Distributed Adversarial Training to Robustify Deep Neural Networks at Scale
(Supplementary Material)

Gaoyuan Zhang1,∗ Songtao Lu1,∗ Yihua Zhang2 Xiangyi Chen3 Pin-Yu Chen1 Quanfu Fan1 Lee Martie1
Lior Horesh1 Mingyi Hong3 Sijia Liu1,2
1IBM Research
2Michigan State University
3University of Minnesota
∗Equal Contribution

1 DAT ALGORITHM FRAMEWORK

Algorithm A1 Distributed adversarial training (DAT) for solving problem ??

1: Initial θ1, dataset D(i) for each of M workers, and T iterations
2: for Iteration t = 1, 2, . . . , T do
3:     for Worker i = 1, 2, . . . , M do
4:         Draw a finite-size data batch B(i) t ⊆ D(i)
5:         For each data sample x ∈ B(i) t, call for an inner maximization oracle:
6:             δ(i) t (x) := arg max
\[\|\delta\|_\infty \leq \epsilon \phi(\theta_t; \delta; x),\]  (A1)
where we omit the label or possible pseudo-label y of x for brevity
7:     end for
8:     (Optional) Call for gradient quantizer Q(·) and transmit Q(g(i) t ) to server
9:     Gradient aggregation at server:
10:    ⊗ Server
11:     ĝt = 1 M M \sum i=1 Q(g(i) t )  (A3)
12:     (Optional) Call for gradient quantizer ĝt ← Q(ĝt), and transmit ĝt to workers:
13:     for Worker i = 1, 2, . . . , M do
14:         (Optional) Call for gradient quantizer A(·) to update θ:
15:     end for
16:     θt+1 = A(θt; ĝt, ηt), ηt is learning rate  (A4)
17: end for

Additional details on gradient quantization Let b denote the number of bits (b ≤ 32), and thus there exists s = 2^b quantization levels. We specify the gradient quantization operation Q(·) in Algorithm A1 as the randomized quantizer \[Alistarh et al., 2017, Yu et al., 2019\]. Formally, the quantization operation at the i-th coordinate of a vector g is given by

\[Q(g_i) = ||g||_2 \cdot \text{sign}(g_i) \cdot \xi_i(g_i, s), \quad \forall i \in \{1, 2, \ldots, d\}.\]  (A5)
In \((A5)\), \(\xi_i(g_i, s)\) is a random number drawn as follows. Given \(|g_i|/\|g\|_2 \in [l/s, (l + 1)/s]\) for some \(l \in \mathbb{N}^+\) and \(0 \leq l < s\), we then have
\[
\xi_i(g_i, s) = \begin{cases} 
{l/s} & \text{with probability } 1 - (s|g_i|/\|g\|_2 - l) \\
(l + 1)/s & \text{with probability } (s|g_i|/\|g\|_2 - l),
\end{cases}
\] (A6)
where \(|a|\) denotes the absolute value of a scalar \(a\), and \(|a|_2\) denotes the \(\ell_2\) norm of a vector \(a\). The rationale behind using \((A5)\) is that \(Q(g_i)\) is an unbiased estimate of \(g_i\), namely, \(E_{\xi_i(g_i, s)}[Q(g_i)] = g_i\), with bounded variance. Moreover, we at most need \((32 + d + bd)\) bits to transmit the quantized \(Q(g)\), where 32 bits for \(\|g\|_2\), 1 bit for sign of \(g_i\) and \(b\) bits for \(\xi_i(g_i, s)\), whereas it needs \(32d\) bits for a single-precision \(g\). Clearly, a small \(b\) saves the communication cost. We note that if every worker performs as a server in DAT, then the quantization operation at Step 10 of Algorithm A1 is no longer needed. In this case, the communication network becomes fully connected. With synchronized communication, this is favored for training DNNs under the All-reduce operation.

## 2 THEORETICAL RESULTS

In this section, we will quantify the convergence behaviour of the proposed DAT algorithm. First, we define the following notations:
\[
\Phi_i(\theta, x) = \max_{\|\delta^{(i)}\|_\infty \leq \epsilon} \phi(\theta, \delta^{(i)}; x), \quad \text{and} \quad \Phi_i(\theta) = E_{x \in D^{(i)}} \Phi_i(\theta; x).
\] (A7)
We also define
\[
l_i(\theta) = E_{x \in D^{(i)}} l(\theta; x),
\] (A8)
where the label \(y\) of \(x\) is omitted for labeled data. Then, the objective function of problem (7) can be expressed in the compact way
\[
\Psi(\theta) = \frac{1}{M} \sum_{i=1}^{M} l_i(\theta) + \Phi_i(\theta)
\] (A9)
and the optimization problem is then given by \(\min_\theta \Psi(\theta)\).

Therefore, it is clear that if a point \(\theta^*\) satisfies
\[
\|\nabla_\theta \Psi(\theta^*)\| \leq \xi,
\] (A10)
then we say \(\theta^*\) is a \(\xi\) approximate first-order stationary point (FOSP) of problem (7).

Prior to delving into the convergence analysis of DAT, we make the following assumptions.

