情報処理学会論文誌　数理モデル化と応用　Vol.8　No.2　31–43 (July 2015)

差分プライベート弱学習器の統合

南 賢太郎[1,a)]　荒井 ひろみ[2,b)]　佐藤 一誠[2,c)]　中川 裕志[2,d)]

受付日 2015年2月3日，再受付日 2015年3月19日，
採録日 2015年4月17日

概要：データが複数の組織にわたり分散して存在しているとき，それらを互いに共有することで，各組織におけるデータ解析の精度向上が期待できる．しかし，保護すべき個人情報がデータに含まれている場合には，異なる組織間での情報の交換は，プライバシ保護を考慮したうえで行われなければならない．本研究では，差分プライバシをみたす弱学習器を互いに交換し，それらを統合する枠組みを提案する．これによって，複数の組織に分散したデータからの学習を，個人情報を保護しつつ効率的に行うことができる．また，特に学習タスクが二値分類である場合について計算機実験を行い，提案手法の性能を評価する．

キーワード：差分プライバシ，統計的学習理論，学習器統合

Aggregating Differentially Private Weak Learners

Kentaro Minami[1,a)]　Hiromi Arai[2,b)]　Issei Sato[2,c)]　Hiroshi Nakagawa[2,d)]

Received: February 3, 2015, Revised: March 19, 2015,
Accepted: April 17, 2015

Abstract: When the dataset is distributed over a number of organizations, one can expect the improvement of data analysis by sharing the dataset each other. However, if the dataset consists of personal information, data sharing procedures must be performed under privacy-preserving constraints. Recently, differentially private algorithms for some statistical learning problems, such as empirical risk minimization, have been considered by several authors. In this work, we introduce a general framework for exponential weighting aggregation (EWA) of differentially private weak learners. This framework allows us to learn effectively from distributed dataset without leakage of personal information. Especially in the case of the binary classification problem, we evaluate the effectiveness of our approach on synthetic and real dataset.

Keywords: differential privacy, statistical learning theory, weak learner aggregation

1. はじめに

近年，医療データ，移動履歴，購買履歴など，個人情報を含むデータ貯蓄量の増大にともない，それらの利活用に対する関心が高まっている．たとえば，データから得られる要約統計量や，分類や回帰問題など種々の機械学習手法によって得られる知見の利用は重要である．一方，そのような個人情報を含むデータを用いた解析結果を外部に公表する場合には，個人情報漏洩のリスクが発生する．そこで，プライバシを保護しつつデータ解析結果を公開する手法が数多く提案されている．

個人情報を含むデータは複数の組織にわたり分散して存在している場合も多い．その場合，それらを統合してデータ解析を行うことで，1つの組織内では得ることのできなかった知見が得られることが期待される．特に医療データを例にとると，ある特定の症例に関して1つの組織（病院）内で取得可能なデータ数が少ないという状況が考えられ，精度の良いデータ解析のためには異なる組織のデータを統合して解析することの意義が大きい．しかし，組織間で個人データそのもの，あるいは個人データをもとにして得られた解析結果を直接共有することは，個人情報の漏洩につ

1 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo, Tokyo 113–0033, Japan
2 東京大学情報基盤センター
Information Technology Center, The University of Tokyo, Bunkyo, Tokyo 113–8658, Japan
a) kentaro_minami@mist.i.u-tokyo.ac.jp
b) arai@dl.itc.u-tokyo.ac.jp
c) sato@r.dl.itc.u-tokyo.ac.jp
d) nakagawa@dl.itc.u-tokyo.ac.jp

表 1　本研究の位置づけ
Table 1　Our contribution.

		弱学習器のタイプ	
		プライバシ制約なし	プライバシ制約あり
統合手法	ERM	罰則なし　（非最適 [2]）	ロジスティック回帰 [3]
		罰則あり　BIC 型 [4] LASSO [5] Dantzig selector [6]	
	EWA	Averaging Expert [7] データ分割 + Mixing [8] Mirror Averaging [9]	**本研究（4 章）**

ながるため望ましくない．したがって，異なる組織間の情報交換は，何らかのプライバシ保護基準を満たすように加工された形で行われる必要がある．本研究の目的は，そのような情報交換に関するプライバシの制約のもとで，複数の組織に属するデータを効率的に活用して学習を行う手法を考察することである．

考察の対象となるフレームワークは 2 つの要素からなる．1 つは，組織間での情報交換に際して制約条件として課されるプライバシ保護基準であり，これには Dwork [1] によって提案された差分プライバシ（differential privacy）を用いる．このプライバシ制約のもとで，各組織は自分のデータを使って学習した学習器を互いに提供しあう．もう 1 つの要素は，他組織から提供された多数の学習器を統合して 1 つの学習器を作るための方法であり，これには指数型重み付け統合（exponential weighting aggregation, EWA）と呼ばれる手法群を用いる．

このような統合フレームワークの導入によって期待される学習精度への影響の概念図を**図 1** に示す．もし組織間での情報交換が許されていなければ，ある組織 m_0 で学習された学習器 $\hat{f}^{(0)}$ の性能は，全データ D_n を使って学習された理想的な学習器 $\hat{f}_{D_n}$ の性能に劣る．しかし，差分プライバシ制約を満たしたうえでの学習器交換が許されるならば，それらを統合して作った学習器 $\hat{f}_{\mathrm{agg}}$ の性能はより理想的な学習器の性能に近づけることができると考えられる．一般に，共有できる情報量とプライバシ保護強度との間にはトレードオフの関係があるため，プライバシ保護を強くすると個々の弱学習器の性能は下がる．そのため統合学習器の性能もプライバシ強度の設定に応じて変化するが，もしデータ増加によって見込まれる解析精度の向上が大きければ，比較的プライバシ保護尺度を強めに設定しても，統合学習器は元の学習器より高い精度を達成することが期待できる．

学習器統合の手法としてみたときの本研究の位置づけを**表 1** に示す．学習器を統合する手法として主要なものは 2 つに大別することができ，1 つは経験リスク最小化（ERM）に基づく統合手法，もう 1 つは指数型重み付け統合である．罰則つきの経験リスク最小化による統合には BIC（統

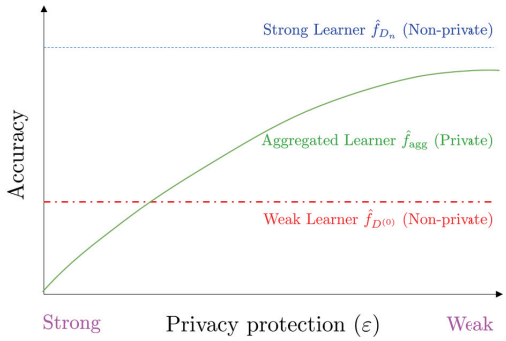


図 1　自組織データのみを使った非プライベート学習器 $\hat{f}^{(0)}$，全データを使った理想的な非プライベート学習器 $\hat{f}_{D_n}$，およびプライベート学習器を統合して作られた学習器 $\hat{f}_{\mathrm{agg}}$ のそれぞれの学習精度と，プライバシ保護強度の関係のイメージ

Fig. 1　An intuitive picture of the relationship between the strength of privacy protection and learning accuracy of estimators. $\hat{f}^{(0)}$ is a non-private local estimator trained over the node m_0's local data set, $\hat{f}_{D_n}$ is a non-private global estimator trained over the whole data set, and $\hat{f}_{\mathrm{agg}}$ is an aggregated estimator made of private weak learners, respectively.

合手法としての扱いは文献 [4] など），LASSO [5]，Dantzig selector [6] などが含まれ，主に回帰問題のモデル選択などで利用されることが多い．本研究に深く関係するものとして，二値分類のタスクに対して，差分プライバシを満たす学習器を統合する枠組みが最初に提案されたのは文献 [3] である．文献 [3] では，差分プライバシを満たすように学習された SVM を，その出力を新しい入力とした L_2-正則化ロジスティック回帰によって統合するという方法がとられており，これは経験リスク最小化による統合手法の一種と見なせる．一方，指数関数型の重み付けを用いて多数の弱学習器を統合する手法も機械学習分野では古くから提案されており，二値分類，線形回帰，密度推定，ノンパラメトリック回帰などへの適用が考察されている．本稿で提案されるフレームワークは，指数型重み付け統合において，弱学習器として分散データから学習された差分プライベート学習器を利用した場合を統括的に含んでいるといえる．指数型重み付けにおいて，分割したデータから弱学習器を

作成するという手法が採用されている例として文献 [8] をあげておく．本来，プライバシを考慮しない通常の学習器統合において，要素となる弱学習器は任意の関数でよく，必ずしもデータの一部から学習されたものである必要はない．しかし，本稿で扱うような各組織に関するプライバシ保護を必要とする状況では，組織間に分散したデータそれぞれで作成した学習器をエキスパートとして統合するという方法が自然に適合する．指数型重み付け統合についてのより詳細なことは 3 章で再び述べる．

プライバシ制約下での統計的学習の問題の取り扱いは，近年さかんに研究されている分野である．文献 [10], [11], [12], [13] では，差分プライバシ制約のもとでの経験リスク最小化について詳細な議論がなされた．また，文献 [14], [15] では，局所プライバシと呼ばれる少し異なるプライバシ制約のもとで，プライバシ保護と汎化能力のトレードオフの関係が理論的に示された．

差分プライバシにおける基本的なアイデアは，本来得たい量に適当なノイズを加えることによって，各個人のデータの変化に対する出力の分布の変化を小さくすることである．一方，医療データ解析などの本来の用途を顧みると，病気の診断や投薬量の決定など，解析結果がきわめてリスクの高い決断に影響する可能性がある．そのような例では，データ解析に対して厳格な精度が要請され，ノイズ付与によってデータを撹乱するという差分プライバシの理念とはしばしば相反しうる．たとえば文献 [16] では，量を誤ることによる致死性が高い薬品の投薬量決定タスクにおいて，既存の差分プライベートな学習器では望まれた性能を達成しないことが指摘された．以上のような背景があり，差分プライバシを保証した場合に，どれだけ信頼のおける解析結果を得ることができるかという問題は非常に関心度が高いものとなっている．

