2023/03/06

フランク・ウルフ法に基づく 1-ノルム正則化ソフトマージン最適化

[春の OR 学会]

◎ 九州大学/理研 AIP　三星　諒太朗

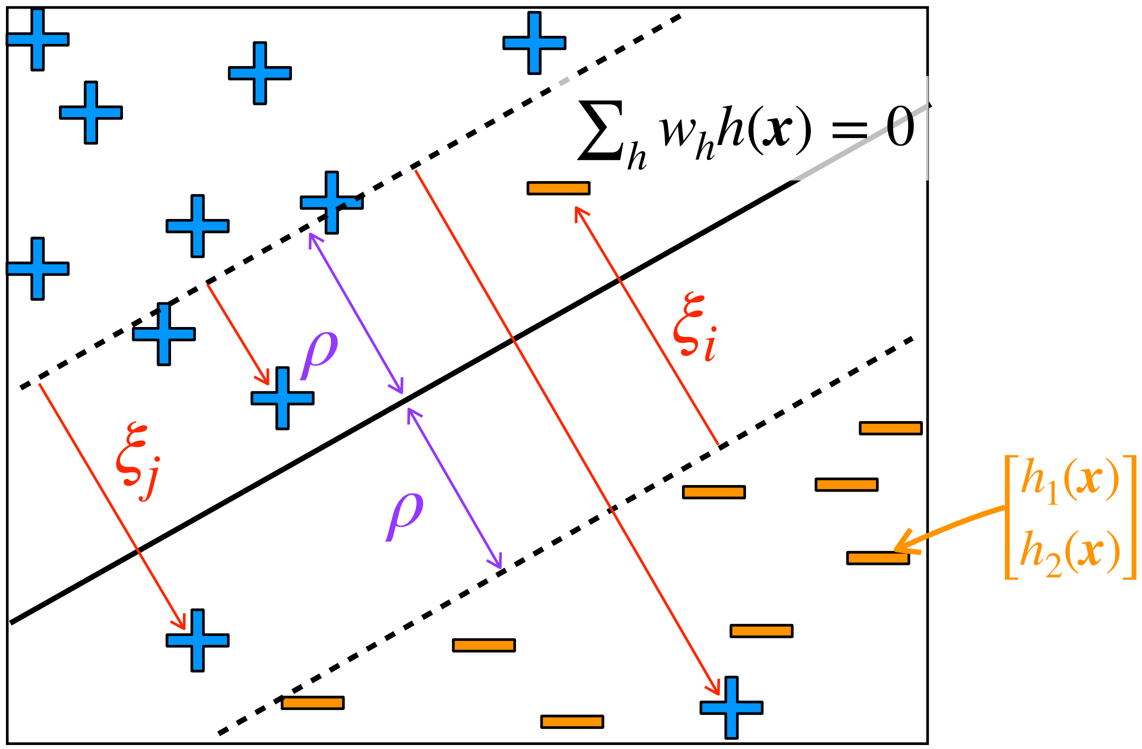
九州大学/理研 AIP　畑埜　晃平

九州大学　瀧本　英二

「ソフト」マージン最適化

$H \subset [-1,+1]^X$ 中の関数の凸結合 $\mathrm{sign} \circ \sum_{h \in H} w_h h$ で

サンプル $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m \subset X \times \{\pm 1\}$ を分類.



「ソフト」マージン最適化問題

$$\max_{\rho, \boldsymbol{w}, \boldsymbol{\xi}} \rho - \frac{1}{\nu} \sum_{i=1}^{m} \xi_i$$

$$\text{s.t. } y_i \sum_{h \in H} w_h h(\boldsymbol{x}_i) \geq \rho - \xi_i, \quad \forall i = 1,2,\ldots,m$$

$$\sum_{h \in H} w_h = 1, \quad \boldsymbol{w} \geq \boldsymbol{0}, \boldsymbol{\xi} \geq \boldsymbol{0}.$$

- ✓ 高々 $\nu \leq m$ 個の点を外れ値とみなし，残りの点でマージン ρ の最大化を目指す.
- ✓ 未知のデータに対しても高い汎化性能を保証 [Bertlett, '98].
- ✓ 線形計画問題だが，一般に H は非常に大きいので直接解くのは非現実的.
  - ✓ H 中の「有用な」関数だけを使うことはできないか？
- ➡ ブースティング（列生成法）によって解く.

ブースティングのプロトコル

ブースティング …「**ブースター**」と「**弱学習者**」との間の繰り返しゲーム.

各ラウンド $t = 1,2,\dots,T$ において：

① S 上の分布 $\boldsymbol{d}_t \in \Delta_{m,\nu}$ を決める.

ブースター → **弱学習者**

ブースター ← **弱学習者**

② 関数 $h_t \in H$ を返す.

