3B1-R-2-8

オンライン学習におけるデータ適応的正則化手法

Data-incentive Weight Truncation method for sequential learning

大岩 秀和[*1] 松島 慎[*1] 中川 裕志[*2]
Hidekazu Oiwa Shin Matsushima Hiroshi Nakagawa

[*1]東京大学大学院情報理工学系研究科
Graduate School of Informatinal Science and Techonology, The University of Tokyo

[*2]東京大学情報基盤センター
Information Technology Center, The University of Tokyo

We propose new weight-truncation methods for the setting that we receive data sequentially. This method realizes the dynamic adjustment for weight-truncation in accordance with the occurrence counts of data. In the conventional sequential learning, weakly occurrence features are removed from the classifier on a priority basis even if these features are essential for classification. For overcoming this deficit, our proposed algorithms integrate the dynamic adjustment ability inside the regularized term. Moreover, we propose the generalized framework of these truncation adjustment methods. We analyze the upper bound of error between the optimal solution and the derived solution and evaluate the computational cost of the generalized framework. Experimental results show that the proposed methods achieve better performances than the conventional method. Especially, we illustrate that rare features are obtained in our proposed methods in the tasks of natural language processing.

1. はじめに

本研究では，大規模データから学習を行う際により小規模で高精度な予測モデルを構築するため，データの性質を動的に予測モデルに織り込む正則化手法を提案する．

本稿では，教師あり学習における新たな手法を提案する．教師あり学習は入力が与えられた際に適切な出力を返す予測モデルを構築する手法の一つであり，予測モデルの学習には正解データを用いる．正解データは入力と対応する出力が共に明らかとなっているデータの集合で構成される．天候等の統計情報から将来の電力使用量を推定する回帰問題やメールのスパム判定タスク等の分類問題が教師あり学習が対象とするタスクとして代表的である．本研究では，学習や予測に用いるデータ数やデータの次元数が非常に大きいケースに高速かつ省メモリで学習を行うことを目的とする．上記の性質を持つデータの事を本稿では大規模データと呼ぶ．

1.1 オンライン学習

大規模データからの学習を高速かつ省メモリに実現する方法として近年注目を集めている枠組みの一つに，オンライン学習がある．オンライン学習は，訓練データを一つ受け取るたびに逐次的に予測モデルを更新する手法である．オンライン学習では，現在与えられているデータのみを陽に扱い，最適化問題を解くことを目的とする．全データを一度に用いて最良の予測モデルを探索するバッチ学習の枠組みでは，特に全データを一度にメモリに載せることが難しい量の大規模データに対してはデータを部分的にメモリに載せた最適化を繰り返す必要があり，速度及び消費メモリ量の観点から効率的に解を得るには様々な工夫が必要になる．高速化・計算の効率化のための研究は，近年においても多数なされている [Yu 10]．また，データが逐次的にやってくるストリームデータ環境では新たな正解データが集められるたびに全データでの最適化を必要とするため効率が悪い．一方でオンライン学習では元よりデータ一つ一つを逐次的に扱うため，データが大規模化しても上記の問題は発生しにくい．このためバッチ学習の場合と比べ，メモリ消費量の削減と計算の高速化が期待される．また，データが一つ与えられる度に逐次的に予測モデルを構築するため，学習の初期段階から高速に性能の高い予測モデルが得られる．

連絡先: 大岩秀和，東京大学大学院，東京大学文京区本郷 7-3-1 総合図書館 4F，03-5841-2729，03-5841-2745，oiwa@r.dl.itc.u-tokyo.ac.jp

1.2 L_1 正則化

クエリログ解析やスパムメール研究などのストリームデータ環境では，予測に用いられるデータ量も莫大となる．この場合，予測モデルが複雑化し予測速度が低下するのを防ぐ必要がある．高速かつ省メモリな予測器の導出のため，より少数の特徴で精度の高い予測モデルを導出する方法として，L_1 正則化に近年注目が集まっている．L_1 正則化は，損失最小化問題に重みベクトルの L_1 ノルムを加え，それらの総和を最小化する最適化問題を解くことで予測に不要な特徴に関する重みを零化する手法である．予測への貢献度が低い特徴を予測モデルから排除し，予測器のコンパクト化を促進する．L_1 正則化を適用することで，学習・予測の高速化と省メモリ化が期待される．また，予測モデルが簡略化されるため，解釈も容易となる．

