A. Supplementary Materials

In this section we provide the proofs of the results stated in the main paper. Throughout the section, we denote the objective function of ℓ_2 and ℓ_1 regularized problems with $J_\lambda(\mathrm{M})$ and $J_\mu(\mathrm{M})$ respectively, where

$$J_\mu(\mathrm{M}) := -\mathbb{E}[\mathrm{x}^\top \mathrm{M}\mathrm{x}] + \mu \mathrm{Tr}\,(\mathrm{M})\,,$$

$$J_\lambda(\mathrm{M}) := -\mathbb{E}[\mathrm{x}^\top \mathrm{M}\mathrm{x}] + \frac{\lambda}{2}\|\mathrm{M}\|_F^2.$$

Recall that the eigendecomposition of C is given by $\mathrm{C} = \sum_{i=1}^d \lambda_i \mathrm{u}_i \mathrm{u}_i^\top$.

A.1. Proofs of Section 2

Proof of Lemma 2.2. Note that the feasible set of Problem 6 is convex, and the objective J_λ is λ-strongly convex. Hence, the optimum is unique. Since Slater condition is satisfied, strong duality holds and KKT conditions are necessary and sufficient for optimality. Let $(\lambda_i(\mathrm{M}), \mathrm{v}_i),\ i \in [d]$ denote the eigenvalues and associated eigenvectors of M. KKT first-order optimality condition yields

$$0 = -\mathrm{C} + \lambda \mathrm{M} - \sum_{i=1}^d \gamma_i \mathrm{v}_i \mathrm{v}_i^\top + \sum_{i=1}^d \omega_i \mathrm{v}_i \mathrm{v}_i^\top + \beta \mathrm{I},$$

where $\gamma_i, \omega_i, \beta \geq 0$ are Lagrange multipliers for the constraints $\mathrm{M} \succeq 0$, $\mathrm{M} \preceq \mathrm{I}$, and $\mathrm{Tr}\,(\mathrm{M}) \leq k$ respectively. Note that except for C, every other term in the above equation has the same set of eigenvectors as M. We conclude that C and M have the same set of eigenvectors, i.e. $\mathrm{u}_i = \mathrm{v}_i,\ i \in [d]$. Rearranging, we get:

$$\mathrm{C} = \sum_{i=1}^d (\lambda\lambda_i(\mathrm{M}) - \gamma_i + \omega_i + \beta)\mathrm{u}_i \mathrm{u}_i^\top$$

Complementary slackness implies that $\gamma_i \lambda_i(\mathrm{M})$, $\omega_i(\lambda_i(\mathrm{M}) - 1)$, and $\beta(\mathrm{Tr}\,(\mathrm{M}) - k)$ all vanish. Now, for an admissible λ, we can verify that the following satisfy the KKT conditions:

$$\begin{aligned}
\mathrm{M} &= \Pi_k(\mathrm{C})\\
\omega_i &= (\lambda_i - \lambda_k)1_{i \leq k}\\
\gamma_i &= (\lambda_k - \lambda - \lambda_j)1_{i > k}\\
\beta &= \lambda_k - \lambda,
\end{aligned}$$

where $\Pi_k(\mathrm{C})$ returns the projection matrix corresponding to top-k eigenspace of C and $1_{i \in \mathscr{A}}$ is the indicator of set $\mathscr{A}$. In particular, $\gamma_i = \lambda_i - \lambda_k \geq 0$ for $i \in [k]$. Also, since $\lambda < g_k$, for $i \in \{k+1, \ldots, d\}$ it holds that $\gamma_i = \lambda_k - \lambda - \lambda_j \geq 0$. It is easy to verify that the complementary slackness conditions are also satisfied for the above assignments of the primal and dual variables. □

Proof of Lemma 2.3. As discussed in the proof of Lemma 2.2, since the feasible set is convex and the objective is strongly convex, the optimum is unique. Furthermore, Salter condition is satisfied. Therefore, strong duality holds and KKT conditions are necessary and sufficient for optimality. Following Lemma 2.2, KKT first-order optimality condition yields:

$$\mathrm{C} = \sum_{i=1}^d (\lambda\lambda_i(\mathrm{M}) - \gamma_i + \omega_i + \beta)\mathrm{u}_i \mathrm{u}_i^\top$$

