Supplementary Material: Accelerating Adaptive Online Learning by Matrix Approximation

Yuanyu Wan and Lijun Zhang

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{wanyy,zhanglj}@lamda.nju.edu.cn

A Theoretical Analysis

In this supplementary material, we provide proof of Theorems 1 and 2.

A.1 Supporting Results

The following results are used throughout our analysis.

Lemma 1. *(Proposition 3 of [1]). Let sequence $\{\boldsymbol{\beta}_t\}$ be generated by ADA-RP. We have*

$$R(T) \leq \frac{1}{\eta}\sum_{t=1}^{T-1}\left[B_{\Psi_{t+1}}(\boldsymbol{\beta}^*, \boldsymbol{\beta}_{t+1}) - B_{\Psi_t}(\boldsymbol{\beta}^*, \boldsymbol{\beta}_{t+1})\right] + \frac{1}{\eta}B_{\Psi_1}(\boldsymbol{\beta}^*, \boldsymbol{\beta}_1) + \frac{\eta}{2}\sum_{t=1}^{T}\|f_t'(\boldsymbol{\beta}_t)\|_{\Psi_t^*}^2.$$

Lemma 2. *Let $\mathrm{X}_t = \sum_{i=1}^t \mathbf{x}_i\mathbf{x}_i^\top$ and $\mathrm{A}^\dagger$ denote the pseudo-inverse of A, then*

$$\sum_{t=1}^{T}\left\langle \mathbf{x}_t, (\mathrm{X}_t^{1/2})^\dagger \mathbf{x}_t\right\rangle \leq 2\sum_{t=1}^{T}\left\langle \mathbf{x}_t, (\mathrm{X}_T^{1/2})^\dagger \mathbf{x}_t\right\rangle = 2\,\mathrm{tr}(\mathrm{X}_T^{1/2}).$$

Lemma 2 can be proved in the same way as Lemma 10 of [1].

Theorem 3. *(Theorem 2.3 of [11]). Let $0 < \epsilon, \delta < 1$ and $\mathrm{S} = \frac{1}{\sqrt{k}}\mathrm{R} \in \mathbb{R}^{k\times n}$ where the entries $\mathrm{R}_{i,j}$ of R are independent standard normal random variables. Then if $k = \Theta(\frac{d+\log(1/\delta)}{\epsilon^2})$, then for any fixed $n \times d$ matrix A, with probability $1-\delta$, simultaneously for all $\mathbf{x} \in \mathbb{R}^d$,*

$$(1-\epsilon)\|\mathrm{A}\mathbf{x}\|_2^2 \leq \|\mathrm{SA}\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathrm{A}\mathbf{x}\|_2^2.$$

Based on the above theorem, we derive the following corollary.

Corollary 1. *Let $0 < \epsilon, \delta < 1$ and each entry of $\mathbf{r}_t \in \mathbb{R}^\tau$ is a Gaussian random variable drawn from $\mathcal{N}(0, 1/\sqrt{\tau})$ independently. Then, if $\tau = \Omega(\frac{r+\log(T/\delta)}{\epsilon^2})$, with probability $1-\delta$, simultaneously for all $t = 1, ..., T$,*

$$(1-\epsilon)\mathrm{C}_t^\top\mathrm{C}_t \preceq \mathrm{S}_t^\top\mathrm{S}_t \preceq (1+\epsilon)\mathrm{C}_t^\top\mathrm{C}_t.$$

Theorem 4. *(Theorem 10 of [20]). Let* R *be a Gaussian random matrix of size* $p \times n$*. Let* $\mathrm{C} = \mathrm{diag}(c_1, \ldots, c_p)$ *and* $\mathrm{S} = \mathrm{diag}(s_1, \ldots, s_p)$ *be* $p \times p$ *diagonal matrices, where* $c_i \neq 0$ *and* $c_i^2 + s_i^2 = 1$ *for all* i*. Let* $\mathrm{M} = \mathrm{C}^2 + \frac{1}{n}\mathrm{S}\mathrm{R}\mathrm{R}^\top\mathrm{S}$ *and* $r = \sum_i s_i^2$*.*

$$Pr(\lambda_1(M) \geq 1 + t) \leq q \cdot \exp\left(-\frac{cnt^2}{\max_i(s_i^2)r}\right),$$
$$Pr(\lambda_p(M) \leq 1 - t) \leq q \cdot \exp\left(-\frac{cnt^2}{\max_i(s_i^2)r}\right),$$

where the constant c *is at least* $1/32$*, and* q *is the rank of* S*.*

Based on the above theorem, we derive the following corollary.

