4U-07

局所強凸性を利用した双対座標上昇法の高速化
Using Local Strong Convexity for SDCA Algorithm

中島 直也 * 廣橋 義寛 † 太田 直哉 * ‡ 加藤 毅 * ‡ §
Naoya Nakajima Yoshihiro Hirohashi Naoya Ohta Tsuyoshi Kato

1. はじめに

SVM やロジスティック回帰など多くの２クラス分類器の学習は，**正則化経験リスク**の最小化問題に帰着される．線形識別器 $\langle \boldsymbol{w}, \cdot \rangle$ のパラメータ $\boldsymbol{w} \in \mathbb{R}^d$ の値を決定するために，n 個の訓練用例題 $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{\pm 1\}$ を収集したとする．すると，最小化すべき正則化経験リスクは

$$P_0(\boldsymbol{w}) := P(\boldsymbol{w}\,;\,\phi) = \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \phi(y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \tag{1}$$

で与えられる．ただし，$\lambda > 0$ は正則化定数である．$\phi : \mathbb{R} \to \mathbb{R}_+$ は凸で単調減少する**損失関数**である．損失関数 $\phi(z)$ をヒンジ損失 $\phi_{1\mathrm{h}}(z) := \max(0, 1-z)$ に設定すると SVM になり，ロジスティック損失 $\phi_\ell(z) := \log(1+\exp(-z))$ にするとロジスティック回帰になる．

正則化経験リスク $P(\cdot\,;\,\phi)$ の最小化問題を解く最適化算法の一つに**確率的座標上昇法** (**SDCA**) [1] がある．SDCA は，損失関数 ϕ が**リプシッツ連続**であるか，**平滑**であるときの理論的な収束解析がなされており，特に，損失関数が平滑であるとき ϵ-最適解に**線形収束**するという理論保証がある（諸定義は付録参照）．

本研究では，**厳罰損失**，すなわちリプシッツ連続でも平滑でもない損失関数に注目する（図 1）．厳罰損失 ϕ は，z が 0 から負の方向へ離れていくと，$\phi(z)$ が急激に大きくなるため，大きな誤りを看過せずに学習することができる．しかし，SDCA は厳罰損失による正則化経験リスクの最小化問題に対して，線形収束は保証されない．これに対して，本研究では，損失関数を部分的に線形近似することによって，厳罰損失でも一定の条件を満たせば線形収束を保証する最適化算法を開発した．

仮定: $\forall i,\ \|\boldsymbol{x}_i\| \le 1$．$\phi : \mathbb{R} \to \mathbb{R}_+$ は凸，単調減少，連続的微分可能．$1 \ge \phi(0) > 0$.

厳罰損失の例: 指数損失

$$\phi_{\mathrm{e}}(z) := \exp(-z) \tag{2}$$

は厳罰損失である．また，**高次ヒンジ損失**

$$\phi_{p\mathrm{h}}(z) := \frac{1}{p}\max(0, 1-z)^p \tag{3}$$

は，$p \ge 3$ では厳罰損失になる．ただし，$p = 1$ なら標準的なヒンジ損失である．確かに，指数損失や $p \ge 3$ の高次ヒンジ損失はリプシッツ連続でも平滑でもない．

*群馬大学理工学部
†株式会社デンソー
‡群馬大学次世代モビリティ社会実装研究センター (CRANTS)
§早稲田大学規範科学総合研究所 (IIRS)

2. 従来の最適化算法： SDCA [1]

従来の SDCA [1] は主問題を解く代わりに，正則化経験リスク $P(\cdot\,;\,\phi)$ の**フェンシェル双対**

$$D_0(\boldsymbol{\alpha}) := D(\boldsymbol{\alpha}\,;\,\phi^*) = -\frac{\lambda}{2}\|\boldsymbol{w}(\boldsymbol{\alpha})\|^2 - \frac{1}{n}\sum_{i=1}^{n}\phi^*(-\alpha_i)$$