本稿の以下の部分は次のような構成になっている．2 章では，まず差分プライバシの定義を与え，実際に差分プライバシを満たすような学習器の構成方法に関する先行研究の結果を紹介する．3 章では，学習器の統合に用いる指数型重み付けについて説明し，具体例として mirror averaging [9] のアルゴリズムを紹介する．本稿における主要な部分は 4 章であり，差分プライベート学習器の統合手法を導入する．4.1 節では差分プライバシ制約のもとで学習器を交換しあうフレームワークについて詳細を述べる．4.2 節では，問題が二値分類の場合に具体的な統合アルゴリズムを述べ，その理論的な性能について説明する．5 章では，計算機実験によって提案手法の性能を検証する．6 章では本稿全体の内容のまとめを行う．

2. 差分プライバシ

2.1 定義と基本的性質

本節では差分プライバシの定義といくつかの基本的な性質について説明する．データセット D を入力として，要約統計量や学習器など，何らかのデータに依存する値を出力する状況を考える．差分プライバシとは，1 要素のみで異なるデータセットに対する出力の確率分布があまり変わらないということを定式化したものである．

入力となるデータセットを D で表し，データセットの全体を $\mathcal{D}$ と書く．$\mathcal{D}$ に属するデータセットには対称的な隣接関係 $D \sim D'$ が定義されているとする．たとえば $D = (d_1, \ldots, d_n)$ は個人情報を含むデータ d_i $(i = 1, \ldots, n)$ の集まりとし，$\mathcal{D}$ は要素数 n のデータセットの全体とすると，$D \sim D'$ であるとは，添字の適当な置換のもとで，1 要素のみで $d_i \neq d'_i$ となり，それ以外の $n-1$ 個の要素では $d_j = d'_j$ $(j \neq i)$ であることと定義する．隣接関係をどのように定義するかは考えている状況によって異なりうる．各 D に対して確率変数 $f_D : \Omega \to \mathcal{X}$ が与えられれているとする．f_D は D から得られる要約統計量，あるいは D から学習された学習器などに対応する．f_D によって値域の空間 $(\mathcal{X}, \mathcal{A})$ に誘導される確率分布を P_D と表す．差分プライバシとは，$D \sim D'$ の場合に分布 P_D どうしの近さを定義したものである．

定義 2.1（Differential Privacy）．確率変数の集合 $\{f_D : D \in \mathcal{D}\}$ が (ε, δ)-差分プライバシを満たすとは，任意の $D \sim D' \in \mathcal{D}$ と $A \in \mathcal{A}$ に対して

$$P_D(A) \leq e^{\varepsilon} P_{D'}(A) + \delta \tag{1}$$

が成り立つことをいう．ここで，$\varepsilon, \delta \geq 0$ は非負のパラメータである．$\delta = 0$ のときは特に ε-差分プライバシを満たすという．

差分プライバシの直感的な意味は次のとおりである．ある個人 i のデータが d_i から d'_i に変更されたという事実が外部から何らかの手段で知られた場合，変更の前後での出力 f_D と $f_{D'}$ の相違から d_i の値が有意に特定されないようにしたい．そのためには，f_D と $f_{D'}$ の分布が何らかの意味で近くなることが要請される．

もし $f.$ が ε-差分プライバシを満たすとすると，

$$\sup_{D \sim D'} \sup_{A \in \mathcal{A}} \frac{P_D(A)}{P_{D'}(A)} \leq e^{\varepsilon} \tag{2}$$

であるので，差分プライバシとは言い換えれば確率値の比率 $P_D(A)/P_{D'}(A)$ が一様に e^{ε} 以下に抑えられていることである．簡単のため f_D（$D \in \mathcal{D}$）が密度 p_D を持つと仮定すると，尤度比 $p_D(z)/p_{D'}(z)$ が一様に e^{ε} 以下であれば ε-差分プライバシを満たすことが分かる．したがって，仮説検定（尤度比検定）の検出力をある閾値より高くできなくなるため，個人のデータ d_i は「有意には」特定されなくなる．以上は直感的な議論であるが，確率分布間の他の距離尺度（全変動距離，統計的ダイバージェンス）での近接性 [17]，仮説検定の検出力の上界 [18]，任意の事前分布を

与えたときの事後分布の近接性 [19] といった意味での正当化がそれぞれ与えられている．

本節の最後に，差分プライバシの基本的な性質として，合成定理（composition theorem）と呼ばれているものの 1 つを説明する．これは直感的には，一度プライベートに共有された情報は，元のデータセットと独立な情報を使ってどのように加工してもプライバシが保たれるということである．

命題 2.1. 確率変数 $f_D : \Omega \to \mathcal{X}$ の集合 $\{f_D;\ D \in \mathcal{D}\}$ は (ε, δ)-差分プライバシを満たすとする．$g:\ \Omega' \to \mathcal{Y}^{\mathcal{X}}$ は $\{f_D\}$ と独立な確率変数で，$\mathcal{X}$ 上の可測関数に値をとるものとする．このとき，$\{g(f_D)\}$ は (ε, δ)-差分プライバシを満たす．

証明. 可測関数 $h : \mathcal{X} \to \mathcal{Y}$ を固定すると，任意の $D \sim D'$ と $A \in \mathcal{B}(\mathcal{Y})$ に対して

$$\begin{aligned}\Pr(h \circ f_D \in A) &= \Pr(f_D \in h^{-1}(A)) \\ &\le e^{\varepsilon}\Pr(f_{D'} \in h^{-1}(A)) + \delta \\ &= e^{\varepsilon}\Pr(h \circ f_{D'} \in A) + \delta \qquad (3)\end{aligned}$$

が成り立つ．よって，g の分布を P_g とすれば

$$\begin{aligned}\Pr(g \circ f_D \in A) &= \mathbb{E}_{h\sim P_g}[\Pr(g \circ f_D \in A \mid g = h)] \\ &\le \mathbb{E}_{h\sim P_g}[e^{\varepsilon}\Pr(g \circ f_{D'} \in A \mid g = h) + \delta] \\ &= e^{\varepsilon}\Pr(g \circ f_{D'} \in A) + \delta \qquad (4)\end{aligned}$$

であるから主張を得る．

□

2.2　差分プライバシを考慮した学習

本節では，差分プライバシをみたすように学習器を公開する方法について，先行研究で知られている結果を紹介する．本稿で考察する問題の一般的な形は次のように述べられる．データセット $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ は，説明変数 $x_i \in \mathcal{X}$ とラベル変数 $y_i \in \mathcal{Y}$ の n 個の組からなる．ラベルの空間 $\mathcal{Y}$ は，考えている問題が二値分類問題であれば $\mathcal{Y} = \{-1, 1\}$，回帰の問題であれば $\mathcal{Y} = \mathbb{R}$ である．学習の問題とは，データセット D_n が与えられたとき，$f(x)$ が未知の説明変数 x に対するラベル変数 y の予測となるように関数 $f:\ \mathcal{X} \to \mathcal{Y}$ を構成する問題である．この f のことを本稿では学習器という．本稿では簡単のため，二値分類問題に限定して述べる．

学習器 f を外部に公開するとは，ユーザが選んだ $x \in \mathcal{X}$ に対して，ラベルの予測値 $f(x)$ を返す機能（オラクル）を提供することである．この公開規則が，D の変化に関して差分プライバシを満たすようにしたい．2.2.1 項では，二値分類の問題を経験リスク最小化問題としての定式化を述べる．2.2.2 項では，2.2.2 項の定式化のもとで，その解を差分プライバシを満たすように公開する手法について説明する．

2.2.1　経験リスク最小化としての二値分類問題

データ (x_i, y_i)（$i = 1, \ldots, n$）は $\mathcal{X} \times \{-1, 1\}$ 上の未知の確率分布 P からの独立な n 個のサンプルとする．データセット $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ が与えられたとき，誤分類リスク $\mathbb{E}_P\left[1_{\{f(x)\neq y\}}\right]$ を最小化するように学習器（分類器）$f:\ \mathcal{X} \to \{-1, 1\}$ を構成したい．このリスクを最小化する理想的な分類器は $f(x) = \mathrm{sgn}(\mathbb{E}[Y|X = x])$ であることが知られているが，未知である真の条件付き分布 $P(Y|X)$ による期待値を含むため，実際に作ることはできない．

そこで，本来の誤分類損失 $1_{\{f(x)\neq y\}}$ を凸関数で緩和し，理想的な分類器に対して一致性を持つ分類器を作ることを考える．有限次元のパラメータ $\theta \in \mathbb{R}^p$ で添字付けられた学習器 $f_\theta:\ \mathcal{X} \to \{-1, 1\}$ を，次の経験リスク

$$\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; d_i) \qquad (5)$$

を最小化するような θ を求めることで構成する．ここで，d_i はデータの組 (x_i, y_i) であり，$\ell(\cdot; d)$ は損失関数である．損失関数は θ に関する凸関数とする．

線形分類器 $f_\theta(x) = \mathrm{sgn}(\langle\theta, x\rangle)$ を ϕ-risk $\ell_\phi(\theta; d) = \phi(-y\langle\theta, x\rangle)$ に関して最適化する問題は経験リスク最小化問題である．たとえば，サポートベクタマシン（SVM）では損失関数はヒンジ損失 $\ell(\theta; d) = (1 - y\langle\theta, x\rangle)_+$ であり，ロジスティック回帰では $\ell(\theta; d_i) = \log(1 + \exp(-y\langle\theta, x\rangle))$ である．また，L_2-正則化や L_1-正則化など，凸関数の正則化項 $r(\theta)$ を加えたリスクを考えることも多い．この場合は損失関数を $\hat{\ell}(\theta; d_i) = \ell(\theta; d_i) + \frac{1}{n}r(\theta)$ と置き換えればよい．