Corollary 2. *Let* $c \geq 1/32$*,* $\alpha > 0$*,* $\sigma_{ti}^2 = \lambda_i(\mathrm{C}_t^\top \mathrm{C}_t)$*,* $\tilde{r}_t = \sum_i \frac{\sigma_{ti}^2}{\alpha + \sigma_{ti}^2}$*,* $\tilde{r}_* = \max_{k \leq t \leq T} \tilde{r}_t$ *and* $\sigma_{*1}^2 = \max_{1 \leq t \leq T} \sigma_{t1}^2$*. Let* $\mathrm{K}_t = \alpha \mathrm{I}_d + \mathrm{C}_t^\top \mathrm{C}_t$*,* $\tilde{\mathrm{K}}_t = \alpha \mathrm{I}_d + \mathrm{S}_t^\top \mathrm{S}_t$*, and* $\tilde{\mathrm{I}}_t = \mathrm{K}_t^{-1/2} \tilde{\mathrm{K}}_t \mathrm{K}_t^{-1/2}$*. If* $\tau \geq \frac{\tilde{r}_* \sigma_{*1}^2}{c\epsilon^2(\alpha + \sigma_{*1}^2)} \log \frac{2dT}{\delta}$*, then with probability at least* $1 - \delta$*, simultaneously for all* $t = 1, \cdots, T$*,*

$$(1 - \epsilon)\mathrm{I}_d \preceq \tilde{\mathrm{I}}_t \preceq (1 + \epsilon)\mathrm{I}_d.$$

A.2 Proof of Theorem 1

Let $\widetilde{\mathrm{X}}_t$ denote $\mathrm{S}_t^\top \mathrm{S}_t$. First, we consider bounding the first term in the upper bound of Lemma 1. With probability $1 - \delta$, we have

$$\begin{aligned}
&B_{\Psi_{t+1}}(\boldsymbol{\beta}^*, \boldsymbol{\beta}_{t+1}) - B_{\Psi_t}(\boldsymbol{\beta}^*, \boldsymbol{\beta}_{t+1}) \\
=&\frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, (\widetilde{\mathrm{X}}_{t+1}^{1/2} - \widetilde{\mathrm{X}}_t^{1/2})(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
\leq&\frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, \sqrt{1+\epsilon}\mathrm{X}_{t+1}^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
&- \frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, \sqrt{1-\epsilon}\mathrm{X}_t^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
\leq&\frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, (\mathrm{X}_{t+1}^{1/2} - \mathrm{X}_t^{1/2})(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
&+ \frac{\epsilon}{4}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, (\mathrm{X}_{t+1}^{1/2} + \mathrm{X}_t^{1/2})(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
\leq&\frac{1}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}\|_2^2 \|(\mathrm{X}_{t+1}^{1/2} - \mathrm{X}_t^{1/2})\| \\
&+ \frac{\epsilon}{4}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, (\mathrm{X}_{t+1}^{1/2} + \mathrm{X}_t^{1/2})(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
\leq&\frac{1}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}\|_2^2 \,\mathrm{tr}(\mathrm{X}_{t+1}^{1/2} - \mathrm{X}_t^{1/2}) \\
&+ \frac{\epsilon}{4}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, (\mathrm{X}_{t+1}^{1/2} + \mathrm{X}_t^{1/2})(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle
\end{aligned}$$

where the first inequality is due to Corollary 1.

