cpSGD: 通信効率が良く、差分的にプライベートな分散 SGD

Naman AgarwalAnanda Theertha SureshFelix Yu

Google BrainGoogle ResearchGoogle Research

Princeton, NJ 08540New York, NYNew York, NY

namanagarwal@google.comtheertha@google.comfelixyu@google.com

サンジブ・クマール H.Brendan McMahan

Google リサーチ Google リサーチ ニューヨーク、NY シアトル、WA sanjivk@google.commcmahan@google.com

# アブストラクト

分散確率的勾配降下法は、分散学習における重要なサブルーチンである。

特に興味深いのは、クライアントがモバイル機器の場合で、通信効率とクライアントのプライバシーの 2 つが重要な問題となる。最近のいくつかの研究では、通信コストの削減やプライバシー保証の導入に焦点が当てられているが、提案された通信効率の良い方法のどれもがプライバシーを保持しておらず、既知のプライバシーメカニズムのどれもが通信効率が良いことは知られていない。このため、本研究では、通信効率と差分プライバシーの両方を実現するアルゴリズムを研究する。*d* 個の変数と *n* ⇡ *d* 個のクライアントに対して、提案手法は座標ごとにクライアントあたり O(loglog(*nd*)) ビットの通信を使用し、一定のプライバシーを保証する。

また、これまでの二項式メカニズムの解析結果を改善し、ガウス式メカニズムとほぼ同等の実用性を達成する一方で、必要な表現ビット数が少ないことを示し、独立した関心を集めることができました。

# 1 はじめに

1.1 背景

分散型確率的勾配降下法(SGD)は、現代の機械学習の基本的な構成要素である[25, 11, 9, 28, 1, 27, 5]。同期型分散学習の典型的なシナリオでは、各ラウンドにおいて、各クライアントがグローバルモデルのコピーを取得し、ローカルデータに基づいて更新します。この更新（通常は勾配の形で）はパラメータサーバに送られ，そこで平均化され，グローバルモデルの更新に使用される．また、中央サーバを介さずに、各クライアントがグローバルモデルを維持し、他のクライアントのすべてまたは一部にグラジエントをブロードキャストし、集約されたグラジエントでモデルを更新する方法もある。本稿では、特に中央集権的な設定を検討しているが、非中央集権的なケースについては、[36]およびその中の参考文献を参照されたい。しばしば，勾配を送信するための通信コストがボトルネックとなる[30, 23, 22]．この問題を解決するために，最近のいくつかの研究では，勾配の量子化とスパース化を介して分散学習アルゴリズムの通信コストを削減することに焦点を当てている[32, 17, 33, 20, 21, 4, 34]．これらのアルゴリズムは、分散学習における通信コスト、ひいては通信時間を改善することが示されている。これは，クライアントが高価なアップリンク通信コストを持つモバイルデバイスである連携学習の設定において，特に効果的である[26, 20]．

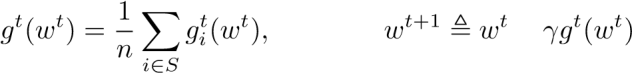
32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

クライアントベースの分散型機械学習では、通信が重要な問題となりますが、同様に重要な問題として、参加しているクライアントのプライバシーとその機密情報を保護することが挙げられます。機械学習アプリケーションに厳密なプライバシー保証を提供することは、最近活発に行われている分野です[6, 35, 31]。特に、微分積分法による勾配降下法は、[2]で研究されています。これらのメカニズムを分散環境に直接適用すると、通信コストの高いアルゴリズムになってしまう。本論文では、厳密なプライバシー保証を実現し、かつ通信効率の良いメカニズムを解析することに重点を置いている。

1.2 通信の 効率化

まず、同期分散 SGD を形式的に説明します。*F*(*w*) : *Rd* !R は、各 *fi* が *i* 番目のクライアントに存在する形式であるとします。例えば、*w* の重みは次のようにな  ります。

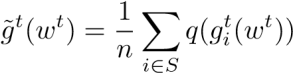
ニューラルネットワーク、*fi*(*w*)はクライアント *i* にあるデータに対するネットワークの損失で、*w*0 を初期値とする。ラウンド *t* において、サーバはすべてのクライアントに *wt* を送信し、*n*（バッチサイズ／ロットサイズ）個のクライアントのランダムなセットに、クライアントのサブセットであるローカル勾配推定値を送信サーバは次のように更新する。



を適当に選択します 。上記の SGD ステップの代わりに、モメンタム、Adagrad、

Adam などの他の最適化アル ゴリズムを使用することもできます。

上記のプロトコルでは、直観的には、log1/⌘）ビットの各ビットを使用します。例えば 1，中規模の *PennTreeBankn* では，クライアントは *d* 個のリアルデータを送信する必要があり，通常は O(*d* - language model [39]) を使用しますが，パラメータ数 *d* は1000 万を超え，したがって，総コストは⇠ 38MB (32 bit float を仮定) となり，携帯電話からサーバにラウンドごとに送信するには大きすぎます．通信効率の良いプロトコルの必要性から，通信コストを削減するための様々な量子化アルゴリズムが提案されている[33, 20, 21, 38, 37, 34, 5]．これらのプロトコルでは，クライアントは勾配を関数 *q* で量子化し，実際のローカル勾配の代わりに *q*(*git*(*wt*))の効率的な表現を送信する．サーバーは，その勾配を次のように計算します．

