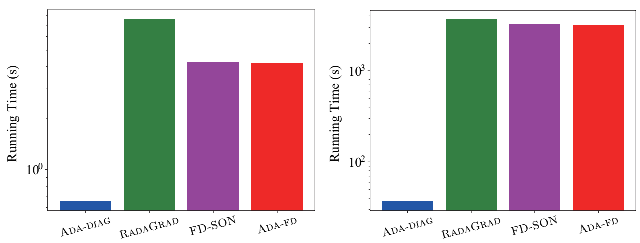
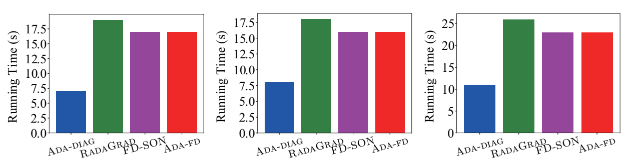
Supplementary Material: Efficient Adaptive Online Learning via Frequent Directions

Yuanyu Wan, Nan Wei, Lijun Zhang
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
{wanyy, zhanglj}@lamda.nju.edu.cn, nwei@smail.nju.edu.cn



(a) Gisette (b) Epsilon

Figure 5: The comparison of running time among different algorithms for composite mirror descent (CMD) method



(a) MNIST (b) CIFAR10 (c) SVHN

Figure 6: The comparison of running time cost by one epoch of each algorithm

A Additional Comparison of Running Time

In Section 4.2, we have performed online classification to evaluate the performance of our ADA-FD with two real world datasets: Gisette and Epsilon which are high-dimensional and dense. Figure 5 shows the comparison of running time among different algorithms for composite mirror descent method on both datasets. We find that our ADA-FD is faster than RADAGRAD and as fast as FD-SON when $d = 5000$ and $d = 2000$.

In Section 4.3, we have compared ADA-FD against ADA-DIAG, RADAGRAD and FD-SON on training the classical convolutional neural networks (CNN). Figure 6 shows the comparison of running time cost by one epoch of each algorithm. We verify that our ADA-FD is faster than RADAGRAD and as fast as FD-SON when applied to training CNN.

B Theoretical Analysis

In this section, we provide omitted proofs.

B.1 Supporting Results

The following results are used throughout our analysis.

Lemma 1. *(Variant of Proposition 2 in Duchi* et al. *[2011]). Let sequence $\{\boldsymbol{\beta}_t\}$ be generated by Algorithm 1. We have*

$$R(T) \leq \frac{1}{\eta}\Psi_T(\boldsymbol{\beta}^*) + \frac{\eta}{2}\sum_{t=1}^{T}\|f_t'(\boldsymbol{\beta}_t)\|_{\Psi_{t-1}^*}^2 + \frac{\sum_{t=1}^{T}\sqrt{\sigma_t}}{2\eta}\max_{t\leq T}\|\boldsymbol{\beta}_{t+1}\|_2^2.$$

Lemma 1 can be regard as an variant of Proposition 2 in Duchi *et al.* [2011], when the condition $\Psi_{t+1}(\boldsymbol{\beta}) \geq \Psi_t(\boldsymbol{\beta})$ cannot be met due to $\mathrm{H}_{t+1} \not\succeq \mathrm{H}_t$ in this work. Lemma 1 can be derived from the proof of Proposition 2 in Duchi *et al.* [2011] with slight modification to deal with $\Psi_{t+1}(\boldsymbol{\beta}) \not\geq \Psi_t(\boldsymbol{\beta})$. We include the proof for completeness.

