２次損失サポートベクトルマシンの非線形正則化パスに関する一考察

烏山 昌幸, 竹内 一郎

名古屋工業大学

2010 年 6 月 14 日

はじめに

正則化パス (Regularization path [Hastie, et al: 04])

- サポートベクトルマシン (SVM) において正則化係数に対する最適解の変化を連続的に追跡
- 効率的なモデル選択

ただし，損失関数が 2 次だと区分線形性失われるため適用不可

はじめに

正則化パス (Regularization path [Hastie, et al: 04])

- サポートベクトルマシン (SVM) において正則化係数に対する最適解の変化を連続的に追跡
- 効率的なモデル選択

ただし，損失関数が 2 次だと区分線形性失われるため適用不可

- 2 次の損失関数を用いた SVM の非線形正則化パスを提案
- 有理近似 [Bunch, et al: 79] というアプローチを用いることで厳密解の追跡が効率よく行えることを示す

サポートベクトルマシン

2 値分類器の学習

- 訓練データセット:$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ $(\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d,\ \ y_i \in \{1, -1\})$
- 判別モデル:$f(\boldsymbol{x}) = \boldsymbol{w}^\top \Phi(\boldsymbol{x})$
- Regularized risk minimization により $\boldsymbol{w}$ を学習:

$$\min_{\boldsymbol{w}} \quad \underbrace{\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2}_{\text{正則化項}} + \underbrace{\sum_{i=1}^n \ell_{\text{hinge}}(y_i, f(\boldsymbol{x}_i))}_{\text{誤差項}},$$

ただし，λ は正則化係数，ℓ_{hinge} はヒンジ損失:

$$\ell_{\text{hinge}}(y, f(\boldsymbol{x})) = \max(0, 1 - yf(\boldsymbol{x})).$$

- 最適な λ は事前にわからないため，通常は様々な値で最適化を行う

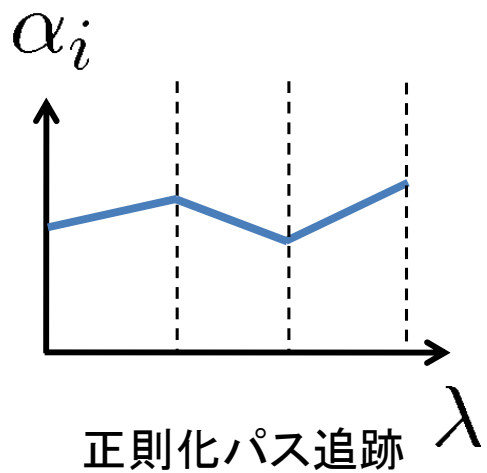
SVM の最適化問題

- 双対問題

$$\max_{\{\alpha_i\}_{i=1}^n} -\frac{1}{2\lambda}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j K_{ij} + \sum_{i=1}^n \alpha_i \quad \text{s.t. } 0 \le \alpha_i \le 1,$$

  ただし，$K_{ij} = \Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_j)$(カーネル関数)

- たくさんの λ に対してこの問題を解くのは大変
- 正則化パス [Hastie, et al: 04]:
  λ が変化した時に最適化問題を解きなおさずに最適な α の変化を効率的に追いかける（区分線形）.

