ID: 1-094

フランク・ウルフとしてのブースティング

https://arxiv.org/abs/2209.10831

三星 諒太朗 九州大学/理研 AIP ryotaro.mitsuboshi@inf.kyushu-u.ac.jp
畑埜 晃平 九州大学/理研 AIP hatano@inf.kyushu-u.ac.jp
瀧本 英二 九州大学 eiji@inf.kyushu-u.ac.jp

エッジ最小化とソフトマージン最適化

入力： サンプル $S = ((x_i, y_i))_{i=1}^m \in (\mathcal{X} \times \{\pm 1\})^m$，パラメータ $\nu \in [1, m]$，仮説集合 $\mathcal{H} \subset [-1, +1]^{\mathcal{X}}$．

エッジ最小化問題

$$\min_{d} \left[f(d) + \max_{h \in \mathcal{H}} (d^\top A)_h \right], \quad f(d) = \begin{cases} 0 & d \in \Delta_{m,\nu} \\ +\infty & \text{otherwise} \end{cases} \quad (1)$$

ただし，$\Delta_{m,\nu} = \{ d \in [0, 1/\nu]^m \mid \|d\|_1 = 1 \}$ で，$A = (y_i h(x_i))$．
$(d^\top A)_h = \sum_i d_i y_i h(x_i)$ は仮説 $h \in \mathcal{H}$ の，分布 d に関するエッジ（正答率）．

$\Updownarrow$ Fenchel 双対問題（双対ギャップ 0）

ソフトマージン最適化問題

$$\max_{w \in \Delta_{\mathcal{H},1}} \left[-f^\star(-Aw) := \min_{d \in \Delta_{m,\nu}} d^\top A w \right] \quad (2)$$

ここで，$f^\star(\theta) = \sup_d \left[\theta^\top d - f(d) \right]$ は f の Fenchel 共役関数．
マージンが大きい $\implies$ 汎化誤差の保証が良い．

出力： $\sum_{h \in \mathcal{H}} \bar{w}_h h$，ただし $\bar{w}$ は (2) の最適解．
$\mathcal{H}$ が非常に大きいとき，LP ソルバを使っても (2) を解くのは困難．

ブースティング

各ラウンド $t = 0, 1, 2, \ldots, T$ において，
1. サンプル上の分布 $d_t \in \Delta_{m,\nu}$ を決める．
2. 弱学習器から $(d_t^\top A)_{h_{t+1}} \geq g$ を満たす仮説 $h_{t+1} \in \mathcal{H}$ を得る．（g は未知）

(2) の ϵ-近似解 $H_T = \sum_{t=1}^T w_{T,t} h_t$ を出力．

- LPBoost

$$d_t^L \leftarrow \arg\min_{d} \max_{h \in \{h_1, h_2, \ldots, h_t\}} (d^\top A)_h + f(d)$$

- ERLPBoost
  $\eta = \frac{2}{\epsilon} \ln \frac{m}{\nu}$ として，

$$d_t^E \leftarrow \arg\min_{d} \max_{h \in \{h_1, h_2, \ldots, h_t\}} (d^\top A)_h + \underbrace{f(d) + \frac{1}{\eta} \sum_{i=1}^m d_i \ln(m d_i)}_{=: \tilde{f}(d)}$$

- C-ERLPBoost
  $w_t^C = w_{t-1}^C + \lambda_{t-1}(e_{h_t} - w_{t-1}^C) \in CH(\{e_{h_1}, e_{h_2}, \ldots, e_{h_t}\})$ として，

$$d_t^C \leftarrow \arg\min_{d} d^\top A w_t^C + \tilde{f}(d)$$

	LPBoost	ERLPBoost	C-ERLPBoost	提案スキーム
T	$\Omega(m)$	$O(\frac{1}{\epsilon^2} \ln \frac{m}{\nu})$	$O(\frac{1}{\epsilon^2} \ln \frac{m}{\nu})$	$O(\frac{1}{\epsilon^2} \ln \frac{m}{\nu})$
反復一回	LP	CP	Sorting	$\geq$ Sorting

