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

Gaoyuan Zhang[1,*] **Songtao Lu**[1,*] **Yihua Zhang**[2] **Xiangyi Chen**[3] **Pin-Yu Chen**[1] **Quanfu Fan**[1] **Lee Martie**[1]
Lior Horesh[1] **Mingyi Hong**[3] **Sijia Liu**[1,2]

[1]IBM Research
[2]Michigan State University
[3]University of Minnesota
[*]Equal Contribution

1 DAT ALGORITHM FRAMEWORK

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

Initial $\boldsymbol{\theta}_1$, dataset $\mathcal{D}^{(i)}$ for each of M workers, and T iterations
for Iteration $t = 1, 2, \ldots, T$ **do**
  for Worker $i = 1, 2, \ldots, M$ **do** ▷ Worker
    Draw a finite-size data batch $\mathcal{B}_t^i \subseteq \mathcal{D}^{(i)}$
    For each data sample $\mathbf{x} \in \mathcal{B}_t^i$, call for an *inner maximization oracle*:

$$\boldsymbol{\delta}_t^{(i)}(\mathbf{x}) := \underset{\|\boldsymbol{\delta}\|_\infty \leq \epsilon}{\arg\max}\ \phi(\boldsymbol{\theta}_t, \boldsymbol{\delta}; \mathbf{x}), \tag{A1}$$

    where we omit the label or possible pseudo-label y of $\mathbf{x}$ for brevity
    Computing local gradient of f_i in (**??**) with respect to $\boldsymbol{\theta}$ given perturbed samples:

$$\mathbf{g}_t^{(i)} = \lambda \mathbb{E}_{\mathbf{x} \in \mathcal{B}_t^{(i)}}[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_t; \mathbf{x})] + \mathbb{E}_{\mathbf{x} \in \mathcal{B}_t^{(i)}}[\nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}_t; \mathbf{x} + \boldsymbol{\delta}_t^{(i)}(\mathbf{x}))] \tag{A2}$$

    (*Optional*) Call for *gradient quantizer* $Q(\cdot)$ and transmit $Q(\mathbf{g}_t^{(i)})$ to server
  end for
  Gradient aggregation at server: ▷ Server

$$\hat{\mathbf{g}}_t = \frac{1}{M} \sum_{i=1}^{M} Q(\mathbf{g}_t^{(i)}) \tag{A3}$$

  (*Optional*) Call for *gradient quantizer* $\hat{\mathbf{g}}_t \leftarrow Q(\hat{\mathbf{g}}_t)$, and transmit $\hat{\mathbf{g}}_t$ to workers:
  for Worker $i = 1, 2, \ldots, M$ **do** ▷ Worker
    Call for an *outer minimization oracle* $\mathcal{A}(\cdot)$ to update $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_{t+1} = \mathcal{A}(\boldsymbol{\theta}_t \hat{\mathbf{g}}_t, \eta_t), \qquad \eta_t \text{ is learning rate} \tag{A4}$$

  end for
end for

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

$$Q(g_i) = \|\mathbf{g}\|_2 \cdot \operatorname{sign}(g_i) \cdot \xi_i(g_i, s), \quad \forall i \in \{1, 2, \ldots, d\}. \tag{A5}$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

In (A5), $\xi_i(g_i, s)$ is a random number drawn as follows. Given $|g_i|/\|\mathbf{g}\|_2 \in [l/s, (l+1)/s]$ for some $l \in \mathbb{N}^+$ and $0 \le l < s$, we then have

$$\xi_i(g_i, s) = \begin{cases} l/s & \text{with probability } 1 - (s|g_i|/\|\mathbf{g}\|_2 - l) \\ (l+1)/s & \text{with probability } (s|g_i|/\|\mathbf{g}\|_2 - l), \end{cases} \tag{A6}$$

where $|a|$ denotes the absolute value of a scalar a, and $\|\mathbf{a}\|_2$ denotes the ℓ_2 norm of a vector $\mathbf{a}$. The rationale behind using (A5) is that $Q(g_i)$ is an *unbiased* estimate of g_i, namely, $\mathbb{E}_{\xi_i(g_i,s)}[Q(g_i)] = g_i$, with bounded variance. Moreover, we at most need $(32 + d + bd)$ bits to transmit the quantized $Q(\mathbf{g})$, where 32 bits for $\|\mathbf{g}\|_2$, 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 $\mathbf{g}$. 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:

$$\Phi_i(\boldsymbol{\theta}, \mathbf{x}) = \max_{\|\boldsymbol{\delta}^{(i)}\|_\infty \le \epsilon} \phi(\boldsymbol{\theta}, \boldsymbol{\delta}^{(i)}; \mathbf{x}), \quad \text{and} \quad \Phi_i(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\in\mathcal{D}^{(i)}} \Phi_i(\boldsymbol{\theta}; \mathbf{x}). \tag{A7}$$

We also define

$$l_i(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\in\mathcal{D}^{(i)}} l(\boldsymbol{\theta}; \mathbf{x}), \tag{A8}$$

where the label y of $\mathbf{x}$ is omitted for labeled data. Then, the objective function of problem (**??**) can be expressed in the compact way

$$\Psi(\boldsymbol{\theta}) = \frac{1}{M} \sum_{i=1}^{M} \lambda l_i(\boldsymbol{\theta}) + \Phi_i(\boldsymbol{\theta}) \tag{A9}$$

and the optimization problem is then given by $\min_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta})$.