2.2.2　差分プライベート経験リスク最小化

経験リスク $\mathcal{L}(\cdot; D_n)$ は凸関数であり，したがって経験リスク最小化問題は凸最適化問題である．いま，「差分プライバシを満たすように経験リスク最小化問題の解を公開する」とは，差分プライバシを満たし，かつ目的関数の期待値 $\mathbb{E}_\theta[\mathcal{L}(\tilde{\theta}; D_n)]$ をなるべく小さくするような確率変数 $\tilde{\theta}$ をサンプルすることとして定式化できる．この問題を差分プライベート経験リスク最小化ということにする．

本項では，学習器（関数）そのものではなく，対応する有限次元のベクトル θ を差分プライベートに公開することを考える．ただし，もしそれが可能であるならば，関数 $f_\theta : \mathcal{X} \to \{-1, 1\}$ を差分プライベートに公開することも実は可能である．実際，$D_n \mapsto \tilde{\theta}$ が差分プライベートであり，かつパラメータの値と学習器との対応 $\theta \mapsto f_\theta$ が可測ならば，2.1 節の合成定理によって，$D_n \mapsto f_{\tilde{\theta}}$ という対応は差分プライベートとなる．

差分プライベート経験リスク最小化に対するアプローチとしては，(a) 解に対する出力摂動法 [1]，(b) 目的関数 $\mathcal{L}(\theta; D)$ の摂動法 [10], [11], [20]，(c) exponential sampling [12], [21],

(d) 差分プライベート確率的勾配降下法 [12], [13] などが提案されている．(a) の出力摂動法は最も単純な方法であり，プライバシを考慮しない場合の真の解 $\theta^* = \inf_{\theta\in\Theta} \mathcal{L}(\theta; D_n)$ の各次元ごとに Laplace 分布に従うノイズを加える [1], [10]．しかし，出力摂動法は，他の手法と比較すると理論的な性能は一般に劣る．

本項の以下の部分では (b) 目的関数摂動法と (d) 差分プライベート確率的勾配降下法について説明する．アルゴリズム 1 は目的関数摂動法 [10], [11]，アルゴリズム 2 は差分プライベート確率的勾配降下法 [12] である．

Algorithm 1 Objective Perturbation [10], [11]

Require: dataset D_n, privacy parameter $\varepsilon > 0$ and $\delta \geq 0$, convex loss function $\mathcal{L}(\theta; D_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; d_i)$, $\| \nabla\ell(\theta; d) \|_2 \leq \zeta$ and upper bound $\lambda_{\max}$ of the eigenvalues of $\nabla^2\ell(\theta; d)$

1: set $\Delta \geq \frac{2\lambda_{\max}}{\varepsilon}$
2: **if** $\delta = 0$ (require ε-DP) **then**
3: sample $b \in \mathbb{R}^p$ from the probability distribution with density $\nu_1(b; \varepsilon, \zeta) \propto \exp(-\varepsilon \| b \|_2 / 2\zeta)$
4: **else if** (require (ε, δ)-DP) **then**
5: sample $b \in \mathbb{R}^p$ from $\nu_2(b; \varepsilon, \delta, \zeta) = \mathcal{N}\left(0, \frac{\zeta^2(8\log\frac{2}{\delta} + 4\varepsilon)}{\varepsilon^2} I\right)$
6: **end if**
7: **return** $\theta_{\text{priv}} = \arg\min_{\theta\in\Theta} \mathcal{L}(\theta; D_n) + \frac{\Delta}{2n} \| \theta \|_2^2 + \frac{1}{n} b^\top \theta$

Algorithm 2 Differentially Private Gradient Descent [12]

Require: dataset D_n, privacy parameter $\varepsilon > 0$ and $\delta > 0$, convex and L-Lipschitz loss function $\mathcal{L}(\theta; D_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; d_i)$, and the learning rate η_t $(t = 1, \ldots, n^2)$

1: set $\sigma^2 = \frac{32L^2n^2\log\left(\frac{n}{\delta}\right)\log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$
2: choose any point from $\theta_1 \in \Theta$
3: **for** $t = 1$ to $n^2 - 1$ **do**
4: sample $d \in D_n$ uniformly at random
5: $\theta_{t+1} = \Pi_\Theta\left(\theta_t - \eta_t(n\nabla\ell(\theta_t; d) + b_t)\right)$, $b_t \sim \mathcal{N}(0, \sigma^2 I)$
6: **end for**
7: **return** $\theta_{\text{priv}} = \theta_{n^2}$

これらのアルゴリズムの出力 θ_{priv} は実際に (ε, δ)-差分プライバシを満たすことが示されている [10], [11], [12]．また，θ_{priv} にともなう学習器の性能について，プライバシを考慮しない場合の経験リスクの真の最小値との差（プライバシリスク [13]）$\mathbb{E}_\theta[\mathcal{L}(\theta_{\text{priv}}; D_n)] - \inf_{\theta\in\Theta} \mathcal{L}(\theta; D_n)$ を評価することは自然である．プライバシリスクが評価できれば，たとえば汎化誤差については文献 [22] の方法を用いて得ることができる．アルゴリズム 1 および 2 が達成するプライバシリスクの上界ついて，それぞれ以下の結果が知られている．

定理 2.1（文献 [11], Thorem 4）．$\zeta > 0$ が存在して，$\forall\theta \in \Theta$ と $\forall d \in \mathcal{X}$ について $\| \ell(\theta; d) \|_2 \leq \zeta$ であるとする．$\lambda_{\max}$ は損失関数の Hessian $\nabla^2\ell$ の最大固有値とする．このとき，

(a) $\delta = 0$ のとき，アルゴリズム 1 によって得られた θ_{priv} は次を満たす．

$$\mathbb{E}_\theta\left[\mathcal{L}(\theta_{\text{priv}}; D_n)\right] - \inf_{\theta\in\Theta} \mathcal{L}(\theta; D_n) = O\left(\frac{\zeta \| \theta^* \|_2 \, p\log p}{\varepsilon n}\right) \tag{6}$$

ただし，θ^* は $\inf_{\theta\in\Theta} \mathcal{L}(\theta; D_n)$ を達成する点であり，期待値 $\mathbb{E}_\theta$ は θ_{priv} についてとる（以下同様）．

(b) $\delta > 0$ のとき，アルゴリズム 1 によって得られた θ_{priv} は次を満たす．

$$\mathbb{E}_\theta\left[\mathcal{L}(\theta_{\text{priv}}; D_n)\right] - \inf_{\theta\in\Theta} \mathcal{L}(\theta; D_n) = O\left(\frac{\zeta \| \theta^* \|_2 \sqrt{p\log(1/\delta)}}{\varepsilon n}\right) \tag{7}$$

定理 2.2（文献 [12]，Theorem 2.4）．(a) 損失関数 ℓ は L-Lipshitz であるとする．このとき，ステップ幅 $\eta_t = \frac{\|\Theta\|_2}{\sqrt{t(n^2L^2 + p\sigma^2)}}$ としたときのアルゴリズム 2 の出力 θ_{priv} は次を満たす．

$$\mathbb{E}_\theta\left[\mathcal{L}(\theta_{\text{priv}}; D_n)\right] - \inf_{\theta\in\Theta} \mathcal{L}(\theta; D_n) = O\left(\frac{L \| \Theta \|_2 \log^{3/2}(n/\delta)\sqrt{p\log(1/\delta)}}{\varepsilon}\right) \tag{8}$$

(b) 損失関数 ℓ は L-Lipshitz かつ β-強凸であるとする．このとき，ステップ幅 $\eta_t = \frac{1}{\beta nt}$ としたときのアルゴリズム 2 の出力 θ_{priv} は次を満たす．

$$\mathbb{E}_\theta\left[\mathcal{L}(\theta_{\text{priv}}; D_n)\right] - \inf_{\theta\in\Theta} \mathcal{L}(\theta; D_n) = O\left(\frac{L^2\log^2(n/\delta)p\log(1/\delta)}{n\beta\varepsilon^2}\right) \tag{9}$$

3. 学習器統合

3.1 指数型重み付けによる統合

本節では指数型重み付け統合（Exponential Weighted Aggregate, EWA）の原理について説明する．有限個の学習器 $\hat{f}^{(m)}$（$m = 1, \ldots, M$）を凸結合（またはより一般に線形結合）して学習器 $\hat{f}$ を作ることを考える．つまり，$\Lambda = \{\lambda \in \mathbb{R}^M_{\geq 0};\ \sum_{i=1}^{M} \lambda_i = 1\}$ を確率単体として，学習器の混合率 $\lambda \in \Lambda$ を適切に選ぶことで，

$$\hat{f}_\lambda(x) = \operatorname{sgn}\left(\sum_{m=1}^{M} \lambda_m \hat{f}^{(m)}(x)\right), \tag{10}$$

として新たに学習器 $\hat{f} = \hat{f}_\lambda$ を作る．

指数型重み付け統合の基本的なアイデアは，凸結合の重み $\lambda = (\lambda_1, \ldots, \lambda_M)$ を

$$\lambda_m \propto \exp\left(-\frac{1}{\beta}\sum_{i=1}^{n} \ell(\hat{f}^{(m)}(x_i), y_i)\right) \tag{11}$$

に比例するように混合率を定めることである．言い換えると，学習器の集合 $\{\hat{f}^{(m)}\}$ を，温度パラメータ $\beta > 0$，エネルギーが経験リスク $n \times \mathcal{L}(\hat{f}^{(m)}; D_n)$ であるような Gibbs

分布で平均化して新たな学習器 $\hat{f}$ を作る．

指数型重み付け統合は，学習器を統合する一般的な手法として古くから考察されてきた [9], [23], [24]．二値分類器の統合については文献 [7], [25], [26] などを参照されたい．また，回帰推定量の統合については文献 [8], [27], [28], [29], [30], [31] などを参照されたい．