$\Delta_{m,\nu} = \{\boldsymbol{d} \in [0,1/\nu]^m \mid \|\boldsymbol{d}\|_1 = 1\}$

直感的には

- **ブースター**は $\{h_1, h_2, \dots, h_{t-1}\}$ 上の
  どの仮説も正答率が小さくなる
  分布を選ぶ.
- **弱学習者**は
  サンプル上の分布 $\boldsymbol{d}_t \in \Delta_{m,\nu}$ に対して
  正答率が一定より大きい h_t を返す.

$$
\begin{aligned}
&\max_{\rho,\boldsymbol{w},\boldsymbol{\xi}} \rho - \frac{1}{\nu}\sum_{i=1}^{m} \xi_i \\
&\text{s.t. } y_i \sum_{h\in H} w_h h(\boldsymbol{x}_i) \geq \rho - \xi_i, \quad \forall i = 1,2,\dots,m \\
&\qquad \sum_{h\in H} w_h = 1, \quad \boldsymbol{w} \geq \boldsymbol{0}, \boldsymbol{\xi} \geq \boldsymbol{0}.
\end{aligned}
$$

⇔ **双対**

$$
\min_{\boldsymbol{d}\in\Delta_{m,\nu}} \max_{h\in H} \sum_{i=1}^{m} d_i y_i h(\boldsymbol{x}_i)
$$

関連研究と主結果

	LPBoost [Demiriz+, 02]	ERLPBoost [Warmuth+, '08]	Cor. ERLPBoost [Shalev-Shwartz+, '10]	本研究
反復回数	$\Omega(m)$	$O\left(\frac{1}{\epsilon^2}\ln\frac{m}{\nu}\right)$	$O\left(\frac{1}{\epsilon^2}\ln\frac{m}{\nu}\right)$	$O\left(\frac{1}{\epsilon^2}\ln\frac{m}{\nu}\right)$
1 反復の計算	LP	CP	LP (Sorting)	$\geq$ LP (Sorting)
実用面	非常に高速	遅い	ERLPB. より遅い	LPB. と同等

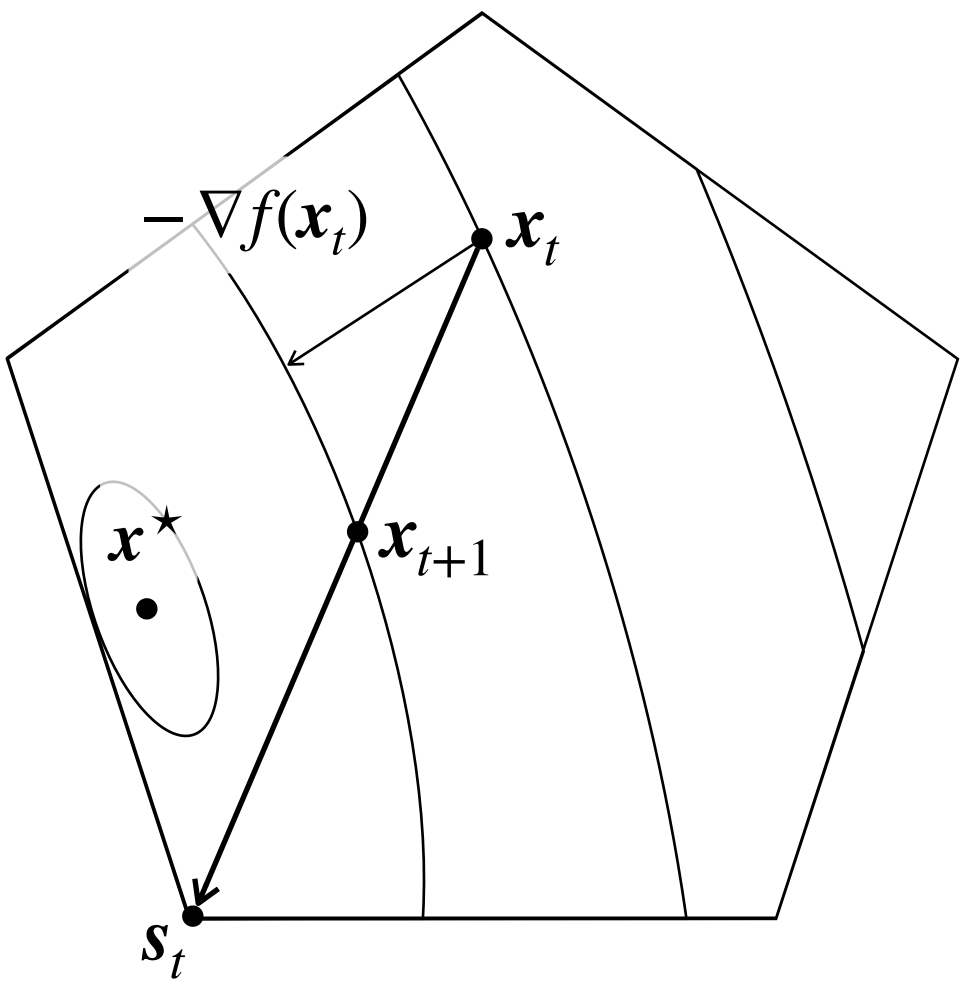
動機：理論保証があり，かつ実用上も高速なアルゴリズムが欲しい

主結果

- 上記のアルゴリズム（後で詳述）は
  すべて フランク・ウルフのアルゴリズムと見做せることを示した.
- 理論保証がある，**一般的な枠組**を提案.