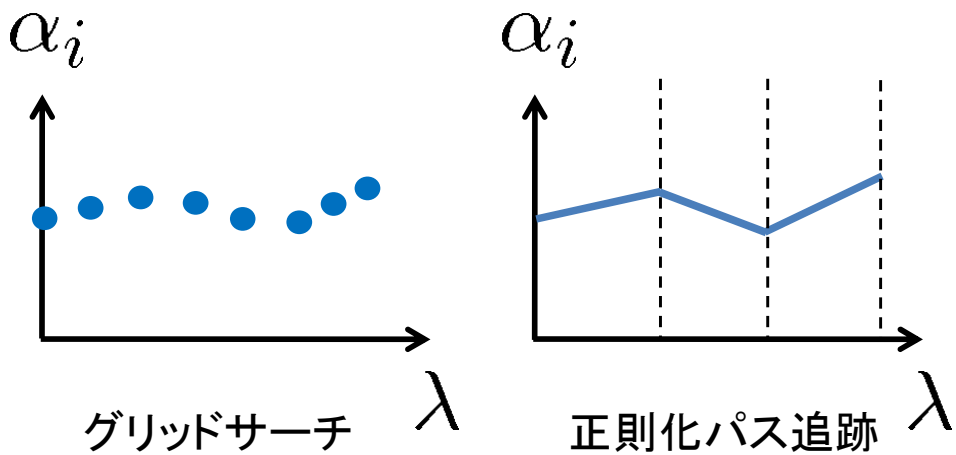


正則化パス追跡

正則化パスの特徴

- 最適解を解析的に求めて区分線形パスを追跡
  以下の step を繰り返す:

  step1 直線の方程式を求める（解析的に解を求める）
  step2 直線の変化点（ブレイクポイント）を見つける
- step2 は「1 変数の 1 次方程式を複数解き，最小の解を見つける」という問題に帰着
- たくさん最適化問題を解きなおすグリッドサーチより効率的かつ厳密



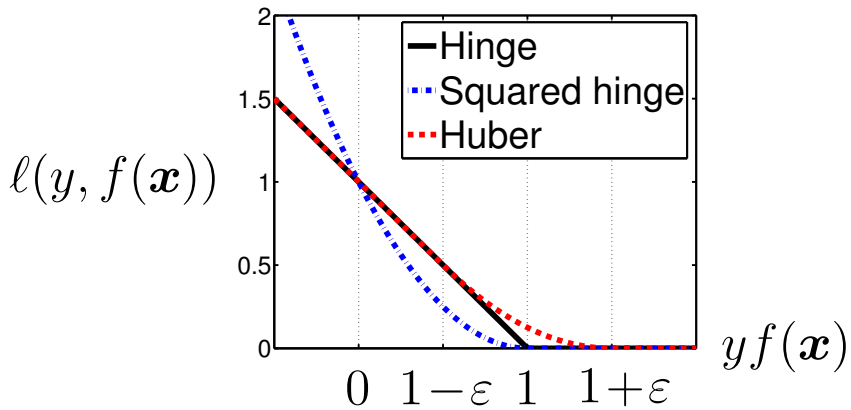
- ただし，２次の損失関数を使うと α が区分線形にならず適用不可

２次の損失関数 [Cristianini, et al: 00, Zhang: 04, 等]

- 2 乗ヒンジ損失: $\ell(y, f(\boldsymbol{x})) = \max(0, 1 - yf(\boldsymbol{x}))^2$.
- フーバー型損失:

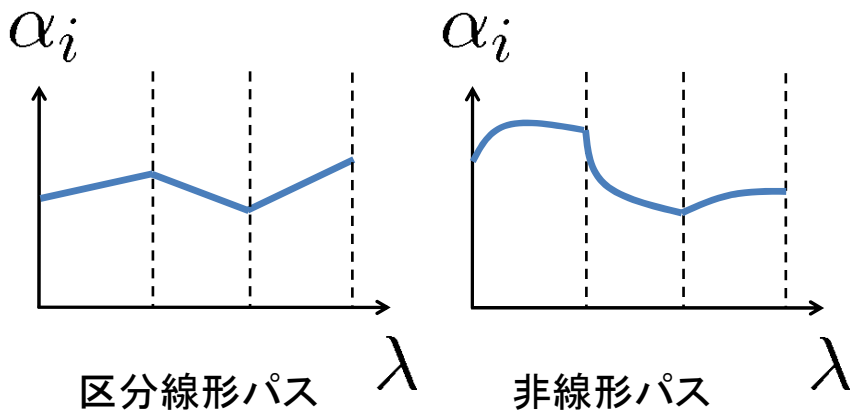
$$\ell(y, f(\boldsymbol{x})) = \begin{cases} 0, & yf(\boldsymbol{x}) > 1 + \varepsilon, \\ \frac{(1+\varepsilon-yf(\boldsymbol{x}))^2}{4\varepsilon}, & |1 - yf(\boldsymbol{x})| \leq \varepsilon, \\ 1 - yf(\boldsymbol{x}), & yf(\boldsymbol{x}) < 1 - \varepsilon, \end{cases}$$

