

分散型フェデレート平均法

Tao Sun、Dongsheng Li、Bao Wang

Bao Wangは、米国ユタ大学Scientific Computing & Imaging Instituteに所属しています。(電子メール: wangbaonj@gmail.com)
Dongsheng LiとBao Wangが共著者です。

概要-Federated Averaging (FedAvg) は、膨大な数のクライアントが存在する分散学習において、通信効率の良いアルゴリズムである。FedAvgでは、クライアントはプライバシー保護のためにデータをローカルに保持し、クライアント間の通信には中央のパラメータサーバを使用する。この中央サーバは各クライアントにパラメータを配布し、クライアントから最新のパラメータを収集する。FedAvgは中央集権的に研究されることがほとんどであり、各通信においてサーバとクライアントの間で大規模な通信が必要となる。さらに、中央のサーバを攻撃すると、システム全体のプライバシーが破られる可能性がある。本論文では、無向グラフで接続されたクライアント上で実装された脱集中型FedAvg with momentum (DFedAvgM) を研究する。DFedAvgMでは、全てのクライアントがモメンタムを用いた確率的勾配降下を行い、隣人とのみ通信を行う。通信コストをさらに削減するために、量子化されたDFedAvgMも考慮する。損失関数がPL特性を満たす場合、収束率は改善される。最後に、DFedAvgMの有効性を数値的に検証する。

索引用語-分散型最適化, 連邦型平均化, モメンタム, 確率的勾配降下法

1 イントロダクション

Federated Learning (FL) はプライバシーを保護する分散型機械学習 (ML) パラダイムである [1]。FLでは、中央のサーバが膨大なクライアント (携帯電話、パッドなど) と接続し、クライアントはサーバとデータを共有することなく、自分のデータを保持する。各通信ラウンドにおいて、クライアントはサーバから現在のグローバルモデルを受け取り、一部のクライアントを選択して、ローカルデータを用いて確率的勾配降下法 (SGD) [2]を複数回繰り返し実行し、グローバルモデルを更新する。そして、中央のサーバはこれらの更新されたパラメータを集約して、更新されたグローバルモデルを得る。上記の学習アルゴリズムは、Federated Average (FedAvg) [1]として知られている。特に、クライアントが同種の場合、FedAvgはローカルSGDと等価である[3]。FedAvgは各通信ラウンドにおいて、複数のローカルSGDの更新とサーバによる1回の集約を行うため、従来のローカルSGDの更新と1回の通信による分散学習と比較して、サーバとクライアント間の通信コストが大幅に削減される。

FLアプリケーションでは、通常、大企業や政府機関がセントラルサーバの役割を担っています。オン

本研究は、助成金 (2018YFB0204300) の下での中国国家重点研究開発プログラムおよび助成金 (61932001および61906200) の下での国家自然科学基金によって一部支援されています。

Tao SunとDongsheng LiはNational University of Defense Technology, Changsha, 410073, Hunan, ChinaのCollege of Computerに所属しています。(電子メール: nudtsuntao@163.com, dsli@nudt.edu.cn)

一方、FLはクライアント数が膨大であるため、サーバとクライアント間の通信コストがボトルネックとなることがある[4]。一方、クライアントから収集した更新モデルはローカルデータの個人情報を内包しているため、ハッカーが中央サーバを攻撃してシステム全体のプライバシーを侵害する可能性があり、プライバシー問題は深刻な課題として残されている。そこで、全てのクライアントが無向グラフで接続された分散型フィード・エラー学習が提案されている[5], [6]。分散型FLは、FLにおけるサーバ-クライアント間の通信を、クライアント-クライアント間の通信に置き換えるものである。

1) 分散型FLではサーバとクライアント間の高価な通信が不要であるが、MLモデル自体が大きくなるとローカルクライアント間の通信が高価になることである。そのため、クライアント間の通信コストを削減できないか、ということが重要である。2) MomentumはSGDの高速化手法としてよく知られている[7]。モメンタムを用いたSGDにより、分散型FLにおけるMLモデルの学習を理論的に収束保証して改善できないか、という問いは自然なものである。

1.1 私たちの貢献

我々は、分散型FedAvg with momentum (DFedAvgM)を提案することにより、上記の質問に肯定的に答えます。クライアント間の通信コストをさらに削減するために、我々はDFedAvgMに量子化も統合する。本論文における我々の貢献は、以下の3つの点で詳しく説明される。

- アルゴリズム的には、全てのクライアントが無向グラフで結ばれているような分散設定にFedAvgを拡張する。我々はDFedAvgMを分散型SGD (DSGD) アルゴリズムから動機付ける。特に、各クライアントでMLモデルを学習するために、運動量付きのSGDを用いる。各クライアント間の通信コストを削減するために、我々はさらにDFedAvgMの量子化バージョンを導入し、各クライアントが量子化されたモデルを送受信する。
- 理論的には、(量子化) DFedAvgMの収束を証明する。理論的な結果から、(量子化) DFedAvgMの収束率はSGDやDSGDの収束率に劣らないことが示された。具体的には、DFedAvgMと量子化DFedAvgMの収束率は、局所学習と全クライアント間を結ぶグラフに依存することを示す。また、非凸の仮定での収束結果の他に、非凸最適化で広く研究されているPolyak-Łojasiewicz (PL) 条件での収束保証も確立している。PL条件のもとでは

PL条件により、(量子化された)DFedAvgMの収束速度が速くなることを立証する。さらに、通信コストの削減を保証する十分条件を提示する。

- 経験的に、我々は大規模な数値実験を行っています。

IIDとNon-IIDの両方で様々なデータセットでディープニューラルネットワーク (DNN) を学習させることについて。その結果

量子化されたDFedAvgMが、MLモデルの学習、通信コストの削減、学習データのメンバーズプライバシー保護に有効であることを示す。

1.2 その他の関連作品

本論文に最も関連する3つの研究ライン、すなわち、連合学習、分散学習、分散連合学習について簡単にレビューする。

フェデレーション学習。FedAvgの多くのバリエーションが理論的な保証を伴って開発されてきた。[8]はFedAvgのローカルクライアントにmomentum法を用いている。[9]は適応型FedAvgを提案し、その中央パラメータサーバーは適応的学習率 η を使用してローカルモデルを集約する。通信量を減らすために、遅延勾配と量子化勾配が使用される。

[10], [11], [12]はFLのためのニュートン型スキームを提案している。ヘテロジニアスデータに対するFedAvgの収束解析は

[13], [14]によって議論されている。FLの進歩や未解決の問題については、2つのサーベイ論文 [15], [16]がある。

分散型学習。分散型アルゴリズムはもともと、複数のセンサーに分散したデータの平均を計算するために開発されたものである[17], [18], [19], [20]。最も単純で効率的な分散アルゴリズムの一つである分散型(サブ)勾配降下法(DGD)は、[21], [22], [23], [24], [25]で研究されている。DGDでは凸の仮定が不要であるため[26]、DGDは非凸最適化に有効である。証明可能収束DSGDは、[27], [28], [4]で提案されている。[27]は確率的分散化アルゴリズムの複雑さの結果を示している。[28]は確率的分散アルゴリズムを双対情報を用いて設計し、理論的な収束性を保証している。

[4]は、DSGDが通信効率においてSGDを上回ることを証明する。非同期DSGDは[29]で解析されている。また、[30], [31]では、運動量を用いたDGDが提案されている。量子化DSGDは[32]で提案されている。

分散型フェデレーション・ラーニング。分散型FLは

エッジデバイスが中央サーバーのプライバシー保護を信頼していない場合に選択される学習パラダイムである [33]。[34]の著者らは、医療アプリケーションのための中央サーバーを持たない新しいFLフレームワークを提案し、この新しい方法は、高度に動的なピアツーピア環境を提供する。また、[6]では、パラメータ空間の信念を導入することで、ベイジックなアプローチをとるノードが接続されたネットワークでMLモデルの学習を行うことを提案している。

1.3 組織・団体

本論文は以下のように構成される：第2節では、問題の数学的定式化といくつかの必要な仮定を提示する。第3節では、DFedAvgMとその量子化アルゴリズムを紹介する。第4節では、提案アルゴリズムの収束について述べる。セクショ

1.4 表記方法

スカラーとベクトルをそれぞれ小文字と太字で表し、行列を大文字で表す。

太字ベクトル $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$ に対して、以下のようにします。

はその ℓ_p ノルム ($p \geq 1$) を $\mathbf{x}_p = (\sum_{i=1}^d |x_i|^p)^{1/p}$ で表す。 $|\mathbf{x}|_1$ および

\mathbf{x} の ℓ_∞ ノルムを $\mathbf{x}_\infty = \max_{i=1}^d |x_i|$ で表し、次のように表す。

\mathbf{x} として ℓ_2 ノルム。行列 \mathbf{A} に \mathbf{A}^T であり、その転置を表す対して

を \mathbf{A} とする。2つの系列 $\mathbf{a}_n, \mathbf{b}_n$ があるとき、次のように書き表す。

$\mathbf{a}_n \leq \mathbf{b}_n$ となるような正の定数 $0 < C < +\infty$ が存在する場合は $\mathbf{a}_n \leq C \mathbf{b}_n$ と書き、 $\mathbf{a}_n \asymp \mathbf{b}_n$ とする。

\mathbf{b}_n と $\mathbf{b}_n \asymp C_2 \mathbf{a}_n$ となるような正の定数 C_1 と C_2 が存在する場合は $\mathbf{a}_n \asymp \mathbf{b}_n$ と記すことにしている。(\mathbf{a}_n) は \mathbf{a}_n の対数係数を隠している。

関数 $f(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$ に対して、その勾配を $\nabla f(\mathbf{x})$ と表し、そのヘシアンを $\nabla^2 f(\mathbf{x})$ と表し、その最小値を $\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ と表します。に関する期待値を $\mathbb{E}[\cdot]$ で表す。

を基礎とする確率空間に変換する。

2 問題設定と仮定

次のような最適化タスクを考える。

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x}), f(\mathbf{x}) = \mathbb{E}_{\xi \sim D_i} f_i(\mathbf{x}; \xi), \quad (1)$$

ここで、 i 番目のクライアントにおけるデータ分布を表し $F_i(\mathbf{x}; \xi)$ は学習データに関連する損失関数

ξ 。問題(1)はMLにおける多くのアプリケーションをモデル化しており、経験的リスク最小化(ERM)として知られている。その後の解析のためにいくつかの仮定を列挙する。

前提1. 関数 f_i は微分可能であり、 ∇f_i は L -リプシッツ連続、 $\forall i \in \{1, 2, \dots, m\}$ 、すなわち、 $\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y}) \leq L \|\mathbf{x} - \mathbf{y}\|$ 、すべての $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ に対して。

$\nabla f_i(\mathbf{y}) \leq L \|\mathbf{x} - \mathbf{y}\|$ 、すべての $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ に対して。

一次リプシッツの仮定はMLコミュニティでよく使われている。ここでは、簡単のために、全ての関数が同じリプシッツ定数 L を持つと仮定する。また、これらの関数が一様でないリプシッツ定数を持つと仮定しても、収束分析には影響がない。

前提2. 関数 f の勾配 σ_1 - 有界分散、すなわち、 $\mathbb{E}[F_i(\mathbf{x}; \xi) f_i(\mathbf{x})^2] \leq \sigma_1^2$ 、 $\forall i \in \{1, 2, \dots, m\}$ 、 $\forall \mathbf{x} \in \mathbb{R}^d$ 、 $\forall \xi \sim D_i$ 。

また、(大域的な)分散が有界であることを仮定する。この論文の

$$\frac{1}{m} \sum_{i=1}^m \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \sigma_g^2 \quad \text{すべての } \mathbf{x} \in \mathbb{R}^d \text{ に対してである。}$$

ン6では、DFedAvgMの広範な数値的検証を行う。本論文は結論の指摘で終わる。技術的な証明とより詳細な実験結果は付録として提供される。

また、局所分散が一様であるとの仮定は、非一様な場合への適用が容易であるため、プレゼンテーションの容易さのために使用されている。グローバル分散の仮定は[9]、[35]で用いられている。定数 σ_g は、データセット $(i)_{1 \leq i \leq m}$ の不均質性を反映しており、 $(i)_{1 \leq i \leq m}$ が同じ分布に従うとき、 $\sigma_g = 0$ となる。

DD

前提3. [36], [4] 任意の $i \in \{1, 2, \dots, m\}$ とし、 $\mathbf{x} \in \mathbf{R}^d$ を

\mathbf{R}^d では、ある $B > 0$ に対して、 $\forall f_i(\mathbf{x}) \leq B$ が成立する。

分散最適化における重要な概念は混合行列であり、これは通常、頂点集合 $= 1, \dots, m$ と辺集合を持つ連結グラフ $= (V, E)$ と関連付けられる。任意のエッジ $(i, l) \in E$ はノード i と l の間の通信リンクを表す。

GV
 $E \subseteq V \times V$

$E \subseteq V \times V$ の
 E

定義1 (混合行列). 混合行列 $\mathbf{W} = [w_{ij}] \in \mathbf{R}^{m \times m}$ は以下の性質を持つものとする。1. (グラフ) $i \neq j$ かつ $(i, j) \in E$ のとき、 $w_{i,j} > 0$ である。

それ以外の場合、 $w_{ij} > 0$; 2. (対称性) $\mathbf{W} = \mathbf{W}^T$; 3. (スカル空間特性) $\text{null}\{\mathbf{I} - \mathbf{W}\} = \text{span}\{\mathbf{1}\}$; 4. (スペクトル特性) $\mathbf{1}^T \mathbf{W} > -\mathbf{1}$.