Proof. The conjugate dual of $t\varphi(\boldsymbol{\beta}) + \frac{1}{\eta}\Psi_t(\boldsymbol{\beta})$ is defined by

$$\Phi_t^*(\mathbf{g}) = \sup_{\boldsymbol{\beta}}\left\{\langle \mathbf{g}, \boldsymbol{\beta}\rangle - t\varphi(\boldsymbol{\beta}) - \frac{1}{\eta}\Psi_t(\boldsymbol{\beta})\right\}.$$

Thus, the gradient of $\Phi_t^*(\mathbf{g})$ can be calculated as

$$\nabla\Phi_t^*(\mathbf{g}) = \arg\min_{\boldsymbol{\beta}}\left\{-\langle \mathbf{g}, \boldsymbol{\beta}\rangle + t\varphi(\boldsymbol{\beta}) + \frac{1}{\eta}\Psi_t(\boldsymbol{\beta})\right\}. \quad (8)$$

Because $\frac{1}{\eta}\Psi_t(\boldsymbol{\beta})$ is $\frac{1}{\eta}$-strongly convex with respect to the norm $\|.\|_{\Psi_t}$, we have

$$\|\nabla\Phi_t^*(\mathbf{x}) - \nabla\Phi_t^*(\mathbf{y})\|_{\Psi_t} \leq \eta\|\mathbf{x}-\mathbf{y}\|_{\Psi_t^*}$$

which means the function Φ_t^* has η-Lipschitz continuous gradients with respect to $\|.\|_{\Psi_t^*}$. Further, we have

$$\Phi_t^*(\mathbf{y}) \leq \Phi_t^*(\mathbf{x}) + \langle\nabla\Phi_t^*(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{\eta}{2}\|\mathbf{y}-\mathbf{x}\|_{\Psi_t^*}^2. \quad (9)$$

Both the identity (8) and the bound (9) were used in the proof of Proposition 2 in Duchi *et al.* [2011]. In order to complete the proof, we introduce an inequality

$$\begin{aligned}&\sum_{t=1}^{T} f_t(\boldsymbol{\beta}_t) + \varphi(\boldsymbol{\beta}_t) - f_t(\boldsymbol{\beta}^*) - \varphi(\boldsymbol{\beta}^*)\\ \leq& \frac{1}{\eta}\Psi_T(\boldsymbol{\beta}^*) + \sum_{t=1}^{T}\{\langle \mathbf{g}_t, \boldsymbol{\beta}_t\rangle + \varphi(\boldsymbol{\beta}_t)\} + \Phi_T^*(-\bar{\mathbf{g}}_T).\end{aligned} \quad (10)$$

from the proof of Proposition 2 in Duchi *et al.* [2011] again.

Due to

$$
\begin{aligned}
&(\mathrm{S}_t^\top \mathrm{S}_t)^{1/2} - (\mathrm{S}_{t-1}^\top \mathrm{S}_{t-1})^{1/2} + \sqrt{\sigma_t} V V^\top \\
=&\mathrm{V}\Sigma'\mathrm{V}^\top + \sqrt{\sigma_t} V V^\top - (\mathrm{S}_{t-1}^\top \mathrm{S}_{t-1})^{1/2} \\
\succeq&\mathrm{V}\Sigma\mathrm{V}^\top - (\mathrm{S}_{t-1}^\top \mathrm{S}_{t-1})^{1/2} \\
=&(\mathrm{S}_{t-1}^\top \mathrm{S}_{t-1} + \mathbf{g}_t \mathbf{g}_t^\top)^{1/2} - (\mathrm{S}_{t-1}^\top \mathrm{S}_{t-1})^{1/2} \\
\succeq&0
\end{aligned}
$$

we have

$$
-\Psi_t(\mathbf{x}) \le -\Psi_{t-1}(\mathbf{x}) + \frac{\sqrt{\sigma_t}}{2}\|\mathbf{x}\|_2^2.
$$