### 2.1 ASSUMPTIONS

A1. Assume objective function has layer-wise Lipschitz continuous gradients with constant \(L_i\) for each layer
\[
\|\nabla_i \Psi(\theta, i) - \nabla_i \Psi(\theta', i)\| \leq L_i \|\theta_i - \theta_i'\|, \forall i \in [h].
\] (A11)
where \(\nabla_i \Psi(\theta, i)\) denotes the gradient w.r.t. the variables at the \(i\)th layer. Also, we assume that \(\Psi(\theta)\) is lower bounded, i.e., \(\Psi^* := \min_\theta \Psi(\theta) > -\infty\) and bounded gradient estimate, i.e., \(\|\nabla_i \Psi(\theta)\| \leq G\).

A2. Assume that \(\phi(\theta, \delta; x)\) is strongly concave with respect to \(\delta\) with parameter \(\mu\) and has the following gradient Lipschitz continuity with constant \(L_\phi\):
\[
\|\nabla_\theta \phi(\theta, \delta; x) - \nabla_\theta \phi(\theta, \delta'; x)\| \leq L_\phi \|\delta - \delta'\|.
\] (A12)

A3. Assume that the gradient estimate is unbiased and has bounded variance, i.e.,
\[
E_{x \in B^{(i)}} \nabla_\theta l(\theta; x) = \nabla_\theta l(\theta), \forall i,
\] (A13)
\[
E_{x \in B^{(i)}} \nabla_\theta \Phi_i(\theta; x) = \nabla_\theta \Phi_i(\theta), \forall i,
\] (A14)
where recall that \(B^{(i)}\) denotes a data batch used at worker \(i\), \(\nabla_\theta l(\theta) := \frac{1}{M} \sum_{i=1}^{M} \nabla_\theta l_i(\theta)\) and \(\nabla_\theta \Phi_i(\theta) := \frac{1}{M} \sum_{i=1}^{M} \nabla_\theta \Phi_i(\theta)\); and
\[
E_{x \in B^{(i)}} \|\nabla_\theta l(\theta; x) - \nabla_\theta l(\theta)\|^2 \leq \sigma^2, \forall i
\] (A15)
\[
E_{x \in B^{(i)}} \|\nabla_\theta \Phi_i(\theta; x) - \nabla_\theta \Phi_i(\theta)\|^2 \leq \sigma^2, \forall i.
\] (A16)
Further, we define a component-wise bounded variance of the gradient estimate

\[
E_{x \in B^{(i)}} \left[ \left\| \nabla g_l(\theta; x) \right\|_{jk} - \left\| \nabla g_l(\theta) \right\|_{jk} \right|^2 \leq \sigma_{jk}^2, \forall i, \tag{A17}
\]

\[
E_{x \in B^{(i)}} \left[ \left\| \nabla \Phi(\theta; x) \right\|_{jk} - \left\| \nabla \Phi(\theta) \right\|_{jk} \right|^2 \leq \sigma_{jk}^2, \forall i, \tag{A18}
\]

where \(j\) denotes the index of the layer, and \(k\) denotes the index of entry at each layer. Under A3, we have

\[
\sum_{j=1}^h \sum_{k=1}^{d_j} \max \{\sigma_{jk}^2, \sigma_{jk}^2\} \leq \sigma^2
\]

A4. Assume that the component wise compression error has bounded variance

\[
E[Q(\mathbf{g}^{(i)}(\theta)_{jk}) - \mathbf{g}^{(i)}(\theta)_{jk}] \leq \delta_{jk}, \forall i.
\]

The assumption A4 is satisfied as the randomized quantization is used [Alistarh et al., 2017, Lemma 3.1].

2.2 ORACLE OF MAXIMIZATION

In practice, \(\Phi_i(\theta; x), \forall i\) may not be obtained, since the inner loop needs to iterate by the infinite number of iterations to achieve the exact maximum point. Therefore, we allow some numerical error term resulted in the maximization step at (A1). This consideration makes the convergence analysis more realistic.

First, we have the following criterion to measure the closeness of the approximate maximizer to the optimal one.

**Definition 1.** Under A2, if point \(\delta(x)\) satisfies

\[
\max_{\delta \leq ||x||} \left( \delta - \delta^*(x), \nabla \delta \phi(\theta, \delta^*(x); x) \right) \leq \varepsilon \tag{A20}
\]

then, it is a \(\varepsilon\) approximate solution to \(\delta^*(x)\), where

\[
\delta^*(x) := \arg \max_{\delta \leq ||x||} \phi(\theta, \delta; x).
\]

and \(x\) denotes the sampled data.

Condition [A20] is standard for defining approximate solutions of an optimization problem over a compact feasible set and has been widely studied in [Wang et al., 2019, Lu et al., 2020].

In the following, we can show that when the inner maximization problem is solved accurately enough, the gradients of function \(\phi(\theta, \delta(x); x)\) at \(\delta(x)\) and \(\delta^*(x)\) are also close. A similar claim of this fact has been shown in [Wang et al., 2019, Lemma 2]. For completeness of the analysis, we provide the specific statement for our problem here and give the detailed proof as well.

