# Boosting as Frank-Wolfe

Ryotaro Mitsuboshi
Kyushu University/Riken AIP
ryotaro.mitsuboshi@inf.kyushu-u.ac.jp

Kohei Hatano
Kyushu University/Riken AIP
hatano@inf.kyushu-u.ac.jp

Eiji Takimoto
Kyushu University
eiji@inf.kyushu-u.ac.jp

KYUSHU UNIVERSITY

## Edge minimization & Soft margin optimization

**Input:** A sample $S = ((x_i, y_i))_{i=1}^m \in (\mathcal{X} \times \{\pm 1\})^m$, a parameter $\nu \in [1, m]$, and a hypothesis set $\mathcal{H} \subset [-1, +1]^{\mathcal{X}}$.

### Edge minimization

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

where $\Delta_{m,\nu} = \{d \in [0, 1/\nu]^m \mid \|d\|_1 = 1\}$ and $A = (y_i h(x_i))$.
$(d^\top A)_h = \sum_i d_i y_i h(x_i)$ is the *edge* of $h \in \mathcal{H}$ w.r.t. the distribution $d$.

$\Updownarrow$ **Fenchel dual problem (zero duality gap)**

### Soft margin optimization

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

where $f^\star(\theta) = \sup_d \left[ \theta^\top d - f(d) \right]$ is Fenchel conjugate function of $f$.

**Output:** $\sum_{h \in \mathcal{H}} \bar{w}_h h$, where $\bar{w}$ is an optimal solution of (2).
Hard for off-the-shelf solvers when $\mathcal{H}$ is a huge set.

## Boosting

Boosting is a protocol between *Booster* and *Weak Learner (WL)*.

At each round $t = 0, 1, 2, \dots, T$,
1. Send a distribution $d_t \in \Delta_{m,\nu}$ over $S$ to WL.
2. WL returns a hypothesis $h_{t+1} \in \mathcal{H}$ such that $(d_t^\top A)_{h_{t+1}} \geq g$ to Booster ($g > 0$ is unknown).

Booster outputs a combined hypothesis $H_T = \sum_{t=1}^T w_{T,t} h_t$, where $w_T$ is an $\epsilon$-approximate solution of (2).

- LPBoost

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

- ERLPBoost

$$d_t^E \leftarrow \arg\min_d \max_{h \in \{h_1, h_2, \dots, h_t\}} (d^\top A)_h + \underbrace{f(d) + \frac{1}{\eta} \sum_{i=1}^m d_i \ln(m d_i)}_{=:\tilde{f}(d)}, \quad \eta = \frac{2}{\epsilon} \ln \frac{m}{\nu}$$

- C-ERLPBoost

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

where $w_t^C = w_{t-1}^C + \lambda_{t-1}(e_{h_t} - w_{t-1}^C) \in CH(\{e_{h_1}, e_{h_2}, \dots, e_{h_t}\})$

$$\lambda_{t-1} = \operatorname{clip}_{[0,1]} \frac{d_{t-1}^\top A (e_{h_t} - w_{t-1}^C)}{\eta \|A(e_{h_t} - w_{t-1}^C)\|_\infty^2}$$

| | LPBoost | ERLPBoost | C-ERLPBoost | **One of our work** |
|---|---|---|---|---|
| $T$ | $\Omega(m)$ | $O(\frac{1}{\epsilon^2} \ln \frac{m}{\nu})$ | $O(\frac{1}{\epsilon^2} \ln \frac{m}{\nu})$ | $O(\frac{1}{\epsilon^2} \ln \frac{m}{\nu})$ |
| Sub-problem | LP | CP | Sorting | **LP** |

## Goal.

Design a practical boosting algorithm with a theoretical guarantee.