フランク・ウルフのアルゴリズム [Marguerite+, '56]

閉凸集合上での凸関数の最小化を行う，一次の最適化アルゴリズム.



$$\min_{x \in S} f(\boldsymbol{x})$$

各ラウンド $t = 1,2,\ldots,T$ において：

1. $\boldsymbol{s}_t \in \arg\max_{s\in S} \boldsymbol{s} \cdot [-\nabla f(\boldsymbol{x}_t)]$
2. $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \lambda_t(\boldsymbol{s}_t - \boldsymbol{x}_t)$, $\lambda_t \in [0,1]$

λ_t の例

- $2/(t+1)$
- $f(\boldsymbol{x}_{t+1})$ の $\boldsymbol{x}_t$ 周りの二次近似の最小解
- 線形探索

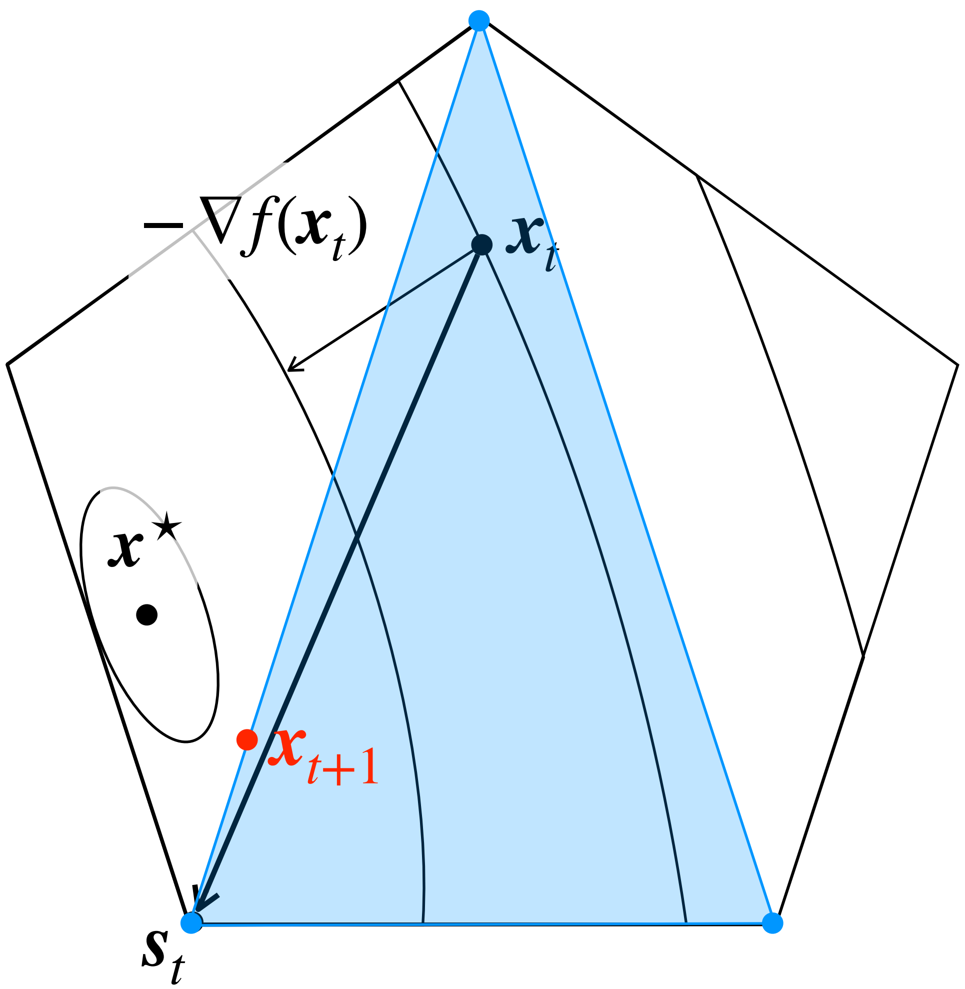
定理 [Jaggi, '13]　$\forall \boldsymbol{x}, \boldsymbol{y},\quad f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + (\boldsymbol{y}-\boldsymbol{x})^\top \nabla f(\boldsymbol{x}) + \frac{\eta}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2$

f が **η-平滑**であるとき，適切な $\{\lambda_t\}_t$ のもと

高々 $T = O(\eta/\epsilon)$ 回の反復で ϵ-近似解に収束.

フランク・ウルフのアルゴリズム [Marguerite+, '56]

閉凸集合上での凸関数の最小化を行う，一次の最適化アルゴリズム.



$$\min_{x \in S} f(\boldsymbol{x})$$

各ラウンド $t = 1,2,\ldots,T$ において：

1. $\boldsymbol{s}_t \in \arg\max_{s \in S} \boldsymbol{s} \cdot [-\nabla f(\boldsymbol{x}_t)]$
2. $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \lambda_t(\boldsymbol{s}_t - \boldsymbol{x}_t),\ \lambda_t \in [0,1]$

その他の λ_t の例：FCFW

- 過去に求めた端点 $\{\boldsymbol{s}_k\}_{k=1}^t \subset S$ の凸包上での f の最小解

定理 [Jaggi, '13]　$\forall \boldsymbol{x}, \boldsymbol{y},\quad f(\boldsymbol{y}) \le f(\boldsymbol{x}) + (\boldsymbol{y} - \boldsymbol{x})^\top \nabla f(\boldsymbol{x}) + \frac{\eta}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2$

f が η-**平滑**であるとき，適切な $\{\lambda_t\}_t$ のもと

高々 $T = O(\eta/\epsilon)$ 回の反復で ϵ-近似解に収束.