1.3 L_1 正則化項付きオンライン学習

ストリームデータ等の大規模なデータから効率的に学習や予測を行うため，オンライン学習と L_1 正則化を組み合わせた最適化手法が近年盛んに研究されている．L_1 正則化項付きオンライン学習の最適化手法として，

- Composite Objective MIrror Descent (COMID) [Duchi 10]
- Regularized Dual Averaging (RDA) [Xiao 10]
- Follow-The-Proximally-Regularized-Leader (FTPRL) [McMahan 10b]

の三種類の代表的な枠組みが提案されている．これらの手法では，計算のさらなる高速化のため与えられたデータに関する損

表 1: 記号の定義

a	スカラー
$\mathbf{a}$	ベクトル
$\mathbf{A}$	行列
$a^{(i)}$	ベクトルの第 i 成分
$A^{(i,j)}$	行列の第 i,j 成分
$\|\lambda\|$	スカラーの絶対値
$\|\mathbf{a}\|_p$	L_p ノルム
$\langle \mathbf{a}, \mathbf{b} \rangle$	ベクトルの内積

失関数を陽には用いず，損失関数の一次近似となる劣勾配を用いて予測モデルの更新を行う[*1]．これらの手法は全て，最適解における目的関数値と毎回の最適化で得られる目的関数値のデータ単位での差の上限は，データ数を増やすごとに減少し零に収束することが示されている．この事実から，これら既存手法はバッチ学習で求められる最適解と同等のパフォーマンスを示す事が証明される．上記の解析手法はリグレット解析と呼ばれる．また実験的にも，大規模なデータセットから省メモリな環境で高速に学習可能であり，高精度かつコンパクトな予測モデルが結果として得られることが知られている．

1.4 本稿の概要

これら既存の L_1 正則化項付きオンライン学習手法は，学習および予測の高速化，省メモリ化を実現することで大規模データを容易に扱うことを可能にした．しかし，これら既存手法は予測に用いられる可能性のある特徴の出現頻度が不均一なデータから学習を行う時，低頻度特徴が優先的に予測モデルから排除される．このため，予測に有用な特徴であっても低頻度特徴であれば零化される可能性が非常に高くなる．

本研究では特徴の出現頻度からの影響を動的に補正し，予測に対する重要度による特徴選択を実現する新たな正則化手法を提案する．提案手法は，損失関数の劣勾配情報を正則化項に組み込み，L_1 正則化項をデータが与えられる度変形させる．この改良により，特徴の出現頻度情報が既知でなくとも低頻度特徴が零化されやすくなる影響を自動的に補正し，より予測に重要な特徴を予測モデルに組み入れる事が可能になる．この提案手法は既存手法と同等の計算速度が保証される．さらにリグレット解析により，提案手法は最適解と同等のパフォーマンスを示す事も保証される．最後に実データを用いた評価実験を行い，低頻度だが予測に有用な特徴をモデルに組み入れる事を実現できていること，その結果としてよりコンパクトかつ高精度な予測モデルを導出できる事を確認する．以降の章では，既存手法のうち RDA のみに焦点をおいて議論を行うが，COMID/FTPRL においても同様の性質が成立する．

2. 定式化：正則化項付きオンライン学習

本稿で使用する記号の定義を表 1 に列挙する．

本稿では，正則化項付きオンライン学習による教師あり学習を取り扱う．教師あり学習の目的は，入力 $\mathbf{x}$ から対応する出力 y を適切に予測する関数を，入力 $\mathbf{x}$ と対応する出力 y のペアで構成される正解データの集合 $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_T, y_T)\}$ を用いて設計する事である．本稿では出力を予測する関数を線形予測器に限定する．そのため，入力から出力への関数を入力ベクトルと重みベクトル $\mathbf{w} \in \Re^n$ の内積で定義する．従って，出力データの予測値 $\hat{y}$ は $\langle \mathbf{w}, \mathbf{x} \rangle$ で計算される．

正則化項付きオンライン学習による教師あり学習では，正解データが一つ与えられるたびに重みベクトル $\mathbf{w}$ が逐次的に更新される．更新は以下の枠組みに従って行われる．