ヒンジ損失に対する利点: 条件付き確率の推定 [Zhang: 04], 微分可能性



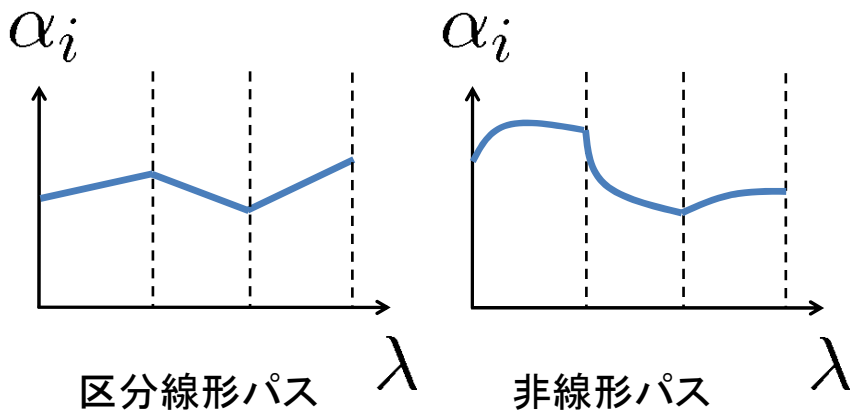
この時，SVM の解は区分非線形に

非線形パス追跡



- 各区間の解の変化を解析的に求められない場合が多い
- ブレイクポイントの発見が難しい

非線形パス追跡



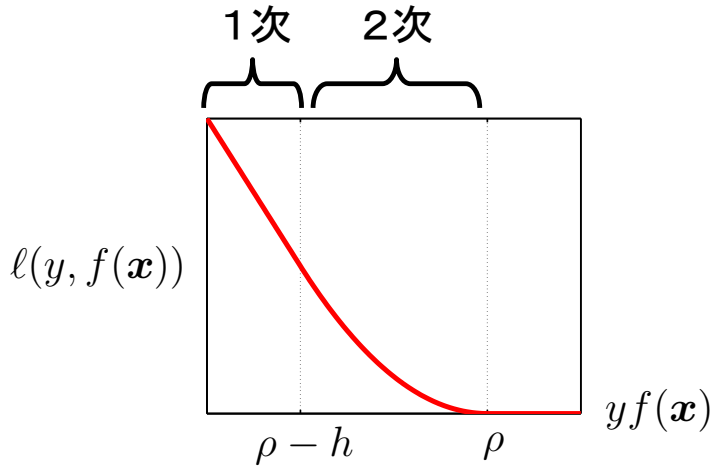
- 各区間の解の変化を解析的に求められない場合が多い
- ブレイクポイントの発見が難しい

- 2 次損失 SVM では α は非線形となるが解析的に λ の関数として導ける
- 関数の形が有理関数と呼ばれる形であること利用して効率的なブレイクポイント検出法を提案し正則化パスを行う

２次損失を用いたSVM

- 損失関数

$$\ell(y, f(\boldsymbol{x})) = \begin{cases} 0, & yf(\boldsymbol{x}) > \rho, \\ (\rho - yf(\boldsymbol{x}))^2, & yf(\boldsymbol{x}) \in [\rho - h, \rho], \\ 2h(\rho - yf(\boldsymbol{x})) - h^2, & yf(\boldsymbol{x}) < \rho - h, \end{cases}$$



(ρ, h) の設定で様々な 2 次損失関数を表現可能

- 解の解析的表現 (σ_k は $[Q_{ij}]_{i,j \in \mathcal{C}}$ の固有値, ζ_{ik} は区間内で定数):

$$\begin{aligned} \alpha_i &= \rho - \textstyle\sum_{k=1}^{\ell_C} \frac{\zeta_{ik}}{\sigma_k + \lambda}, && \text{for } i \in \{i \mid y_i f(\boldsymbol{x}_i) = \rho - \alpha\}, \\ \alpha_i &= 0, && \text{for } i \in \{i \mid y_i f(\boldsymbol{x}_i) \geq \rho\}, \\ \alpha_i &= h, && \text{for } i \in \{i \mid y_i f(\boldsymbol{x}_i) \leq \rho - h\}. \end{aligned}$$

