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 $H_T \succeq ... \succeq H_1 \succeq 0$, the regret of online mirror descent holds that
\[
R(T) \leq \frac{1}{2\eta} \sum_{t=1}^{T} (||w_t - w^*||_{H_t}^2 - ||w_{t+1} - w^*||_{H_t}^2) + \frac{\eta}{2} \sum_{t=1}^{T} (||g_t||_{H_t}^2).
\] (1)

If we further assume $D = \max_{t \leq T} ||w_t - w^*||_2$, then we have
\[
R(T) \leq \frac{D^2}{2\eta} \text{tr}(H_T) + \frac{\eta}{2} \sum_{t=1}^{T} (||g_t||_{H_t}^2).\] (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
\[
2\sqrt{x_2} + \frac{x_1 - x_2}{\sqrt{x_1}} \leq 2\sqrt{x_1}.
\] (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
\[
f(x_2) \leq f(x_1) + f(x_1)'(x_2 - x_1),
\] (4)

which is
\[
\sqrt{x_2} \leq \sqrt{x_1} + \frac{x_2 - x_1}{2\sqrt{x_1}}.
\] (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 $F_T(S)$ on $\Psi$ as
\[
F_T(S) = \sum_{t=1}^{T} ||g_t||_S^2,
\] (6)
and then define the matrix $H_T$ as
\[ H_T = C_T S_T, \quad S_T = \arg \min_{S \in \Psi, \text{Tr}(S) \leq 1} F_T(S), \]
where $C_T = \sqrt{F_T(S_T)}$. The regret of online mirror descent holds that
\[ R(T) \leq \left( \frac{D^2}{2\eta} + \eta \right) C_T = \left( \frac{D^2}{2\eta} + \eta \right) \min_{S \in \Psi, \text{Tr}(S) \leq 1} F_T(S). \]

Proof. According to Lemma 1, we have
\[ R(T) \leq \frac{D^2}{2\eta} \text{Tr}(H_T) + \frac{\eta}{2} \sum_{t=1}^{T} (\|g_t\|_{H_t})^2. \]

For the first term on the right side of Eq. (9), according to the definition of $H_T$ and $S_T$, we have
\[ \text{Tr}(H_T) = \text{Tr}(C_T S_T) = C_T \text{Tr}(S_T) \leq C_T. \]

Then we only need to prove
\[ \sum_{t=1}^{T} (\|g_t\|_{H_t})^2 \leq 2C_T = 2 \sqrt{\sum_{t=1}^{T} (\|g_t\|_{S_T})^2}. \]

Since
\[ \sum_{t=1}^{T} (\|g_t\|_{H_T})^2 = \sum_{t=1}^{T} (\|g_t\|_{C_T S_T})^2 = \frac{1}{C_T} \sum_{t=1}^{T} (\|g_t\|_{S_T})^2 = \sqrt{\sum_{t=1}^{T} (\|g_t\|_{S_T})^2}, \]
in order to prove Eq. (11), we need to prove that
\[ \sum_{t=1}^{T} (\|g_t\|_{H_t})^2 \leq \frac{1}{C_T} \sum_{t=1}^{T} (\|g_t\|_{S_T})^2. \]

The above equation can be proved by mathematical induction. For $T = 1$, $(\|g_t\|_{H_t})^2 \leq 2 (\|g_t\|_{H_t})^2$ holds obviously. Suppose it holds that
\[ \sum_{t=1}^{T-1} (\|g_t\|_{H_t})^2 \leq \frac{1}{C_T} \sum_{t=1}^{T-1} (\|g_t\|_{S_T})^2, \]
then we have
\[ \sum_{t=1}^{T} (\|g_t\|_{H_t})^2 = \sum_{t=1}^{T-1} (\|g_t\|_{H_t})^2 + (\|g_t\|_{H_T})^2 \leq 2 \sum_{t=1}^{T-1} (\|g_t\|_{H_{T-1}})^2 + (\|g_t\|_{H_T})^2 \]
\[ \leq 2 \sqrt{\sum_{t=1}^{T-1} (\|g_t\|_{S_{T-1}})^2} + \frac{1}{C_T} (\|g_t\|_{S_T})^2 \]
\[ \leq 2C_{T-1} + \frac{1}{C_T} (\|g_t\|_{S_T})^2. \]
Meanwhile, we can prove that