1. t 番目の入力データ $\mathbf{x}_t$ を受け取る
2. 現在の重みベクトル $\mathbf{w}_t$ と入力ベクトル $\mathbf{x}_t$ の内積より，出力の予測値 $\hat{y}_t$ を求める
3. 予測値 $\hat{y}_t$ と真の出力値 y_t を用いて，重みベクトルを $\mathbf{w}_{t+1}$ に更新する
4. 次の正解データが存在する場合，1. に戻る

重みベクトルの評価は，損失関数 $\ell_t(\cdot) : \Re^n \to \Re_+$ と正則化項 $\Phi_t(\cdot) : \Re^n \to \Re_+$ の重み付け和で定められる．教師あり学習はこの評価値の総和を最小化する事を目標とする．

損失関数 ℓ_t は，予測値 $\hat{y}_t$ と真の値 y_t との乖離度に応じて，値が大きくなる関数である．本稿で扱う損失関数は

$$\ell_t(\mathbf{w}) = \hat{\ell}_t(\langle \mathbf{w}, \mathbf{x}_t \rangle; y_t) = \hat{\ell}_t(\hat{y}_t; y_t) \,. \tag{1}$$

を満たす関数 $\hat{\ell}(\cdot)$ が存在する損失関数に限定する[*2]．(1) 式が成立するとき損失関数の重みベクトルに関する勾配は，$\mathbf{x}_t$ のスカラー倍で表すことが出来る．さらに，損失関数は凸性を持つ関数に限定する．ここで，凸性を持つ関数とは (2) 式を満たす関数 f のことである．

$$\forall \mathbf{a}_1, \mathbf{a}_2 \in \Re^n \quad \forall \lambda \in [0,1]$$
$$\lambda f(\mathbf{a}_1) + (1-\lambda) f(\mathbf{a}_2) \geq f\left(\lambda \mathbf{a}_1 + (1-\lambda)\mathbf{a}_2\right) \,. \tag{2}$$

ヒンジ損失関数 $\ell_t(\mathbf{w}) = [1 - y_t \langle \mathbf{w}, \mathbf{x}_t \rangle]_+$，二乗損失関数 $\ell_t(\mathbf{w}) = (y_t - \langle \mathbf{w}, \mathbf{x}_t \rangle)^2$ は，これらの制約を全て満たす損失関数である．

損失関数 ℓ_t における劣勾配の定義は，(3) 式を満足するベクトル $\mathbf{g} \in \Re^n$ で表現される．

$$\forall \mathbf{y} \quad \ell_t(\mathbf{w}_t) \geq \ell_t(\mathbf{y}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{w}_t \rangle \,. \tag{3}$$

ヒンジ損失関数のように微分不可能な点が存在しても関数が凸性を持っていれば，全ての点で劣勾配は存在する．

正則化項 Φ は，重みベクトルにスパース化や汎化性能の向上等，ある種の構造を導入する際に用いる．特に，重みベクトルを疎な形に直し予測モデルのコンパクト化を促進する，L_1 正則化と呼ばれる手法がある．L_1 正則化項は，重みベクトルの L_1 ノルムによって定義される．式で表すと，以下の通りである．

$$\Phi(\mathbf{w}) = \lambda \|\mathbf{w}\|_1 \,. \tag{4}$$

ここで，λ は最適化問題において損失関数と正則化項の重要度を調節するパラメータである．

正則化項付きオンライン学習における教師あり学習の目的は，学習過程で生じる損失関数と正則化項の総和を最小化することである．これを式で表すと，

$$\min_{\mathbf{w}_1, \mathbf{w}_2, \ldots} \sum_t \left(\ell_t(\mathbf{w}_t) + \lambda \Phi_t(\mathbf{w}_t)\right) \tag{5}$$

となる．この目的関数を最小化するように重みベクトルを逐次的に更新することがオンライン学習の目標となる．このような重みベクトルを求める最適化手法として RDA などのアルゴリズムが提案されている．

*1 各回の損失関数を陽に扱い最適化問題を解く方法 (Implicit Update) も議論されている [McMahan 10a] が，本稿では損失関数の劣勾配のみを用いた手法に限定して議論する．

