Supplementary Materials to: “A General Regret Bound of Preconditioned Gradient Method for DNN Training”

Hongwei Yong Ying Sun Lei Zhang
The Hong Kong Polytechnic University
hongwei.yong@polyu.edu.hk, {csysun, cslzhang}@comp.polyu.edu.hk

The following materials are provided in this supplementary file:

- The proofs of Theorem 1 and Lemmas presented in the main paper (*cf.* Section 3 in the main paper).
- The detailed algorithms of SGDM_BK and AdamW_BK (*cf.* Section 4 in the main paper).
- The hyper-parameter settings of different optimizers and some ablation studies of the proposed method (*cf.* Section 5 in the main paper).

A. Proofs of Theorem 1 and Lemmas

Lemma 1 [1, 2]. *For any sequence of matrices $\{\boldsymbol{H}_t \succeq \mathbf{0}\}_{t=1}^T$, the regret of online mirror descent holds that*

$$R(T) \leq \frac{1}{2\eta}\sum\nolimits_{t=1}^{T}\left(||\boldsymbol{w}_t - \boldsymbol{w}^*||^2_{\boldsymbol{H}_t} - ||\boldsymbol{w}_{t+1} - \boldsymbol{w}^*||^2_{\boldsymbol{H}_t}\right) + \frac{\eta}{2}\sum\nolimits_{t=1}^{T}\left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_t}\right)^2. \tag{1}$$

If we further assume $D = \max_{t\leq T}||\boldsymbol{w}_t - \boldsymbol{w}^*||_2$, then we have

$$R(T) \leq \frac{D^2}{2\eta}Tr(\boldsymbol{H}_T) + \frac{\eta}{2}\sum\nolimits_{t=1}^{T}\left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_t}\right)^2. \tag{2}$$

The proof of **Lemma 1** can be found in [1, 2].

A1. Proof of Theorem 1

Before proving Theorem 1, let’s first prove the following Proposition 1.

Proposition 1. *For any $x_1 \geq 0$ and $x_2 \geq 0$, it holds that*

$$2\sqrt{x_2} + \frac{x_1 - x_2}{\sqrt{x_1}} \leq 2\sqrt{x_1}. \tag{3}$$

Proof. Let $f(x) = \sqrt{x}$. Because it is a concavity function, for any $x_1 \geq 0$ and $x_2 \geq 0$, we have

$$f(x_2) \leq f(x_1) + f(x_1)'(x_2 - x_1), \tag{4}$$

which is

$$\sqrt{x_2} \leq \sqrt{x_1} + \frac{x_2 - x_1}{2\sqrt{x_1}}. \tag{5}$$

Therefore, Eq. (3) holds. The proof is completed. ■

Theorem 1. *For any cone constraint $\Psi \subseteq \mathbb{R}^{d\times d}$, we define a guide function $F_T(\boldsymbol{S})$ on Ψ as*

$$F_T(\boldsymbol{S}) = \sum\nolimits_{t=1}^{T}(||\boldsymbol{g}_t||^*_{\boldsymbol{S}})^2, \tag{6}$$

and then define the matrix $\boldsymbol{H}_T$ as

$$\boldsymbol{H}_T = C_T \boldsymbol{S}_T, \quad \boldsymbol{S}_T = \arg \min_{\boldsymbol{S} \in \Psi, \boldsymbol{S} \succeq \boldsymbol{0}, Tr(\boldsymbol{S}) \leq 1} F_T(\boldsymbol{S}), \tag{7}$$

where $C_T = \sqrt{f(\boldsymbol{S}_T)}$*. The regret of online mirror descent holds that*

$$R(T) \leq (\frac{D^2}{2\eta} + \eta) C_T = (\frac{D^2}{2\eta} + \eta) \sqrt{\min_{\boldsymbol{S} \in \Psi, \boldsymbol{S} \succeq \boldsymbol{0}, Tr(\boldsymbol{S}) \leq 1} F_T(\boldsymbol{S})}. \tag{8}$$

Proof. According to **Lemma 1**, we have

$$R(T) \leq \frac{D^2}{2\eta} \text{Tr}(\boldsymbol{H}_T) + \frac{\eta}{2} \sum_{t=1}^{T} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_t} \right)^2. \tag{9}$$

For the first term on the right side of Eq. (9), according to the definition of $\boldsymbol{H}_T$ and $\boldsymbol{S}_T$, we have

$$\text{Tr}(\boldsymbol{H}_T) = \text{Tr}(C_T \boldsymbol{S}_T) = C_T \text{Tr}(\boldsymbol{S}_T) \leq C_T. \tag{10}$$

Then we only need to prove

$$\sum_{t=1}^{T} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_t} \right)^2 \leq 2C_T = 2\sqrt{\sum_{t=1}^{T} (||\boldsymbol{g}_t||^*_{\boldsymbol{S}_T})^2}. \tag{11}$$

Since

$$\sum_{t=1}^{T} (||\boldsymbol{g}_t||^*_{\boldsymbol{H}_T})^2 = \sum_{t=1}^{T} (||\boldsymbol{g}_t||^*_{C_T \boldsymbol{S}_T})^2 = \frac{1}{C_T} \sum_{t=1}^{T} (||\boldsymbol{g}_t||^*_{\boldsymbol{S}_T})^2 = \sqrt{\sum_{t=1}^{T} (||\boldsymbol{g}_t||^*_{\boldsymbol{S}_T})^2}, \tag{12}$$

in order to prove Eq. (11), we need to prove that

$$\sum_{t=1}^{T} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_t} \right)^2 \leq 2 \sum_{t=1}^{T} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_T} \right)^2. \tag{13}$$

The above equation can be proved by mathematical induction. For $T = 1$, $\left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_1} \right)^2 \leq 2 \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_1} \right)^2$ holds obviously. Suppose it holds that

$$\sum_{t=1}^{T-1} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_t} \right)^2 \leq 2 \sum_{t=1}^{T-1} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_{T-1}} \right)^2, \tag{14}$$

then we have

$$\begin{aligned} \sum_{t=1}^{T} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_t} \right)^2 &= \sum_{t=1}^{T-1} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_t} \right)^2 + \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_T} \right)^2 \\ &\leq 2 \sum_{t=1}^{T-1} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_{T-1}} \right)^2 + \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_T} \right)^2 \\ &= 2\sqrt{\sum_{t=1}^{T-1} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{S}_{T-1}} \right)^2} + \frac{1}{C_T} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{S}_T} \right)^2 \\ &\leq 2C_{T-1} + \frac{1}{C_T} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{S}_T} \right)^2. \end{aligned} \tag{15}$$