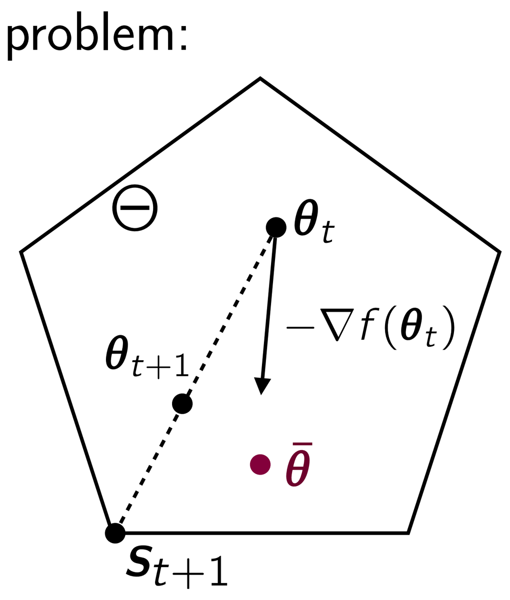
## The Frank-Wolfe algorithm

A first-order algorithm for solving the following class of problem:

$$\min_{\theta \in \Theta} f(\theta)$$

$\Theta \subset \mathbb{R}^m$ is a bounded & closed convex set, and $f : \Theta \to \mathbb{R}$ is an $\eta$-smooth convex function.

At each round $t = 0, 1, 2, \dots, T$,
1. Compute $s_{t+1} \leftarrow \arg\min_{s \in \Theta} s^\top \nabla f(\theta_t)$.
2. $\theta_{t+1} = \theta_t + \lambda_t (s_{t+1} - \theta_t)$, $\lambda_t \in [0, 1]$.




- Converges to an $\epsilon$-approximate solution in $O(\eta/\epsilon)$ rounds.
- FW algorithms assume an LP oracle over $\Theta$.
- Fast update per round.

**Fully-corrective FW:** $\theta_{t+1} \in \arg\min_{\theta \in CH(\{s_1, s_2, \dots, s_{t+1}\})} f(\theta)$
- Optimizes $f$ over the convex hull of $\{s_1, s_2, \dots, s_t\}$.
- Highly improves the objective value.
- Slow update per round.

## A unified view of the boosting algorithms

Consider the Fenchel dual problem of $\min_d \tilde{f}(d) + \max_{h \in \mathcal{H}} (d^\top A)_h$.

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

Here, we denote $-A\Delta_{\mathcal{H},1} = \{-Aw \mid w \in \Delta_{\mathcal{H},1}\}$.
- $\tilde{f}^\star$ is $\eta$-smooth w.r.t. $L_\infty$-norm ($\tilde{f}$ is $1/\eta$-strongly convex w.r.t. $L_1$-norm).
- Distribution $d$ over $S$ is a/the (sub-)gradient of $f^\star$ / $\tilde{f}^\star$ at some point $\theta$.
  - **C-ERLPBoost** $d_t^C = \nabla \tilde{f}^\star(-Aw_t^C)$.
  - **ERLPBoost** $d_t^E = \nabla \tilde{f}^\star(-Aw_t^E)$, where $w_t^E = \arg\min_{w \in CH(\{e_{h_1}, e_{h_2}, \dots, e_{h_t}\})} \tilde{f}^\star(-Aw)$.
  - **LPBoost** $d_t^L \in \partial f^\star(-Aw_t^L)$, where $w_t^L \in \arg\min_{w \in CH(\{e_{h_1}, e_{h_2}, \dots, e_{h_t}\})} f^\star(-Aw)$.
- A max-edge WL corresponds to the LP oracle in FW:

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

**Theorem.**
LPBoost, ERLPBoost, and C-ERLPBoost are instances of the FW algorithm.

## A new boosting scheme

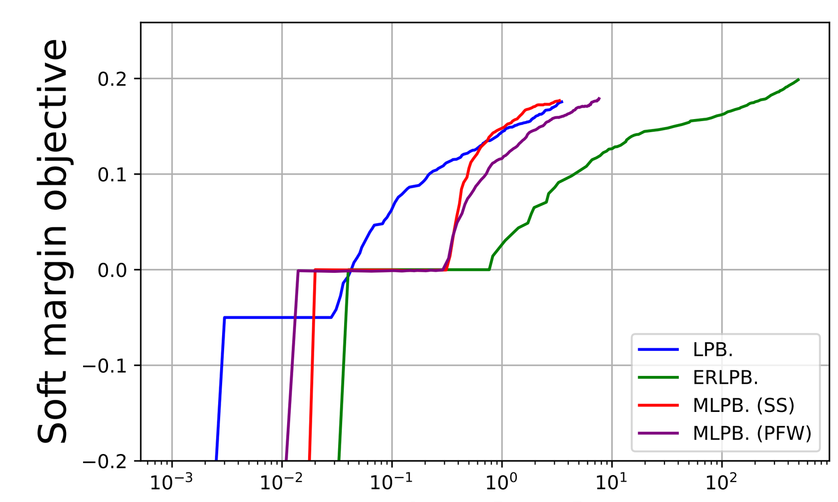
At each round $t = 0, 1, 2, \dots, T$,
1. Compute $d_t = \nabla \tilde{f}^\star(-Aw_t) = \arg\min_{d \in \Delta_{m,\nu}} d^\top A w_t + \frac{1}{\eta} \sum_{i=1}^m d_i \ln(m d_i)$.
2. Obtain a hypothesis $h_{t+1} \in \mathcal{H}$.
3. $\min_{q \leq t} (d_q^\top A)_{h_{q+1}} + \tilde{f}^\star(-Aw_t) \leq \epsilon/2 \implies$ **break**.
4. Primary (Frank-Wolfe) update: $w_{t+1}^F \in \Delta_{\mathcal{H},1}$.
5. Secondary update: $w_{t+1}^B \in \Delta_{\mathcal{H},1}$.
6. $w_{t+1} \in \arg\min_{w \in \{w_{t+1}^F, w_{t+1}^B\}} \tilde{f}^\star(-Aw)$

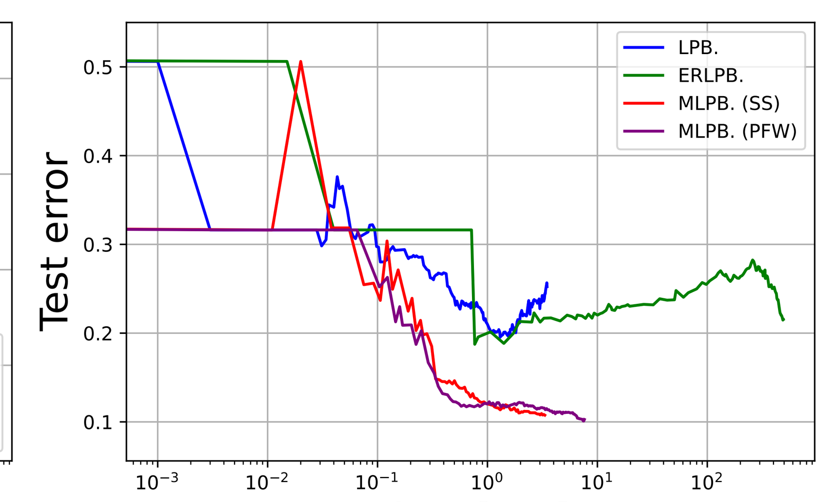
**Theorem.**
With an appropriate primary update $w_{t+1}^F$, the proposed scheme outputs a $w_T$ such that $-f^\star(-Aw_T) \geq g - \epsilon$ after at most $T = O(\frac{1}{\epsilon^2} \ln \frac{m}{\nu})$ rounds.