*2 損失関数は予測値 $\hat{y}_t$ と真の値 y_t にのみ値が依存する．

2.1 RDA

RDA では，正解データが一つ与えられるたびに (6) 式に基づいて重みベクトルを更新する．

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \sum_{\tau=1}^{t} \langle \mathbf{g}_\tau, \mathbf{w} \rangle + t\Phi(\mathbf{w}) + \beta_t h(\mathbf{w}) \ . \quad (6)$$

(6) 式から明らかなように，RDA では三種類の項の総和を最小化するように，毎回重みベクトルを更新するアルゴリズムである．

第一項は重みベクトルとこれまでの劣勾配を総和したベクトルとの内積で表現される．ここで，$\mathbf{g}_t$ は，t 個目のデータに対する $\mathbf{w}_t$ における損失関数の劣勾配を表す．劣勾配は各正解データに対して損失関数を最大化する方角を示すベクトルである．従って，劣勾配との内積を最小化する事は，損失関数を最小化する方角の重みベクトルを導出することになる．第二項は，正則化項 Φ を正解データの数だけ総和した項である．第三項は初期点との距離を測る項になる．ここで，$\{\beta_t\}_{t\geq 1}$ は非減少な正値の数列で，アルゴリズムの収束速度に影響を与える重要な値の列になる．$h(\cdot) : \Re^n \to \Re_+$ は，重みベクトルの初期点との Bregman 距離の性質を満たす関数を利用する．

正解データが新しく与えられるたび，現在の重みベクトルを用いて新たな劣勾配を導出し，(6) 式に従って重みベクトルを更新する操作を行うアルゴリズムが RDA である．例として，重みベクトルにスパースな構造を導入するため正則化項に L_1 正則化項，Bregman 距離の項に重みベクトルの二乗ノルムを用いた時，重みベクトルの更新式は閉じた形で求める事が可能である．この時，入力ベクトルの非零要素数に線形の速度で毎回の重みベクトル更新を実現できる．

3. 提案手法：データ適応的正則化

正則化項付きオンライン学習の既存手法は，特徴の出現回数が不均一なデータ集合に対して低頻度特徴が零化されやすくなる性質を持つ事を先に述べた．この章では，はじめに L_1 正則化項月オンライン学習手法の一つである RDA を例に挙げ，既存手法では低頻度特徴が必要以上に零化されやすい特性を持つ事を示す．その後，この性質を補正するため特徴頻度の不均一性を動的に調節する新たな正則化手法を提案する．

3.1 特徴の出現頻度の不均一性

特徴 A の出現頻度が 1/100，特徴 B の出現頻度が 1/2 のデータセットから，一般的な L_1 正則化項を導入した RDA で学習を行うと仮定する．正則化項 $\Phi(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$，近接項 $h(\mathbf{w}) = \|\mathbf{w}\|_2^2/2$，また $\beta_t = 1/\sqrt{t}$ と定義する．この条件下でパラメータ更新を行った時，特徴 A に対応する次元の劣勾配の値が非零である数値を集めた時，平均値が 100λ を超えていない場合，特徴 A に対応する重みは必ず 0 になる．一方で，特徴 B では平均値が 2λ を超えていれば，零化されるとは限らない．したがって，特徴間で頻度が大きく異なるデータから学習を行う場合，低頻度特徴は予測に有用であっても予測モデルから排除されやすい．自然言語処理やパターン認識で教師あり学習が対象とするタスクでは特徴の出現頻度が不均一なデータが多く，この性質はアルゴリズムの精度に強い影響を与える．

3.2 データ適応的正則化

上記の問題が発生するのは，既存手法では特徴の出現頻度や重みの更新頻度とは無関係に全特徴に共通の零化を施しているためである．この作用により，予測に有用な特徴であっても出現頻度が低ければ自動的に予測モデルから排除される (図 1)．

