Vol.2020-MPS-127 No.20
2020/3/3

河川水中病原体予測のための符号制約 SVM と双対学習算法

土田 康平[1,a)]　田島 賢哉[1]　佐野 大輔[2]　加藤 毅[1,3,b)]

概要：線形識別器の訓練において，ある特徴に正の相関があるというドメイン知識がある場合，訓練用データの個数が十分にあれば，対応する重み係数は正になることが好ましい．しかし，訓練用データ数が不十分，もしくは，データにノイズが多いとき，対応する重み係数が負に学習されてしまうことがある．我々は重み係数の符号を制約して線形識別器を学習する方法を考案し，河川水中病原体の予測への応用において有効性を確認してきた．本稿は，符号制約下で SVM を学習するための新しい最適化アルゴリズムを提案する．本研究で開発したアルゴリズムは，フランクウルフ法に基づいており，次の 3 点の長所を持つ：(i) 劣線形収束する; (ii) 各反復の計算コストは $O(nd)$; (iii) 停止条件が明確．すなわち，射影勾配法と同等の計算時間を持ちながら，さらに，反復を停止したときの解の精度を保証する算法となる．これらの理論保証は公開データセットを使った数値例で例示し，有効性を示す．

1. はじめに

線形サポートベクトルマシン (SVM) $\langle \boldsymbol{w}, \cdot \rangle$ のモデルパラメータ $\boldsymbol{w} := [w_1, \dots, w_d]^\top \in \mathbb{R}^d$ の学習タスクにおいて，高い汎化能力を得るには，十分な個数の訓練用例題が必要である．しかし，応用によっては，十分な訓練用例題が得られない場合がしばしばある．特に，医学や生物学などにおいて，1 個のデータ点を得るのに，高価な試薬や，少なからぬ労力を要するような応用は珍しくない [2], [5], [8]．訓練用例題の個数の不足による訓練精度の低下を抑制するアプローチとしては，事前知識の活用が有用である．

応用分野によっては，ドメイン知識として，ある説明変量 x_h が目的変量 y と正の相関があることが分かっている場合がある．負の相関があると分かっている場合もある．正の相関があることが分かっている場合，対応するモデルパラメータ w_h は正になるべきである．負の相関があることが分かっている場合，w_h は負になるべきである．しかし，訓練用例題が十分ではなく，相関が弱い場合は，対応するモデルパラメータは，既知の相関の符号と逆の符号に学習されてしまうことがしばしばおこり，その結果，そのモデルパラメータが正しい予測を妨げてしまう [5]．

本論文では，一部の説明変量と目的変量との相関の符号があらかじめわかっているときに，そのドメイン知識を SVM 学習に組み込むことができる，新しい学習アルゴリズムを提案する．SVM 学習とは，後述する正則化経験リスク $P(\boldsymbol{w})$ を最小にするモデルパラメータの値 $\boldsymbol{w}$ を見つけることである．従来の SVM 学習のアルゴリズムとしては，**ペガサス法** [6] が有名である．ペガサス法は主問題を**勾配法**で解く方法である．このアルゴリズムは，最適値に**劣線形収束** [1] することが示されている．すなわち，$P(\boldsymbol{w}) - P(\boldsymbol{w}_\star) \le \epsilon$ に達するまでに必要な反復数が $O(1/\epsilon)$ であることが保証されている．ただし，$\boldsymbol{w}_\star$ は最適解である．また，各反復の計算量は $O(nd)$ でおさまる．ただし，d は説明変量の個数，n は訓練用例題数である．よって，ペガサス法は，理論的にも実用的にも使用しやすい最適化アルゴリズムとなっていた．

本研究では，$P(\boldsymbol{w})$ を符号制約下で最小化するアルゴリズムを探求してきた．我々は，これまでにペガサス法をベースにした最適化アルゴリズムとして，**符号制約ペガサス法**を開発していた [9]．符号制約ペガサス法は，従来のペガサス法の各反復において，**実行可能領域**への射影を行うステップを挿入したものになっている．すなわち，**射影勾配法** [1] である．射影ステップの計算量は $O(d)$ で済むことから，各反復の計算量は $O(nd)$ を維持する．さらに，射影ステップを挿入しても，最適値に劣線形収束することを示した．しかし，符号制約ペガサス法は従来のペガサス法と同じ短所を抱えている．それは，解が **ϵ-最適解**に達したか判断できないことである．すなわち，最小値 $P(\boldsymbol{w}_\star)$ が不明なため，**目的誤差**が $P(\boldsymbol{w}) - P(\boldsymbol{w}_\star) \le \epsilon$ に達したか判定できない (図 1(a))．

本稿では，最適解への劣線形収束が保証されており，かつ，解が ϵ-最適解に達したか判定するための明確な停止条

1 群馬大学
2 東北大学
3 早稲田大学
a) tsuchida-kouhei@kato-lab.cs.gunma-u.ac.jp
b) katotsu@cs.gunma-u.ac.jp