*,*

1⌘は、座標ごとの量子化精度です。*d*次元のベクトル*X*をユークリッド距離で一定の精度で表現

するためには、通常、各座標は⌘=*1/pd* の精度で量子化されます。

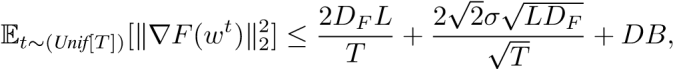
とし、以前のように *wt* を更新する。具体的には、[33]では、完全な（または浮動小数点の）

算術精度の要求を、平均して値あたり 1 ビットまたは数ビットに低減する量子化アルゴリズムを提案しています。その後、例えば[21]のような多くの研究がなされ、特に[5]では、ストキャスティック量子化とエリアス符号化[15]を用いて、凸関数に対する通信最適 SGD を得ることができることが示されています。すべてのラウンド *t* における期待される通信コストが *c* で制限される場合、修正勾配降下法の総通信コストは最大でも

*T* - *c.* (1)

これまでの論文はすべて、勾配圧縮の誤差を SGD の収束に関連づけています。まず、非凸関数の場合には、そのような結果の 1 つを完全性のために述べ、付録 A でそれを証明します。同様の(そしてより強い)結果は、[16]および[29]の結果を用いて(強)凸関数に対して得ることができます。

補論 1 ([16])。*F* を*L* 平滑とし、8*x krnF*(*x*)k2p2*DDF* とする。ここで( *pwLT*0 )が1o を満たすとすると、*Fthen after*(*w*0) *FT*(*roundw*⇤)



*DF* です。*q* を量子化スキームとし、, min *L* 1*,*

ここで、 *B* ＝max1*tT kEq*[*gt*(*wt*) *g*〜*t*(*wt*)]k とする。上の式の期待値は、グラデーションと量子化のランダム性に対するものです。

上記の結果は、非凸関数に対する分散型 SGD の収束性を、式(2)における最悪のケースの平均二乗誤差(MSE)と勾配平均推定値のバイアスに関連づけています。したがって、勾配推定における平均二乗誤差が小さいほど、収束性が高くなります。そこで、分散平均推定(DME)の問題に注目する。ここでの目的は、ベクトルの集合の平均を推定することである。

1.3 ディファレンシャル・プライバシー

上記の方式は、通信コストを削減する一方で、プライバシーを保証するものではありませんでした。我々は、差分プライバシー(DP)のレンズからプライバシーを研究する。ディファレンシャル・プライバシーの概念[13]は，機械学習タスクにおける有用なデータ分析を可能にする一方で，個人のプライバシーの強力な概念を提供する．詳細は[14]を参照されたい。非公式には、出力が差分プライバシーであるためには、推定されたモデルは、特定のクライアントのデータが考慮されたかどうかを区別できない必要がある。これを正式に定義するのがセクション 2 である。

クライアントベースの分散学習の文脈では、我々はクライアントから集約された勾配のプライバシーに興味がある。平均勾配の差分プライバシーは、DP が後処理によって保存されるので、結果として得られるモデルのプライバシーを意味する。標準的なアプローチは、サーバが平均化された勾配にノイズを加えることである（例えば、[14, 2]およびその中の参考文献を参照）。しかし、上記は、クライアントがサーバを信頼できるという制限的な仮定の下でしか機能しません。我々の目標は、クライアントが中央のアグリゲータを信頼する必要性を最小限にすることであり、それゆえ我々は以下のモデルを提案する。クライアントは，送信前に自分のグラデーション*git* にノイズのシェアを追加する．クライアントは送信前に自分のグラデーションに自分の分のノイズを加える。サーバーでグラデーションを集約すると、各クライアントで加えられたノイズの合計に等しいノイズを持つ推定値が得られる。

このアプローチは、いくつかのシナリオにおいて、サーバー制御によるノイズ追加よりも優れています。クライアントがサーバを信頼していない。サーバが信頼できない場合でも、上記の方式は、暗号化された安全な集約方式を介して実装することができます[7]。これにより、サーバが知る個々のユーザに関する情報は、合計から推測されるものだけになります。したがって、集約の差分プライバシーにより、パラメータサーバが個々のユーザの情報を学習しないことが保証されました。これにより、クライアントがサーバを完全には信頼していなくても、プロトコルへの参加を促すことができます。安全な集約方式は通信コストを増加させますが（例えば、[7]では *k* に対して log2(*k* - = 1024*n*)を追加していますが、4 ビットの量子化プロトコルであれば通信コストのレベルを下げることができます）、我々が提案した通信上の利点はまだ有効であることに注意してください。例えば、 *n* が32 ビット表現に比べて 67％も減少した場合。

サーバは怠慢だが悪意はない：サーバはノイズを追加することを「忘れる」かもしれないが、悪意はなく、個々のユーザの特性を学習することには興味がない。ただし、サーバーが学習したモデルを公開する場合は、差分非公開にする必要があります。

14, 2]の結果を拡張する自然な方法は，個々のユーザが送信前に自分のグラデーションにガウスノイズを加えることです．ガウシアンの和はガウシアンそのものなので、プライバシーに関する差分結果が得られます。しかし、送信される値は実数となり、勾配圧縮の利点は失われてしまいます。さらに、安全な集約プロトコル[7]は離散的な入力を必要とする。これらの問題を解決するために、我々は、クライアントが適切にパラメータ化された二項分布から引き出されたノイズを追加することを提案します。これを「二項メカニズム」と呼びます。二項分布の確率変数は離散的であるため、効率的に送信することができます。さらに、二項分布の選択は、二項の和も二項分布となるため、分散型の設定では便利です。

*Z*1 ⇠ Bin(*N*1*,p),Z*2 ⇠ Bin(*N*2*,p*) then *Z*1 + *Z*2 ⇠ Bin(*N*1 + *N*2*,p*)となります。