**Lemma 1.** Let \(\delta_t^{(k)}\) be the \((\mu\varepsilon)/L^2_\phi\) approximate solution of the inner maximization problem for worker \(k\), i.e.,

\[
\max_{\delta_t} \phi(\theta_t, \delta_t^{(k)}(x_t); x_t) = \phi(\theta_t, \delta_t^{(k)}(x_t); x_t), \forall t.
\]

Under A2, we have

\[
\left\| \nabla \phi \left( \theta_t, \delta_t^{(k)}(x_t); x_t \right) - \nabla \phi \left( \theta_t, \delta^*(x_t); x_t \right) \right\|^2 \leq \varepsilon. \tag{A22}
\]

Throughout the convergence analysis, we assume that \(\delta_t^{(k)}(x_t), \forall k, t\) are all the \((\mu\varepsilon)/L^2_\phi\) approximate solutions of the inner maximization problem. Let us define

\[
\left\| \left[ \nabla \phi(\theta_t, \delta_t^{(k)}(x_t); x_t) \right]_{ij} - \left[ \nabla \phi(\theta_t, \delta^*(x_t); x_t) \right]_{ij} \right\|^2 = \varepsilon_{ij}. \tag{A23}
\]

From Lemma 1, we know that when \(\delta_t^{(k)}(x_t)\) is a \((\mu\varepsilon)/L^2_\phi\) approximate solution, then

\[
\sum_{i=1}^h \sum_{j=1}^{d_i} \varepsilon_{ij} = \sum_{i=1}^h \sum_{j=1}^{d_i} \left\| \left[ \nabla \phi(\theta_t, \delta_t^{(k)}(x_t); x_t) \right]_{ij} - \left[ \nabla \phi(\theta_t, \delta^*(x_t); x_t) \right]_{ij} \right\|^2 \leq \varepsilon. \tag{A24}
\]
2.3 FORMAL STATEMENTS OF CONVERGENCE RATE GUARANTEES

In what follows, we provide the formal statement of convergence rate of DAT. In our analysis, we focus on the 1-sided quantization, namely, Step 10 of Algorithm [A1] is omitted, and specify the outer minimization oracle by LAMB [You et al., 2019], see Algorithm [A2]. The addition and multiplication operations in LAMB are component-wise.

**Theorem 1.** Under A1-A4, suppose that \{θ_t\} is generated by DAT for a total number of T iterations, and let the problem dimension at each layer be \(d_i = d/h\). Then the convergence rate of DAT is given by

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \| \nabla \theta \Psi(\theta_t) \|^2 \leq \frac{\Delta_{\psi}}{\eta_t c_t C T} + 2 \left( \varepsilon + \frac{(1 + \lambda)\sigma^2}{MB} \right) + 4\delta^2 + \frac{\kappa \sqrt{\beta} \epsilon}{C} \| \chi \|_1 + \frac{\eta_t c_u \kappa \| L \|_1}{2C}.
\]  
(A25)

where \( \Delta_{\psi} := \mathbb{E}[\Psi(\theta_t) - \Psi^*] \), \( \eta_t \) is the learning rate, \( \kappa = c_u/c_t \), \( c_t \) and \( c_u \) are constants used in LALR (?), \( \chi \) is an error term with the \((ih+j)\)th entry being \( \sqrt{\frac{(1+\lambda)\sigma_i^2}{MB}} + \varepsilon_{ij} + \delta_{ij}^2 \), \( \varepsilon \) and \( \varepsilon_{ij} \) were given in (A24). \( L = [L_1, \ldots, L_h]^T \), \( C = \frac{1}{4} \sqrt{\frac{h(1+\beta_2)}{G^2 \text{d}/\text{a}}} \), 0 < \( \beta_2 < 1 \) is increased in LAMB, \( B = \min\{|B|, \forall i\} \), and \( G \) is given in A1.

**Remark 1.** When the batch size is large, i.e., \( B \sim \sqrt{T} \), then the gradient estimate error will be \( O(\sigma^2/\sqrt{T}) \). Further, it is worth noting that different from the convergence results of LAMB, there is a linear speedup of decreasing the gradient estimate error in DAT with respect to \( M \), i.e., \( O(\sigma^2/(M \sqrt{T})) \), which is the advantage of using multiple computing nodes.

**Remark 2.** Note that A4 implies \( \mathbb{E}[|Q([g^{(k)}(\theta)]_{ij}) - [g^{(k)}(\theta)]_{ij}|^2] \leq \sum_{l=1}^{h} \sum_{j=1}^{d} \delta_{ij}^2 : = \delta^2 \). From [Alistarh et al., 2017, Lemma 3.1], we know that \( \delta^2 \leq \min\{d/s^2, \sqrt{d}/s\} G^2 \). Recall that \( s = 2^b \), where \( b \) is the number of quantization bits. Therefore, with a proper choice of the parameters, we can have the following convergence result that has been shown in Theorem 2.

**Corollary 1.** Under the same conditions of Theorem [T] if we choose

\[
\eta_t \sim O\left(1/\sqrt{T}\right), \quad \varepsilon \sim O\left(\xi^2\right),
\]  
(A66)

we then have

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \| \nabla \theta \Psi(\theta_t) \|^2 \leq \frac{\Delta_{\psi}}{c_t C \sqrt{T}} + \frac{(1 + \lambda)\sigma^2}{MB} + \frac{c_u \kappa \| L \|_1}{2C \sqrt{T}} + O\left(\xi, \frac{\sigma}{\sqrt{MT}} \min\left\{\frac{d}{4\xi}, \frac{\sqrt{d}}{2\xi}\right\}\right).
\]  
(A27)

In summary, when the batch size is large enough, DAT converges to a first-order stationary point of problem (??) and there is a linear speed-up in terms of \( M \) with respect to \( \sigma^2 \). Next, we provide the details of the proof.