Therefore, it is clear that if a point $\boldsymbol{\theta}^\star$ satisfies

$$\|\nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}^\star)\| \le \xi, \tag{A10}$$

then we say $\boldsymbol{\theta}^\star$ is a ξ approximate first-order stationary point (FOSP) of problem (**??**).

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

$$\|\nabla_i \Psi(\boldsymbol{\theta}_{\cdot,i}) - \nabla_i \Psi(\boldsymbol{\theta}'_{\cdot,i})\| \le L_i \|\boldsymbol{\theta}_{\cdot,i} - \boldsymbol{\theta}'_{\cdot,i}\|, \forall i \in [h]. \tag{A11}$$

where $\nabla_i \Psi(\boldsymbol{\theta}_{\cdot,i})$ denotes the gradient w.r.t. the variables at the ith layer. Also, we assume that $\Psi(\boldsymbol{\theta})$ is lower bounded, i.e., $\Psi^\star := \min_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}) > -\infty$ and bounded gradient estimate, i.e., $\|\nabla \hat{\mathbf{g}}_t^{(i)}\| \le G$.

A2. Assume that $\phi(\boldsymbol{\theta}, \boldsymbol{\delta}; \mathbf{x})$ is strongly concave with respect to $\boldsymbol{\delta}$ with parameter μ and has the following gradient Lipschitz continuity with constant L_ϕ:

$$\|\nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}, \boldsymbol{\delta}; \mathbf{x}) - \nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}, \boldsymbol{\delta}'; \mathbf{x})\| \le L_\phi \|\boldsymbol{\delta} - \boldsymbol{\delta}'\|. \tag{A12}$$

A3. Assume that the gradient estimate is unbiased and has bounded variance, i.e.,

$$\mathbb{E}_{\mathbf{x}\in\mathcal{B}^{(i)}}[\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}; \mathbf{x})] = \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}), \forall i, \tag{A13}$$

$$\mathbb{E}_{\mathbf{x}\in\mathcal{B}^{(i)}}[\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}; \mathbf{x})] = \nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}), \forall i, \tag{A14}$$

where recall that $\mathcal{B}^{(i)}$ denotes a data batch used at worker i, $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) := \frac{1}{M}\sum_{i=1}^{M} \nabla_{\boldsymbol{\theta}} l_i(\boldsymbol{\theta})$ and $\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}) := \frac{1}{M}\sum_{i=1}^{M} \nabla_{\boldsymbol{\theta}} \Phi_i(\boldsymbol{\theta})$; and

$$\mathbb{E}_{\mathbf{x}\in\mathcal{B}^{(i)}} \|\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}; \mathbf{x}) - \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})\|^2 \le \sigma^2, \forall i \tag{A15}$$

$$\mathbb{E}_{\mathbf{x}\in\mathcal{B}^{(i)}} \|\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}; \mathbf{x}) - \nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta})\|^2 \le \sigma^2, \forall i. \tag{A16}$$

Further, we define a component-wise bounded variance of the gradient estimate

$$\mathbb{E}_{\mathbf{x}\in\mathcal{B}^{(i)}}\|[\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta};\mathbf{x})]_{jk} - [\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})]_{jk}\|^2 \leq \sigma_{jk}^2, \forall i, \tag{A17}$$

$$\mathbb{E}_{\mathbf{x}\in\mathcal{B}^{(i)}}\|[\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta};\mathbf{x})]_{jk} - [\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta})]_{jk}\|^2 \leq \sigma_{jk}'^2, \forall i, \tag{A18}$$

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

$$\mathbb{E}[(Q([\mathbf{g}^{(i)}(\boldsymbol{\theta})]_{jk}) - [\mathbf{g}^{(i)}(\boldsymbol{\theta})]_{jk})^2] \leq \delta_{jk}^2, \forall i. \tag{A19}$$

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, $\Phi_i(\boldsymbol{\theta};\mathbf{x}), \forall i$ 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 $\boldsymbol{\delta}(\mathbf{x})$ satisfies*

$$\max_{\boldsymbol{\delta}\leq\|\epsilon\|}\langle\boldsymbol{\delta}-\boldsymbol{\delta}^*(\mathbf{x}), \nabla_{\boldsymbol{\delta}}\phi(\boldsymbol{\theta},\boldsymbol{\delta}^*(\mathbf{x});\mathbf{x})\rangle \leq \varepsilon \tag{A20}$$

then, it is a ε approximate solution to $\boldsymbol{\delta}^*(\mathbf{x})$, where

$$\boldsymbol{\delta}^*(\mathbf{x}) := \arg\max_{\boldsymbol{\delta}\leq\|\epsilon\|}\phi(\boldsymbol{\theta},\boldsymbol{\delta};\mathbf{x}). \tag{A21}$$

and $\mathbf{x}$ 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 $\phi(\boldsymbol{\theta},\boldsymbol{\delta}(\mathbf{x});\mathbf{x})$ at $\boldsymbol{\delta}(\mathbf{x})$ and $\boldsymbol{\delta}^*(\mathbf{x})$ 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 $\boldsymbol{\delta}_t^{(k)}$ be the $(\mu\varepsilon)/L_\phi^2$ approximate solution of the inner maximization problem for worker k, i.e., $\max_{\boldsymbol{\delta}^{(k)}}\phi(\boldsymbol{\theta},\boldsymbol{\delta}^{(k)};\mathbf{x}_t)$, where $\mathbf{x}_t$ denotes the sampled data at the tth iteration of DAT. Under A2, we have*

$$\left\|\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}_t,\boldsymbol{\delta}_t^{(k)}(\mathbf{x}_t);\mathbf{x}_t\right) - \nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}_t,(\boldsymbol{\delta}^*)_t^{(k)}(\mathbf{x}_t);\mathbf{x}_t\right)\right\|^2 \leq \varepsilon. \tag{A22}$$

