数理最適化から機械学習へのアプローチ

武田 朗子　　数理・推論研究系 教授

● **数理最適化法**とは「与えられた制約の下で，よりよい目的を達成するための数理最適化問題」

$$\min\ f(\boldsymbol{x})$$
$$\text{s.t.}\quad g_i(\boldsymbol{x}) \le 0,\quad i = 1, \ldots, r$$

に定式化してそれを解く計算手順までを含めて呼ぶ．

● **機械学習**の分野では，データから規則性やパターンを発見するため，しばしば数理最適化法が用いられる．

ex) 二値判別問題

データ $\boldsymbol{x}_i \in \mathbb{R}^n,\ y_i \in \{+1, -1\}\ (i = 1, 2, \ldots, m)$

から予測誤差 $\Pr(h(X) \neq Y)$ がなるべく小さくなるように，

判別関数 $h(X) = X^\top \boldsymbol{w}^* + b^*$ を定める．

最近の数理最適化法がどのように機械学習に使われるか，いくつかの例を挙げて紹介する

【研究1：ロバスト最適化法の二値判別への適用】

様々な二値判別モデルに対して，ロバスト最適化による統一的な定式化（UCM）を与えた．

max-min型の定式化により
データの変動に頑健な解が得られる

$$\max_{\|\boldsymbol{w}\|^2=1}\ \min_{\boldsymbol{x}_+ \in \mathcal{U}_+, \boldsymbol{x}_- \in \mathcal{U}_-} (\boldsymbol{x}_+ - \boldsymbol{x}_-)^\top \boldsymbol{w}$$

集合の境界

Briec ('97)の双対性

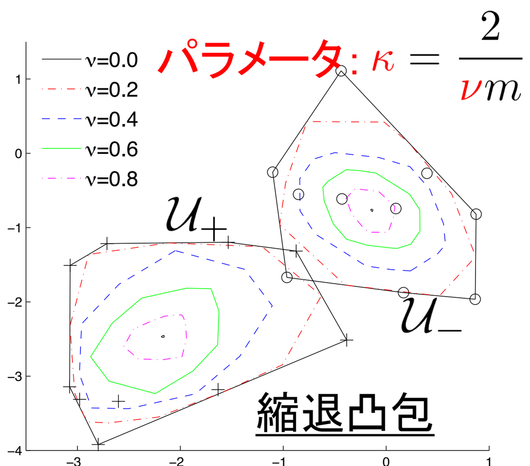
$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|^2 \quad \text{s.t.}\quad \boldsymbol{x} \in \mathrm{bd}(\mathcal{U}_+ \ominus \mathcal{U}_-) \quad \ldots\text{(UCM)}$$

- $\boldsymbol{x}_+, \boldsymbol{x}_-$: 各クラスの平均ベクトル（真の分布が未知なので不確定）
- $\mathcal{U}_+, \mathcal{U}_-$ ：有界閉凸集合（$\boldsymbol{x}_+, \boldsymbol{x}_-$ の動きうる範囲）
- $\mathcal{U}_+ \ominus \mathcal{U}_- := \{\boldsymbol{x}_+ - \boldsymbol{x}_- :\ \boldsymbol{x}_+ \in \mathcal{U}_+,\ \boldsymbol{x}_- \in \mathcal{U}_-\}$: $\mathcal{U}_\pm$ のMinkovski差
- b^*: $h(X) = 0$ が最適解 $\boldsymbol{x}_+^*, \boldsymbol{x}_-^*$ の中点を通るように設定
- 2つの集合 $\mathcal{U}_+, \mathcal{U}_-$ が交わる → (UCM)は非凸最適化問題
  交わらない → 凸最適化問題
- 様々な二値判別モデルの差異は $\mathcal{U}_+ \ominus \mathcal{U}_-$ にある

ex.1) 楕円： $\mathcal{U}_\pm = \{\bar{\boldsymbol{x}}_\pm + \Sigma_\pm^{1/2} \boldsymbol{\lambda} : \|\boldsymbol{\lambda}\| \le \kappa\}$

ex.2) 縮退凸包：

$$\mathcal{U}_\pm = \{\sum_{i \in M_\pm} \lambda_i \boldsymbol{x}_i : \boldsymbol{e}^\top \boldsymbol{\lambda} = 1,\ \boldsymbol{0} \le \boldsymbol{\lambda} \le \kappa \boldsymbol{e}\}$$