\[
C_T^2 - C_{T-1}^2 = \sum_{t=1}^{T} (\|g_t\|_{s_T}^2)^2 - \sum_{t=1}^{T-1} (\|g_t\|_{s_{T-1}}^2)^2 \\
= \sum_{t=1}^{T-1} (\|g_t\|_{s_T}^2)^2 - \sum_{t=1}^{T-1} (\|g_t\|_{s_{T-1}}^2)^2 + (\|g_T\|_{s_T}^2)^2 \\
= \sum_{t=1}^{T-1} (\|g_t\|_{s_T}^2)^2 - \min_{S \in \Psi, S \geq 0, Tr(S) \leq 1} \sum_{t=1}^{T-1} (\|g_t\|_S^2)^2 + (\|g_T\|_{s_T}^2)^2 \\
\geq (\|g_T\|_{s_T}^2)^2. \tag{16}
\]

Therefore, for Eq. (15), we have

\[
\sum_{t=1}^{T} (\|g_t\|_{u_t}^2)^2 \leq 2C_{T-1} + \frac{1}{C_T} (\|g_T\|_{s_T}^2)^2 \\
\leq 2C_{T-1} + \frac{C_T^2 - C_{T-1}^2}{C_T}. \tag{17}
\]

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

\[
\sum_{t=1}^{T} (\|g_t\|_{u_t}^2)^2 \leq 2C_{T-1} + \frac{C_T^2 - C_{T-1}^2}{C_T} \\
\leq 2C_T \\
= 2\sqrt{\sum_{t=1}^{T} (\|g_t\|_{s_T}^2)^2} \\
= 2 \sum_{t=1}^{T} (\|g_t\|_{s_T}^2)^2. \tag{18}
\]

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 ~ 5.

Proposition 2. It holds that

\[
\frac{1}{n} \sum_{i=1}^{n} g_i g_i^\top \succeq g g^\top, \quad \text{where} \quad g = \frac{1}{n} \sum_{i=1}^{n} g_i. \tag{19}
\]

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

\[
x^\top (\frac{1}{n} \sum_{i=1}^{n} g_i g_i^\top) x \geq x^\top g g^\top x.
\]

Hence, we have

\[
\frac{1}{n} \sum_{i=1}^{n} g_i g_i^\top \succeq g g^\top.
\]

The proof is completed. ■
According to **Proposition 2**, we also have

\[
\frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} g_{ti} g_{ti}^\top \geq \sum_{t=1}^{T} g_{ti} g_{ti}^\top , \text{ where } g_{ti} = \frac{1}{n} \sum_{i=1}^{n} g_{ti} .
\]  

(20)

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

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

**Proposition 4.** If \( x_i \geq 0 \) and \( y_i \geq 0 \) for \( i = 1,2,...,n \), we have \( \sum_{i=1}^{n} x_i y_i \leq \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right) \).

*Proof.* \( \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right) = \sum_{i,j=1}^{n} x_i y_j = \sum_{i=1}^{n} x_i y_j + \sum_{i \neq j}^{n} x_i y_j \geq \sum_{i=1}^{n} x_i y_i . \) The proof is completed. \( \blacksquare \)

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

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

1. \( (A \otimes B)^\top = A^\top \otimes B^\top \), \( (A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \) (if \( A \) and \( B \) are invertible);
2. \( (A \otimes B)(A' \otimes B') = (AA') \otimes (BB') \);
3. If \( A \succeq 0 \) and \( B \succeq 0 \), \( A \otimes B \succeq 0 \);
4. \( \text{Tr}(A \otimes B) = \text{Tr}(A) \text{Tr}(B) \).