Throughout the convergence analysis, we assume that $\boldsymbol{\delta}_t^{(k)}(\mathbf{x}_t), \forall k, t$ are all the $(\mu\varepsilon)/L_\phi^2$ approximate solutions of the inner maximization problem. Let us define

$$\left\|[\nabla\phi(\boldsymbol{\theta}_t,\boldsymbol{\delta}_t^{(k)}(\mathbf{x}_t);\mathbf{x}_t)]_{ij} - [\nabla\phi(\boldsymbol{\theta}_t,(\boldsymbol{\delta}^*)_t^{(k)}(\mathbf{x}_t);\mathbf{x}_t]_{ij}\right\|^2 = \varepsilon_{ij}. \tag{A23}$$

From Lemma 1, we know that when $\boldsymbol{\delta}_t^{(k)}(\mathbf{x}_t)$ is a $(\mu\varepsilon)/L_\phi^2$ approximate solution, then

$$\sum_{i=1}^{h}\sum_{j=1}^{d_i}\varepsilon_{ij} = \sum_{i=1}^{h}\sum_{j=1}^{d_i}\left\|[\nabla\phi(\boldsymbol{\theta}_t,\boldsymbol{\delta}_t^{(k)}(\mathbf{x}_t);\mathbf{x}_t)]_{ij} - [\nabla\phi(\boldsymbol{\theta}_t,(\boldsymbol{\delta}^*)_t^{(k)}(\mathbf{x}_t);\mathbf{x}_t]_{ij}\right\|^2 \leq \varepsilon. \tag{A24}$$

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 $\{\boldsymbol{\theta}_t\}$ 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*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta}_t)\|^2 \leq \frac{\Delta_\Psi}{\eta_t c_l C T} + 2\left(\varepsilon + \frac{(1+\lambda)\sigma^2}{MB}\right) + 4\delta^2 + \frac{\kappa\sqrt{3}}{C}\|\boldsymbol{\chi}\|_1 + \frac{\eta_t c_u \kappa \|L\|_1}{2C}. \tag{A25}$$

*where $\Delta_\Psi := \mathbb{E}[\Psi(\boldsymbol{\theta}_1)] - \Psi^\star]$, η_t is the learning rate, $\kappa = c_u/c_l$, c_l and c_u are constants used in LALR (**??**), $\boldsymbol{\chi}$ is an error term with the $(ih+j)$th entry being $\sqrt{\frac{(1+\lambda)\sigma_{ij}^2}{MB} + \varepsilon_{ij} + \delta_{ij}^2}$, ε and ε_{ij} were given in (A24), $L = [L_1, \ldots, L_h]^T$, $C = \frac{1}{4}\sqrt{\frac{h(1-\beta_2)}{G^2 d}}$, $0 < \beta_2 < 1$ is given in LAMB, $B = \min\{|\mathcal{B}^{(i)}|, \forall i\}$, 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 $\mathcal{O}(\sigma^2/\sqrt{T})$. Further, it is worth noting that different from the convergence results of LAMB, there is a linear speedup of deceasing the gradient estimate error in DAT with respect to M, i.e., $\mathcal{O}(\sigma^2/(M\sqrt{T}))$, which is the advantage of using multiple computing nodes.

Remark 2. Note that A4 implies $\mathbb{E}[(Q([\mathbf{g}^{(k)}(\boldsymbol{\theta})]_{ij}) - [\mathbf{g}^{(k)}(\boldsymbol{\theta})]_{ij}\|^2] \leq \sum_{i=1}^{h}\sum_{j=1}^{d_i}\delta_{ij}^2 := \delta^2$. 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 **??**.

Corollary 1. *Under the same conditions of Theorem 1, if we choose*

$$\eta_t \sim \mathcal{O}(1/\sqrt{T}), \quad \varepsilon \sim \mathcal{O}(\xi^2), \tag{A26}$$

we then have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta}_t)\|^2 \leq \frac{\Delta_\Psi}{c_l C\sqrt{T}} + \frac{(1+\lambda)\sigma^2}{MB} + \frac{c_u\kappa\|L\|_1}{2C\sqrt{T}} + \mathcal{O}\left(\xi, \frac{\sigma}{\sqrt{MT}}, \min\left\{\frac{d}{4^b}, \frac{\sqrt{d}}{2^b}\right\}\right). \tag{A27}$$

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 σ^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 ϵ is

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \frac{1}{2\epsilon}\|\mathbf{x}\|^2 + \frac{\epsilon}{2}\|\mathbf{y}\|^2, \tag{A28}$$

where $\mathbf{x}, \mathbf{y}$ are two vectors.

2. Define the historical trajectory of the iterates as $\mathcal{F}_t = \{\boldsymbol{\theta}_{t-1}, \ldots, \boldsymbol{\theta}_1\}$.

3. We denote vector $[\mathbf{x}]_i$ as the parameters at the ith layer of the neural net and $[\mathbf{x}]_{ij}$ represents the jth entry of the parameter at the ith layer.