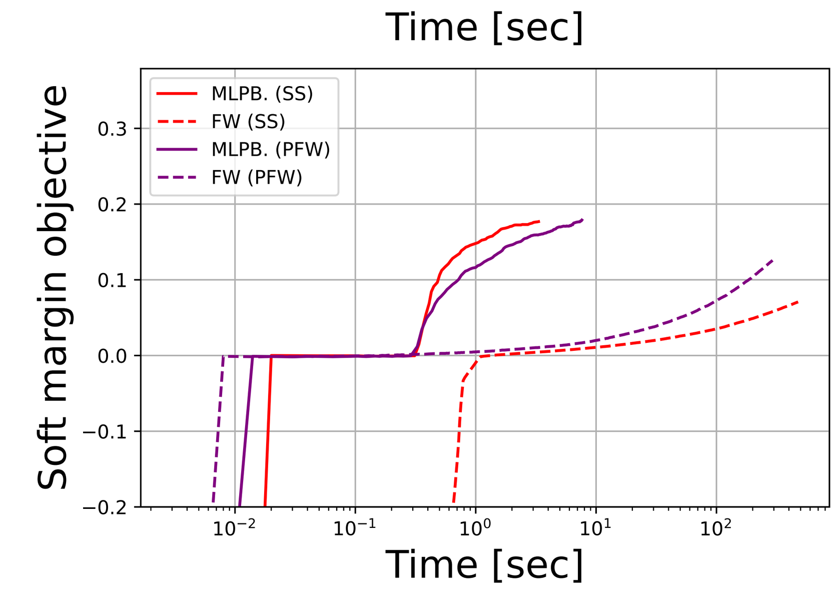
- One can compute the distribution $d_t = \nabla \tilde{f}^\star(-Aw_t)$ in $O(m \ln m)$ time.
- One can choose $w_{t+1}^F$ that guarantee some improvement per round.
- Choosing $w_{t+1}^B = \arg\min_{w \in CH(\{e_{j_1}, e_{j_2}, \dots, e_{j_{t+1}}\})} \tilde{f}^\star(-Aw)$ yields ERLPBoost.

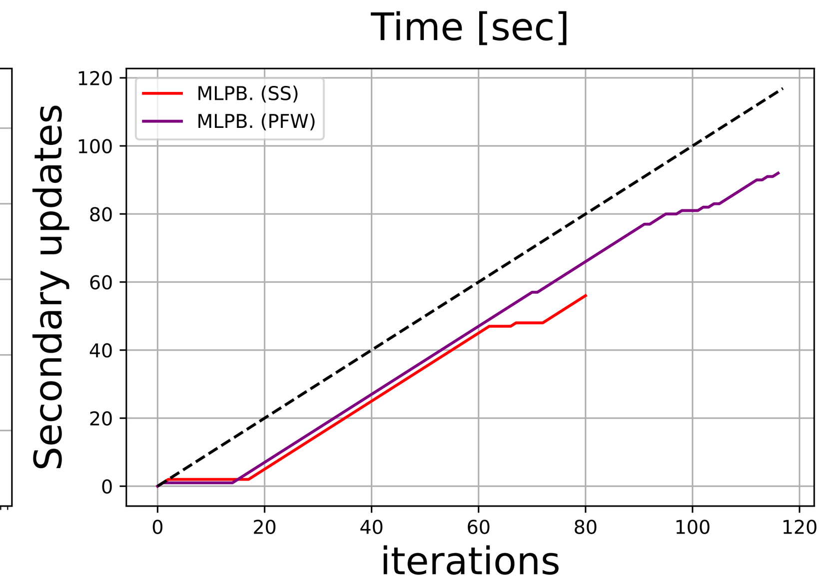
## Experiments

- MLPBoost: Uses $w_t^B \in \arg\min_{w \in CH(\{e_{h_1}, e_{h_2}, \dots, e_{h_t}\})} f^\star(-Aw)$.
- Datasets: Gunnar Rätsch's benchmark datasets [a].
- Parameters: $\epsilon = 0.01$, $\nu = 0.1m$.
- Max-edge Weak Learner for decision tree with depth 2.




Average running times (seconds) for 5-fold CV.

| | $m$ | LPB. | ERLPB. | C-ERLPB. | MLPB. (SS) | MLPB. (PFW) |
|---|---|---|---|---|---|---|
| R.norm | 7,400 | 22.09 | 1,148.16 | $> 10^4$ | 26.76 | 36.73 |
| Twonorm | 7,400 | 105.40 | $> 10^4$ | $> 10^4$ | 478.22 | 397.91 |
| Waveform | 5,000 | 437.29 | 9,018.54 | $> 10^4$ | 2,243.07 | 1,619.56 |

## Summary & Future work

- We provide a unified view of the boosting algorithms.
- We propose a new scheme for boosting based on the FW algorithm.
  - By adopting LPBoost as the secondary update, our scheme yields a practical boosting algorithm with a theoretical guarantee.
- LPBoost is still faster in practice. $\implies$ Any other secondary update?

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