One can assert that the following set of primal and dual variables satisfy the KKT conditions:

$$\begin{aligned}
\mathrm{M} &= \sum_{i=1}^p \mathrm{u}_i \mathrm{u}_i^\top + \frac{k-p}{q-p}\sum_{j=p+1}^q \mathrm{u}_j \mathrm{u}_j^\top\\
\omega_i &= (\lambda_i - \lambda_{p+1} + \frac{k-q}{q-p}\lambda)1_{i \leq p}\\
\gamma_i &= (\lambda_{p+1} - \lambda_i - \frac{k-p}{q-p}\lambda)1_{i > q}\\
\beta &= \lambda_{p+1} - \frac{k-p}{q-p}\lambda,
\end{aligned}$$

where $1_{i \in \mathscr{A}}$ is the indicator of set $\mathscr{A}$. □

Proof of Theorem 2.4. Let's denote the stochastic gradient at time t by $\widehat{\mathrm{g}}_t := -\mathrm{x}_t \mathrm{x}_t^\top + \lambda \mathrm{M}_t$. Note that

$$\|\widehat{\mathrm{g}}_t\|_F = \|-\mathrm{x}_t \mathrm{x}_t^\top + \lambda \mathrm{M}_t\|_F \leq \|\mathrm{x}_t\|^2 + \lambda\|\mathrm{M}_t\|_F \leq 1 + \lambda\sqrt{k}.$$

Hence, $G^2 := (1 + \lambda\sqrt{k})^2 \geq \mathbb{E}[\|\widehat{\mathrm{g}}_t\|^2]$ for all $t \in \{1, \ldots, T\}$. Let M^* be the global optimum of Problem 6. Since λ is admissible, it follows from Lemmas 2.2 and 2.3 that M^* is also an optimum for Problem 4. Using the following Lemma from (Rakhlin et al., 2012), we bound the distance between M_T and M^*.

Lemma A.1 (Lemma 1 of (Rakhlin et al., 2012))**.** Suppose J is λ-strongly convex over a convex set $\mathscr{M}$, and that $\mathbb{E}[\|\widehat{\mathrm{g}}_t\|_F^2] \leq G^2$. Then if we pick $\eta_t = \frac{1}{\lambda t}$, it holds for any T that

$$\mathbb{E}[\|\mathrm{M}_T - \mathrm{M}^*\|_F^2] \leq \frac{4G^2}{\lambda^2 T}. \tag{14}$$

Now, we note that

$$\begin{aligned}
\|\mathrm{M}^* - \tilde{\mathrm{M}}\|_F &= \|\mathrm{M}^* - \mathrm{M}_T + \mathrm{M}_T - \tilde{\mathrm{M}}\|_F\\
&\leq \|\mathrm{M}^* - \mathrm{M}_T\|_F + \|\mathrm{M}_T - \tilde{\mathrm{M}}\|_F\\
&\leq 2\|\mathrm{M}^* - \mathrm{M}_T\|_F
\end{aligned}$$

where the second inequality holds because by definition of $\tilde{\mathrm{M}}$, $\|\mathrm{M}_T - \tilde{\mathrm{M}}\|_F \leq \|\mathrm{M}^* - \mathrm{M}_T\|_F$. Plugging back the above into Equation (14) completes the proof. □

Proof of Theorem 2.5. Let $\mathrm{M}_* = \sum_{i=1}^k \mathrm{u}_i\mathrm{u}_i^\top$ and $\tilde{\mathrm{M}} = \sum_{i=1}^k \tilde{\mathrm{u}}_i\tilde{\mathrm{u}}_i^\top$ be the eigendecompositions of M_* and $\tilde{\mathrm{M}}$ respectively. Denote the suboptimality gap by $\epsilon := \mathbb{E}[\mathrm{x}^\top\mathrm{M}_*\mathrm{x} - \mathrm{x}^\top\tilde{\mathrm{M}}\mathrm{x}]$, we have