非線形パス追跡におけるブレイクポイント検出

「1 変数非線形方程式を複数解き，最小の解を見つける」という問題に帰着

- 今，ある $\lambda_0 (= -t_0)$ から λ を減らすとする.
- 非線形方程式の例 (最小の $t = -\lambda$ を見つけたい):

$$\rho - \sum_{k=1}^{\ell_C} \frac{\zeta_{ik}}{\sigma_k - t} = 0,\ t \in (t_0, 0).$$

たくさんある非線形方程式を如何に効率的に解くか？
本研究では有理近似 [Bunch, et al: 79] を導入

非線形パス追跡におけるブレイクポイント検出

「1 変数非線形方程式を複数解き，最小の解を見つける」という問題に帰着

- 今，ある $\lambda_0 (= -t_0)$ から λ を減らすとする.
- 非線形方程式の例 (最小の $t = -\lambda$ を見つけたい):

$$\rho - \sum_{k=1}^{\ell_C} \frac{\zeta_{ik}}{\sigma_k - t} = 0, \ t \in (t_0, 0).$$

たくさんある非線形方程式を如何に効率的に解くか？
本研究では有理近似 [Bunch, et al: 79] を導入

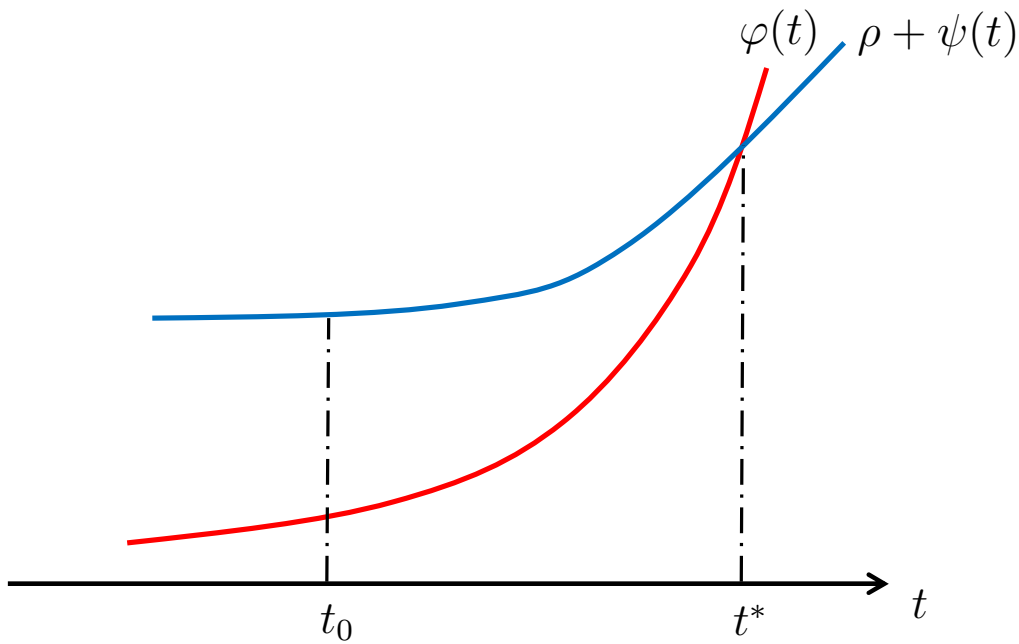
- ζ_{ik} の符号で分解:

$$\rho + \underbrace{\sum_{k \in \{k | \zeta_{ik} < 0\}} \frac{-\zeta_{ik}}{\sigma_k - t}}_{\equiv \psi(t)} = \underbrace{\sum_{k \in \{k | \zeta_{ik} > 0\}} \frac{\zeta_{ik}}{\sigma_k - t}}_{\equiv \varphi(t)}.$$

$\sigma_k \geq 0$(カーネルの半正定値性より) から，
$\psi(t)$ と $\varphi(t)$ は $t \in (t_0, 0)$ において凸単調増加

有理近似

- 非線形方程式 $\rho + \psi(t) = \varphi(t), t \in (t_0, 0)$ の最も小さい解を求める.