目標．

理論保証があり，かつ実用的なブースティングアルゴリズムが欲しい．

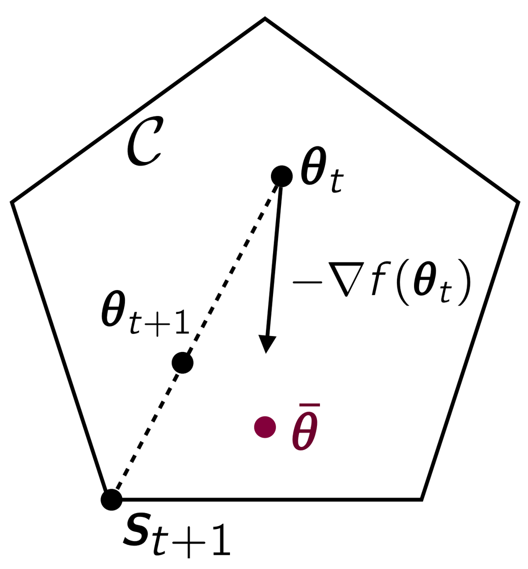
フランク・ウルフ（FW）のアルゴリズム

次の形の最適化問題を解く一次の繰り返し法．

$$\min_{\theta \in \mathcal{C}} f(\theta)$$

$\mathcal{C} \subset \mathbb{R}^m$ は閉凸集合，
$f : \mathcal{C} \to \mathbb{R}$ は η-平滑な凸関数．

各ラウンド $t = 0, 1, 2, \ldots, T$ において，
1. $s_{t+1} \leftarrow \arg\min_{s \in \mathcal{C}} s^\top \nabla f(\theta_{t+1})$ を計算．
2. $\theta_{t+1} = \theta_t + \lambda_t (s_{t+1} - \theta_t), \quad \lambda_t \in [0, 1]$．



$\lambda_t \in [0, 1]$ には様々な選択肢がある．
例：$\lambda_t = \frac{2}{t+2}$ (Classic step)，$\lambda_t = \mathrm{clip}_{[0,1]} \frac{\nabla f(\theta_t)^\top (\theta_t - s_{t+1})}{\eta \|\theta_t - s_{t+1}\|^2}$ (Short step)，等．

- $f(\theta_T) - f(\bar{\theta}) = O(\eta / T)$．
- 一反復あたりの計算が非常に高速．$\mathcal{C}$ が凸多面体→ LP．

Fully-corrective FW: $\theta_{t+1} \in \arg\min_{\theta \in CH(\{s_1, s_2, \ldots, s_{t+1}\})} f(\theta)$
- 過去に選んだ端点集合の凸包上で最適化問題を解く．
- 最も目的関数値を改善．

フランク・ウルフ（FW）としてのブースティング

(1) で f の代わりに $\tilde{f}$ を採用し，その Fenchel 双対問題を考える．

$$\max_{\theta \in -A\Delta_{\mathcal{H},1}} -\tilde{f}^\star(\theta) := \max_{w \in \Delta_{\mathcal{H},1}} -\tilde{f}^\star(-Aw) = \max_{w \in \Delta_{\mathcal{H},1}} - \left[\max_d -d^\top A w - \tilde{f}(d) \right]$$