Meanwhile, we can prove that

$$
\begin{aligned}
C_T^2 - C_{T-1}^2 &= \sum_{t=1}^{T} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{S}_T}\right)^2 - \sum_{t=1}^{T-1} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{S}_{T-1}}\right)^2 \\
&= \sum_{t=1}^{T-1} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{S}_T}\right)^2 - \sum_{t=1}^{T-1} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{S}_{T-1}}\right)^2 + \left(||\boldsymbol{g}_T||^*_{\boldsymbol{S}_T}\right)^2 \\
&= \sum_{t=1}^{T-1} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{S}_T}\right)^2 - \min_{\boldsymbol{S}\in\Psi, \boldsymbol{S}\succeq \boldsymbol{0}, \mathrm{Tr}(\boldsymbol{S})\leq 1} \sum_{t=1}^{T-1} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{S}}\right)^2 + \left(||\boldsymbol{g}_T||^*_{\boldsymbol{S}_T}\right)^2 \\
&\geq \left(||\boldsymbol{g}_T||^*_{\boldsymbol{S}_T}\right)^2 .
\end{aligned}
\tag{16}
$$

Therefore, for Eq. (15), we have

$$
\begin{aligned}
\sum_{t=1}^{T} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_t}\right)^2 &\leq 2C_{T-1} + \frac{1}{C_T}\left(||\boldsymbol{g}_t||^*_{\boldsymbol{S}_T}\right)^2 \\
&\leq 2C_{T-1} + \frac{C_T^2 - C_{T-1}^2}{C_T}.
\end{aligned}
\tag{17}
$$

According to **Proposition 1** and let $x_1 = C_T^2$, $x_2 = C_{T-1}^2$, we have

$$
\begin{aligned}
\sum_{t=1}^{T} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{H}_t}\right)^2 &\leq 2C_{T-1} + \frac{C_T^2 - C_{T-1}^2}{C_T} \\
&\leq 2C_T \\
&= 2\sqrt{\sum_{t=1}^{T} \left(||\boldsymbol{g}_t||^*_{\boldsymbol{S}_T}\right)^2} \\
&= 2\sum_{t=1}^{T} (||\boldsymbol{g}_t||^*_{\boldsymbol{H}_T})^2 .
\end{aligned}
\tag{18}
$$

Now Eq. (13) is proved. Combining it with Eqs. (9), (10) and (11), we obtain the regret bound Eq. (8). The proof is completed. ■

A2. Proof of Lemma 3

We then prove **Lemma 3** in the main paper. To prove it, we first present the following **Propositions** $2 \sim 5$.

Proposition 2. *It holds that*

$$
\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i\boldsymbol{g}_i^\top \succeq \boldsymbol{g}\boldsymbol{g}^\top, \ \textit{where} \ \ \boldsymbol{g} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i. \tag{19}
$$

Proof. For any $\boldsymbol{x}$, it holds that $\boldsymbol{x}^\top(\frac{1}{n}\sum_{i=1}^n \boldsymbol{g}_i\boldsymbol{g}_i^\top)\boldsymbol{x} = \frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}^\top\boldsymbol{g}_i)^2$ and $\boldsymbol{x}^\top\boldsymbol{g}\boldsymbol{g}^\top\boldsymbol{x} = (\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}^\top\boldsymbol{g}_i)^2$. By using the convexity of $\alpha \mapsto \alpha^2$, we have $(\frac{1}{n}\sum_{i=1}^n \alpha_i)^2 \leq \frac{1}{n}\sum_{i=1}^n \alpha_i^2$. Then there is $(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}^\top\boldsymbol{g}_i)^2 \leq \frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}^\top\boldsymbol{g}_i)^2$, which means

$$
\boldsymbol{x}^\top(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i\boldsymbol{g}_i^\top)\boldsymbol{x} \geq \boldsymbol{x}^\top\boldsymbol{g}\boldsymbol{g}^\top\boldsymbol{x}.
$$

Hence, we have

$$
\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i\boldsymbol{g}_i^\top \succeq \boldsymbol{g}\boldsymbol{g}^\top.
$$

The proof is completed. ■

According to **Proposition 2**, we also have

$$\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\boldsymbol{g}_{ti}\boldsymbol{g}_{ti}^{\top} \succeq \sum_{t=1}^{T}\boldsymbol{g}_{t}\boldsymbol{g}_{t}^{\top}, \text{ where } \boldsymbol{g}_t = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_{ti}. \tag{20}$$

Proposition 3. *If $\boldsymbol{A} \succeq \boldsymbol{B}$, then for any $\boldsymbol{S} \succeq \boldsymbol{0}$, Tr$(\boldsymbol{SA}) \geq$ Tr$(\boldsymbol{SB})$.*

Proof. $\mathrm{Tr}(\boldsymbol{SA}) - \mathrm{Tr}(\boldsymbol{SB}) = \mathrm{Tr}(\boldsymbol{S}(\boldsymbol{A}-\boldsymbol{B}))$. Let $\boldsymbol{C} = \boldsymbol{A} - \boldsymbol{B} \succeq \boldsymbol{0}$, then $\boldsymbol{C}$ is PSD. We can find a matrix $\boldsymbol{Q}$, which meets $\boldsymbol{C} = \boldsymbol{QQ}^{\top}$. Therefore, $\mathrm{Tr}(\boldsymbol{SC}) = \mathrm{Tr}(\boldsymbol{SQQ}^{\top}) = \mathrm{Tr}(\boldsymbol{Q}^{\top}\boldsymbol{SQ}) = \sum_{i=1}^{d}\boldsymbol{q}_i^{\top}\boldsymbol{S}\boldsymbol{q}_i \geq 0$. The proof is completed. ■

Proposition 4. *If $x_i \geq 0$ and $y_i \geq 0$ for i=1,2,...,n, we have $\sum_i^n x_i y_i \leq (\sum_i^n x_i)(\sum_i^n y_i)$.*

Proof. $(\sum_i^n x_i)(\sum_i^n y_i) = (\sum_i^n x_i)(\sum_j^n y_j) = \sum_{i=j}^n x_i y_j + \sum_{i\neq j}^n x_i y_j \geq \sum_i^n x_i y_i$. The proof is completed. ■

The following proposition summarizes some properties of the Kronecker product, which can be found at [3].