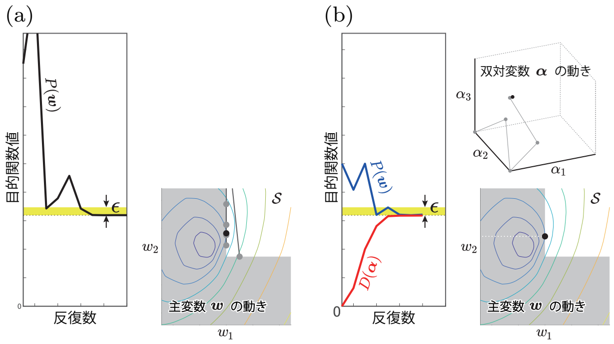

図 1 符号制約ペガサス法と提案法の違い．(a) 符号制約ペガサス法は主問題に対する最適化を行うので，目的誤差 $P(\boldsymbol{w}) - P(\boldsymbol{w}_\star)$ が ϵ 以下に収束したか判定することができない．(b) 提案法では，双対問題に対して最適化を行うので，目的誤差の上限となる双対ギャップ $P(\boldsymbol{w}(\boldsymbol{\alpha})) - D(\boldsymbol{\alpha})$ はモニタリング可能である．双対ギャップが ϵ 以下になったときに反復を停止すれば，目的誤差 $P(\boldsymbol{w}(\boldsymbol{\alpha})) - P(\boldsymbol{w}_\star)$ も ϵ 以下になることが保証される．

件を利用できる最適化アルゴリズムを提案する．提案する最適化アルゴリズムは，SVM 学習の主問題を符号制約下で直接解くのではなく，**双対問題**を解く方法である．双対問題を解くために，**フランクウルフ法** [3], [4] を採用し，その枠組みから逸脱しないように開発した．これによって，本稿で提案する最適化アルゴリズムでも，フランクウルフ法が保証する劣線形収束の性質を継承することになる．

反復数を低く抑える理論保証があったとしても，各反復の計算が重ければ実用的なアルゴリズムにならない．フランクウルフ法の各反復は，**方向探索ステップ**と**線探索ステップ**の 2 ステップからなる．本稿におけるもっとも大きな発見は次の 2 点からなる：

- 符号制約下での SVM 学習問題の双対問題を解くフランクウルフ法において，方向探索ステップは計算量 $O(nd)$ で計算できる；
- 線探索ステップの計算量も $O(nd)$ におさまる．

すなわち，フランクウルフ法における各反復は $O(nd)$ の計算量で済むことになる．よって，本稿で提案する最適化アルゴリズムは我々がこれまでに開発していた符号制約ペガサス法と同レベルの計算量と反復数で最適解に収束させることが出来る．加えて，明確な停止条件を持つアルゴリズムになっている (図 1(b)) ため，停止条件が不明確という符号制約ペガサス法の短所は克服される．

2. 従来のサポートベクトルマシン

従来の線形 SVM は，線形識別器 $\langle \boldsymbol{w}, \cdot \rangle$ を作り出す．線形識別器は，未知データ $\boldsymbol{x} \in \mathbb{R}^d$ に対して，$\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ が 0 以上ならば陽性と予測し，$\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ が 0 以下ならば陰性と予測する．閾値としては，0 以外の数値を使うこともある．

例えば，水質データから大腸菌数を予測する場合，説明変量として，水温 (略称 WT)，電気伝導度 (EC)，懸濁物質含有量 (SS)，生物化学的酸素要求量 (BOD)，全窒素量 (TN)，全リン量 (TP)，水素イオン指数 (pH)，溶存酸素 (DO)，流量 (FR) を用いることができる．このうち，水素イオン指数は，$\mathrm{pH}_+ := \max(0, \mathrm{pH} - 7)$ および $\mathrm{pH}_- := \max(0, 7 - \mathrm{pH})$ の 2 変量に分解すると中性からどれほど離れているか表現できるようになる [5]．また，バイアス項を表現するため，常に 1 の値を加えるとすると，特徴ベクトル $\boldsymbol{x}$ は $d = 11$ 次元のベクトルになる．大腸菌が水中に一定以上存在するか否か予測する問題では，特徴ベクトル $\boldsymbol{x}_i \in \mathbb{R}^d$ および水中に一定以上存在するか否かを表すクラスラベル $y_i \in \{\pm 1\}$ の対 $(\boldsymbol{x}_i, y_i)$ からなる訓練用例題の集合 $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{\pm 1\}$ から $\boldsymbol{w} \in \mathbb{R}^d$ の値を決定する．

従来の SVM は次のように定式化されている：

$$\begin{aligned} &\min \quad \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i \quad \text{wrt} \quad \boldsymbol{w} \in \mathbb{R}^d, \boldsymbol{\xi} \in \mathbb{R}^n, \\ &\text{subj to} \quad \forall i \in [n], \quad y_i \langle \boldsymbol{x}_i, \boldsymbol{w} \rangle \geq 1 - \xi_i, \quad \xi_i \geq 0. \end{aligned} \tag{1}$$

ただし，$\lambda > 0$ は正則化定数と呼ばれる定数である．式 (1) のような表現は，**2 次計画問題**の標準形に近いため，SVM が 2 次計画問題の範疇にあることを示すためには便利な表現である．**統計的学習理論** [7] の視点から見ると，SVM は，次式に定義する**正則化経験リスク** $P(\boldsymbol{w})$ を最小化する方法論と見るほうが理解しやすい：

$$P(\boldsymbol{w}) := \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y_i \langle \boldsymbol{x}_i, \boldsymbol{w} \rangle). \tag{2}$$