$$\begin{aligned}\epsilon &= \sum_{i=1}^k \mathrm{u}_i^\top\mathbb{E}[\mathrm{x}\mathrm{x}^\top]\mathrm{u}_i - \mathbb{E}[\sum_{i=1}^k \tilde{\mathrm{u}}_i^\top\mathbb{E}[\mathrm{x}\mathrm{x}^\top]\tilde{\mathrm{u}}_i] \\ &= \sum_{i=1}^k \lambda_i - \mathbb{E}[\sum_{j=1}^k \tilde{\mathrm{u}}_j^\top \sum_{i=1}^d \lambda_i\mathrm{u}_i\mathrm{u}_i^\top\tilde{\mathrm{u}}_j] \\ &= \sum_{i=1}^k \lambda_i - \sum_{j=1}^k\sum_{i=1}^d \lambda_i\mathbb{E}[(\mathrm{u}_i^\top\tilde{\mathrm{u}}_j)^2]\end{aligned}$$

Noting that $\mathbb{E}[(\mathrm{u}_i^\top\tilde{\mathrm{u}}_j)^2] \geq 0$, we get

$$\begin{aligned}\epsilon &\leq \sum_{i=1}^k \lambda_i - \sum_{j=1}^k\sum_{i=1}^k \lambda_i\mathbb{E}[(\mathrm{u}_i^\top\tilde{\mathrm{u}}_j)^2] \\ &= \sum_{i=1}^k \lambda_i(1 - \sum_{j=1}^k \mathbb{E}[(\mathrm{u}_i^\top\tilde{\mathrm{u}}_j)^2])\end{aligned}$$

Since $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ and $1 - \sum_{j=1}^k \mathbb{E}[(\mathrm{u}_i^\top\tilde{\mathrm{u}}_j)^2] \geq 0$, we get that

$$\epsilon \leq \lambda_1 \sum_{i=1}^k (1 - \mathbb{E}[\sum_{j=1}^k (\mathrm{u}_i^\top\tilde{\mathrm{u}}_j)^2]) = \lambda_1(k - \mathbb{E}[\|\mathrm{U}^\top\tilde{\mathrm{U}}\|_F^2]) \tag{15}$$

Now, we show that $k - \mathbb{E}[\|\tilde{\mathrm{U}}^\top\mathrm{U}_*\|_F^2] = \frac{1}{2}\mathbb{E}[\|\mathrm{M}_* - \tilde{\mathrm{M}}\|_F^2]$.

$$\begin{aligned}\|\mathrm{M}_* - \tilde{\mathrm{M}}\|_F^2 &= \|\mathrm{M}_*\|_F^2 + \|\tilde{\mathrm{M}}\|_F^2 - 2\langle\mathrm{M}_*, \tilde{\mathrm{M}}\rangle \\ &= 2k - 2\langle\mathrm{U}\mathrm{U}^\top, \tilde{\mathrm{U}}\tilde{\mathrm{U}}^\top\rangle \\ &= 2\left(k - \|\tilde{\mathrm{U}}^\top\mathrm{U}_*\|_F^2\right).\end{aligned}$$

Plugging back this result in Equation (15), we get

$$\epsilon \leq \frac{\lambda_1}{2}\mathbb{E}[\|\mathrm{M}_* - \tilde{\mathrm{M}}\|_F^2] \leq \frac{8\lambda_1(1+\lambda\sqrt{k})^2}{\lambda^2 T}$$

where the last inequality follows from Equation (8). □

A.2. Proofs of Section 3

Proof of Lemma 3.1. Since the problem is convex and Slater condition is satisfied, the KKT conditions are necessary and sufficient for optimality. Let $(\lambda_i(\mathrm{M}), \mathrm{v}_i),\ i \in [d]$ denote the eigenvalues and associated eigenvectors of M. KKT first-order optimality condition yields

$$0 = -\mathrm{C} + \mu\mathrm{I} - \sum_{i=1}^d \gamma_i\mathrm{v}_i\mathrm{v}_i^\top + \sum_{i=1}^d \omega_i\mathrm{v}_i\mathrm{v}_i^\top + \beta\mathrm{I},$$

where $\gamma_i, \omega_i, \beta \geq 0$ are Lagrange multipliers for the constraints $\mathrm{M} \succeq 0$, $\mathrm{M} \preceq \mathrm{I}$, and $\mathrm{Tr}\,(\mathrm{M}) \leq k$ respectively. Note that except for C, every other term in the above equation has the same set of eigenvectors as M. We conclude that C and M have the same set of eigenvectors, i.e. $\mathrm{u}_i = \mathrm{v}_i,\ i \in [d]$. Rearranging, we get:

$$\mathrm{C} = \sum_{i=1}^d (\mu - \gamma_i + \omega_i + \beta)\mathrm{u}_i\mathrm{u}_i^\top. \tag{16}$$