を最大化する $\boldsymbol{\alpha}$ を求める反復的最適化算法である．ただし，$\boldsymbol{w}(\boldsymbol{\alpha}) := \tilde{\boldsymbol{X}}\boldsymbol{\alpha}/(\lambda n)$，$\tilde{\boldsymbol{X}} := [y_1\boldsymbol{x}_1, \ldots, y_1\boldsymbol{x}_n]$．主問題の最適解は，$D_0$ を最大にする解 $\boldsymbol{\alpha}_\star$ から，$\boldsymbol{w}_\star = \boldsymbol{w}(\boldsymbol{\alpha}_\star)$ のように復元できる (算法の詳細は付録参照)．

線形収束の理論: 反復数が

$$t \ge \left(n + \frac{1}{\lambda\gamma_{\mathrm{gl}}}\right)\log\left(\frac{1}{\epsilon}\right) \tag{4}$$

を満たすとき，$\boldsymbol{\alpha}^{(t)}$ が ϵ-最適解となること (i.e. $D_0(\boldsymbol{\alpha}_\star) - \mathbb{E}[D_0(\boldsymbol{\alpha}^{(t)})] \le \epsilon$) が保証されていた．これは，SDCA は損失関数が平滑なら線形収束することを意味する．しかし，厳罰損失では，強凸係数 γ_{gl} が 0 になるため，反復数を下から抑えられない．

3. 提案する最適化算法

本研究で開発した方法論の線形収束性は，次の補助定理が根幹をなしている．

補助定理 3.1. 正則化経験リスク $P(\cdot\,;\,\phi)$ の最小解 $\boldsymbol{w}_\star$ は次の線形制約を満たす：$\tilde{\boldsymbol{X}}^\top \boldsymbol{w}_\star \ge b_u \mathbf{1}_n$．ただし，$b_u := \phi^{-1}(n\phi(0))$ とする．

上記 b_u を使って，元の損失関数 ϕ を

$$\tilde{\phi}(z) := \begin{cases} -u_u z + c_u & \text{if } z < b_u, \\ \phi(z) & \text{if } b_u \le z \end{cases}$$

$$(\text{ただし} \quad u_u := -\nabla\phi(b_u), \quad c_u := u_u b_u + \phi(b_u))$$

と部分的に**線形近似**する．正則化経験リスク $P_0 = P(\cdot\,;\,\phi)$ の最小化問題の最適解 $\boldsymbol{w}_\star$ は，$P(\cdot\,;\,\phi)$ の損失関数を $\tilde{\phi}$ に置き換えた $\tilde{P} := P(\cdot\,;\,\tilde{\phi})$ の最小解にもなる．提案法は，基本的には $\tilde{P}$ のフェンシェル双対 $\tilde{D} := D(\cdot\,;\,\tilde{\phi}^*)$ を最大化する SDCA を適用するものである．提案法では，各反復の更新則において近似する前の損失関数 ϕ の**局所強凸係数** $\gamma_t := \gamma(\alpha^{(t-1)}, u_t, \phi)$ を算出し，$J_t^2(\cdot\,;\,\gamma_t)$ を最大化する $\Delta\alpha$ を求めるように変更する．ただし，

$$\gamma(u, u', \phi) := \max\left\{\gamma \in \mathbb{R}_+ \,\middle|\, \textsc{Sci}(u, u', \phi, \gamma) \ge 0\right\}$$

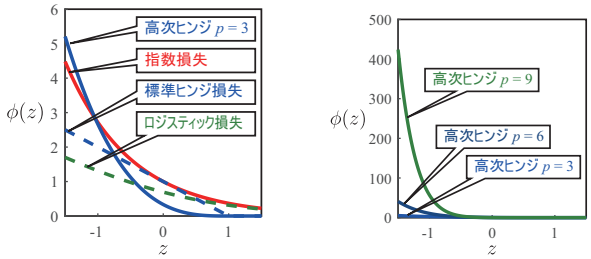

図 1: **損失関数.** 厳罰損失 (指数損失や高次ヒンジ損失) は 0 から負の方向に離れると急激に損失が大きくなる.