そのため、集計後のトータルノイズを簡単に分析することができ、分散型の設定には便利2です。二項のメカニズムは、離散的な出力を持つ他のアプリケーションでも、独立した興味を持つことができます。さらに、ガウス型とは異なり、浮動小数点表現の問題を回避することができます。

1. 4 今回の結果のまとめ

二項式メカニズムまず、離散値データを解放する一般的なメカニズムとして、二項メカニズムを研究します。これまでの二項機構（ノイズ Bin(*N,p*)を加える場合）の解析は[12] によるもので、*p* = 1/2 の 1 次元の場合を解析し、（", ）の差分プライバシーを実現するためには、*N* は64log(2*/ )/"*2 である必要があることを示していました。

* ガウス分布とは異なり、二項分布は回転不変ではないため、解析はより多次元的になります。我々は、1 次元の二項式メカニズムの分析を *d* 次元に拡張する。この解析で利用される重要な事実は、二項分布が平均の周りで局所的に回転不変であることです。
* これは、二項法とガウス法が同じように機能することを意味しています。私たちは先

の結果を改良し、*N* 8log(2*/ )/"*2 で十分であることを示しました。これは一定の改善で

2 もう一つの選択肢はポアソン分布です。ポアソン分布とは異なり，二項分布は有界の支持率を持ち，通信の複雑さも常に有界であるため，容易に分析することができます．

はありますが、差分プライバシーを実用化する上で非常に重要であることに注意して

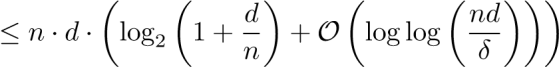
ください。

Differentially-private distributed mean estimation (DME):ガウス型メカニズムを直接適用するには、*n* - *d* 個のリアルが必要であり、したがって、*n* - *d* - log(*nd*) ビットの通信が必要とな

ります。これは、33]や二項式のメカニズムと実際には禁止されている可能性があります。

我々はまず、量子化の直接適用を提案する[

本論文では、そのプライバシー／エラー保証と通信コストを明らかにした。さらに、この方式をランダムローテーションと組み合わせることで、通信量を大幅に改善できることを示します。特に、" = *O*(1) "の場合、ガウス型メカニズムと同じプライバシーとエラーのトレードオフを達成するアルゴリズムを提供します。

 ビットを使用しています。

は、分散 SGD のラウンドごとにしたがって、*d* ⇡ *n* のとき、ビット数は *n* - *d* - log(log(*nd*)/ ) となります。

本論文の残りの部分は以下のように構成されている。第 2 節では、差分プライバシーの概念を復習し、二項式メカニズムに関する結果を述べる。また、SGD の収束がラウンドごとの勾配推定計算の誤差に還元できるという事実に触発され、セクション 3 で DME の問題を正式に記述し、セクション 4 で我々の結果を述べる。

セクション 4.2 では、DME のコンテキストで量子化と組み合わせた二項機構の実装を提供し、分析します。主なアイデアは、各クライアントが、サーバーに送信する前に、適切にパラメータ化された二項分布から引き出されたノイズを各量子化値に加えることです。サーバはさらに、ノイズによってもたらされたバイアスを差し引き、不偏の平均推定値を得る。さらに、セクション 4.3 では、MSE を削減する[33]で提案された回転手順が、差分プライバシーによる追加の誤差を削減するのに役立つことを示しています。

# 2差分のプライバシー

2.1 Notation

まず、差分プライバシーの概念を定義します。形式的には、隣接するデータセットの概

念 MND⇢D⇥D とクエリ関数(", )が与えられたデータセット D のセットが与えられた場合、 differentially private if for *anyf* :D !X に対して、あるメカニズム

:X !O を用いて、クエリの答えを公開することを、測定可能な部分集合 *S* ✓ O と 2 つの隣接するデータセット(D1*,*D2)2 ND と定義します。

Pr(M(*f*(D1))2 *S*) e*"* Pr(M(*f*(D2))2 *S) + .* (3)

特に断りのない限り、本稿の残りの部分では、出力空間を想定し、メカニズム MX の誤差を測る指標として平均二乗誤差を考慮する。形式的には、*,*O ✓ *Rd*.我々は

*.* 多くのメカニズムにおける差分プライバシーを特徴づける重要な量は、クエリ *f* の感度です。D !*Rd* の感度である。形式的には次のように定義される。

*.* (4)

(", )差分プライバシーを実現する正統的なメカニズムは、ガウス・メカニズム 2 *Mg*

［14］です。Mof the Gaussian mechanism. *g*(*f*(*D*)) , *f*(*D*) + *Z,* ここで *Z* ⇠ N(0*, Id*)である。こ

こで、よく知られているプライバシー保証を述べます。

Lemma 12p ([14])。任意の *, `*2 感度境界 3 と誤差が 2 で境界され、*d*2-p22*.*log1*.*25*/ , Mg* は

( 2log1*.*25/ , ) 差動的秘密である。

3すべてのログは、特に記載がない限り、ベース *e* に対するものです。

2.2 二項対立メカニズム

ここで、出力空間の量子化スケール *s*=*1/j* をある *j* 2 N で与えた場合の二項機構を定義し

ます。 問い合わせ N の X*,p* 2 (0*,f*1)is,*Zd* である。

二項式メカニズムは、3 つの量 *N,p,s* でパラメータ化されます。

· (5)