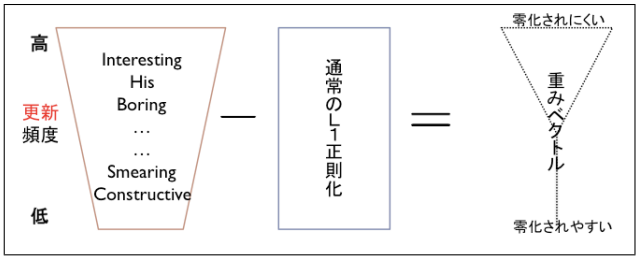

図 1: 通常の L_1 正則化

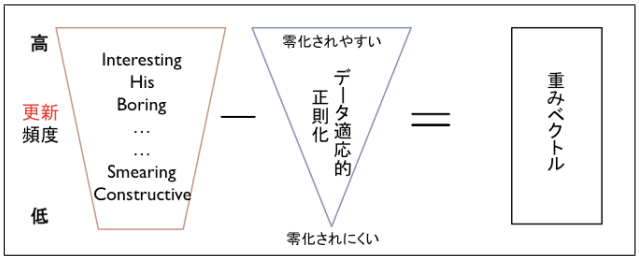

図 2: 特徴の更新頻度に応じた L_1 正則化

稀な特徴の重みを強調するように前処理を施す TF-IDF 等の手法が既存手法として提案されている．このように，予測モデルを構築する際に，特徴の出現頻度は重要な情報となりうる．しかし，データの出現頻度の不均一性を補正するために前処理を施すには一度全データを走査して全特徴の出現回数を算出しなければならない．そのため，データを一度走査するだけで学習が可能なオンライン学習の長所を失ってしまう．また，ストリームデータのように時系列とともにデータの性質が変化する環境下では，前処理によって特徴の出現頻度を補正しても，その後，計算速度を落とさず，収束性を満たす条件を破らず，継続的に頻度補正を加える事は困難である．動的に特徴の出現頻度を補正する手法としては，FOBOS[*3]への拡張となる正則化手法として [大岩 11] が提案されている．本提案手法は [大岩 11] の FOBOS に対する拡張を，RDA や FTPRL に適応可能な形で一般化した正則化モデルになる．

本研究では，損失関数の劣勾配の更新情報を用いて，データの性質に応じた零化調節が動的に実現する手法を提案する．本提案手法では，損失関数の劣勾配の更新情報を L_1 正則化項に組み込み，正則化項にデータの性質を導入する (図 2)．この拡張により，予測に重要であっても特徴の出現頻度が低いために零化されていた重みが優先的に排除される現象を防ぐことが出来る．同時に，重要性の低い頻出特徴の予測モデルからの排除も可能になる．本研究の最大の貢献は，重みの更新頻度に応じた L_1 正則化を実現するための一般的な数理モデル化を示し，それらの適用可能性，理論解析そして実データを用いた実験によるパフォーマンス性能の比較検証を行った事である．

4. 提案手法の定式化

重みの更新頻度に応じた L_1 正則化を実現するため，損失関数の劣勾配の情報を正則化項に導入する．劣勾配情報を組み込んだ行列 $\mathbf{R}_{t,q}$ を定義し，その行列をノルムの内側に導入する．式として表すと，L_1 正則化項が (7) 式の形に変形する．

$$\Phi_t(\mathbf{w}) = \lambda\|\mathbf{R}_{t,q}\mathbf{w}\|_1 \ , \quad (7)$$

*3 FOBOS は COMID の特殊系である．

ここで，行列 $\mathbf{R}_{t,q}$ は損失関数の劣勾配を用いて以下のように定義する．

$$\mathbf{R}_{t,q} = \begin{pmatrix} r_{t,q}^{(1)} & 0 & \dots & 0 \\ 0 & r_{t,q}^{(2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_{t,q}^{(d)} \end{pmatrix}$$

$$s.t.\ \ r_{t,q}^{(i)} = \sqrt[q]{\sum_{\tau=1}^{t} \left| g_{\tau}^{(i)} \right|^{q}}\ .$$

$r_{t,q}^{(i)}$ の項は，各データの劣勾配の第 i 成分を並べたベクトルの q ノルムで定義され，重みの更新頻度に対して正の相関を持つ．従って，更新頻度に応じた L_1 正則化が実現される．

既存手法の COMID/RDA/FTPRL 全てに提案手法のモデル化を適用可能である．本提案手法は計算コスト，リグレットの面で既存手法と同等の性能を示す事を理論的に示せる[*4]．