学習器を統合する手法のその他の枠組みとしては，罰則つき経験リスク最小化（BIC 型罰則，LASSO [5]，Dantzig selector [6]）がある．特に，線形回帰の文脈ではこれらの手法の有用性が知られており，たとえば BIC 型の罰則は後述のオラクル不等式の意味での理論最適を達成することが知られている [4]．一般に，BIC 型の罰則つき最適化は高次元の組合せ最適化となり，指数型重み付けと比較すると計算量が多くなる傾向にある．

さて，統合手法の学習理論的な評価指標を与えるために，オラクル不等式の概念について説明しておく．いま，M 個の弱学習器 $\{\hat{f}^{(m)}\}_{m=1}^{M}$ が与えられたとして，リスク

$$\mathcal{R}(\hat{f}_{\mathrm{agg}}) = \mathbb{E}_{(X,Y)}[\ell(\hat{f}_{\mathrm{agg}}(X), Y)] \tag{12}$$

によって統合された学習器 $\hat{f}_{\mathrm{agg}}$ の性能を評価する．このとき，$\{\hat{f}^{(m)}\}$ の中で最良のものに対する $\hat{f}_{\mathrm{agg}}$ のリスクを評価する不等式

$$\mathcal{R}(\hat{f}_{\mathrm{agg}}) \le \min_{1\le m\le M} \mathcal{R}(\hat{f}^{(m)}) + \Delta_{n,M} \tag{13}$$

をモデル選択型オラクル不等式という．右辺の残差項 $\Delta_{n,M}$ は，統合された学習器 $\hat{f}_{\mathrm{agg}}$ のリスクが，理想的に選ばれた最良の弱学習器のリスクに対してどれだけ悪くなりうるかという，ワーストケースの評価として解釈できる．

文献 [28] では，回帰問題において達成可能な $\Delta_{n,M}$ のオーダの下界として

$$\Delta_{n,M} = \frac{C\log M}{n} \tag{14}$$

が与えられた．この下界は，たとえば次節で説明する mirror averaging によって達成される．

経験リスク最小化に基づく統合を用いる場合，問題に応じた適切な罰則項が選択されることが不可欠である．たとえば，正則化項のない単純な経験リスク最小化 $\hat{f} = \arg\min_{\lambda\in\Lambda} \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{f}_\lambda(x_i), y_i)$ によって混合率 λ を決定する場合，残差項は $\Delta_{n,M} = O\left(\sqrt{\frac{\log M}{n}}\right)$ であることが知られており，これは最適なレートを達成しない [2]．一方，指数型重み付けによる統合手法は，理論最適性の保証のために損失関数や弱学習器に要請される正則条件は比較的少ないといわれている [29]．

3.2　Mirror averaging による統合

本節では指数型重み付けに区分されるアルゴリズムの具体例として，mirror averaging [9] によって二値分類の学習器を統合する手法について説明する．まず $t = 1,\dots,n$ について，t 個のデータ $\{d_1,\dots,d_t\}$ を用いた混合率 $\lambda^{(t)} = (\lambda_1^{(t)},\dots,\lambda_M^{(t)})^\top$ を

$$\lambda_m^{(t)} = \frac{\exp\left(-\beta^{-1}\sum_{i=1}^{t}\ell(\hat{f}^{(m)}(x_i), y_i)\right)}{\sum_{l=1}^{M}\exp\left(-\beta^{-1}\sum_{i=1}^{t}\ell(\hat{f}^{(l)}(x_i), y_i)\right)} \tag{15}$$

として計算する．ただし $\beta > 0$ は温度パラメータであり，ℓ に応じて十分大きく定める．次に，学習器の混合率 $\lambda = (\lambda_1,\dots,\lambda_M)^\top$ を $\lambda^{(t)}$ の平均

$$\lambda_m = \frac{1}{n}\sum_{t=1}^{n}\lambda_m^{(t)} \tag{16}$$

で求める．したがって，統合された学習器 $\hat{f}_{\mathrm{agg}}$ は次のように表される．

$$\begin{aligned}\hat{f}_{\mathrm{agg}}(x) &= \mathrm{sgn}\left(\sum_{m=1}^{M}\lambda_m\tilde{f}^{(m)}(x)\right)\\ &= \mathrm{sgn}\left(\sum_{m=1}^{M}\sum_{t=1}^{n}\lambda_m^{(t)}\tilde{f}^{(m)}(x)\right).\end{aligned} \tag{17}$$

分類器 $\tilde{f}(x)$ に対する ϕ-リスクは

$$\mathcal{R}(\tilde{f}) = \mathbb{E}_{(X,Y)}[\ell(\tilde{f}(X), Y)] \tag{18}$$

で与えられる．ここで，期待値 $\mathbb{E}_{(X,Y)}$ はデータ (X_1, Y_1) の独立なコピーに関してとる．式 (17) で定まる mirror averaging 推定量の ϕ-リスクに関する理論的性能については次の定理が知られている．

定理 3.1（文献 [9]，Corollary 5.3）．二値分類問題において，分類器 $\hat{f}(x) = \mathrm{sgn}\tilde{f}(x)$ に対する損失関数が ϕ-損失 $\ell(\tilde{f}(x), y) = \phi(-y\tilde{f}(x))$ である場合を考える．ここで $\phi : \mathbb{R} \to \mathbb{R}_{\ge 0}$ は 2 回微分可能な凸関数であって，$\beta_\phi > 0$ が存在して

$$\forall|x| \le 1,\ \{\phi'(x)\}^2 \le \beta_\phi\phi''(x) \tag{19}$$

を満たすものとする．

このとき，$\beta \ge \beta_\phi$ に対する mirror averaging 推定量 (16) について次が成立する．

$$\mathbb{E}^n_{(X,Y)}[\mathcal{R}(\tilde{f}_{\mathrm{agg}})] \le \min_{1\le m\le M}\mathcal{R}(\tilde{f}^{(m)}) + \frac{\beta\log M}{n} \tag{20}$$

ただし，期待値 $\mathbb{E}^n_{(X,Y)}$ は D_n に含まれるデータの同時分布に関してとる

たとえば，ロジスティック回帰の損失 $\phi(z) = \log(1+\exp(z))$ は条件 (19) を満たし，その場合 $\beta_\phi = \mathrm{e}$ ととればよい．また，モデル選択型オラクル不等式 (20) の残差項 $\frac{\beta\log M}{n}$ のオーダーは理論的な下界と一致する．

4. 差分プライベート学習器の統合

4.1 統合フレームワークの概略

本節では，複数組織に分散したデータからの学習の問題を定式化し，それを扱うための差分プライベート学習器統合の一般的なフレームワークを提案する．全データ $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ が，実際には $M+1$ の異なる組織に分散しているとする．各組織 m はそれぞれデータセット $D^{(m)}$（$m = 0, 1, \ldots, m$）を持ち，全データ D_n は

$$D_n = \coprod_{m=0}^{M} D^{(m)} \tag{21}$$

のように $D^{(m)}$ の非交和となっている．ただし，$\coprod$ は非交和を表す記号であり，

$$D^{(m)} = \{(x_1^{(m)}, y_1^{(m)}), \ldots, (x_{n_m}^{(m)}, y_{n_m}^{(m)})\}, \tag{22}$$

および $\sum_{m=0}^{M} n_m = n$ である．

各組織 m は，データセット D_n の情報を用いて学習の問題を解きたい．しかし，他組織のデータ $D^{(l)}$（$l \neq m$）を閲覧することはできず，統合されたデータセット D_n の情報を直接利用して学習することはできないものとする．一方，他組織 l（$l \neq m$）から，$D^{(l)}$ に関して (ε, δ)-差分プライバシを満たすように公開された情報を譲り受けることは許されているとする．以上のような状況で，各組織 m が全データ D_n の情報をなるべく効率的に利用して学習器を作る問題を考える．この問題に対する一般的なアプローチとして，各組織 l から (ε, δ)-差分プライバシを満たすような学習器 $\hat{f}^{(l)}$ を受け取り，それらを弱学習器として統合することが考えられる．

以下では，$m = m_0 = 0$ は自組織を表す添字とし，$D^{(0)}$ はプライバシを考慮せずに使える自組織内のデータとする．**図 2** は差分プライベート弱学習器統合のフレームワークの概念図である．

組織 m_0 は，自分以外の各組織から弱学習器 $\hat{f}^{(m)}$（$m = 1, \ldots, M$）を受け取る．このとき，各 $\hat{f}^{(m)}$ はデータセット $D^{(m)}$ に関して (ε, δ)-差分プライバシを満たすように作られる．

ここで，個々の問題設定に対して，そのような学習器が具体的に作れることが必要である．汎化性能を問わなければ，差分プライベートに学習器を作ること自体は一般に可能であると考えられる．もし学習の問題が経験リスク最小化として定式化できる場合は，2.2.2 項の差分プライベート経験リスク最小化の手法によって作られたものを用いる．そうでない場合，すなわちノンパラメトリック回帰や密度推定の問題では，たとえば文献 [32] の Gaussian process による出力摂動法などを用いて学習器を構成する．

次に，組織 m_0 は，自組織で自由に使えるデータ $D^{(0)}$ に基づき，提供された M 個の弱学習器 $\{\hat{f}^{(m)}\}_{m=1}^{M}$ を指数型重み付けによって統合する．これによって，組織 m_0 は統合された学習器 $\hat{f}_{\mathrm{agg}}$ を得る．

上記のフレームワークの効果の直感的な説明は次のとおりである．組織 m_0 が受け取る弱学習器の集合 $\hat{f}^{(m)}$ は，理想的には他組織のデータセット $\{D^{(m)}\}$ が持つ情報のうち，個人情報に起因する成分を取り除いて学習器の汎化能力に影響する部分だけを抽出したものと考えられる．したがって，差分プライバシのノイズによる性能の劣化よりもデータ数増加による学習器の性能の向上が大きく見込めるならば，統合した学習器 $\hat{f}_{\mathrm{agg}}$ は組織 m_0 で独自に学習した学習器よりも良くなることが期待できる．