Proposition 5 [3]. *Let $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{A}'$, $\boldsymbol{B}'$ be the matrices with appropriate dimensions. Then the following properties hold:*

(1) $(\boldsymbol{A}\otimes\boldsymbol{B})^{\top} = \boldsymbol{A}^{\top}\otimes\boldsymbol{B}^{\top}$, $(\boldsymbol{A}\otimes\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1}$ *(if $\boldsymbol{A}$ and $\boldsymbol{B}$ are invertible);*

(2) $(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{A}'\otimes\boldsymbol{B}') = (\boldsymbol{AA}')\otimes(\boldsymbol{BB}')$;

(3) if $\boldsymbol{A} \succeq \boldsymbol{0}$ and $\boldsymbol{B} \succeq \boldsymbol{0}$, $\boldsymbol{A}\otimes\boldsymbol{B} \succeq \boldsymbol{0}$;

(4) Tr$(\boldsymbol{A}\otimes\boldsymbol{B}) =$ Tr$(\boldsymbol{A})$Tr$(\boldsymbol{B})$.

We then prove **Lemma 3** in the main paper.

Lemma 3. *Denote by $\boldsymbol{L}_T = \sum_{t=1}^{T}\sum_{i=1}^{n}\boldsymbol{\delta}_{ti}\boldsymbol{\delta}_{ti}^{\top}$ and $\boldsymbol{R}_T = \sum_{t=1}^{T}\sum_{i=1}^{n}\boldsymbol{x}_{ti}\boldsymbol{x}_{ti}^{\top}$, there is*

$$\begin{aligned} F_T(\boldsymbol{S}) &\leq Tr\left((\boldsymbol{S}_1^{-1}\otimes\boldsymbol{S}_2^{-1})\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\boldsymbol{g}_{ti}\boldsymbol{g}_{ti}^{\top}\right) \\ &\leq \frac{1}{n}Tr(\boldsymbol{S}_1^{-1}\boldsymbol{L}_T)Tr(\boldsymbol{S}_2^{-1}\boldsymbol{R}_T). \end{aligned} \tag{21}$$

Proof. From **Proposition 2**, we have

$$\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\boldsymbol{g}_{ti}\boldsymbol{g}_{ti}^{\top} \succeq \sum_{t=1}^{T}\boldsymbol{g}_{t}\boldsymbol{g}_{t}^{\top}, \text{ where } \boldsymbol{g}_t = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_{ti}. \tag{22}$$

Together with **Proposition 3**, we have

$$\mathrm{Tr}\left((\boldsymbol{S}_1^{-1}\otimes\boldsymbol{S}_2^{-1})\sum_{t=1}^{T}\boldsymbol{g}_t\boldsymbol{g}_t^{\top}\right) \leq \mathrm{Tr}\left((\boldsymbol{S}_1^{-1}\otimes\boldsymbol{S}_2^{-1})\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\boldsymbol{g}_{ti}\boldsymbol{g}_{ti}^{\top}\right). \tag{23}$$

Finally, according to the properties of Kronecker Product in **Proposition 5**, we have

$$
\begin{aligned}
\text{Tr}\left((\boldsymbol{S}_1^{-1}\otimes\boldsymbol{S}_2^{-1})\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\boldsymbol{g}_{ti}\boldsymbol{g}_{ti}^{\top}\right) &= \text{Tr}\left((\boldsymbol{S}_1^{-1}\otimes\boldsymbol{S}_2^{-1})\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}(\boldsymbol{\delta}_{ti}\otimes\boldsymbol{x}_{ti})(\boldsymbol{\delta}_{ti}\otimes\boldsymbol{x}_{ti})^{\top}\right)\\
&= \text{Tr}\left((\boldsymbol{S}_1^{-1}\otimes\boldsymbol{S}_2^{-1})\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}(\boldsymbol{\delta}_{ti}\boldsymbol{\delta}_{ti}^{\top})\otimes(\boldsymbol{x}_{ti}\boldsymbol{x}_{ti}^{\top})\right)\\
&= \text{Tr}\left(\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}(\boldsymbol{S}_1^{-1}\boldsymbol{\delta}_{ti}\boldsymbol{\delta}_{ti}^{\top})\otimes(\boldsymbol{S}_2^{-1}\boldsymbol{x}_{ti}\boldsymbol{x}_{ti}^{\top})\right)\\
&= \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\text{Tr}(\boldsymbol{S}_1^{-1}\boldsymbol{\delta}_{ti}\boldsymbol{\delta}_{ti}^{\top})\text{Tr}(\boldsymbol{S}_2^{-1}\boldsymbol{x}_{ti}\boldsymbol{x}_{ti}^{\top}) \qquad (24)\\
&\leq \frac{1}{n}(\sum_{t=1}^{T}\sum_{i=1}^{n}\text{Tr}(\boldsymbol{S}_1^{-1}\boldsymbol{\delta}_{ti}\boldsymbol{\delta}_{ti}^{\top}))(\sum_{t=1}^{T}\sum_{i=1}^{n}\text{Tr}(\boldsymbol{S}_2^{-1}\boldsymbol{x}_{ti}\boldsymbol{x}_{ti}^{\top})), \quad (\textbf{Proposition 4})\\
&= \frac{1}{n}(\text{Tr}(\boldsymbol{S}_1^{-1}\sum_{t=1}^{T}\sum_{i=1}^{n}\boldsymbol{\delta}_{ti}\boldsymbol{\delta}_{ti}^{\top}))(\text{Tr}(\boldsymbol{S}_2^{-1}\sum_{t=1}^{T}\sum_{i=1}^{n}\boldsymbol{x}_{ti}\boldsymbol{x}_{ti}^{\top}))\\
&= \frac{1}{n}\text{Tr}(\boldsymbol{S}_1^{-1}\boldsymbol{L}_T)\text{Tr}(\boldsymbol{S}_2^{-1}\boldsymbol{R}_T).
\end{aligned}
$$

The proof is completed. ■

A3. Proof of Lemma 4

We first present the following **Propositions** $6 \sim 7$ before we prove **Lemma 4**.

Proposition 6. *Suppose $\boldsymbol{D}\in\mathbb{R}^{d\times d}$ is a diagonal matrix and $\boldsymbol{D}\succeq\boldsymbol{0}$, then*

$$
\min_{\boldsymbol{U}\in\mathbb{R}^{d\times d},\boldsymbol{U}\boldsymbol{U}^{\top}=\boldsymbol{I}} ||\boldsymbol{U}\boldsymbol{D}||_{12} = Tr(\boldsymbol{D}), \qquad (25)
$$

where $||\boldsymbol{A}||_{12}=\sum_i\sqrt{\sum_j A_{ij}^2}$ is the matrix L_{12}-norm, and $\boldsymbol{U}=\boldsymbol{I}$ is the optimal point.