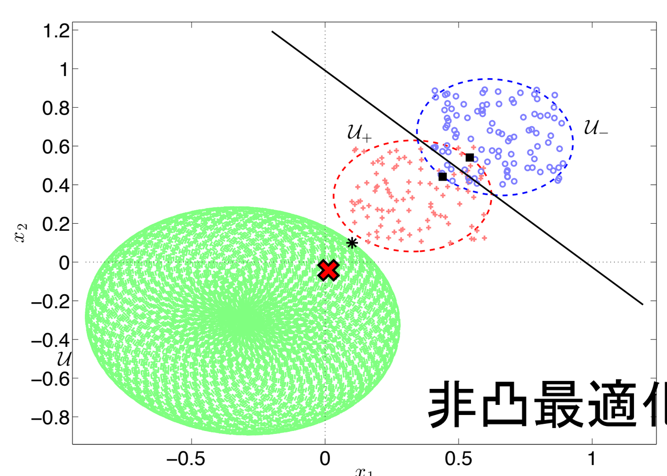
《非凸最適化が必要となる状況》

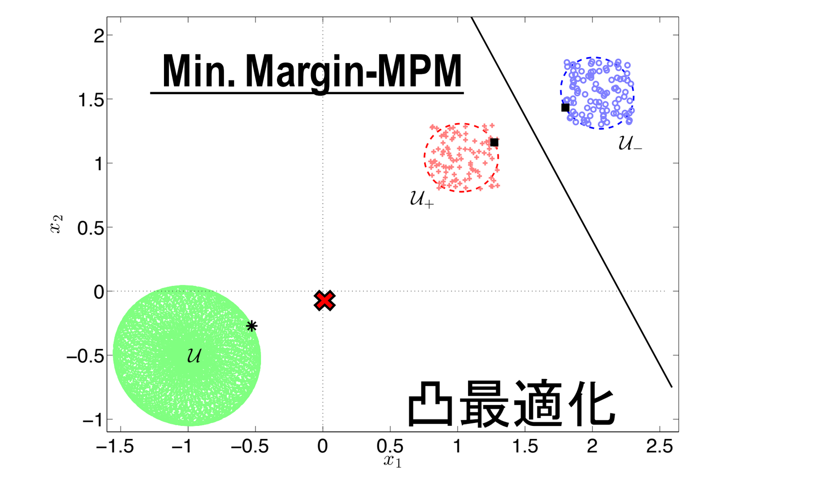
●パラメータ κ が大きい → $\mathcal{U}_+, \mathcal{U}_-$ が交わる → (UCM)は非凸最適化

●多くの既存モデルは，集合が交わらない範囲内に κ をとる．しかし，

*各クラスの平均ベクトルが近い & クラス間のデータ数に偏り
→ 交わらないような κ の範囲はとても小さい（範囲のない場合も）

*パラメータの範囲を広げると，よい判別関数が得られる可能性↑





$\mathcal{U}_+, \mathcal{U}_-$	交わりあり	接する	交わりなし
$\mathcal{U}_+ \ominus \mathcal{U}_-$ 楕円	非凸FLD	Fisher Linear Discriminant (FLD)	Sparse Feature Selection Bhattacharyya ('04)
$\mathcal{U}_+, \mathcal{U}_-$ 楕円	非凸MPM	Minimax Prob. Machine Lanckriet et al. ('02)	Min. Margin-MPM Nath&Bhattacharyya ('07)
$\mathcal{U}_+, \mathcal{U}_-$ 縮退凸包	Eν-SVM Perez-Cruz + ('03)	ν-SVM (= C-SVM) Scholkopf et al. ('00)	
$\mathcal{U}_+, \mathcal{U}_-$ 凸包	—	—	Hard Margin SVM Boser et al. ('92)

研究2：非凸最適化の厳密解法　　研究3：凸最適化の汎用解法

【研究2：非凸FLDのための非凸2次最適化】

非凸FLD（$\mathcal{U}_+ \ominus \mathcal{U}_-$：原点を含む楕円）を非凸最適化問題に変形：

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|^2 \quad \text{s.t.}\quad (\boldsymbol{x} - \boldsymbol{c})^\top A^{-1} (\boldsymbol{x} - \boldsymbol{c}) = \kappa^2$$

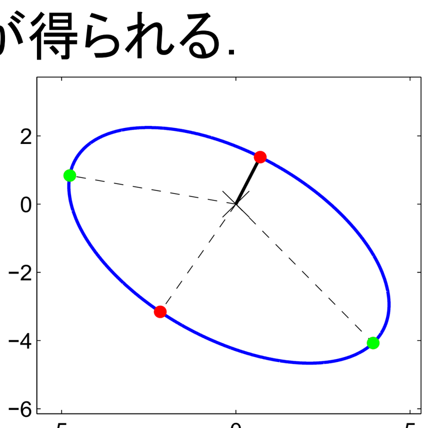
（ただし，$A := \Sigma_+ + \Sigma_-$，$\boldsymbol{c} := \bar{\boldsymbol{x}}_+ - \bar{\boldsymbol{x}}_-$ ）.

