kappa(\mathbf{B})^{\frac{1}{2}} \tag{29}$$

The last step is true since both $\mathbf{V}_\perp^\top\mathbf{B}^{\frac{1}{2}}$ and $\mathbf{V}^\top\mathbf{B}^{\frac{1}{2}}$ are orthonormal matrix.

For the second term, we know $\|\tilde{\mathbf{V}_\perp}^\top\tilde{\mathbf{W}}_0\| \sim O(\sqrt{d}+\sqrt{k})$ with high probability, and by equation 3.2 in (Rudelson & Vershynin, 2010) we know $\sigma_k(\tilde{\mathbf{V}}^\top\tilde{\mathbf{W}}_0) \ge \frac{\eta}{\sqrt{k}}$ with probability at least $1-\eta$, which finishes the proof. □

C.3. CCA Setting

Since our approach to CCA directly calls Algorithm 2 for solving generalized eigenvalue problem as subroutine, most of the theoretical property should be clear other than random projection step in Algorithm 3. Here, we give following lemma. The proof of Theorem 7 easily follow from the combination of this lemma and Theorem 6.

Lemma 14. *If the $\begin{pmatrix}\bar{\mathbf{W}}_x\\ \bar{\mathbf{W}}_y\end{pmatrix}$ as constructed in Algorithm 3 has angle at most θ with the true top-$2k$ generalized eigenspace of $\mathbf{A}, \mathbf{B}$, then with probability $1-\zeta$, both $\mathbf{W}_x$, $\mathbf{W}_y$ has angle at most $O(k^2\theta/\zeta^2)$ with the true top-k canonical space of* $\mathbf{X}, \mathbf{Y}$.

Proof. We will prove this for $\mathbf{W}_y$, the proof for $\mathbf{W}_x$ follows directly from same strategy.

Recall $\mathbf{B} = \begin{pmatrix}\mathbf{S}_{xx} & 0\\ 0 & \mathbf{S}_{yy}\end{pmatrix}$. Let $\Phi \in \mathbb{R}^{d_1\times k}$ be the true top k subspace of $\mathbf{X}$ and $\Psi \in \mathbb{R}^{d_2\times k}$ be the true top k subspace of $\mathbf{Y}$.Then by construction we know the top $2k$ subspace should be $\frac{1}{\sqrt{2}}\begin{pmatrix}\Phi & -\Phi\\ \Psi & \Psi\end{pmatrix}$.

By properties of principal angle, we know there exists an orthonormal matrix $\mathbf{R} \in \mathbb{R}^{2k\times 2k}$ such that

$$\|\frac{1}{\sqrt{2}}\mathbf{B}^{1/2}\begin{pmatrix}\Phi & -\Phi\\ \Psi & \Psi\end{pmatrix}\mathbf{R} - \mathbf{B}^{1/2}\begin{pmatrix}\bar{\mathbf{W}}_x\\ \bar{\mathbf{W}}_y\end{pmatrix}\| \le 2\sin\frac{\theta}{2}.$$

In particular, if we only look at the last d_2 rows, we have

$$\|\frac{1}{\sqrt{2}}\mathbf{S}_{yy}^{1/2}\begin{pmatrix}\Psi & \Psi\end{pmatrix}\mathbf{R} - \mathbf{S}_{yy}^{1/2}\bar{\mathbf{W}}_y\| \le 2\sin\frac{\theta}{2}.$$

Let $\mathbf{U}$ be the random Gaussian projection we used, and let $\mathbf{R}\mathbf{U} = \begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}$, we know

$$\begin{aligned}\mathbf{S}_{yy}^{1/2}\bar{\mathbf{W}}_y\mathbf{U} &= \frac{1}{\sqrt{2}}\mathbf{S}_{yy}^{1/2}\begin{pmatrix}\Psi & \Psi\end{pmatrix}\begin{pmatrix}\mathbf{U}_1\\ \mathbf{U}_2\end{pmatrix}+\mathbf{E}\\ &= \frac{1}{\sqrt{2}}\mathbf{S}_{yy}^{1/2}\Psi(\mathbf{U}_1+\mathbf{U}_2)+\mathbf{E},\end{aligned}$$

where $\mathbf{E}$ is the error (after multipled by random matrix $\mathbf{U}$), with $\|\mathbf{E}\| \le O(2\sqrt{k}\sin\frac{\theta}{2}) \le O(\sqrt{k}\theta)$.