Complementary slackness implies that $\gamma_i\lambda_i(\mathrm{M})$, $\omega_i(\lambda_i(\mathrm{M}) - 1)$, and $\beta(\mathrm{Tr}\,(\mathrm{M}) - k)$ all vanish. Now, for an admissible μ, we can assert that the following satisfy the KKT conditions:

$$\begin{aligned}\mathrm{M} &= \Pi_k(\mathrm{C}) \\ \omega_i &= (\lambda_i - \lambda_k)1_{i\leq k} \\ \gamma_i &= (\lambda_k - \lambda_i)1_{i>k} \\ \beta &= \lambda_k - \mu,\end{aligned}$$

where $\Pi_k(\mathrm{C})$ returns the projection matrix corresponding to top-k eigenspace of C and $1_{i\in\mathscr{A}}$ is the indicator of set $\mathscr{A}$. We shall now prove that M is the unique global optimum. Assume $\mathrm{M}_0 \neq \mathrm{M}$ achieves the optimum as well. Observe that all extreme points of the feasible set are rank-k projection matrices. Since the objective is linear, it must be the case that M_0 is a convex combination of some extreme points of the feasible set, i.e. $\mathrm{M}_0 = \sum_{i=1}^r \alpha_i\mathrm{M}_i$ where $\mathrm{M}_1, \ldots, \mathrm{M}_r$ are all rank-k projection matrices and $\sum_{i=1}^r \alpha_i = 1$ and $\alpha_i \geq 0, \forall i \in 1, \ldots, r$. Hence, $\mathrm{Tr}\,(\mathrm{M}_0) = \sum_{i=1}^r \alpha_i\mathrm{Tr}\,(\mathrm{M}_i) = k$. Since $-\mathbb{E}_\mathrm{x}[\mathrm{x}^\top\mathrm{M}\mathrm{x}] + \mu\mathrm{Tr}\,(\mathrm{M}) = -\mathbb{E}_\mathrm{x}[\mathrm{x}^\top\mathrm{M}_0\mathrm{x}] + \mu\mathrm{Tr}\,(\mathrm{M}_0)$ and $\mathrm{Tr}\,(\mathrm{M}) = \mathrm{Tr}\,(\mathrm{M}_0) = k$, it should be the case that $-\mathbb{E}_\mathrm{x}[\mathrm{x}^\top\mathrm{M}\mathrm{x}] = -\mathbb{E}_\mathrm{x}[\mathrm{x}^\top\mathrm{M}_0\mathrm{x}]$, which contradicts with the eigengap assumption ($g_k > 0$). □

Key insights on the admissibility condition ($\mu \leq \lambda_k$): The stationary condition in Equation (16) gives a system of linear equations with d equations:

$$\lambda_i = \mu - \gamma_i + \omega_i + \beta,\ \forall i \in [d]. \tag{17}$$

If $\mu > \lambda_k$, since $\beta \geq 0$, it should hold for all $j \in \{k, \ldots, d\}$ that $-\gamma_j + \omega_j < 0$. Since $\omega_j \geq 0$, we should have $\gamma_j > 0$. By complementary slackness, we conclude that $\lambda_j(\mathrm{M}^*) = 0$. This implies that $\mathrm{Tr}(\mathrm{M}^*) < k$, because at least $d - k + 1$ eigenvalues of M^* are equal to zero, and the rest are at most one. In this case, M^* cannot be a solution of Problem 4. Hence, the admissibility condition in the statement of Lemma 3.1.