4. We define

$$\mathbf{g}_t := \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{\mathbf{x}_t \in \mathcal{B}^{(i)}}\left(\lambda\nabla l(\boldsymbol{\theta}_t; \mathbf{x}_t) + \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}_t, \boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t)\right) = \frac{1}{M}\sum_{i=1}^{M}\mathbf{g}_t^{(i)}. \tag{A29}$$

3.2 DETAILS OF LAMB ALGORITHM

Algorithm A2 LAMB [You et al., 2019]

Input: learning rate η_t, $0 < \beta_1, \beta_2 < 1$, scaling function $\tau(\cdot)$, $\zeta > 0$
for $t = 1, \dots$ **do**
  $\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\hat{\mathbf{g}}_t$, where $\hat{\mathbf{g}}_t$ is given by (A3)
  $\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\hat{\mathbf{g}}_t^2$
  $\mathbf{m}_t = \mathbf{m}_t/(1-\beta_1^t)$
  $\mathbf{v}_t = \mathbf{v}_t/(1-\beta_2^t)$
  Compute ratio $\mathbf{u}_t = \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t}+\zeta}$
end for
Update

$$\boldsymbol{\theta}_{t+1,i} = \boldsymbol{\theta}_{t,i} - \frac{\eta_t \tau(\|\boldsymbol{\theta}_{t,i}\|)}{\|\mathbf{u}_{t,i}\|}\mathbf{u}_{t,i}. \tag{A30}$$

3.3 PROOF OF LEMMA 1

Proof. From A2, we have

$$\left\|\nabla\phi\left(\boldsymbol{\theta}_t, \boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t\right) - \nabla\phi\left(\boldsymbol{\theta}_t, (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t\right)\right\| \le L_\phi \|\boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t) - (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t)\|. \tag{A31}$$

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

$$\begin{aligned}&\mu\|\boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t) - (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t)\| \\ &\qquad\le \left\langle \nabla_{\boldsymbol{\delta}}\phi(\boldsymbol{\theta}_t, (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t) - \nabla_{\boldsymbol{\delta}}\phi(\boldsymbol{\theta}_t, \boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t), \boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t) - (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t)\right\rangle.\end{aligned} \tag{A32}$$

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

$$\left\langle (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t) - \boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t), \nabla_{\boldsymbol{\delta}}\phi(\boldsymbol{\theta}_t, \boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t)\right\rangle \le \varepsilon. \tag{A33}$$

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

$$\left\langle (\boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t) - (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t), \nabla_{\boldsymbol{\delta}}\phi(\boldsymbol{\theta}_t, (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t)\right\rangle \le 0. \tag{A34}$$

Adding them together, we can obtain

$$\left\langle \boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t) - (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t), \nabla_{\boldsymbol{\delta}}\phi(\boldsymbol{\theta}_t, (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t) - \nabla_{\boldsymbol{\delta}}\phi(\boldsymbol{\theta}_t, \boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t)\right\rangle \le \varepsilon. \tag{A35}$$

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

$$\mu\|\boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t) - (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t)\|^2 \le \varepsilon. \tag{A36}$$

Combining (A31), we have

$$\left\|\nabla\phi(\boldsymbol{\theta}_t, \boldsymbol{\delta}_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t) - \nabla\phi(\boldsymbol{\theta}_t, (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t)\right\|^2 \le L_\phi^2 \frac{\varepsilon}{\mu}. \tag{A37}$$

□

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 $\{\boldsymbol{\theta}_t\}$ is generated by DAT. Then, we have*

$$\mathbb{E}[-\langle \nabla\Psi(\boldsymbol{\theta}_t), \hat{\mathbf{g}}_t\rangle] \leq -\frac{\mathbb{E}\|\nabla\Psi(\boldsymbol{\theta}_t)\|^2}{2} + \varepsilon + \frac{(1+\lambda)\sigma^2}{MB}. \tag{A38}$$

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

$$\nabla_{\boldsymbol{\theta}}\Phi_i(\boldsymbol{\theta}, \mathbf{x}) = \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}, (\boldsymbol{\delta}^*)^{(i)}(\mathbf{x}); \mathbf{x}), \tag{A39}$$

so we can get

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta}) =& \frac{1}{M}\sum_{i=1}^{M} \lambda\nabla_{\boldsymbol{\theta}} l_i(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\Phi_i(\boldsymbol{\theta}) && \text{(A40)}\\
=& \lambda\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) + \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{\mathbf{x}\in\mathcal{D}^{(i)}}\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}, (\boldsymbol{\delta}^*)^{(i)}(\mathbf{x}); \mathbf{x}) && \text{(A41)}\\
:=& \bar{\mathbf{g}}(\boldsymbol{\theta}). && \text{(A42)}
\end{aligned}$$