Let $\mathbf{V} = (\mathbf{S}_{yy}^{1/2}\bar{\mathbf{W}}_y\mathbf{U})^\top \mathbf{S}_{yy}^{1/2}\bar{\mathbf{W}}_y\mathbf{U}$, the orthonormalization step gives a matrix $\mathbf{W}_y$ that is equivalent (up to rotation) to $\bar{\mathbf{W}}_y\mathbf{U}\mathbf{V}^{-1/2}$. Our goal is to show $\mathbf{V}^{-1/2} \approx ((\mathbf{U}_1+\mathbf{U}_2)^\top(\mathbf{U}_1+\mathbf{U}_2))^{-1/2}$ so we get roughly Ψ.

Note that $\Psi^\top\mathbf{S}_{yy}\Psi = \mathbf{I}$, therefore $\mathbf{V} = \frac{1}{2}(\mathbf{U}_1+\mathbf{U}_2)^\top(\mathbf{U}_1+\mathbf{U}_2)+\mathbf{E}'$ where the error $\mathbf{E}' = (\frac{1}{\sqrt{2}}\mathbf{S}_{yy}^{1/2}\Psi(\mathbf{U}_1+\mathbf{U}_2))^\top\mathbf{E} + \frac{1}{\sqrt{2}}\mathbf{E}^\top(\mathbf{S}_{yy}^{1/2}\Psi(\mathbf{U}_1+\mathbf{U}_2)) + \mathbf{E}^\top\mathbf{E})$. We know with high probability $\|\mathbf{U}_1+\mathbf{U}_2\| \le O(\sqrt{k})$, with probability at least $1-\zeta$, $\sigma_{min}(\mathbf{U}_1+\mathbf{U}_2) \ge \Omega(\zeta/\sqrt{k})$. Therefore we know $\sigma_{min}[(\mathbf{U}_1+\mathbf{U}_2)^\top(\mathbf{U}_1+\mathbf{U}_2)] \ge \Omega(\zeta^2/k)$ and $\|\mathbf{E}'\| \le O(k\theta)$. By matrix perturbation for inverse we know $\|\mathbf{V}^{-1/2} - \sqrt{2}((\mathbf{U}_1+\mathbf{U}_2)^\top(\mathbf{U}_1+\mathbf{U}_2))^{-1/2}\| \le O(k^2\theta/\zeta^2)$. Since $(\mathbf{U}_1+\mathbf{U}_2)((\mathbf{U}_1+\mathbf{U}_2)^\top(\mathbf{U}_1+\mathbf{U}_2))^{-1/2} = \mathbf{R}'$ is an orthonormal matrix, we know there's some orthonormal matrix $\mathbf{R}''$ so that:

$$\begin{aligned}&\|\mathbf{S}_{yy}^{1/2}\mathbf{W}_y - \mathbf{S}_{yy}^{1/2}\Psi\mathbf{R}''\| = \|\mathbf{S}_{yy}^{1/2}\bar{\mathbf{W}}_y\mathbf{U}\mathbf{V}^{-1/2} - \mathbf{S}_{yy}^{1/2}\Psi\mathbf{R}'\|\\ \le&\|\mathbf{S}_{yy}^{1/2}\bar{\mathbf{W}}_y\mathbf{U}\mathbf{V}^{-1/2} - \sqrt{2}\mathbf{S}_{yy}^{1/2}\bar{\mathbf{W}}_y\mathbf{U}((\mathbf{U}_1+\mathbf{U}_2)^\top(\mathbf{U}_1+\mathbf{U}_2))^{-1/2}\|\\ &+\|\sqrt{2}\mathbf{S}_{yy}^{1/2}\bar{\mathbf{W}}_y\mathbf{U}((\mathbf{U}_1+\mathbf{U}_2)^\top(\mathbf{U}_1+\mathbf{U}_2))^{-1/2} - \mathbf{S}_{yy}^{1/2}\Psi\mathbf{R}'\| \le O(k^2\theta/\zeta^2)\end{aligned}$$

Therefore the angle between the $\mathbf{W}_y$ and the truth Ψ is bounded by $O(k^2\theta/\zeta^2)$.

□