- 最適解はKKT条件： $\left.\begin{array}{l} A\boldsymbol{x} = \lambda(\boldsymbol{x} - \boldsymbol{c}) \\ (\boldsymbol{x} - \boldsymbol{c})^\top A^{-1}(\boldsymbol{x} - \boldsymbol{c}) = \kappa^2 \end{array}\right\} \ldots\text{(KKT)}$

  を満たすから，(KKT) を満たす $(\boldsymbol{x}, \lambda)$ を列挙すればよい．

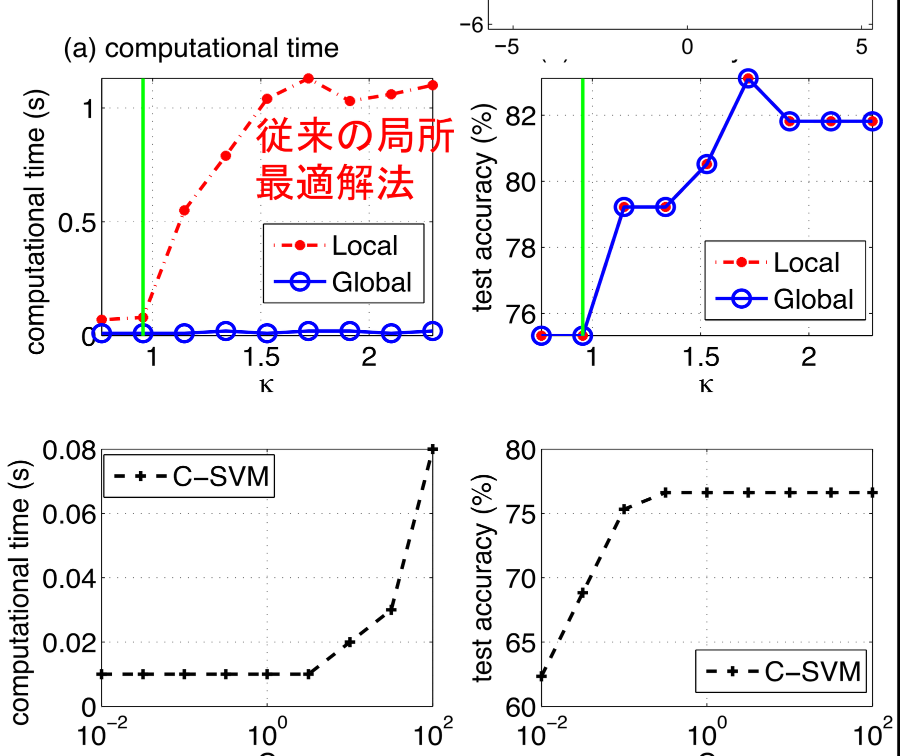
- $\det M(\lambda) = 0$ を解けば，λ が得られ，(KKT)から $\boldsymbol{x}$ が得られる．
  一般化固有値計算を利用した解法（計算量：$O(n^3)$）

  ただし，$M(s) = \begin{bmatrix} \kappa^2 & O & \boldsymbol{c}^\top \\ O & -A & sI - A \\ \boldsymbol{c} & sI - A & O \end{bmatrix}$



- 非凸MPM（2次等式制約が二本の非凸2次最適化モデル）に同様の解法（計算量：$O(n^6)$）
- より一般のCDT問題（多項式時間解法の有無は最近まで未解決．実装可能な解法はない）に対して，$O(n^6)$ の解法を構築．

