DISTRIBUTED ASYNCHRONOUS OPTIMIZATION WITH UNBOUNDED DELAYS: HOW SLOW CAN YOU GO? [SUPPLEMENTARY MATERIAL]

ZHENGYUAN ZHOU, PANAYOTIS MERTIKOPOULOS, NICHOLAS BAMBOS, PETER GLYNN, YINYU YE, LI-JIA LI, AND LI FEI-FEI

Abstract. We provide missing proofs and necessary auxiliary results in this appendix.

1. Auxiliary Results

We state two useful auxiliary results that will be used later. The first one is from Bubeck et al. (2015):

Lemma 1.1. *Let $\mathcal{X}$ be a compact and convex subset of $\mathbb{R}^d$. Then for any $x \in \mathcal{X}, y \in \mathbb{R}^d$:*

$$\langle \mathbf{proj}_{\mathcal{X}}(y) - x, \mathbf{proj}_{\mathcal{X}}(y) - y \rangle \leq 0. \tag{1.1}$$

The second result we use is an L_p-bounded martingale convergence theorem:

Lemma 1.2 (Hall & Heyde, 1980)**.** *Let S_n be a martingale adapted to the filtration $\mathcal{S}_n$. If for some $p \geq 1$, $\sup_{t \geq 0} \mathbf{E}[|S_n|^p] < \infty$, then S_n converges almost surely to a random variable S_∞ with $\mathbf{E}[|S_\infty|^p] < \infty$.*

Remark 1.1. Note that $\mathbf{E}[|S_\infty|^p] < \infty$ obviously implies S_∞ is finite almost surely.

2. Problem Setup

For convenience, we restate here our blanket assumptions:

Assumption 1. F satisfies the following:

(1) $F(x;\omega)$ is differentiable in x for $\mathbb{P}$-almost all $\omega \in \Omega$.[1]
(2) $\nabla F(x;\omega)$ has bounded second moment, that is, $\mathbb{E}[\|\nabla F(x;\omega)\|_2^2] < \infty$ for all $x \in \mathcal{X}$.[2]
(3) $\nabla F(x;\omega)$ is Lipschitz continuous in the mean: $\mathbb{E}[\nabla F(x;\omega)]$ is Lipschitz on $\mathcal{X}$.

Assumption 2. The optimization problem is *variationally coherent in the mean*, i.e.,

$$\mathbb{E}[\langle \nabla F(x;\omega), x - x^* \rangle] > 0, \tag{VC}$$

for all $x^* \in \mathcal{X}^*$ and all $x \notin \mathcal{X}^*$.

[1]The results in this paper can be generalized to non-smooth objectives by using subgradient devices instead of gradients. For ease of exposition, we stick with smooth objectives to avoid cumbersome notation needed to deal with subgradients.

[2]It is understood here that the gradient $\nabla F(x;\omega)$ is only taken with respect to x: no differential structure is assumed on Ω.

Assumption 3. The gradient delay process d_n and the step-size sequence α_n of satisfy one of the following conditions:

(1) *Bounded delays:* $\sup_n d_n < \infty$ and $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$, $\sum_{n=1}^{\infty} \alpha_n = \infty$.
(2) *Linearly growing delays:* $d_n = \mathcal{O}(n)$ and $\alpha_n \propto 1/(n \log n)$ for large n.
(3) *Polynomially growing delays:* $d_n = \mathcal{O}(n^q)$ for some $q \geq 1$ and $\alpha_n \propto 1/(n \log n \log\log n)$ for large n.

We also define the following constants that will be handy later in the proofs:

(1) $C_1 = \sup_{x \in \mathcal{X}} \mathbb{E}[\|\nabla F(x;\omega)\|_2]$.
(2) $C_2 = \sup_{x \in \mathcal{X}} \mathbb{E}[\|\nabla F(x;\omega)\|_2^2]$.
(3) C_3 is the Lipschitz constant for $\nabla f(x) (= \mathbb{E}[\nabla F(x;\omega)])$:

$$\| \nabla f(x) - \nabla f(x')\|_2 \leq C_3 \|x - x'\|_2. \tag{2.1}$$

(4) $C_4 = \sup_{x,x' \in \mathcal{X}} \|x - x'\|_2$.

Since $\nabla F(x,\omega)$ has bounded second moment, it must also have bounded first moment. Consequently, since $\mathcal{X}$ is compact and $\mathbb{E}[\nabla F(x;\omega)]$ is continuous, $C_1 < \infty, C_2 < \infty$.

3. Deterministic Convergence Analysis

We begin with the deterministic case, i.e., the distributed asynchronous gradient descent (DAGD) scheme with perfect gradient information.

3.1. Energy Function.

Recall the energy function defined in the main text:

$$E(y) = \|x\|_2^2 - \|\mathbf{proj}_{\mathcal{X}}(y)\|_2^2 + 2\langle y, \mathbf{proj}_{\mathcal{X}}(y) - x\rangle, \tag{3.1}$$

where x is a fixed base point in $\mathcal{X}$. Here we emphasize explicitly its dependence on x and write $E(x,y)$ instead: even though it's more cumbersome notation compared to before, it would also make clear the role played by the specific choice of x later.

Lemma 3.1. *Pick any $x \in \mathcal{X}, y \in \mathbb{R}^d$. We have:*

(1) $E(x,y) \geq 0$ with equality if and only if $\mathbf{proj}_{\mathcal{X}}(y) = x$.
(2) Let $\{y_n\}_{n=1}^{\infty}$ be a sequence. Then for any $x \in \mathcal{X}$, $E(x,y_n) \to 0$ if and only if $\mathbf{proj}_{\mathcal{X}}(y_n) \to x$.

Proof. Per the definition of the energy function, we have:

$$\begin{aligned} E(x,y) - \|\mathbf{proj}_{\mathcal{X}}(y) - x\|_2^2 &= \|x\|_2^2 - \|\mathbf{proj}_{\mathcal{X}}(y)\|_2^2 + 2\langle y, \mathbf{proj}_{\mathcal{X}}(y) - x\rangle - \Big\{\|\mathbf{proj}_{\mathcal{X}}(y)\|_2^2 - 2\langle \mathbf{proj}_{\mathcal{X}}(y), x\rangle + \|\\ &= -2\|\mathbf{proj}_{\mathcal{X}}(y)\|_2^2 + 2\langle y - \mathbf{proj}_{\mathcal{X}}(y), \mathbf{proj}_{\mathcal{X}}(y) - x\rangle \\ &\quad + 2\langle \mathbf{proj}_{\mathcal{X}}(y), \mathbf{proj}_{\mathcal{X}}(y) - x\rangle + 2\langle \mathbf{proj}_{\mathcal{X}}(y), x\rangle \\ &= 2\langle y - \mathbf{proj}_{\mathcal{X}}(y), \mathbf{proj}_{\mathcal{X}}(y) - x\rangle \geq 0, \end{aligned} \tag{3.2}$$

where the last inequality follows from Lemma 1.1. Consequently, $E(x,y) - \|\mathbf{proj}_{\mathcal{X}}(y) - x\|_2^2 = 0$. And if $E(x,y) = 0$, we must have $\|\mathbf{proj}_{\mathcal{X}}(y) - x\|_2^2$, therefore implying $\mathbf{proj}_{\mathcal{X}}(y) = x$. Similarly, if $E(x,y_n) \to 0$, then $\|\mathbf{proj}_{\mathcal{X}}(y_n) - x\|_2^2 \to 0$, thereby implying $\mathbf{proj}_{\mathcal{X}}(y_n) \to x$. ∎

3.2. **Main Convergence Result.**

Proposition 3.2. *Under Assumptions 1–3, DAGD admits a subsequence x_{n_k} that converges to $\mathcal{X}^*$ as $k \to \infty$.*

Proof. We provide the details for all the steps.