Then, we have

$$\begin{aligned}
\mathbb{E}\langle\nabla\Psi(\boldsymbol{\theta}_t), \mathbf{g}_t\rangle =& \mathbb{E}\langle\nabla\Psi(\boldsymbol{\theta}_t), \bar{\mathbf{g}}_t\rangle + \mathbb{E}\langle\nabla\Psi(\boldsymbol{\theta}_t), \mathbf{g}_t - \bar{\mathbf{g}}_t\rangle && \text{(A43)}\\
=& \mathbb{E}_{\mathcal{F}_t}\mathbb{E}_{\mathbf{x}_t|\mathcal{F}_t}\langle\nabla\Psi(\boldsymbol{\theta}_t), \bar{\mathbf{g}}_t\rangle + \mathbb{E}\langle\nabla\Psi(\boldsymbol{\theta}_t), \mathbf{g}_t - \bar{\mathbf{g}}_t\rangle && \text{(A44)}\\
\overset{(A42)}{=}& \mathbb{E}\|\nabla\Psi(\boldsymbol{\theta}_t)\|^2 + \mathbb{E}\langle\nabla\Psi(\boldsymbol{\theta}_t), \mathbf{g}_t - \bar{\mathbf{g}}_t\rangle && \text{(A45)}\\
=& \mathbb{E}\|\nabla\Psi(\boldsymbol{\theta}_t)\|^2 + \mathbb{E}\langle\nabla\Psi(\boldsymbol{\theta}_t), \mathbf{g}_t - \mathbf{g}_t^*\rangle + \mathbb{E}\langle\nabla\Psi(\boldsymbol{\theta}_t), \mathbf{g}_t^* - \bar{\mathbf{g}}_t\rangle && \text{(A46)}
\end{aligned}$$

where

$$\bar{\mathbf{g}}_t := \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{\mathbf{x}_t\in\mathcal{D}^{(i)}}\left(\lambda\nabla l(\boldsymbol{\theta}_t, \mathbf{x}_t) + \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}_t, (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t)\right) = \lambda\nabla l(\boldsymbol{\theta}_t) + \nabla\Phi(\boldsymbol{\theta}_t), \tag{A47}$$

and

$$\mathbf{g}_t^* := \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{\mathbf{x}_t\in\mathcal{B}^{(i)}}\left(\lambda\nabla l(\boldsymbol{\theta}_t, \mathbf{x}_t) + \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}_t, (\boldsymbol{\delta}^*)_t^{(i)}(\mathbf{x}_t); \mathbf{x}_t)\right). \tag{A48}$$

Next, we can quantify the different between $\mathbf{g}_t$ and $\mathbf{g}_t^*$ by gradient Lipschitz continuity of function $\tau(\cdot)$ as the following

$$\mathbb{E}\|\mathbf{g}_t - \mathbf{g}_t^*\|^2 \overset{(a)}{\leq} \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{\mathcal{F}_t}\mathbb{E}_{\mathbf{x}_t|\mathcal{F}_t}\left[\|\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}_t, (\boldsymbol{\delta}^*)^{(i)}(\mathbf{x}_t); \mathbf{x}_t) - \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}_t, \boldsymbol{\delta}^{(i)}(\mathbf{x}_t); \mathbf{x}_t)\|^2\right] \overset{(A24)}{\leq} \varepsilon \tag{A49}$$

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

And the difference between $\bar{\mathbf{g}}_t$ and $\mathbf{g}_t^*$ can be upper bounded by

$$\begin{aligned}
\mathbb{E}\|\bar{\mathbf{g}}_t - \mathbf{g}_t^*\|^2 =& \mathbb{E}_{\mathcal{F}_t}\left\|\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{\mathbf{x}_t|\mathcal{F}_t}\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}_t, (\boldsymbol{\delta}^*)^{(i)}(\mathbf{x}_t); \mathbf{x}_t) - \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}_t)\right\|^2\\
&+ \lambda\mathbb{E}_{\mathcal{F}_t}\left\|\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{\mathbf{x}_t|\mathcal{F}_t}\nabla l(\boldsymbol{\theta}_t; \mathbf{x}_t) - \nabla l(\boldsymbol{\theta}_t)\right\|^2 && \text{(A50)}\\
\overset{A3}{=}& \frac{(1+\lambda)\sigma^2}{MB}. && \text{(A51)}
\end{aligned}$$

Applying Young's inequality with parameter 2, we have

$$\mathbb{E}[-\langle \nabla\Psi(\boldsymbol{\theta}_t), \mathbf{g}_t\rangle] \leq -\mathbb{E}\|\nabla\Psi(\boldsymbol{\theta}_t)\|^2 + \frac{\mathbb{E}\|\nabla\Psi(\boldsymbol{\theta}_t)\|^2}{2} + \mathbb{E}\|\bar{\mathbf{g}}_t - \mathbf{g}_t^*\|^2 + \mathbb{E}\|\mathbf{g}_t^* - \mathbf{g}_t\|^2 \tag{A52}$$

$$\overset{(A49)}{\leq} -\frac{\mathbb{E}\|\nabla\Psi(\boldsymbol{\theta}_t)\|^2}{2} + \varepsilon + \frac{(1+\lambda)\sigma^2}{MB}. \tag{A53}$$

□

3.5 PROOF OF THEOREM 1

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

$$\Psi(\boldsymbol{\theta}_{t+1}) \overset{A1}{\leq} \Psi(\boldsymbol{\theta}_t) + \sum_{i=1}^{h}\langle[\nabla_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta}_t)]_i, \boldsymbol{\theta}_{t+1,i} - \boldsymbol{\theta}_{t,i}\rangle + \sum_{i=1}^{h}\frac{L_i}{2}\|\boldsymbol{\theta}_{t+1,i} - \boldsymbol{\theta}_{t,i}\|^2 \tag{A54}$$

$$\overset{(a)}{\leq} \Psi(\boldsymbol{\theta}_t) \underbrace{-\eta_t \sum_{i=1}^{h}\sum_{j=1}^{d_i}\tau(\|\boldsymbol{\theta}_{t,i}\|)\left\langle [\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}, \frac{[\mathbf{u}_t]_{ij}}{\|\mathbf{u}_{t,i}\|}\right\rangle}_{:=\boldsymbol{\mathcal{R}}} + \sum_{i=1}^{h}\frac{\eta_t^2 c_u^2 L_i}{2}, \tag{A55}$$