Proof. For any orthogonal matrix $\boldsymbol{U}\in\mathbb{R}^{d\times d}$, denote by $\{\boldsymbol{u}_i\}_{i=1}^{d}$ the row vectors of $\boldsymbol{U}$, we have

$$
\begin{aligned}
\text{Tr}(\boldsymbol{D}) &= \text{Tr}(\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\top})\\
&= \sum_{i=1}^{d}\boldsymbol{u}_i^{\top}\boldsymbol{D}\boldsymbol{u}_i\\
&= \sum_{i=1}^{d}\langle\boldsymbol{D}\boldsymbol{u}_i,\boldsymbol{u}_i\rangle\\
&= \sum_{i=1}^{d}||\boldsymbol{D}\boldsymbol{u}_i||_2\text{cos}\langle\boldsymbol{D}\boldsymbol{u}_i,\boldsymbol{u}_i\rangle \qquad (26)\\
&\leq \sum_{i=1}^{d}||\boldsymbol{D}\boldsymbol{u}_i||_2\\
&= \sum_{i=1}^{d}||\boldsymbol{u}_i^{\top}\boldsymbol{D}||_2\\
&= ||\boldsymbol{U}\boldsymbol{D}||_{12}.
\end{aligned}
$$

When $\boldsymbol{U}=\boldsymbol{I}$, $\cos\langle\boldsymbol{D}\boldsymbol{u}_i,\boldsymbol{u}_i\rangle=1$ for $i=1,2,...,d$, and the equality holds. The proof is completed. ■

Proposition 7. *Suppose $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, $\boldsymbol{A} \succeq \boldsymbol{0}$, $\boldsymbol{D} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{D}$ is a diagonal matrix, then*

$$\arg \min_{\boldsymbol{D}\succeq\boldsymbol{0},Tr(\boldsymbol{D})\leq 1} Tr(\boldsymbol{D}^{-1}\boldsymbol{A}) = \frac{1}{||\boldsymbol{A}^{\frac{1}{2}}||_{12}} Diag((\boldsymbol{A}^{\frac{1}{2}})^{\odot 2}\boldsymbol{1})^{\odot\frac{1}{2}} \tag{27}$$

and

$$\min_{\boldsymbol{D}\succeq\boldsymbol{0},Tr(\boldsymbol{D})\leq 1} Tr(\boldsymbol{D}^{-1}\boldsymbol{A}) = ||\boldsymbol{A}^{\frac{1}{2}}||_{12}^2. \tag{28}$$

Proof. Let $\boldsymbol{D} = \mathrm{Diag}(\boldsymbol{d})$ and $\boldsymbol{B} = \boldsymbol{A}^{\frac{1}{2}}$, then we have

$$\min_{\boldsymbol{D}\succeq\boldsymbol{0},\mathrm{Tr}(\boldsymbol{D})\leq 1} \mathrm{Tr}(\boldsymbol{D}^{-1}\boldsymbol{A}) = \min_{\boldsymbol{D}\succeq\boldsymbol{0},\mathrm{Tr}(\boldsymbol{D})\leq 1} \mathrm{Tr}(\boldsymbol{D}^{-1}\boldsymbol{B}\boldsymbol{B}^\top) = \sum_{i=1}^d\sum_{j=1}^d \frac{b_{ij}^2}{d_i}. \tag{29}$$

By introducing multipliers $\boldsymbol{\lambda} \succeq \boldsymbol{0}$ and $\theta \geq 0$, we can write the Lagrangian of the constrained problem in Eq. (29) as

$$L(\boldsymbol{d},\boldsymbol{\lambda},\theta) = \sum_{i=1}^d\sum_{j=1}^d \frac{b_{ij}^2}{d_i} - \langle\boldsymbol{\lambda},\boldsymbol{d}\rangle + \theta(\boldsymbol{1}^\top\boldsymbol{d} - 1). \tag{30}$$

Obviously, $d_i \neq 0$. According to the complementarity conditions, we know $\lambda_i = 0$. Then, we have

$$d_i = \theta^{-\frac{1}{2}}(\sum_{j=1}^d b_{ij}^2)^{\frac{1}{2}}. \tag{31}$$

With the constraint $\boldsymbol{1}^\top\boldsymbol{d} \leq 1$, we can choose a proper θ so that

$$d_i = \frac{(\sum_{j=1}^d b_{ij}^2)^{\frac{1}{2}}}{\sum_{i=1}^d(\sum_{j=1}^d b_{ij}^2)^{\frac{1}{2}}} \tag{32}$$

meets the constraint. Therefore, $\boldsymbol{d} = \frac{(\boldsymbol{B}^{\odot 2}\boldsymbol{1})^{\odot\frac{1}{2}}}{\boldsymbol{1}^\top(\boldsymbol{B}^{\odot 2}\boldsymbol{1})^{\odot\frac{1}{2}}}$, $\boldsymbol{D} = \frac{1}{||\boldsymbol{A}^{\frac{1}{2}}||_{12}} Diag((\boldsymbol{A}^{\frac{1}{2}})^{\odot 2}\boldsymbol{1})^{\odot\frac{1}{2}})$ and the minimum value of the objective function is

$$\begin{aligned}\mathrm{Tr}(\boldsymbol{D}^{-1}\boldsymbol{A}) &= \sum_{i=1}^d\sum_{j=1}^d \frac{b_{ij}^2}{d_i} \\ &= (\sum_{i=1}^d(\sum_{j=1}^d b_{ij}^2)^{\frac{1}{2}})\sum_{i=1}^d\sum_{j=1}^d \frac{b_{ij}^2}{(\sum_{j=1}^d b_{ij}^2)^{\frac{1}{2}}} \\ &= (\sum_{i=1}^d(\sum_{j=1}^d b_{ij}^2)^{\frac{1}{2}})^2 \\ &= ||\boldsymbol{B}||_{12}^2 \\ &= ||\boldsymbol{A}^{\frac{1}{2}}||_{12}^2.\end{aligned} \tag{33}$$

The proof is completed. ■

Then we prove **Lemma 4** in the main paper.