制約付き最適化問題 (1) と正則化経験リスク $P(\boldsymbol{w})$ の最小化という制約なし最適化問題の最適解が等しいことは容易に示すことができる．

3. 符号制約サポートベクトルマシン

符号制約の下で SVM 学習を行う方法を**符号制約** *SVM* と呼ぶ．符号制約は，目的変量との間の相関関係が既知の説明変量に対する係数 w_h に課す制約とする．

例えば，前述の水質データから大腸菌数を予測するタスクでは，水温，電気伝導度，懸濁物質含有量，生物化学的酸素要求量，全窒素量，全リン量は大腸菌数と正の相関があることが水質工学においては周知の事実として知られている [5]．また，pH_+，pH_-，溶存酸素，流量は大腸菌数と負の相関があることが知られている．このような情報を捨てずに SVM 学習に有効利用するために符号制約を課す．大腸菌数と正の相関のある説明変量の添え字集合を $\mathcal{I}_+$，大腸菌数と負の相関のある説明変量の添え字集合を $\mathcal{I}_-$ とすると，それらに対応する係数 w_h には，

$$\forall h \in \mathcal{I}_+, \; w_h \geq 0, \quad 及び \quad \forall h' \in \mathcal{I}_-, \; w_{h'} \leq 0 \tag{3}$$

を課すことにする．ここで，$\mathcal{I}_-$ に属する説明変量の値を，

あらかじめ，符号を逆転させてしまうことにより，$\mathcal{I}_- \cup \mathcal{I}_+$ を改めて $\mathcal{I}_+$ とし，$\mathcal{I}_-$ は空集合としてよいことになる．以降，そのような前処理を施すことを仮定し，非負制約 $w_h \geq 0$ のみを扱うことにする．各要素を次のようにおいた定数 $\boldsymbol{\sigma} := [\sigma_1, \ldots, \sigma_d]^\top \in \{0,1\}^d$ を導入する：

$$\sigma_h := \begin{cases} 1 & \text{for } h \in \mathcal{I}_+, \\ 0 & \text{for } h \in \mathcal{I}_0 := [d] \setminus \mathcal{I}_+. \end{cases} \tag{4}$$

すると，符号制約は $\boldsymbol{\sigma} \odot \boldsymbol{w} \geq \boldsymbol{0}_d$ のように簡潔に表現できる．ただし，演算子 $\odot$ はアダマール積を表す．符号制約 SVM は実行可能領域

$$\mathcal{S} := \left\{ \boldsymbol{w} \in \mathbb{R}^d \,\middle|\, \boldsymbol{\sigma} \odot \boldsymbol{w} \geq \boldsymbol{0}_d \right\} \tag{5}$$

の中から正則化経験リスク $P(\boldsymbol{w})$ を最小にする $\boldsymbol{w}$ を見つける．すなわち，次のような最適化問題として書き表すことができる：

$$\min \quad P(\boldsymbol{w}) \quad \text{wrt} \quad \boldsymbol{w} \in \mathcal{S}. \tag{6}$$

我々が以前開発した符号制約ペガサス法 [9] では，従来のペガサス法の各反復において，解 $\boldsymbol{v}$ を実行可能領域に射影する，すなわち，

$$\Pi_{\mathcal{S}}(\boldsymbol{v}) := \underset{\boldsymbol{w} \in \mathcal{S}}{\operatorname{argmin}} \|\boldsymbol{v} - \boldsymbol{w}\| = \boldsymbol{v} + \max(\boldsymbol{0}, -\boldsymbol{\sigma} \odot \boldsymbol{v}) \tag{7}$$

を施すというステップを挿入することによって実現した．1 節で述べたように，符号制約ペガサス法は劣線形収束の理論保証は得られているが，最適化問題 (6) の最小値 $P(\boldsymbol{w}_\star)$ が分からないため，現在の解 $\boldsymbol{w}$ に対する目的誤差 $P(\boldsymbol{w}) - P(\boldsymbol{w}_\star)$ が十分小さくなったか判定することができないという短所を患っていた．

4. 双対問題

本研究で新たに開発したアルゴリズムは双対問題 [1] を介して符号制約 SVM を学習するというアプローチを採用した．符号制約 SVM の学習問題 (6) の双対問題は，以下のように与えられる：

$$\begin{aligned} &\max \quad D(\boldsymbol{\alpha}) \quad \text{wrt} \quad \boldsymbol{\alpha} \in [0,1]^n, \\ &\text{where} \quad D(\boldsymbol{\alpha}) := -\frac{\lambda}{2}\|\boldsymbol{w}(\boldsymbol{\alpha})\|^2 + \frac{1}{n}\langle \boldsymbol{1}, \boldsymbol{\alpha} \rangle \\ &\qquad \boldsymbol{w}(\boldsymbol{\alpha}) := \Pi_{\mathcal{S}}\left(\frac{1}{\lambda n}\boldsymbol{X}\boldsymbol{\alpha}\right). \end{aligned} \tag{8}$$