グラフの場合、対応する混合行列は一意ではなく、グラフの隣接行列が与えられると、その最大次数行列とメトロポリス・ヘイスティングス [37] はともに混合行列になる。 \mathbf{W} の対称性は、その固有値が実数であり、非増加順に並べられることを示す。 $\lambda_1(\mathbf{W})$ を \mathbf{W} の i 番目の最大固有値とすると、 $\lambda_1(\mathbf{W}) = 1 > \lambda_2(\mathbf{W})$ $\lambda_m(\mathbf{W}) > -1$ である。¹混合行列はマルコフ連鎖の確率遷移行列の役割も果たす。 \mathbf{W} のかなり重要な定数は $\lambda = \lambda(\mathbf{W}) := \max \lambda_2(\mathbf{W}), \lambda_m(\mathbf{W})$ で、これは混合行列によって導入されるマルコフ連鎖が安定状態に収束する速度を記述しています。

3 分散型フェデレートアベレージング

3.1 モメンタムを持つ分散型FedAvg

まず、分散学習に関する先行研究を簡単に振り返る。分散学習は以下のような流れで実施される。

- 1) クライアント i はパラメータ $\mathbf{x}(i) \in \mathbb{R}^d$ の近似コピーを保持し、 $\mathbf{x}(i)$ における $\nabla f_i := \mathbf{g}(i)$ の非バイアス推定値を計算する。 $(\mathbf{x}(i))_{i \in \mathcal{N}}$ は非合意であってもよい。
- 2) (通信) クライアントはローカルパラメータを更新する $\mathbf{x}(i)$ をその近傍の加重平均として、 $\mathbf{x}(i) = \mathbf{0}$.

$$\mathbf{x}(i) = \sum_{l \in \mathcal{N}(i)} w_{i,l} \mathbf{x}(l)$$

- 3) (トレーニング) クライアント i はパラメータを $\mathbf{x}(i)$ として更新する。
 $\eta \mathbf{g}(i)$ を学習率 $\eta > 0$ とする。

アルゴリズム1 DFedAvgM

```

1: パラメータ:  $\eta > 0, KZ + \epsilon, 0\theta < 1$ .
2: 初期化:  $\mathbf{x}^0 = \mathbf{0}$ 
3: for  $t = 1, 2, \dots$  do.
4:   for  $i = 1, 2, \dots, m$ 
5:     ノード  $i$  が局所学習を行う (4)  $K$  回、送信します
6:    $\mathbf{z}^t(i) = \mathbf{y}^{t,K}(i)$ 
7:   ノード  $i$  は (5) のように更新します。
8: end for

```

図1(a)に示すように、従来の分散化では、学習反復のたびに通信ステップが必要であった。このことは、上記のバニラ分散化アルゴリズムがFedAvgとは異なり、FedAvgは通信の前に複数の局所的な学習ステップを行うことを示している。このため、分散化アルゴリズムの方式を若干変更する必要がある。簡単のために、我々の分散型FedAvgアルゴリズムの動機付けのためにDSGDを修正することを考える。元のDGDを適用する場合、以下の点に注意する必要がある。

を解くと、次のような反復に行き着く。

$$\begin{aligned} \mathbf{x}^{t+1}(i) &= \sum_{l \in \mathcal{N}(i)} w_{i,l} \mathbf{x}^t(l) - \gamma \mathbf{g}^t(i) \\ &= \sum_{l \in \mathcal{N}(i)} w_{i,l} [\mathbf{x}^t(l) - \gamma \mathbf{g}^t(l)] + \gamma \mathbf{g}^t(i) \end{aligned} \quad (2)$$

ここで、 $\sum_{l \in \mathcal{N}(i)} w_{i,l} = 1$ であることを利用した。(2)において、もし $\mathbf{x}^t(l)$ を $\mathbf{x}^t(i)$ に置き換えると、アルゴリズムは次のように

(3)では、クライアントは1回の学習反復の後に隣人と通信するが、これを連合最適化の設定に一般化することが可能である。(3)の1回のSGD反復を、heavy-ball[38]反復を含む複数のSGDに置き換えます。従って、DFedAvgMは以下のように表現されます。各 $t \in \mathbb{Z}_+$ において、各クライアント $i \in \{1, 2, \dots, m\}$, $\mathbf{y}^{t,0}(i) = \mathbf{x}^t(i)$ とする。そして、各ノードの内部反復は次のように実行される。

$$\mathbf{y}^{t,k+1}(i) = \mathbf{y}^{t,k}(i) - \eta \nabla f_i(\mathbf{y}^{t,k}(i)) + \theta (\mathbf{y}^{t,k}(i) - \mathbf{y}^{t,k-1}(i)), \quad (4)$$

where $\nabla f_i(\mathbf{y}^{t,k}(i)) = \nabla f_i(\mathbf{y}^{t,k}(i))$. After K inner iterations in each local client, the resulting parameters $\mathbf{z}^t(i)$ is sent to its neighbors $\mathcal{N}(i)$. Every client then updates its parameters by taking the local averaging as follows

$$\mathbf{x}^{t+1}(i) = \sum_{l \in \mathcal{N}(i)} w_{i,l} \mathbf{z}^t(l). \quad (5)$$

DFedAvgMの手順は、図1 (b) のようになります。DFedAvgMはローカルコンピューティングと通信のトレードオフの関係にあることがわかる。通信コストは通常、計算コストよりもはるかに高いことがよく知られており[39]、DFedAvgMはDSGDよりも効率的であることを示しています。

アルゴリズム2 量子化されたDFedAvgM

```

1: パラメータ:  $\eta > 0, KZ + \epsilon, 0\theta < 1, s, b$ .
2: 初期化:  $\mathbf{x}^0 = \mathbf{0}$ 
3: for  $t = 1, 2, \dots$  do.
4:   for  $i = 1, 2, \dots, m$ 
5:     ノード  $i$  は局所学習 (4) を  $K$  回行い、 $\mathbf{q}(i)$  を送信する。
6:    $\mathbf{Q}[\mathbf{y}^{t,K}(i), \mathbf{x}^t(i)]$  to  $(i)$ 
7:   ノード  $i$  は (7) のように更新します。

```

```

7:   end for
8: end for

```

3.2 量子化による効率的な通信

DFedAvgMでは、クライアント i は、 $\mathbf{x}^t(i)$ をその隣接するノード $\mathcal{N}(i)$ へ送信する。したがって、隣接数 $|\mathcal{N}(i)|$ のとき

が大きくなると、クライアント-クライアント間の通信がアルゴリズム効率の主要なボトルネックとなる。我々は、通信コストを削減するために、量子化トリックを利用する[40], [41]。具体的には、以下のような量子化手法を考える。定数 $s > 0$ と限定されたビット数 $b \in \mathbb{Z}_+$ の場合、表現可能な範囲は、 $2^{b-1}s, (2^{b-1}-1)s, \dots, 0, s, 2s, \dots, (2^{b-1}-1)s$ となる。任意の $a \in \mathbb{R}$ に対して $2^{b-1}s a < (2^{b-1}-1)s$ となる整数 $k \in \mathbb{Z}$ を求め、 ks で a を置き換える。上記の量子化方式は決定論的で、 $a \in \mathbb{R}$ に対して $q(a) := as$ と書ける。決定論のルールに加え、確率的量子化は次の式を用いる。

$$q(a) = \begin{cases} KS, & \text{W.P. } 1 - \frac{A-KS}{s}, \\ (k+1)s, & \text{w.p. } \frac{a-ks}{s}. \end{cases}$$

確率的量子化は不偏であること、すなわち、任意の $a \in \mathbb{R}$ に対して $E[q(a)] = a$ であることは容易に理解できる。座標がすべて32ビットで格納されているベクトル $\mathbf{x} \in \mathbb{R}^d$ に対して、すべての座標を量子化することを考えます。

反復される。

$$\mathbf{x}^{t+1}(i) = \sum_{l \in \mathcal{N}(i)} w_{i,l} [\mathbf{x}^t(l) - \gamma \mathbf{g}^t(l)] \quad (3)$$

$$l \in N(i) \text{ 。$$

1. これは、混合行列のスペクトル特性に基づいています。

の $\mathbf{x} = [x_1, x_2, \dots, x_d] \in \mathbb{R}^d$ である。このとき、多次元量子化演算子は次のように定義されます。

$$Q(\mathbf{x}) := [q(x)_1, q(x)_2, \dots, q(x)_d] \text{ である。} \quad (6)$$

決定論的量子化方式、確率論的量子化方式ともに、 $x_i \in [-2^{b-1}s, (2^{b-1} - 1)s]$ ならば、 $\|Q(\mathbf{x}) - \mathbf{x}\|^2 \leq d s^2$ が成立する。

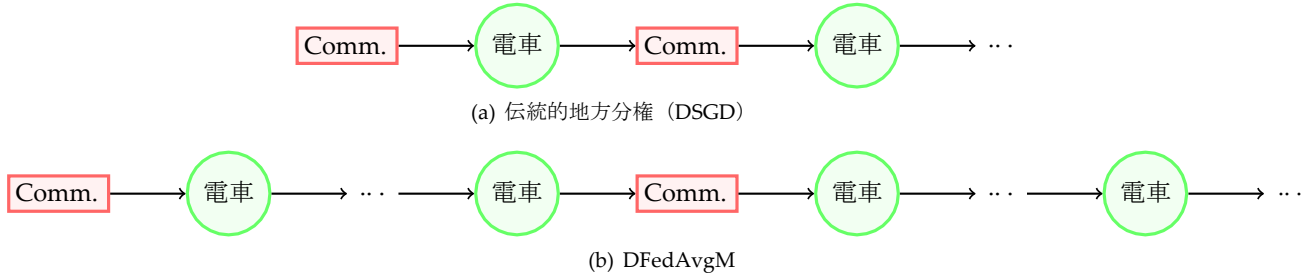


図1.従来の分散型確率的勾配降下法 (DSGD) と提案する運動量付き分散連合平均法 (DFedAvgM) の通信・学習様式の比較。DSGDでは、各クライアントは1回の学習ステップの後、近隣のクライアントと通信を行う。しかし、DFedAvgMでは、各クライアントは複数回の学習反復の後、近隣のクライアントと通信を行う。

$i \in \{1, 2, \dots\}$ の場合、 d 。この論文では、量子化された $\mathbf{z}^t(i)$ に対して成立する以下の仮定を持つ演算子です。

の2つの量子化方式があります。

仮定4. 量子化演算子 $Q: \mathbb{R}^d \rightarrow \mathbb{R}^d$

は、任意の $\mathbf{x} \in \mathbb{R}^d$ に対して $s > 0$ で $\mathbb{E} \|Q(\mathbf{x}) - \mathbf{x}\|^2 \leq s^2$ を満たしている。

パラメータを直接量子化することは、十分に滑らかな損失関数では実現可能であるが、DNNでは不可能な場合がある。このため、パラメータの差分を量子化することを考える。量子化されたDFedAvgMは、以下のことが可能である。