Lemma 4. *If $\boldsymbol{A} \succ \boldsymbol{0}$, we have*

$$\arg \min_{\boldsymbol{S}\succeq\boldsymbol{0},Tr(\boldsymbol{S})\leq 1} Tr(\boldsymbol{S}^{-1}\boldsymbol{A}) = \boldsymbol{A}^{\frac{1}{2}}/Tr(\boldsymbol{A}^{\frac{1}{2}}). \tag{34}$$

Proof. Because

$$\begin{aligned}\min_{\boldsymbol{S}\succeq\boldsymbol{0},\mathrm{Tr}(\boldsymbol{S})\leq 1} \mathrm{Tr}(\boldsymbol{S}^{-1}\boldsymbol{A}) &= \min_{\boldsymbol{D}=\mathrm{Diag}(\boldsymbol{d}),\boldsymbol{d}\succeq\boldsymbol{0},\boldsymbol{1}^\top\boldsymbol{d}\leq 1,\boldsymbol{U}\boldsymbol{U}^\top=\boldsymbol{I}} \mathrm{Tr}(\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{U}^\top\boldsymbol{A}) \\ &= \min_{\boldsymbol{U}\boldsymbol{U}^\top=\boldsymbol{I}} \min_{\boldsymbol{D}=\mathrm{Diag}(\boldsymbol{d}),\boldsymbol{d}\succeq\boldsymbol{0}\boldsymbol{1}^\top\boldsymbol{d}\leq 1,} \mathrm{Tr}(\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{U}^\top\boldsymbol{A}),\end{aligned} \tag{35}$$

we can find the optimal diagonal matrix $\boldsymbol{D}$ and orthogonal matrix $\boldsymbol{U}$ to obtain the optimal $\boldsymbol{S}$ by $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$. We first fix $\boldsymbol{U}$ to find the optimal $\boldsymbol{D}$. Since

$$\min_{\boldsymbol{D}=\text{Diag}(\boldsymbol{d}),\boldsymbol{d}\succeq\boldsymbol{0},\boldsymbol{1}^\top\boldsymbol{d}\leq 1,} \text{Tr}(\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{U}^\top\boldsymbol{A}) = \min_{\boldsymbol{D}=\text{Diag}(\boldsymbol{d}),\boldsymbol{d}\succeq\boldsymbol{0},\boldsymbol{1}^\top\boldsymbol{d}\leq 1,} \text{Tr}(\boldsymbol{D}^{-1}\boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{U}), \tag{36}$$

according to **Proposition 7**, we know the optimal $\boldsymbol{D}$ is

$$\boldsymbol{D} = \frac{1}{||\boldsymbol{U}^\top\boldsymbol{A}^{\frac{1}{2}}||_{12}} Diag((\boldsymbol{U}^\top\boldsymbol{A}^{\frac{1}{2}})^{\odot 2}\boldsymbol{1})^{\odot\frac{1}{2}}. \tag{37}$$

Meanwhile, we have

$$\min_{\boldsymbol{D}=\text{Diag}(\boldsymbol{d}),\boldsymbol{d}\succeq\boldsymbol{0},\boldsymbol{1}^\top\boldsymbol{d}\leq 1,} \text{Tr}(\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{U}^\top\boldsymbol{A}) = ||\boldsymbol{U}^\top\boldsymbol{A}^{\frac{1}{2}}||_{12}^2. \tag{38}$$

We then minimize Eq (38) w.r.t. $\boldsymbol{U}$. Suppose the SVD decomposition of $\boldsymbol{A}$ is $\boldsymbol{A} = \boldsymbol{U}_A\boldsymbol{D}_A\boldsymbol{U}_A^\top$, there is

$$||\boldsymbol{U}^\top\boldsymbol{A}^{\frac{1}{2}}||_{12}^2 = ||\boldsymbol{U}^\top\boldsymbol{U}_A\boldsymbol{D}_A^{\frac{1}{2}}\boldsymbol{U}_A^\top||_{12}^2 = ||\boldsymbol{U}^\top\boldsymbol{U}_A\boldsymbol{D}_A^{\frac{1}{2}}||_{12}^2. \tag{39}$$

According to **Proposition 6**, we know that when $\boldsymbol{U}^\top\boldsymbol{U}_A = \boldsymbol{I}$, *i.e.*, $\boldsymbol{U} = \boldsymbol{U}_A$, Eq (39) reaches its minimal value. Therefore, we have the optimal $\boldsymbol{D}$ as follows

$$\begin{aligned}\boldsymbol{D} &= \frac{1}{||\boldsymbol{U}_A^\top\boldsymbol{A}^{\frac{1}{2}}||_{12}} Diag((\boldsymbol{U}_A^\top\boldsymbol{A}^{\frac{1}{2}})^{\odot 2}\boldsymbol{1})^{\odot\frac{1}{2}} \\ &= \frac{1}{||\boldsymbol{U}_A^\top\boldsymbol{U}_A\boldsymbol{D}_A^{\frac{1}{2}}\boldsymbol{U}_A^\top||_{12}} Diag((\boldsymbol{U}_A^\top\boldsymbol{U}_A\boldsymbol{D}_A^{\frac{1}{2}}\boldsymbol{U}_A^\top)^{\odot 2}\boldsymbol{1})^{\odot\frac{1}{2}} \\ &= \frac{1}{||\boldsymbol{D}_A^{\frac{1}{2}}\boldsymbol{U}_A^\top||_{12}} Diag((\boldsymbol{D}_A^{\frac{1}{2}}\boldsymbol{U}_A^\top)^{\odot 2}\boldsymbol{1})^{\odot\frac{1}{2}} \\ &= \frac{1}{||\boldsymbol{D}_A^{\frac{1}{2}}||_{12}} Diag((\boldsymbol{D}_A^{\frac{1}{2}})^{\odot 2}\boldsymbol{1})^{\odot\frac{1}{2}} \\ &= \frac{1}{\text{Tr}(\boldsymbol{D}_A^{\frac{1}{2}})} Diag(\boldsymbol{D}_A^{\frac{1}{2}}).\end{aligned} \tag{40}$$

Then, the optimal $\boldsymbol{S}$ is

$$\begin{aligned}\boldsymbol{S} &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top \\ &= \frac{1}{\text{Tr}(\boldsymbol{D}_A^{\frac{1}{2}})}\boldsymbol{U}_A Diag(\boldsymbol{D}_A^{\frac{1}{2}})\boldsymbol{U}_A^\top \\ &= \frac{1}{\text{Tr}(\boldsymbol{U}_A\boldsymbol{D}_A^{\frac{1}{2}}\boldsymbol{U}_A^\top)}\boldsymbol{U}_A Diag(\boldsymbol{D}_A^{\frac{1}{2}})\boldsymbol{U}_A^\top \\ &= \frac{1}{\text{Tr}(\boldsymbol{A}^{\frac{1}{2}})}\boldsymbol{A}^{\frac{1}{2}}.\end{aligned} \tag{41}$$

The proof is completed. ■

A4. Proof of Lemma 2

Finally, we prove **Lemma 2** in the main paper.