(1) Defining $b_n = \nabla f(x_{s(n)}) - \nabla f(x_n)$, we can rewrite the gradient update in DAGD as:

$$\begin{aligned} y_{n+1} &= y_n - \alpha_{n+1} \nabla f(x_{s(n)}) \\ &= y_n - \alpha_{n+1} \nabla f(x_n) - \alpha_{n+1}\{\nabla f(x_{s(n)}) - \nabla f(x_n)\} \\ &= y_n - \alpha_{n+1}(\nabla f(x_n) + b_n). \end{aligned} \tag{3.3}$$

Recall here once again that $s(n)$ denotes the previous iteration count whose gradient becomes available only at the current iteration n. By bounding b_n's magnitude using the delay sequence through a careful analysis, we establish that under each one of the three conditions, $\lim_{n\to\infty} \|b_n\|_2 = 0$. The analysis here, particularly the one for the last two conditions, reveals the following pattern: as the magnitude of the delays gets larger and larger in the order of growth, one needs to use a more and more mild step-size in order to mitigate the damage done by the stale gradient information. Intuitively, smaller step-size is more helpful in larger delays because it has a better "averaging" effect, such that it is more tolerant of the delays.

To see this, we start by expanding b_n as follows:

$$\begin{aligned} \|b_n\|_2 &= \| \nabla f(x_{s(n)}) - \nabla f(x_n)\|_2 \leq L\|x_{s(n)} - x_n\|_2 \\ &= C_3 \| \mathbf{proj}_{\mathcal{X}}(y_{s(n)}) - \mathbf{proj}_{\mathcal{X}}(y_n)\|_2 \leq L\|y_{s(n)} - y_n\|_2 \\ &\leq C_3 \Big\{ \|y_{s(n)} - y_{s(n)+1}\|_2 + \|y_{s(n)+1} - y_{s(n)+2}\|_2 + \cdots + \|y_{n-1} - y_n\|_2 \Big\} \\ &= C_3 \sum_{r=s(n)}^{n-1} \|\alpha_{r+1} \nabla f(x_{s(r)})\|_2 \leq C_3 \sup_{x\in\mathcal{X}} \| \nabla f(x)\|_2 \sum_{r=s(n)}^{n-1} \alpha_{r+1} = C_3 V_{\max} \sum_{r=s(n)}^{n-1} \alpha_{r+1}. \end{aligned} \tag{3.4}$$

We now consider two cases, depending on whether the delays are bounded or not.

(a) If $\{\alpha_n\}_{n=1}^{\infty}$ and $d_n \leq D, \forall n$ satisfy Assumption 3, then $d_{s(n)} = n - s(n) \leq D$. Consequently,

$$0 < \sum_{r=s(n)}^{n-1} \alpha_{r+1} = \sum_{r=s(n)+1}^{n} \alpha_r \leq \sum_{r=n-D}^{n} \alpha_r \leq D \max_{r\in\{n-D,\ldots,n\}} \alpha_r \to 0 \quad \text{as } n \to \infty, \tag{3.5}$$

where the limit approaching 0 follows from $\lim_{n\to\infty} \alpha_n = 0$, which itself is a consequence of Assumption 3. This implies $\lim_{n\to\infty} C_3 V_{\max} \sum_{r=s(n)}^{n-1} \alpha_{r+1} = 0$ and consequently, $\lim_{n\to\infty} \|b_n\|_2 = 0$.

(b) We consider each of the two conditions in turn. When $\alpha_{n-1} = \frac{1}{n \log n}$ and $d_n = O(n)$, it is easy to verify (by integration) that this particular choice of sequence satisfies $\sum_{n=1}^{\infty} \alpha_n^2 < \infty, \sum_{n=1}^{\infty} \alpha_n = \infty$. Since $d_n = O(n)$, we have $n - s(n) \leq K s(n)$ for some universal constant $K > 0$,

which means $n \leq s(n) + Ks(n)$. Consequently, we have:

$$\begin{aligned} 0 < \sum_{r=s(n)}^{n-1} \alpha_{r+1} &= \sum_{r=s(n)}^{s(n)+Ks(n)} \alpha_r \leq \int_{s(n)}^{s(n)+Ks(n)} \frac{1}{r \log r} dr \\ &= \log \frac{\log(s(n) + Ks(n))}{\log s(n)} = \log \frac{\log(K+1) + \log s(n)}{\log s(n)} \to 0 \quad \text{as } n \to \infty, \end{aligned} \tag{3.6}$$

where the last limit follows from $s(n) \to \infty$ as $n \to \infty$ because $n \leq s(n) + Ks(n)$.

Next, When $\alpha_{n-1} = \frac{1}{n \log n \log\log n}$ and $d_n = O(n^a), a > 1$, it is again easy to verify (by integration) that this particular choice of sequence satisfies Assumption 3. Since $d_n = O(n^a)$, we have $n - s(n) \leq Ks(n)^a$ for some universal constant $K > 0$, which means $n \leq s(n) + Ks(n)^a$. Consequently, we have:

$$\begin{aligned} 0 < \sum_{r=s(n)}^{n-1} \alpha_{r+1} &= \sum_{r=s(n)}^{s(n)+Ks(n)^a} \alpha_r \leq \int_{s(n)}^{s(n)+Ks(n)^a} \frac{1}{r \log r \log\log r} dr \\ &= \log \frac{\log\log(s(n) + Ks(n)^a)}{\log\log s(n)} \leq \log \frac{\log\log(s(n)^a + Ks(n)^a)}{\log\log s(n)} \\ &= \log \frac{\log(\log(K+1) + a\log s(n))}{\log\log s(n)} < \log \frac{\log((K+1)\log s(n) + a\log s(n))}{\log\log s(n)} \\ &= \log \frac{\log(K+1+a) + \log\log s(n)}{\log\log s(n)} \to 0, \text{ as } n \to \infty, \end{aligned} \tag{3.7}$$

where the last limit follows from $s(n) \to \infty$ as $n \to \infty$, again because $n \leq s(n) + Ks(n)^a$.