LPBoost [Demiriz+, '02] のアイデア

ソフトマージン最適化問題の双対問題を行生成法で解く

$$\min_{\gamma, \boldsymbol{d} \in \Delta_{m,\nu}} \gamma \qquad \text{— (♤)}$$

$$\text{s.t.} \sum_{i=1}^{m} d_i y_i h(\boldsymbol{x}_i) \leq \gamma, \quad \forall h \in H$$

Lagrange
双対問題

$\Longleftrightarrow$

$$\max_{\rho, \boldsymbol{w}, \boldsymbol{\xi}} \rho - \frac{1}{\nu} \sum_{i=1}^{m} \xi_i$$

$$\text{s.t. } y_i \sum_{h \in H} w_h h(\boldsymbol{x}_i) \geq \rho - \xi_i, \quad \forall i = 1,2,\ldots,m$$

$$\boldsymbol{w} \in \Delta_H, \quad \boldsymbol{\xi} \geq \boldsymbol{0}.$$

$\Delta_H = \{\boldsymbol{w} \in [0,1]^H \mid \|\boldsymbol{w}\|_1 = 1\}$

- (♤) も H が大きいときは解くのが難しい.
- $H_0 = \emptyset$ から始め, t ラウンド目において
  $H_{t-1} = \{h_1, h_2, \ldots, h_{t-1}\}$ 上での (♤) の最適解 $\boldsymbol{d}_t$ を弱学習者に渡す.

各ラウンド $t = 1,2,\ldots,T$ において：

① H_{t-1} 上の (♤) の解 $\boldsymbol{d}_t \in \Delta_{m,\nu}$.

ブースター $\rightleftarrows$ 弱学習者

② 関数 $h_t \in H$ を返す.

ERLPBoost [Warmuth+, '08] のアイデア

LPBoost + エントロピー正則化

$$\min_{\gamma, \boldsymbol{d} \in \Delta_{m,\nu}} \gamma + \frac{1}{\eta} \sum_{i=1}^{m} d_i \ln(m d_i) \qquad \text{— } (\diamondsuit)$$

$$\text{s.t. } \sum_{i=1}^{m} d_i y_i h(\boldsymbol{x}_i) \leq \gamma, \quad \forall h \in H$$

- $\eta > 0$ を十分大きく取れば，正則化項は目的関数値にあまり影響を与えない
- $H_0 = \emptyset$ から始め，t ラウンド目において

  $H_{t-1} = \{h_1, h_2, \ldots, h_{t-1}\}$ 上での $(\diamondsuit)$ の最適解 $\boldsymbol{d}_t$ を弱学習者に渡す.

各ラウンド $t = 1,2,\ldots,T$ において：

① H_{t-1} 上の $(\diamondsuit)$ の解 $\boldsymbol{d}_t \in \Delta_{m,\nu}$.

ブースター → 弱学習者

ブースター ← 弱学習者

② 関数 $h_t \in H$ を返す.

提案手法のアイデア

$A \in [-1, +1]^{m \times H}$ を，第 (i, h) 成分が $A_{i,h} = y_i h(\boldsymbol{x}_i)$ である行列とする．

$$\min_{\gamma, \boldsymbol{d} \in \Delta_{m,\nu}} \gamma + \frac{1}{\eta} \sum_{i=1}^{m} d_i \ln(m d_i)$$

s.t. $\sum_{i=1}^{m} d_i y_i h(\boldsymbol{x}_i) \le \gamma, \quad \forall h \in H$

$$f(\boldsymbol{d}) \triangleq \begin{cases} 0, & \boldsymbol{d} \in \Delta_{m,\nu} \\ +\infty, & \boldsymbol{d} \notin \Delta_{m,\nu} \end{cases}$$

$$\tilde{f}(\boldsymbol{d}) \triangleq f(\boldsymbol{d}) + \frac{1}{\eta} \sum_{i=1}^{m} d_i \ln(m d_i)$$

① 制約を目的関数に

$$\min_{\boldsymbol{d}} \max_{h \in H} (\boldsymbol{d}^\top A)_h + \tilde{f}(\boldsymbol{d})$$

② **Fenchel** 双対問題

$$\max_{\boldsymbol{w} \in \Delta_H} -\tilde{f}^\star(-A\boldsymbol{w})$$

$$\min_{\boldsymbol{w} \in \Delta_H} \tilde{f}^\star(-A\boldsymbol{w})$$

③ フランク・ウルフ法で解く

最適解にのみ興味がある

Point!

- $\tilde{f}^\star$ はフランク・ウルフ法の仮定を満たす

f の Fenchel 共役関数 $f^\star$： $f^\star(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta}} \boldsymbol{\mu}^\top \boldsymbol{\theta} - f(\boldsymbol{\theta})$

提案手法のアイデア

$$\min_{w \in \Delta_H} \tilde{f}^\star(-Aw)$$

$(Aw)_i = y_i \sum_{h \in H} w_h h(x_i)$

1. 勾配ベクトルの計算

$$\begin{aligned} d_t &= \nabla \tilde{f}^\star(-Aw_{t-1}) \\ &= \arg\max_d d^\top(-Aw_{t-1}) - \tilde{f}(d) \\ &= \arg\min_{d \in \Delta_{m,\nu}} d^\top A w_{t-1} + \frac{1}{\eta} \sum_{i=1}^{m} d_i \ln(m d_i) \end{aligned}$$

2. 端点の計算

$$\begin{aligned} h_t &\in \arg\min_{h \in H} d_t^\top(-Ae_h) \\ &= \arg\max_{h \in H} (d_t^\top A)_h \end{aligned}$$

arg max でなくともある値以上の h が得られるならば収束性を言える（後述）

3. 重みの更新

$$w_t = w_{t-1} + \lambda_t(e_{h_t} - w_{t-1})$$

通常の FW だけでなく，任意のアルゴリズムを組み込むことが可能！

フランク・ウルフ法

$$\min_{x \in S} f(x)$$

各ラウンド $t = 1,2,\ldots,T$ において：

1. **Get** $\nabla f(x_{t-1})$
2. $s_t \in \arg\min_{s \in S} s \cdot \nabla f(x_{t-1})$
3. $x_t = x_{t-1} + \lambda_t(s_t - x_{t-1})$, $\lambda_t \in [0,1]$

フランク・ウルフ法でブースティングを説明

ブースティング

$$\min_{w \in \Delta_{m,\nu}} \tilde{f}^\star(-A\boldsymbol{w})$$

各ラウンド $t = 1,2,\ldots,T$ において：

1. S 上の分布 $\boldsymbol{d}_t \in \Delta_{m,\nu}$ を決める.

ブースター → 弱学習者

2. 関数 $h_t \in H$ を返す.