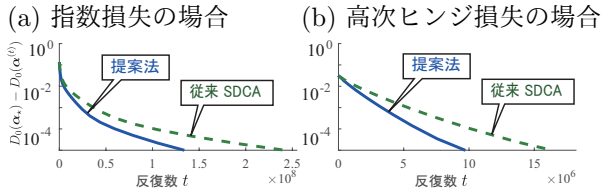

図 2: **収束の速さの比較.** どちらの損失関数でも提案法は従来法より早くギャップが小さくなり，最適値に近づいている.

とした．J_t^2 の定義は (9) 参照.

線形収束の保証: 上述の最適化算法では次の収束理論が成り立つ：

定理 1. 区間 $[0, u_u]$ における局所強凸係数の下限を $\tilde{\gamma}$ とおく．提案算法は，反復数が

$$t \geq \left(n + \frac{1}{\lambda\tilde{\gamma}}\right) \log\left(\frac{1}{\epsilon}\right) \tag{5}$$

に達したとき，**期待ギャップ**は $D_0(\boldsymbol{\alpha}_\star) - \mathbb{E}[D_0(\boldsymbol{\alpha}^{(t)})] \leq \epsilon$ を満たす．(証明は付録参照)

定理 1 では，提案算法は $\tilde{D}$ の最大化を行うが，その算法で生成される $\boldsymbol{\alpha}^{(t)}$ でもとの双対関数 D_0 の期待ギャップが抑えられることに注意されたい.

局所強凸係数の具体例: 指数損失の局所強凸係数および半直線 $[b_u, +\infty)$ における下限は

$$\gamma(u, u', \phi_{\mathrm{e}}) = \frac{1}{\max(u, u')}, \quad \tilde{\gamma} = \frac{1}{n}$$

となる．高次ヒンジ損失では，

$$\gamma(u, u', \phi_{\mathrm{e}}) = \frac{\max(u, u')^{1/(p-1)-1}}{p-1}, \quad \tilde{\gamma} = \frac{n^{(2-p)/p}}{p-1}$$

となる．よって，これらでは提案する最適化算法の反復数の下限を形成でき，線形収束が保証される.

4. 実験

収束速度: 提案最適化法と従来の SDCA を公開データセット cov1 に適用した ($n = 581,012$, $d = 54$).

表 1: **正解率.** 太字は最高性能，下線は最高性能と有意差がないもの．p-ヒンジは高次ヒンジ損失を表す.

	従来 SVM	指数損失	3-ヒンジ	9-ヒンジ
Arrhythmia	0.611	0.604	0.668	**0.716**
Segment	0.812	0.877	0.871	**0.901**
Soybean Large	0.854	0.847	0.886	**0.907**
Spambase	0.905	0.906	0.916	**0.926**
Spect	0.830	**0.839**	0.832	**0.839**

正則化定数に $\lambda = 1/n$ を用いて，各反復の**ギャップ** $D_0(\boldsymbol{\alpha}_\star) - D_0(\boldsymbol{\alpha}^{(t)})$ を観測した．損失関数には，指数損失および高次ヒンジ ($p = 3$) を使用した．図 2 に収束の様子を示す．提案最適化法は従来 SDCA より，高速に最適解に収束していることがわかる.

パターン認識の汎化性能: 表 1 に 5 個のデータセットにおける 2 クラス分類の性能を示す．幾つかのデータセットでは厳罰損失が統計的有意に高い正解率を得た.

A. 付録

リプシッツ連続の定義: 損失関数 ϕ がリプシッツ連続であるとは，$\forall z, \forall z' \in \mathbb{R}$, $|z - z'| \leq L|\phi(z) - \phi(z')|$ なる $L \in \mathbb{R}$ が存在することである.