と要約される。(4)をK回実行した後、クライアント i は $\mathbf{q}^t(i) \leftarrow Q(\mathbf{y}^{t,K}(i) - \mathbf{x}^t(i))$ を量子化し、 $\mathbf{N}(i)$ に送信する。

$\mathbf{q}^t(j) |_{j \in \mathbf{N}(i)}$ を受信した後、各クライアントはそのローカルを更新する。

というパラメータがあります。

$$\mathbf{x}^{t+1}(i) = \mathbf{x}^t(i) + \frac{1}{|\mathbf{N}(i)|} \sum_{j \in \mathbf{N}(i)} \mathbf{q}^t(j). \quad (7)$$

各通信において、クライアント i はペア $(s, \mathbf{q}^t(i))$ を $\mathbf{N}(i)$ に送るだけでよく、その表現には $(32 + \alpha)$ が必要である。

$\deg(\mathbf{N}(i))$ ビットでなく、 $32 \deg(\mathbf{N}(i))$ ビットで送信する。これは、量子化されていないバージョンです。 d が大きく、 $b < 32$ の場合は

通信量を大幅に削減することができます。

4 C 収束分析

本節では、提案する手法の収束性を分析する。

(量子化)DFedAvgM。DFedAvgM の収束分析は、SGD、運動量付き SGD、DSGD よりもはるかに複雑です。技術的に難しいのは、 $\mathbf{z}^t(i) - \mathbf{x}^t(i)$ が勾配の不偏推定に失敗することで、勾配のを用いたSGDまたはSGDを複数回繰り返した後、 $\nabla f_i(\mathbf{x}^t(i))$ となる。

の運動量は、各クライアントで以下では、以下のことを考慮される平均点の収束を表す。

$\mathbf{x}^t := \frac{1}{m} \sum_{i=1}^m \mathbf{x}^t(i)$ の収束をまず示す。における一般的な非凸目的関数に対するDFedAvgM+。以下の定理。

$$\mathbb{E} \nabla f(\bar{\mathbf{x}}) \leq \frac{22f(\mathbf{x}_1) - 2 \min f}{\gamma(K, \eta)T}$$

+ $\alpha(K, \eta) + \beta(K, \eta)$ である。

ここで、 T は通信ラウンドの総数、定数は次のように与えられる。

$$\gamma(K, \eta) = \frac{\eta(K - \theta)}{(1 - \theta)} - \frac{64(1 - \theta)^2 L^2 K \eta^{43}}{K - \theta - 64LK} \eta^2,$$

$$\alpha(K, \eta) = \frac{((1 - \theta)L\eta + L\eta^2)(8K\sigma^2 + 32K^2\sigma^2 + \frac{64K\theta^2(\sigma^2 + B^2)}{(1 - \theta)^2})}{(K - \theta) \lg \frac{1}{(1 - \theta)^2}}.$$

$$\beta(K, \eta, \lambda) = \frac{64(1 - \theta)L^4 K \eta^{45}}{(K - \theta)} + 64L^3 K^2 \eta^4 \times \frac{(8K\sigma^2 + 32K^2\sigma^2 + 32K^2 B^2 + \frac{64K^2\theta^2}{(1 - \theta)^2}(\sigma^2 + B^2))}{[(1 - \lambda)(\eta \frac{(K - \theta)}{(1 - \theta)} - \frac{64(1 - \theta)L^2 K \eta^3}{64LK^2 \eta^2})]}.$$

$$\frac{64(1 - \theta)L^4 K \eta^{45}}{(K - \theta)} + 64L^3 K^2 \eta^4 \times \frac{(8K\sigma^2 + 32K^2\sigma^2 + 32K^2 B^2 + \frac{64K^2\theta^2}{(1 - \theta)^2}(\sigma^2 + B^2))}{[(1 - \lambda)(\eta \frac{(K - \theta)}{(1 - \theta)} - \frac{64(1 - \theta)L^2 K \eta^3}{64LK^2 \eta^2})]}.$$

$$\eta = \Theta\left(\frac{1}{\sqrt{LK}T}\right) \text{ となる。} T \text{ が十分に大きく、} 64L^2 K^2 \eta^4 + 64LK \eta < 1 \text{ とすると、} \gamma(K, \eta) = \Theta\left(\frac{1}{((1 - \theta)\sqrt{T})}\right) \text{ となります。}$$

定理1から T に関する上界を得るために、以下のよう $\eta = \Theta\left(\frac{1}{\sqrt{LK}T}\right)$ となる。 T が十分に大きく、 $64L^2 K^2 \eta^4 + 64LK \eta < 1$ とすると、 $\gamma(K, \eta) = \Theta\left(\frac{1}{((1 - \theta)\sqrt{T})}\right)$ となります。

であり、 $\alpha(K, \eta) = \Theta\left(\frac{(1 - \theta)\sigma^2 + (1 - \theta)K\sigma^2 + \frac{\theta^2}{2}K(\sigma^2 + B^2)}{K\sqrt{T}}\right)$ 、および $\beta(K, \eta, \lambda) = \Theta\left(\frac{(1 - \theta)(\frac{\sigma^2}{L} + K\sigma^2 + KB^2) + \frac{\theta^2}{2}K(\sigma^2 + B^2)}{(1 - \lambda)KT^{3/2}}\right)$ 。ベース

この η の選択と定理1について、次のようになります。DFedAvgM の収束率。

命題1. 通信ラウンド数 T が十分大きいので、次のようになる。

$$\mathbb{E} \nabla f(\bar{\mathbf{x}}) = \frac{(1 - \theta)(f(\mathbf{x}_1) - \min f)}{\sqrt{T}} \frac{1}{\frac{(1 - \theta)\sigma^2 + (1 - \theta)K\sigma^2 + \frac{\theta^2}{2}K(\sigma^2 + B^2)}{K\sqrt{T}}}$$

定理1 (一般的な非凸性)。 数列を
 $\mathbf{x}^t(i)_{t \geq 0} \{i \in \text{DFedAvgM}\}$ によって*i*のために生成される

$\{i \in \{1, 2, \dots, m\}\}$ とし、前提条件1、2、3が成立するとする。
 さらに、クライアントモデルの学習に用いる、momentumを用いたSGDのステップサイズ η が、以下を満たすとする。

$$0 < \eta \leq \frac{1}{\sqrt{L}} \text{かつ } 64L^2 K^2 \eta^2 + 64LK \eta < 1.$$

ここで、 L は ∇f のリプシッツ定数、 K は各通信の前に、 η クライアントの反復回数を設定します。
 次に

$$(1 - \theta)(\sigma^2 + K\sigma + KB) + (1 - \theta)K(\sigma^2 + B) + \frac{(1 - \theta)(\sigma^2 + K\sigma + KB) + (1 - \theta)K(\sigma^2 + B)}{(1 - \lambda)KT^{3/2}}.$$

命題1より、DFe- dAvgMの速度は局所iterationの数Kを増やすと改善されることがわかる。また、運動量 θ が0でKが十分大きいとき、その境界は次のように支配される。

$\sqrt{\frac{\sigma^2 + B^2}{(1 - \lambda)^2 T^{3/2}}}$ であり、このとき局所分散が減少する。この現象は、我々の直感的な理解と一致している：ローカルクライアントでは、大きな

Kは局所最小化することができ、その場合、局所分散バウンドしても何の支障もありません。任意の $E > 0$ エラーに到達するために $O(\frac{1}{E})$ 回の通信が必要であり、これはは、SGD、DSGDと同じです。特筆すべきは、+（プラス）です。はシミュレーションの結果よりも高速化できるかどうかは

の $f(\mathbf{x})$ 関係 $\min f + \frac{\sigma^2}{g}$ と $\frac{\sigma^2}{g} + B^2$ すなわち、もし $f(\mathbf{x}_1) - \min f + \frac{\sigma^2}{g} \leq \frac{\sigma^2}{g} + B^2$ 、 $\theta \in [0, 1]$ が大きくなると率が向上する。 $f(\mathbf{x}_1) - \min f + \frac{\sigma^2}{g} \leq \frac{\sigma^2}{g} + B^2$ 、大きな θ は、以下のようになる可能性がある。
DFedAvgMのパフォーマンスを低下させる。

上記で確立された収束の結果は、単に目的関数に滑らかな仮定が必要なだけで、非常に一般的であり、余分な性質が欠落しているために、なぜかシャープではない。例えば、最近の（非）凸の研究 [42], [43], [44] は、Polyak と Łojasiewicz [45], [46] にちなんで名付けられた PL 特性の下でアルゴリズム性能を利用している。滑らかな関数 f に対して、以下の条件でPL- ν 特性を満たすと言う。

$$\nabla f(\mathbf{x}) \succeq 2\nu(f(\mathbf{x}) - \min f), \forall \mathbf{x} \in \text{dom}(f). \quad (8)$$

よく知られた強い凸性はPL条件を含意するが、その逆はない。以下では、PL 条件の下での DFedAvgM の収束を示す。

定理2 (PL条件)。 関数 f がPL- ν 条件を満たすと仮定すると、以下の収束率が成り立つ。

$$E f(\bar{\mathbf{x}}^T) - \min f \leq [1 - \nu \gamma(K, \eta)]^T (f(\bar{\mathbf{x}}^0) - \min f) + \frac{\alpha(K, \eta) + \beta(K, \eta, \lambda)}{2\nu 2\nu}.$$

$f(\mathbf{x}^0) - \min f \geq 0$ であることから、右辺は

は $\alpha(K, \eta) + \beta(K, \eta, \lambda)$ よりも大 $= O(\eta)$ となる。まだ、 $\eta = \frac{1}{LKT}$ であれば、収束率は少なくとも $O(1/T)$ である。しかし、非常に小さな η を選ぶことはできません。さもなければ、支配項 $[1 - \nu \gamma(K, \eta)]^T (f(\mathbf{x}^0) - \min f)$ は非常にゆっくりと減衰します。もし学習率は $\eta = c_1 \ln^{c_2} T / (LKT^{c_2})$ のような形を楽しむことができる。

で、 $c_1, c_2 > 0$ とする。 c_1, c_2, c_3 の最適な選択について、以下の結果を証明できる。

命題2. $\eta = c_1 \ln^{c_2} T / (LKT^{c_2})$ とし、 $c_1, c_2 > 0$ とする。DFedAvgMの最適レートは $O(1/T)$ であり、この場合 $c_1 = L/\nu, c_2 = 1, c_3 = 1$ - , ということになります。 $\{ \}$
 $\eta = 1/(\nu KT \ln T)$ とする。 $\in \{ \}$

この発見は、強い凸性を持つSGDの最適レートに関する既存の結果[47], [48]と一致する。PL条件では、DFedAvgMの収束率が改善される。次に、量子化されたDFedAvgMの収束保証を示すと、以下になる。
という定理があります。

$$\begin{aligned} \min_{1 \leq t \leq T} E \nabla f(\bar{\mathbf{x}}^t) &= \frac{(1 - \theta)(f(\bar{\mathbf{x}}^1) - \min f)}{\sqrt{T}} \\ &\leq \frac{(1 - \theta)(\sigma^2 + K\sigma^2 + KB^2) + \frac{\theta^2}{g} K(\sigma^2 + B^2)}{K\sqrt{T} + \frac{(1 - \theta)l}{(1 - \lambda)KT^{3/2}}} + Ts. \end{aligned}$$

もし関数 f がさらにPL条件を満たし $\eta = \frac{1}{\sqrt{TK \ln T}}$ ということになる。
 $E(f(\bar{\mathbf{x}}^T) - \min f) = O(\frac{1}{T} + Ts)$ とする。

According to Theorem 3, to reach any given $E > 0$ error in general case, we need to set $s = O(E^2)$ and set the number of communication round as $T = \Theta(\frac{1}{E})$. While with PL condition, we set $T = \Theta(\frac{1}{E})$ and $s = O(E^2)$. It follows $E(f(\mathbf{x}^T) - \min f) = O(E)$. Therefore, under the PL condition, the number of communication round is reduced.