Thus, we can get

$$
\begin{aligned}
&\sum_{t=1}^{T-1}\left[B_{\Psi_{t+1}}(\boldsymbol{\beta}^*,\boldsymbol{\beta}_{t+1})-B_{\Psi_t}(\boldsymbol{\beta}^*,\boldsymbol{\beta}_{t+1})\right]\\
\leq&\frac{1}{2}\sum_{t=1}^{T-1}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_{t+1}\|_2^2\operatorname{tr}(\mathrm{X}_{t+1}^{1/2}-\mathrm{X}_t^{1/2})\\
&+\frac{\epsilon}{4}\sum_{t=1}^{T-1}\left\langle\boldsymbol{\beta}^*-\boldsymbol{\beta}_{t+1},(\mathrm{X}_{t+1}^{1/2}+\mathrm{X}_t^{1/2})(\boldsymbol{\beta}^*-\boldsymbol{\beta}_{t+1})\right\rangle\\
\leq&\frac{1}{2}\max_{t\leq T}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_t\|_2^2\operatorname{tr}(\mathrm{X}_T^{1/2})-\frac{1}{2}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_1\|_2^2\operatorname{tr}(\mathrm{X}_1^{1/2})\\
&+\frac{\epsilon}{2}\max_{t\leq T}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_t\|_2^2\sum_{t=1}^{T}\|\mathrm{X}_t^{1/2}\|-\frac{\epsilon}{4}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_1\|_2^2\operatorname{tr}(\mathrm{X}_1^{1/2}).
\end{aligned}
\tag{3}
$$

Note that $\boldsymbol{\beta}_1=\mathbf{0}$, then

$$
\begin{aligned}
B_{\Psi_1}(\boldsymbol{\beta}^*,\boldsymbol{\beta}_1)=&\frac{1}{2}\left\langle\boldsymbol{\beta}^*,(\sigma\mathrm{I}_d+\widetilde{\mathrm{X}}_1^{1/2})\boldsymbol{\beta}^*\right\rangle\\
\leq&\frac{1}{2}\sigma\|\boldsymbol{\beta}^*\|_2^2+\frac{2+\epsilon}{4}\|\boldsymbol{\beta}^*\|_2^2\operatorname{tr}(\mathrm{X}_1^{1/2})
\end{aligned}
\tag{4}
$$

where the inequality is due to Corollary 1.

Then, we consider the upper bound of $\sum_{t=1}^{T}\|f_t'(\boldsymbol{\beta}_t)\|_{\Psi_t^*}^2$. With probability $1-\delta$, we have

$$
\begin{aligned}
&\frac{1}{2}\|f_t'(\boldsymbol{\beta}_t)\|_{\Psi_t^*}^2=\left\langle\mathbf{g}_t,(\sigma\mathrm{I}_d+\widetilde{\mathrm{X}}_t^{1/2})^{-1}\mathbf{g}_t\right\rangle\\
\leq&\frac{1}{\sqrt{1-\epsilon}}\left\langle\mathbf{g}_t,(\mathrm{X}_t^{\dagger})^{1/2}\mathbf{g}_t\right\rangle=\frac{l'(\boldsymbol{\beta}_t^\top\mathbf{x}_t)^2}{\sqrt{1-\epsilon}}\left\langle\mathbf{x}_t,(\mathrm{X}_t^{\dagger})^{1/2}\mathbf{x}_t\right\rangle
\end{aligned}
$$

where the inequality is due to Corollary 1. According to Lemma 2, we have

$$
\begin{aligned}
\sum_{t=1}^{T}\|f_t'(\boldsymbol{\beta}_t)\|_{\Psi_t^*}^2\leq&\sum_{t=1}^{T}\frac{2l'(\boldsymbol{\beta}_t^\top\mathbf{x}_t)^2}{\sqrt{1-\epsilon}}\left\langle\mathbf{x}_t,(\mathrm{X}_t^{\dagger})^{1/2}\mathbf{x}_t\right\rangle\\
\leq&\max_{t\leq T}l'(\boldsymbol{\beta}_t^\top\mathbf{x}_t)^2\frac{2}{\sqrt{1-\epsilon}}\sum_{t=1}^{T}\left\langle\mathbf{x}_t,(\mathrm{X}_t^{\dagger})^{1/2}\mathbf{x}_t\right\rangle\\
\leq&\frac{4}{\sqrt{1-\epsilon}}\max_{t\leq T}l'(\boldsymbol{\beta}_t^\top\mathbf{x}_t)^2\operatorname{tr}(\mathrm{X}_T^{1/2}).
\end{aligned}
\tag{5}
$$

We complete the proof by substituting (3), (4), and (5) into Lemma 1.

A.3 Proof of Theorem 2

Inspired by the proof of Theorem 1, we can derive Theorem 2 by respectively bounding each term in the upper bound of Lemma 1. Before that, we need to derive the lower and upper bounds of $(\mathrm{S}_t^\top\mathrm{S}_t)^{1/2}$ based on Corollary 2.