ここで、各座標について、*i*, *Zi* ⇠ Bin(*N,p*)が独立である場合について、[12]で紹介されてい

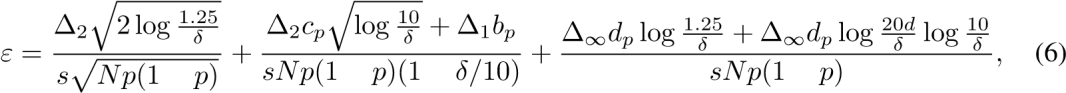
ます。1 次元 2 項機構=1/2。二項機構は回転不変ではないので、この解析は重要である。

平均付近の局所的な回転不変構造を注意深く利用することで、以下のことを示す。

Theorem 1.任意の *,* パラメータ *N,p,s* と感度限界1*,* 2*,* 1 のような場合

*Np*(1 *p*) max(23log(10*d/ ),*2 *1/s*)*,*

二項式メカニズムは、（"、）のために異なるプライベートなものです。

 ここで、*bp*、*cp*、*dp* はそれぞれ*(16)*、*(11)*、*(15)*で定義されており、*p*=1/2 の場合、*bp*=1/3、 *cp*=5/2、*dp*=2/3 となります。機構の誤差は，*d* - *s*2 - *Np*(1 *p*)である．

その証明を付録 B に示す。二項式メカニズムの設計と保証について、いくつか指摘しておく。なお、二項機構のプライバシー保証は、3 つの感度パラメータ depends on 2

2*,* 1,1 に依存するのに対し，Gaussian メカニズムでは， . 1 および 1 と

1 の項は、離散化による複雑さの追加と見ることができます。

次に、式(6)において *s*=1（ノイズにスケールを与えない）とすると、1 と 2 の項がノイズの分散に対して異なるスケールを持つことが容易に理解できます。このことが、メカニズムに付随する量子化スケール *s* を使用する動機となっています。実際には、二項ノイズによって提供される整数の解像度が問題に対して大きすぎて、保証が悪くなる可能性があります。この設定では、量子化パラメータ *s* は、ノイズを正しく正規化するのに役立ちます。また、確率変数の分散がガウス型になっていれば問題ありません。形式的には、*Z* を固定し、*Np*(1 *p*)を増加させ、 *s(*1makes the Binomial mechanism *closerp*)と *s pd*

)を減少させると、Cauchy-Schwartz 不等式を用いて、「保証(6)」は、/(*c*

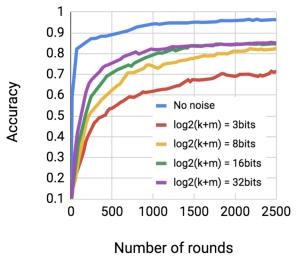
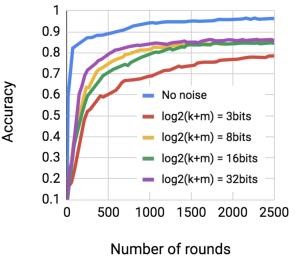
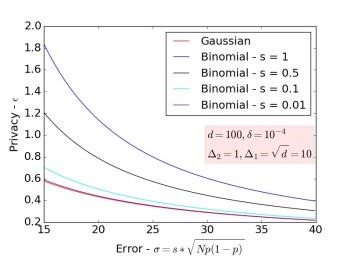
" = (2/ )p2log1*.*25/ (1 + O (*1/c*))となります。

二項分布の分散は *Np*(1 *p*)であり、「」の先頭の項は、ガウス型メカニズムの「」の項と正確に一致します。さらに、*s* が *o*(*1/pd* )であれば、このメカニズムはガウス型メカニズムに近づくことになります。この結果は、1 次元 2 項分布のガウス分布への収束に関する Berry-Esseen 型の中心極限定理と一致する。図 1 は、ガウス型と二項型の誤差をプロットしたものです。スケールを小さくすると，二項式メカニズムの誤差対プライバシーのトレードオフは，ガウス式メカニズムのそれに近づくことがわかる．

最後に、固定の通信予算で分散を最大化するためには、一般的には *p* = 1/2 が最適な選択となりますが、必要な分散が非常に小さく、*N* と *p* = 1/2 の整数選択では達成できないようなコーナーケースがあるかもしれないことに注意してください。一般的な *p* を用いた我々の結果は、このようなコーナーケースにも対応しています。

# 3分布平均推定（DME

補論 1 では、SGD の収束率を、各ステップで勾配を近似する際の MSE と関連づけました。式(1)は、SGD の通信コストを、勾配手段を推定する際の通信コストに関連づけています。高度な合成定理 (Thm. 3.5 [19]) またはモーメント・アカウンティング [2] を使用して、SGD のプライバシー保証と各インスタンス *t* における勾配平均推定のプライバシー保証とを関連付けることができます。また、SGD ではクライアントをサンプリングすることが多いことに注意してください。



(a) " *=* 4.0 (b) " *=* 2.0

図 1：ガウス型と二項型の誤差とプライバシーの比較 図 2：無限の MNIST データセットにおける回転を伴う cpSGD。*k* は量子化レベルの数、*m* は二項ノイズの量子化と差分プライバシーなしのパラメータスケール（*p* = 0*.*5, *s* = 1）。ベースラインは= 10 9 です。したがって、[33]と同様に、本稿の残りの部分では、DME の MSE とプライバシーの保証にのみ焦点を当てる。同期分散 GD の結果は、Corollary 1（収束）、Advanced composition theorem（プライバシー）、Eq.1（コミュニケーション）から導かれる。

形式的には，DME の問題は，*n* 個のベクトル *X* , {*X*1 *...Xn*} が与えられ，*Xi* 2 *Rd* がクライアント *i* 上にあるとき，中央のサーバで平均値を計算したいと定義されます．各ラウンド *t* での勾配降下のために，*Xi* は*git* に設定される．DME は，分散 PCA/クラスタリングを含む多くの分散学習アルゴリズムの基本的な構成要素である[24]．