In the following, we provide a sufficient condition for communications-saving of the two quantization rules men-

命題3. のステップサイズを使って、ビットでストキャステイックまたはデターミニスティック量子化規則を使用すると仮定する。

$\eta = \frac{1}{LKT}$ で学習させたパラメータが

は、すべてのクライアントがオーバーフローしない、つまり、すべての座標が $[2^{b-1}s, (2^{b-1}-1)s]$ に含まれる。仮定1、2、3が成り立つとする。もし、所望の誤差

$$t > (1 - \theta) \frac{3LBSd}{4} \times \frac{\sigma^2 + \frac{8}{2(f(\mathbf{x}^0) - \min f) + \frac{l}{K}} + \frac{32\sigma^2}{g} + \frac{64\theta^2(\sigma^2 + B^2)}{(1 - \theta)^2}}{1}$$

と $b < \frac{128}{9} + \frac{32}{9}$, 量子化されたDFedAvgMは以下のようになります。

は、32ビットでDFedAvgMに勝ります。

命題3は、クアンの優劣を示すものである。の場合、DFedAvgMは ν を保持する。

$O((1 - \theta)s)$ よりも小さい。また、 K

が増加すると、 E の下限保証値が減少するため、局所的な反復を複数回行う必要があることがわかる。さらに、 θ を大きくしても下限は減少する。

とき **5P**

定理3. シーケンス \mathbf{x}^t ($t \geq 0$) が量子化されたDFedAvgMによって生成されるとすると、すべての $i, 1, 2, \dots, m$ とし、定理1および仮定4のすべての仮定が成立する。

ROOFS

5.1 テクニカルレmmas

$\mathbf{1} := [1, 1, \dots, 1]^T \in \mathbb{R}^m$ と定義する。

$$\mathbf{P} := \frac{\mathbf{1}\mathbf{1}^T \mathbf{R} \mathbf{m} \times \mathbf{m}}{m}.$$

行列 \mathbf{A} に対して、そのスペクトルノルムを \mathbf{A}_{op} と表
 $\eta = \Theta()$ と \implies が成立し、 T が十分に大きい場合に
 は、以下ようになります。

する。¹ T

LK

2. この学習率は、MLコミュニティでよく使われている。

記する。また、 $\mathbf{X} := \mathbf{x}(1), \mathbf{x}(2), \dots$ を定義する。、 $\mathbf{x}(m)$
 $\in \mathbb{R}^*$.

Lemma 1. [Lemma 4, [4]]. 任意の $k \in \mathbb{Z}_+$ に対して、混合行列は
 $\mathbf{W} \in \mathbb{R}^m$ を満たす。

$$\overline{\mathbf{W}} - \mathbf{P}_{\text{op}} \preceq \lambda,^k$$

ここで、 $\lambda := \max\{|\lambda_2|, |\lambda_m(W)|\}$ である。

また、筆者は[命題1、[21]]において、 \mathbf{W}^k -

$\mathbf{P}_{\text{op}} \preceq C\lambda^k$ 行列に依存するある $C > 0$ の場合。

Lemma 2. 仮定2と仮定3が成立し、 $0 \leq \theta < 1$ であるとする。 $(\mathbf{y}^{t,k}(\mathbf{i}))_{t \geq 0}$ を(量子化)DFedAvgMによって生成されたものであるとする。すると、以下のようになる。

$$\mathbb{E} \|\mathbf{y}^{t,k+1}(\mathbf{i}) - \mathbf{y}^{t,k}(\mathbf{i})\|^2 \leq \frac{1}{(1-\theta)^2} (2\eta^2 \sigma^2 + 2\eta^2 B^2)$$

$0 \leq k \leq K-1$ のとき

Lemma 3. ステップサイズ $0 < \eta \leq \frac{1}{L}$ 、 $\mathbf{i} \in \{1, 2, \dots, m\}$ とし、 $(\mathbf{y}^{t,k}(\mathbf{i}))_{t \geq 0}$ 、 $(\mathbf{x}^t(\mathbf{i}))_{t \geq 0}$ は、すべての $\mathbf{i} \in \{1, 2, \dots, m\}$ について、(量子化)DFedAvgMによって生成されるものとする。

$\{1, 2, \dots, m\}$. 仮定3が成立する場合、次が成り立つ。

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E} \|\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{x}^t(\mathbf{i})\|^2 \leq C_1 \eta^2 + 32K^2 \eta^2 \mathbb{E} \|\nabla f(\mathbf{x}^t(\mathbf{i}))\|^2$$

ここで、 $C := 8K\sigma^2 + 32K^2\sigma^2 + \frac{m}{64K\theta}(\sigma^2 + B^2)$ の

$0 \leq k \leq K$ とする。

$\mathbf{y}^{t,K}(\mathbf{i}) = \mathbf{z}^t(\mathbf{i})$ であることから、Lemma 3 も以下のようになり立つ。

Lemma 4. ステップサイズ $\eta > 0$ を与え、 $\{\mathbf{x}^t(\mathbf{i})\}_{t \geq 0}$ とする。

すべての $\mathbf{i} \in \{1, 2, \dots, m\}$ に対してDFedAvgMによって生成される。 m . 仮定3が成立する場合、以下のような境界がある。

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E} \|\mathbf{x}^t(\mathbf{i}) - \mathbf{x}^t(\mathbf{i})\|^2 \leq \frac{\eta^2}{1-\lambda}, \quad (9)$$

ここで、 $C := 8K\sigma^2 + 32K^2\sigma^2 + \frac{m}{64K\theta}(\sigma^2 + B^2)$

$$32K^2 B^2.$$

Lemma 5. ステップサイズ $\eta > 0$ を与え、 $\{\mathbf{x}^t(\mathbf{i})\}_{t \geq 0}$ とする。

は、すべての $\mathbf{i} \in \{1, 2, \dots, m\}$ に対して量子化されたDFedAvgMによって生成される。

$\{1, 2, \dots, m\}$. 仮定3が成立する場合、以下のようになる。

$$\frac{1}{m} \sum_{i=1}^m \mathbb{E} \|\mathbf{x}^t(\mathbf{i}) - \mathbf{x}^t(\mathbf{i})\|^2 \leq \frac{\eta^2}{1-\lambda} + \frac{2\eta^2}{1-\lambda}. \quad (10)$$

5.2 技術的レマの証明

5.2.1 レマ2の証明

任意の $\psi > 0$ が与えられると、Cauchyの不等式により、以下のようになる。

$$\mathbb{E} \|\mathbf{y}^{t,k+1}(\mathbf{i}) - \mathbf{y}^{t,k}(\mathbf{i})\|^2$$

$\mathbf{b} = \theta(\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{y}^{t,k-1}(\mathbf{i}))$ とする。一般性を損なわず、 $\theta \neq 0$ とする。 $\psi = \frac{1}{\theta} - 1$ とすると、次が得られる。

$$\mathbb{E} \|\mathbf{y}^{t,k+1}(\mathbf{i}) - \mathbf{y}^{t,k}(\mathbf{i})\|^2 \leq \theta \mathbb{E} \|\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{y}^{t,k-1}(\mathbf{i})\|^2 + \frac{2\eta^2}{1-\theta} + \frac{2\eta^2 B^2}{1-\theta}.$$

数学的帰納法を用いて、任意の整数 $0 \leq k$

K に対して、以下のようになる。

$$\mathbb{E} \|\mathbf{y}^{t,k+1}(\mathbf{i}) - \mathbf{y}^{t,k}(\mathbf{i})\|^2 \leq \frac{2\eta^2}{1-\theta} + \frac{2\eta^2 B^2}{1-\theta} + \frac{2\eta^2 \sigma^2 + 2\eta^2 B^2}{(1-\theta)^2}.$$

5.2.2 レマ3の証明

なお、任意の $k \in \{0, 1, \dots, K-1\}$ のとき、ノード \mathbf{i} では次が成り立つ。

$$\begin{aligned} \mathbb{E} \|\mathbf{y}^{t,k+1}(\mathbf{i}) - \mathbf{x}^t(\mathbf{i})\|^2 &= \mathbb{E} \|\mathbf{y}^{t,k}(\mathbf{i}) - \eta \tilde{\mathbf{g}}^k(\mathbf{i}) - \mathbf{x}^t(\mathbf{i}) + \theta(\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{y}^{t,k-1}(\mathbf{i}))\|^2 \\ &\leq \mathbb{E} \|\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{x}^t(\mathbf{i}) - \eta \tilde{\mathbf{g}}^k(\mathbf{i}) - \nabla f(\mathbf{y}^{t,k}(\mathbf{i})) + \nabla f(\mathbf{y}^{t,k}(\mathbf{i}))\|^2 \\ &\quad + \theta \mathbb{E} \|\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{y}^{t,k-1}(\mathbf{i})\|^2 \leq \text{I} + \text{II}, \end{aligned}$$

ここで、 $\text{I} = (1 - \frac{1}{2K-1}) \mathbb{E} \|\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{x}^t(\mathbf{i}) - \eta \tilde{\mathbf{g}}^k(\mathbf{i}) - \nabla f(\mathbf{y}^{t,k}(\mathbf{i}))\|^2$ と $\text{II} = 2K\eta^2 \mathbb{E} \|\nabla f(\mathbf{y}^{t,k}(\mathbf{i})) - \nabla f(\mathbf{x}^t(\mathbf{i}))\|^2$ と $\text{III} = \theta \mathbb{E} \|\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{y}^{t,k-1}(\mathbf{i})\|^2$ とする。 $\tilde{\mathbf{g}}^k(\mathbf{i})$ の不偏期待特性から、次のようになる。

$$\begin{aligned} \text{I} &= (1 - \frac{1}{2K-1}) \mathbb{E} \|\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{x}^t(\mathbf{i}) - \eta \tilde{\mathbf{g}}^k(\mathbf{i}) - \nabla f(\mathbf{y}^{t,k}(\mathbf{i}))\|^2 \\ &\quad + \eta^2 \mathbb{E} \|\tilde{\mathbf{g}}^k(\mathbf{i}) - \nabla f(\mathbf{y}^{t,k}(\mathbf{i}))\|^2. \end{aligned}$$

一方、レマ2を用いると、次のような境界が得られる。

$$\begin{aligned} \text{II} &\leq 8K\eta^2 \mathbb{E} \|\nabla f(\mathbf{y}^{t,k}(\mathbf{i})) - \nabla f(\mathbf{x}^t(\mathbf{i}))\|^2 \\ &\quad + 8K\eta^2 \mathbb{E} \|\nabla f(\mathbf{x}^t(\mathbf{i})) - \nabla f(\mathbf{x}^t(\mathbf{i}))\|^2 \\ &\quad + 8K\eta^2 \mathbb{E} \|\nabla f(\mathbf{x}^t(\mathbf{i}))\|^2 + 8K\theta^2 \mathbb{E} \|\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{y}^{t,k-1}(\mathbf{i})\|^2 \\ &\leq 8L K \eta^2 \mathbb{E} \|\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{x}^t(\mathbf{i})\|^2 + 8K\eta \sigma g \\ &\quad + 16K\theta^2 \mathbb{E} \|\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{y}^{t,k-1}(\mathbf{i})\|^2 \end{aligned}$$

$$= \mathbb{E} \left[\eta \mathbf{g}^{t,k}(\mathbf{i}) + \frac{\theta(\mathbf{y})^{t,k}}{\mathbf{y}} (\mathbf{i}) - \mathbf{t}, k-1(\mathbf{i}) \right]^2$$

$$\stackrel{a)}{\leq} (1 + \phi) \mathbb{E} \left[\mathbf{y}^k(\mathbf{i}) - \mathbf{t}, k-1(\mathbf{i}) \right]^2$$

$$+ (1 + \frac{1}{\psi}) \eta^2 \mathbb{E} \left[\mathbf{g}^k(\mathbf{i}) - \nabla \mathbf{f}(\mathbf{y}^{t,k}(\mathbf{i})) + \nabla_{\mathbf{i}} \mathbf{f}(\mathbf{y}^{t,k}(\mathbf{i})) \right]^2$$

$$\leq (1 + \phi) \theta^2 \mathbb{E} \left[\mathbf{y}^{t,k}(\mathbf{i}) - \mathbf{y}^{t,k-1}(\mathbf{i}) \right]^2$$

$$+ (2 + \frac{2}{\psi}) \eta^2 \nabla \mathbf{f}(\mathbf{y}^{t,k}(\mathbf{i}))$$

$$\bar{\psi} \mathbf{i} + 8K \eta^2 \sigma^2 + 8K \eta^2 \mathbb{E} \left[\nabla \mathbf{f}(\mathbf{x}^t(\mathbf{i})) \right]^2 +$$

$$+ 2(1 + \frac{1}{\psi}) \eta^2 \mathbb{E} \left[\mathbf{g}^k(\mathbf{i}) - \nabla_{\mathbf{i}} \mathbf{f}(\mathbf{y}^{t,k}(\mathbf{i})) \right]^2,$$

$$\bar{\psi} \leq$$

ここで、 $a)$ はCauchyの不等式 $\mathbb{E} \left[\mathbf{a} + \mathbf{b} \right]^2 \leq 16K \theta^2$