また，統合した学習器 $\hat{f}_{\mathrm{agg}}$ は，次の命題の意味で差分プライバシを満たす．

命題 4.1. データセット D_n に含まれる点は，ある未知の確率分布 P にそれぞれ従い，互いに独立であるとする．各組織 $m = 1, \ldots, M$ が公開した弱学習器 $\hat{f}^{(m)}$ は，データセット $D^{(m)}$ に関して (ε, δ)-差分プライバシを満たすとする．統合学習器 $\hat{f}_{\mathrm{agg}}$ は，自組織のデータセット $D^{(0)}$ に対する損

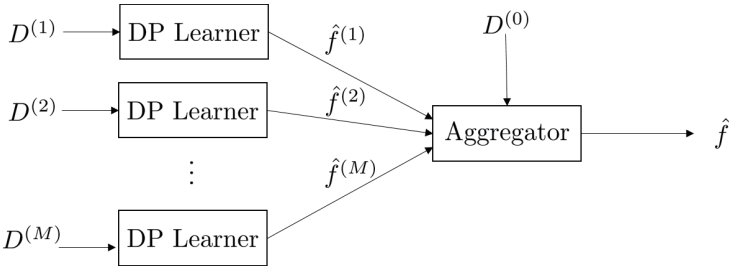


図 2　プライベート学習器統合フレームワーク．各組織 m は学習器 $\hat{f}^{(m)}$ を差分プライバシを満たすように構成し，互いに共有する．組織 m_0 は，他組織 m（$m = 1, \ldots, M$）から提供された学習器を自組織のデータ $D^{(0)}$ を用いて統合し，学習器 $\hat{f}_{\mathrm{agg}}$ を構成する

Fig. 2　Framework for aggregation of private weak learners. Each organization m train a differentially private learner $\hat{f}^{(m)}$ on its local data $D^{(m)}$, and disclose them each other. Then an organization m_0 combines the weak learners using its local data $D^{(0)}$ and obtains an aggregated learner $\hat{f}_{\mathrm{agg}}$.

失の情報 $\{\ell(\hat{f}^{(m)}(x_i^{(0)}), y_i^{(0)}); m = 1, \ldots, M,\ (x_i^{(0)}, y_i^{(0)}) \in D^{(0)}\}$ を用いて構成されるとする．このとき，$\hat{f}_{\mathrm{agg}}$ は他組織のデータセット $D_n \setminus D^{(0)} = \coprod_{m=1}^{M} D^{(m)}$ に関して (ε, δ)-差分プライバシを満たす．

命題 4.1 により，mirror averaging などの典型的な学習器統合のアルゴリズムによって作られた統合学習器は $D_n \setminus D^{(0)} = \coprod_{m=1}^{M} D^{(m)}$ について (ε, δ)-差分プライバシを満たすことが分かる．つまり，組織 m_0 が本来知りえない，他組織 m（$m \neq m_0$）に属する個人ユーザの情報は，学習器の統合を行ったとしても依然として (ε, δ)-差分プライバシで保護されている．証明は，データの独立性の仮定と命題 2.1 の合成定理から従う．

なお，学習タスクが二値分類問題である場合には，枠組みとして同等のものが文献 [3] によって提案されている．文献 [3] では，統合フェーズにおける差分プライベート線形分類器の混合率を，L_2-正則化ロジスティック回帰で決定する方法がとられており，経験リスク最小化による混合の一種であると見なせる．

4.2　ロジスティック回帰の例

ここでは具体例として，ロジスティック回帰を統合した場合の性能について考察する．ロジスティック回帰の損失関数は，線形分類器 $f_\theta(x) = \mathrm{sgn}(\langle \theta, x \rangle)$ に対して

$$\ell(\theta, d) = \ell(f_\theta(x), y) = \log(1 + \exp(-y\langle \theta, x \rangle)) \qquad (23)$$

で定義される．通常の非プライベートな設定においては，L_2-正則化ロジスティック回帰推定量は

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \langle \theta, x_i \rangle)) + \lambda_{\mathrm{reg}} \parallel \theta \parallel_2^2 \right\} \qquad (24)$$

で与えられる．ここで，$\Theta \subset \mathbb{R}^p$ は推定量の候補全体の凸集合である．すべてのデータ $(X, Y) \in \mathbb{R}^p \times \{-1, 1\}$ はある確率分布 P から独立に生成されているとする．簡単のため，本節では X の周辺分布の台は $\mathbb{R}^p$ の単位球に含まれると仮定する．また，Θ もコンパクトとする．

4.1 節で導入したフレームワークに従い，各組織は目的関数摂動法（アルゴリズム 1）によって ε-差分プライベートな推定量を学習し，互いに公開する．次に，組織 m_0 は他組織から公開された M 個の線形分類器 $\{\hat{\theta}^{(m)}\}$ を mirror averaging によって統合し，新たな線形分類器 $\hat{\theta}_{\mathrm{agg}}$ を作る．

$\hat{\theta}_{\mathrm{agg}}$ については，次の評価が得られる．ただし，リスク関数は $\mathcal{R}(\theta) = \mathcal{R}(f_\theta) = \mathbb{E}_{(X,Y)}[\ell(f_\theta(X), Y)]$ で与えられる．

定理 4.1. 上記のように統合推定量 $\hat{\theta}_{\mathrm{agg}}$ 構成する．このとき，自組織 m_0 のデータ $D^{(0)}$ を除いたデータ $D_n \setminus D^{(0)}$ の同時分布に関して確率 $1 - \alpha$（$0 < \alpha < 1$）で，次の不等式が成り立つ．

$$\begin{aligned} &\mathbb{E}_{(X,Y)}^{n_0} \mathbb{E}_\theta^M [\mathcal{R}(\hat{\theta}_{\mathrm{agg}})] - \inf_{\theta \in \Theta} \mathcal{R}(\theta) \\ &\leq A_2(\lambda_{\mathrm{reg}}) + \frac{1}{\lambda_{\mathrm{reg}}} \left\{ \sqrt{\frac{M \log(2M/\alpha)}{n - n_0}} + \sqrt{\frac{4M}{n - n_0}} \right. \\ &\left. + \frac{C_1 M \log(2M/\alpha)}{n - n_0} \right\} + \frac{C_2 M}{\varepsilon(n - n_0)} + \frac{\beta \log M}{n_0} \end{aligned} \qquad (25)$$

ここで，式中に現れる期待値について，$\mathbb{E}_{(X,Y)}^{n_0}$ は $D^{(0)}$ に含まれるデータの同時分布 P に関してとり，$\mathbb{E}_\theta^M$ は各組織 $m \in \{1, \ldots, M\}$ の目的関数摂動法によって付与されたノイズに関してとる．また，$A_2(\lambda_{\mathrm{reg}})$ は正則化パラメータ λ_{reg} に依存する値である．C_1，C_2 はそれぞれ正の定数である．$A_2(\lambda_{\mathrm{reg}})$，$C_1$，$C_2$ はいずれも，データ数 n_0，n，組織数 M，プライバシパラメータ ε にはよらない値である．

定理 4.1 の導出は付録 A.1 で示す．

5.　計算機実験

提案手法の性能を評価するため，人工データ（5.1 節）および実データ（5.2 節）に対して計算機実験を行った．

各実験に共通する設定を以下で説明する．まず，各組織に対応する $M + 1$ 個のノードに対して $n/(M + 1)$ 個ずつのデータが均等に配置されるようにする．各ノードは自分のデータを用いて，ε-差分プライベートなロジスティック回帰分類器を学習する．差分プライベートロジスティック回帰の学習アルゴリズムとしては文献 [10] の目的関数摂動法を採用した．

次に，各ノードは自分以外の M ノードから受け取った弱学習器を mirror averaging (17) によって統合する．統合時の損失関数としてはロジスティック回帰の損失関数を用いる．この損失に対しては，温度パラメータを $\beta \geq \beta_\phi = e \approx 2.718$ とすると条件 (19) を満たす．そこで，本章の実験では一律に $\beta = 3$ とした．

5.1　人工データ

本節では人工的に生成したデータに対する実験を行い，提案手法の性能について考察する．

対象となる分類器は線形分類器であるため，線形分離可能なデータを正しく分類できるかどうかが重要な評価尺度の 1 つであると考えられる．

そこで，文献 [20] にならい，次のようにしてデータを生成し，テストデータに対する正答率で分類器の性能を評価した．まず，10 次元の単位球面上の一様分布から法線ベクトル 1 点をサンプルし，原点を含む分離平面を生成した．次に，同じく 10 次元の単位球面上の一様分布からデータをサンプルし，分離平面で区切ることで二値のラベルを付与した．この際に分離平面から距離 0.03 以内をマージンとし，マージン内に含まれたデータは取り除いた．以上のようにして二値のラベルと 10 次元の単位ベクトルの組か

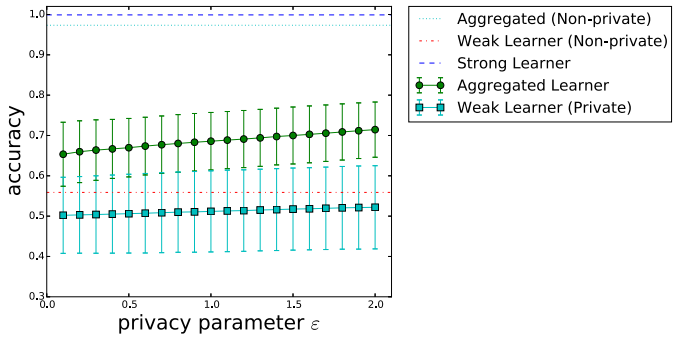


図 3　プライバシパラメータを変化させた場合の正答率．データ数 $n = 5,000$，ノード数 $M + 1 = 100$

Fig. 3　Average accuracies on synthetic data with $n = 5,000$ and $M + 1 = 100$.

表 2　統合学習器の正答率が自組織データによるロジスティック回帰の正答率を上回ったノード数

Table 2　Number of nodes where the aggregated classifiers exceeds the local classifier in accuracy.