最初の実装可能な多項式時間解法



【研究3：加速近接勾配法に基づく二値判別の汎用解法】

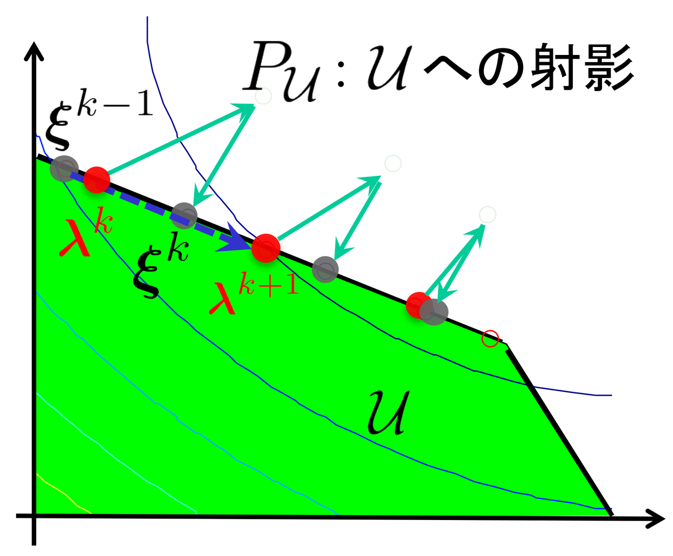
二値判別モデルの統一的な定式化(UCM)に対して汎用解法を提案した．様々な判別モデルを同じプラットフォームで試すことが可能となり，データセットに応じて適切なモデルを選べる．また，汎用解法は個々の既存解法に比べても高速である．

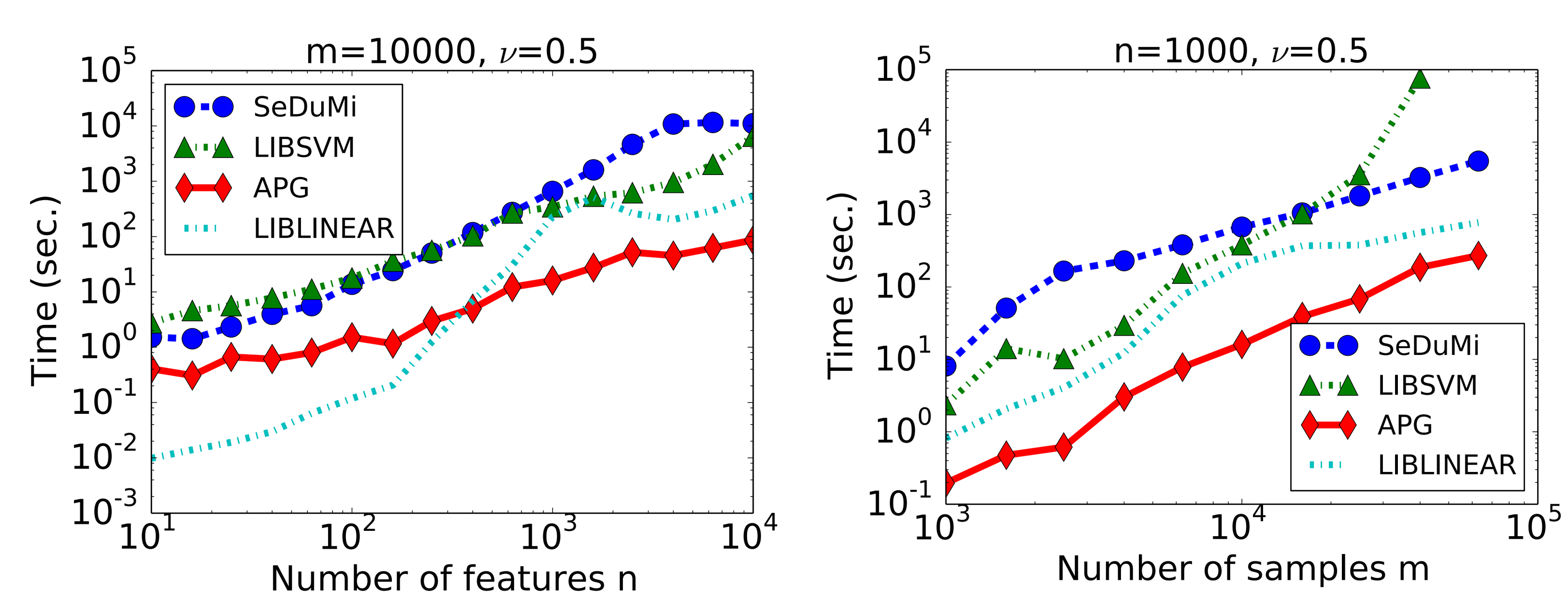
加速近接勾配法 (APG) (Beck &Teboulle, '09) + 高速化テクニック

Step 1. $\xi^k = P_{\mathcal{U}}\left(\boldsymbol{\lambda}^k - \frac{1}{L}\nabla f(\boldsymbol{\lambda}^k)\right)$.

Step 2. $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$.

Step 3. $\boldsymbol{\lambda}^{k+1} = \xi^k + \underbrace{\frac{t_k - 1}{t_{k+1}}(\xi^k - \xi^{k-1})}_{\text{慣性}}$.





人工データに対する計算時間 (SVM)