フランク・ウルフ法

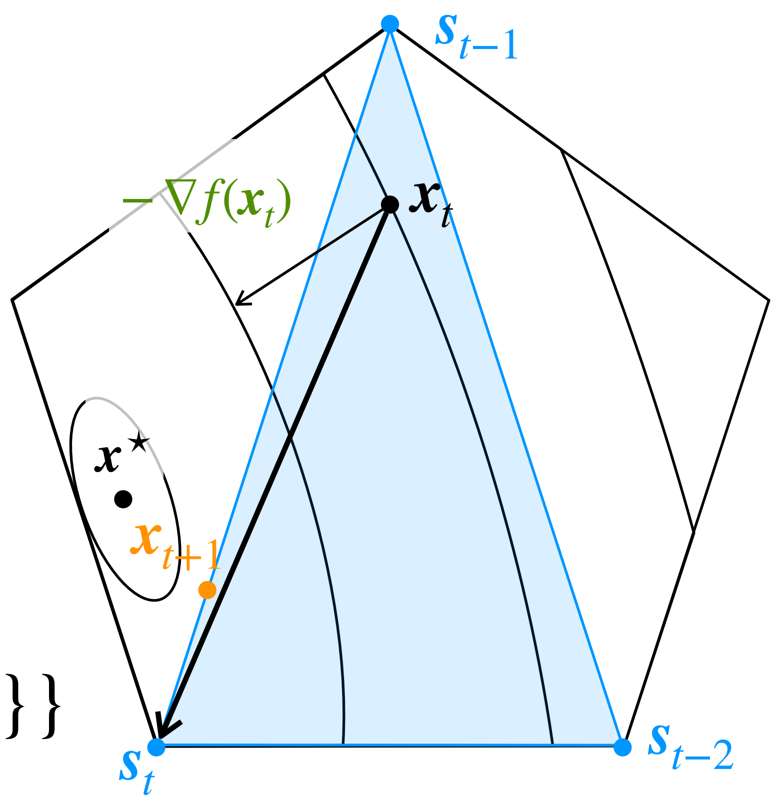
$$\min_{\boldsymbol{x} \in S} f(\boldsymbol{x})$$

各ラウンド $t = 1,2,\ldots,T$ において：

1. **Get** $\nabla f(\boldsymbol{x}_{t-1})$
2. $\boldsymbol{s}_t \in \arg\max_{\boldsymbol{s} \in S} \boldsymbol{s} \cdot [-\nabla f(\boldsymbol{x}_{t-1})]$
3. $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \lambda_t(\boldsymbol{s}_t - \boldsymbol{x}_{t-1})$, $\lambda_t \in [0,1]$

例：ERLPBoost [Warmuth+, '08]

1. $\boldsymbol{d}_t = \arg\min_{\boldsymbol{d}} \max_{h \in \{h_1,h_2,\ldots,h_{t-1}\}} (\boldsymbol{d}^\top A)_h + \tilde{f}(\boldsymbol{d})$
   $= \nabla \tilde{f}^\star(\boldsymbol{\theta}_{t-1})$, 但し $\boldsymbol{\theta}_{t-1} = -A\boldsymbol{w}_{t-1}$
2. $\boldsymbol{e}_{h_t} \in \arg\min_{\boldsymbol{e}_h : h \in H} \boldsymbol{d}^\top(-A\boldsymbol{e}_h) = \arg\max_{\boldsymbol{e}_h : h \in H} \boldsymbol{d}^\top A\boldsymbol{e}_h$
3. $\boldsymbol{w}_t = \arg\min_{\boldsymbol{w} \in \Delta_{H_t}} \tilde{f}^\star(-A\boldsymbol{w})$,
   但し $\Delta_{H_t} = \{\boldsymbol{w} \in \Delta_H \mid w_h > 0 \implies h \in \{h_1, h_2, \ldots, h_t\}\}$



フランク・ウルフ法でブースティングを説明

ブースティング

$$\min_{w \in \Delta_{m,\nu}} f^\star(-A\boldsymbol{w})$$

各ラウンド $t = 1,2,\ldots,T$ において：

1. S 上の分布 $\boldsymbol{d}_t \in \Delta_{m,\nu}$ を決める.

ブースター ⇄ 弱学習者

2. 関数 $h_t \in H$ を返す.