Thus, we have

$$
\begin{aligned}
&\Phi_T^*(-\bar{\mathbf{g}}_T) \\
=&-\langle \bar{\mathbf{g}}_T, \boldsymbol{\beta}_{T+1}\rangle - T\varphi(\boldsymbol{\beta}_{T+1}) - \frac{1}{\eta}\Psi_T(\boldsymbol{\beta}_{T+1}) \\
\le&-\langle \bar{\mathbf{g}}_T, \boldsymbol{\beta}_{T+1}\rangle - T\varphi(\boldsymbol{\beta}_{T+1}) - \frac{1}{\eta}\Psi_{T-1}(\boldsymbol{\beta}_{T+1}) \\
&+ \frac{\sqrt{\sigma_T}}{2\eta}\|\boldsymbol{\beta}_{T+1}\|_2^2 \\
\le&\sup_{\boldsymbol{\beta}}\left(-\langle \bar{\mathbf{g}}_T, \boldsymbol{\beta}\rangle - (T-1)\varphi(\boldsymbol{\beta}) - \frac{1}{\eta}\Psi_{T-1}(\boldsymbol{\beta})\right) \\
&- \varphi(\boldsymbol{\beta}_{T+1}) + \frac{\sqrt{\sigma_T}}{2\eta}\|\boldsymbol{\beta}_{T+1}\|_2^2 \\
=&\Phi_{T-1}^*(-\bar{\mathbf{g}}_T) - \varphi(\boldsymbol{\beta}_{T+1}) + \frac{\sqrt{\sigma_T}}{2\eta}\|\boldsymbol{\beta}_{T+1}\|_2^2
\end{aligned}
$$

which contains an additional term $\frac{\sqrt{\sigma_T}}{2\eta}\|\boldsymbol{\beta}_{T+1}\|_2^2$ caused by $\mathrm{H}_T \not\succeq \mathrm{H}_{T-1}$ compared with Duchi *et al.* [2011].

Using the identity (8), the bound (9) and the inequality (10), we have

$$
\begin{aligned}
&\sum_{t=1}^T f_t(\boldsymbol{\beta}_t) + \varphi(\boldsymbol{\beta}_{t+1}) - f_t(\boldsymbol{\beta}^*) - \varphi(\boldsymbol{\beta}^*) \\
\le&\frac{1}{\eta}\Psi_T(\boldsymbol{\beta}^*) + \sum_{t=1}^T \{\langle \mathbf{g}_t, \boldsymbol{\beta}_t\rangle + \varphi(\boldsymbol{\beta}_{t+1})\} + \Phi_{T-1}^*(-\bar{\mathbf{g}}_T) \\
&- \varphi(\boldsymbol{\beta}_{T+1}) + \frac{\sqrt{\sigma_T}}{2\eta}\|\boldsymbol{\beta}_{T+1}\|_2^2 \\
\le&\frac{1}{\eta}\Psi_T(\boldsymbol{\beta}^*) + \sum_{t=1}^T \{\langle \mathbf{g}_t, \boldsymbol{\beta}_t\rangle + \varphi(\boldsymbol{\beta}_{t+1})\} + \Phi_{T-1}^*(-\bar{\mathbf{g}}_{T-1}) \\
&- \left\langle \nabla\Phi_{T-1}^*(-\bar{\mathbf{g}}_{T-1}), \mathbf{g}_T\right\rangle + \frac{\eta}{2}\|\mathbf{g}_T\|_{\Psi_{T-1}^*}^2 - \varphi(\boldsymbol{\beta}_{T+1}) \\
&+ \frac{\sqrt{\sigma_T}}{2\eta}\|\boldsymbol{\beta}_{T+1}\|_2^2 \\
=&\frac{1}{\eta}\Psi_T(\boldsymbol{\beta}^*) + \sum_{t=1}^{T-1} \{\langle \mathbf{g}_t, \boldsymbol{\beta}_t\rangle + \varphi(\boldsymbol{\beta}_{t+1})\} + \Phi_{T-1}^*(-\bar{\mathbf{g}}_{T-1}) \\
&+ \frac{\eta}{2}\|\mathbf{g}_T\|_{\Psi_{T-1}^*}^2 + \frac{\sqrt{\sigma_T}}{2\eta}\|\boldsymbol{\beta}_{T+1}\|_2^2.
\end{aligned}
$$