3 PROOF DETAILS

3.1 PRELIMINARIES

In the proof, we use the following inequality and notations.

1. Young’s inequality with parameter \( \epsilon \) is

\[
\langle x, y \rangle \leq \frac{1}{2\epsilon} \| x \|^2 + \frac{\epsilon}{2} \| y \|^2,
\]  
(A28)

where \( x, y \) are two vectors.

2. Define the historical trajectory of the iterates as \( F_t = \{\theta_{t-1}, \ldots, \theta_1\} \).

3. We denote vector \([x]_i\) as the parameters at the \( i \)th layer of the neural net and \([x]_{ij}\) represents the \( j \)th entry of the parameter at the \( i \)th layer.

4. We define

\[
g_t := \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{x_t \in B(i)} \left( \lambda \nabla l(\theta_t; x_t) + \nabla \theta \phi(\theta_t, \delta_t^{(i)}(x_t); x_t) \right) = \frac{1}{M} \sum_{i=1}^{M} g_t^{(i)}.
\]  
(A29)
3.2 DETAILS OF LAMB ALGORITHM

Algorithm A2 LAMB [You et al. 2019]

Input: learning rate $\eta_t$, $0 < \beta_1, \beta_2 < 1$, scaling function $\tau(\cdot)$, $\zeta > 0$

for $t = 1, \ldots$ do

\[ m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \ \text{where } g_t \text{ is given by (A3)} \]

\[ v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \]

\[ m_t = m_t / (1 - \beta_1^t) \]

\[ v_t = v_t / (1 - \beta_2^t) \]

Compute ratio $u_t = \frac{m_t}{\sqrt{v_t + \zeta}}$

end for

Update

\[ \theta_{t+1,i} = \theta_{t,i} - \frac{\eta_t \tau(\|\theta_{t,i}\|)}{\|u_{t,i}\|} u_{t,i}. \]  \hspace{1cm} (A30)

3.3 PROOF OF LEMMA 1

Proof. From A2, we have

\[ \| \nabla \phi(\theta_t, \delta^{(i)}_t(x_t); x_t) - \nabla \phi(\theta_t, (\delta^*)^{(i)}_t(x_t); x_t) \| \leq L_\phi \| \delta^{(i)}_t(x_t) - (\delta^*)^{(i)}_t(x_t) \|. \]  \hspace{1cm} (A31)

Also, we know that function $\phi(\theta, \delta, x)$ is strongly concave with respect to $\delta$, so we have

\[ \mu \| \delta^{(i)}_t(x_t) - (\delta^*)^{(i)}_t(x_t) \| \]

\[ \leq \left< \nabla_\delta \phi(\theta_t, (\delta^*)^{(i)}_t(x_t); x_t) - \nabla_\delta \phi(\theta_t, \delta^{(i)}_t(x_t); x_t), \delta^{(i)}_t(x_t) - (\delta^*)^{(i)}_t(x_t) \right>. \]  \hspace{1cm} (A32)

Next, we have two conditions about the qualities of solutions $\delta^{(i)}_t(x_t)$ and $(\delta^*)^{(i)}_t(x_t)$. First, we know that $\delta^{(i)}_t(x_t)$ is a-$\varepsilon$ approximate solution to $(\delta^*)^{(i)}_t(x_t)$, so we have

\[ \left< (\delta^*)^{(i)}_t(x_t) - \delta^{(i)}_t(x_t), \nabla_\delta \phi(\theta_t, (\delta^*)^{(i)}_t(x_t); x_t) \right> \leq \varepsilon. \]  \hspace{1cm} (A33)

Second, since $(\delta^*)^{(i)}_t(x_t)$ is the optimal solution, it satisfies

\[ \left< (\delta^*)^{(i)}_t(x_t) - (\delta^*)^{(i)}_t(x_t), \nabla_\delta \phi(\theta_t, (\delta^*)^{(i)}_t(x_t); x_t) \right> \leq 0. \]  \hspace{1cm} (A34)

Adding them together, we can obtain

\[ \left< \delta^{(i)}_t(x_t) - (\delta^*)^{(i)}_t(x_t), \nabla_\delta \phi(\theta_t, (\delta^*)^{(i)}_t(x_t); x_t) - \nabla_\delta \phi(\theta_t, (\delta^*)^{(i)}_t(x_t); x_t) \right> \leq \varepsilon. \]  \hspace{1cm} (A35)

Substituting (A35) into (A32), we can get

\[ \mu \| \delta^{(i)}_t(x_t) - (\delta^*)^{(i)}_t(x_t) \|^2 \leq \varepsilon. \]  \hspace{1cm} (A36)

Combining (A31), we have

\[ \| \nabla \phi(\theta_t, (\delta^*)^{(i)}_t(x_t); x_t) - \nabla \phi(\theta_t, (\delta^*)^{(i)}_t(x_t); x_t) \|^2 \leq L_\phi^2 \varepsilon / \mu. \]  \hspace{1cm} (A37)
3.4 DESCENT OF QUANTIZED LAMB

First, we provide the following lemma as a stepping stone for the subsequent analysis.