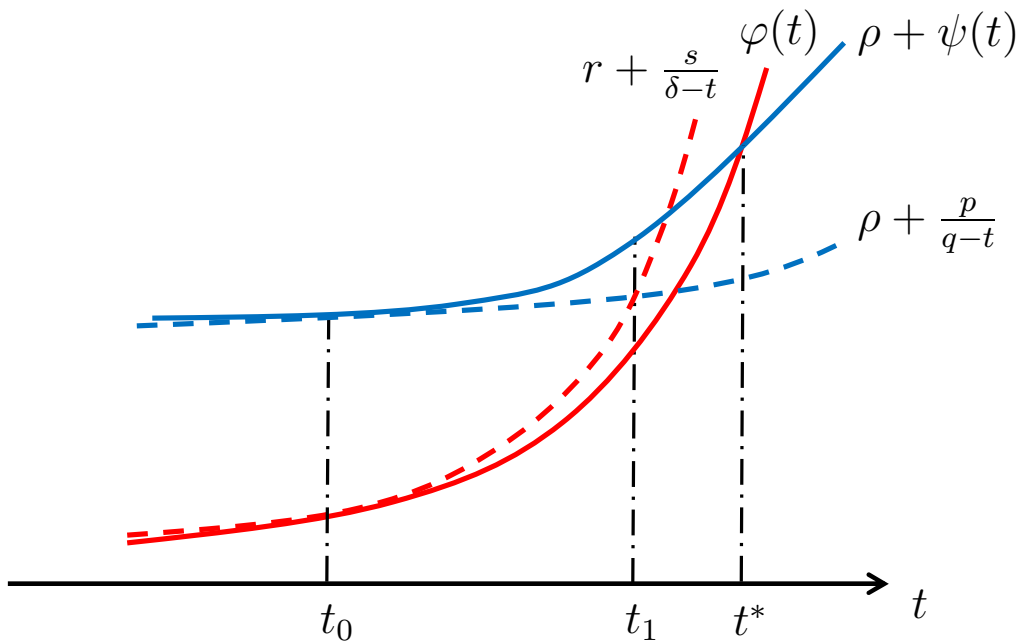


有理近似

- 非線形方程式 $\rho + \psi(t) = \varphi(t), t \in (t_0, 0)$ の最も小さい解を求める.

  step1 $\psi(t)$ を下限 $\frac{p}{q-t}$, $\varphi(t)$ を上限 $\frac{s}{\delta-t}$ で近似
  (δ, p, q, r, s は近似が t_0 において一次微分まで一致するよう定める)

  step2 近似式同士の交点を求める（２次方程式）

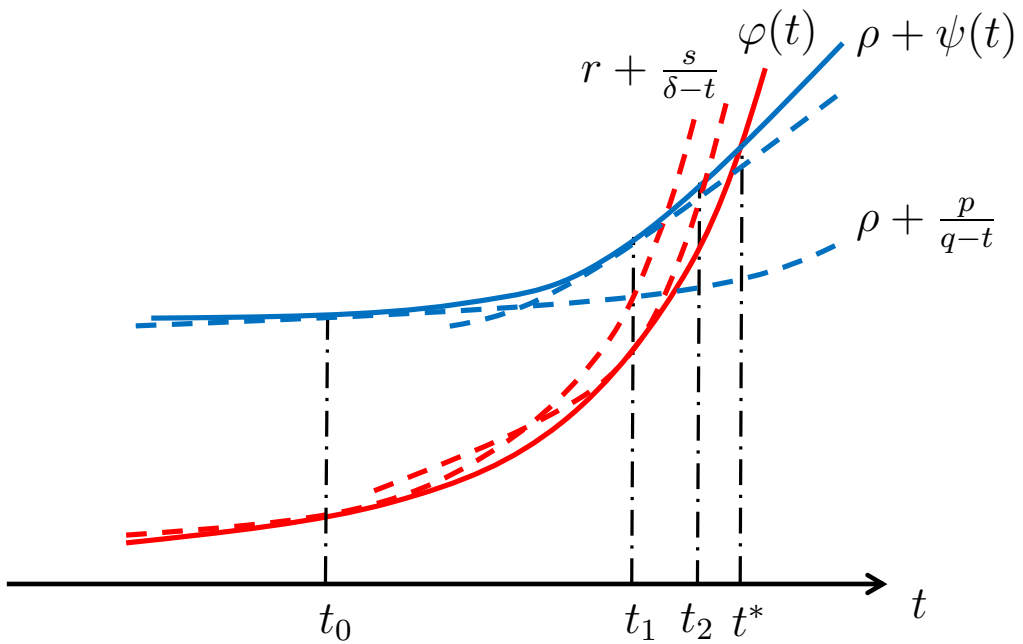


有理近似

- 非線形方程式 $\rho + \psi(t) = \varphi(t), t \in (t_0, 0)$ の最も小さい解を求める.