Lemma 2. *Suppose Ψ is the set of either diagonal matrices or full-matrices, according to the definition of $\boldsymbol{S}_T$ and $\boldsymbol{H}_T$ in Eq. (7), we have*

$$\boldsymbol{H}_T = Diag\big((\sum_{t=1}^T \boldsymbol{g}_t \odot \boldsymbol{g}_t)^{\odot\frac{1}{2}}\big), \quad \boldsymbol{H}_T = \big(\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top\big)^{\frac{1}{2}}. \tag{42}$$

Proof. Because $\boldsymbol{H}_T = C_T \boldsymbol{S}_T$, we see that we only need to solve $\boldsymbol{S}_T$. We first prove the case when Ψ is the set of diagonal matrices. Let $\boldsymbol{S} = \text{Diag}(\boldsymbol{s})$ and $\boldsymbol{H} = \text{Diag}(\boldsymbol{h})$, where $\boldsymbol{s}$ and $\boldsymbol{h}$ are the diagonal vectors of $\boldsymbol{S}$ and $\boldsymbol{H}$, respectively, we have

$$\boldsymbol{s}_T = \arg \min_{\boldsymbol{s} \succeq 0, \mathbf{1}^\top \boldsymbol{s} \leq 1} \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{g_{ti}^2}{s_i}, \tag{43}$$

where $\boldsymbol{s} \succeq 0$ means all the coefficients of vector $\boldsymbol{s}$ are non-negative. By introducing multipliers $\boldsymbol{\lambda} \succeq \mathbf{0}$ and $\theta \geq 0$, we can have the Lagrangian of the above constrained optimization problem:

$$L(\boldsymbol{s}, \boldsymbol{\lambda}, \theta) = \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{g_{ti}^2}{s_i} - \langle \boldsymbol{\lambda}, \boldsymbol{s} \rangle + \theta(\mathbf{1}^\top \boldsymbol{s} - 1). \tag{44}$$

Taking the partial derivatives w.r.t. s_i, we have

$$\frac{\partial L(\boldsymbol{s}, \boldsymbol{\lambda}, \theta)}{\partial s_i} = -\sum_{t=1}^{T} \frac{g_{ti}^2}{s_i^2} - \lambda_i + \theta = 0. \tag{45}$$

Obviously, $s_i \neq 0$, and according to the complementarity conditions, we know $\lambda_i = 0$. Then, we have

$$s_i = \theta^{-\frac{1}{2}} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}. \tag{46}$$

With the constraint $\mathbf{1}^\top \boldsymbol{s} \leq 1$, we can choose a proper θ so that

$$s_{Ti} = \frac{(\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}}{\sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}} \tag{47}$$

meets the constraint. Meanwhile, we can derive that

$$\begin{aligned} C_T &= \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{d} \frac{g_{ti}^2}{s_{Ti}}} \\ &= \sqrt{\sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}} \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{g_{ti}^2}{(\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}}} \\ &= \sqrt{\sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}} \sum_{i=1}^{d} \frac{\sum_{t=1}^{T} g_{ti}^2}{(\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}}} \\ &= \sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}. \end{aligned} \tag{48}$$

Therefore,

$$\boldsymbol{h}_{Ti} = C_T \boldsymbol{s}_{Ti} = \sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}} \frac{(\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}}{\sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}} = (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}, \tag{49}$$

and finally we have $\boldsymbol{H}_T = \text{Diag}\big(\big(\sum_{t=1}^{T} \boldsymbol{g}_t \odot \boldsymbol{g}_t \big)^{\odot \frac{1}{2}} \big)$.

When Ψ is the set of full matrices, we have

$$\boldsymbol{S}_T = \arg \min_{\boldsymbol{S} \succeq \mathbf{0}, \text{Tr}(\boldsymbol{S}) \leq 1} \sum_{t=1}^{T} (||\boldsymbol{g}_t||_{\boldsymbol{S}}^*)^2 = \arg \min_{\boldsymbol{S} \succeq \mathbf{0}, \text{Tr}(\boldsymbol{S}) \leq 1} \text{Tr} \left(\boldsymbol{S}^{-1} \sum_{t=1}^{T} \boldsymbol{g}_t \boldsymbol{g}_t^\top \right). \tag{50}$$

Algorithm 1: SGDM_BK

```
Input: $T_s, T_{ir}, \alpha, \epsilon, \beta, \boldsymbol{W}_0, \boldsymbol{L}_0, \boldsymbol{R}_0, \eta$
Output: $\boldsymbol{W}_T$
for t=1:T do
    $\boldsymbol{X}_t = [\boldsymbol{x}_{ti}]_{i=1}^n, \Delta_t = [\boldsymbol{\delta}_{ti}]_{i=1}^n, \boldsymbol{G}_t = \nabla_{\boldsymbol{W}_t}\mathcal{L}$;
    if $t\%T_s = 0$ then
        $\boldsymbol{L}_t = \alpha\boldsymbol{L}_{t-1} + (1-\alpha)\Delta_t\Delta_t^\top$;
        $\boldsymbol{R}_t = \alpha\boldsymbol{R}_{t-1} + (1-\alpha)\boldsymbol{X}_t\boldsymbol{X}_t^\top$
    else
        $\boldsymbol{L}_t = \boldsymbol{L}_{t-1}, \boldsymbol{R}_t = \boldsymbol{R}_{t-1}$
    end
    if $t\%T_{ir} = 0$ then
        Compute $\lambda_{max}^L$ and $\lambda_{max}^R$ by Power Iteration;
        Compute $\widehat{\boldsymbol{L}}_t = (\boldsymbol{L}_t + \lambda_{max}^L\epsilon\boldsymbol{I})^{-\frac{1}{2}}$ and
          $\widehat{\boldsymbol{R}}_t = (\boldsymbol{R}_t + \lambda_{max}^R\epsilon\boldsymbol{I})^{-\frac{1}{2}}$ by Schur-Newton Iteration;
    else
        $\widehat{\boldsymbol{L}}_t = \widehat{\boldsymbol{L}}_{t-1}$ and $\widehat{\boldsymbol{R}}_t = \widehat{\boldsymbol{L}}_{t-1}$
    end
    $\widehat{\boldsymbol{G}}_t = \widehat{\boldsymbol{L}}_t\boldsymbol{G}_t\widehat{\boldsymbol{R}}_t,\ \tilde{\boldsymbol{G}}_t = \widehat{\boldsymbol{G}}_t\frac{||\boldsymbol{G}_t||_2}{||\widehat{\boldsymbol{G}}_t||_2}$ $\boldsymbol{M}_t = \beta\boldsymbol{M}_t + (1-\beta)\tilde{\boldsymbol{G}}_t$;
    $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta\boldsymbol{M}_t$;
end
```