Proof of Theorem 3.2. Lets denote the stochastic gradient at time t by $\widehat{\mathrm{g}}_t := -\mathrm{x}_t\mathrm{x}_t^\top + \mu\mathrm{I}$. Note that

$$\|\widehat{\mathrm{g}}_t\|_F = \|-\mathrm{x}_t\mathrm{x}_t^\top + \mu\mathrm{I}\|_F \leq \|\mathrm{x}_t\|^2 + \mu\|\mathrm{I}\|_F \leq 1 + \mu\sqrt{d}.$$

Hence, $\mathbb{E}[\|\widehat{\mathrm{g}}_t\|^2] \leq (1+\mu\sqrt{d})^2 =: G^2$ for all iterates. Also note that the diameter of the feasible set is at most $2\sqrt{k}$, that is, $\|\mathrm{M}-\mathrm{M}'\|_F \leq 2\sqrt{k} =: D$ for all M and M′ in the feasible set. By analysis of SGD for convex smooth problems (see Theorem 2 of (Shamir & Zhang, 2013)), we get that for $\eta_t = \frac{c}{\sqrt{t}}$,

$$\begin{aligned}\mathbb{E}[J_\mu(\mathrm{M}_T)] - J_\mu(\mathrm{M}^*) &\leq (\frac{D^2}{c} + cG^2)\frac{2+\log T}{\sqrt{T}}\\ &\leq (\frac{4k}{c} + c(1+\mu\sqrt{d})^2)\frac{2+\log T}{\sqrt{T}}\end{aligned}$$

For $T > 1$, we have $2 + \log T \leq 4\log T$. Choosing $c = \frac{\sqrt{4k}}{1+\mu\sqrt{d}}$ gives

$$\mathbb{E}[J_\mu(\mathrm{M}_T)] - J_\mu(\mathrm{M}_*) \leq \frac{16\sqrt{k}(1+\mu\sqrt{d})\log T}{\sqrt{T}}. \quad (18)$$

To prove the claim of the theorem, we start from its left hand side where we have for any $T > 1$:

$$\begin{aligned}\mathbb{E}[\mathrm{x}^\top\mathrm{M}^*\mathrm{x}] - \mathbb{E}[\mathrm{x}^\top\mathrm{M}_T\mathrm{x}] &= \mathbb{E}[J_\mu(\mathrm{M}_T)] - J_\mu(\mathrm{M}^*)\\ &\quad + \mu\mathrm{Tr}\left(\mathrm{M}^* - \mathbb{E}[\mathrm{M}_T]\right)\end{aligned} \quad (19)$$

We first bound the quantity $\mathrm{Tr}\left(\mathrm{M}^*-\mathrm{M}'\right)$ in terms of the difference in objectives achieved by M^* and M'. Eigendecomposition of $\mathrm{C}-\mu\mathrm{I}$ is given by $\mathrm{C}-\mu\mathrm{I} = \sum_{i=1}^d(\lambda_i-\mu)\mathrm{u}_i\mathrm{u}_i^\top$. By definition of the principal subspace we have $\mathrm{M}^* = \sum_{i=1}^k \mathrm{u}_i\mathrm{u}_i^\top$. Since $\mathrm{u}_1,\ldots,\mathrm{u}_d$ form an orthogonal basis for $\mathbb{R}^d$, we have the following equalities for the trace of the difference:

$$\begin{aligned}\mathrm{Tr}\left(\mathrm{M}^*-\mathrm{M}'\right) &= \sum_{i=1}^d \mathrm{u}_i^\top(\mathrm{M}^*-\mathrm{M}')\mathrm{u}_i\\ &= \sum_{i=1}^k \underbrace{\mathrm{u}_i^\top(\mathrm{M}^*-\mathrm{M}')\mathrm{u}_i}_{\geq 0} + \sum_{i=k+1}^d \underbrace{\mathrm{u}_i^\top(\mathrm{M}^*-\mathrm{M}')\mathrm{u}_i}_{\leq 0}\end{aligned}$$