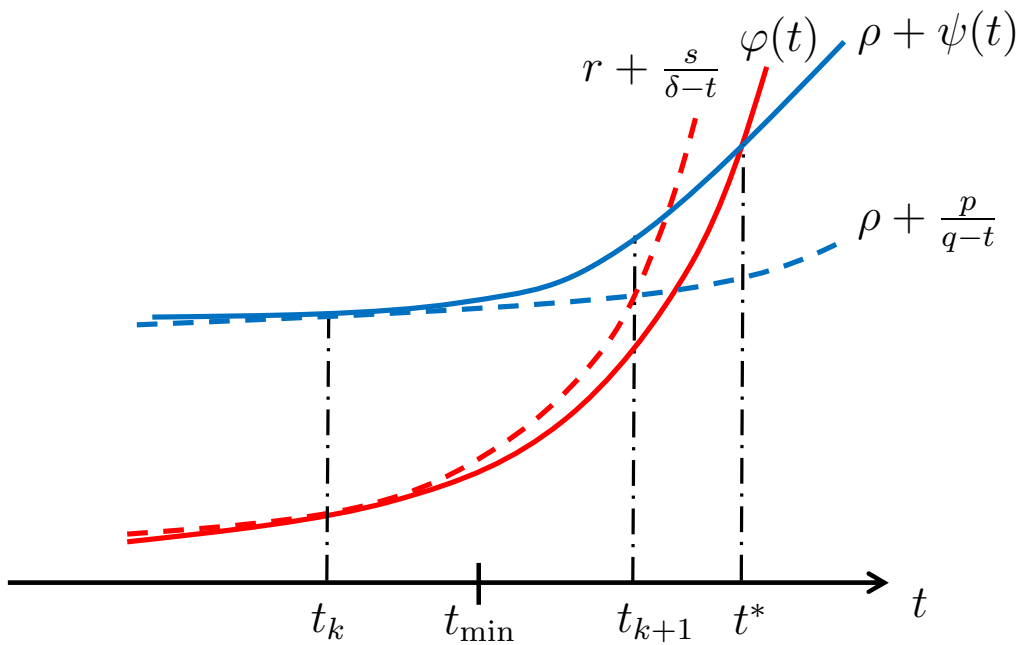
  step1 $\psi(t)$ を下限 $\frac{p}{q-t}$, $\varphi(t)$ を上限 $\frac{s}{\delta - t}$ で近似
  (δ, p, q, r, s は近似が t_0 において一次微分まで一致するよう定める)