(2) With the defintion of b_n, DAGD can be written as:

$$\begin{aligned} x_n &= \mathbf{proj}_{\mathcal{X}}(y_n), \\ y_{n+1} &= y_n - \alpha_{n+1}(\nabla f(x_n) + b_n). \end{aligned} \tag{3.8}$$

We then use the energy function to study the behavior of y_n and x_n. More specifically, we look at the quantity $E(\mathcal{X}^*, Y_{n+1}) - E(\mathcal{X}^*, Y_n)$ and bound this one-step change using the step size α_n, the b_n sequence and the defining quantity $\langle \nabla f(x_n), x_n - x^* \rangle$ of a variationally coherent function (as well as another term that will prove inconsequential). We then telescope on $E(\mathcal{X}^*, Y_{n+1}) - E(\mathcal{X}^*, Y_n)$ to obtain an upper bound for $E(\mathcal{X}^*, Y_{n+1}) - E(\mathcal{X}^*, Y_0)$. Since the energy function is always non-negative (Lemma 3.1), $E(\mathcal{X}^*, Y_{n+1}) - E(\mathcal{X}^*, Y_0)$ is at least $-E(\mathcal{X}^*, Y_0)$ for every n. However, utilizing the fact that b_n converges to 0 and that $\langle \nabla f(x_n), x_n - x^* \rangle$ is always positive (unless the iterate is exactly an optimal solution), we show that the upper bound will approach $-\infty$ if X_n only enters $\mathcal{N}(\mathcal{X}^*, \epsilon)$, an open ϵ-neighborhood of $\mathcal{X}^*$, a finite number of times (for an arbitrary $\epsilon > 0$). This generates an immediate contradiction, and thereby establishes that X_n will get arbitrarily close to $\mathcal{X}^*$ for an infinite number of times. This then implies that there exists a subsequence of the DAGD iterates that converges to the set of global optima: $x_{n_k} \to \mathcal{X}^*$, as $k \to \infty$.

To prove the claim, for simplicity and without loss of generality, we assume $\mathcal{X}^* = \{x^*\}$. If $\mathcal{X}^*$ contains more than one point, the analysis is

identical provided we change all the point-to-point distance to point-to-set distance. We now bound the energy change in one step as follows:

$$
\begin{aligned}
&E(x^*, y_{n+1}) - E(x^*, y_n) = \|x^*\|_2^2 - \|\,\mathbf{proj}_{\mathcal{X}}(y_{n+1})\|_2^2 + 2\langle y_{n+1}, \mathbf{proj}_{\mathcal{X}}(y_{n+1}) - x^*\rangle \\
&- \Big\{\|x^*\|_2^2 - \|\,\mathbf{proj}_{\mathcal{X}}(y_n)\|_2^2 + 2\langle y_n, \mathbf{proj}_{\mathcal{X}}(y_n) - x^*\rangle\Big\} \\
&= \|\,\mathbf{proj}_{\mathcal{X}}(y_n)\|_2^2 - \|\,\mathbf{proj}_{\mathcal{X}}(y_{n+1})\|_2^2 - 2\langle y_n, \mathbf{proj}_{\mathcal{X}}(y_n) - x^*\rangle \\
&+ 2\langle y_n - \alpha_{n+1}(\nabla f(x_n) + b_n), \mathbf{proj}_{\mathcal{X}}(y_{n+1}) - x^*\rangle \\
&= \|\,\mathbf{proj}_{\mathcal{X}}(y_n)\|_2^2 - \|\,\mathbf{proj}_{\mathcal{X}}(y_{n+1})\|_2^2 + 2\langle y_n, \mathbf{proj}_{\mathcal{X}}(y_{n+1}) - \mathbf{proj}_{\mathcal{X}}(y_n)\rangle \\
&- 2\langle \alpha_{n+1}(\nabla f(x_n) + b_n), \mathbf{proj}_{\mathcal{X}}(y_n) - x^* + \mathbf{proj}_{\mathcal{X}}(y_{n+1}) - \mathbf{proj}_{\mathcal{X}}(y_n)\rangle \\
&= -2\langle \alpha_{n+1}(\nabla f(x_n) + b_n), \mathbf{proj}_{\mathcal{X}}(y_n) - x^*\rangle + 2\langle y_{n+1}, \mathbf{proj}_{\mathcal{X}}(y_{n+1}) - \mathbf{proj}_{\mathcal{X}}(y_n)\rangle \\
&+ \|\,\mathbf{proj}_{\mathcal{X}}(y_n)\|_2^2 - \|\,\mathbf{proj}_{\mathcal{X}}(y_{n+1})\|_2^2 \\
&= -2\langle \alpha_{n+1}(\nabla f(x_n) + b_n), x_n - x^*\rangle + \|\,\mathbf{proj}_{\mathcal{X}}(y_n) - y_{n+1}\|_2^2 - \|\,\mathbf{proj}_{\mathcal{X}}(y_{n+1}) - y_{n+1}\|_2^2 \\
&\leq -2\langle \alpha_{n+1}(\nabla f(x_n) + b_n), x_n - x^*\rangle + \|y_n - y_{n+1}\|_2^2
\end{aligned}
\tag{3.9}
$$

where the last equality follows from completing the squares and the last inequality follows from projection onto convex sets.

Now telescoping yields:

$$
\begin{aligned}
&E(x^*, y_{n+1}) - E(x^*, y_0) = \sum_{r=0}^{n} \{E(x^*, y_{r+1}) - E(x^*, y_r)\} \\
&\leq \sum_{r=0}^{n} \{-2\alpha_{r+1}\langle \nabla f(x_r) + b_r, x_r - x^*\rangle + \alpha_{r+1}^2 \|\nabla f(x_r) + b_r\|_2^2\} \\
&\leq -2\sum_{r=0}^{n} \alpha_{r+1}\{\langle \nabla f(x_r), x_r - x^*\rangle - \|b_r\|_2 \|x_r - x^*\|_2\} + 2\sum_{r=0}^{n} \alpha_{r+1}^2 \{\|\nabla f(x_r)\|_2^2 + \|b_r\|_2^2\} \\
&\leq -2\sum_{r=0}^{n} \alpha_{r+1}\{\langle \nabla f(x_r), x_r - x^*\rangle - C_4\|b_r\|_2\} + 2\sum_{r=0}^{n} \alpha_{r+1}^2 \{C_2 + B\},
\end{aligned}
\tag{3.10}
$$

where the last inequality follows from the fact that b_n's must be bounded (since $\lim_{n\to\infty} \|b_n\|_2 = 0$) and hence let $B \triangleq \sup_n \|b_n\|_2$. By Assumption 3, we have $2\sum_{r=0}^{n} \alpha_{r+1}^2\{C_2 + B\} = \overline{B} < \infty$. Now fix any positive number ϵ. Assume for contradiction purposes x_n only enters $\mathcal{N}(x^*, \epsilon)$ a finite number of times and let t_1 be the last time this occurs. This means that for all $r > t_1$, x_r is outside the open set $\mathcal{N}(x^*, \epsilon)$. Therefore, since a continuous function always achieves its minimum on a compact set, we have: $\langle \nabla f(x_r), x_r - x^*\rangle \geq \min_{x \in \mathcal{X} - \mathcal{N}(x^*, \epsilon)} \langle \nabla f(x), x - x^*\rangle \triangleq a > 0, \forall r > t_1$ (note that a depends on ϵ). Further, since $b_r \to 0$ as $r \to \infty$, pick t_2 such that $\|b_r\|_2 < \frac{a}{2C_4}, \forall r \geq t_2$. Denoting $t = \max(t_1, t_2)$, we continue the chain

of inequalities in Equation (3.10) below:

$$
\begin{aligned}
&- E(x^*, y_0) \leq E(x^*, y_{n+1}) - E(x^*, y_0) \leq -2\sum_{r=0}^{t} \alpha_{r+1}\{\langle \nabla f(x_r), x_r - x^* \rangle - C_4\|b_r\|_2\} \\
&- 2\sum_{r=t+1}^{n} \alpha_{r+1}\{\langle \nabla f(x_r), x_r - x^* \rangle - C_4\|b_r\|_2\} + 2\sum_{r=0}^{n} \alpha_{r+1}^2\{C_2 + B\} \\
&\leq -2\sum_{r=0}^{t} \alpha_{r+1}\{\langle \nabla f(x_r), x_r - x^* \rangle - C_4\|b_r\|_2\} - 2\sum_{r=t+1}^{n} \alpha_{r+1}\{a - C_4\|b_r\|_2\} + \overline{B} \\
&\leq 2C_4\sum_{r=0}^{t} \alpha_{r+1}|b_r\|_2 + \overline{B} - 2\sum_{r=t+1}^{n} \alpha_{r+1}\{a - \frac{a}{2}\} \\
&= \overline{\overline{B}} - a\sum_{r=t+1}^{n} \alpha_{r+1} \to -\infty, \text{ as } n \to \infty
\end{aligned}
\tag{3.11}
$$

where the first inequality follows from the energy function always being positive (Lemma 3.1), the second-to-last inequality follows from variational coherence and the limit on the last line follows from Assumption 3 and that $\overline{\overline{B}} \triangleq 2C_4 \sum_{r=0}^{t} \alpha_{r+1}|b_r\|_2 + \overline{B}$ is just some finite constant. This yields an immediate contradiction and the claim is therefore established.

■

We are now in a position to prove our main convergence result for DAGD:

Theorem 3.3. *Under Assumptions 1–3, the global state variable x_n of DAGD converges to $\mathcal{X}^*$.*

Proof. Fix a given $\delta > 0$. Since $\alpha_n \to 0, b_n \to 0$ as $n \to \infty$, for any $a > 0$, we can pick an N large enough (depending on δ and a) such that $\forall n \geq N$, the following three statements all hold true:

$$
\begin{aligned}
&2BC_4\alpha_{n+1} + 2\alpha_{n+1}^2(C_2 + B^2) \leq \frac{\delta}{2}, \\
&C_4\|b_n\|_2 \leq \frac{a}{2}, \\
&\alpha_{n+1}(C_2 + B^2) < \frac{a}{2}.
\end{aligned}
\tag{3.12}
$$

We show that under either of the (exhaustive) following possibilities, if $E(x^*, y_n)$ is less than δ, $E(x^*, y_{n+1})$ is less than δ as well, where $n \geq N$.

(1) Case 1: $E(x^*, y_n) < \frac{\delta}{2}$.
(2) Case 2: $\frac{\delta}{2} \leq E(x^*, y_n) < \delta$.

Under Case 1, it follows from Equation (3.9):

$$\begin{aligned}
&E(x^*, y_{n+1}) - E(x^*, y_n) \leq -2\alpha_{n+1}\langle \nabla f(x_n) + b_n, x_n - x^* \rangle + \alpha_{n+1}^2 \| \nabla f(x_n) + b_n \|_2^2 \\
&\leq -2\alpha_{n+1}\langle b_n, x_n - x^* \rangle + \alpha_{n+1}^2 \| \nabla f(x_n) + b_n \|_2^2 \\
&\leq 2\alpha_{n+1}\|b_n\|_2 \|x_n - x^*\|_2 + 2\alpha_{n+1}^2 (C_2 + B^2) \\
&\leq 2BC_4\alpha_{n+1} + 2\alpha_{n+1}^2 (C_2 + B^2) \\
&\leq \frac{\delta}{2},
\end{aligned} \tag{3.13}$$

where the second inequality follows from variational coherence. This then implies that $E(x^*, y_{n+1}) \leq E(x^*, y_n) + \frac{\delta}{2} < \delta$.

Under Case 2, Eq. (3.9) readily yields:

$$\begin{aligned}
&E(x^*, y_{n+1}) - E(x^*, y_n) \leq -2\alpha_{n+1}\langle \nabla f(x_n) + b_n, x_n - x^* \rangle + \alpha_{n+1}^2 \| \nabla f(x_n) + b_n \|_2^2 \\
&= -2\alpha_{n+1}\langle \nabla f(x_n), x_n - x^* \rangle - 2\alpha_{n+1}\langle b_n, x_n - x^* \rangle + \alpha_{n+1}^2 \| \nabla f(x_n) + b_n \|_2^2 \\
&\leq -2\alpha_{n+1} a + 2\alpha_{n+1}\|b_n\|_2 \|x_n - x^*\|_2 + 2\alpha_{n+1}^2 (C_2 + B^2) \\
&\leq -2\alpha_{n+1}\Big\{ a - C_4\|b_n\|_2 - \alpha_{n+1}(C_2 + B^2) \Big\} \\
&\leq -2\alpha_{n+1}\Big\{ a - \frac{a}{2} - \alpha_{n+1}(C_2 + B^2) \Big\} \\
&= -2\alpha_{n+1}\Big\{ \frac{a}{2} - \alpha_{n+1}(C_2 + B^2) \Big\} \\
&< 0,
\end{aligned} \tag{3.14}$$

where the second inequality follows from $\langle \nabla f(x_n), x_n - x^* \rangle \geq a$ under Case 2[3]. Consequently, $E(x^*, y_{n+1}) < E(x^*, y_n) < \delta$. ∎

4. Stochastic Convergence Analysis

4.1. Recurrence of DASGD.

Proposition 4.1. *Under Assumptions 1–3, DAGD admits a subsequence X_{n_k} that converges to $\mathcal{X}^*$ almost surely; concretely, $X_{n_k} \to \mathcal{X}^*$ with probability 1 as $k \to \infty$.*

Proof. For streamlining purposes, we break up the proof in two distinct steps below.

(1) We rewrite the gradient update in distributed asynchronous stochastic gradient descent (DASGD) as:

$$\begin{aligned}
Y_{n+1} &= Y_n - \alpha_{n+1}\nabla F(X_{s(n)}, \omega_{s(n)+1}) \\
&= Y_n - \alpha_{n+1}\Big\{ \nabla f(X_n) + \nabla f(X_{s(n)}) - \nabla f(X_n) \\
&\quad + \nabla F(X_{s(n)}, \omega_{s(n)+1}) - \nabla f(X_{s(n)}) \Big\}.
\end{aligned} \tag{4.1}$$

By defining $B_n = \nabla f(X_{s(n)}) - \nabla f(X_n)$ and $U_{n+1} = \nabla F(X_{s(n)}, \omega_{s(n)+1}) - \nabla f(X_{s(n)})$, we can rewrite DASGD as:

$$Y_{n+1} = Y_n - \alpha_{n+1}\{\nabla f(X_n) + B_n + U_{n+1}\}. \tag{4.2}$$

[3]Here a is a constant that only depends on δ: if $E(x^*, y) \geq \frac{\delta}{2}$, then by part 2 of Lemma 3.1, $\mathbf{proj}_{\mathcal{X}}(y)$ must be outside an ϵ-neighborhood of x^*, for some $\epsilon > 0$. On this neighborhood, the strictly positive continuous function $\langle \nabla f(x), x - x^* \rangle$ must achieve a minimum value $a > 0$.