フランク・ウルフ法

$$\min_{\boldsymbol{x} \in S} f(\boldsymbol{x})$$

各ラウンド $t = 1,2,\ldots,T$ において：

1. Get $\nabla f(\boldsymbol{x}_{t-1})$
2. $\boldsymbol{s}_t \in \arg\max_{s \in S} \boldsymbol{s} \cdot [-\nabla f(\boldsymbol{x}_{t-1})]$
3. $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \lambda_t(\boldsymbol{s}_t - \boldsymbol{x}_{t-1}),\ \lambda_t \in [0,1]$

例：LPBoost [Demiriz+, '02]

$$f(\boldsymbol{d}) = \begin{cases} 0 & \boldsymbol{d} \in \Delta_{m,\nu} \\ +\infty & \boldsymbol{d} \notin \Delta_{m,\nu} \end{cases}$$

1. $\boldsymbol{d}_t \in \arg\min_{\boldsymbol{d}} \max_{h \in \{h_1,h_2,\ldots,h_{t-1}\}} (\boldsymbol{d}^\top A)_h + f(\boldsymbol{d})$
   $= \nabla \tilde{f}^\star(\boldsymbol{\theta}_{t-1})$, 但し $\boldsymbol{\theta}_{t-1} = -A\boldsymbol{w}_{t-1}$
2. $\boldsymbol{e}_{h_t} \in \arg\min_{\boldsymbol{e}_h : h \in H} \boldsymbol{d}^\top(-A\boldsymbol{e}_h) =$
3. $\boldsymbol{w}_t = \arg\min_{\boldsymbol{w} \in \Delta_{H_t}} f^\star(-A\boldsymbol{w})$,
   但し $\Delta_{H_t} = \{\boldsymbol{w} \in \Delta_H \mid w_h > 0 \implies h \in \{h_1, h_2, \ldots, h_t\}\}$

$f^\star$ は平滑性を持たないので,
フランク・ウルフの保証を使えない.

提案手法

フランク・ウルフ法の解析をよくみると次のことがわかる.

観察 各ラウンド t において

$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \frac{2}{t+1}(\boldsymbol{s}_t - \boldsymbol{x}_{t-1})$ よりも関数値が小さくなる $\hat{\boldsymbol{x}}_t$ を使った場合,

$\boldsymbol{x}_t$ を使った場合と同じ収束率が保証される.

1. $\boldsymbol{d}_t = \nabla \tilde{f}^\star(-A\boldsymbol{w}_{t-1}) \in \Delta_{m,\nu}$

ブースター → 弱学習者

2. 関数 $h_t \in H$ を返す.

$\sum_{i=1}^m d_i y_i h(\boldsymbol{x}_i) \geq g$ を満たす h を返す

3-1. $\boldsymbol{w}_t^{\mathcal{F}} = \boldsymbol{w}_{t-1} + \frac{2}{t+1}(\boldsymbol{s}_t - \boldsymbol{w}_{t-1})$

3-2. 任意に $\boldsymbol{w}_t^{\mathcal{A}}$ を決める.

3-3. $\boldsymbol{w}_t^{\mathcal{F}}, \boldsymbol{w}_t^{\mathcal{A}}$ のうち, $\tilde{f}^\star \circ (-A)$ をより小さくする方を $\boldsymbol{w}_t$ として採用.

定理 提案手法は $T = O\left(\frac{2}{\epsilon^2} \ln \frac{m}{\nu}\right)$ 反復後 $-f^\star(-A\boldsymbol{w}_T) \geq g - \epsilon$ を達成.

提案手法: 実用例

フランク・ウルフ法の解析をよくみると次のことがわかる.

観察 各ラウンド t において

$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \frac{2}{t+1}(\boldsymbol{s}_t - \boldsymbol{x}_{t-1})$ よりも関数値が小さくなる $\hat{\boldsymbol{x}}_t$ を使った場合,

$\boldsymbol{x}_t$ を使った場合と同じ収束率が保証される.

1. $\boldsymbol{d}_t = \nabla \tilde{f}^\star(-A\boldsymbol{w}_{t-1}) \in \Delta_{m,\nu}$

ブースター

弱学習者

2. 関数 $h_t \in H$ を返す.

3-1. $w_t^{\mathcal{F}} = w_{t-1} + \frac{2}{t+1}(s_t - w_{t-1})$

$\sum^m \ d_i y_i h(x_i) \geq g$ を満たす h を返す

3-2. 任意に $w_t^{\mathcal{A}}$ を決める.

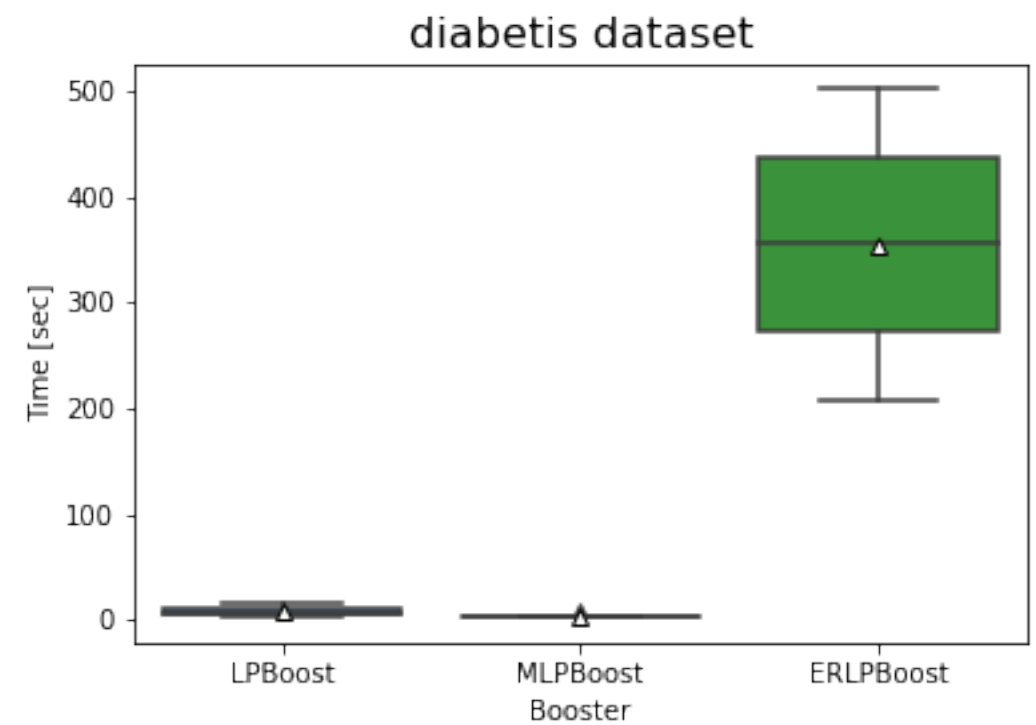
3-3. $w_t^{\mathcal{F}}, w_t^{\mathcal{A}}$ のうち, $\tilde{f}^\star \circ (-A)$