ε	0.1	0.3	0.5	0.7	0.9	1.1	1.3	1.5	1.7	1.9
Improved	88	88	92	97	97	97	97	99	99	100

らなるデータを 6,000 点生成した．そのうち $n = 5,000$ 点を訓練データ，残りの 1,000 点をテストデータとして用いた．なお，テストデータには二値のラベルが均等に 500 点ずつ含まれるようにした．よって特に，どちらか一方のラベルのみを出力する単純な分類器の正答率は 0.5 となる．

まず予備実験として，全 5,000 データを利用してロジスティック回帰を行った場合の正答率を計算すると 0.999 であった．これは，プライバシをまったく考慮することなくデータを自由に共有できる場合の正答率に対応している．

統合学習器の性能に対するプライバシ保護の尺度 ε の寄与について考察する．ノードの数は 100 とし，各ノードに 5,000 個の訓練データを 50 個ずつランダムに割り当てた．各ノードは公開用の ε-差分プライベートなロジスティック回帰分類器と，比較用の通常のロジスティック回帰分類器をそれぞれ学習した．差分プライベートな学習器を互いに交換したのち，mirror averaging によってそれらを統合した．

図 3 は横軸に ε，縦軸に正答率をプロットしたものである．ただし，正答率は 100 個のノード間で平均をとり，差分プライバシを考慮した手法の場合はさらに標準偏差をプロットしている．まず，自分のデータのみで通常のロジスティック回帰を学習した場合の平均正答率（一点鎖線）は 0.559 であった．これは，情報をいっさい共有することができず，各ノードが孤立している場合に達成可能な精度に相当する．また，プライバシを考慮しなくてもよいと仮定した場合（あるいは $\varepsilon = \infty$ と見なした場合），通常の学習器を互いに交換した場合の統合学習器の平均正答率（鎖線）は 0.973 であった．これにより，データそのものを共有しなくても，弱学習器さえ交換することができれば，mirror averaging によって他組織のデータを効率的に利用できることを示唆している．

次に，本稿の枠組みに従い，ε-差分プライバシを満たす学習器を統合したものの平均正答率が丸いマーカで示したプロットである．平均正答率は，ε が大きくなり，差分プライバシの保護強度が弱まるにつれて増加することが見てとれる．一方，$\varepsilon = 0.1$ の場合であっても，統合学習器の平均正答率は自分のデータのみを使った学習器の平均正答率を上回っている．なお，四角形のマーカで示したプロットは個々の差分プライベート学習器の平均正答率であり，これらはプライバシ保護のためのノイズを付与しているため，通常のロジスティック回帰分類器よりも正答率が低くなっている．しかし，それらを統合した場合にはデータ増加による寄与によって，より理想的な学習器の性能に近づくことができていると考えられる．また，**表 2** は，100 ノードのうち，統合学習器の正答率が自組織データのみで学習した分類器の正答率を上回ったノード数である．大部分のノードが，フレームワークに加入することによって元より精度の高い学習器が得ることができていることが分かる．

なお，本節の実験において，ロジスティック回帰の正則化パラメータ λ_{reg} は，文献 [20] にならい一律に 0.01 と設定した．この設定方法に関する注意点を述べておく．プライバシを考慮しない通常の学習の場合には，正則化パラメータは訓練データを用いた交差検定などの手法によって決定することが可能である．一方，今回のような設定では，チューニングしたパラメータそのものがデータに依存するため，その値を代入して最適化された学習器は差分プライバシを満たすことが保証されなくなるという問題が生じる．これに対するアプローチとしては，たとえば，有限個の候

補の中から差分プライバシを満たすようにパラメータを選択する方法が提案されている [10]．しかし，差分プライバシ制約下での学習の問題におけるこのようなハイパーパラメータ選択の問題は，それ自体が 1 つの大きなテーマとして研究されているものであり [33]，今回のフレームワークに適合した効果的なチューニング手法の考察は本稿で取り扱う範疇を超える．そこで本稿では，学習器統合による効果をより単純化して比較することを考え，前もって決定した正則化パラメータを全ノードで一律に用いることとした．

5.2　医療関連データ

次に，より実用的な例において提案手法の性能を評価するため，実データに対する実験を行った．実験には UCI Machine Learning Repository [34] で公開されている Pima Indians Diabetes および Breast Cancer Wisconsin（Diagnostic）データセットを用いた（以下，それぞれ Diabetes / Breast Cancer とする）．

Breast Cancer のタスクは，細胞核の画像特徴量から，がん細胞の良性（B）あるいは悪性（M）の二値ラベルを判定するものである．Breast Cancer の各データは実数 30 次元からなる特徴量ベクトルと二値のラベルの組であり，全体で 569 点からなる．そのうち B ラベルがついたデータは 357 点，M ラベルは 212 点である．実験では，M ラベルから 85 点，B ラベルから 84 点をランダムに取得し，計 169 点をテストデータとした．残りの 400 点を訓練データとして，10 個のノードに 40 点ずつランダムに配分した．

Diabetes のタスクは，患者の年齢，血糖値，血圧，妊娠回数などのデータから糖尿病の診断結果を推定するものである．Diabetes の各データは 8 次元の整数値および実数値からなる患者の属性ベクトルと糖尿病であるか否かのラベルからなり，データ数は 768 である．そのうち，糖尿病と診断された正例データは 268 点，負例は 500 点である．実験では，正例と負例からそれぞれ 84 点ずつランダムに取得し，計 168 点をテストデータとした．残りの 600 点を訓練データとし，10 個のノードに 60 点ずつランダムに配分した．

図 4，**図 5** はそれぞれ Breast Cancer および Diabetes データセットに対する実験結果である．Breast Cancer データセットでは，プライバシ保護強度が強い場合（$\varepsilon < 1.5$）には，統合学習器の平均精度は自分のデータのみを使った学習器の精度を改善していない．しかし，$\varepsilon \leq 2.5$ のとき，統合学習器は自分のデータのみを使った学習器の精度を上回った．そこで，この場合には，プライバシ強度を $\varepsilon \approx 2.5$ 程度に設定しても情報共有のメリットがあるといえる．また，Diabetes データセットでは，$\varepsilon \geq 0.3$ のとき，統合学習器は自組織のデータを用いた学習器と比較して平均的に同程度か，上回る精度を示した．

なお，文献 [3] において，経験リスク最小化に基づく統合手法によっても改善がみられるという実験結果が示されている．しかし，文献 [3] ではプライバシ強度を $\varepsilon = 10$ に設定しており，これは 1 データのみ異なるデータセットからの出力分布の確率密度関数の値が各点で $e^{10} \approx 2.2 \times 10^5$ 倍程度異なることを許容することを意味する．本節の実験では，$\varepsilon \approx 2.5$ 程度でも改善が確認でき，より現実的なプライバシ強度でも統合のメリットが期待できることを示唆している．

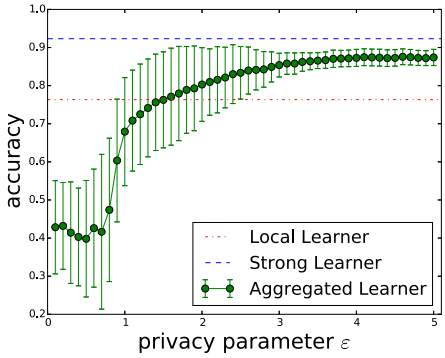


図 4　Breast Cancer における平均正答率
Fig. 4　Breast Cancer.

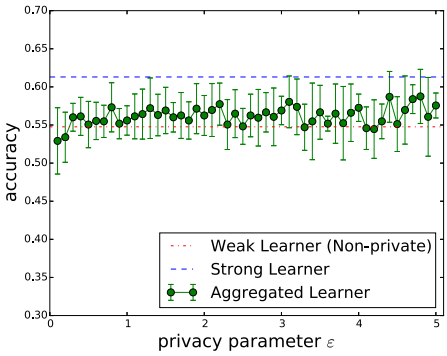


図 5　Diabetes における平均正答率
Fig. 5　Diabetes.

6.　まとめと考察

本稿では，差分プライベート学習の手法と弱学習器統合の手法を組み合わせることにより，差分プライバシ制約のもとで複数組織に分散したデータを効率的に利用して学習を行うフレームワークを提案した．また，二値分類問題における具体的なアルゴリズムとして，差分プライベート経験リスク最小化によって得られた弱学習器を mirror averaging によって統合する手法を提案した．さらに具体的な場合として，弱学習器がロジスティック回帰である場合の理論的な性能について議論し，計算機実験によって提案手法の有用性を確認した．

本稿の枠組みそのものは，二値分類以外にもより広いクラスの統計的学習の問題に適用可能である．たとえば，線形回帰の問題に対しては，本稿と同様にして差分プライベート経験リスク最小化および mirror averaging が適用できると考えられ，その性能の検証は本研究の今後の課題で

ある．5 章の計算機実験の結果は，比較的厳しいプライバシ制約のもとでも統合によって学習器の精度向上が見込めることを示した．このことは，文献 [16] で指摘されているような，差分プライバシを保証すると医療データ解析で要請される精度が得られないというジレンマ的問題が部分的に解決できる可能性を示唆している．

謝辞　本研究は科学研究費補助金基盤研究（B）課題番号 15H02700 の支援を受けて行った．

参考文献

[1] Dwork, C.: Differential privacy, *Proc. 33rd International Colloquium on Automata, Languages and Programming* (*ICALP*), pp.1–12 (2006).

[2] Lecué, G. and Mendelson, S.: Aggregation via empirical risk minimization, *Probability Theory and Related Field*, Vol.145, No.3-4, pp.591–613 (2009).

[3] Sarwate, A.D., Plis, S.M., Turner, J.A., Arbabshirani, M.R. and Calhoun, V.D.: Sharing privacy-sensitive access to neuroimaging and genetics data: A review and preliminary validation, *Frontiers in Neuroinformatics*, Vol.8 (2014).

[4] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H.: Aggregation for Gaussian regression, *The Annals of Statistics*, Vol.35, No.4, pp.1674–1697 (2007).