We then establish the following two facts in this step. First, we verify that $\sum_{r=0}^{n} U_{n+1}$ is a martingale adapted to $Y_1, Y_2 \ldots, Y_{n+1}$. Second, we show that $\lim_{n\to\infty} \|B_n\|_2 = 0, a.s.$.

The second claim is done by first giving an upper bound on $\|B_n\|_2$ by writing $\nabla f(X_{s(n)}) - \nabla f(X_n)$ as a sum of one-step changes $(\nabla f(X_{s(n)}) - \nabla f(X_{s(n)+1}) + \nabla f(X_{s(n)+1}) - \cdots + \nabla f(X_{n-1}) - \nabla f(X_n))$ and analyzing each such successive change. We then break that upper bound into two parts, one deterministic and one stochastic. For the deterministic part, the same analysis in the proof to Proposition 3.2 yields convergence to 0. The stochastic part turns out to be the tail of a martingale. By leveraging the property of the step-size and a crucial property of martingale differences (two martingale differences at different time steps are uncorrelated), we establish that the martingale is L_2-bounded. Therefore we can apply some variant of Doob's martingale convergence theorem to establish that the martingale converges almost surely to a limiting random variable that has finite second moment (and hence almost surely finite). Consequently, writing the tail as a difference between two terms (each of which converges to the same limiting random almost surely), we know the tail converges to 0 almost surely.

To see that $\sum_{r=0}^{n} U_{n+1}$ is a martingale adapted to $Y_0, Y_1 \ldots, Y_{n+1}$, first note that, by defintion, B_n is adapted to $Y_0, Y_1 \ldots, Y_n$ (since X_n is a deterministic function of Y_n) and Y_{n+1}, Y_n, B_n together determine U_{n+1}. We then check that their first moments are bounded:

$$
\begin{aligned}
&\mathbb{E}[\| \sum_{r=0}^{n} \|U_{r+1}\|_2] \leq \sum_{r=0}^{n} \mathbb{E}[\|U_{r+1}\|_2] = \sum_{r=0}^{n} \mathbb{E}[\|\nabla F(X_{s(r)}, \omega_{s(r)+1}) - \nabla f(X_{s(r)})\|_2] \\
&\leq \sum_{r=0}^{n} \Big\{ \mathbb{E}[\|\nabla F(X_{s(r)}, \omega_{s(r)+1})\|_2] + \mathbb{E}[\| \nabla f(X_{s(r)})\|_2] \Big\} \\
&\leq \sum_{r=0}^{n} \Big\{ \sup_{x\in\mathcal{X}} \mathbb{E}[\|\nabla F(x, \omega)\|_2] + \sup_{x\in\mathcal{X}} \| \nabla f(x)\|_2 \Big\} \\
&= \sum_{r=0}^{n} \Big\{ \sup_{x\in\mathcal{X}} \mathbb{E}[\|\nabla F(x, \omega)\|_2] + \sup_{x\in\mathcal{X}} \| \mathbb{E}[\nabla F(x, \omega)]\|_2 \Big\} \\
&\leq \sum_{r=0}^{n} 2 \sup_{x\in\mathcal{X}} \mathbb{E}[\|\nabla F(x, \omega)\|_2] = \sum_{r=0}^{n} 2C_1 = 2(n+1)C_1 < \infty,
\end{aligned}
\tag{4.3}
$$

where the last inequality follows from Jensen's inequality (since $\|\cdot\|_2$ is a convex function). Finally, the martingale property holds because: $\mathbb{E}[\sum_{r=0}^{n} U_{r+1} \mid Y_1, \ldots, Y_{n+1}] = \mathbb{E}[\nabla F(X_{s(n)}, \omega_{s(n)+1}) - \nabla f(X_{s(n)}) \mid Y_1, \ldots, Y_n] + \sum_{r=0}^{n-1} U_{r+1} = \sum_{r=0}^{n-1} U_{r+1}$.

Therefore, $\sum_{r=0}^{n} U_{r+1}$ is a martingale, and U_{n+1} is a martingale difference sequence adapted to $Y_0, Y_1 \ldots, Y_{n+1}$.

Next, we show that $\lim_{n\to\infty} \|B_n\|_2 = 0, a.s.$. By definition, we can expand B_n as follows:

$$\begin{aligned}
\|B_n\|_2 &= \| \nabla f(X_{s(n)}) - \nabla f(X_n)\|_2 \leq C_3 \|X_{s(n)} - X_n\|_2 = C_3 \| \mathbf{proj}_{\mathcal{X}}(Y_{s(n)}) - \mathbf{proj}_{\mathcal{X}}(Y_n)\|_2 \\
&\leq C_3 \|Y_{s(n)} - Y_n\|_2 = C_3 \Big\| Y_{s(n)} - Y_{s(n)+1} + Y_{s(n)+1} - Y_{s(n)+2} + \cdots + Y_{n-1} - Y_n \Big\|_2 \\
&= C_3 \Big\| \sum_{r=s(n)}^{n-1} \alpha_{r+1} \Big\{ Y_r - Y_{r+1} \Big\} \Big\|_2 = C_3 \Big\| \sum_{r=s(n)}^{n-1} \alpha_{r+1} \nabla F(X_{s(r)}, \omega_{s(r)+1}) \Big\|_2 \\
&= C_3 \Big\| \sum_{r=s(n)}^{n-1} \alpha_{r+1} \Big\{ \nabla f(X_{s(r)}) + \nabla F(X_{s(r)}, \omega_{s(r)+1}) - \nabla f(X_{s(r)}) \Big\} \Big\|_2 \\
&= C_3 \Big\| \sum_{r=s(n)}^{n-1} \alpha_{r+1} \nabla f(X_{s(r)}) + \sum_{r=s(n)}^{n-1} \alpha_{r+1} U_{r+1} \Big\|_2 \\
&\leq C_3 \sum_{r=s(n)}^{n-1} \alpha_{r+1} \| \nabla f(X_{s(r)})\|_2 + C_3 \Big\| \sum_{r=s(n)}^{n-1} \alpha_{r+1} U_{r+1} \Big\|_2 \\
&\leq C_3 C_1 \sum_{r=s(n)}^{n-1} \alpha_{r+1} + C_3 \| \sum_{r=s(n)}^{n-1} \alpha_{r+1} U_{r+1}\|_2 \\
&= C_3 C_1 \sum_{r=s(n)}^{n-1} \alpha_{r+1} + C_3 \| \sum_{r=0}^{n-1} \alpha_{r+1} U_{r+1} - \sum_{r=0}^{s(n)-1} \alpha_{r+1} U_{r+1}\|_2, \qquad (4.4)
\end{aligned}$$

where the first inequality follows from ∇f being Liptichz-continuous (Assumption 3) and the second inequality follows from $\mathbf{proj}_{\mathcal{X}}$ is a non-expansive map.