ただし，行列 $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ の第 i 列には，第 i 例題の特徴ベクトル $\boldsymbol{x}_i \in \mathbb{R}^d$ にクラスラベル $y_i \in \{\pm 1\}$ をかけた値 $y_i \boldsymbol{x}_i$ が格納されているとする．符号制約を課さない場合は $\mathcal{S} = \mathbb{R}^d$ となり，$\Pi_{\mathcal{S}}(\boldsymbol{v}) = \boldsymbol{v}$, $\boldsymbol{w}(\boldsymbol{\alpha}) = \boldsymbol{X}\boldsymbol{\alpha}/(\lambda n)$ になることから，これらを (8) に代入すると，従来の SVM の双対問題が復元できることが分かる．よって，従来の SVM と符号制約 SVM の双対問題における差異は，実行可能領域への射影の有無といえる．従来の SVM の双対目的関数は，単純な 2 次関数であった．符号制約 SVM の双対目的関数 $D(\cdot)$ には，実行可能領域への射影が含まれていることで，従来の SVM の双対問題を解くためのアプローチが直接利用できなくなり，非自明な問題となっている．

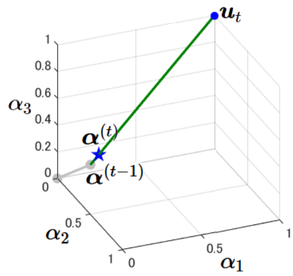

図 2　フランクウルフ法の 1 反復．フランクウルフ法では，第 t 反復において，現在の解 $\boldsymbol{\alpha}^{(t-1)}$ のまわりで双対目的関数 $D(\cdot)$ を線形近似した関数を目的関数とした部分問題を考え，その部分問題の最適解 $\boldsymbol{u}_t$ を求める．次に 2 点 $\boldsymbol{\alpha}^{(t-1)}$ と $\boldsymbol{u}_t$ を結ぶ線分上で最も双対目的関数の値が大きくなる点 $\boldsymbol{\alpha}^{(t)}$ に解を移動させる．

双対問題 (8) の最適解を $\boldsymbol{\alpha}_\star$ とすると，主問題 (6) の最適解は $\boldsymbol{w}_\star = \boldsymbol{w}(\boldsymbol{\alpha}_\star)$ を満たす．よって，双対問題 (8) の最適解を求めたのちに，$\boldsymbol{w}(\cdot)$ を介して最適な主変数 $\boldsymbol{w}_\star$ を復元すればよいことになる．

主問題を最小化するのではなく，双対問題を最大化することによって，明確な停止条件を得ることができる．双対目的関数 $D(\cdot)$ の性質 [1]

$$P(\boldsymbol{w}(\boldsymbol{\alpha})) - P(\boldsymbol{w}_\star) \leq P(\boldsymbol{w}(\boldsymbol{\alpha})) - D(\boldsymbol{\alpha}) \tag{9}$$

から，目的誤差 $P(\boldsymbol{w}) - P(\boldsymbol{w}_\star)$ が ϵ 以内の解を得るためには，**双対ギャップ**が $P(\boldsymbol{w}(\boldsymbol{\alpha})) - D(\boldsymbol{\alpha}) \leq \epsilon$ に達した時に反復を停止させればよい．

5. フランクウルフ法の適用

本節では，符号制約 SVM の双対問題を解くためのアルゴリズムを提案する．本研究では，前節で定義した $D(\boldsymbol{\alpha})$ を最大化するためフランクウルフ法を採用した．フランクウルフ法は**凸多面体**内で最適化を行う枠組みである．フランクウルフ法の枠組み内でアルゴリズムを構成すれば，最適解に劣線形収束することが理論的に保証されている．符号制約 SVM の双対問題 (8) の場合，超立方体 $[0,1]^n$ 内で最適化を行うので，フランクウルフ法を適用するための前提条件を満たしていることが分かる．

フランクウルフ法の各反復は方向探索ステップと線探索ステップからなる (図 2)．方向探索ステップでは，現在の解 $\boldsymbol{\alpha}^{(t-1)}$ まわりで双対目的関数 $D(\cdot)$ を**線形近似**した関数

$$\boldsymbol{u} \mapsto \left\langle \nabla D(\boldsymbol{\alpha}^{(t-1)}), \boldsymbol{u} \right\rangle + D(\boldsymbol{\alpha}^{(t-1)}) \tag{10}$$

を実行可能領域内で最適化する部分問題を解く．この部分

Algorithm 1: Frank-Wolfe algorithm for solving the dual problem (8).

begin
  Let $\boldsymbol{\alpha}^{(0)} \in [0,1]^n$;
  for $t := 1$ **to** T **do**
    $\boldsymbol{u}_t \in \underset{\boldsymbol{u}\in[0,1]^n}{\operatorname{argmax}} \left\langle \nabla D(\boldsymbol{\alpha}^{(t-1)}), \boldsymbol{u} \right\rangle$;
    $\boldsymbol{q}_t := \boldsymbol{u}_t - \boldsymbol{\alpha}^{(t-1)}$;
    $\eta_t \in \arg\max_{\eta\in[0,1]} D(\boldsymbol{\alpha}^{(t-1)} + \eta \boldsymbol{q}_t)$;
    $\boldsymbol{\alpha}^{(t)} := \boldsymbol{\alpha}^{(t-1)} + \eta_t \boldsymbol{q}_t$;
  end
end