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

**Lemma 3.** Denote by \( L_T = \sum_{t=1}^{T} \sum_{i=1}^{n} g_{ti} g_{ti}^\top \) and \( R_T = \sum_{t=1}^{T} \sum_{i=1}^{n} x_{ti} x_{ti}^\top \), there is

\[
F_T(S) \leq \text{Tr}\left((S_1^{-1} \otimes S_2^{-1}) \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} g_{ti} g_{ti}^\top \right)
\]

\[
\leq \frac{1}{n} \text{Tr}(S_1^{-1} L_T) \text{Tr}(S_2^{-1} R_T) .
\]

(21)

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

\[
\frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} g_{ti} g_{ti}^\top \geq \sum_{t=1}^{T} g_{ti} g_{ti}^\top , \text{ where } g_{ti} = \frac{1}{n} \sum_{i=1}^{n} g_{ti} .
\]

(22)

Together with **Proposition 3**, we have

\[
\text{Tr}\left((S_1^{-1} \otimes S_2^{-1}) \sum_{t=1}^{T} g_{ti} g_{ti}^\top \right) \leq \text{Tr}\left((S_1^{-1} \otimes S_2^{-1}) \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} g_{ti} g_{ti}^\top \right) .
\]

(23)

Finally, according to the properties of Kronecker Product in **Proposition 5**, we have
\[
\text{Tr} \left( (S_1^{-1} \otimes S_2^{-1}) \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} g_t g_i^\top \right) = \text{Tr} \left( (S_1^{-1} \otimes S_2^{-1}) \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} (\delta_{t,i} \otimes x_{t,i}) (\delta_{t,i} \otimes x_{t,i}^\top) \right)
\]
\[
= \text{Tr} \left( \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} (S_1^{-1} \delta_{t,i} \delta_{t,i}^\top) \otimes (S_2^{-1} x_{t,i} x_{t,i}^\top) \right)
\]
\[
= \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \text{Tr}(S_1^{-1} \delta_{t,i}) \text{Tr}(S_2^{-1} x_{t,i} x_{t,i}^\top)
\]
\[
\leq \frac{1}{n} \left( \sum_{t=1}^{T} \sum_{i=1}^{n} \text{Tr}(S_1^{-1} \delta_{t,i} \delta_{t,i}^\top) \right) \left( \sum_{t=1}^{T} \sum_{i=1}^{n} \text{Tr}(S_2^{-1} x_{t,i} x_{t,i}^\top) \right) \tag{Proposition 4}
\]
\[
= \frac{1}{n} \text{Tr}(S_1^{-1} L_T) \text{Tr}(S_2^{-1} R_T).
\]

The proof is completed. ■

A3. Proof of Lemma 4

We first present the following Propositions 6 ∼ 7 before we prove Lemma 4.

**Proposition 6.** Suppose \(D \in \mathbb{R}^{d \times d}\) is a diagonal matrix and \(D \succeq 0\), then
\[
\min_{U \in \mathbb{R}^{d \times d}, \text{tr}(U) = 1} \|UD\|_{12} = \text{Tr}(D),
\]
where \(\|A\|_{12} = \sum_i \sqrt{\sum_j A_{ij}^2}\) is the matrix \(L_{12}\)-norm, and \(U = I\) is the optimal point.

**Proof.** For any orthogonal matrix \(U \in \mathbb{R}^{d \times d}\), denote by \(\{u_i\}_{i=1}^{d}\) the row vectors of \(U\), we have
\[
\text{Tr}(D) = \text{Tr}(UDU^\top)
\]
\[
= \sum_{i=1}^{d} u_i^\top Du_i
\]
\[
= \sum_{i=1}^{d} \langle Du_i, u_i \rangle
\]
\[
= \sum_{i=1}^{d} ||Du_i||_2 \cos \langle Du_i, u_i \rangle \tag{26}
\]
\[
\leq \sum_{i=1}^{d} ||Du_i||_2
\]
\[
= \sum_{i=1}^{d} ||u_i^\top D||_2
\]
\[
= ||UD||_{12}.
\]