By the same analysis as in the deterministic case, the first part of the last line of Equation (4.4) converges to 0 (under each one of the three conditions on step-size and delays):

$$\lim_{n\to\infty} C_3 C_1 \sum_{r=s(n)}^{n-1} \alpha_{r+1} = 0. \qquad (4.5)$$

We then analyze the limit of $\| \sum_{r=0}^{n-1} \alpha_{r+1} U_{r+1} - \sum_{r=0}^{s(n)-1} \alpha_{r+1} U_{r+1}\|_2$. Define:

$$M_n = \sum_{r=0}^{n-1} \alpha_{r+1} U_{r+1}.$$

Since U_{r+1}'s are martingale differences, M_n is a martingale. Further, in each of the three conditions, $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$. This implies that M_n is an

L_2-bounded martingale because:

$$\sup_n \mathbb{E}[\|M_n\|_2^2] = \sup_n \mathbb{E}[\Big\| \sum_{r=0}^{n-1} \alpha_{r+1} U_{r+1} \Big\|_2^2] = \sup_n \mathbb{E}[\langle \sum_{r=0}^{n-1} \alpha_{r+1} U_{r+1}, \sum_{r=0}^{n-1} \alpha_{r+1} U_{r+1} \rangle]$$
$$= \sup_n \mathbb{E}[\sum_{i,j} \langle \alpha_{i+1} U_{i+1}, \alpha_{j+1} U_{j+1} \rangle] = \sup_n \sum_{r=0}^{n-1} \mathbb{E}[\langle \alpha_{r+1} U_{r+1}, \alpha_{r+1} U_{r+1} \rangle]$$
$$= \sup_n \sum_{r=0}^{n-1} \alpha_{r+1}^2 \, \mathbb{E}[\|U_{r+1}\|_2^2] \leq \sup_n 4C_2 \sum_{r=0}^{n-1} \alpha_{r+1}^2 \leq 4C_2 \sum_{r=0}^{\infty} \alpha_{r+1}^2 < \infty, \tag{4.6}$$

where the last inequality in the second line follows from the martingale property as follows:

$$\mathbb{E}[\langle \alpha_{i+1} U_{i+1}, \alpha_{j+1} U_{j+1} \rangle] = \alpha_{i+1}\alpha_{j+1} \, \mathbb{E}[\langle U_{i+1}, U_{j+1} \rangle]$$
$$= \alpha_{i+1}\alpha_{j+1} \, \mathbb{E}[\mathbb{E}[\langle U_{i+1}, U_{j+1} \rangle \mid Y_0, Y_1, \ldots, Y_{i+1}]]$$
$$= \alpha_{i+1}\alpha_{j+1} \, \mathbb{E}[\langle U_{i+1}, \mathbb{E}[U_{j+1} \mid Y_0, Y_1, \ldots, Y_{i+1}] \rangle] = \alpha_{i+1}\alpha_{j+1} \, \mathbb{E}[\langle U_{i+1}, 0 \rangle] = 0, \tag{4.7}$$

where the second equality follows from the tower property (and without loss of generality, we have assumed $i < j$, the third equality follows from U_{i+1} is adapted to $Y_0, Y_1, \ldots, Y_{i+1}$ and the second-to-last equality follows from U_{n+1} is a martingale difference. Consequently, all the cross terms in the second line of Equation (4.6) are 0. Therefore, by Lemma 1.2, by taking $p = 2$ $\lim_{n\to\infty} M_n = M_\infty$, a.s., where M_∞ has finite second-moment. Further, since in all three cases $s(n) \to \infty$ as $n \to \infty$ (because there is at most a polynomial lag between $s(n)$ and n), we have $\lim_{n\to\infty} M_{s(n)} = M_\infty$, a.s.. Therefore

$$\lim_{n\to\infty} \Big\{ \sum_{r=0}^{n-1} \alpha_{r+1} U_{r+1} - \sum_{r=0}^{s(n)-1} \alpha_{r+1} U_{r+1} \Big\} = \lim_{n\to\infty} \Big\{ M_n - M_{s(n)} \Big\} = 0, \text{ a.s.},$$

thereby implying:

$$\lim_{n\to\infty} C_3 \Big\| \sum_{r=0}^{n-1} \alpha_{r+1} U_{r+1} - \sum_{r=0}^{s(n)-1} \alpha_{r+1} U_{r+1} \Big\|_2 = 0. \tag{4.8}$$

Combining Equation (4.5) and Equation (4.8) yields $\lim_{n\to\infty} \|B_n\|_2 = 0, a.s.$.

(2) The full DASGD update is then:

$$X_n = \mathbf{proj}_{\mathcal{X}}(Y_n) \tag{4.9}$$
$$Y_{n+1} = Y_n - \alpha_{n+1}\{\nabla f(X_n) + B_n + U_{n+1}\}. \tag{4.10}$$

Similar to the deterministic case before, we again bound the one-step change of the energy function $E(\mathcal{X}^*, Y_{n+1}) - E(\mathcal{X}^*, Y_n)$ and then telescope the differences. The two distinctions from the determinstic case are: 1) Everything is now a random variable. 2) We have three terms: in addition to B_n, we also have a martingale term U_{n+1}. Since B_n converges to 0 almost surely (as shown in the previous step), its effect is the same as b_n in the deterministic case. Futher, the analysis utilizes law of large numbers for martingale as well as Doob's martingale convergence theorem to bound the effect of the

various martingale terms and to establish that the final dominating term is still the same term as in the deterministic case: a term that converges to $-\infty$ (which generates a contradiction since the energy function is always positive) unless a subsequence X_{n_k} converges almost surely to $\mathcal{X}^*$.

Similar to Equation (3.9), we now have:

$$\begin{aligned}
&E(x^*, Y_{n+1}) - E(x^*, Y_n) \leq -2\alpha_{n+1}\langle \nabla f(X_n) + B_n + U_{n+1}, X_n - x^* \rangle + \|Y_n - Y_{n+1}\|_2^2 \\
&= -2\alpha_{n+1}\langle \nabla f(X_n) + B_n + U_{n+1}, X_n - x^* \rangle + \alpha_{n+1}^2 \| \nabla f(X_n) + B_n + U_{n+1}\|_2^2 \\
&\leq -2\alpha_{n+1}\langle \nabla f(X_n) + B_n + U_{n+1}, X_n - x^* \rangle + 3\alpha_{n+1}^2 \Big\{ \| \nabla f(X_n)\|_2^2 + \|B_n\|_2^2 + \|U_{n+1}\|_2^2 \Big\} \\
&\leq -2\alpha_{n+1}\langle \nabla f(X_n) + B_n + U_{n+1}, X_n - x^* \rangle + 3\alpha_{n+1}^2 \Big\{ C_2 + \|B_n\|_2^2 + \|U_{n+1}\|_2^2 \Big\}.
\end{aligned} \tag{4.11}$$