By repeating the above steps, we have

$$
\begin{aligned}
&\sum_{t=1}^T f_t(\boldsymbol{\beta}_t) + \varphi(\boldsymbol{\beta}_{t+1}) - f_t(\boldsymbol{\beta}^*) - \varphi(\boldsymbol{\beta}^*) \\
\le&\frac{1}{\eta}\Psi_T(\boldsymbol{\beta}^*) + \frac{\eta}{2}\sum_{t=1}^T \|\mathbf{g}_t\|_{\Psi_{t-1}^*}^2 + \sum_{t=1}^T \frac{\sqrt{\sigma_t}}{2\eta}\|\boldsymbol{\beta}_{t+1}\|_2^2 \\
&+ \Phi_0^*(\bar{\mathbf{g}}_0).
\end{aligned}
$$

Note that $\varphi(\boldsymbol{\beta}) = 0$ and $\Phi_0^*(0) = 0$. We complete the proof. □

Lemma 2. *(Proposition 3 in Duchi* et al. *[2011]). Let sequence $\{\boldsymbol{\beta}_t\}$ be generated by Algorithm 2. We have*

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

Lemma 3. *(Lemma 10 in Duchi* et al. *[2011]) Let $\mathrm{G}_t = \sum_{i=1}^t \mathbf{g}_i\mathbf{g}_i^\top$ and $\mathrm{A}^\dagger$ denote the pseudo-inverse of A, then*

$$
\sum_{t=1}^T \langle \mathbf{g}_t, (\mathrm{G}_t^{1/2})^\dagger \mathbf{g}_t\rangle \le 2\sum_{t=1}^\top \langle \mathbf{g}_t, (\mathrm{G}_T^{1/2})^\dagger \mathbf{g}_t\rangle = 2\,\mathrm{tr}(\mathrm{G}_T^{1/2}).
$$

Lemma 4. *(Derived From Theorem 3.1 and its Proof in Ghashami* et al. *[2016]) Let $\Delta_t = \sum_{i=1}^t \sigma_i$. In Algorithm 1 and 2, S_t is the sketch of the input C_t produced by frequent directions. Then for any t and $k < \tau$,*

$$
\mathrm{C}_t^\top \mathrm{C}_t \succeq \mathrm{S}_t^\top \mathrm{S}_t \succeq \mathrm{C}_t^\top \mathrm{C}_t - \Delta_t \mathrm{I}_p
$$

and

$$
\Delta_t \le \|\mathrm{C}_t - \mathrm{C}_t^k\|_F^2/(\tau - k)
$$

where C_t^k denotes the minimizer of $\|\mathrm{C}_t - \mathrm{C}_t^k\|_F$ over all rank k matrices.

B.2 Proof of Theorem 1

We first consider $\frac{1}{\eta}\Psi_T(\boldsymbol{\beta}^*)$ in the upper bound of Lemma 1. We have

$$
\begin{aligned}
\frac{1}{\eta}\Psi_T(\boldsymbol{\beta}^*) =& \frac{1}{2\eta}\left\langle \boldsymbol{\beta}^*, (\delta\mathrm{I}_d + (\mathrm{S}_T^\top \mathrm{S}_T)^{1/2})\boldsymbol{\beta}^*\right\rangle \\
\le& \frac{\delta}{2\eta}\|\boldsymbol{\beta}^*\|_2^2 + \frac{1}{2\eta}\left\langle \boldsymbol{\beta}^*, (\mathrm{C}_T^\top \mathrm{C}_T)^{1/2}\boldsymbol{\beta}^*\right\rangle \\
\le& \frac{\delta}{2\eta}\|\boldsymbol{\beta}^*\|_2^2 + \frac{1}{2\eta}\lambda_{\max}(\mathrm{G}_T^{1/2})\|\boldsymbol{\beta}^*\|_2^2 \\
\le& \frac{\delta}{2\eta}\|\boldsymbol{\beta}^*\|_2^2 + \frac{1}{2\eta}\,\mathrm{tr}(\mathrm{G}_T^{1/2})\|\boldsymbol{\beta}^*\|_2^2.
\end{aligned}
\tag{11}
$$