Let the SVD of $\mathrm{C}_t^\top$ be $\mathrm{C}_t^\top = \mathrm{U}\Sigma\mathrm{V}^\top$ where $\mathrm{U} \in \mathbb{R}^{d\times d}, \Sigma \in \mathbb{R}^{d\times d}, \mathrm{V} \in \mathbb{R}^{t\times d}$. According to Corollary 2, with probability at least $1-\delta$, simultaneously for all $t = 1, ..., T$,

$$\begin{aligned}
\mathrm{S}_t^\top \mathrm{S}_t &= \tilde{\mathrm{K}}_t - \alpha \mathrm{I}_d = \mathrm{K}_t^{1/2} \tilde{\mathrm{I}}_t \mathrm{K}_t^{1/2} - \alpha \mathrm{I}_d \\
&\preceq (1+\epsilon)\mathrm{K}_t - \alpha \mathrm{I}_d = (1+\epsilon)\mathrm{C}_t^\top \mathrm{C}_t + \epsilon\alpha \mathrm{I}_d \\
&= \mathrm{U}\big((1+\epsilon)\Sigma\Sigma + \epsilon\alpha\mathrm{I}_d\big)\mathrm{U}^\top
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{S}_t^\top \mathrm{S}_t + \epsilon\alpha \mathrm{I}_d &= \tilde{\mathrm{K}}_t - \alpha \mathrm{I}_d + \epsilon\alpha \mathrm{I}_d \\
&= \mathrm{K}_t^{1/2} \tilde{\mathrm{I}}_t \mathrm{K}_t^{1/2} - \alpha \mathrm{I}_d + \epsilon\alpha \mathrm{I}_d \\
&\succeq (1-\epsilon)\mathrm{K}_t - \alpha \mathrm{I}_d + \epsilon\alpha \mathrm{I}_d \\
&= (1-\epsilon)\mathrm{C}_t^\top \mathrm{C}_t.
\end{aligned}$$

Then simultaneously for all $t = 1, ..., T$, we have

$$(\mathrm{S}_t^\top \mathrm{S}_t)^{1/2} \preceq \sqrt{1+\epsilon}\,\mathrm{U}(\Sigma\Sigma)^{1/2}\mathrm{U}^\top + \sqrt{\epsilon\alpha}\,\mathrm{U}\mathrm{I}_d\mathrm{U}^\top = \sqrt{1+\epsilon}\,\mathrm{X}_t^{1/2} + \sqrt{\epsilon\alpha}\,\mathrm{I}_d \tag{6}$$

and

$$\begin{aligned}
(\mathrm{S}_t^\top \mathrm{S}_t)^{1/2} &= (\mathrm{S}_t^\top \mathrm{S}_t)^{1/2} + \sqrt{\epsilon\alpha}\,\mathrm{I}_d - \sqrt{\epsilon\alpha}\,\mathrm{I}_d \\
&\succeq \big((\mathrm{S}_t^\top \mathrm{S}_t) + \epsilon\alpha \mathrm{I}_d\big)^{1/2} - \sqrt{\epsilon\alpha}\,\mathrm{I}_d \\
&\succeq \sqrt{1-\epsilon}\,\mathrm{X}_t^{1/2} - \sqrt{\epsilon\alpha}\,\mathrm{I}_d.
\end{aligned} \tag{7}$$

Then we consider bounding the first term in the upper bound of Lemma 1. Let $\widetilde{\mathrm{X}}_t$ denote $\mathrm{S}_t^\top \mathrm{S}_t$. Simultaneously for all $t = 1, ..., T$, we have

$$\begin{aligned}
& B_{\Psi_{t+1}}(\boldsymbol{\beta}^*, \boldsymbol{\beta}_{t+1}) - B_{\Psi_t}(\boldsymbol{\beta}^*, \boldsymbol{\beta}_{t+1}) \\
=& \frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, (\widetilde{\mathrm{X}}_{t+1}^{1/2} - \widetilde{\mathrm{X}}_t^{1/2})(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
\leq& \frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, \sqrt{1+\epsilon}\,\mathrm{X}_{t+1}^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
& - \frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, \sqrt{1-\epsilon}\,\mathrm{X}_t^{1/2})(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
& + \frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, 2\sqrt{\epsilon\alpha}\,\mathrm{I}_d(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
=& \frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, \sqrt{1+\epsilon}\,\mathrm{X}_{t+1}^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
& - \frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, \sqrt{1-\epsilon}\,\mathrm{X}_t^{1/2})(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
& + \sqrt{\epsilon\alpha}\|(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\|_2^2 \\
\leq& \frac{1}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}\|_2^2 \operatorname{tr}(\mathrm{X}_{t+1}^{1/2} - \mathrm{X}_t^{1/2}) \\
& + \frac{\epsilon}{4}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, (\mathrm{X}_{t+1}^{1/2} + \mathrm{X}_t^{1/2})(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\right\rangle \\
& + \sqrt{\epsilon\alpha}\|(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1})\|_2^2
\end{aligned}$$