Algorithm 2: AdamW_BK

```
Input: $T_s, T_{ir}, \alpha, \epsilon, \epsilon', \beta_1, \beta_2, \boldsymbol{W}_0, \boldsymbol{L}_0, \boldsymbol{R}_0, \eta$
Output: $\boldsymbol{W}_T$
for t=1:T do
    $\boldsymbol{X}_t = [\boldsymbol{x}_{ti}]_{i=1}^n, \Delta_t = [\boldsymbol{\delta}_{ti}]_{i=1}^n, \boldsymbol{G}_t = \nabla_{\boldsymbol{W}_t}\mathcal{L}$;
    if $t\%T = 0$ then
        $\boldsymbol{L}_t = \alpha\boldsymbol{L}_{t-1} + (1-\alpha)\Delta_t\Delta_t^\top$;
        $\boldsymbol{R}_t = \alpha\boldsymbol{R}_{t-1} + (1-\alpha)\boldsymbol{X}_t\boldsymbol{X}_t^\top$
    else
        $\boldsymbol{L}_t = \boldsymbol{L}_{t-1}, \boldsymbol{R}_t = \boldsymbol{R}_{t-1}$
    end
    if $t\%T_{ir} = 0$ then
        Compute $\lambda_{max}^L$ and $\lambda_{max}^R$ by Power Iteration;
        Compute $\widehat{\boldsymbol{L}}_t = (\boldsymbol{L}_t + \lambda_{max}^L\epsilon\boldsymbol{I})^{-\frac{1}{2}}$ and
          $\widehat{\boldsymbol{R}}_t = (\boldsymbol{R}_t + \lambda_{max}^R\epsilon\boldsymbol{I})^{-\frac{1}{2}}$ by Schur-Newton Iteration;
    else
        $\widehat{\boldsymbol{L}}_t = \widehat{\boldsymbol{L}}_{t-1}$ and $\widehat{\boldsymbol{R}}_t = \widehat{\boldsymbol{L}}_{t-1}$
    end
    $\widehat{\boldsymbol{G}}_t = \widehat{\boldsymbol{L}}_t\boldsymbol{G}_t\widehat{\boldsymbol{R}}_t,\ \tilde{\boldsymbol{G}}_t = \widehat{\boldsymbol{G}}_t\frac{||\boldsymbol{G}_t||_2}{||\widehat{\boldsymbol{G}}_t||_2}$
    $\boldsymbol{M}_t = \beta_1\boldsymbol{M}_{t-1} + (1-\beta_1)\tilde{\boldsymbol{G}}_t$;
    $\boldsymbol{V}_t = \beta_2\boldsymbol{V}_{t-1} + (1-\beta_2)\tilde{\boldsymbol{G}}_t \odot \tilde{\boldsymbol{G}}_t$;
    $\widehat{\boldsymbol{M}}_t = \frac{\boldsymbol{M}_t}{1-\beta_1^\top},\ \widehat{\boldsymbol{V}}_t = \frac{\boldsymbol{V}_t}{1-\beta_2^\top}$;
    $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta\frac{\widehat{\boldsymbol{M}}_t}{\sqrt{\widehat{\boldsymbol{V}}_t}+\epsilon'}$;
end
```

According to **Lemma 4**, we have

$$\boldsymbol{S}_T = (\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top)^{\frac{1}{2}}/\mathrm{Tr}\left((\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top)^{\frac{1}{2}}\right). \tag{51}$$

Meanwhile, there is

$$\begin{aligned} C_T &= \sqrt{\sum_{t=1}^T(||\boldsymbol{g}_t||_{\boldsymbol{S}_T}^*)^2} \\ &= \sqrt{\mathrm{Tr}\left(\boldsymbol{S}_T^{-1}\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top\right)} \\ &= \sqrt{\mathrm{Tr}\left((\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top)^{-\frac{1}{2}}\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top\right)\mathrm{Tr}\left((\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top)^{\frac{1}{2}}\right)} \\ &= \mathrm{Tr}\left((\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top)^{\frac{1}{2}}\right). \end{aligned} \tag{52}$$

Therefore,

$$\begin{aligned} \boldsymbol{H}_T &= C_T\boldsymbol{S}_T \\ &= \mathrm{Tr}\left((\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top)^{\frac{1}{2}}\right)(\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top)^{\frac{1}{2}}/\mathrm{Tr}\left((\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top)^{\frac{1}{2}}\right) \\ &= (\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top)^{\frac{1}{2}}. \end{aligned} \tag{53}$$

The proof is completed. ■

B. The Algorithms of SGDM_BK and AdamW_BK

By embedding our proposed AdaBK into the commonly used algorithms SGDM and AdamW, we obtain two new optimizers, namely **SGDM_BK and AdamW_BK**, which are described in **Algorithm 1** and **Algorithm 2**, respectively.

Table 1. Settings of learning rate (LR), weight decay (WD) and WD methods for different optimizers on CIFAR10/100. Here, the WD methods include L_2 regularization weight decay (L_2 in short) and weight decouple (decouple in short).

Optimizer	SGDM	AdamW	Adagrad	RAdam	Adabelief	Shampoo	KFAC	SGDM_BK	AdamW_BK
LR	0.1	0.001	0.01	0.001	0.001	0.001	0.01	0.05	0.001
WD	0.0005	0.5	0.0005	0.5	0.5	0.0005	0.005	0.001	0.5
WD method	L_2	decouple	L_2	decouple	decouple	L_2	decouple	L_2	decouple

Table 2. Settings of learning rate (LR), weight decay (WD) and WD methods (L_2 and decouple) for different optimizers on ImageNet.