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

Next, we split term R as two parts by leveraging $\text{sign}([\nabla\Psi(\boldsymbol{\theta}_t)]_{ij})$ and $\text{sign}([\mathbf{u}_t]_{ij})$ as follows.

$$\begin{aligned}
\boldsymbol{\mathcal{R}} =& -\eta_t \sum_{i=1}^{h}\sum_{j=1}^{d_i}\tau(\|\boldsymbol{\theta}_{t,i}\|)[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}\frac{[\mathbf{u}_t]_{ij}}{\|\mathbf{u}_{t,i}\|}\mathbb{1}\left(\text{sign}([\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}) = \text{sign}([\mathbf{u}_t]_{ij})\right) \\
& -\eta_t \sum_{i=1}^{h}\sum_{j=1}^{d_i}\tau(\|\boldsymbol{\theta}_{t,i}\|)[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}\frac{[\mathbf{u}_t]_{ij}}{\|\mathbf{u}_{t,i}\|}\mathbb{1}\left(\text{sign}([\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}) \neq \text{sign}([\mathbf{u}_t]_{ij})\right) && \text{(A56)} \\
\overset{(a)}{\leq}& -\eta_t c_l \sum_{i=1}^{h}\sum_{j=1}^{d_i}\sqrt{\frac{1-\beta_2}{G^2 d_i}}[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}[\hat{\mathbf{g}}_t]_{ij}\mathbb{1}\left(\text{sign}([\nabla[\Psi(\boldsymbol{\theta}_t)]_{ij}) = \text{sign}([\hat{\mathbf{g}}_t]_{ij})\right) \\
& -\eta_t \sum_{i=1}^{h}\sum_{j=1}^{d_i}\tau(\|\boldsymbol{\theta}_{t,i}\|)[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}\frac{[\mathbf{u}_t]_{ij}}{\|\mathbf{u}_{t,i}\|}\mathbb{1}\left(\text{sign}([\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}) \neq \text{sign}([\mathbf{u}_t]_{ij})\right) && \text{(A57)} \\
\overset{(b)}{\leq}& -\eta_t c_l \sum_{i=1}^{h}\sum_{j=1}^{d_i}\sqrt{\frac{1-\beta_2}{G^2 d_i}}[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}[\hat{\mathbf{g}}_t]_{ij} \\
& -\eta_t \sum_{i=1}^{h}\sum_{j=1}^{d_i}\tau(\|\boldsymbol{\theta}_{t,i}\|)[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}\frac{[\mathbf{u}_t]_{ij}}{\|\mathbf{u}_{t,i}\|}\mathbb{1}\left(\text{sign}([\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}) \neq \text{sign}([\mathbf{u}_t]_{ij})\right). && \text{(A58)}
\end{aligned}$$

where in (a) we use the fact that $\|\mathbf{u}_{t,i}\| \leq \sqrt{\frac{d_i}{1-\beta_2}}$ and $\sqrt{\mathbf{v}_t} \leq G$, and in (b) we add

$$-\eta_t c_l \sum_{i=1}^{h}\sum_{j=1}^{d_i}\sqrt{\frac{1-\beta_2}{G^2 d_i}}[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}[\hat{\mathbf{g}}_t]_{ij}\mathbb{1}\left(\text{sign}([\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}) \neq \text{sign}([\hat{\mathbf{g}}_t]_{ij})\right) \geq 0. \tag{A59}$$

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

$$\mathbb{E}[\boldsymbol{\mathcal{R}}] \leq \underbrace{-\eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \sum_{i=1}^{h} \sum_{j=1}^{d_i} \mathbb{E}[[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}[\hat{\mathbf{g}}_t]_{ij}}_{:=\boldsymbol{\mathcal{U}}}$$
$$\underbrace{+\eta_t c_u \sum_{i=1}^{h} \sum_{j=1}^{d_i} \mathbb{E}\left[[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij} \mathbb{1}\left(\text{sign}([\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}) \neq \text{sign}([\mathbf{u}_t]_{ij})\right)\right]}_{:=\boldsymbol{\mathcal{V}}}. \tag{A60}$$

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

$$\boldsymbol{\mathcal{U}} \leq -\eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \sum_{i=1}^{h} \mathbb{E}\left\langle [\nabla\Psi(\boldsymbol{\theta})]_i, [\hat{\mathbf{g}}_t]_i \right\rangle \tag{A61}$$

$$\leq -\eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \sum_{i=1}^{h} \mathbb{E}\left\langle [\nabla\Psi(\boldsymbol{\theta}_t)]_i, [\hat{\mathbf{g}}_t]_i - [\mathbf{g}_t]_i + [\mathbf{g}_t]_i \right\rangle \tag{A62}$$

$$\leq -\eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \left(\mathbb{E}\left\langle \nabla\Psi(\boldsymbol{\theta}), \mathbf{g}_t \right\rangle + \sum_{i=1}^{h} \mathbb{E}\left\langle [\nabla\Psi(\boldsymbol{\theta}_t)]_i, [\hat{\mathbf{g}}_t]_i - [\mathbf{g}_t]_i \right\rangle \right). \tag{A63}$$