where the first inequality is due to (6), (7) and the last inequality has been proved in the proof of Theorem 1.

Thus, we can get

$$\begin{aligned}&\sum_{t=1}^{T-1}\left[B_{\Psi_{t+1}}(\boldsymbol{\beta}^*,\boldsymbol{\beta}_{t+1})-B_{\Psi_t}(\boldsymbol{\beta}^*,\boldsymbol{\beta}_{t+1})\right]\\ \leq&\frac{1}{2}\max_{t\leq T}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_t\|_2^2\operatorname{tr}(\mathrm{X}_T^{1/2})-\frac{1}{2}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_1\|_2^2\operatorname{tr}(\mathrm{X}_1^{1/2})\\ &+\frac{\epsilon}{2}\max_{t\leq T}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_t\|_2^2\sum_{t=1}^{T}\|\mathrm{X}_t^{1/2}\|\\ &-\frac{\epsilon}{4}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_1\|_2^2\operatorname{tr}(\mathrm{X}_1^{1/2})\\ &+\sqrt{\epsilon\alpha}(T-1)\max_{t\leq T}\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_t\|_2^2.\end{aligned}\tag{8}$$

Note that $\boldsymbol{\beta}_1=\mathbf{0}$, then

$$\begin{aligned}B_{\Psi_1}(\boldsymbol{\beta}^*,\boldsymbol{\beta}_1)=&\frac{1}{2}\left\langle\boldsymbol{\beta}^*,(\sigma\mathrm{I}_d+\widetilde{\mathrm{X}}_1^{1/2})\boldsymbol{\beta}^*\right\rangle\\ \leq&\frac{1}{2}\sigma\|\boldsymbol{\beta}^*\|_2^2+\frac{2+\epsilon}{4}\|\boldsymbol{\beta}^*\|_2^2\operatorname{tr}(\mathrm{X}_1^{1/2})+\frac{1}{2}\sqrt{\epsilon\alpha}\|\boldsymbol{\beta}^*\|_2^2.\end{aligned}\tag{9}$$

Before considering the upper bound of $\sum_{t=1}^{T}\|f_t'(\boldsymbol{\beta}_t)\|_{\Psi_t^*}^2$, we need to derive the upper bound of H_t^{-1}.

Let the SVD of $\mathrm{S}_t^\top$ be $\mathrm{S}_t^\top=\mathrm{U}\Sigma\mathrm{V}^\top$ where $\mathrm{U}\in\mathbb{R}^{d\times d},\Sigma\in\mathbb{R}^{d\times d},\mathrm{V}\in\mathbb{R}^{t\times d}$. We also have, for all $t=1,...,T$,

$$\begin{aligned}\mathrm{H}_t&=\sigma\mathrm{I}_d+(\mathrm{S}_t^\top\mathrm{S}_t)^{1/2}=\mathrm{U}\big(\sigma\mathrm{I}_d+(\Sigma\Sigma)^{1/2}\big)\mathrm{U}^\top\\ &\succeq\mathrm{U}\big(\alpha\mathrm{I}_d+(\Sigma\Sigma)\big)^{1/2}\mathrm{U}^\top=(\alpha\mathrm{I}_d+\mathrm{S}_t^\top\mathrm{S}_t)^{1/2}\end{aligned}$$

due to $\sigma\geq\sqrt{\alpha}\geq\sqrt{\lambda_i(\mathrm{S}_t^\top\mathrm{S}_t)+\alpha}-\sqrt{\lambda_i(\mathrm{S}_t^\top\mathrm{S}_t)}$ for all $i=1,\cdots,d$.