Note that by definition of M^*, and the fact that $\|\mathrm{M}'\|_2 \leq 1$, we get that $\mathrm{u}_i^\top(\mathrm{M}^*-\mathrm{M}')\mathrm{u}_i \geq 0$ for all $i \in \{1,\ldots,k\}$. On the other hand, since $\lambda_i\mathrm{M}^* = 0$ for $i \in \{k+1,\ldots,d\}$, we have that $\mathrm{u}_i^\top(\mathrm{M}^*-\mathrm{M}')\mathrm{u}_i \leq 0$. We use this property to upper bound the trace by scaling up the ≥ 0 part and scaling down the ≤ 0 part:

$$\begin{aligned}\mathrm{Tr}\left(\mathrm{M}^*-\mathrm{M}'\right) &\leq \sum_{i=1}^k \frac{\lambda_i-\mu}{\lambda_k-\mu}\mathrm{u}_i^\top(\mathrm{M}^*-\mathrm{M}')\mathrm{u}_i\\ &\quad + \sum_{i=k+1}^d \frac{\lambda_i-\mu}{\lambda_k-\mu}\mathrm{u}_i^\top(\mathrm{M}^*-\mathrm{M}')\mathrm{u}_i\\ &= \frac{1}{\lambda_k-\mu}\sum_{i=1}^d(\lambda_i-\mu)\mathrm{u}_i^\top(\mathrm{M}^*-\mathrm{M}')\mathrm{u}_i\\ &= \frac{1}{\lambda_k-\mu}\langle \mathrm{C}-\mu\mathrm{I}, \mathrm{M}^*-\mathrm{M}'\rangle\\ &= \frac{1}{\lambda_k-\mu}(J_\mu(\mathrm{M}')-J_\mu(\mathrm{M}^*))\end{aligned}$$

where the first inequality holds because for $i \in \{1,\ldots,k\}$, $0 < \lambda_k-\mu \leq \lambda_i-\mu$ so that $\frac{\lambda_i-\mu}{\lambda_k-\mu} \geq 1$. Furthermore, for $i \in \{k+1,\ldots,d\}$, $0 < \lambda_k-\mu \geq \lambda_i-\mu$ so that $\frac{\lambda_i-\mu}{\lambda_k-\mu} \leq 1$. Observe that $\mu\mathrm{Tr}\left(\mathrm{M}^*-\mathrm{M}'\right) \leq \frac{\mu}{\lambda_k-\mu}\left(J_\mu(\mathrm{M}')-J_\mu(\mathrm{M}^*)\right)$ whenever μ is an admissible regularization parameter, which together with Equation 19 imply

$$\begin{aligned}\mathbb{E}[\mathrm{x}^\top\mathrm{M}^*\mathrm{x}] - \mathbb{E}[\mathrm{x}^\top\mathrm{M}_T\mathrm{x}] &\leq (\mathbb{E}[J_\mu(\mathrm{M}_T)] - J_\mu(\mathrm{M}^*))\\ &\quad + \frac{\mu}{\lambda_k-\mu}(\mathbb{E}[J_\mu(\mathrm{M}_T)] - J_\mu(\mathrm{M}^*))\\ &\leq \frac{\lambda_k}{\lambda_k-\mu}(\mathbb{E}[J_\mu(\mathrm{M}_T)] - J_\mu(\mathrm{M}^*))\end{aligned}$$

Plugging back the above in Equaition (18), we get

$$\mathbb{E}[\mathrm{x}^\top\mathrm{M}_*\mathrm{x}] - \mathbb{E}[\mathrm{x}^\top\mathrm{M}_T\mathrm{x}] \leq \frac{16\lambda_k\sqrt{k}(1+\mu\sqrt{d})\log T}{(\lambda_k-\mu)\sqrt{T}}.$$

When $\mu \leq \min\{\frac{\lambda_k}{2}, \frac{1}{\sqrt{d}}\}$, it is easy to see that the right hand side above is bounded by $\frac{64\sqrt{k}\log T}{\sqrt{T}}$, which completes the proof.

□

Proof of Lemma 3.3. Let denote the tail at t-th iterate by $\delta_t = \sum_{i=k+1}^d \lambda_i(\mathrm{M}_t)$. Expand the right hand side to get

$$\begin{aligned}\langle \mathrm{C}-\mu\mathrm{I}, \mathrm{M}_*-\mathrm{M}_t\rangle &= \sum_{i=1}^k \lambda_i - \langle \mathrm{C}, \mathrm{M}_t\rangle\\ &\quad - \mu(\mathrm{Tr}\left(\mathrm{M}_*\right) - \mathrm{Tr}\left(\mathrm{M}_t\right))\end{aligned}$$