Applying Lemma 2, we can get

$$\boldsymbol{\mathcal{U}} \overset{(A38)}{\leq} -\eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \frac{1}{2} \mathbb{E}\|\nabla\Psi(\boldsymbol{\theta}_t)\|^2 + \eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \left(\varepsilon + \frac{(1+\lambda)\sigma^2}{MB}\right)$$
$$-\eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \sum_{i=1}^{h} \mathbb{E}\left\langle [\nabla\Psi(\boldsymbol{\theta}_t)]_i, [\hat{\mathbf{g}}_t]_i - [\mathbf{g}_t]_i \right\rangle \tag{A64}$$

$$\overset{(a)}{\leq} -\eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \frac{1}{2} \mathbb{E}\|\nabla\Psi(\boldsymbol{\theta}_t)\|^2 + \eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \left(\varepsilon + \frac{(1+\lambda)\sigma^2}{MB}\right)$$
$$+\frac{\eta_t c_l}{4} \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \mathbb{E}\|\nabla\Psi(\boldsymbol{\theta}_t)\|^2 + c_l \eta_t \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \mathbb{E}\|\hat{\mathbf{g}}_t - \mathbf{g}_t\|^2 \tag{A65}$$

$$\overset{(b)}{\leq} -\frac{\eta_t c_l}{4} \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \frac{1}{2} \mathbb{E}\|\nabla\Psi(\boldsymbol{\theta}_t)\|^2 + \eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \left(\varepsilon + \frac{(1+\lambda)\sigma^2}{MB}\right)$$
$$+\eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}} \delta^2 \tag{A66}$$

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

$$\mathbb{E}\|\hat{\mathbf{g}}_t - \mathbf{g}_t\|^2 = \mathbb{E}\left\| \frac{1}{M} \sum_{i=1}^{M} Q(\mathbf{g}_t^{(i)}) - \mathbf{g}_t^{(i)} \right\|^2 \overset{A4}{\leq} \delta^2. \tag{A67}$$

Second, we give the upper of $\boldsymbol{\mathcal{V}}$:

$$\boldsymbol{\mathcal{V}} \leq \eta_t c_u \sum_{i=1}^{h} \sum_{j=1}^{d_i} [\nabla\Psi(\boldsymbol{\theta}_t)]_{ij} \underbrace{\mathbb{P}\left(\text{sign}([\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}) \neq \text{sign}([\hat{\mathbf{g}}_t]_{ij})\right)}_{:=\boldsymbol{\mathcal{W}}} \tag{A68}$$

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

$$
\begin{aligned}
\boldsymbol{\mathcal{W}} =& \mathbb{P}\left(\text{sign}([\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}) \neq \text{sign}([\hat{\mathbf{g}}_t]_{ij})\right) \\
\leq& \mathbb{P}[|[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij} - [\hat{\mathbf{g}}_t]_{ij}| > [\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}] \qquad \text{(A69)} \\
\leq& \frac{\mathbb{E}[[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij} - [\hat{\mathbf{g}}_t]_{ij}]}{|[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}|} \qquad \text{(A70)} \\
\leq& \frac{\sqrt{\mathbb{E}[([\nabla\Psi(\boldsymbol{\theta}_t)]_{ij} - [\hat{\mathbf{g}}_t]_{ij})^2]}}{|[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}|} \qquad \text{(A71)} \\
\overset{(A42)}{\leq}& \frac{\sqrt{\mathbb{E}[([\bar{\mathbf{g}}_t]_{ij} - [\mathbf{g}_t^*]_{ij} + [\mathbf{g}_t^*]_{ij} - [\mathbf{g}_t]_{ij} + [\mathbf{g}_t]_{ij} - [\hat{\mathbf{g}}_t]_{ij})^2]}}{|[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}|} \qquad \text{(A72)} \\
\overset{(a)}{\leq}& \sqrt{3}\frac{\sqrt{\frac{(1+\lambda)\sigma_{ij}^2}{M|\mathcal{B}|} + \epsilon_{ij} + \delta_{ij}^2}}{|[\nabla\Psi(\boldsymbol{\theta}_t)]_{ij}|} \qquad \text{(A73)}
\end{aligned}
$$

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

$$\mathbb{E}[([\bar{\mathbf{g}}_t]_{ij} - [\mathbf{g}_t^*]_{ij})^2] \leq \frac{(1+\lambda)\sigma_{ij}^2}{MB}; \qquad \text{(A74)}$$

ii) from (A49), we can get

$$\mathbb{E}[([\mathbf{g}_t]_{ij} - [\mathbf{g}_t^*]_{ij})^2] \leq \varepsilon_{ij}; \qquad \text{(A75)}$$

and *iii*) from (A67), we know

$$\mathbb{E}[([\hat{\mathbf{g}}_t]_{ij} - [\mathbf{g}_t]_{ij})^2] \leq \delta_{ij}^2. \qquad \text{(A76)}$$