問題は一般に線形計画問題になる．第 t 反復における，その部分問題の解を $\boldsymbol{u}_t$ とし，$\boldsymbol{q}_t := \boldsymbol{u}_t - \boldsymbol{\alpha}^{(t-1)}$ とおく．

線探索ステップでは，現在の解 $\boldsymbol{\alpha}^{(t-1)}$ と線形近似したときの最適解 $\boldsymbol{u}_t$ とを結ぶ線分上で双対目的関数の値が最大となる点に解を更新する．すなわち，新しい解は

$$\boldsymbol{\alpha}^{(t)} := \boldsymbol{\alpha}^{(t-1)} + \eta_t \boldsymbol{q}_t \tag{11}$$

と表され，η_t は

$$\eta_t := \underset{\eta\in[0,1]}{\operatorname{argmax}}\, D(\boldsymbol{\alpha}^{(t-1)} + \eta \boldsymbol{q}_t) \tag{12}$$

のように定める．以上をまとめると，フランクウルフ法は Algorithm 1 のようにあらわされる．

符号制約ペガサス法は，最適解に劣線形収束し，かつ，各反復は $O(nd)$ で計算できる．しかし，明確な停止条件を持たない．双対問題にフランクウルフ法を適用すると，そのアルゴリズムは最適解に劣線形収束し，かつ，明確な停止条件を持つことになる．もし，フランクウルフ法の各反復も計算量 $O(nd)$ 以内で実現できれば，符号制約ペガサス法の長所を保持したまま，短所を克服した方法論となる．フランクウルフ法の各反復を計算量 $O(nd)$ 以内におさえるためには，方向探索ステップも線探索ステップも計算量を $O(nd)$ 以内に抑えるアルゴリズムが必要になる．

6. 方向探索ステップ

本節では，フランクウルフ法 (Algorithm 1) における方向探索ステップが計算量 $O(nd)$ でおさまることを示す．符号制約 SVM の双対問題 (8) に対する方向探索ステップでは，次の部分問題を解くことが要求される：

$$\min \quad \left\langle \nabla D(\boldsymbol{\alpha}^{(t-1)}), \boldsymbol{u} \right\rangle \quad \text{wrt} \quad \boldsymbol{u} \in [0,1]^n. \tag{13}$$

部分問題 (13) は線形計画問題になる．もし汎用の線形計画ソルバーに頼ってしまうと 1 回の方向探索ステップで $O(n^3)$ の計算量が生じてしまう．これは，例題数 n が大きくなると，許容しがたい計算コストになる．

本研究では，線形計画問題 (13) の最適解が閉形式で与えられることを発見した．その閉形式解 $\boldsymbol{u}_t \in [0,1]^n$ の第 i 要素は

$$u_{i,t} = \begin{cases} 1 & \text{if } y_i \left\langle \boldsymbol{w}(\boldsymbol{\alpha}^{(t-1)}), \boldsymbol{x}_i \right\rangle < 1, \\ 0 & \text{if } y_i \left\langle \boldsymbol{w}(\boldsymbol{\alpha}^{(t-1)}), \boldsymbol{x}_i \right\rangle \geq 1 \end{cases} \tag{14}$$

と表される．線形計画問題 (13) はユニークであるとは限らない．式 (14) で与えた解は，最適解の一つである．最適解のいずれをとっても劣線形収束の理論保証は維持できる．

方向探索ステップの計算量.

線形計画問題 (13) の閉形式解 (14) を得るための計算手順とそれぞれの計算コストは次のようになる：

- $\boldsymbol{v}^{(t-1)} := \boldsymbol{X}\boldsymbol{\alpha}^{(t-1)}/(\lambda n)$ を計算する; $O(nd)$.
- $\boldsymbol{w}^{(t-1)} := \Pi_{\mathcal{S}}(\boldsymbol{v}^{(t-1)})$ を計算する; $O(d)$.
- $\boldsymbol{z}^{(t-1)} := \boldsymbol{X}^\top \boldsymbol{w}^{(t-1)}$ を計算する； $O(nd)$.
- $\forall i \in [n]$, $u_{i,t} := \mathbb{1}(1 > z_i^{(t-1)})$ を計算する； $O(n)$.

ただし，$z_i^{(t-1)}$ は $\boldsymbol{z}^{(t-1)}$ の第 i 要素である．$\mathbb{1}(\cdot)$ は真偽値を引数に取り，真ならば 1，偽ならば 0 を返す演算子である．これより，方向探索ステップは $O(nd)$ の計算量となることが示された．

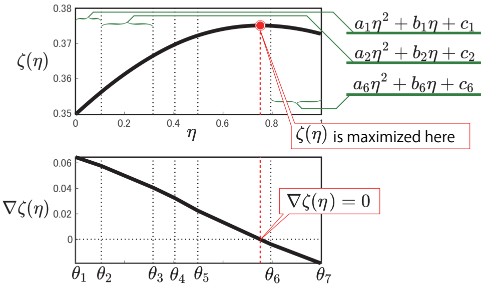

図 3 線探索問題の目的関数 $\zeta(\eta)$ は区分 2 次関数になる．第 h 区間 $[\theta_h, \theta_{h+1}]$ において，この関数は $\zeta(\eta) = a_h\eta^2 + b_h\eta + c_h$ の形式で表される．