By Von Neumann's trace inequality, we know $\langle \mathrm{C}, \mathrm{M}_t\rangle \leq \sum_{i=1}^d \lambda_i\lambda_i(\mathrm{M}_t)$. Substituting in the above equality and expanding, we get

$$
\begin{aligned}
&\langle \mathrm{C} - \mu \mathrm{I}, \mathrm{M}_* - \mathrm{M}_t \rangle \\
&\geq \sum_{i=1}^{k} \lambda_i - \sum_{i=1}^{d} \lambda_i \lambda_i(\mathrm{M}_t) - \mu\Big(\sum_{i=1}^{k} 1 - \sum_{i=1}^{d} \lambda_i(\mathrm{M}_t)\Big) \\
&= \sum_{i=1}^{k} \lambda_i (1 - \lambda_i(\mathrm{M}_t)) - \sum_{i=k+1}^{d} \lambda_i \lambda_i(\mathrm{M}_t) \\
&- \mu \sum_{i=1}^{k} (1 - \lambda_i(\mathrm{M}_t)) + \mu \sum_{i=k+1}^{d} \lambda_i(\mathrm{M}_t) \\
&\geq \lambda_k \sum_{i=1}^{k} (1 - \lambda_i(\mathrm{M}_t)) - \lambda_{k+1} \sum_{i=k+1}^{d} \lambda_i(\mathrm{M}_t) \\
&- \mu \sum_{i=1}^{k} (1 - \lambda_i(\mathrm{M}_t)) + \mu \sum_{i=k+1}^{d} \lambda_i(\mathrm{M}_t) \\
&= (\lambda_k - \mu) \sum_{i=1}^{k} (1 - \lambda_i(\mathrm{M}_t)) - (\lambda_{k+1} - \mu) \sum_{i=k+1}^{d} \lambda_i(\mathrm{M}_t) \\
&= (\lambda_k - \mu) \left(k - \sum_{i=1}^{k} \lambda_i(\mathrm{M}_t) \right) - (\lambda_{k+1} - \mu)\delta_t \\
&\geq (\lambda_k - \mu)\delta_t - (\lambda_{k+1} - \mu)\delta_t = (\lambda_k - \lambda_{k+1})\delta_t
\end{aligned}
$$

where the second inequality holds because λ_i's are sorted in descending order. □

Proof of Lemma 3.4. Simply follows from the following inequalities:

$$
\begin{aligned}
\sum_{i=k+1}^{d} \lambda_i(\mathrm{M}_t + \eta_t \mathrm{x}_t \mathrm{x}_t^\top) &\leq \sum_{i=k+1}^{d} \lambda_i(\mathrm{M}_t) + \sum_{i=1}^{d-k} \lambda_i(\eta_t \mathrm{x}_t \mathrm{x}_t^\top) \\
&= \sum_{i=k+1}^{d} \lambda_i(\mathrm{M}_t) + \eta_t \|\mathrm{x}_t\|^2
\end{aligned}
$$

where the first inequality is due to Lidskii's (see, e.g. (Tao, 2012)). Taking expectation of both sides

$$
\begin{aligned}
\mathbb{E}[\sum_{i=k+1}^{d} \lambda_i(\mathrm{M}_t + \eta_t \mathrm{x}_t \mathrm{x}_t^\top)] &\leq \mathbb{E}[\sum_{i=k+1}^{d} \lambda_i(\mathrm{M}_t) + \eta_t \|\mathrm{x}_t\|^2] \\
&\leq \sum_{i=k+1}^{d} \lambda_i(\mathrm{M}_t) + \eta_t
\end{aligned}
$$

where the last inequality holds because $\mathbb{E}[\|\mathrm{x}_t\|^2] \leq 1$ by assumptions of Theorem 3.2. □