$(1 + \frac{1}{\psi}) \mathbb{E} \left[\mathbf{a}^2 + (1 + \phi) \mathbb{E} \left[\mathbf{b}^2 \right] \right]$ で $\mathbf{a} = \eta \mathbf{g}^{t,k}(\mathbf{i})$ とする。

$$+ 8K \eta \mathbb{E} \left[\nabla \mathbf{f}(\mathbf{x}(\mathbf{i})) \right] + (1 - \theta)^2 (\eta \sigma l + \eta B) \text{ とする}$$

したがって、次のようになります。

$$\mathbb{E} \left[\mathbf{y}^{t,k+1}(\mathbf{i}) - \mathbf{x}^t(\mathbf{i}) \right]^2$$

$$\leq (1 + \frac{1}{8L K \eta}) \mathbb{E} \left[\mathbf{y}^{22t}(\mathbf{i}) - \mathbf{x}(\mathbf{i}) + \frac{K t^2 \sigma^2}{2 \eta} \right]$$

$$2K - 1$$

$$16K \theta^2 \quad l$$

$$g$$

$$(1 - \theta)^2 (\eta^2 \sigma^2 + \eta^2 B^2)$$

$$\frac{1}{t,k} \quad t22$$

$$22 \ 2$$

$$(1 + \frac{K-1}{22}) \mathbb{E} \left[\mathbf{y}(\mathbf{i}) - \mathbf{x}(\mathbf{i}) \right]^2 + 2 \eta \sigma l + 8K \eta \sigma g$$

$$+ \frac{1}{(1 - \theta)^2} (\eta \sigma l + \eta B) + 8K \eta \mathbb{E} \left[\nabla \mathbf{f}(\mathbf{x}(\mathbf{i})) \right],$$

$$\begin{aligned} &= \sum_{i=1}^m \mathbf{z}_i^2 \\ &+ \\ &\theta \\ &\eta \end{aligned}$$

$$\leq \left(\sum_{j=0}^{t-1} \lambda^{t-1-j} \right) \left(\sum_{j=0}^{t-1} \lambda^{t-1-j} \mathbb{E} \zeta_j^2 \right)$$

直接計算すると、以下のようになります。
。

$$\mathbb{E} \zeta_j^2 \leq \mathbf{W}^2 - \mathbb{E} \mathbf{X}^j - \mathbf{Z}^j^2 \leq \mathbb{E} \mathbf{X}^j - \mathbf{Z}^j^2.$$

$$= -\eta \sum_{i=1}^m \frac{\sum_{k=0}^K \nabla_{\mathbf{f}_i}^T(\mathbf{y}^{t,k-1}(i))}{m} + \theta (\mathbf{z}^t - \mathbf{x}^t)$$

したがって、次のようになります。

$$\begin{aligned} \mathbf{z}^t - \mathbf{x}^t &= \frac{1}{1-\theta} \left(-\eta \sum_{i=1}^m \frac{\sum_{k=0}^K \nabla_{\mathbf{f}_i}^T(\mathbf{y}^{t,k}(i))}{m} \right) \\ &= \frac{1}{1-\theta} \left(\eta \sum_{i=1}^m \frac{\sum_{k=0}^K \nabla_{\mathbf{f}_i}^T(\mathbf{y}^{t,k-1}(i))}{m} \right) \end{aligned} \tag{18}$$

のリップシッツ連続性から、 ∇f が得られる。

$$\begin{aligned} & \mathbf{f}(\mathbf{x}) \leq \mathbf{E}f(\mathbf{x}) + \mathbf{E}(\nabla f(\mathbf{x}), \mathbf{z} - \mathbf{x}) \\ & \quad + \frac{L}{2} \mathbf{E} \|\mathbf{z} - \mathbf{x}\|^2, \quad (19) \end{aligned}$$

ここで、(17)を用いた。とする。

$$K = \frac{K - \theta}{1 - \theta},$$

が導き出される。

$$\begin{aligned} & \mathbf{E}K \nabla f(\bar{\mathbf{x}}t) - (\mathbf{z}t - \mathbf{x}t) / K \\ & = \mathbf{E}K \nabla f(\bar{\mathbf{x}}t) - \eta \nabla f(\bar{\mathbf{x}}t) + \eta \nabla f(\bar{\mathbf{x}}t) - (\mathbf{z}t - \mathbf{x}t) / K \\ & = -\eta K \mathbf{E} \nabla f(\bar{\mathbf{x}}t)^2 + \mathbf{E}(\nabla f(\bar{\mathbf{x}}t), \eta \nabla f(\bar{\mathbf{x}}t) + (\mathbf{z}t - \mathbf{x}t) / K) \\ & \stackrel{a)}{\leq} -\eta K \mathbf{E} \nabla f(\bar{\mathbf{x}}t)^2 \\ & \quad + \eta \mathbf{E} \nabla f(\bar{\mathbf{x}}t) \cdot \left[\frac{\sum_{i=1}^m \mathbf{E} \nabla f_i(\mathbf{x}^t) \cdot \nabla f_i(\mathbf{y}^{t, K-1}(i))}{m} \right] \\ & \quad + \frac{\sum_{i=1}^m (1 - \theta) [\nabla f_i(\mathbf{x}^t) - \nabla f_i(\mathbf{y}^{t, K-1}(i))] }{m} \\ & \leq -\eta K \mathbf{E} \nabla f(\bar{\mathbf{x}}t)^2 + \eta \frac{1}{m} \sum_{i=1}^m \mathbf{E} \nabla f(\bar{\mathbf{x}}t) \cdot (\mathbf{x}^t - \mathbf{y}^{t, K}(i)) \\ & \leq -\eta K \mathbf{E} \nabla f(\bar{\mathbf{x}}t)^2 + \frac{\eta K}{2} \mathbf{E} \nabla f(\bar{\mathbf{x}}t)^2 \\ & \quad + \frac{\eta L^2}{2K} \sum_{i=1}^m \mathbf{E} \nabla f(\mathbf{x}^t(i))^2 \\ & \quad + \frac{(C_1 \eta^2 + 32K^2 \eta^2 t^{i-1})}{2K}, \end{aligned}$$

ここで、a)は(18)を用いている。同様に、以下のようになる。

$$\begin{aligned} & \mathbf{E}(\bar{\mathbf{x}}t+1 - \mathbf{x}t)^2 = \mathbf{E}(\mathbf{z}t - \mathbf{x}t)^2 \\ & \leq 2m \sum_{i=1}^m \mathbf{E}(\mathbf{z}^t(i) - \mathbf{x}^t)^2 \\ & \leq \frac{L}{2} C_1^2 + 16LK^2 \eta^2 \sum_{i=1}^m \mathbf{E} \nabla f(\mathbf{x}^t(i))^2 \end{aligned}$$

表が(19)は次のように

$$\begin{aligned} \mathbf{E}f(\bar{\mathbf{x}}t+1) & \leq \mathbf{E}f(\bar{\mathbf{x}}t) + \frac{\eta K}{2} \mathbf{E} \nabla f(\bar{\mathbf{x}}t)^2 + \frac{L2K}{2} \frac{C_1^2}{\eta^2} \\ & + \frac{L}{2} C_1 \eta^2 + 16L^2 K^4 \cdot \frac{2}{2} + 16LK^2 \eta^2 \frac{\eta^3 t^{i-1}}{2} \frac{1}{K} \end{aligned}$$

したがって、次のようになります。

$$\mathbf{E}f(\bar{\mathbf{x}}t+1) \leq \mathbf{E}f(\bar{\mathbf{x}}t)$$

$$\begin{aligned} & \eta(K - \theta) \frac{32(1 - \theta)L^2 K \eta^{43}}{2(1 - \theta)} - 32LK \eta^{22} \\ & \times \mathbf{E} \nabla f(\bar{\mathbf{x}}t)^2 + \frac{(1 - \theta)L^2 L \eta^2}{K^2} + \frac{2(K - \theta)^2}{64K^2 \theta^2} \\ & \times (8K \sigma + 32K \sigma \theta + 32K B + \frac{1}{(1 - \theta)^2} (\sigma + B)) \\ & + (32(1 - \theta)L^4 K^4 \eta^5 / (K - \theta) + 32L^3 K^2 \eta^4) / (1 - \lambda) \\ & \times (8K \sigma + 32K \sigma \theta + 32K B + \frac{1}{(1 - \theta)^2} (\sigma + B)) \end{aligned} \quad (20)$$

不等式(20)を $t=1$ から T まで和集合させるとが証明された。

5.4 定理2の証明

PLの条件付き。

$$\mathbf{E} \nabla f(\bar{\mathbf{x}}t)^2 \geq 2 \nu \mathbf{E}(f(\bar{\mathbf{x}}t) - \min f)$$

(20)からスタートする。

$$\begin{aligned} \mathbf{E}f(\bar{\mathbf{x}}t) & \leq \mathbf{E}f(\mathbf{x}t) - \nu(K, \eta) \mathbf{E}(f(\mathbf{x}t) - \min f) \\ & \quad + \frac{\gamma(K, \eta) \alpha(K, \lambda)}{2} + \frac{\gamma(K, \eta) \beta(K, \eta, \lambda)}{2}. \end{aligned} \quad (21)$$

$\xi_t := \mathbf{E}(f(\mathbf{x}t) - \min f)$ を定義することにより、次が成り立つ。

$$\begin{aligned} \xi_t & \leq [1 - \nu(K, \eta)] \xi_t \\ & \quad + \frac{\gamma(K, \eta) \alpha(K, \eta)}{2} + \frac{\gamma(K, \eta) \beta(K, \eta, \lambda)}{2}. \end{aligned} \quad (22)$$

22

したがって、次に導かれるのは

$$\begin{aligned} \xi_T & \leq [1 - \nu(K, \eta)] \xi_{T_0} \\ & \quad + \gamma \left(\frac{K, \eta}{T-1} \alpha(K, \eta) + \frac{\gamma(K, \eta) \beta(K, \eta, \lambda)}{22} \right) \\ & \quad \times \left(\sum_{t=0}^{T-1} [1 - \nu(K, \eta)]^t \right) \\ & \leq [1 - \nu(K, \eta)] \xi_0 \\ & \quad + \gamma \left(\frac{K, \eta}{T\alpha} \alpha(K, \eta) + \frac{\gamma(K, \eta) \beta(K, \eta, \lambda)}{(K, \eta) \beta(K, \eta, \lambda)} \frac{1}{22} \right) \\ & \quad + \frac{2}{K} \sum_{i=1}^m \mathbf{E} \nabla f(\mathbf{x}^t(i))^2 \end{aligned}$$

$= [1 - \nu \gamma(K, \eta)]$ となる。 $\xi_0 + 2\nu + 2\nu$ 。

直接計算と定理4により、以下ようになります。

$$\begin{aligned} & \frac{m \mathbb{E}_{i=1}^m \|\nabla f(\mathbf{x}^t(i))\|^2}{m} \\ &= \frac{m \mathbb{E}_{i=1}^m \|\nabla f(\mathbf{x}^t(i)) - \nabla f(\mathbf{x}^t) + \nabla f(\mathbf{x}^t)\|^2}{m} \\ &\leq \frac{m \mathbb{E}_{i=1}^m \|\nabla f(\mathbf{x}^t(i)) - \nabla f(\mathbf{x}^t)\|^2 + 2 \mathbb{E} \|\nabla f(\mathbf{x}^t)\|^2}{m} \\ &\leq 2L \frac{m \|\mathbf{x}^t(i) - \mathbf{x}^t\|^2}{m} + 2 \mathbb{E} \|\nabla f(\mathbf{x}^t)\|^2 \\ &\leq \frac{2L2C2\eta22c}{-\lambda} + 2 \mathbb{E} \|\nabla f(\mathbf{x}^t)\|. \end{aligned}$$

そして、その結果が証明される。

5.5 命題2の証明

簡単に計算すると

$$\gamma(K, \eta) = \Theta\left(\frac{1}{T^c}\right) \\ \frac{\alpha(K, \eta)}{2} + \frac{\beta(K, \eta, \lambda)}{2} = O\left(\frac{1}{T^c}\right) \text{ となる。}$$