When \(U = I\), \(\cos \langle Du_i, u_i \rangle = 1\) for \(i = 1, 2, ..., d\), and the equality holds. The proof is completed. ■
Proposition 7. Suppose \( A \in \mathbb{R}^{d \times d}, A \succeq 0, D \in \mathbb{R}^{d \times d} \) and \( D \) is a diagonal matrix, then

\[
\arg \min_{D \succeq 0, \text{tr}(D) \leq 1} \text{Tr}(D^{-1}A) = \frac{1}{||A^{\frac{1}{2}}||_{12}} \text{Diag}((A^{\frac{1}{2}})^{\odot 2}1)^{\odot \frac{1}{2}}
\] (27)

and

\[
\min_{D \succeq 0, \text{tr}(D) \leq 1} \text{Tr}(D^{-1}A) = ||A^{\frac{1}{2}}||_{12}^2.
\] (28)

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

\[
\min_{D \succeq 0, \text{tr}(D) \leq 1} \text{Tr}(D^{-1}A) = \min_{D \succeq 0, \text{tr}(D) \leq 1} \text{Tr}(D^{-1}BB^\top) = \sum_{i=1}^d \sum_{j=1}^d \frac{b_{ij}^2}{d_i}.
\] (29)

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

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

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

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

With the constraint \( 1^\top d \leq 1, \) we can choose a proper \( \theta \) so that

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

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

\[
\text{Tr}(D^{-1}A) = \sum_{i=1}^d \sum_{j=1}^d \frac{b_{ij}^2}{d_i}
\]

\[
= \left( \sum_{i=1}^d \left( \sum_{j=1}^d b_{ij}^2 \right)^{\frac{1}{2}} \right) \sum_{i=1}^d \sum_{j=1}^d \frac{b_{ij}^2}{ \left( \sum_{j=1}^d b_{ij}^2 \right)^{\frac{1}{2}} }
\]

\[
= \left( \sum_{i=1}^d \left( \sum_{j=1}^d b_{ij}^2 \right)^{\frac{1}{2}} \right)^2
\]

\[
= ||B||_{12}^2
\]

\[
= ||A^{\frac{1}{2}}||_{12}^2.
\] (33)

The proof is completed. \( \blacksquare \)

Then we prove Lemma 4 in the main paper.

**Lemma 4.** If \( A \succ 0, \) we have

\[
\arg \min_{S \succeq 0, \text{tr}(S) \leq 1} \text{Tr}(S^{-1}A) = A^{\frac{1}{2}}/\text{Tr}(A^{\frac{1}{2}}).
\] (34)

**Proof.** Because

\[
\min_{S \succeq 0, \text{tr}(S) \leq 1} \text{Tr}(S^{-1}A) = \min_{D = \text{Diag}(d), d \succeq 0.1^\top d \leq 1, UU^\top = I} \text{Tr}(UD^{-1}U^\top A)
\]