ここで，$-A\Delta_{\mathcal{H},1} = \{ -Aw \mid w \in \Delta_{\mathcal{H},1} \} \subset \mathbb{R}^m$ とおいた．
- 双対ギャップ 0．
- $\tilde{f}^\star$ は L_∞-ノルムに関して η-平滑（$\tilde{f}$ は L_1-ノルムに関して $1/\eta$-強凸）．
- 分布 d は，ある $\theta = -Aw$ における $f^\star$，$\tilde{f}^\star$ の（劣）勾配ベクトル．
  C-ERLPB. $d_t^C = \nabla \tilde{f}^\star(-Aw_t^C)$．
  ERLPB. $d_t^E = \nabla \tilde{f}^\star(-Aw_t^E)$，ただし $w_t^E = \arg\min_{w \in CH(\{e_{h_1}, e_{h_2}, \ldots, e_{h_t}\})} \tilde{f}^\star(-Aw)$．
  LPB. $d_t^L \in \partial f^\star(-Aw_t^L)$，ただし $w_t^L \in \arg\min_{w \in CH(\{e_{h_1}, e_{h_2}, \ldots, e_{h_t}\})} f^\star(-Aw)$．
- FW の LP オラクルは最大エッジを返す弱学習器．

$$h_{t+1} \in \arg\max_{h \in \mathcal{H}} (d_t^\top A)_h \iff e_{h_{t+1}} \in \arg\max_{e \in \Delta_{\mathcal{H},1}} d_t^\top A e = \arg\min_{\theta \in -A\Delta_{\mathcal{H},1}} \theta^\top d_t$$

定理．
LPBoost，ERLPBoost，そして C-ERLPBoost は FW アルゴリズム．

提案スキーム

各ラウンド $t = 0, 1, 2, \ldots, T$ において，
1. 分布 $d_t = \nabla \tilde{f}^\star(-Aw_t) = \arg\min_{d \in \Delta_{m,\nu}} d^\top A w_t + \frac{1}{\eta} \sum_{i=1}^m d_i \ln(m d_i)$ を計算．
2. 仮説 $h_{t+1} \in \mathcal{H}$ を得る．
3. $\min_{q \leq t} (d_q^\top A)_{h_{q+1}} + \tilde{f}^\star(-Aw_t) \leq \epsilon/2 \implies$ **break**．
4. $w_{t+1}^F \in \Delta_{\mathcal{H},1}$ を FW的に計算．
5. $w_{t+1}^B \in \Delta_{\mathcal{H},1}$ を 任意に計算．
6. $w_{t+1} \in \arg\min_{w \in \{w_{t+1}^F, w_{t+1}^B\}} \tilde{f}^\star(-Aw)$