したがって、定理2の第1項を束縛すればよいことになる。と

T が大きい場合、 $\gamma(K, \eta) \rightarrow 0$ となり、その

対数値は $T \log[1 - \nu \gamma(K, \eta)] = T \log[1 - \nu \gamma(K, \eta)] = \Theta(-T \nu \gamma(K, \eta))$ です。

。

この設定により、以下の

ようになります。

$$T \nu \gamma(K, \eta) \approx \frac{\nu c_1 \ln^3 T}{L T c_2 - 1}.$$

。

すると、次の
ようになりま
す。

$$\mathbb{E}f(\mathbf{x}^T) - \min f = O\left(\exp\left(-\frac{vc \ln^3 T}{LKTc_2 - 1Tc_2}\right) + \frac{1}{LKTc_2 - 1Tc_2}\right)$$

まず、 c_2 をどのように選ぶかを考える。L'Hospital's64LK

のルールは、任意の $\delta > 0$ に対
して、 $T \rightarrow +\infty$ として

$$\exp\left(-\frac{1}{T\delta}\right) \rightarrow 1.$$

従って、以下のよう よりも速い速度が遅くなり
に設定する必要があ
ります。 $c_2 \leq 1$
 $O(T)$ である。このため、 $c_1 = \nu \frac{L}{c_2}$ 、 $c_2 = 1$ 、 $c_3 = -1$ を
選択する。

5.6 定理3の証明

\mathbf{y}^t として $i=1 \vee \dots \vee m$ 量子化されたDFedAvgMでは、以下の
ようになります。

下つ $\mathbf{x}^{t+1} - \mathbf{x}^t = Q(\mathbf{y}^t - \mathbf{x}^t)$ リプシッツの連続性で
腹 $\nabla f(\mathbf{x})$ とな
りま
す。

$$\mathbb{E}f(\mathbf{x}^{t+1}) \leq \mathbb{E}f(\mathbf{x}^t) - \frac{\mathbb{E} \|\nabla f(\mathbf{x}^t), Q(\mathbf{y}^t - \mathbf{x}^t)\|^2}{2} - \frac{\mathbb{E} \|\mathbf{x}^t - \mathbf{x}\|^2}{2},$$

私たちは

$$\begin{aligned} & \mathbb{E}(\nabla f(\mathbf{x}^t), Q(\mathbf{y}^t - \mathbf{x}^t)) \\ &= \mathbb{E}(\nabla f(\mathbf{x}^t), \mathbf{y}^t - \mathbf{x}^t) + \mathbb{E}(\nabla f(\mathbf{x}^t), \mathbf{y}^t - \mathbf{x}^t - Q(\mathbf{y}^t - \mathbf{x}^t)) \\ &\leq (\nabla f(\mathbf{x}^t), \mathbf{y}^t - \mathbf{x}^t) + B \int ds \end{aligned}$$

$$\begin{aligned} \frac{L}{2} \mathbb{E} \|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 &= \frac{L}{2} \mathbb{E} \|Q(\mathbf{y}^t - \mathbf{x}^t)\|^2 \\ &\leq L \mathbb{E} \|\mathbf{y}^t - \mathbf{x}^t\|^2 + L \mathbb{E} \|Q(\mathbf{y}^t - \mathbf{x}^t) - (\mathbf{y}^t - \mathbf{x}^t)\|^2 \\ &\leq L \mathbb{E} \|\mathbf{y}^t - \mathbf{x}^t\|^2 + \frac{L}{2} \mathbb{E} \|\mathbf{y}^t - \mathbf{x}^t\|^2 \\ &\text{以下略} \end{aligned}$$

なお、 $\mathbb{E}(\nabla f(\mathbf{x}^t), \mathbf{y}^t - \mathbf{x}^t)$ と $\mathbb{E} \|\mathbf{y}^t - \mathbf{x}^t\|^2$ はともに、以下の
ことが可能です。
は $\mathbb{E}(\nabla f(\mathbf{x}^t), \mathbf{z}^t - \mathbf{x}^t)$ と $\mathbb{E} \|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2$ の境界を継承して
いる。

を、定理1の証明においてしたがって、次
のようになる。

$$\begin{aligned} \mathbb{E}f(\mathbf{x}^{t+1}) &\leq \mathbb{E}f(\mathbf{x}^t) - \frac{\eta K}{2} \mathbb{E} \|\nabla f(\mathbf{x}^t)\|^2 \\ &\quad + \frac{L^2 K}{2K} C_1 \eta^2 + \frac{L C_1}{2K} \eta^2 + \frac{L}{2K} \eta^2 + B \end{aligned}$$

ここで、 $\zeta(K, \eta, \lambda, s) := L^{2K^2} C \eta^3 + LC \eta^2 + (L^{2K^4} \eta^3 +$

$$32LK^2 \eta^2) \left(\frac{2L^2 C_3 \eta^2}{1-\lambda} + \frac{4L^2}{1-\lambda m} \int ds \right) + Lds^2 \sqrt{\frac{K}{ds}}$$

ステップサイズ $\eta = \frac{1}{\sqrt{K}}$ から、 $\eta K - 64L^2 K^4 \eta^3 -$ となることがわ
 $\Theta\left(\frac{1}{\sqrt{K}}\right)$ かる。

$\eta^2 > 0$ から、 T が大きくなると、L'Hospital's64LK を選ぶ。
 $s > 0$ が小さいとき、 $s^2 =$

$O(s)$ です。そして、次のよ $-64L^2 K^4 \eta^3 - 64LK^2 \eta^2 =$

うになる。 ηK
 $\Theta\left(\frac{1}{\sqrt{K}}\right)$ 。ここで、次のことを考える。

$$\begin{aligned} & \frac{32L^2 K^4}{\eta^3} + 32LK^2 \frac{2L^2 C_3 \eta^2}{1-\lambda} + \frac{4L^2 ds}{1-\lambda m} \\ & \left(\frac{K}{1-\lambda} \eta \right) \left(\frac{1-\lambda}{1-\lambda} \frac{L^2 K}{C_1 \eta^3} + \frac{Lds^2}{m} + \frac{B}{ds} \right) \\ & + LC_1 \eta^2 + \frac{L^2 K}{2K} C_1 \eta^3 + \frac{Lds^2}{m} + \frac{B}{ds} \\ & + \frac{64L^2 K^4}{\eta^3} + \frac{64LK^2 \eta^2}{\eta^3} \end{aligned} \quad (23)$$

というオーダーで、 $O\left(\frac{1}{\sqrt{K}}\right)$ $\frac{1}{K} \left(\frac{1}{1+B} \right) \frac{1}{TKT}$
 $\sigma^2 + K \sigma^2 + KB^2 + \frac{\theta^2}{2} K(\sigma^2 + B^2) \sqrt{\cdot}$
 $\frac{Lg}{(1-\theta)^{21}}$ がPL-vを満たすことを示す。もし、関数 f

$$\mathbb{E}(f(\mathbf{x}^t) - \min f)$$

$$\begin{aligned} & \leq [1 - \nu \gamma(K, \eta)]^T (f(\mathbf{x}^0) - \min f) \frac{\zeta(K, \eta, \lambda, s)}{\nu \gamma(K, \eta)} \\ & , \eta] \text{ となる。} \end{aligned}$$

$\eta =$ のと $\frac{1}{\sqrt{K}} \ln \frac{1}{\epsilon}$ 、 $[1 - \nu \gamma(K, \eta)]^T = \epsilon$ とす
き

$$\frac{\zeta(K, \eta, \lambda, s)}{\nu \gamma(K, \eta)} = \frac{1}{T} + Ts \text{ となる。}$$

5.7 命題3の証明

同じ誤差になるように、両アルゴリズムの通信コストを計
算する。 η の1より大きい次数を省略し、(20)より、以下の
ようになる。

$$f(\mathbf{x}^0) - \min f$$

$$\begin{aligned} & \mathbb{E} \|\nabla f(\mathbf{x})\|^2 \approx \frac{\eta K T}{64K \theta^2} \\ & + L\eta(8\sigma^2 + 32K\sigma^2 + (1-\theta)^2(\sigma^2 + B)) \end{aligned}$$

$$2(1 - \theta) \left(\frac{f(\mathbf{x}^t(i)) - \min f}{\sqrt{T}} \right)^2 + \frac{32L^2 K^4 \eta^2 \mathbb{E} \|\nabla f(\mathbf{x}^t(i))\|^2}{m} + \left(\frac{K}{\eta} + 32LK^2 \eta^2 \right) \frac{\sum_{i=1}^m \mathbb{E} \|\nabla f(\mathbf{x}^t(i))\|^2}{m}$$

レンマ5を用いると、以下のようになります。

$$\begin{aligned} & \frac{m \mathbb{E} \|\nabla f(\mathbf{x}^t(i))\|^2}{m \mathbb{E} \|\nabla f(\mathbf{x}^t(i))\|^2 + 2 \mathbb{E} \|\nabla f(\mathbf{x}^t(i))\|^2} \\ & \leq \frac{\sum_{i=1}^m \mathbb{E} \|\nabla f(\mathbf{x}^t(i))\|^2}{m} \\ & \leq 2L^2 \frac{\sum_{i=1}^m \|\mathbf{x}^t(i) - \mathbf{x}^t\|^2}{m} + 2 \mathbb{E} \|\nabla f(\mathbf{x}^t(i))\|^2 \\ & \leq \frac{2L C_3}{\eta} + \frac{4L ds}{1-\lambda} + 2 \mathbb{E} \|\nabla f(\mathbf{x}^t(i))\|^2. \end{aligned}$$

不等号を一緒に組み合わせる

$$\mathbb{E} f(\mathbf{x}^{t+1}) \leq \mathbb{E} f(\mathbf{x}^t) + \zeta(K, \eta, \lambda, s) \cdot \frac{64L^2 K^4 \eta^3}{2} - \frac{64LK \eta}{2} \mathbb{E} \|\nabla f(\mathbf{x}^t)\|^2$$

$$= \frac{f(\mathbf{x}^0) - \min f}{\sqrt{T}} + \frac{8(1-\theta) \sigma^2 + 32(1-\theta) K \sigma^2 + 64K \theta^2 (\sigma^2 + B^2)}{K \sqrt{T}} + \frac{g}{(1-\theta) l} \text{ である。}$$

をDFedAvgMとする。(23)より。

$$\begin{aligned} & \mathbb{E} \|\nabla f(\mathbf{x}^t)\|^2 \approx \frac{2(1-\theta)(f(\mathbf{x}^0) - \min f)}{T} \\ & \leq \frac{8(1-\theta) \sigma^2 + 32(1-\theta) K \sigma^2 + 64K \theta^2 (\sigma^2 + B^2)}{K \sqrt{T}} + \frac{g}{(1-\theta) l} \\ & + 2(1-\theta) L B \sqrt{T} \cdot \frac{K}{T} \end{aligned}$$

を考える $E > 0$ とすると 遵奉と

$$\frac{8(1-\theta) \sigma^2 + 32(1-\theta) K \sigma^2 + 64K \theta^2 (\sigma^2 + B^2)}{K \sqrt{T}} + \frac{g}{(1-\theta) l} = E.$$

つまり、DFedAvgMは T 回の反復で E の誤差を出力することができる。しかし、量子化により誤差が生じるため、量子化DFedAvgMの反復回数を増やさなければならない。反復回数を T とする。量子化DFedAvgMの E 誤差を求めるには、 $3(1-\theta)\sqrt{LB} \sqrt{ds} \leq E$ も必要であり、次のようになる。

$$3(1-\theta)^2 LB \sqrt{ds} 2(f(\bar{\mathbf{x}}^0) - \min f) + \frac{8}{K} + 32\sigma_g^2 + \frac{64\theta^2}{(1-\theta)^2} (\sigma^2 + B^2) \leq E.$$