  step2 近似式同士の交点を求める（２次方程式）



有理近似の利点

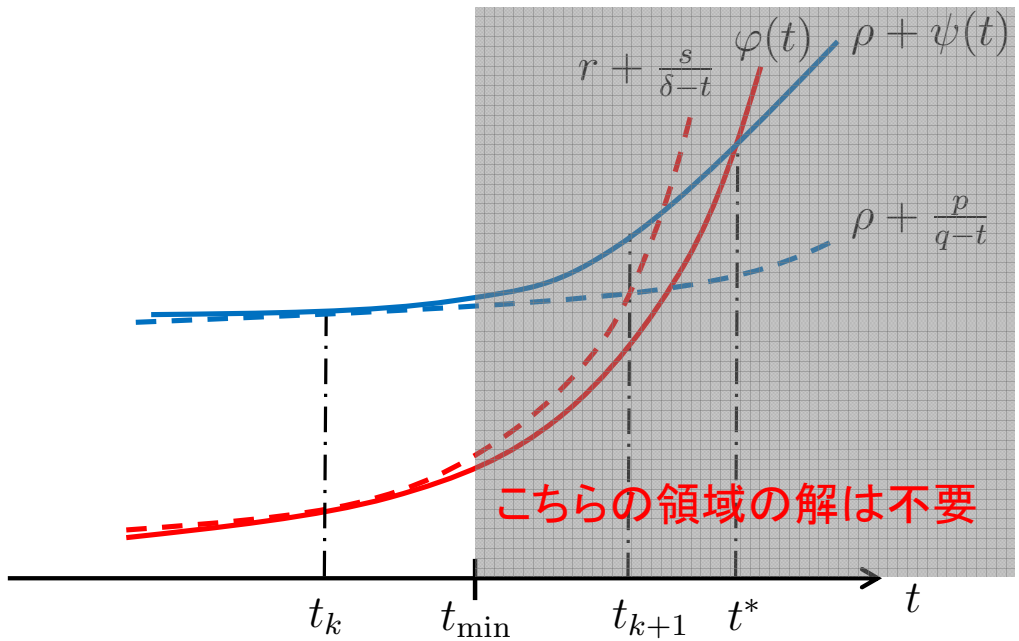
- 適当な仮定のもとで 2 次収束が保証
- 複数の方程式のなかの最小の t （最大の λ）だけあればよいため全ての方程式を収束させる必要はない



$t_{\min}$ を別の方程式で得られた解とすると $t_k > t_{\min}$ となったら探索を打ち切ってもよい（その t^* が最小となることはないため）

有理近似の利点

- 適当な仮定のもとで 2 次収束が保証
- 複数の方程式のなかの最小の t （最大の λ）だけあればよいため全ての方程式を収束させる必要はない



$t_{\min}$ を別の方程式で得られた解とすると $t_k > t_{\min}$ となったら探索を打ち切ってもよい（その t^* が最小となることはないため）

計算機実験

- 損失関数はフーバー型 $\rho = 1.1, h = 0.2$,
  RBF カーネル:$\exp(-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2), \gamma = 1/d$ (d は $\boldsymbol{x}$ の次元).
- ある値以上大きな λ では自明な解が計算できるため,
  その値を λ_0 として，$\lambda \leq 10^{-5}$ になるまでパス追跡
  (有理近似の停止条件は $|c + \psi(t) - \varphi(t)| < 10^{-12}$)
- 様々な λ で解を計算する場合と計算時間を比較
  - SMO アルゴリズム [Platt: 99] を停止基準 10^{-6} で使用
  - ブレイクポイントの数だけ解を計算
- 正規分布による人工データ ($\boldsymbol{x} \in \mathbb{R}^2$) と実データ

人工データ

計算時間比較 (sec.)

n	λ-path	λ-path (打ち切り無)	SMO
100	0.28 (0.03)	2.80 (0.33)	37.73 (12.27)
200	1.10 (0.07)	13.43 (1.08)	271.89 (59.33)
400	4.72 (0.33)	57.06 (2.87)	1635.52 (224.54)

イベント回数と有理近似の平均繰り返し回数.

n	#events L	iteration	iteration (打ち切り無)
100	150.00 (14.79)	1.20 (0.04)	11.99 (0.51)
200	286.80 (15.45)	1.12 (0.01)	14.72 (0.46)
400	532.80 (33.02)	1.07 (0.01)	16.28 (0.44)

* 全てのイベント点でパス追跡の解の精度は SMO より高いことを確認

- SMO と比較して効率良く最適解の追跡が可能であった
- 有理近似打ち切りの効果が非常に大きい

実データ

- australian, $n = 552, d = 14$
  german, $n = 800, d = 24$
  各 10 回平均

計算時間比較 (sec.)

	λ-path	SMO
australian	82.81 (7.13)	2519.12 (366.11)
german	313.91 (15.62)	2409.60 (164.76)

イベント回数と有理近似の平均繰り返し回数.

	#events L	iteration
australian	1274.90 (30.90)	1.05 (0.00)
german	1821.80 (25.07)	1.04 (0.01)

まとめ

- 2 次の損失関数を用いた SVM の非線形パス追跡法を提案した.
- 有理近似を利用したブレイクポイント検出法により効率的にパス追跡が可能であることを実験的に示した.