**Lemma 2.** Under A1–A3, suppose that sequence \( \{ \theta_t \} \) is generated by DAT. Then, we have

\[
\mathbb{E}[-(\nabla \Psi(\theta_t), \dot{g}_t)] \leq -\frac{\mathbb{E}\|\nabla \Psi(\theta_t)\|^2}{2} + \varepsilon + \frac{(1 + \lambda)\sigma^2}{MB}.
\]

**(A38)**

**Proof.** From (A21), (A7) and A2, we know that

\[
\nabla_{\theta} \Phi_t(\theta, x) = \nabla_{\theta} \phi(\theta, (\delta^*)^{(i)}(x); x),
\]

so we can get

\[
\nabla_{\theta} \Psi(\theta) = \frac{1}{M} \sum_{i=1}^{M} \lambda \nabla_{\theta} l_i(\theta) + \nabla_{\theta} \Phi_t(\theta)
\]

\[
= \lambda \nabla_{\theta} l(\theta) + \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{x \in D(i)} \nabla_{\theta} \phi(\theta, (\delta^*)^{(i)}(x); x)
\]

\[
:= \bar{g}(\theta).
\]

Then, we have

\[
\mathbb{E}(\nabla \Psi(\theta_t), g_t) = \mathbb{E}(\nabla \Psi(\theta_t), \dot{g}_t) + \mathbb{E}(\nabla \Psi(\theta_t), g_t - \dot{g}_t)
\]

\[
= \mathbb{E}_{x \in \mathcal{D}} \mathbb{E}_{x \in \mathcal{D}} \langle \nabla \Psi(\theta_t), \dot{g}_t \rangle + \mathbb{E}(\nabla \Psi(\theta_t), g_t - \dot{g}_t)
\]

\[
= \mathbb{E}(\nabla \Psi(\theta_t), g_t - g^*_t) + \mathbb{E}(\nabla \Psi(\theta_t), g_t^* - \dot{g}_t)
\]

where

\[
g_t := \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{x_i \in \mathcal{D}(i)} \left( \lambda \nabla l(\theta_t, x_i) + \nabla_{\theta} \phi(\theta_t, (\delta^*)^{(i)}(x_i); x_i) \right) = \lambda \nabla l(\theta_t) + \nabla \Phi(\theta_t),
\]

and

\[
g_t^* := \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{x_i \in \mathcal{B}(i)} \left( \lambda \nabla l(\theta_t, x_i) + \nabla_{\theta} \phi(\theta_t, (\delta^*)^{(i)}(x_i); x_i) \right).
\]

Next, we can quantify the difference between \( g_t \) and \( g_t^* \) by gradient Lipschitz continuity of function \( \tau(\cdot) \) as the following

\[
\mathbb{E}\|g_t - g_t^*\|^2 \leq \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{x_i \in \mathcal{D}(i)} \left\| \nabla_{\theta} \phi(\theta_t, (\delta^*)^{(i)}(x_i); x_i) - \nabla_{\theta} \phi(\theta, (\delta^*)^{(i)}(x_i); x_i) \right\|^2 \leq \varepsilon
\]

where in (a) we use Jensen’s inequality.

And the difference between \( \bar{g}_t \) and \( g_t^* \) can be upper bounded by

\[
\mathbb{E}\|\bar{g}_t - g_t^*\|^2 = \mathbb{E}_{x_i \in \mathcal{D}(i)} \left\| \lambda \nabla l(\theta_t; x_i) + \nabla_{\theta} \phi(\theta_t, (\delta^*)^{(i)}(x_i); x_i) - \nabla_{\theta} \phi(\theta_t) \right\|^2
\]

\[
+ \lambda \mathbb{E}_{x_i \in \mathcal{D}(i)} \lambda \sum_{i=1}^{M} \mathbb{E}_{x_i \in \mathcal{D}(i)} \nabla l(\theta_t; x_i) - \nabla l(\theta_t) \right\|^2
\]

\[
\leq \frac{1}{MB} (1 + \lambda)\sigma^2.
\]

**(A3)**
Applying Young’s inequality with parameter 2, we have

\[
E[-\langle \nabla \Psi(\theta_t), g_t \rangle] \leq -E\|\nabla \Psi(\theta_t)\|^2 + \frac{E\|\nabla \Psi(\theta_t)\|^2}{2} + E\|g_t - g_t^*\|^2 + E\|g_t^* - g_t\|^2
\]  

(A52)

\[
\leq -E\|\nabla \Psi(\theta_t)\|^2 - \frac{E\|\nabla \Psi(\theta_t)\|^2}{2} + \varepsilon + \frac{(1 + \lambda)\sigma^2}{MB}.
\]  

(A53)

3.5 PROOF OF THEOREM 1

Proof. We set \(\beta_1 = 0\) in LAMB for simplicity. From gradient Lipschitz continuity, we have

\[
\Psi(\theta_{t+1}) \leq \Psi(\theta_t) + \sum_{i=1}^{h} \|\nabla \theta \Psi(\theta_t)\|_{ij} \left[ \frac{u_{ij}}{|u_{ij}|} \right] \mathbb{1} (\text{sign}(\nabla \Psi(\theta_t))_{ij} = \text{sign}(u_{ij}))
\]

\[
- \eta_t \sum_{i=1}^{h} \sum_{j=1}^{d_i} \tau(\|\theta_{t,i}\|) \left[ \frac{u_{ij}}{|u_{ij}|} \right] \mathbb{1} (\text{sign}(\nabla \Psi(\theta_t))_{ij} \neq \text{sign}(u_{ij}))
\]  

(A54)

where in (a) we use (A30), and the upper bound of \(\tau(\|\theta_{t,i}\|)\).