5. 実験評価

最後に分類タスクを用いた実データ実験を提案手法に適用し，正則化拡張によるパフォーマンス性能への効果を検証した．実験には，Amazon.com のデータセット [Blitzer 07] である四種類の評価分類タスク (books，dvd，electronics，kitchen) とニュース記事のカテゴリ分類タスクである 20 NewsGroups[Lang 95](news20) データセットの二種類のサブセット (ob-2-1, sb-2-1) を用いた．λ の選択のため 10 分割の交差検定を用い，20 回反復計算を行い予測モデルを構築した．

実験結果より，提案手法は既存手法と比べて，大部分のデータセットでより高精度かつコンパクトな予測モデルが得られている事を確認した (表 2)．次に，これらのデータセットで学習された予測モデルを比較し，保持された特徴の違いを検証した．各手法で構築された予測モデルにおいて，重要と判定された語句を表 3 に示した．この結果から，本研究で提案したデータ適応的な正則化は特徴の出現頻度の影響を低減し，低頻度であっても予測に重要な特徴を捉えられている事を確認した．これらの実験結果から，提案手法において低頻度かつ予測に有用な特徴が保持できる事が示される．また，提案手法がより高精度でコンパクトな解を導出している事が示される．

表 2: 各分類手法の正識別率 [%] (重みの零率 [%])

	データ適応的 RDA ($q = \infty$)	RDA
books	**86.50** (**91.13**)	86.00 (88.69)
dvd	**86.31** (**94.74**)	85.58 (93.23)
electronics	**89.30** (88.93)	88.94 (**91.93**)
kitchen	**90.95** (**97.49**)	90.23 (91.23)
ob-2-1	96.20 (**82.30**)	**96.90** (76.96)
sb-2-1	**98.80** (**93.69**)	97.70 (89.74)

表 3: 各手法で判定された重要語句．() 内は出現回数．

データ適応的 RDA ($q = \infty$)	RDA
”some interesting” (117)	”his” (1491)
”a constructive” (29)	”more” (877)
”be successful” (64)	”time” (1161)
”was blatantly” (106)	”almost” (376)
”smearing” (30)	”say” (2407)

6. おわりに

本研究では，予測に有効な低頻度特徴が予測モデルから排除されやすい既存の L_1 正則化付きオンライン学習手法の問題点を解決するため，重みの更新頻度に応じて動的に変形するデータ適応的な L_1 正則化手法を提案した．さらに，提案手法の計算コスト・収束性を証明し，既存手法と同等の性質を持つことを確認した．最後に実データを用いた実験を行い，提案手法が既存手法よりも効率的な学習が可能なことを示した．

参考文献

[Blitzer 07] Blitzer, J., Dredze, M., and Pereira, F.: Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification, in *ACL*, pp. 440–447, Association for Computational Linguistics (2007)

[Duchi 10] Duchi, J. C., Shalev-Shwartz, S., Singer, Y., and Tewari, A.: Composite Objective Mirror Descent, in *COLT*, pp. 14–26, Omnipress (2010)

[Lang 95] Lang, K.: Newsweeder: Learning to filter netnews, in *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 331–339 (1995)

[McMahan 10a] McMahan, H. B.: A Unified View of Regularized Dual Averaging and Mirror Descent with Implicit Updates, *CoRR*, Vol. abs/1009.3240, (2010)

[McMahan 10b] McMahan, H. B. and Streeter, M. J.: Adaptive Bound Optimization for Online Convex Optimization, in *COLT*, pp. 244–256, Omnipress (2010)

[Xiao 10] Xiao, L.: Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization, *Journal of Machine Learning Research*, Vol. 11, pp. 2543–2596 (2010)

[Yu 10] Yu, H.-F., Hsieh, C.-J., Chang, K.-W., and Lin, C.-J.: Large linear classification when data cannot fit in memory, in *KDD*, pp. 833–842, ACM (2010)

[大岩 11] 大岩 秀和, 松島 慎, 中川 裕志：特徴の出現回数に応じた L1 正則化を実現する教師ありオンライン学習手法, 情報処理学会論文誌数理モデル化と応用（TOM）, Vol. 4, No. 3, pp. 84–93 (2011)

∗4 分量の制約から証明は省く．