Then according to Corollary 2, with probability at least $1-\delta$, simultaneously for all $t=1,...,T$,

$$\begin{aligned}\mathrm{H}_t^{-1}&\preceq\big((\alpha\mathrm{I}_d+\mathrm{S}_t^\top\mathrm{S}_t)^{1/2}\big)^{-1}=\big((\mathrm{K}_t^{1/2}\tilde{\mathrm{I}}_t\mathrm{K}_t^{1/2})^{-1}\big)^{1/2}\\ &\preceq\frac{1}{\sqrt{1-\epsilon}}(\mathrm{K}_t^{-1})^{1/2}=\frac{1}{\sqrt{1-\epsilon}}\big((\alpha\mathrm{I}_d+\mathrm{X}_t)^{-1}\big)^{1/2}.\end{aligned}$$

Thus, we can get

$$\begin{aligned}\|f_t'(\boldsymbol{\beta}_t)\|_{\Psi_t^*}^2=&2\left\langle\mathbf{g}_t,\mathrm{H}_t^{-1}\mathbf{g}_t\right\rangle\leq\frac{2}{\sqrt{1-\epsilon}}\left\langle\mathbf{g}_t,\big((\alpha\mathrm{I}_d+\mathrm{X}_t)^{-1}\big)^{1/2}\mathbf{g}_t\right\rangle\\ =&\frac{2l'(\boldsymbol{\beta}_t^\top\mathbf{x}_t)^2}{\sqrt{1-\epsilon}}\left\langle\mathbf{x}_t,(\mathrm{X}_t^\dagger)^{1/2}\mathbf{x}_t\right\rangle.\end{aligned}$$

According to Lemma 2, we have

$$\begin{aligned}\sum_{t=1}^{T}\|f_t'(\boldsymbol{\beta}_t)\|_{\Psi_t^*}^2 &\le \frac{2}{\sqrt{1-\epsilon}}\max_{t\le T} l'(\boldsymbol{\beta}_t^\top \mathbf{x}_t)^2 \sum_{t=1}^{T}\left\langle \mathbf{x}_t, (\mathrm{X}_t^\dagger)^{1/2}\mathbf{x}_t\right\rangle \\ &\le \frac{4}{\sqrt{1-\epsilon}}\max_{t\le T} l'(\boldsymbol{\beta}_t^\top \mathbf{x}_t)^2 \,\mathrm{tr}(\mathrm{X}_T^{1/2}).\end{aligned} \tag{10}$$

We complete the proof by substituting (8), (9), and (10) into Lemma 1.

A.4 Proof of Corollary 1

Let $\mathrm{C}_t = \mathrm{U}\Sigma\mathrm{V}^\top$ be the singular value decomposition of C_t. Notice that $\mathrm{U} \in \mathbb{R}^{t\times r}, \Sigma\mathrm{V}^\top \in \mathbb{R}^{r\times d}$. According to Theorem 3, we have if $\tau = \Theta(\frac{r+\log(1/\delta)}{\epsilon^2})$, then simultaneously $\forall \mathbf{x} \in \mathbb{R}^r$, with probability $1-\delta$,

$$(1-\epsilon)\|\mathrm{U}\mathbf{x}\|_2^2 \le \|\mathrm{R_t}\mathrm{U}\mathbf{x}\|_2^2 \le (1+\epsilon)\|\mathrm{U}\mathbf{x}\|_2^2$$

Let $\mathbf{y} \in \mathbb{R}^d$ be arbitrary vector, then $\mathrm{C}_t\mathbf{y} = \mathrm{U}\Sigma\mathrm{V}^\top\mathbf{y} = \mathrm{U}\mathbf{x}$ where $\mathbf{x} = \Sigma\mathrm{V}^\top\mathbf{y} \in \mathbb{R}^r$.
Then we have

$$\mathbf{y}^\top \mathrm{S}_t^\top \mathrm{S}_t\mathbf{y} = \mathbf{y}^\top \mathrm{C}_t^\top \mathrm{R}_t^\top \mathrm{R}_t\mathrm{C}_t\mathbf{y} = \|\mathrm{R_t}\mathrm{U}\mathbf{x}\|_2^2 \le (1+\epsilon)\|\mathrm{U}\mathbf{x}\|_2^2 = (1+\epsilon)\mathbf{y}^\top\mathrm{C}_t^\top\mathrm{C}_t\mathbf{y}$$

and

$$\mathbf{y}^\top \mathrm{S}_t^\top \mathrm{S}_t\mathbf{y} = \mathbf{y}^\top \mathrm{C}_t^\top \mathrm{R}_t^\top \mathrm{R}_t\mathrm{C}_t\mathbf{y} = \|\mathrm{R_t}\mathrm{U}\mathbf{x}\|_2^2 \ge (1-\epsilon)\|\mathrm{U}\mathbf{x}\|_2^2 = (1-\epsilon)\mathbf{y}^\top\mathrm{C}_t^\top\mathrm{C}_t\mathbf{y}.$$