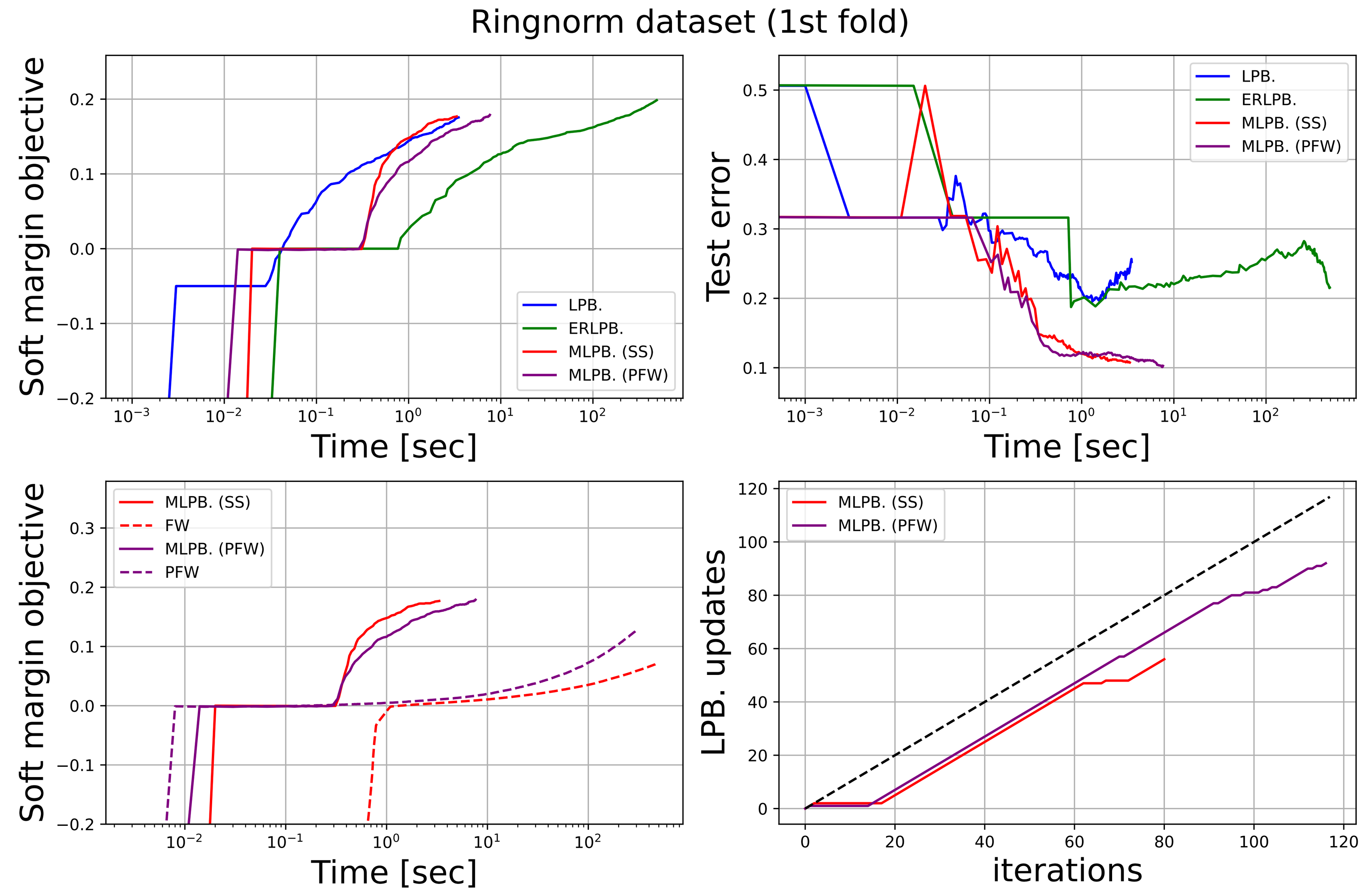
定理．
適切な w_{t+1}^F の更新の下，提案スキームは高々 $T = O(\frac{1}{\epsilon^2} \ln \frac{m}{\nu})$ ラウンド後に $-f^\star(-Aw_T) \geq g - \epsilon$ を達成．

- $d_t = \nabla \tilde{f}^\star(-Aw_t)$ は $O(m \ln m)$ 時間で計算可能．
- 各反復 t において弱学習器が $h_{t+1} \in \arg\max_{h \in \mathcal{H}} (d_t^\top A)_h$ を返す．
  $\implies$ (2) の ϵ-近似解を出力．
- w_{t+1}^F の計算は各ラウンドで一定量の改善を保証する FW．
  $\implies$ FW にも自由度がある．
- $w_{t+1}^B = \arg\min_{w \in CH(\{e_{j_1}, e_{j_2}, \ldots, e_{j_{t+1}}\})} \tilde{f}^\star(-Aw)$ で ERLPBoost になる．

実験

Gunnar Rätsch のベンチマークデータセット [a] で比較．
- パラメータ：$\epsilon = 0.01$，$\nu = 0.1m$．
- 弱学習器：エッジが最大となる深さ 2 の決定木を返す．

Ringnorm dataset (1st fold)



まとめと今後の課題

- ブースティングを FW として捉えた．
- FW に基づく新たなブースティングの枠組みを提案．
  - LPBoost を組み込み，理論保証付きで高速なブースティングを実現．
- 最悪ケースを除き，LPBoost は高速．
  $\implies$ 収束保証つきでより高速なブースティングはできるか？

[a] http://theoval.cmp.uea.ac.uk/~gcc/matlab/default.html#benchmarks.