Optimizer		SGDM	AdamW	Adagrad	RAdam	Adabelief	Shampoo	KFAC	SGDM_BK	AdamW_BK
ResNet18	LR	0.1	0.001	0.01	0.001	0.001	0.001	0.01	0.1	0.001
	WD	0.0001	0.1	0.0001	0.1	0.05	0.0001	0.001	0.0001	0.1
ResNet50	LR	0.1	0.001	0.01	0.001	0.001	0.001	0.01	0.05	0.0005
	WD	0.0001	0.1	0.0001	0.05	0.1	0.0001	0.001	0.0003	0.3
WD method		L_2	decouple	L_2	decouple	decouple	L_2	decouple	L_2	decouple

Table 3. Testing accuracies (%) of DNNs with different dampening ϵ.

$T_s = 50$ and $T_{ir} = 500$							
	ϵ	0.1	0.01	0.001	0.0001	0.00001	0.000001
ResNet18	SGDM_BK	78.60 ± .23	79.26 ± .12	79.21 ± .22	79.53 ± .22	79.35 ± .29	79.36 ± .22
	AdamW_BK	77.80 ± .23	78.38 ± .10	78.43 ± .15	78.61 ± .26	78.78 ± .15	78.55 ± .20
ResNet50	SGDM_BK	79.89 ± .31	80.66 ± .30	80.89 ± .27	81.00 ± .17	81.10 ± .19	81.15 ± .23
	AdamW_BK	79.57 ± .15	80.11 ± .21	80.10 ± .14	79.97 ± .31	80.13 ± .15	80.11 ± .19
$T_s = 200$ and $T_{ir} = 2000$							
	ϵ	0.1	0.01	0.001	0.0001	0.00001	0.000001
ResNet18	SGDM_BK	78.47 ± .17	78.97 ± .22	79.31 ± .23	79.24 ± .05	79.30 ± .07	79.17 ± .16
	AdamW_BK	77.84 ± .14	78.39 ± .18	78.63 ± .16	78.39 ± .17	78.66 ± .34	78.57 ± .29
ResNet50	SGDM_BK	80.07 ± .16	80.80 ± .09	80.94 ± .30	80.95 ± .31	81.26 ± .20	81.04 ± .15
	AdamW_BK	79.36 ± .11	79.78 ± .16	80.06 ± .23	80.11 ± .05	80.15 ± .19	79.95 ± .29

Table 4. Testing accuracies (%) and training time (h) with different updating intervals.

ResNet18									
Baseline		T_s T_{ir}	5 50	10 100	20 200	50 500	100 1000	200 2000	500 5000
SGDM	77.20 ± .30	SGDM_BK	79.35 ± .20	79.23 ± .18	79.37 ± .23	79.47 ± .24	79.37 ± .11	79.30 ± .07	79.29 ± .13
Time	1.12	Time	3.66	2.85	2.08	1.62	1.46	1.39	1.34
AdamW	77.23 ± .10	AdamW_BK	78.43 ± .17	78.58 ± .32	78.36 ± .15	78.38 ± .23	78.62 ± .16	78.66 ± .34	78.53 ± .10
Time	1.16	Time	3.68	2.87	2.10	1.65	1.49	1.42	1.36
ResNet50									
SGDM	77.78 ± .43	SGDM_BK	81.21 ± .21	81.09 ± .18	81.10 ± .18	81.06 ± .14	80.86 ± .10	81.26 ± .20	81.00 ± .26
Time	3.78	Time	7.57	6.35	5.23	4.58	4.33	4.21	4.16
AdamW	78.10 ± .17	AdamW_BK	80.02 ± .07	80.08 ± .18	80.00 ± .13	80.07 ± .29	80.06 ± .13	80.15 ± .19	80.06 ± .30
Time	3.83	Time	7.57	6.36	5.26	4.60	4.38	4.26	4.20

C. Hyper-parameter Settings and Ablation Studies

We first give the hyper-parameter settings of all optimizers in the image classification task, then give the tuning results of the hyper-parameters of AdaBK, including the dampening parameter ϵ and the statistics updating intervals T_s and T_{ir}. Meanwhile, we provide some ablation studies of SGDM_BK and AdamW_BK on memory usage and training time.

The CIFAR100 dataset is employed for the ablation studies of AdaBK. The initial learning rate (LR) and weight decay (WD) of SGDM_BK and AdamW are 0.05 and 0.001, and 0.001 and 0.5, respectively. The training schedule is the same as that in the main paper. Our experiments are conducted with NVIDIA GeForce RTX 2080Ti GPUs under the PyTorch 1.11 framework. All the experiments, if not specified, are repeated 4 times, with the performance reported in a "mean ± std" format and the training time reported in average.

LR and WD Settings. We first introduce the hyper-parameters of different optimizers we evaluated in **Section 5** of our main paper. We tune the LR and WD of all optimizers by grid search. On CIFAR100/10, we tune the LR in $\{1e^{-4}, 5e^{-4}5e^{-4}, 1e^{-3}, 5e^{-3}, 1e^{-2}, 5e^{-2}, 0.1\}$ and WD in $\{1e^{-4}, 3e^{-4}, 5e^{-4}, 1e^{-3}, 3e^{-3}, 5e^{-3}, 1e^{-2}, 3e^{-2}, 5e^{-2}, 0.1, 0.3, 0.5\}$, and choose the best combination of them for all optimizers. The final settings are described in Table 1. While for SGDM_BK, we use a learning rate of 0.1 and weight decay of 0.0005 for DenseNet. On ImageNet, we refer to the strategies in [4] to tune the LR and WD on ResNet18 and ResNet50, respectively.

The final settings are described in Table 2. For Swin transformer in ImageNet, AdamW uses the default LR (0.001) and WD (0.05) of MMClassification, while AdamW_BK uses an LR of 0.002 and WD of 0.025.

Dampening. Table 3 shows the testing results for different dampening parameters under different updating intervals, *i.e.*, $T_s = 50$ with $T_{ir} = 500$, and $T_s = 200$ with $T_{ir} = 2000$. From the testing results, we can see that our optimizer is relatively stable for different choices of dampening. The maximum performance fluctuation does not exceed 1.19%. We then set ϵ to 0.00001 in the experiments.

Statistics Updating Intervals. The testing accuracies and training time of different settings of intervals T_s, T_{ir} are reported in Table 4. In these experiments, we set the dampening parameter ϵ to 0.00001. We can see that the increase of statistics update interval can greatly reduce the time required for training DNNs while keeping similar accuracy. We then set $T_s = 200$ with $T_{ir} = 2000$ in the experiments.

References

[1] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011. 1

[2] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018. 1

[3] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008. 4

[4] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in neural information processing systems*, 33:18795–18806, 2020. 11