DFedAvgMが E に到達するまでの総通信コストは

$$32dT \sum_{i=1}^n [\deg(N(i))] \text{である。}$$

一方、量子化バージョンでは、総通信コストは

$$(32+db) \frac{9}{4} \sum_{i=1}^n [\deg(N(i))] \text{である。}$$

したがって、以下のような場合、通信量を減らすことができます。

$$(32+db) \frac{9}{4} < 32d.$$

6数値結果

提案する通信量子化付きDFedAvgMを画像分類と言語モデリングの両方に適用し、単純なリング構造の通信ネットワークを考慮したDNNの学習を行う。DFedAvgMがDNNを効率的に学習できること、特に通信効率について検証することを目的とする。さらに、DFedAvgMをDNNの学習に用いた場合のメンバーシップのプライバシー保護について考察する。我々はメンバーシップ推論攻撃(MIA) [49]を適用し、(量子化)DFedAvgMが学習データのMPを保護する効率性を検証する。MIAでは、攻撃モデルとして二値分類器³あるデータ点がターゲットモデルの学習セットに含まれるかどうかを判断するものである。以下の各データセットについて、MIAを行うために、まず、その学習セットを同じ大きさのDshadowとDtargetに分割する。さらに、Dshadowを同じ大きさの2つに分割し、以下のように表す。

をDとする。とD_{train}で、Dtargetを半分に分割し、D_{train}とD_{target}を分割しています。

寛げ目標、アウ目標、MIAは以下のように進行します。

1) 電車
2) 学習したシャドウモデルを適用して、Dshadow内のすべてのデータ点を予測し、各クラスに属する対応する分類確率を学習する。次に、上位3つの分類確率(2値分類の場合は2つ)を取り出して、各データ点の特徴ベクトルを形成する。特徴ベクトルは、対応するデータ点がD^{train}にある場合は1、それ以外は0としてタグ付けされる。次に、すべてのラベル付けされた特徴ベクトルを利用して、攻撃モデルを学習する。

Dによるトレーニングの各点に対する特徴ベクトルを求めます。

Dtargetとする。最後に、攻撃モデルを利用してデータ点がD^{train}にあること。我々が構築する攻撃モデル

を適用し、ターゲットモデルの学習セットに属する確率(p)を予測する。任意の固定閾値tが与えられたとき、pがtならは訓練集合のメンバーとして分類し(正のサンプル)、p<tならは訓練集合に属しないと結論づける(負のサンプル)；従って、異なる閾値で異なる攻撃結果を得ることができる。このように、閾値を変えることで異なる攻撃結果を得ることができます。

は、メンバーシップ推論攻撃の評価として、ROC曲線下面積(AUC)を用いている。

AUCが0.5であれば、完全なメンバーシッププライバシーを確保できる(攻撃推論を行わず)、AUCが高いほど

ターゲットモデルを非公開とする。

6.1 MNISTの分類

6.1.0.1 DFedAvgMの効率：100クライアントを用いて、MNIST分類のための2つのDNNを学習させる。1) ReLU活性化を用いた、各200ユニットからなる2つの隠れ層を持つ単純な多層パーセプトロン(全パラメータ199,210個)。

であり、これを2NNと呼ぶ。2) 2つのコンボリネーション層(1層目：32チャンネル、2層目：32チャンネル)と22の最大プーリングで続く512個のユニットとReLUの活性化を持つ接続層と

最終出力層(総パラメータ数1,663,370)。我々は、MNISTデータのクライアントに対する2つの分割、すなわち、IIDと非IIDを研究する。IIDでは、データはシャッフルされ、20クライアントに分割され、各クライアントは3000例を受け取る。IIDでは

非IIDの場合、まずデータを桁ラベルでソートし、1500サイズの40シャードに分割し、20クライアントに2シャードずつ割り当てる。学習では、ローカルバッチサイズ(クライアントの学習データのバッチサイズ)を50、学習率0.01、運動量0.9とした。図2と図3は、異なる通信ビットと異なるローカルエポックを用いて、DFedAvgMによりMNIST分類のCNNを学習した結果(図2: IID、図3 Non-IID)である。これらの結果は、DNNの学習におけるDFedAvgMの効率性を確認するものであり、特にクライアントのデータがIIDである場合に有効である。IIDとNon-IIDの両方において、通信ビットはDFedAvgMの性能に影響を与えない；メンバーシップ推論攻撃における学習損失、テスト精度、AUCがほぼ同じであることが分かる。局所的な学習エポックを増やすことで、IID設定での学習を加速させることができるが、その代償としてプライバシー漏洩が速くなる。しかし、Non-IIDの場合、局所学習エポックを増やしても

DFedAvgMのトレーニングや個人情報保護に役立たない。

DFedAvgMによる2NNの学習も同様で、図4と図5を参照されたい。

6.1.0.2 DFedAvgM、FedAvg、およびDSGDの比較：次に、MNISTの分類のための2NNの学習におけるDFedAvgM、FedAvg、DSGDを比較する。FedAvg、DSGDとともに同じローカルバッチサイズ50を用い、学習率はともに0.1としている⁴。FedAvgでは、各ラウンドで学習と通信に参与する全てのクライアントを選択した。図6は3つのアルゴリズムの比較である。を超えるIID MNISTのテストロスとテスト精度を実現しました。

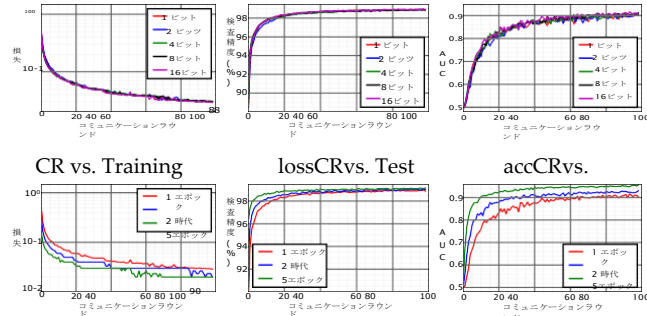
ルは2値分類器であり、データ点がターゲットモデルの訓練集合に含まれるかどうかを決定するものであることに注意

されたい. 任意のデータ $\xi \in D_{target}$ について
。

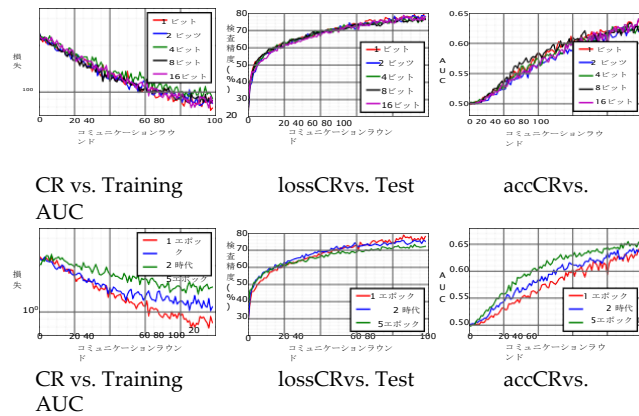
3. 攻撃モデルとしては, [49]から引用した, 64ノードの隠れ層とそれに続くソフトマックス出力関数を持つ多層パーセプトロンを使用する。

通信ラウンドと通信コストという点では
DFedAvgMはFedAvgと同程度の速度で収束し、DSGDよりもはるかに高速です。DFedAvgMはFedAvgとDSGDに対して通信コストで大きな優位性を持っています。非IIDのMNISTでは、2NNの学習は

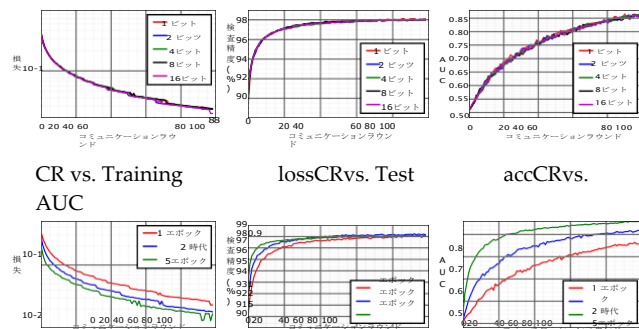
4. DFedAvgはFedAvgやDSGDよりも小さな学習率で数値的な収束が可能であることに注目したい。



CR vs. Training lossCRvs. Test accCRvs. AUC
図2.DFedAvgMを用いたIID MNIST分類のためのCNNの訓練：通信ビットは異なるがローカルエポックは1に固定（1段目）、ローカルエポックは異なるが通信ビットは16に固定（2段目）。異なる量子化DFedAvgMはほぼ同様の性能を示し、より多くのローカルエポックは、より速いプライバシーリーク率の代償として、学習を加速することができる。CR: 通信ラウンド。

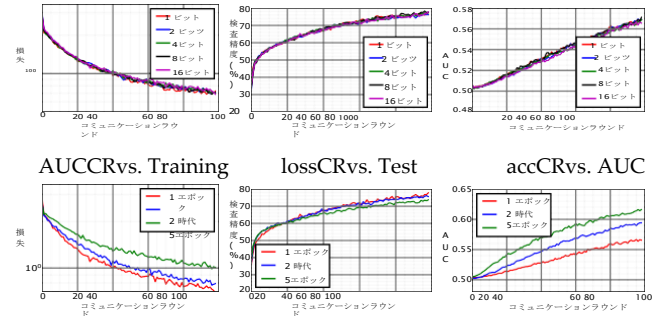


CR vs. Training lossCRvs. Test accCRvs. AUC
図3.DFedAvgMを用いた非IID MNIST分類のための学習CNN：通信ビットは異なるがローカルエポックは1に固定（1段目）、ローカルエポックは異なるが通信ビットは16に固定（2段目）。量子化されたDFedAvgMが異なっても、性能に大きな差は生じない。ローカルエポックを増やしても、学習の高速化やデータプライバシーの保護には役立たない。

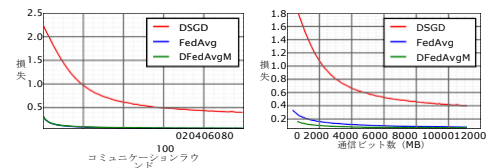


CR vs. Training lossCRvs. Test accCRvs. AUC
図4.DFedAvgMを用いたIID MNIST分類のための2NN学習：通信ビットは異なるがローカルエポックは1に固定（1段目）、ローカルエポックは異なるが通信ビットは16に固定（2段目）。量子化されたDFedAvgMの性能はほぼ同等であり、ローカルエポックを増やすと、プライバシー漏洩が速くなる代償として、学習を加速することができます。

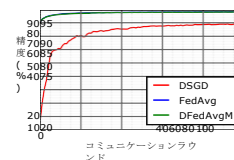
FedAvgは96.81%のテスト精度を達成できるが、DFedAvgとDSGDは共に85%を超えることはできない。この欠点は、DSGDとDFedAvgMがともに隣人としか通信しないため、隣人とそれ自身はすべての可能なクラスをカバーするのに十分な学習データを含んでいない可能性があります。



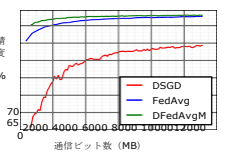
CR vs. Training lossCRvs. Test accCRvs. AUC
図5.DFedAvgMを用いた非IID MNIST分類の学習2NNの場合：通信ビットは異なるがローカルエポックは1に固定（1行目）、ローカルエポックは異なるが通信ビットは16に固定（2行目）。量子化されたDFedAvgMが異なっても、性能に大きな差は生じない。ローカルエポックを増やしても、学習の高速化やデータプライバシーの保護には役立たない。



CR vs. Test



lossCBvs. Test loss



CR vs. Test

accCBvs. Test acc

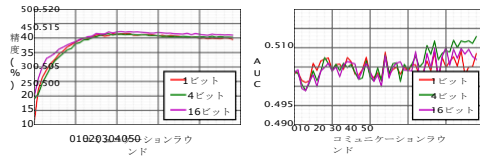
図6.MNIST分類のための2NN学習におけるDSGD、FedAvg、DFedAvgMの効率比較。(a)と(c): テスト損失とテスト精度 vs 通信ラウンド。(b), (d): 通信ビット数に対するテスト損失とテスト精度。DFedAvgMは通信ラウンド数ではFedAvgと同等であるが、通信コストの観点からはDFedAvgMはFedAvgよりも大幅に効率的であることがわかる。CR: 通信ラウンド、CB: 通信ビット。

6.2 言語モデリング用LSTM

SHAKESPEAREデータセットについて、[1]と同様の処理を行い、その結果、データセット

を1146クライアントに非IID方式で分散させた。このデータに対して、DFedAvgMを用いてスタック型文字レベルLSTM言語モデルを学習し、次の文字を予測する。非IID環境におけるDFedAvgMの問題点を解決する一つの可能な方法は、より効率的なグローバル通信のための新しいグラフ構造を設計することである。