[5] Tibshirani, R.: Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society: Series B*, Vol.58, No.1, pp.267–288 (1996).

[6] Candes, E. and Tao, T.: The Dantzig selector: Statistical estimation when p is much larger than n, *The Annals of Statistics*, Vol.35, No.6, pp.2313–2351 (2007).

[7] Vovk, V.: Aggregating strategies, *Proc. 3rd Annual Conference on Learning Theory* (*COLT*), pp.371–386 (1990).

[8] Yang, Y.: Adaptive regression by mixing, *Journal of the American Statistical Association*, Vol.96, pp.574–588 (2001).

[9] Juditsky, A., Rigollet, P. and Tsybakov, A.B.: Learning by mirror averaging, *The Annals of Statistics*, Vol.36, No.5, pp.2183–2206 (2008).

[10] Chaudhuri, K., Monteleoni, C. and Sarwate, A.: Differentially private empirical risk minimization, *Journal of Machine Learning Research*, Vol.12, pp.1069–1109 (2011).

[11] Kifer, D., Smith, A. and Thakurta, A.: Private convex empirical risk minimization and high-dimensional regression, *Proc. 25th Annual Conference on Learning Theory* (*COLT*), pp. 25.1–25.40 (2012).

[12] Bassily, R., Smith, A. and Thakurta, A.: Differentially Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds, *IEEE 55th Annual Symposium on Foundations of Computer Science* (*FOCS*) (2014).

[13] Talwar, K., Thakurta, A. and Zhang, L.: Private Empirical Risk Minimization Beyond the Worst Case: The Effect of the Constraint Set Geometry, *ArXiv e-prints* (2014).

[14] Duchi, J., Jordan, M. and Wainwright, M.: Local privacy and statistical minimax rates, *IEEE 54th Annual Symposium on Foundations of Computer Science* (*FOCS*), pp.429–438 (2013).

[15] Duchi, J., Jordan, M. and Wainwright, M.: Privacy aware learning, *J. ACM*, Vol.61, No.6, Article No.38 (2014).

[16] Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D. and Rintenpart, T.: Privacy in Pharmocogenetics: An End-to-End Case Study of Personalized Warfalin Dosing, *23rd USENIX Security Symposium* (2014).

[17] Dwork, C. and Roth, A.: *The Algorithmic Foundations of Differential Privacy*, Now Publishers (2014).

[18] Wasserman, L. and Zhou, S.: A statistical framework for differential privacy, *The Journal of The Americal Statistical Association*, Vol.105, pp.375–289 (2010).

[19] Kasiviswanathan, S. and Smith, A.: On the 'semantics' of differential privacy: A Bayesian formulation, *Journal of Privacy and Confidentiality*, Vol.6, No.1, pp.1–16 (2014).

[20] Chaudhuri, K. and Moteleoni, C.: Privacy-preserving Logistic Regression, *Proc. 22nd Annual Conference on Neural Information Processing System* (*NIPS*) (2008).

[21] McSherry, F. and Talwar, K.: Mechanism design via differential privacy, *IEEE 48th Annual Symposium on Foundations of Computer Science* (*FOCS*), pp.94–103 (2007).

[22] Shalev-Shwartz, S., Shamir, O., Srebro, N. and Sridharan, K.: Stochastic Convex Optimization, *Proc. 22nd Annual Conference on Learning Theory* (*COLT*) (2009).

[23] Catoni, O.: *Statistical Learning Theory and Stochastic Optimization*, Springer (2004).

[24] Cesa-Bianchi, N. and Lugosi, G.: *Prediction, Learning and Games*, Cambridge University Press (2006).

[25] Kivinnen, J. and Warmuth, M.K.: Averaging expert predictions, *4th European Conference on Computational Learning Theory* (*EuroCOLT*), pp.153–167 (1999).

[26] Juditsky, A.B., Nazin, A.V., Tsybakov, A.B. and Vayatis, N.: Recursive aggregation of estimators by the mirror descent algorithm with averaging, *Problems of Information Transmission*, Vol.41, No.4, pp.368–384 (2005).

[27] Nemirovski, A.: *Topics in non-parametric statistics*, Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998, Lecture Notes in Mathematics, Vol.1738, Springer, New York (2000).

[28] Tsybakov, A.B.: Optimal rates of aggregation, Technical report (2003).

[29] Dalalyan, A. and Tsybakov, A.B.: Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity, *Machine Learning*, Vol.72, pp.39–61 (2008).

[30] Dalalyan, A. and Tsybakov, A.B.: Mirror averaging with sparsity priors, *Bernoulli*, Vol.18, No.3, pp.914–944 (2012).

[31] Alquier, P. and Lounici, L.: PAC-Bayesian bounds for sparse regression estimation with exponential weights, *Electronic Journal of Statistics*, Vol.5, pp.127–145 (2011).

[32] Hall, R., Rinaldo, A. and Wasserman, L.: Differential privacy for functions and functional data, *Journal of Machine Learning Research*, Vol.14, pp.703–727 (2013).

[33] Chaudhuri, K. and Vinterbo, S.: A Stability-based Validation Procedure for Differentially Private Machine Learning, *Proc. 28th Annual Conference on Neural Information Processing System* (*NIPS*) (2013).

[34] Asuncion, A. and Newman, D.J.: UCI Machine Learning Repository.

[35] Steinwart, I. and Christmann, A.: *Support Vector Machines*, Springer (2008).

付　　録

A.1　定理 4.1 の証明

最初に個々の組織 $m = 1, \ldots, M$ について，差分プライベート弱学習器 $\theta_{\mathrm{priv}} = \theta_{\mathrm{priv}}^{(m)}$ の性能を評価しよう．θ_{priv} による期待値をとった汎化誤差 $\mathbb{E}_\theta[\mathcal{R}(\theta_{\mathrm{priv}})] - \inf_{\theta\in\Theta}\mathcal{R}(\theta)$ の確率的上界を計算する．まず，経験リスクおよびリスクと正則化項との和を最小化する θ をそれぞれ

$$\theta_{n_m,\lambda_{\mathrm{reg}}} = \operatorname*{argmin}_{\theta\in\Theta}\left\{\mathcal{L}(\theta; D^{(m)}) + \lambda_{\mathrm{reg}} \parallel \theta \parallel_2^2\right\}, \quad \text{(A.1)}$$

$$\theta_{P,\lambda_{\mathrm{reg}}} = \operatorname*{argmin}_{\theta\in\Theta}\left\{\mathcal{R}(\theta) + \lambda_{\mathrm{reg}} \parallel \theta \parallel_2^2\right\} \quad \text{(A.2)}$$

とおく．すると

$$\begin{aligned}
&\mathbb{E}_\theta[\mathcal{R}(\theta_{\mathrm{priv}})] - \inf_{\theta\in\Theta}\mathcal{R}(\theta)\\
&\le \mathbb{E}_\theta[\mathcal{R}(\theta_{\mathrm{priv}}) + \lambda_{\mathrm{reg}} \parallel \theta_{\mathrm{priv}} \parallel_2^2] - \inf_{\theta\in\Theta}\mathcal{R}(\theta) \quad \text{(A.3)}\\
&= \mathbb{E}_\theta\left[\left\{\mathcal{R}(\theta_{\mathrm{priv}}) + \lambda_{\mathrm{reg}} \parallel \theta_{\mathrm{priv}} \parallel_2^2\right\}\right.\\
&\quad \left. - \left\{\mathcal{R}(\theta_{P,\lambda_{\mathrm{reg}}}) + \lambda_{\mathrm{reg}} \parallel \theta_{P,\lambda_{\mathrm{reg}}} \parallel_2^2\right\}\right]\\
&\quad + \mathcal{R}(\theta_{P,\lambda_{\mathrm{reg}}}) + \lambda_{\mathrm{reg}} \parallel \theta_{P,\lambda_{\mathrm{reg}}} \parallel_2^2 - \inf_{\theta\in\Theta}\mathcal{R}(\theta)\\
&\le \mathbb{E}_\theta\left[\left\{\mathcal{R}(\theta_{\mathrm{priv}}) + \lambda_{\mathrm{reg}} \parallel \theta_{\mathrm{priv}} \parallel_2^2\right\}\right.\\
&\quad \left. - \left\{\mathcal{R}(\theta_{P,\lambda_{\mathrm{reg}}}) + \lambda_{\mathrm{reg}} \parallel \theta_{P,\lambda_{\mathrm{reg}}} \parallel_2^2\right\}\right]\\
&\quad + \mathcal{R}(\theta_{n_m,\lambda_{\mathrm{reg}}}) + \lambda_{\mathrm{reg}} \parallel \theta_{n_m,\lambda_{\mathrm{reg}}} \parallel_2^2 - \inf_{\theta\in\Theta}\mathcal{R}(\theta) \quad \text{(A.4)}
\end{aligned}$$

である．ここで，第 1 の不等式 (A.3) は自明であり，第 2 の不等式 (A.4) は (A.2) の右辺の目的関数の凸性から，最小値を達成する点が一意であることから分かる．この最右辺は，プライバシリスクの期待値と，プライバシを考慮しない場合の汎化誤差の和の形になっていることに注意する．