このような仮定をする理由は、プライバシー保証 *Xi* を定義し、分析することができるようにするためであり、すなわち、kXik *D.*は、各クライアントで勾配クリッピングを採用することによって、実際に強制されることが多い。この仮定は、勾配降下法や differentially private gradient descent に関する過去の研究（例えば[2]）に登場することに注意してください。我々の結果は、統計的な仮定なしにすべての勾配に対しても成立するため、SGD に対して望ましい収束結果とプライバシー結果を得ることができます。

3.1 通信プロトコル

我々が提案する通信アルゴリズムは、同時かつ独立したものであり、すなわち、クライアントは独立して同時にサーバにデータを送信するものである。また、プライベート・ランダムネスとパブリック・ランダムネスの両方を使用することができます。プライベート・ランダムネスとは，各クライアントが個別に生成した乱数値を指し，パブリック・ランダムネスとは，すべての当事者間で共有される乱数値の列を指す3．クライアント *Xi* に *Xi* 2 *Rd* が存在する *n* 個のベクトル *X* , {*X*1 *...Xn*}が与えられます（*q*(*Xi*)*i* とします）。任意の独立したセントラルサーバ通信プロトコルにおいて、各クライアントは、以下の関数を送信します。

は、*q*(*X*1*),q*(*X*2*),...,q*(*Xn*)のある関数によって平均値を推定する。⇡をそのようなプロトコルとし、*Ci(*⇡*,Xi*)をプロトコル⇡中に *i* 番目のクライアントが送信するビットの期待値とする。 ⇡とし、本稿では期待値はプロトコルのランダム性に対するものとします。

プロトコル *Ci(*⇡*,Xi*)を持つすべてのクライアントを、クライアント *i* が送信したビット数とする。推定平均値をˆ¯とする。

 *X*

プロトコル⇡の場合、推定値の MSE は E*(*⇡*,X*1*n*)=*EhkX*ˆ¯ *X*¯k22i となる。の境界があることに注意してください。

は、式(2)の勾配推定値の境界に変換され、Corollary 1 によって収束が保証されることになります。

3. 2 差分のプライバシー

DME のプライバシー結果を述べるために、データセットとネイバーの概念を以下のように定義する。A

3 公開されたランダム性は、サーバーがランダムシードを通信することでエミュレートできます。

データセットとは、1 つのベクトルに違いがあるベクトルの集まりである。なお、DME の隣人の概念は、分散型グラデーションの文脈では、*X* = {*X*1*,...Xn*}となります。隣り合うデータセットの概念は、典型的には、*X,X*⌦*i* は、一人のユーザーの情報だけが異なるものに対応する場合、つまり、次のような場合に隣り合う。

降下法は、2 つのデータセット *F* = *f*1*,f*2*,...fn* と *F*0 = *f*10*,f*20*,...fn*0 が、1 つの関数 *fi* で異なっていれば隣り合っていると訳され、個々のクライアントのデータのプライバシーを保証することに相当します。境界 kXik2 *D* は、勾配クリッピングによって確保されていると仮定することに変換されます。

# 4 分布平均推定（DME）の結果

このセクションでは、DME のコンテキストにおけるアルゴリズム、関連する MSE、およびプライバシー保証について説明します。まず、各クライアントのベクトルにガウスノイズを加えることでガウスメカニズムを実装した場合の結果を述べることで、ベースラインを確立します。

4.1 ガウス型プロトコル

ガウス型メカニズムでは、各クライアントは、ベクトル *Yi* = *Xi* + *Zi* を送信します。

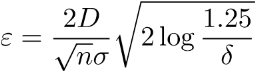
 N(0*,* 2*Id*)である。サーバーは平均値を *X*ˆ¯=*1/n* -で推定する。この

プロトコルを⇡*g* と呼ぶ。 N(0*,* 2*Id/n*)と分布している場合、上記のメカニズムは、出力にガウスメカニズムを適用して分散を 2*/n*.*Xi* のいずれかを変更すると、*X*¯のノルムは最大で

*2D/n* だけ変化するので、次の定理は定理 1 から直接導かれます。

Theorem 2.ガウス型メカニズムの下では、平均推定値は不偏であり、通信コストは*n* - *d* 個のリアルである。さらに、任意の*and* に対して、それは（"、）差分的にプライベートである。

となって



実数を量子化してO(*logdn*/")ビットにすることで，プライバシーに大きな影響を与えない 4可能性があることを指摘します．しかし、これは漸近的なものであり、固定通信予算が少なく、*d* が数百万のオーダーである場合、実際には法外なものになる可能性があります [20]。通信コストを削減する自然な方法は量子化であり，各クライアントは送信する前に Yis を量子化します．しかし、ガウシアンメカニズムの量子化によってプライバシー保証がどのように低下するかは、特にアグリゲーションの下では解析が困難です。その代わりに、我々は次に述べる 2 項メカニズムを使用することを提案します。

4.2 トカスティックな *k* レベル量子化＋二項メカニズム

ここで、[33]で提案された *k* ビット確率的量子化⇡*sk* に基づくメカニズム⇡*sk*(Bin(*m,p*))を二項メカニズムと組み合わせて定義する。これは、3 つの量 *k,m,p* でパラメータ化されます。

4続いて、 すべての値を *1/*poly(*n,d,1/",*log1/)の精度で量子化することで、プライバシーの損失を最小限に抑えることができることを説明しています。実際には、浮動小数点表現による 32 ビットの量子化を用いて実装されることが多い。