\[
= \min_{UU^\top = I D = \text{Diag}(d), d \succeq 0.1^\top d \leq 1} \text{Tr}(UD^{-1}U^\top A),
\] (35)
we can find the optimal diagonal matrix $D$ and orthogonal matrix $U$ to obtain the optimal $S$ by $S = UDU^T$. We first fix $U$ to find the optimal $D$. Since
\[
\min_{D=\text{Diag}(d), d \succeq 0, d^\top d \leq 1} \text{Tr}(UD^{-1}U^T A) = \min_{D=\text{Diag}(d), d \succeq 0, d^\top d \leq 1} \text{Tr}(D^{-1}U^T A U),
\]
(36)
according to Proposition 7, we know the optimal $D$ is
\[
D = \frac{1}{||U^T A ||_{12}} \text{Diag}((U^T A^2)^{\otimes 2} 1)^{\otimes \frac{1}{2}}.
\]
(37)
Meanwhile, we have
\[
\min_{D=\text{Diag}(d), d \succeq 0, d^\top d \leq 1} \text{Tr}(UD^{-1}U^T A) = ||U^T A ||_{12}^2.
\]
(38)
We then minimize Eq (38) w.r.t. $U$. Suppose the SVD decomposition of $A$ is $A = U_A D_A U_A^\top$, there is
\[
||U^T A ||_{12}^2 = ||U^T U_A D_A^\frac{1}{2} U_A^\top ||_{12}^2 = ||U^T U_A D_A^\frac{1}{2} U_A^\top ||_{12}^2.
\]
(39)
According to Proposition 6, we know that when $U^T U_A = I$, i.e., $U = U_A$, Eq (39) reaches its minimal value. Therefore, we have the optimal $D$ as follows
\[
D = \frac{1}{||U_A^\top A^\frac{1}{2} ||_{12}} \text{Diag}((U_A^\top A^\frac{1}{2})^{\otimes 2} 1)^{\otimes \frac{1}{2}} = \frac{1}{||U_A^\top A^\frac{1}{2} D_A^\frac{1}{2} U_A^\top ||_{12}} \text{Diag}((D_A^\frac{1}{2} U_A^\top)^{\otimes 2} 1)^{\otimes \frac{1}{2}} = \frac{1}{||D_A^\frac{1}{2} ||_{12}} \text{Diag}((D_A^\frac{1}{2})^{\otimes 2} 1)^{\otimes \frac{1}{2}} = \frac{1}{\text{Tr}(D_A^\frac{1}{2})} \text{Diag}(D_A^\frac{1}{2}).
\]
(40)
Then, the optimal $S$ is
\[
S = UDU^T = \frac{1}{\text{Tr}(D_A^\frac{1}{2})} U_A \text{Diag}(D_A^\frac{1}{2}) U_A^\top = \frac{1}{\text{Tr}(U_A D_A^\frac{1}{2} U_A^\top)} U_A \text{Diag}(D_A^\frac{1}{2}) U_A^\top = \frac{1}{\text{Tr}(A^\frac{1}{2})} A^\frac{1}{2}.
\]
(41)
The proof is completed. ■

A4. Proof of Lemma 2

Finally, we prove Lemma 2 in the main paper.

Lemma 2. Suppose $\Psi$ is the set of either diagonal matrices or full-matrices, according to the definition of $S_T$ and $H_T$ in Eq. (7), we have
\[
H_T = \text{Diag}\left(\sum_{t=1}^{T} g_t \otimes g_t^\top\right)^{\otimes \frac{1}{2}}, \quad H_T = \left(\sum_{t=1}^{T} g_t g_t^\top\right)^{\otimes \frac{1}{2}}.
\]
(42)
Proof. Because \( H_T = C_T S_T \), we see that we only need to solve \( S_T \). We first prove the case when \( \Psi \) is the set of diagonal matrices. Let \( S = \text{Diag}(s) \) and \( H = \text{Diag}(h) \), where \( s \) and \( h \) are the diagonal vectors of \( S \) and \( H \), respectively, we have

\[
s_T = \arg \min_{s \succeq 0, \, 1^T s \leq 1} \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{g_{ti}^2}{s_i},
\]

(43)

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

\[
L(s, \lambda, \theta) = \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{g_{ti}^2}{s_i} - \langle \lambda, s \rangle + \theta (1^T s - 1).
\]

(44)

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

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

(45)

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

\[
s_i = \theta^{-\frac{1}{2}} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}.
\]

(46)

With the constraint \( 1^T s \leq 1 \), we can choose a proper \( \theta \) so that

\[
s_T = \frac{(\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}}{\sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}}
\]

meets the constraint. Meanwhile, we can derive that

\[
C_T = \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{d} \frac{g_{ti}^2}{s_i}}
\]

\[
= \sqrt{\sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}} \sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}} \sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}
\]

(48)

\[
= \sqrt{\sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}} \sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}} \sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}
\]

\[
= \sum_{i=1}^{d} (\sum_{t=1}^{T} g_{ti}^2)^{\frac{1}{2}}.
\]