Before considering $\frac{\eta}{2}\sum_{t=1}^T \|f_t'(\boldsymbol{\beta}_t)\|_{\Psi_{t-1}^*}^2$, we need derive the lower bound of H_{t-1}. Let $c = \frac{\delta}{\|\mathbf{g}_t\|_2 + \sqrt{\Delta_{t-1}}}$. If $c < 1$, we have

$$
\begin{aligned}
\mathrm{H}_{t-1} &= \delta \mathrm{I}_d + (\mathrm{S}_{t-1}^\top \mathrm{S}_{t-1})^{1/2} \\
&\succeq c(\|\mathbf{g}_t\|_2 \mathrm{I}_d + \sqrt{\Delta_{t-1}}\mathrm{I}_d + (\mathrm{S}_{t-1}^\top \mathrm{S}_{t-1})^{1/2}) \\
&\succeq c(\|\mathbf{g}_t\|_2 \mathrm{I}_d + (\Delta_{t-1}\mathrm{I}_d + \mathrm{S}_{t-1}^\top \mathrm{S}_{t-1})^{1/2}) \\
&\succeq c(\|\mathbf{g}_t\|_2 \mathrm{I}_d + (\mathrm{C}_{t-1}^\top \mathrm{C}_{t-1})^{1/2}) \\
&\succeq c(\mathrm{C}_{t-1}^\top \mathrm{C}_{t-1} + \|\mathbf{g}_t\|_2^2 \mathrm{I}_d)^{1/2} \\
&\succeq c(\mathrm{C}_t^\top \mathrm{C}_t)^{1/2}
\end{aligned}
$$

where the second inequality is due to $\sqrt{\Delta_t} + x \geq \sqrt{\Delta_t + x^2}$ for any $x \geq 0$ and the third inequality is due to Lemma 4. And in the other case $\delta \geq \sqrt{\Delta_{t-1}} + \|\mathbf{g}_t\|_2$, we have

$$
\begin{aligned}
\mathrm{H}_{t-1} &= \delta \mathrm{I}_d + (\mathrm{S}_{t-1}^\top \mathrm{S}_{t-1})^{1/2} \\
&\succeq \|\mathbf{g}_t\|_2 \mathrm{I}_d + \sqrt{\Delta_{t-1}}\mathrm{I}_d + (\mathrm{S}_{t-1}^\top \mathrm{S}_{t-1})^{1/2} \\
&\succeq \|\mathbf{g}_t\|_2 \mathrm{I}_d + (\Delta_{t-1}\mathrm{I}_d + \mathrm{S}_{t-1}^\top \mathrm{S}_{t-1})^{1/2} \\
&\succeq \|\mathbf{g}_t\|_2 \mathrm{I}_d + (\mathrm{C}_{t-1}^\top \mathrm{C}_{t-1})^{1/2} \\
&\succeq (\mathrm{C}_t^\top \mathrm{C}_t)^{1/2}.
\end{aligned}
$$

Thus for any $\delta > 0$, we have

$$
\mathrm{H}_{t-1} \succeq \min\left(1, \frac{\delta}{\|\mathbf{g}_t\|_2 + \sqrt{\Delta_{t-1}}}\right)(\mathrm{C}_t^\top \mathrm{C}_t)^{1/2}.
$$

Then we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \|f_t'(\boldsymbol{\beta}_t)\|_{\Psi_{t-1}^*}^2 \\
=&\sum_{t=1}^{T} 2\left\langle \mathbf{g}_t, (\mathrm{H}_{t-1})^{-1}\mathbf{g}_t \right\rangle \\
\leq&\sum_{t=1}^{T} 2\max\left(1, \frac{\|\mathbf{g}_t\|_2 + \sqrt{\Delta_{t-1}}}{\delta}\right)\left\langle \mathbf{g}_t, (\mathrm{G}_t^\dagger)^{1/2}\mathbf{g}_t \right\rangle \\
\leq&2\max\left(1, \frac{\max_{t\leq T}\|\mathbf{g}_t\|_2 + \sqrt{\Delta_T}}{\delta}\right)\sum_{t=1}^{T}\left\langle \mathbf{g}_t, (\mathrm{G}_t^\dagger)^{1/2}\mathbf{g}_t \right\rangle \\
\leq&4\max\left(1, \frac{\max_{t\leq T}\|\mathbf{g}_t\|_2 + \sqrt{\Delta_T}}{\delta}\right)\mathrm{tr}(\mathrm{G}_T^{1/2})
\end{aligned} \tag{12}
$$