Then, we have $(1-\epsilon)\mathrm{C}_t^\top\mathrm{C}_t \preceq \mathrm{S}_t^\top\mathrm{S}_t \preceq (1+\epsilon)\mathrm{C}_t^\top\mathrm{C}_t$ with probability $1-\delta$, provided $\tau = \Omega(\frac{r+\log(1/\delta)}{\epsilon^2})$. Using the union bound, we have if $\tau = \Omega(\frac{r+\log(T/\delta)}{\epsilon^2})$, with probability $1-\delta$, simultaneously for all $t = 1, ..., T$,

$$(1-\epsilon)\mathrm{C}_t^\top\mathrm{C}_t \preceq \mathrm{S}_t^\top\mathrm{S}_t \preceq (1+\epsilon)\mathrm{C}_t^\top\mathrm{C}_t.$$

A.5 Proof of Corollary 2

Let the SVD of $\mathrm{C}_t^\top$ be $\mathrm{C}_t^\top = \mathrm{U}\Sigma\mathrm{V}^\top$ where $\mathrm{U} \in \mathbb{R}^{d\times d}, \Sigma \in \mathbb{R}^{d\times d}, \mathrm{V} \in \mathbb{R}^{t\times d}$. Then we have $\mathrm{K}_t = \mathrm{U}(\alpha\mathrm{I}_d + \Sigma\Sigma^\top)\mathrm{U}^\top$ and

$$\begin{aligned}\tilde{\mathrm{I}}_t &= \mathrm{K}_t^{-1/2}\tilde{\mathrm{K}}_t\mathrm{K}_t^{-1/2} = \mathrm{K}_t^{-1/2}(\alpha\mathrm{I}_d + \mathrm{C}_t^\top\mathrm{R}_t^\top\mathrm{R}_t\mathrm{C}_t)\mathrm{K}_t^{-1/2} \\ &= \mathrm{U}\Big(\alpha\mathrm{I}_d(\alpha\mathrm{I}_d+\Sigma\Sigma)^{-1} + (\alpha\mathrm{I}_p+\Sigma\Sigma)^{-1/2}\Sigma\mathrm{V}^\top\mathrm{R}_t^\top\mathrm{R}_t\mathrm{V}\Sigma(\alpha\mathrm{I}_d+\Sigma\Sigma^\top)^{-1/2}\Big)\mathrm{U}^\top \\ &= \mathrm{U}\Big(\alpha\mathrm{I}_d(\alpha\mathrm{I}_d+\Sigma\Sigma)^{-1} + (\alpha\mathrm{I}_p+\Sigma\Sigma)^{-1/2}\Sigma\mathrm{R}\mathrm{R}^\top\Sigma(\alpha\mathrm{I}_d+\Sigma\Sigma^\top)^{-1/2}\Big)\mathrm{U}^\top\end{aligned}$$

where $\mathrm{R} = \mathrm{V}^\top\mathrm{R}_t^\top \in \mathbb{R}^{d\times\tau}$ is a Gaussian random matrix due to that V is an orthogonal matrix and $\mathrm{R}_t^\top$ is a Gaussian random matrix.

Let $c_i^2 = \frac{\alpha}{\alpha+\sigma_{ti}^2}$ and $s_i^2 = \frac{\sigma_{ti}^2}{\alpha+\sigma_{ti}^2}$. Then according to Theorem 4, with probability at least $1-\delta$,

$$(1-\epsilon)\mathrm{I}_d \preceq \tilde{\mathrm{I}}_t \preceq (1+\epsilon)\mathrm{I}_d$$

provided $\tau \geq \frac{\tilde{r}_t \sigma_{t1}^2}{c\epsilon^2(\alpha+\sigma_{t1}^2)} \log \frac{2d}{\delta}$ where the constant c is at least $1/32$. Using the union bound, we complete the proof.