Therefore,

\[
h_T = C_T s_T = \sum_{t=1}^{T} \frac{g_{ti}^2}{s_T}
\]

\[
= \sum_{i=1}^{d} \left( \sum_{t=1}^{T} g_{ti}^2 \right)^{\frac{1}{2}} \left( \sum_{i=1}^{d} \left( \sum_{t=1}^{T} g_{ti}^2 \right)^{\frac{1}{2}} \right)^{-1}
\]

(49)

and finally we have \( H_T = \text{Diag} \left( \left( \sum_{t=1}^{T} g_t \odot g_t \right)^{\frac{1}{2}} \right) \).

When \( \Psi \) is the set of full matrices, we have

\[
S_T = \arg \min_{S \succeq 0, \, \text{tr}(S) \leq \lambda} \sum_{t=1}^{T} \| g_t \|^2_S = \arg \min_{S \succeq 0, \, \text{tr}(S) \leq \lambda} \text{tr} \left( S^{-1} \sum_{t=1}^{T} g_t g_t^T \right).
\]

(50)
Lemma 4. We have

$$S_T = \left( \sum_{t=1}^{T} g_t g_t^\top \right)^{\frac{1}{2}} / \text{Tr} \left( \sum_{t=1}^{T} g_t g_t^\top \right)^{\frac{1}{2}}. $$

\[ (51) \]

Meanwhile, there is

$$C_T = \sqrt{\sum_{t=1}^{T} \left( \|g_t\|_{S_T} \right)^2} = \sqrt{\text{Tr} \left( S_T^{-1} \sum_{t=1}^{T} g_t g_t^\top \right)} = \sqrt{\text{Tr} \left( \sum_{t=1}^{T} g_t g_t^\top \right)^{-\frac{1}{2}} \sum_{t=1}^{T} g_t g_t^\top \right) \text{Tr} \left( \sum_{t=1}^{T} g_t g_t^\top \right)^{\frac{1}{2}}} = \text{Tr} \left( \sum_{t=1}^{T} g_t g_t^\top \right)^{\frac{1}{2}}. $$

\[ (52) \]

Therefore,

$$H_T = C_T S_T = \text{Tr} \left( \sum_{t=1}^{T} g_t g_t^\top \right)^{\frac{1}{2}} / \text{Tr} \left( \sum_{t=1}^{T} g_t g_t^\top \right)^{\frac{1}{2}} = \sum_{t=1}^{T} g_t g_t^\top \right)^{\frac{1}{2}}. $$

\[ (53) \]

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).

<table>
<thead>
<tr>
<th>Optimizer</th>
<th>SGDM</th>
<th>AdamW</th>
<th>Adagrad</th>
<th>RAdam</th>
<th>Adabelief</th>
<th>Shampoo</th>
<th>KFAC</th>
<th>SGDM_BK</th>
<th>AdamW_BK</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.001</td>
<td>0.01</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.01</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>WD</td>
<td>0.0005</td>
<td>0.5</td>
<td>0.0005</td>
<td>0.5</td>
<td>0.0005</td>
<td>0.005</td>
<td>0.001</td>
<td>0.005</td>
<td>0.5</td>
</tr>
<tr>
<td>WD method</td>
<td>$L_2$</td>
<td>decouple</td>
<td>$L_2$</td>
<td>decouple</td>
<td>decouple</td>
<td>$L_2$</td>
<td>decouple</td>
<td>$L_2$</td>
<td>decouple</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Optimizer</th>
<th>SGDM</th>
<th>AdamW</th>
<th>Adagrad</th>
<th>RAdam</th>
<th>Adabelief</th>
<th>Shampoo</th>
<th>KFAC</th>
<th>SGDM_BK</th>
<th>AdamW_BK</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet18</td>
<td>LR</td>
<td>0.1</td>
<td>0.001</td>
<td>0.01</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>WD</td>
<td>0.0001</td>
<td>0.1</td>
<td>0.0006</td>
<td>0.1</td>
<td>0.05</td>
<td>0.0005</td>
<td>0.0001</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>WD method</td>
<td>$L_2$</td>
<td>decouple</td>
<td>$L_2$</td>
<td>decouple</td>
<td>decouple</td>
<td>$L_2$</td>
<td>decouple</td>
<td>$L_2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Optimizer</th>
<th>SGDM</th>
<th>AdamW</th>
<th>Adagrad</th>
<th>RAdam</th>
<th>Adabelief</th>
<th>Shampoo</th>
<th>KFAC</th>
<th>SGDM_BK</th>
<th>AdamW_BK</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50</td>
<td>LR</td>
<td>0.1</td>
<td>0.001</td>
<td>0.01</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>WD</td>
<td>0.0001</td>
<td>0.1</td>
<td>0.0006</td>
<td>0.1</td>
<td>0.05</td>
<td>0.0005</td>
<td>0.0001</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>WD method</td>
<td>$L_2$</td>
<td>decouple</td>
<td>$L_2$</td>
<td>decouple</td>
<td>decouple</td>
<td>$L_2$</td>
<td>decouple</td>
<td>$L_2$</td>
</tr>
</tbody>
</table>