まず、サーバーは、すべての *i,j* に対して Xmax *Xi*(*j*) Xmax となるように、Xmax をすべてのクライアントに送信します。そして、クライアントは自分のベクトルの各座標を[ Xmax*,*Xmax]の範囲にクリップします。範囲[0*,k*]内のすべての整数 *r* に対して、*B*(*r*)はビン(各 *r* に対して

1 つ)を表すとすると、すなわち、次のようになります。

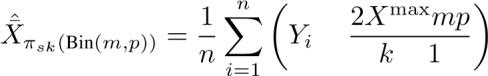
*,* (7) このアルゴリズムは、各座標を確率的にビンの 1 つに量子化し、スケーリングされた二項ノイズを加えます。形式的には、クライアント *i* はすべての *j* に対して以下の量を計算します。

*B*(*r* + 1) w.p. 2Xmax

( *Ui*（*j*）＝*Yi*（*j*）＝*Ui*（*j*）＋。 *- Ti*(*j*)となります。(8) *B*(*r*) でなければならない。 *k* 1



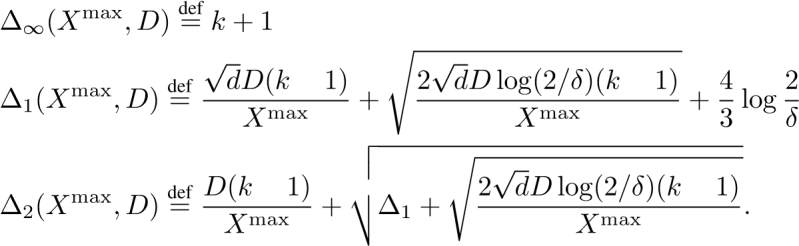
ここで、*r* は、*Xi*(*j*)2 [*B*(*r),B*(*r* + 1)] and *Ti*(*j*)⇠ Bin(*m,p*)となるようなものです。クライアントは、*Yi* をサーバーに送信します。サーバは *X*¯を次のように推定します。

 *.* (9)

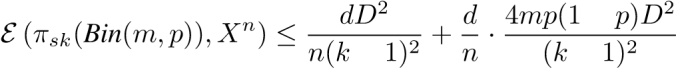
8*j*, *Xi*(*j*) 2 [ Xmax*,*Xmax] とすれば、平均値の不偏推定値となる

。

形式的な保証を述べる前に、量子化プロトコルの感度を表す以下の量を、適切なノルムで定義する必要があります。

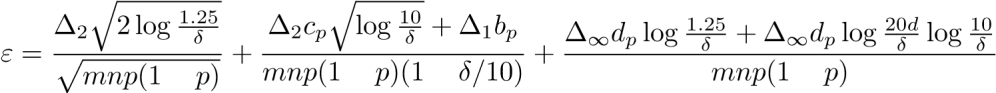
u (10)

表記を簡潔にするために、パラメータ *k,* を省略しています。この Xmax の値を用いて、定理 1 を利用して、⇡*sk*(Bin(*m,p*))の MSE、感度、通信量を以下のように特徴付けます。Theorem 3.Xmax = *D* であれば、平均推定値は不偏であり

*,* さらに、もし

*mnp*(1 *p*) max(23log(10*d/ ),*21(*D,Xmax*))となります。

とすると、任意の *, X*ˆ*¯*⇡*sk*(*Bin*(*m,p*)) は *(",*2 ) 差動私有であり、ここで *" (*定理*1* で与えられる*)* は

*,*

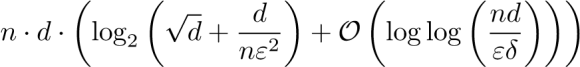
で、感度パラメータ{ 1(*Xmax,D),* 2(*Xmax,D), (*1(*Xmax,D*)}で定義されています。*)*

さらに。

C*(*⇡*sk*(*Bin*(*m,p)),Xn*) = *n* - (dlog2(*k* + *m*) + O〜(1)) となります。5

その証明を付録 D に示す。上の定理における「」の式の第 1 項は、ガウス型メカニズムの場合と同じ保証を回復する（定理 2）。さら~~に~~、*k >> pd* のとき、後続の項は無視できることがわかる。形式的には、ガウス型メカニズムと同じ保証を実現するための「1」の通信コストをまとめた次の公理が導かれる。

冠詞 2⇡*sk*(*Bin*(*m,p*))の実装が存在します。この実装は、全精度ガウス機構と同じプライバシーとエラーを、以下の総通信量で達成します。

ビットを使用しています。

上記のアルゴリズムの通信コストは、クライアントごとに座標ごとに⌦(*logd*)ビットであり、これは法外なものとなります。次のセクションでは、これらの境界が回転によってさらに改善できることを示します。

4.3 無作為化された回転によるエラーの減少

補論 2 にあるように、⇡*sk*(Bin(*m,p*))がガウスメカニズムと同じエラーとプライバシーを持つためには、保証される通信コストの最良の境界は、*n* がどれだけ大きくても、1 座標あたり⌦(log(*d*))ビットである。これは、誤差が *O*(*d*(Xmax)2*/n*)に比例するためであることが証明されました。したがって、Xmax が小さいと MSE は小さくなり、例えば *Xi* が単位球面上で一様な場合、Xmax は O ⇣p(*logd)/d*⌘(whp)[10]となります。[33]は，量子化の前にベクトルをランダムに回転させても同じ効果が得られることを示しました．ここでは，ランダムな回転により，誤差の先行項が減少するとともに，プライバシー保証が向上することを示します．