Next, we split term \(R\) as two parts by leveraging sign\((\nabla \Psi(\theta_t))_{ij}\) and sign\((u_{ij})\) as follows.

\[
R = -\eta_t \sum_{i=1}^{h} \sum_{j=1}^{d_i} \tau(\|\theta_{t,i}\|) \left[ \frac{u_{ij}}{|u_{ij}|} \right] \mathbb{1} (\text{sign}(\nabla \Psi(\theta_t))_{ij} = \text{sign}(u_{ij}))
\]

\[
- \eta_t \sum_{i=1}^{h} \sum_{j=1}^{d_i} \tau(\|\theta_{t,i}\|) \left[ \frac{u_{ij}}{|u_{ij}|} \right] \mathbb{1} (\text{sign}(\nabla \Psi(\theta_t))_{ij} \neq \text{sign}(u_{ij}))
\]  

(A55)

Next, we set \(\beta = 0\) in LAMB for simplicity. From gradient Lipschitz continuity, we have

\[
\Psi(\theta_{t+1}) \leq \Psi(\theta_t) + \sum_{i=1}^{h} \|\nabla \theta \Psi(\theta_t)\|_{ij} \left[ \frac{u_{ij}}{|u_{ij}|} \right] \mathbb{1} (\text{sign}(\nabla \Psi(\theta_t))_{ij} = \text{sign}(u_{ij}))
\]

\[
- \eta_t \sum_{i=1}^{h} \sum_{j=1}^{d_i} \tau(\|\theta_{t,i}\|) \left[ \frac{u_{ij}}{|u_{ij}|} \right] \mathbb{1} (\text{sign}(\nabla \Psi(\theta_t))_{ij} \neq \text{sign}(u_{ij}))
\]  

(A56)

\[
\leq -\eta_t c_l \sum_{i=1}^{h} \sum_{j=1}^{d_i} \sqrt{\frac{1 - \beta_2}{G^2 d_i}} \left[ \nabla \Psi(\theta_t) \right]_{ij} \left[ \frac{g_{t,i}}{|g_{t,i}|} \right] \mathbb{1} (\text{sign}(\nabla \Psi(\theta_t))_{ij} = \text{sign}(g_{t,i}))
\]

\[
- \eta_t \sum_{i=1}^{h} \sum_{j=1}^{d_i} \tau(\|\theta_{t,i}\|) \left[ \frac{u_{ij}}{|u_{ij}|} \right] \mathbb{1} (\text{sign}(\nabla \Psi(\theta_t))_{ij} \neq \text{sign}(u_{ij}))
\]  

(A57)

where in (a) we use the fact that \(\|u_{t,i}\| \leq \sqrt{\frac{d_i}{1 - \beta_2}}\) and \(\sqrt{\lambda} \leq G\), and in (b) we add

\[
-\eta_t c_l \sum_{i=1}^{h} \sum_{j=1}^{d_i} \sqrt{\frac{1 - \beta_2}{G^2 d_i}} \left[ \nabla \Psi(\theta_t) \right]_{ij} \left[ \frac{g_{t,i}}{|g_{t,i}|} \right] \mathbb{1} (\text{sign}(\nabla \Psi(\theta_t))_{ij} \neq \text{sign}(g_{t,i})) \geq 0.
\]  

(A58)
Taking expectation on both sides of (A58), we have the following:

\[
\mathbb{E}[\mathcal{R}] \leq -\eta_c t \sqrt{\frac{h(1 - \beta_2)}{G^2 d}} \sum_{i=1}^{h} \sum_{j=1}^{d_i} \mathbb{E}[[\nabla \Psi(\theta_t)]_i]_j [\hat{g}_t]_{ij} \\
\leq -\eta_c u \sum_{i=1}^{h} \sum_{j=1}^{d_i} \mathbb{E}[[\nabla \Psi(\theta_t)]_i]_j (\text{sign}([\nabla \Psi(\theta_t)]_i)_j) \neq \text{sign}([u_t]_{ij})) \\
\quad := \mathcal{U}
\]

Next, we will get the upper bounds of \( \mathcal{U} \) and \( \mathcal{V} \) separately as follows. First, we write the inner product between \([\nabla \Psi(\theta)]_i]_j \) and \([\hat{g}_t]_{ij} \) more compactly,

\[
\mathcal{U} \leq -\eta_c t \sqrt{\frac{h(1 - \beta_2)}{G^2 d}} \sum_{i=1}^{h} \mathbb{E} \langle [\nabla \Psi(\theta_t)]_i, [\hat{g}_t]_i \rangle \\
\leq -\eta_c t \sqrt{\frac{h(1 - \beta_2)}{G^2 d}} \sum_{i=1}^{h} \mathbb{E} \langle [\nabla \Psi(\theta_t)]_i, [\hat{g}_t]_i - [g_t]_i + [g_t]_i \rangle \\
\leq -\eta_c t \sqrt{\frac{h(1 - \beta_2)}{G^2 d}} \left( \mathbb{E} \langle [\nabla \Psi(\theta)]_i, g_t \rangle + \sum_{i=1}^{h} \mathbb{E} \langle [\nabla \Psi(\theta_t)]_i, [\hat{g}_t]_i - [g_t]_i \rangle \right).
\]  