平滑の定義: 損失関数 ϕ が平滑であるとは，

$$\forall u, \forall u' \in -\mathrm{dom}(\phi^*), \quad \textsc{Sci}(u, u', \phi, \gamma_{\mathrm{gl}}) \geq 0 \tag{6}$$

なる**強凸係数** $\gamma_{\mathrm{gl}} > 0$ が存在することである．ただし，$\mathrm{dom}(f)$ は関数 f の有効ドメインを表し，また，$\textsc{Sci}(u, u', \phi, \gamma)$ は

$$\textsc{Sci}(u, u', \phi, \gamma) := \inf_{s \in [0,1]} \left(s\phi^*(-u) + (1-s)\phi^*(-u') - \phi^*(-u' - s(u-u')) - \frac{(1-s)s(u-u')^2}{2}\gamma \right) \tag{7}$$

と定義される指標である．ϕ^* は ϕ の**凸共役**である.

従来 SDCA の更新則: SDCA の反復 $t-1$ における解を $\boldsymbol{\alpha}^{(t-1)}$ とする．初期値を $\boldsymbol{\alpha}^{(0)} = \mathbf{0}_n$ とする．各反復 t で無作為に選んだ例題 $i \in \{1, \ldots, n\}$ に対し，

$$J_t^0(\Delta\alpha) := D_0(\boldsymbol{\alpha}^{(t-1)} + \Delta\alpha \boldsymbol{e}_i) - D_0(\boldsymbol{\alpha}^{(t-1)}) \geq 0 \tag{8}$$

なる $\Delta\alpha$ を選び，$\boldsymbol{\alpha}^{(t)} := \boldsymbol{\alpha}^{(t-1)} + \Delta\alpha \boldsymbol{e}_i$ のように更新する．一般に $J_t^0(\Delta\alpha)$ を最大化する解は閉形式で求まらないので，その下限となる 2 次関数

$$J_t^2(\Delta\alpha; \gamma_{\mathrm{gl}}) := \frac{\Delta\alpha}{2nq_t}\left(2F_t + \gamma q_t^2\right) - \frac{\Delta\alpha^2}{2n}\frac{\gamma_{\mathrm{gl}}}{\bar{s}_i(\gamma_{\mathrm{gl}})} \tag{9}$$

を用いて，区間 $I_t := [-(-q_t)_+, (q_t)_+]$ 内で $J_t^2(\cdot\,; \gamma_{\mathrm{gl}})$ を最大化する $\Delta\alpha$ を求める．ただし，$z_{t-1} := \left\langle \boldsymbol{w}(\boldsymbol{\alpha}^{(t-1)}), \boldsymbol{x}_i \right\rangle$, $u_t := -\nabla\phi(z_{t-1})$, $q_t := u_t - \alpha_i^{(t-1)}$, $\bar{s}_i(\gamma) := \lambda n\gamma/(\lambda n\gamma + \|\boldsymbol{x}_i\|^2)$, $F_t := -u_t z_{t-1} - \phi_*(-u_t) + \phi_*(-\alpha_i^{(t-1)}) + z_{t-1}\alpha_i^{(t-1)}$ とする.

定理 1 の証明: 補助定理 3.1，文献 [1] の証明技法，および次の性質 (i)–(iii) を組み合わせると示すことができる：(i) $\alpha_i^{(t)} \in [0, u_u]$; (ii) $\forall \Delta\alpha \in I_t$, $J_t^0(\Delta\alpha) \geq J_t^2(\Delta\alpha\,; \gamma_t) \geq J_t^2(\Delta\alpha\,; \tilde{\gamma})$; (iii) $\mathbb{E}[J_t^0(\Delta\alpha_t)] \geq (\tilde{P}(\boldsymbol{w}^{(t-1)}) - D_0(\boldsymbol{\alpha}^{(t-1)}))\bar{s}/n$. ただし $\Delta\alpha_t$ は反復 t で求まる $\Delta\alpha$, $\bar{s} = (\lambda n\tilde{\gamma})/(\lambda n\tilde{\gamma} + 1)$.

参考文献

[1] Shalev-Shwartz, S. and Zhang, T.: Stochastic Dual Coordinate Ascent Methods for Regularized Loss, *J. Mach. Learn. Res.*, Vol. 14, No. 1, pp. 567–599 (2013).