- $\arg\max_{w \in \Delta_{H_t}} -\tilde{f}^\star(-Aw)$ → ERLPBoost
- LPBoost の双対解 → MLPBoost (提案法)

実験

Setting.

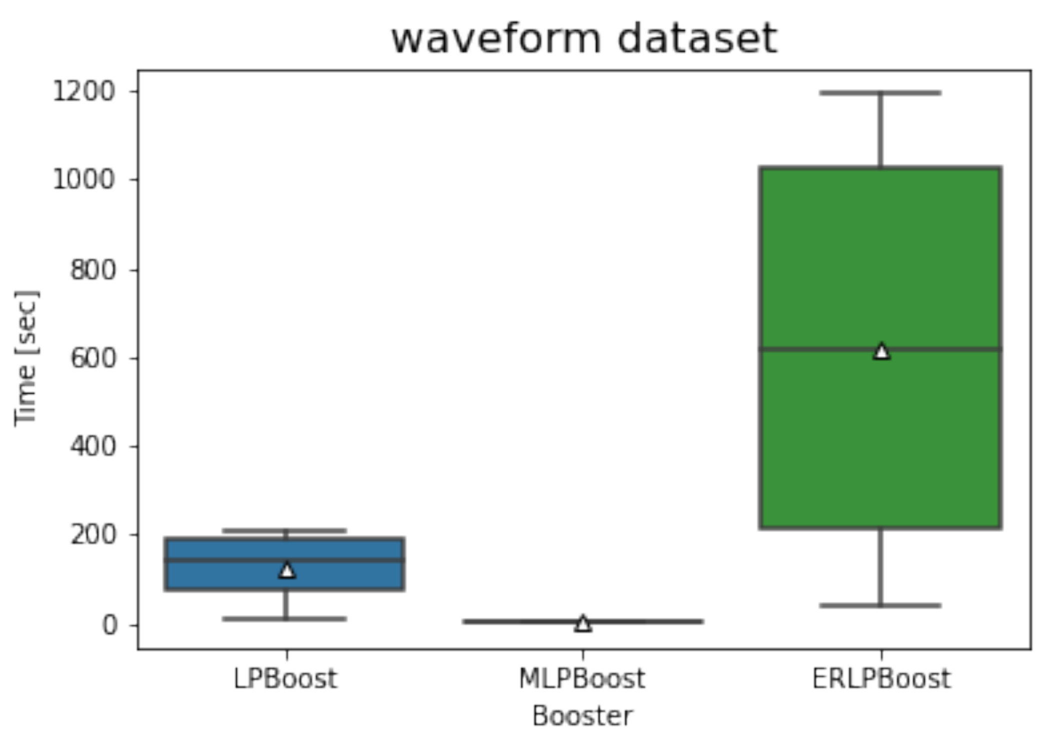
- 最適化ソルバ: **Gurobi 9.0.1**
- 弱学習器: 深さ２の決定木（正答率基準）
- $\epsilon = 0.01$
- $\nu \in \{0.01, 0.02, 0.05, 0.1\}$ で計測