Proof of Theorem 3.5. As a consequence of Lemma 3.3 and Lemma 3.4, we have that

$$
\begin{aligned}
\tilde{\delta}_t := \sum_{i=k+1}^{d} \lambda_i(\mathrm{M}_t + \eta_t \mathrm{x}_t \mathrm{x}_t^\top) &\leq \sum_{i=k+1}^{d} \lambda_i(\mathrm{M}_t) + \eta_t \\
&\leq \frac{1}{g_k} \langle \mathrm{C} - \mu \mathrm{I}, \mathrm{M}_* - \mathrm{M}_t \rangle + \eta_t \\
&\leq \frac{16\sqrt{k}(1 + \mu\sqrt{d}) \log t}{g_k \sqrt{t}} + \frac{4\sqrt{k}}{(1 + \mu\sqrt{d})\sqrt{t}}
\end{aligned}
$$

where the last inequality follows from Equation (18). After shrinking the spectrum by $-\eta_t \mu \mathrm{I}$, at most $\frac{\tilde{\delta}_t}{\mu \eta_t}$ eigenvalues in the tail will not be eliminated.

$$
\begin{aligned}
\frac{\tilde{\delta}_t}{\mu \eta_t} &= \frac{\frac{16\sqrt{k}(1+\mu\sqrt{d}) \log t}{g_k \sqrt{t}} + \frac{4\sqrt{k}}{(1+\mu\sqrt{d})\sqrt{t}}}{\mu \frac{2}{1+\mu\sqrt{d}} \sqrt{\frac{k}{t}}} \\
&\leq \frac{8 \log t (1 + \mu\sqrt{d})^2}{\mu g_k} + \frac{2}{\mu} \leq \frac{33 \log t}{\mu g_k}.
\end{aligned}
$$

where the last inequality holds since $\mu \leq 1/\sqrt{d}$. □

A.3. Proofs of Section 4

Proof of Lemma 4.1. Note that the feasible set of Problem 12 is convex, and the objective function is λ-strongly convex. Hence, the optimum is unique. Since Slater condition is satisfied, strong duality holds and KKT conditions are necessary and sufficient for optimality. Let $(\lambda_i(\mathrm{M}), \mathrm{v}_i),\ i \in [d]$ denote the eigenvalues and associated eigenvectors of M. KKT first-order optimality condition yields

$$
0 = -\mathrm{C} + \mu \mathrm{I} + \lambda \mathrm{M} - \sum_{i=1}^{d} \gamma_i \mathrm{v}_i \mathrm{v}_i^\top + \sum_{i=1}^{d} \omega_i \mathrm{v}_i \mathrm{v}_i^\top + \beta \mathrm{I},
$$

where $\gamma_i, \omega_i, \beta \geq 0$ are Lagrange multipliers for the constraints $\mathrm{M} \succeq 0$, $\mathrm{M} \preceq \mathrm{I}$, and $\mathrm{Tr}\,(\mathrm{M}) \leq k$ respectively. Note that except for C, every other term in the above equation has the same set of eigenvectors as M. We conclude that C and M have the same set of eigenvectors, i.e. $\mathrm{u}_i = \mathrm{v}_i,\ i \in [d]$. Rearranging, we get:

$$
\mathrm{C} = \sum_{i=1}^{d} (\mu + \lambda \lambda_i(\mathrm{M}) - \gamma_i + \omega_i + \beta) \mathrm{u}_i \mathrm{u}_i^\top
$$

Complementary slackness implies that $\gamma_i \lambda_i(\mathrm{M})$, $\omega_i(\lambda_i(\mathrm{M}) - 1)$, and $\beta(\mathrm{Tr}\,(\mathrm{M}) - k)$ all vanish. Now, If (λ, μ) is an admissible regularization pair, we can verify that the following satisfy the KKT conditions:

$$
\begin{aligned}
\mathrm{M} &= \Pi_k(\mathrm{C}) \\
\omega_i &= (\lambda_i - \lambda_k) 1_{i \leq k} \\
\gamma_i &= (\lambda_k - \lambda - \lambda_i) 1_{i > k} \\
\beta &= \lambda_k - \mu - \lambda,
\end{aligned}
$$

where $\Pi_k(\mathrm{C})$ returns the projection matrix corresponding to top-k eigenspace of C and $1_{i \in \mathscr{A}}$ is the indicator of set $\mathscr{A}$. In particular, $\gamma_i = \lambda_i - \lambda_k \geq 0$ for $i \in [k]$. Also, since $\lambda < g_k$, for $i \in \{k+1, \ldots, d\}$ it holds that $\gamma_i = \lambda_k - \lambda - \lambda_j \geq 0$. It is easy to verify that the complementary slackness conditions are also satisfied for the above assignments of the primal and dual variables. □