Applying Lemma 2 we can get

\[
\mathcal{U} \overset{(A58)}{=} -\eta_c t \sqrt{\frac{h(1 - \beta_2)}{G^2 d}} \frac{1}{2} \mathbb{E} \|\nabla \Psi(\theta_t)\|^2 + \eta_c t \sqrt{\frac{h(1 - \beta_2)}{G^2 d}} \left( \varepsilon + \frac{(1 + \lambda)\sigma^2}{MB} \right) \\
\quad - \frac{h(1 - \beta_2)}{G^2 d} \sum_{i=1}^{h} \mathbb{E} \langle [\nabla \Psi(\theta_t)]_i, [\hat{g}_t]_i - [g_t]_i \rangle \\
\overset{(a)}{\leq} -\eta_c t \sqrt{\frac{h(1 - \beta_2)}{G^2 d}} \frac{1}{2} \mathbb{E} \|\nabla \Psi(\theta_t)\|^2 + \eta_c t \sqrt{\frac{h(1 - \beta_2)}{G^2 d}} \left( \varepsilon + \frac{(1 + \lambda)\sigma^2}{MB} \right) \\
\quad + \frac{h(1 - \beta_2)}{G^2 d} \mathbb{E} \|\nabla \Psi(\theta_t)\|^2 + \eta_c t \sqrt{\frac{h(1 - \beta_2)}{G^2 d}} \left( \varepsilon + \frac{(1 + \lambda)\sigma^2}{MB} \right) \\
\overset{(b)}{\leq} -\eta_c t \sqrt{\frac{h(1 - \beta_2)}{G^2 d}} \frac{1}{2} \mathbb{E} \|\nabla \Psi(\theta_t)\|^2 + \eta_c t \sqrt{\frac{h(1 - \beta_2)}{G^2 d}} \left( \varepsilon + \frac{(1 + \lambda)\sigma^2}{MB} \right) \\
\quad + \eta_c t \sqrt{\frac{h(1 - \beta_2)}{G^2 d}} \delta^2
\]

where we use the in \((a)\) we use Young's inequality (with parameter 2), and in \((b)\) we have

\[
\mathbb{E} \|\hat{g}_t - g_t\|^2 = \mathbb{E} \left\| \frac{1}{M} \sum_{i=1}^{M} Q(g_t^{(i)}) - g_t^{(i)} \right\|^2 \overset{A4}{\leq} \delta^2.
\]  

Second, we give the upper of \( \mathcal{V} \):

\[
\mathcal{V} \leq \eta_c u \sum_{i=1}^{h} \sum_{j=1}^{d_i} \mathbb{E} \text{sign}([\nabla \Psi(\theta_t)]_i)_j \neq \text{sign}([\hat{g}_t]_{ij}) \\
\quad := \mathcal{W}
\]

where the upper bound of \( \mathcal{W} \) can be quantified by using Markov's inequality followed by Jensen's inequality as the
following:

\[ \mathcal{W} = \mathbb{P} (\text{sign}([\nabla \Psi(\theta)_t])_{ij} \neq \text{sign}([\hat{g}_t]_{ij})) \]

\[ \leq \mathbb{E} \left[ \frac{||\nabla \Psi(\theta)_t||_{ij}}{||\nabla \Psi(\theta)_t||_{ij}} \right] \]

\[ \leq \sqrt{\mathbb{E} \left[ \frac{(||\nabla \Psi(\theta)_t||_{ij}^2)}{||\nabla \Psi(\theta)_t||_{ij}} \right]} \]

\[ \leq \sqrt{\frac{\sum_{i=1}^{h} \frac{\sigma^2_{d_i} + \varepsilon_{d_i} + \delta^2_{i} + \eta^2\sigma^2}{MB}}{\sum_{i=1}^{h} L_i}} \]  

(A73)

where (a) is true due to the following relations: i) from (A51), we have

\[ \mathbb{E}[\left([\hat{g}_t]_{ij} - [g^*_t]_{ij}\right)^2] \leq \frac{(1 + \lambda)\sigma^2_{ij}}{MB}; \]  

(A74)

ii) from (A49), we can get

\[ \mathbb{E}[\left([g_t]_{ij} - [g^*_t]_{ij}\right)^2] \leq \varepsilon_{ij}; \]  

(A75)

and iii) from (A67), we know

\[ \mathbb{E}[\left([\hat{g}_t]_{ij} - [g_t]_{ij}\right)^2] \leq \delta^2_{ij}. \]  

(A76)

Therefore, combining (A55) with the upper bound of \( \mathcal{U} \) shown in (A66) and \( \mathcal{V} \) shown in (A68)-(A73), we have

\[ \mathbb{E}[\Psi(\theta_{t+1})] \leq \mathbb{E}[\Psi(\theta_t)] - \eta_t c_t \sqrt{h(1 - \beta_2) \frac{1}{4} \mathbb{E}[\nabla \Psi(\theta)^2] + \frac{h(1 - \beta_2)}{G^2 d} \left( \varepsilon + \frac{(1 + \lambda)\sigma^2}{MB} \right)} \]

\[ + \eta_t c_t \sqrt{h(1 - \beta_2) \frac{1}{G^2 d} \delta^2 + \frac{h}{MB} \sum_{i=1}^{h} \sum_{j=1}^{d_i} \sqrt{\frac{(1 + \lambda)\sigma^2_{ij}}{MB}} + \varepsilon_{ij} + \delta^2_{ij} + \frac{\eta_t^2 \sigma^2}{2} \sum_{i=1}^{h} L_i}. \]  