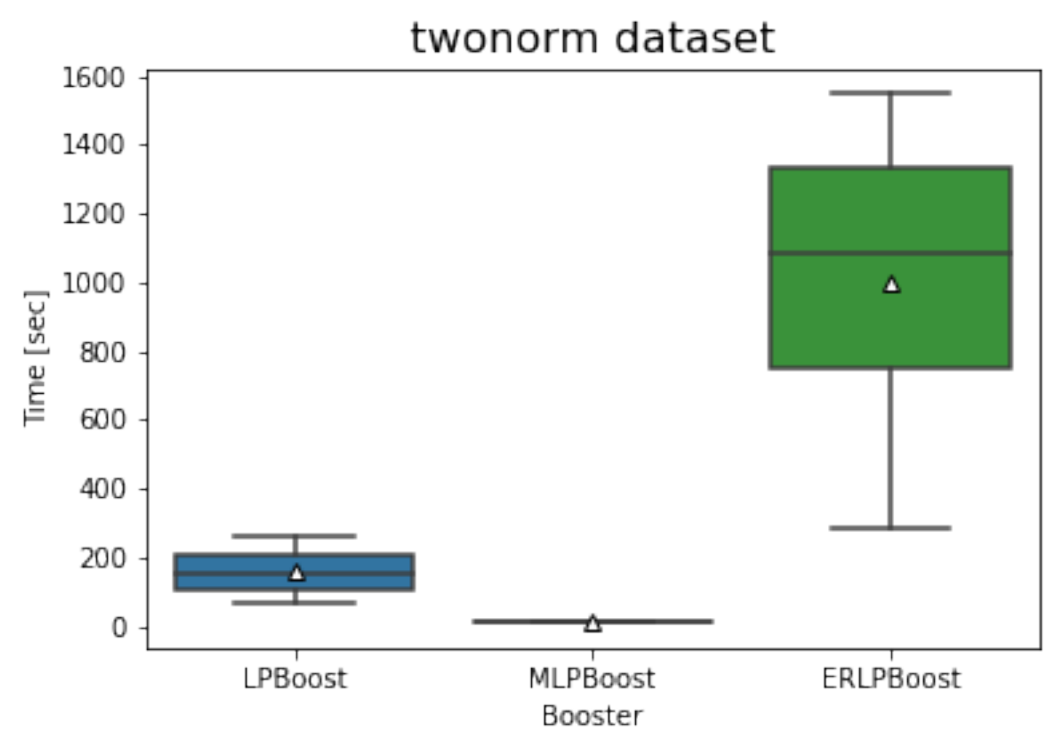


LPBoost

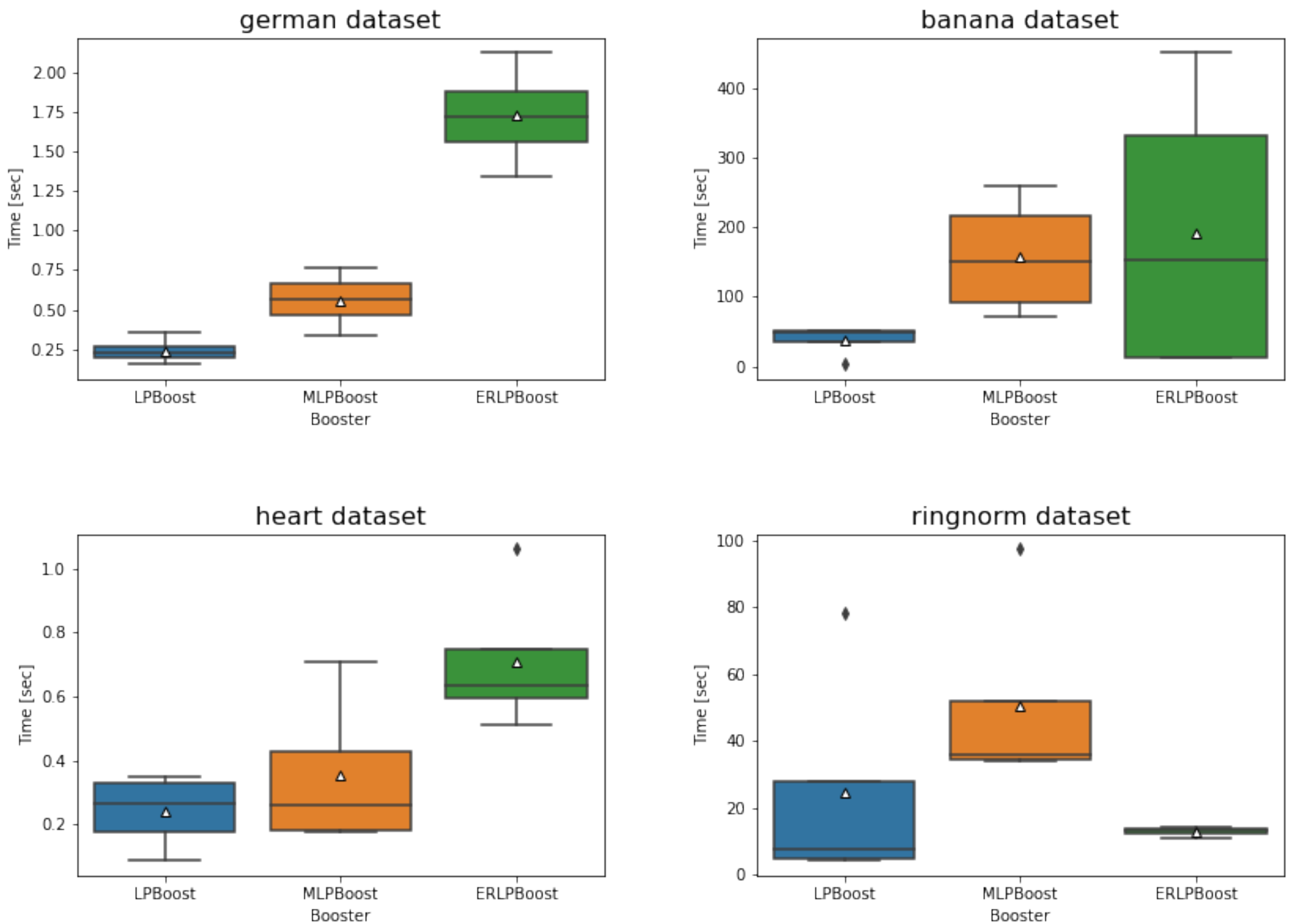
MLPBoost
（本研究）

ERLPBoost





実験



LPBoost

MLPBoost
(本研究)

ERLPBoost

まとめと今後の課題

- いくつかのブースティングアルゴリズムは
  フランク・ウルフ法で説明できることを示した.
- 理論保証があり，かつ高速なブースティングアルゴリズムを提案.
  - LPBoost より（常に）速くできるか？
- 強凸性を仮定すれば線形収束するか？