where the last inequality is due to Lemma 3.

We complete the proof by substituting (11) and (12) into Lemma 1.

B.3 Proof of Theorem 2

According to Algorithm 2 and the property of frequent directions, we have

$$
(\mathrm{S}_t^\top \mathrm{S}_t)^{1/2} - (\mathrm{S}_{t-1}^\top \mathrm{S}_{t-1})^{1/2} + \sqrt{\sigma_t} V V^\top \succeq 0
$$

which has been proved in the proof of Lemma 1.

Let $\widetilde{\mathrm{G}}_t = \mathrm{S}_t^\top \mathrm{S}_t$. Considering the first term in the upper bound of Lemma 2, we have

$$
\begin{aligned}
&B_{\Psi_{t+1}}(\boldsymbol{\beta}^*, \boldsymbol{\beta}_{t+1}) - B_{\Psi_t}(\boldsymbol{\beta}^*, \boldsymbol{\beta}_{t+1}) \\
=&\frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, (\mathrm{H}_{t+1} - \mathrm{H}_t)(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}) \right\rangle \\
\leq&\frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, (\widetilde{\mathrm{G}}_{t+1}^{1/2} - \widetilde{\mathrm{G}}_t^{1/2})(\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}) \right\rangle \\
&+ \frac{1}{2}\left\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}, \sqrt{\sigma_{t+1}} V V^\top (\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}) \right\rangle \\
\leq&\frac{1}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}\|_2^2 \lambda_{\max}(\widetilde{\mathrm{G}}_{t+1}^{1/2} - \widetilde{\mathrm{G}}_t^{1/2} + \sqrt{\sigma_{t+1}} V V^\top) \\
\leq&\frac{1}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{t+1}\|_2^2 \,\mathrm{tr}(\widetilde{\mathrm{G}}_{t+1}^{1/2} - \widetilde{\mathrm{G}}_t^{1/2} + \sqrt{\sigma_{t+1}} V V^\top).
\end{aligned}
$$

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

$$
\begin{aligned}
&\sum_{t=1}^{T-1}[B_{\Psi_{t+1}}(\boldsymbol{\beta}^*, \boldsymbol{\beta}_{t+1}) - B_{\Psi_t}(\boldsymbol{\beta}^*, \boldsymbol{\beta}_{t+1})] + B_{\Psi_1}(\boldsymbol{\beta}^*, \boldsymbol{\beta}_1) \\
\leq&\frac{1}{2}\max_{t\leq T}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_t\|_2^2 \,\mathrm{tr}(\widetilde{\mathrm{G}}_T^{1/2} - \widetilde{\mathrm{G}}_1^{1/2}) \\
&+ \frac{\tau\sum_{t=2}^{\top}\sqrt{\sigma_t}}{2}\max_{t\leq T}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_t\|_2^2 + \frac{1}{2}\left\langle \boldsymbol{\beta}^*, \mathrm{H}_1\boldsymbol{\beta}^* \right\rangle \\
\leq&\frac{1}{2}\max_{t\leq T}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_t\|_2^2 \,\mathrm{tr}(\mathrm{G}_T^{1/2}) \\
&+ \frac{\tau\sum_{t=1}^{\top}\sqrt{\sigma_t}}{2}\max_{t\leq T}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_t\|_2^2 + \frac{\delta}{2}\|\boldsymbol{\beta}^*\|_2^2
\end{aligned} \tag{13}
$$

where we use Lemma 4 in the last inequality.