For contradiction purposes assume X_n enters $\mathcal{N}(x^*, \epsilon)$ only a finite number of times with positive probability. By starting the sequence at a later index if necessary, we can without loss of generality X_n never enters $\mathcal{N}(x^*, \epsilon)$ with positive probability. Then on this event (of X_n never enters $\mathcal{N}(x^*, \epsilon)$), we have $\langle \nabla f(X_n), X_n - x^* \rangle \geq a > 0$ as before. Telescoping Equation (4.11) then yields:

$$\begin{aligned}
&-\infty < -E(x^*, Y_0) \leq E(x^*, Y_{n+1}) - E(x^*, Y_0) = \sum_{r=0}^{n} \{E(x^*, Y_{r+1}) - E(x^*, Y_r)\} \\
&\leq -2\sum_{r=0}^{n} \alpha_{n+1}\langle \nabla f(X_n) + B_n + U_{n+1}, X_n - x^* \rangle + 3\sum_{r=0}^{n} \alpha_{n+1}^2 \Big\{ C_2 + \|B_n\|_2^2 + \|U_{n+1}\|_2^2 \Big\} \\
&\leq -2\sum_{r=0}^{n} \alpha_{n+1} \Big\{ a + \langle B_n + U_{n+1}, X_n - x^* \rangle \Big\} + 3\sum_{r=0}^{n} \alpha_{n+1}^2 \Big\{ C_2 + \|B_n\|_2^2 + \|U_{n+1}\|_2^2 \Big\} \\
&\to -\infty \text{ a.s. } \text{ as } n \to \infty.
\end{aligned} \tag{4.12}$$

We justify the last-line limit of Equation (4.12) by looking at each of its components in turn:

(a) Since $\sum_{r=0}^{n} \alpha_{n+1}^2 < \infty$, and $\lim_{n\to\infty} \|B_n\|_2^2 = 0$, a.s., we have $3\sum_{r=0}^{\infty} \alpha_{n+1}^2 \Big\{ C_2 + \|B_n\|_2^2 \Big\} = C$, a.s., for some constant $C < \infty$.

(b) $\sum_{r=0}^{n} \alpha_{n+1}^2 \|U_{n+1}\|_2^2$ is submartingale that is L_1 bounded since:

$$\begin{aligned}
&\sup_n \mathbb{E}[\sum_{r=0}^{n} \alpha_{n+1}^2 \|U_{n+1}\|_2^2] \leq \sup_n \sum_{r=0}^{n} \alpha_{n+1}^2 \, \mathbb{E}[\|U_{n+1}\|_2^2] \leq \sup_n \sum_{r=0}^{n} \alpha_{n+1}^2 \, \mathbb{E}[\|U_{n+1}\|_2^2] \\
&\leq 2 \sup_n \sum_{r=0}^{n} \alpha_{n+1}^2 \Big\{ \mathbb{E}[\| \nabla f(X_n)\|_2^2] + \mathbb{E}[\| \nabla F(X_n, \omega_{n+1})\|_2^2] \Big\} \\
&\leq 2 \sup_n \sum_{r=0}^{n} 2C_2 \alpha_{n+1}^2 < \infty.
\end{aligned} \tag{4.13}$$

Consequently, by martingale convergence theorem (Lemma 1.2 by taking $p = 1$), $3\sum_{r=0}^{n} \alpha_{n+1}^2 \|U_{n+1}\|_2^2 \to R$, a.s., for some random variable R that is almost surely finite (in fact $\mathbb{E}[|R|] < \infty$).

(c) Since $\|B_n\|_2$ converges to 0 almost surely, its average also converges to 0 almost surely:

$$\sum_{n=0}^{\infty} \frac{\alpha_{n+1}\|B_n\|_2}{\sum_{r=1}^{n}\alpha_{r+1}} = 0, \text{ a.s.},$$

there by implying that

$$\sum_{n=0}^{\infty} \frac{\alpha_{n+1}\langle B_n, X_n - x^*\rangle}{\sum_{r=1}^{n}\alpha_{r+1}} = 0, \text{ a.s.},$$

since $|\langle B_n, X_n - x^*\rangle| \leq \|B_n\|_2\|X_n - x^*\|_2 \leq C_4\|B_n\|_2$.
In addition, $\langle U_{n+1}, X_n - x^*\rangle$ is a martingale difference that is L_1 bounded (because $\mathbb{E}[\|\langle U_{n+1}, X_n - x^*\rangle\|_2^2] \leq \mathbb{E}[\|U_{n+1}\|_2^2\|X_n - x^*\|_2^2] \leq C_4\,\mathbb{E}[\|U_{n+1}\|_2^2] \leq 4C_4C_2 < \infty$), law of large number therefore implies: $\sum_{n=0}^{\infty}\frac{\alpha_{n+1}\langle U_{n+1}, X_n - x^*\rangle}{\sum_{r=1}^{n}\alpha_{r+1}} = 0$, a.s. Combining the above two limits, we have

$$\lim_{n\to\infty}\frac{\sum_{r=0}^{n}\alpha_{n+1}\langle B_n + U_{n+1}, X_n - x^*\rangle}{\sum_{r=0}^{n}\alpha_{r+1}} = 0, \text{ a.s.}$$

Consequently, $-\sum_{r=0}^{n}\alpha_{n+1}\Big\{a+\langle B_n+U_{n+1}, X_n-x^*\rangle\Big\} = -\{\sum_{r=0}^{n}\alpha_{n+1}\}\Big\{a+\frac{\sum_{r=0}^{n}\alpha_{n+1}\langle B_n+U_{n+1}, X_n-x^*\rangle}{\sum_{r=0}^{n}\alpha_{r+1}}\Big\} \to -\infty$, as $n \to -\infty$.

■

4.2. **ODE Approximation of DASGD.**

Lemma 4.2. *Fix any $\delta > 0$.*

(1) If $\mathbf{proj}_{\mathcal{X}}(P(t,y)) \notin \mathcal{X}^*$, then $E(\mathcal{X}^*, P(t,y))$ is a strictly decreasing function in t for every $y \in \mathbb{R}^d$.

(2) There exists some time constant $T(\delta) > 0$ such that $\sup_{y\in\mathbb{R}^d: E(x^*,y)>\frac{\delta}{2}} E(x^*, P(t,y)) - E(x^*, y) < \frac{\delta}{2}$, $\forall t > T(\delta)$.

Proof. The first claim follows by computing the derivative of the energy function with respect to time (for notational simplicity, here we just use $y(t)$ to denote $P(t,y)$):

$$\begin{aligned}
&\frac{d}{dt}E(x^*, y(t)) = \frac{d}{dt}\Big\{\|x^*\|_2^2 - \|\,\mathbf{proj}_{\mathcal{X}}(y(t))\|_2^2 + 2\langle y(t), \mathbf{proj}_{\mathcal{X}}(y(t)) - x^*\rangle\Big\}\\
&= \frac{d}{dt}\Big\{ - \|\,\mathbf{proj}_{\mathcal{X}}(y(t)) - y(t)\|_2^2 + \|y(t)\|_2^2 + 2\langle y(t), -x^*\rangle\Big\}\\
&= 2\langle \dot{y}(t), \mathbf{proj}_{\mathcal{X}}(y(t)) - y(t)\rangle + 2\langle y(t), \dot{y}(t)\rangle + 2\langle \dot{y}(t), -x^*\rangle\Big\}\\
&= -2\langle \nabla f(x(t)), x(t) - y(t)\rangle - 2\langle \nabla f(x(t)), y(t)\rangle - 2\langle \nabla f(x(t)), -x^*\rangle\Big\}\\
&= -\langle \nabla f(x(t)), x(t) - x^*\rangle \leq 0,
\end{aligned} \tag{4.14}$$

where the last inequality is strict unless $\mathbf{proj}_{\mathcal{X}}(y(t)) = x(t) = x^*$ (or $x(t) \in \mathcal{X}^*$ in the set case). Note that even though $\mathbf{proj}_{\mathcal{X}}(y(t)) - y(t)$ is not differentiable, $\|\,\mathbf{proj}_{\mathcal{X}}(y(t)) - y(t)\|_2^2$ is; and in computing its derivative, we applied the envelope theorem.