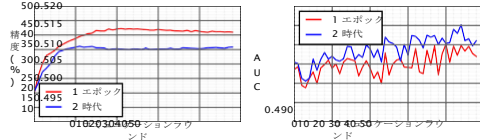
行の各文字を読み取った後このモデルは一連の文字を入力とし、それぞれを学習された8次元空間に埋め込む。次に、埋め込まれた文字は、それぞれ256ノードを持つ2つのLSTM層で処理される。最後に、2番目のLSTM層の出力は、文字ごとに1つのノードを持つソフトマックス出力層に送られる。このモデルは866,578個のパラメータを持ち、80文字のアンロール長で学習を行った。ローカルバッチサイズは10とし、学習レートは[1]と同じ1.47を用いた。また、運動量は0.9を選択した。図7は、量子化とローカルエポックを変えた場合の、MIAの下での通信ラウンド対テスト精度とAUCをプロットしたものである。これらの結果から、1)学習が進むにつれて精度とMIAの両方が向上する、2)通信コストが高くなると収束が早くなる、3)ローカルエポックが増えるとDFedAvgMの性能が悪化することが示された。



CR vs. Test

accuracyCR

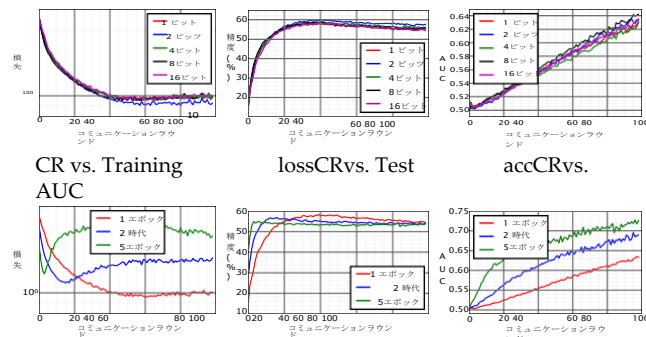
vs. AUC



CR vs. Test

accuracyCR vs. AUC

図7.DFe-dAvgMを用いたSHAKESPEARE分類のためのLSTMの学習：通信ビットは異なるがローカルエポックは1に固定（1段目）、ローカルエポックは異なるが通信ビットは16に固定（2段目）。高精度の通信を使用することで、若干の性能向上が見られます。ローカルエポックを増やしても、学習の高速化やデータプライバシーの保護には役立たない。

CR vs. Training
AUC

lossCR vs. Test

accCR vs.

CR vs. Training

lossCR vs. Test

accCR vs. AUC

図8.DFedAvgMを用いたIID CIFAR10分類のためのCNNの学習：通信ビットは異なるがローカルエポックは1に固定（1段目）、ローカルエポックは異なるが通信ビットは16に固定（2段目）。量子化されたDFedAvgMの性能はほぼ同じで、ローカルエポックを増やすと、最初のうちは学習を加速できるが、学習を続けるとあまり性能が良くならない。

6.3 CIFAR10分類

最後に、DFedAvgMを用いてResNet20をCIFAR10分類のために学習させた。この分類は3チャンネル32枚の画像からなる10クラスである。学習例5万、テスト例1万があり、これらを20クライアントに単式に分割し、[1]に従ってIID設定のみを考慮した。データ拡張とDNNは[1]で用いたものと同じものを使用する。ローカルバッチサイズは50、学習率は0.01、運動量は0.9に設定した。図8は、量子化とローカルエポックを変えた場合のMIAの通信ラウンド対テスト精度とAUCを示したものである。ローカルエポックを1に設定した場合、通信ビットの違いによって、MIAのもとでの学習損失、テスト精度、AUCに大きな差は生じない。しかし、通信ビット16固定の場合、ローカルエポックを1→2→5と増やすと、学習が収束しなくなることがわかった。

7 結語

本論文では、DFedAvgMとその量子化版を提案した。DFedAvgMは既存のFedAvgと比較して、次の2つの大きな利点がある：1) FedAvgでは、通信ラウンドごとに中央パラメータサーバとローカルクライアント間の通信が必要となり、クライアント数が非常に多い場合、この通信が非常に高価となる。これに対し、DFedAvgMでは、通信は

2) FedAvgでは、中央サーバがクライアントから更新されたモデルを収集し、中央サーバを攻撃することでFedAvgを破ることができる。のプライバシーを保護します。これに対し、概念的にはFedAvgよりDFedAvgMの方がプライバシーを破るのが難しい。

さらに、DFedAvgMとその量子化版について、一般的な非凸の仮定下で理論的な収束を確立し、（量子化）DFedAvgMの最悪の場合の収束率がDSGDのそれと同じであることを示した。特に、量子化されたDFedAvgMがDSGDと同じ収束率を持つことを証明した。

量子化された) DFedAvgMの輻輳率。

はPL条件を満たす。DFedAvgMとその量子化版がMLモデルの学習に有効であり、メンバーシッププライバシーを保護することを検証するために、広範な数値実験を行う。

参考文献

- [1] H.B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data", pp. 1273-1282, 2017.
- [2] H. Robbins and S. Monro, "A stochastic approximation method," 数理統計学年報、第22巻、No.3, pp.400-407, 1951.
- [3] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola, "Parallelized 確率的勾配降下法"『神経情報学会の進歩』（日本経済新聞出版社 処理システム, pp.2595-2603, 2010.
- [4] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, および J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems*, pp.5330-5340, 2017.
- [5] A. Lalitha, S. Shekhar, T. Javidi, and F. Koushanfar, "Fully decentralized federated learning," in *Third workshop on Bayesian Deep Learning (NeurIPS)*, 2018.
- [6] A. Lalitha, O. C. Kilinc, T. Javidi, and F. Koushanfar, "Peer-to-peer federated learning on graphs," *arXiv preprint arXiv:1901.11173*, 2019年.
- [7] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On importance of initialization and momentum in deep learning," in *International conference on machine learning*, pp.1139-1147, 2013.
- [8] T.-M. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.
- [9] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konecny, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," in *International Conference on Learning Representations*, 2021.
- [10] T. Chen, G. Giannakis, T. Sun, and W. Yin, "Lag: Lazily aggregated gradient for communication-efficient distributed learning," in *Advances in Neural Information Processing Systems*, pp.5050-5060, 2018.
- [11] J. Sun, T. Chen, G. Giannakis, and Z. Yang, "Communication-efficient distributed learning via lazily aggregated quantized gradients," in *Advances in Neural Information Processing Systems*, pp.3365-3375, 2019.
- [12] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "FedDane: A federated newton-type method," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pp.1227-1231, IEEE, 2019.
- [13] A. Khaled, K. Mishchenko, and P. Richtárik, "First analysis of local gd on heterogeneous data," *arXiv preprint arXiv:1909.04715*, 2019.
- [14] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On convergence of fedavg on non-IID data," in *International Conference on Learning Representations*, 2020.
- [15] H. B. McMahan, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol.14, no.1, 2021.
- [16] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *arXiv: Learning*, 2019.
- [17] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Gossip algorithms: このような場合、「Gossip Algorithm」（ゴシップアルゴリズム）は、「Design, Analysis and Applications」（設計、解析、応

用)、「INFOCOM 2005.24th Annual Joint Conference of the
IEEE Computer and Communications Societies.IEEE 論文集, vol.3,
pp.1653-1664, IEEE, 2005.

- [18] R.また, このような環境下でも, 「社会的責任」を果たすことができるよう, 「社会的責任」を果たすための仕組みづくりを進めている.
- [19] L.また, このような場合, 「無線センサーネットワークにおける時刻同期のための分散型コンセンサスプロトコル」, 2007 46th *IEEE conference on decision and control*, pp.2289-2294, IEEE, 2007.
- [20] T.このため, このような問題点を解決するために, 以下のような対策を講じる必要がある.
- [21] A.本論文では, このような問題点を解決するために必要な手法の一つである「分散型サブグラディエント法」について述べる.
- [22] A.また, このような場合, 「逐次処理」ではなく, 「分散処理」を行うことが望ましい.
- [23] D.Jakovetic', J. Xavier, and J. M. Moura, "Fast distributed gradient methods," *IEEE Transactions on Automatic Control*, vol. 59, no.5, pp.1131-1146, 2014.
- [24] I.Matei and J. S. Baras, "Performance evaluation of the consensus-based distributed subgradient method under random communication topologies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no.4, pp.754-771, 2011.
- [25] K.Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no.3, pp.1835-1854, 2016.
- [26] J.Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Transactions on Signal Processing*, vol.66, no.11, pp.2834-2848, 2018.
- [27] B.Sirb and X. Ye, "Consensus optimization with delayed and stochastic gradients on decentralized networks," in *Big Data (Big Data), 2016 IEEE International Conference on*, pp.76-85, IEEE, 2016.
- [28] また, このような場合, 「通信効率に優れた分散型・確率的最適化アルゴリズム」『数学プログラミング』180 巻, 1 号, 237-284 頁, 2020 年.
- [29] X.Lian, W. Zhang, C. Zhang, and J. Liu, "Asynchronous decentralized parallel stochastic gradient descent," in *Proceedings of the 35th International Conference on Machine Learning*, pp.3043-3052, 2018.
- [30] T.Sun, P. Yin, D. Li, C. Huang, L. Guan, and H. Jiang, "Non-ergodic convergence analysis of heavy-ball algorithms," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.33, pp. 5033-5040, 2019.
- [31] R.Xin and U. A. Khan, "Distributed heavy-ball:A generalization and acceleration of first-order methods with gradient tracking," *IEEE Transactions on Automatic Control*, 2019.
- [32] A.Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "Quantized decentralized consensus optimization," in *2018 IEEE Conference on Decision and Control (CDC)*, pp.5838-5843, IEEE, 2018.
- [33] Q.Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, *Federated learning*. モーガン&クレイプルー出版社, 2019.
- [34] H.Xing, O. Simeone, and S. Bi, "Decentralized federated learning via SGD over wireless D2D networks," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp.1-5, IEEE, 2020.
- [35] T.Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of the 1st Adaptive & Multitask Learning Workshop, Long Beach, California*, 2019.
- [36] S.Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no.4, pp.2341-2368, 2013.
- [37] S.このような状況下で, 「己を律し、己に克つ」ことが重要である. 4, pp.667-689, 2004.
- [38] B.T. Polyak, "Some methods of speeding up the convergence of iteration methods," *Ussr Computational Mathematics and Mathematical Physics*, vol. 4, no.5, pp.1-17, 1964.
- [39] M.Li, D. G. Andersen, A. J. Smola, and K. Yu, "Communication efficient distributed machine learning with the parameter server," in *Advances in Neural Information Processing Systems*, pp.19-27, 2014.
- [40] D.Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Advances in Neural Information Processing Systems*, pp.1709-1720, 2017.
- [41] S.Magnusson, H. Shokri-Ghadikolaei, and N. Li, "On maintaining linear convergence of distributed learning and optimization under limited communication," *IEEE Transactions on Signal Processing*, vol.68, pp. 6101-6116, 2020.
- [42] H.Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-~~prox~~ condition," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795-811, Springer, 2016.
- [43] S.J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *International conference on machine learning*, pp. 314-323, 2016.
- [44] D.J. Foster, A. Sekhari, and K. Sridharan, "Uniform convergence of gradients for non-convex learning and optimization," in *Advances in Neural Information Processing Systems*, pp. 8745-8756, 2018.
- [45] B.T. Polyak, "Gradient methods for minimizing functionals," *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, vol. 3, no.4, pp.643-653, 1963.
- [46] S.また, このような場合にも, 「実解析的部分集合の位相的性質」, *Coll. du CNRS, Les e'quations aux de'rive'es partielles*, vol. 117, pp. 87- 89, 1963.
- [47] S.Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: また, "Mathematical programming, vol.127, no.1, pp.3-30, 2011 "では, "Primal estimated subgradient solver for svm," と題する講演を行いました.
- [48] A.Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol.19, no.1, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol.19, no.4, pp.1574-1609, 2009.
- [49] A.サレム、Y. チャン、M. ハンバート、P. ベラン、M. フリッツ、および M.Backes, "MI-leaks:機械学習モデルにおけるモデルおよびデータに依存しないメンバーシップ推論攻撃と防御、" *In Annual Network and Distributed System Security Symposium (NDSS 2019)*, 2019.