Before considering $\sum_{t=1}^{T}\|f_t'(\boldsymbol{\beta}_t)\|_{\Psi_t^*}^2$, we need derive the lower bound of H_t. If $\delta < \sqrt{\Delta_t}$, we have

$$
\begin{aligned}
\mathrm{H}_t = \delta \mathrm{I}_d + (\mathrm{S}_t^\top \mathrm{S}_t)^{1/2} &\succeq \frac{\delta(\sqrt{\Delta_t}\mathrm{I}_d + (\mathrm{S}_t^\top \mathrm{S}_t)^{1/2})}{\sqrt{\Delta_t}} \\
&\succeq \frac{\delta(\Delta_t \mathrm{I}_d + \mathrm{S}_t^\top \mathrm{S}_t)^{1/2}}{\sqrt{\Delta_t}} \succeq \frac{\delta}{\sqrt{\Delta_t}}(\mathrm{C}_t^\top \mathrm{C}_t)^{1/2}
\end{aligned}
$$

where the second inequality is due to $\sqrt{\Delta_t} + x \;>= \sqrt{\Delta_t + x^2}$ for $x \geq 0$ and the third inequality is due to Lemma 4. And in the other case $\delta \geq \sqrt{\Delta_t}$, we have

$$
\begin{aligned}
\mathrm{H}_t = \delta \mathrm{I}_d + (\mathrm{S}_t^\top \mathrm{S}_t)^{1/2} &\succeq \sqrt{\Delta_t}\mathrm{I}_d + (\mathrm{S}_t^\top \mathrm{S}_t)^{1/2} \\
&\succeq (\Delta_t \mathrm{I}_d + \mathrm{S}_t^\top \mathrm{S}_t)^{1/2} \succeq (\mathrm{C}_t^\top \mathrm{C}_t)^{1/2}.
\end{aligned}
$$

Thus for any $\delta > 0$, we have

$$
\mathrm{H}_t \succeq \min\left(1, \frac{\delta}{\sqrt{\Delta_t}}\right)(\mathrm{C}_t^\top \mathrm{C}_t)^{1/2}.
$$

Then we have

$$\begin{aligned}\sum_{t=1}^{T}\|f_t'(\boldsymbol{\beta}_t)\|_{\Psi_t^*}^2 &= \sum_{t=1}^{T} 2\left\langle \mathbf{g}_t, (\mathrm{H}_t)^{-1}\mathbf{g}_t\right\rangle \\ &\leq \sum_{t=1}^{T} 2\max\left(1, \frac{\sqrt{\Delta_t}}{\delta}\right)\left\langle \mathbf{g}_t, (\mathrm{G}_t^{\dagger})^{1/2}\mathbf{g}_t\right\rangle \\ &\leq 2\max\left(1, \frac{\sqrt{\Delta_T}}{\delta}\right)\sum_{t=1}^{T}\left\langle \mathbf{g}_t, (\mathrm{G}_t^{\dagger})^{1/2}\mathbf{g}_t\right\rangle \\ &\leq 4\max\left(1, \frac{\sqrt{\Delta_T}}{\delta}\right)\mathrm{tr}(\mathrm{G}_T^{1/2})\end{aligned} \tag{14}$$

where the last inequality is due to Lemma 3.

By combining (13) and (14), we have

$$\begin{aligned}R(T) \leq &\frac{\delta}{2\eta}\|\boldsymbol{\beta}_*\|_2^2 + \frac{1}{2\eta}\max_{t\leq T}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_t\|_2^2\,\mathrm{tr}(\mathrm{G}_T^{1/2}) \\ &+ 2\eta\max\left(1, \frac{\sqrt{\Delta_T}}{\delta}\right)\mathrm{tr}(\mathrm{G}_T^{1/2}) \\ &+ \frac{\tau\sum_{t=1}^{\top}\sqrt{\sigma_t}}{2\eta}\max_{t\leq T}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_t\|_2^2.\end{aligned}$$