For the second claim, consider any y that satisfies $E(x^*, P(t,y)) > \frac{\delta}{2}$. Then by the monotonicity property in the first part of the lemma, it follows that $E(x^*, P(s,y)) >$

$\frac{\delta}{2}, \forall 0 \leq s \leq t$. This means that there exists some positive constant $a(\delta)$ such that $\forall 0 \leq s \leq t$:

$$\frac{d}{ds}E(x^*, P(s,y)) = -\langle \nabla f(x(s)), x(s) - x^* \rangle \leq -a(\delta). \tag{4.15}$$

Consequently, pick $T(\delta) = \frac{\delta}{2a(\delta)}$, Equation (4.15) implies that for any $t > T(\delta)$:

$$E(x^*, P(t,y)) \leq E(x^*, P(T(\delta), y)) \leq E(x^*, y) - T(\delta)a(\delta) \leq E(x^*, y) - \frac{\delta}{2}. \tag{4.16}$$

Since Equation 4.16 is true for any y, taking sup over y establishes the claim. ■

4.3. **Main Convergence Result.** For convenience, we restate here the convergence result we wish to prove:

Theorem 4.3. *Under Assumptions 1–3, the global state variable X_n of DASGD converges to $\mathcal{X}^*$ with probability* 1.

Proof. We again give the outline of the main strategy of the proof. By Proposition 4.1, Y_n will get arbitrarily close to $\mathcal{X}^*$ infinitely often. It then suffices to show that, after long enough iterations, if Y_n ever gets ϵ-close to $\mathcal{X}^*$, all the ensuing iterates will be ϵ-close to $\mathcal{X}^*$ almost surely.

The way we show this "trapping" property is to use the energy function. Specifically, we consider $E(x^*, A(t))$ and show that no matter how small ϵ is, for all sufficiently large t, if $E(x^*, A(t_0))$ is less than ϵ for some t_0, then $E(x^*, A(t)) < \epsilon, \forall t > t_0$. This would then complete the proof because $A(t)$ actually contains all the DASGD iterates, and hence if $E(x^*, A(t)) < \epsilon, \forall t > t_0$, then $E(x^*, Y_n) < \epsilon$ for all sufficiently large n. Furthermore, since $A(t)$ contains all the iterates, the hypothesis that " if $E(x^*, A(t_0))$ is less than ϵ for some t_0" will be satisfied due to Proposition 4.1.

We expand on one more layer of detail and defer the rest into appendix. To obtain control $E(x^*, A(t))$, we control two things: the energy on the ODE path $E(x^*, P(t,y))$ and the discrepancy between $E(x^*, P(t,y))$ and $E(x^*, A(t))$. The former can be made arbitrarily small as a result of Lemma 4.2 (we have a direct handle on how the ODE path would behave). The latter can also be made arbitrarily small since $A(t)$ is an asymptotic pseudotrajectory for P, the two paths are close. Therefore, the discrepancy between $E(x^*, P)$ and $E(x^*, A)$ should also be vanishingly small. Consequently, since $E(x^*, A(t)) = E(x^*, P(t,y)) + \{E(x^*, A(t)) - E(x^*, P(t,y))\}$, and both terms on the right can be made arbitrarily small, so can $E(x^*, A(t))$ be made arbitrarily small.

We now flesh out more details of the proof. Fix any $\epsilon > 0$. Since $A(t)$ is an asymptotic pseudotrajectory for P, we have:

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} \|Y(t+h) - P(h, Y(t))\|_2 = 0. \tag{4.17}$$

Consequently, for any $\delta > 0$, there exists some $\tau(\delta, T)$ such that $\|Y(t+h) - P(h, Y(t))\|_2 < \delta$ for all $t \geq \tau$ and all $h \in [0, T]$. We therefore have the following chain of inequalities:

$$E(x^*, A(t+h)) = E(x^*, P(h, A(t)) + A(t+h) - P(h, A(t))) \tag{4.18}$$

$$\leq E(x^*, P(h, A(t))) + \langle A(t+h) - P(h, A(t)), \mathbf{proj}_{\mathcal{X}}(P(h, A(t))) - x^* \rangle + \frac{1}{2}\|A(t+h) - P(h, A(t))\|_2^2$$

$$\leq E(x^*, P(h, A(t))) + C_4\delta + \frac{1}{2}\delta^2 = E(x^*, P(h, A(t))) + \frac{\varepsilon}{2}, \tag{4.19}$$

where in the last step we have choosen δ small enough such that $C_4\delta + \frac{1}{2}\delta^2 = \frac{\varepsilon}{2}$.

Now by Proposition 4.1, there exists some τ_0 such that $E(x^*, A(\tau_0)) < \frac{\varepsilon}{2}$. Our goal is to establish that $E(x^*, A(\tau_0 + h)) < \varepsilon$ for all $h \in [0, \infty)$. To that end, partition the $[0, \infty)$ into disjoint time intervals of the form $[(n-1)T_\varepsilon, nT_\varepsilon)$ for some appropriate T_ε. By Lemma 4.2, we have:

$$E(x^*, P(h, A(\tau_0))) \leq E(x^*, P(0, A(\tau_0))) = E(x^*, A(\tau_0)) < \frac{\varepsilon}{2} \quad \text{for all } h \geq 0. \tag{4.20}$$

Consequently:

$$E(x^*, A(\tau_0 + h)) < E(x^*, P(h, A(\tau_0))) + \frac{\varepsilon}{2} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \tag{4.21}$$

where the last inequality is a consequence of (4.20).

Now, assume inductively that Eq. (4.21) holds for all $h \in [(n-1)T_\varepsilon, nT_\varepsilon)$ for some $n \geq 1$. Then, for all $h \in [(n-1)T_\varepsilon, nT_\varepsilon)$, we have:

$$E(x^*, A(\tau_0 + T_\varepsilon + h)) < E(x^*, P(T_\varepsilon, A(\tau_0 + h))) + \frac{\varepsilon}{2} \leq \max\left\{\frac{\varepsilon}{2}, E(x^*, A(\tau_0 + h)) - \frac{\varepsilon}{2}\right\} + \frac{\varepsilon}{2}$$
$$\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \tag{4.22}$$

Consequently, Eq. (4.21) holds for all $h \in [nT_\varepsilon, (n+1)T_\varepsilon)$. This completes the induction and our proof. ■

References

Bubeck, Sébastien et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Hall, P. and Heyde, C. C. *Martingale Limit Theory and Its Application*. Probability and Mathematical Statistics. Academic Press, New York, 1980.