Table 3. Testing accuracies (%) of DNNs with different dampening $\epsilon$.

<table>
<thead>
<tr>
<th>Optimizer</th>
<th>SGDM</th>
<th>AdamW</th>
<th>Adagrad</th>
<th>RAdam</th>
<th>Adabelief</th>
<th>Shampoo</th>
<th>KFAC</th>
<th>SGDM_BK</th>
<th>AdamW_BK</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet18</td>
<td>$\epsilon$</td>
<td>0.1</td>
<td>0.01</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>ResNet50</td>
<td>$\epsilon$</td>
<td>0.1</td>
<td>0.01</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Optimizer</th>
<th>SGDM</th>
<th>AdamW</th>
<th>Adagrad</th>
<th>RAdam</th>
<th>Adabelief</th>
<th>Shampoo</th>
<th>KFAC</th>
<th>SGDM_BK</th>
<th>AdamW_BK</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet18</td>
<td>$T_e = 50$ and $T_{ir} = 500$</td>
<td>$78.60 \pm 0.23$</td>
<td>$79.26 \pm 0.12$</td>
<td>$79.21 \pm 0.22$</td>
<td>$79.53 \pm 0.22$</td>
<td>$79.35 \pm 0.29$</td>
<td>$79.36 \pm 0.22$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$T_e = 200$ and $T_{ir} = 2000$</td>
<td>$78.49 \pm 0.31$</td>
<td>$80.66 \pm 0.21$</td>
<td>$80.10 \pm 0.51$</td>
<td>$80.13 \pm 0.51$</td>
<td>$80.05 \pm 0.51$</td>
<td>$80.19 \pm 0.51$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet50</td>
<td>$T_e = 50$ and $T_{ir} = 500$</td>
<td>$78.60 \pm 0.23$</td>
<td>$79.26 \pm 0.12$</td>
<td>$79.21 \pm 0.22$</td>
<td>$79.53 \pm 0.22$</td>
<td>$79.35 \pm 0.29$</td>
<td>$79.36 \pm 0.22$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$T_e = 200$ and $T_{ir} = 2000$</td>
<td>$78.49 \pm 0.31$</td>
<td>$80.66 \pm 0.21$</td>
<td>$80.10 \pm 0.51$</td>
<td>$80.13 \pm 0.51$</td>
<td>$80.05 \pm 0.51$</td>
<td>$80.19 \pm 0.51$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 $\epsilon$ and the statistics updating intervals $T_e$ 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 $\{1e^{-4}, 5e^{-4}, 1e^{-3}, 3e^{-3}, 5e^{-3}, 1e^{-2}, 3e^{-2}, 5e^{-2}, 0.1, 0.3, 0.5\}$, 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 $\epsilon$ 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 $\epsilon$ 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**