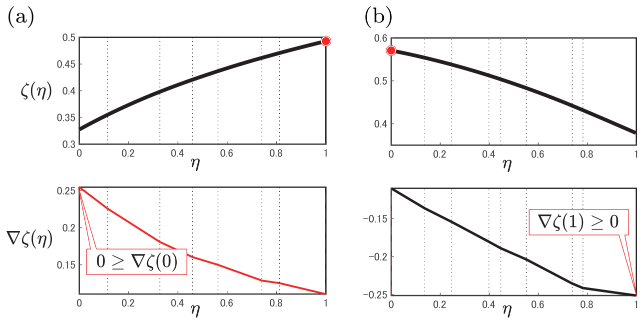

図 4 線探索における 2 ケース．図 3 に示すような，$\nabla\zeta(0) > 0 > \nabla\zeta(1)$ の場合は，$\nabla\zeta(\eta) = 0$ なる η で $\zeta(\eta)$ は最大化される．(a) $0 \geq \nabla\zeta(0)$ の場合，$\zeta(\eta)$ は $\zeta = 1$ で最大化される．(b) $\nabla\zeta(1) \geq 0$ の場合，$\zeta(\eta)$ は $\zeta = 0$ で最大化される．

7. 線探索ステップ

符号制約 SVM の双対問題を解くためのフランクウルフ法の各反復が $O(nd)$ の計算コストでおさえられることを示すために，前節では方向探索ステップが $O(nd)$ で計算できることを示した．本節では，線探索ステップも $O(nd)$

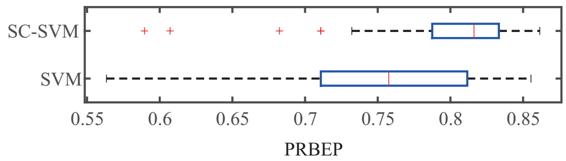

図 5 パターン認識性能の比較．

で計算できることを示す．線探索ステップは次の部分問題を解く：

$$\begin{aligned} &\max \quad \zeta(\eta) \quad \text{wrt} \quad \eta \in [0,1], \\ &\text{where} \quad \zeta(\eta) := D(\boldsymbol{\alpha}+\eta\boldsymbol{q}), \\ &\qquad \boldsymbol{\alpha} \in [0,1]^n,\ \boldsymbol{q} \in [0,1]^n - \boldsymbol{\alpha},\ \boldsymbol{q} \neq \boldsymbol{0}_n. \end{aligned} \tag{15}$$

本研究において部分問題 (15) が $O(nd)$ で解くことができることが判明したのは，次の補助定理を発見したからであった：

Lemma 1. 部分問題 (15) の目的関数 $\zeta : [0,1] \to \mathbb{R}$ は微分可能で上に凸な区分 2 次関数であり，その導関数は連続で単調非増加関数である．すなわち，整数 $d_t \in [d+1]$，$a_k \leq 0$ なる係数 (a_k, b_k, c_k) $(k \in [d_t])$，$0 = \theta_1 < \theta_2 < \cdots < \theta_{d_t+1} = 1$ なる端点 $\theta_1, \ldots, \theta_{d_t+1} \in \mathbb{R}$ が存在し，$\zeta(\cdot)$ は $\forall k \in [d_t]$, $\forall \eta \in [\theta_k, \theta_{k+1}]$,

$$\zeta(\eta) = a_k \eta^2 + b_k \eta + c_k, \tag{16}$$

と表され，また，$\forall k \in [d_t - 1]$,

$$2a_k\theta_{k+1} + b_k = 2a_{k+1}\theta_{k+1} + b_{k+1}. \tag{17}$$

を満たす．

図 3 は Lemma 1 で導入した端点と区分 2 次関数の係数を例示する．図 4 のように 3 個のケースに分解すれば線探索問題の解を導出できる．

線探索問題 (15) の解.

Lemma 1 より，線探索問題 (15) の最適解も次式で表される：

$$\eta_\star = \begin{cases} 0 & \text{if } b_1 \leq 0, \\ 1 & \text{if } 2a_{d_t+1} + b_{d_t+1} \geq 0, \\ -\frac{b_{k_\star}}{2a_{k_\star}} & \text{if } b_1 > 0,\ 2a_{d_t+1} + b_{d_t+1} < 0,\ a_{k_\star} < 0, \\ \theta_{k_\star} & \text{if } b_1 > 0,\ 2a_{d_t+1} + b_{d_t+1} < 0,\ a_{k_\star} = 0 \end{cases} \tag{18}$$

ただし，$k_\star \in [d_t]$ は，d_t 個の区間番号の一つであり，その区間内で導関数が 0 になるような区間の番号である．すなわち，$k_\star \in [d_t]$ は

$$2a_{k_\star}\theta_{k_\star} + b_{k_\star} \geq 0 \qquad \text{及び} \qquad 2a_{k_\star}\theta_{k_\star+1} + b_{k_\star} \leq 0. \tag{19}$$

を満たす．$k_\star$ の定義より，$b_1 > 0$, $2a_{d_t+1} + b_{d_t+1} < 0$ かつ $a_{k_\star} = 0$ の場合，$b_{k_\star} = 0$ を満たす．

端点および区分 2 次関数 ζ の係数.