パブリック・ランダムを用いて、すべてのクライアントとセントラル・サーバーは、ランダムな直交行列を生成します。

*R* 2 *Rd*⇥*d* を、ある既知の分布にしたがって算出する。入力 *X*1 *...Xn* を取る DME のプロトコル ⇡が与えられた場合、各クライアント *i* がまず *Xi*0 = *RXi* を計算し、その上でプロトコルを実行するプロトコルを Rot*(*⇡*,R*)と定義する。次にサーバは、プロトコル⇡を用いて回転空間での平均推定値 *X*ˆ¯0 を求め、*R* 1 を乗じて元の基底での座標、すなわち *X*ˆ¯ = *R*

1*X*ˆ¯0 を求めます。

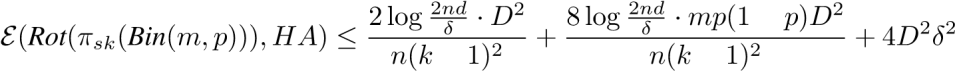
実際には *d* は巨大になることがあるため，高速な行列ベクトル積を可能にする直交行列が必要です．ブロック対角行列のような高速な乗算をサポートしない行列では，

．33]と同様に，我々は特別なタイプの直交±1（確率 0*.*5）を用いることを提案し，*H* は Walsh-Hadamard 行列である[18]．回転とその逆回転の両方を適用すると，

O(*dlogd*)時間と O(1)空間を要する（インプレース・アルゴリズムを使用）．次の定理は、Rot*(*⇡*sk*(Bin(*m,p)),HA*)の MSE とプライバシーの保証を与える。

5O〜は、多対数の因子を表すのに使われます。

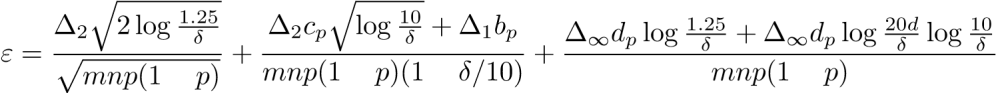
Theorem 4 (Appendix E).任意の場合、Xmax = *2*Dqlog(2*dnd*/ ) とすると

*.*

平均推定値のバイアスは、*2D* で拘束されます。さらに、もし

*mnp*(1 *p*) max(23log(10*d/ ),*21(*D,Xmax*))となります。

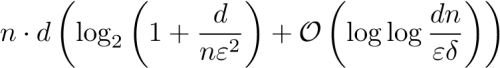
とすると、*X*ˆ¯(*Rot(*⇡*sk*(*Bin*(*m,p*))))) *is (",*3 ) *differentially private)*であり、ここで「（定理*1* で与えられる）」は

*,* を感度パラメータ{1(*Xmax,D)*、2(*Xmax,D)*、1(*Xmax,D*)}で設定します（式（*10*））。さらに

C(*Rot(*⇡*sk*(*Bin*(*m,p))),Xn*) = *n* - (dlog2(*k* + *m*) + O〜(1)) となります。

ここで、次の補論は、Rot*(*⇡*sk*(Bin(*m,p)),HA*)の通信コストを、補論 2 と同様に、*"* 1 *"*の場合に制限する。

推論 3*Rot(*⇡*sk*(*Bin*(*m,p)),HA*)の実装が存在し、全精度ガウス機構と同じ誤差とプライバシーを全通信量で実現しています。

ビットを使用しています。

なお、*k* はもはや⌦(*pd* )に設定する必要はなく、したがって、*d* = *o*(*n "*2")であれば、 Rot*(*⇡*sk*(Bin(*m,p)),HA*)は、ガウス型メカニズムと同じプライバシーと効用を持ちながら、わずか O(ndloglog(*nd*/"))の通信コストで済む。

# 5ディスカッション

25M 個のデータポイントと 25M 個のクライアントを持つ無限の MNIST データセット[8]を用いて，3 層モデル（60 個の隠れノード，ReLU 活性化）を学習した．各ステップにおいて，1 万人のクライアントがサーバにデータを送信する．この設定は，何億人ものユーザがいる連合学習の実世界の設定に近い．結果は図 2 のとおりです。通信コストとプライバシーのパラメータ「」に応じて，モデルが異なるレベルの精度を達成していることに注目してください．正確に 1 つのエポックでモデルをトレーニングしたので，各サンプルはトレーニングで最大 1 回使用されたことに注意してください．この設定では，バッチごとの「」と全体の「」は同じである．

いくつかの興味深い将来の方向性があります。理論的な面では、我々の二項式メカニズムの分析がタイトであるかどうかは明らかではありません。さらに、二項式メカニズムに対して、モーメント法を用いて、より優れたプライバシー保護を行うことができれば、興味深い。実用面では、ニューラルネットワークのトポロジー、過剰なパラメータ化、最適化アルゴリズムが、私的に学習したモデルの精度に与える影響を調査する予定です。

# リファレンス

1. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow:Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.
2. Martín Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang.Deep Learning with differential privacy.In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308-318.ACM, 2016」を参照してください。
3. Nir Ailon and Bernard Chazelle.近似最近傍法と高速 JohnsonLindenstrauss 変換．In *STOC*,

2006.

1. Dan Alistarh, Demjan Grubic, Jerry Liu, Ryota Tomioka, and Milan Vojnovic.通信効率の良い確率的勾配降下法，ニューラルネットワークへの応用．2017.
2. Dan Alistarh, Jerry Li, Ryota Tomioka, and Milan Vojnovic.QSGD: Randomized quantization for communication-optimal stochastic gradient descent. *arXiv:1610.02132*, 2016.
3. Raef Bassily, Adam Smith, and Abhradeep Thakurta.プライベートな経験的リスクの最小

化。Efficient algorithms and tight error bounds.In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 464-473.IEEE, 2014.

1. Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth.Practical secure aggregation for privacy-preserving machine learning. ページ 1175-1191, 2017.
2. Leon Bottou.無限のニストデータセット。
3. Adam Coates, Brody Huval, Tao Wang, David Wu, Bryan Catanzaro, Ng Andrew.COT の HPC システムによる深層学習。国際機械学習会議（*International Conference on Machine Learning*）では、1337-1345 ページ、2013 年に発表されました。
4. Sanjoy Dasgupta and Anupam Gupta.johnson and lindenstrauss の定理の初歩的な証明。ランダム構造とアルゴリズム*, 22*(1):60-65, 2003.
5. Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew

Senior, Paul Tucker, Ke Yang, Quoc V Le, et al. 大規模分散ディープネットワーク.In *Advances in neural information processing systems*, pages 1223-1231, 2012.

1. Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor.私たちのデータ、私たち自身。分散型ノイズ生成によるプライバシー。In *Eurocrypt*, volume

4004, pages 486-503.Springer, 2006.

1. Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith.Cynthia Dwork, Frank McSherry, Kobbi Nissim, Adam Smith, Private Data Analysis におけるノイズと感度の較

正.*TCC*, volume 3876, pages 265-284.Springer, 2006.

1. Cynthia Dwork と Aaron Roth.The algorithmic foundations of differential privacy.*Found.Trends Theor.Comput.Sci*. , 9(3&#8211;4):211-407, August 2014.
2. ピーター・イライアス普遍コードワードセットと整数の表現.*IEEE transactions on information theory*, 21(2):194-203, 1975.
3. Saeed Ghadimi and Guanghui Lan.非凸の確率的計画法に対する確率的一次・零次法。

*SIAM Journal on Optimization*, 23(4):2341-2368, 2013.

1. Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan.限られた数値精度での深層学習。In *Proceedings of the 32nd International Conference on Machine*

*Learning (ICML-15*), pages 1737-1746, 2015.

1. キャシー・J・ホラダムハダマード行列とその応用.プリンストン大学出版局, 2012. [19] Peter Kairouz, Sewoong Oh, and Pramod Viswanath.ディファレンシャル・プライバシーの構成定理。*IEEE Transactions on Information Theory*, 63(6):4037-4049, 2017.
2. Jakub Konecnˇ y, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh,` and Dave Bacon.Federated learning:Strategies for improvement communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.
3. Jakub Konecnˇ y and Peter Richtárik.ランダム化分布平均推定。Accuracy vs`

Communication. *arXiv preprint arXiv:1611.07555*, 2016.

1. Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Amr Ahmed, Vanja Josifovski,

James Long, Eugene J Shekita, and Bor-Yiing Su.パラメータサーバによる分散型機械学習のスケーリング。In *OSDI*, volume 1, page 3, 2014.

1. Mu Li, David G Andersen, Alexander J Smola, and Kai Yu.parameter server を用いた通信効率の良い分散機械学習。*Advances in Neural Information Processing Systems*, pages 19-27,

2014」に掲載されています。

1. Stuart Lloyd.PCM における最小二乗法による量子化.*IEEE Transactions on Information*

*Theory*, 28(2):129-137, 1982.

1. Ryan McDonald, Keith Hall, and Gideon Mann.構造化パーセプトロンの分散学習戦略.*HLT*,

2010.

1. H.Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 分散されたデータからのディープネットワークの通信効率の良い学習。In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS*), 2016.
2. H.Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Aguera y Arcas.Federated learning of deep networks using model averaging. *arXiv:1602.05629*, 2016.
3. Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur.Parallel training of deep neural networks with natural gradient and parameter averaging. *arXiv preprint*, 2014.
4. Alexander Rakhlin, Ohad Shamir, Karthik Sridharan, et al. Making gradient descent optimum for strongly convex stochastic optimization.*ICML* にて。Citeseer, 2012.
5. ベンジャミン・レヒト、クリストファー・レ、スティーブン・ライト、フェン・ニ

ュウ。Hogwild:Hogwild: A lock-free approach to parallelizing stochastic gradient descent.Hogwild: A lock-free approach to parallelizing stochastic gradient descent.In *Advances in neural information processing systems*, pages 693-701, 2011.

1. Anand D Sarwate and Kamalika Chaudhuri.差分プライバシーによる信号処理と機械学習。

Algorithms and challenges for continuous data.*IEEE signal processing magazine*, 30(5):86-94, 2013.

1. Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu.1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns.In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
2. Ananda Theertha Suresh, X Yu Felix, Sanjiv Kumar, and H Brendan McMahan.限られた通信での分散平均推定。In *International Conference on Machine Learning*, pages 3329-3337, 2017.
3. Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li.Terngrad:Ternary gradients to reduce communication in distributed deep learning. *arXiv preprint arXiv:1705.07878*, 2017.
4. Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton.Bolton differential privacy for scalable stochastic gradient descent-based analytics.In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1307-1322.ACM, 2017.
5. Kun Yuan, Qing Ling, and Wotao Yin.On the convergence of decentralized gradient descent.*SIAM Journal on Optimization*, 26(3):1835-1854, 2016.
6. Huizi Mao Yu Wang Bill Dally Yujun Lin, Song Han.深い勾配の圧縮。分散学習のための通信帯域の削減。学習表現に関する国際会議、2018。
7. Takuya Akiba Yusuke Tsuzuku, Hiroto Imachi.Variance-based gradient compression for efficient distributed deep learning, 2018.
8. Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals.リカレントニューラルネットワークの正則化. *arXiv preprint arXiv:1409.2329*, 2014.