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

$$
\begin{aligned}
\mathbb{E}[\Psi(\boldsymbol{\theta}_{t+1})] \leq& \mathbb{E}[\Psi(\boldsymbol{\theta}_t)] - \eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}}\frac{1}{4}\mathbb{E}\|\nabla\Psi(\boldsymbol{\theta}_t)\|^2 + \eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}}\left(\varepsilon + \frac{(1+\lambda)\sigma^2}{MB}\right) \\
&+ \eta_t c_l \sqrt{\frac{h(1-\beta_2)}{G^2 d}}\delta^2 + \eta_t c_u \sqrt{3}\sum_{i=1}^{h}\sum_{j=1}^{d_i}\sqrt{\frac{(1+\lambda)\sigma_{ij}^2}{MB} + \varepsilon_{ij} + \delta_{ij}^2} + \frac{\eta_t^2 c_u^2 \sum_{i=1}^{h} L_i}{2}. \qquad \text{(A77)}
\end{aligned}
$$

Note that the error vector $\boldsymbol{\chi}$ is defined as the following

$$\boldsymbol{\chi} = \begin{bmatrix} \sqrt{\frac{(1+\lambda)\sigma_{11}^2}{M|\mathcal{B}|} + \varepsilon_{11} + \delta_{11}^2} \\ \vdots \\ \sqrt{\frac{(1+\lambda)\sigma_{ij}^2}{M|\mathcal{B}|} + \varepsilon_{ij} + \delta_{ij}^2} \\ \vdots \\ \sqrt{\frac{(1+\lambda)\sigma_{hd_h}^2}{M|\mathcal{B}|} + \varepsilon_{hd_h} + \delta_{hd_h}^2} \end{bmatrix} \in \mathbb{R}^d, \qquad \text{(A78)}$$

and we have

$$L = \begin{bmatrix} L_1 \\ \vdots \\ L_h \end{bmatrix} \in \mathbb{R}^h. \qquad \text{(A79)}$$

Recall

$$\kappa = \frac{c_u}{c_l}. \qquad \text{(A80)}$$

Rearranging the terms, we can arrive at

$$\underbrace{\sqrt{\frac{h(1-\beta_2)}{G^2 d}}\frac{1}{4}}_{:=C}\left(\|\nabla\Psi(\boldsymbol{\theta}_t)\|^2\right) \leq \frac{\mathbb{E}[\Psi(\boldsymbol{\theta}_t)] - \mathbb{E}[\Psi(\boldsymbol{\theta}_{t+1})]}{\eta_t c_l} + 4C\delta^2 + 2C\left(\varepsilon + \frac{(1+\lambda)\sigma^2}{MB}\right) + \sqrt{3}\kappa\|\boldsymbol{\chi}\|_1 + \frac{\eta_t c_u \kappa \|L\|_1}{2}. \tag{A81}$$

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

$$\frac{1}{T}\sum_{t=1}^{\top}\mathbb{E}\|\nabla_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta}_t)\|^2 \leq \frac{\mathbb{E}[\Psi(\boldsymbol{\theta}_1)] - \mathbb{E}[\Psi(\boldsymbol{\theta}_{T+1})]}{\eta_t c_l C T} + 2\left(\varepsilon + \frac{(1+\lambda)\sigma^2}{MB}\right) + 4\delta^2 + \frac{\kappa\sqrt{3}}{C}\|\boldsymbol{\chi}\|_1 + \frac{\eta_t c_u \kappa \|L\|_1}{2C}. \tag{A82}$$

□

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. β_1 and β_2 are set to 0.9 and 0.999. The initial learning rate η_1 is tuned over $\{0.01, 0.05, 0.1, 0.2, 0.4\}$, which is decayed by $\times 1/10$ at the training epoch $20, 25, 28$, respectively. To execute algorithms with the initial learning rate η_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 a 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 Nvida 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

D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *NeurIPS*, 2017.

S. Lu, M. Razaviyayn, B. Yang, K. Huang, and M. Hong. Finding second-order stationary points efficiently in smooth nonconvex linearly constrained optimization problems. In *NeurIPS*, 2020.

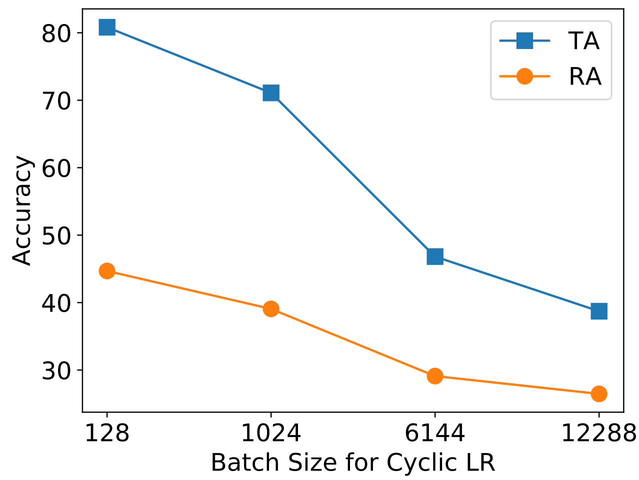

Figure A1: TA/RA of Fast AT with CLR versus batch sizes on (CIFAR-10, ResNet-18).

Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu. On the convergence and robustness of adversarial training. In *ICML*, 2019.

E. Wong, L. Rice, and J. Z. Kolter. Fast is better than free: Revisiting adversarial training. In *ICLR*, 2020.

Y. You, J. Li, et al. Large batch optimization for deep learning: Training bert in 76 minutes. In *ICLR*, 2019.

Y. Yu, J. Wu, and L. Huang. Double quantization for communication-efficient distributed optimization. In *NeurIPS*, 2019.