ここでは, Lemma 1 における端点 $\theta_1, \theta_2, \ldots, \theta_{d_t+1} \in [0,1]$ および区分 2 次関数 ζ の係数 (a_k, b_k, c_k) $(k = 1, \ldots, d_t)$ をどのように決めることができるか述べる．

ベクトル $\boldsymbol{v}_0 := [v_{1,0}, \ldots, v_{d,0}]^\top$ および $\boldsymbol{v}_q := [v_{1,q}, \ldots, v_{d,q}]^\top$ はそれぞれ次の要素を持つとする:$\forall h \in [d]$,

$$v_{h,0} := \frac{1}{\lambda n}\langle \boldsymbol{f}_h, \boldsymbol{\alpha} \rangle, \quad v_{h,q} := \frac{1}{\lambda n}\langle \boldsymbol{f}_h, \boldsymbol{u} - \boldsymbol{\alpha} \rangle. \tag{20}$$

ただし，$\boldsymbol{f}_h \in \mathbb{R}^n$ は $\boldsymbol{X}^\top$ の第 h 列である．離散集合

$$\Theta := \left\{ \theta \in [0,1] \,\middle|\, \exists h \in \mathcal{I}_+ \text{ s.t. } v_h^q \neq 0,\ \theta = -\frac{v_h^0}{v_h^q} \right\} \cup \{0,1\} \tag{21}$$

を導入する．区間の個数 d_t は離散集合 Θ の要素数とする (i.e. $d_t = \mathrm{card}(\Theta) - 1$)．また，$d_t$ 個の区間を分ける端点 $\theta_1, \ldots, \theta_{d_t+1}$ は，Θ の要素を $\theta_1 < \theta_2 < \cdots < \theta_{d_t+1}$ のように昇順に整列したものである．この設定において，式 (16) は，係数を次のように与えると成り立つ：$\forall k \in [d_t]$,

$$\begin{aligned} a_k &:= -\frac{\lambda}{2}\sum_{h \in \mathcal{H}_k} v_{h,q}^2, \quad b_k := \frac{1}{n}\langle \boldsymbol{1}, \boldsymbol{u} - \boldsymbol{\alpha} \rangle - \lambda \sum_{h \in \mathcal{H}_k} v_{h,q} v_{h,0}, \\ c_k &:= \frac{1}{n}\langle \boldsymbol{1}, \boldsymbol{\alpha} \rangle - \frac{\lambda}{2}\sum_{h \in \mathcal{H}_k} v_{h,0}^2. \end{aligned} \tag{22}$$

ただし，

$$\mathcal{H}_k := \mathcal{I}_0 \cup \{ h \in \mathcal{I}_+ \,|\, v_{h,0} + 0.5(\theta_k + \theta_{k+1}) v_{h,q} > 0 \}. \tag{23}$$

線探索ステップの計算量.

線探索問題を解くために必要な手順とそれぞれの計算コストは次のようになる：

$\boldsymbol{v}_0$ および $\boldsymbol{v}_q$ を計算する;	$O(nd)$.
Θ を定める;	$O(d)$.
Θ の要素を昇順に整列する;	$O(d \log d)$.
$\forall k \in [d_t]$ に対して $\mathcal{H}_k$ を計算する;	$O(d^2)$.
$\forall k \in [d_t]$ に対して (a_k, b_k, c_k) を計算する;	$O(d^2)$.
$k_\star$ を見つける;	$O(d)$.
解 (18) を計算する;	$O(1)$.

よって，d が $O(n)$ 以内ならば，線探索ステップは $O(nd)$ の計算コストで抑えられる．

6 節および本節で述べた議論により，次の結果を得るに至った．

Theorem 2. 符号制約 SVM の双対問題を解くためのフランクウルフ法の各反復は，d が $O(n)$ 以内ならば，$O(nd)$ の計算コストでおさえられる．