プライバシリスクを評価する．ロジスティック回帰は一般化線形問題（A.2.1 節参照）であり，$\phi(z) = \log(1 + \mathrm{e}^z)$ は 1-Lipshitz である．よって，$\ell(\theta, d)$ は $\sup\{\parallel x \parallel; x \in \mathrm{supp}P_X\} \le 1$ より θ について 1-Lipshitz である．また，Θ のコンパクト性の仮定より，$R > 0$ が存在して，すべての $\theta \in \Theta$ に対して $\parallel \theta \parallel_2 \le R$ が成立する．そこで，定理 A.2.1 を利用して汎化誤差を経験誤差で抑えることができる．これより，$C_1 > 0$ が存在し，任意の θ_{priv} に対し，確率 $1 - \alpha/2$ で

$$\begin{aligned}
&\left\{\mathcal{R}(\theta_{\mathrm{priv}}) + \lambda_{\mathrm{reg}} \parallel \theta_{\mathrm{priv}} \parallel_2^2\right\}\\
&- \left\{\mathcal{R}(\theta_{P,\lambda_{\mathrm{reg}}}) + \lambda_{\mathrm{reg}} \parallel \theta_{P,\lambda_{\mathrm{reg}}} \parallel_2^2\right\}\\
&\le \left\{\mathcal{L}(\theta_{\mathrm{priv}}; D^{(m)}) + \lambda_{\mathrm{reg}} \parallel \theta_{\mathrm{priv}} \parallel_2^2\right\}\\
&- \left\{\mathcal{L}(\theta_{n_m,\lambda_{\mathrm{reg}}}) + \lambda_{\mathrm{reg}} \parallel \theta_{n_m,\lambda_{\mathrm{reg}}} \parallel_2^2\right\}\\
&+ C_1\frac{R^2\log(2/\alpha)}{\lambda_{\mathrm{reg}} n_m} \quad \text{(A.5)}
\end{aligned}$$

が成立する．ここで，右辺の最初の 2 項を θ_{priv} によって期待値をとったものは，定理 2.1 (a) より $O\left(\frac{Rp\log p}{\varepsilon n_m}\right)$ で上から抑えられる．

次に，プライバシを考慮しない場合の汎化誤差に相当する項

$$\mathcal{R}(\theta_{n_m,\lambda_{\mathrm{reg}}}) + \lambda_{\mathrm{reg}} \parallel \theta_{n_m,\lambda_{\mathrm{reg}}} \parallel_2^2 - \inf_{\theta\in\Theta}\mathcal{R}(\theta) \quad \text{(A.6)}$$

は，カーネル法に関する一般論（文献 [35]，Theorem 6.24）から，確率 $1 - \alpha/2$ で

$$\begin{aligned}
&A_2(\lambda_{\mathrm{reg}}) + \frac{1}{\lambda_{\mathrm{reg}}}\left(\sqrt{\frac{\log(2/\alpha)}{n_m}}\right.\\
&\qquad \left. + \sqrt{\frac{4}{n_m}} + \frac{8\log(2/\alpha)}{n_m}\right) \quad \text{(A.7)}
\end{aligned}$$

で上から抑えられる．ここで，$A_2(\lambda_{\mathrm{reg}})$ は

$$\begin{aligned}
A_2(\lambda_{\mathrm{reg}}) = \inf_{\theta\in\Theta}&\left\{\lambda_{\mathrm{reg}} \parallel \theta \parallel_2^2\right.\\
&\left. + \mathcal{R}(\theta) - \inf_{\theta'\in\Theta}\mathcal{R}(\theta')\right\} \quad \text{(A.8)}
\end{aligned}$$

で定義される近似誤差関数であって，n_m によらない値である．

以上をまとめると，各組織 m_0 における期待汎化誤差は，確率 $1 - \alpha$ で

$$\begin{aligned}
&\mathbb{E}_\theta[\mathcal{R}(\theta_{\mathrm{priv}})] - \inf_{\theta\in\Theta}\mathcal{R}(\theta) \le A_2(\lambda_{\mathrm{reg}})\\
&+ \frac{1}{\lambda_{\mathrm{reg}}}\left(\sqrt{\frac{\log(2/\alpha)}{n_m}} + \sqrt{\frac{4}{n_m}} + \frac{8\log(2/\alpha)}{n_m}\right)\\
&+ C_1\frac{R^2\log(2/\alpha)}{\lambda_{\mathrm{reg}} n_m} + C_2\frac{Rp\log p}{\varepsilon n_m} \quad \text{(A.9)}
\end{aligned}$$

を満たす．ここで，$C_2 > 0$ は n_m に依存しない定数である．

次に，組織 m_0 は，受け取った学習器 $\{\theta_{\mathrm{priv}}^{(m)}\}$ を mirror averaging により統合して θ_{agg} を構成する．定理 4.1 の不等式 (20) より，$\{\theta_{\mathrm{priv}}^{(m)}\}$ が与えられたもとでは

$$\begin{aligned}
&\mathbb{E}_{(X,Y)}^{n_0}[\mathcal{R}(\hat{\theta}_{\mathrm{agg}})] - \inf_{\theta\in\Theta}\mathcal{R}(\theta)\\
&\le \min_{1\le m\le M}\left\{\mathcal{R}(\theta_{\mathrm{priv}}^{(m)}) - \inf_{\theta\in\Theta}\mathcal{R}(\theta)\right\} + \frac{\beta\log M}{n_0} \quad \text{(A.10)}
\end{aligned}$$

が成立するので，右辺第 1 項の上界の評価を考えればよい．

さて，式 (A.9) の右辺の上界は n_m に関して単調減少である．ここで，要求された上界を求めるためのアイデアは，それぞれの弱学習器 $\theta_{\mathrm{priv}}^{(m)}$ に対する上界のうち，最良のものに対する上界は，各組織に対してデータが均等に配られた場合に達成されるということである．一般性を失わずに $n - n_0$ は M の倍数であると仮定する．式 (A.9) の右辺のうち最小のものを最も大きくするために，$n_1 = n_2 = \cdots = n_M = (n - n_0)/M$ とおく．各組織のデータセット $D^{(m)}$ およびプライバシノイズがすべて独立であるから，式 (A.9) の不等式が $m = 1, \ldots, M$ についてそれぞれ確率 $1 - \alpha/M$ で成立するならば，確率 $(1 - \alpha/M)^M$ ですべての不等式が同時に成立する．したがって，確率 $1 - \alpha$ で

$$\min_{1\le m\le M}\left\{\mathcal{R}(\theta^{(m)}_{\mathrm{priv}})-\inf_{\theta\in\Theta}\mathcal{R}(\theta)\right\}\le A_2(\lambda_{\mathrm{reg}})$$
$$+\frac{1}{\lambda_{\mathrm{reg}}}\left(\sqrt{\frac{M\log(\frac{2M}{\alpha})}{n-n_0}}+\sqrt{\frac{4M}{n-n_0}}+\frac{8M\log(\frac{2M}{\alpha})}{n-n_0}\right)$$
$$+C_1\frac{R^2M\log(\frac{2M}{\alpha})}{\lambda_{\mathrm{reg}}(n-n_0)}+C_2\frac{RpM\log p}{\varepsilon(n-n_0)}\qquad(A.11)$$

ただし，$0<\alpha<1$，$M\ge 1$ のとき $(1-\alpha/M)^M\ge 1-\alpha$ に注意する．

最後に，式 (A.10) の両辺をデータ $\{\theta^{(m)}_{\mathrm{priv}}\}$ の同時分布で期待値をとり，式 (A.11) の結果を代入することによって定理の主張を得る．

A.2　補助的な定理など

A.2.1　確率的凸最適化における有用な不等式

ある確率分布 P に従う確率変数 $Z\in\mathcal{Z}$ と，パラメータ z を持つ凸損失関数 $\ell(\theta,z)$ があるとき，リスク関数を $\mathcal{R}(\theta)=\mathbb{E}_Z[f(\theta,Z)]$ と定める．$\Theta\in\mathbb{R}^p$ での最適化問題 $\min_{\theta\in\Theta}\mathcal{R}(\theta)$ を P からの独立同分布なサンプル $D_n=\{z_1,\dots,z_n\}$ を通して考えることを確率的凸最適化（stochastic convex optimization）という．特に，$g:\mathbb{R}\times\mathcal{Z}\to\mathbb{R}$ を第1引数に関して凸である関数，$r:\Theta\to\mathbb{R}$ を凸関数，$\phi:\mathcal{Z}\to\mathbb{R}^p$ を任意の特徴写像として，損失関数が $\ell(\theta,z)=g(\langle\theta,\phi(z)\rangle,z)+r(\theta)$ と表されるとき，上記の確率的凸最適化問題は一般化線形（generalized linear）であるという．

確率的凸最適化問題において，本来の目的関数 $\mathcal{R}(\theta)$ を経験リスク $\mathcal{L}(\theta;D_n)=n^{-1}\sum_{i=1}^n\ell(\theta,z_i)$ によって評価することは重要である．次の定理は，一般化線形問題においては，そのような評価が一様にできることを述べている．

定理 A.2.1（文献 [22]，Theorem 2）．上の一般化線形問題において，r は λ-強凸であり，ϕ の像は有界であって $\mathbb{R}^p$ の半径 R の球に含まれ，さらに g は第1引数に関して L_g-Lipshitz であるとする．このとき，Z の任意の確率分布 P と，任意の $\delta>0$ に対して，P^n に関して $1-\delta$ 以上の確率で次の不等式が成立する

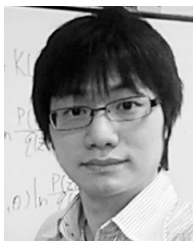
$$\mathcal{R}(\theta)-\inf_{\theta'\in\Theta}\mathcal{R}(\theta')\le 2(\mathcal{L}(\theta;D_n)-\inf_{\theta'\in\Theta}\mathcal{L}(\theta';D_n))$$
$$+O\left(\frac{(RL_g)^2\log(1/\delta)}{\lambda n}\right).\qquad(A.12)$$

南 賢太郎　（学生会員）

2014 年東京大学工学部卒業．2014 年より東京大学大学院情報理工学系研究科修士課程在籍．

荒井 ひろみ

2010 年東京工業大学大学院総合理工学研究科博士課程修了．筑波大学研究員，理化学研究所基礎科学特別研究員を経て，2014 年より東京大学情報基盤センター助教．



佐藤 一誠　（正会員）

2011 年東京大学大学院情報理工学系研究科博士課程修了．2011 年より東京大学情報基盤センター助教．2013 年より科学技術振興機構さきがけ研究員を兼務．統計的機械学習およびデータマイニングの研究に従事．

中川 裕志　（正会員）

1980 年東京大学大学院工学系研究科博士課程修了．1980 年より横浜国立大学工学部勤務．1999 年より東京大学情報基盤センター教授．統計的機械学習の研究に従事．