(A77)

Note that the error vector \( \chi \) is defined as the following

\[ \chi = \begin{bmatrix} \sqrt{\frac{(1 + \lambda)\sigma^2_{11}}{M|B|} + \varepsilon_{11} + \delta^2_{11}} \\ \vdots \\ \sqrt{\frac{(1 + \lambda)\sigma^2_{ij}}{M|B|} + \varepsilon_{ij} + \delta^2_{ij}} \\ \vdots \\ \sqrt{\frac{(1 + \lambda)\sigma^2_{hdh}}{M|B|} + \varepsilon_{hdh} + \delta^2_{hdh}} \end{bmatrix} \in \mathbb{R}^d, \]  

(A78)

and we have

\[ L = \begin{bmatrix} L_1 \\ \vdots \\ L_h \end{bmatrix} \in \mathbb{R}^h. \]  

(A79)

Recall

\[ \kappa = \frac{c_u}{c_l}. \]  

(A80)
Rearranging the terms, we can arrive at

\[
\sqrt{h(1 - \beta^2)} \frac{1}{G^2d} \left( \|\nabla \Psi(\theta_t)\|^2 \right) \leq \frac{E[\Psi(\theta_t)] - E[\Psi(\theta_{t+1})]}{\eta_t c_l} + 4C\delta^2 + 2C \left( \varepsilon + \frac{(1 + \lambda)\sigma^2}{MB} \right) + \sqrt{3}\kappa \|\chi\|_1 + \frac{\eta_t c_w\kappa\|L\|_1}{2C}.
\]

Applying the telescoping sum over \( t = 1, \ldots, T \), we have

\[
\frac{1}{T} \sum_{t=1}^{T} E[\nabla \theta \Psi(\theta_t)]^2 \leq \frac{E[\Psi(\theta_1)] - E[\Psi(\theta_T + 1)]}{\eta_t c_l C T} + 2 \left( \varepsilon + \frac{(1 + \lambda)\sigma^2}{MB} \right) + 4\delta^2 + \frac{\kappa\sqrt{3}}{C} \|\chi\|_1 + \frac{\eta_t c_w\kappa\|L\|_1}{2C}.
\]

\[\text{(A82)}\]

4 ADDITIONAL EXPERIMENTS

4.1 TRAINING DETAILS

ImageNet AT and Fast AT experiments are conducted at a single computing node with dual 22-core CPU, 512GB RAM and 6 Nvidia V100 GPUs. The training epoch is 30 by calling for the momentum SGD optimizer. The weight decay and momentum parameters are set to 0.0001 and 0.9. The initial learning rate is set to 0.1 (tuned over \{0.01, 0.05, 0.1, 0.2\}), which is decayed by \( \times 1/10 \) at the training epoch 20, 25, 28, respectively.

ImageNet DAT experiments are conducted at \{1, 3, 6\} computing nodes with dual 22-core CPU, 512GB RAM and 6 Nvidia V100 GPUs. The training epoch is 30 by calling for the LAMB optimizer. The weight decay is set to 0.0001. \( \beta_1 \) and \( \beta_2 \) are set to 0.9 and 0.999. The initial learning rate \( \eta_1 \) is tuned over \{0.01, 0.05, 0.1, 0.2\}, which is decayed by \( \times 1/10 \) at the training epoch 20, 25, 28, respectively. To execute algorithms with the initial learning rate \( \eta_1 \) greater than 0.2, we choose the model weights after 5-epoch warm-up as its initialization for DAT, where each warm-up epoch \( k \) uses the linearly increased learning rate \( (k/5)\eta_1 \).

4.2 ADDITIONAL RESULTS

Discussion on cyclic learning rate. It was shown in [Wong et al., 2020] that the use of a cyclic learning rate (CLR) trick can further accelerate the Fast AT algorithm in the small-batch setting [Wong et al., 2020]. In Figure A1, we present the performance of Fast AT with CLR versus batch sizes. We observe that when CLR meets the large-batch setting, it becomes significantly worse than its performance in the small-batch setting. The reason is that CLR requires a certain number of iterations to proceed with the cyclic schedule. However, the use of large data batch only results in small amount of iterations by fixing the number of epochs.

Additional details on HPC setups. To further reduce communication cost, we also conduct DAT at a HPC cluster. The computing nodes of the cluster are connected with InfiniBand (IB) and PCIe Gen4 switch. To compare with results in Table ??, we use 6 of 57 nodes of the cluster. Each node has 6 Nvidia V100s which are interconnected with NVLink. We use Nvidia NCCL as communication backend. In Table ??, we have presented the performance of DAT for ImageNet, ResNet-50 with use of HPC compared to standard (non-HPC) distributed system.

References


Figure A1: TA/RA of Fast AT with CLR versus batch sizes on (CIFAR-10, ResNet-18).