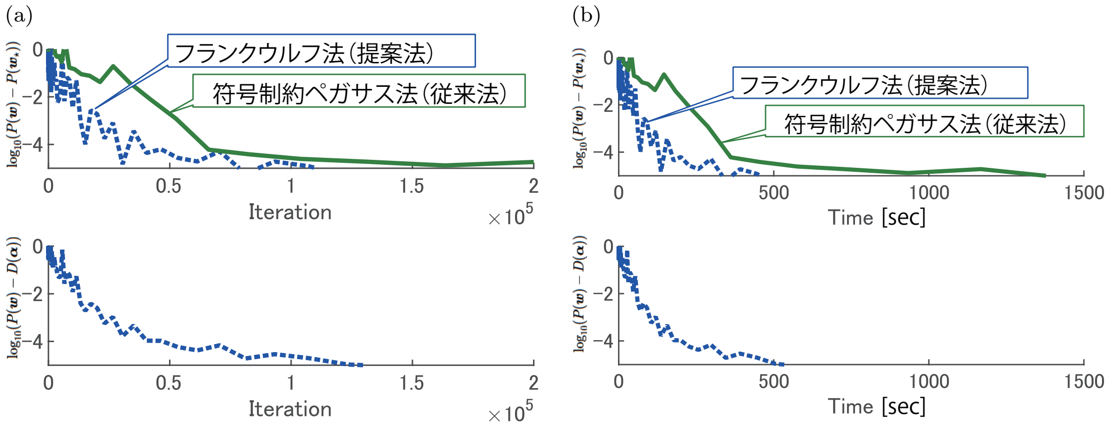

図 6　収束のふるまい．上段：目的誤差 $P(\boldsymbol{w}) - P(\boldsymbol{w}_\star)$ の比較．下段：フランクウルフ法における双対ギャップ $P(\boldsymbol{w}) - D(\boldsymbol{\alpha})$ のふるまい．

8. 実験

河川における大腸菌が一定以上存在するか水質データから予測する問題を考える．3 節で述べた $d = 11$ 次元の水質データを用いる．本研究では，177 回大腸菌数を測定し，その日時における水質データを収集した．このうち，10 個を無作為に選び，符号制約 SVM と従来の SVM の比較を行った．残りの 167 個を評価用データとして，PRBEP(Precision Recall Break Even Point) を算出した．これを 50 回繰り返し，箱ひげ図をプロットしたのが，図 5 である．符号制約の効果が明白に表れており，符号制約 SVM は強力なアプローチであることが実証された．

さらに，libsvm のウェブサイトで公開されているデータセット USPS を使って，本稿で提案した最適化アルゴリズムのスケーラビリティを検証した．このデータセットは例題数 $n = 4,374$, 次元数 $d = 256$ からなる．図 6 上段において提案するフランクウルフ法と，従来法である符号制約ペガサス法の目的誤差を比較した．フランクウルフ法と符号制約ペガサス法は要した反復数や CPU 時間に大きな違いはないことが示された．図 6 下段に停止条件を提供する双対ギャップをプロットした．双対ギャップも目的誤差とほぼ変わらない計算速度で 0 に収束していることが分かる．

9. おわりに

本稿では，フランクウルフ法に基づく符号制約 SVM の学習法を提案した．提案法は，双対ギャップ $P(\boldsymbol{w}(\boldsymbol{\alpha})) - D(\boldsymbol{\alpha})$ を停止条件に用いることで，解の精度を保証できるという強みがある．加えて，これまでに開発していた射影勾配法と同じく劣線形収束し，かつ，各反復の計算量も $O(nd)$ を維持している．このように，提案法も射影勾配法も同等な計算量で学習できることを理論的に示した．また，実データによる数値実験により，符号制約の効果を実証し，かつ，フランクウルフ法の双対ギャップも目的誤差も，従来の符号制約ペガサス法の目的誤差と同等な反復数や CPU 時間で 0 に収束することを確認した．誌面の制約から本稿には記載できなかった証明や導出は，別の機会に報告する予定である．

参考文献

[1] Bertsekas, D. P.: *Nonlinear Programming, second edition*, Athena Scientific (1999).
[2] Ito, E., Sato, T., Sano, D., Utagawa, E. and Kato, T.: Virus Particle Detection by Convolutional Neural Network in Transmission Electron Microscopy Images, *Food and Environmental Virology*, Vol. 10, No. 2, pp. 201–208 (2018).
[3] Jaggi, M.: Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization, *Proceedings of the 30th International Conference on Machine Learning* (Dasgupta, S. and McAllester, D., eds.), Proceedings of Machine Learning Research, Vol. 28, No. 1, Atlanta, Georgia, USA, PMLR, pp. 427–435 (2013).
[4] Kato, T. and Hirohashi, Y.: Learning Weighted Top-k Support Vector Machine, *Proceedings of The 10th Asian Conference on Machine Learning* (Lee, W. S. and Suzuki, T., eds.), Proceedings of Machine Learning Research, Vol. 101, PMLR, pp. – (2019).
[5] Kato, T., Kobayashi, A., Oishi, W., Kadoya, S., Okabe, S., Ohta, N., Amarasiri, M. and Sano, D.: Sign-constrained linear regression for prediction of microbe concentration based on water quality datasets, *Journal of Water and Health*, Vol. 17, No. 3, pp. 404–415 (2019).
[6] Shalev-Shwartz, S., Singer, Y., Srebro, N. and Cotter, A.: Pegasos: primal estimated sub-gradient solver for SVM, *Math. Program.*, Vol. 127, No. 1, pp. 3–30 (2011).
[7] Vapnik, V. N.: *Statistical Learning Theory*, Wiley-Interscience (1998).
[8] Varela, M., Ouardani, I., Kato, T., Kadoya, S., Aouni, M., Sano, D., and Romalde, J.: Sapovirus in wastewater treatment plants in Tunisia: Prevalence, removal, and genetic characterization, *Applied and Environmental Microbiology*, Vol. 84, No. 6, pp. 1–11 (2018).
[9] 小林美里，宮村明帆，佐野大輔，加藤　毅：符号制限線形識別器の開発と河川水中大腸菌数予測への応用，第 15 回情報科学技術フォーラム FIT2016，第 1 分冊，pp. 149–150 (2016).