Scale-Free Algorithms for Online Linear Optimization

Francesco Orabona **Dávid Pál**

Yahoo Labs NYC

October 4, 2015
ALT 2015

Online Linear Optimization

For $t = 1, 2, \ldots$

- predict $w_t \in K \subseteq \mathbb{R}^d$
- receive loss vector $\ell_t \in \mathbb{R}^d$
- suffer loss $\langle \ell_t, w_t \rangle$

Competitive analysis w.r.t. static strategy $u \in K$:

$$\text{Regret}_T(u) = \underbrace{\sum_{t=1}^{T} \langle \ell_t, w_t \rangle}_{\text{algorithm's loss}} - \underbrace{\sum_{t=1}^{T} \langle \ell_t, u \rangle}_{\text{comparator's loss}}$$

Goal: Design algorithms with sublinear Regret_T.

Applications

- Offline and stochastic convex optimization
  - Logistic regression ($K = \mathbb{R}^d$)
- Online combinatorial problems
  - learning with expert advice (K = probability simplex)
  - shortest path (K = flow polytope)
  - bipartite matching (K = doubly stochastic matrices)
  - spanning tree (K = spanning tree polytope)
  - k-subset, etc.

Standard Regret Bound

Theorem (Abernethy et al. ’08; Rakhlin ’09)

For any bounded convex $K \subseteq \mathbb{R}^d$ and any norm $\|\cdot\|$, there exists an algorithm that ***receives*** *T* ***and*** *$\sum_{t=1}^{T} \|\ell_t\|_*^2$* ***before the first round*** *and satisfies*

$$\forall u \in K \qquad \mathrm{Regret}_T(u) \le C_{K,\|\cdot\|} \sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2} \, .$$

(Mirror Descent, Follow The Regularized Leader)

Corollary

If $\|\ell_t\|_* \le B$ then $\mathrm{Regret}_T(u) \le C_{K,\|\cdot\|} B \sqrt{T}$.

Adaptive Regret Bound

Theorem (Orabona & P.)

*For any bounded convex $K \subseteq \mathbb{R}^d$ and any norm $\| \cdot \|$, there exists an algorithm that ~~**receives** T **and** $\sum_{t=1}^T \|\ell_t\|_*^2$ **before the first round and**~~ satisfies*

$$\forall T \quad \forall u \in K \qquad \text{Regret}_T(u) \leq C'_{K,\|\cdot\|} \sqrt{\sum_{t=1}^T \|\ell_t\|_*^2} \,.$$

- The value of $C'_{K,\|\cdot\|}$ later in the talk.
- Similar result for unbounded K.

Adaptivity

Adaptivity to unknown T is easy:

- Doubling trick. Try $T = 1, 2, 4, 8, 16, 32, \ldots$

Adaptivity to unknown $\|\ell_t\|_*$:

- AdaHedge for K = probability simplex

  [de Rooij, van Erven, Grünwald, Koolen '14]
- AdaGrad, FTRL Proximal for $\|\cdot\|_2$ and $\|\ell_t\|_2 \geq 1$

  [Duchi, Hazan, Singer '11; McMahan & Streeter '10]
- AdaFTRL for any bounded K, any norm

  [this paper]
- SOLO FTRL for any K (bounded or unbounded), any norm

  [this paper]

Strong Convexity

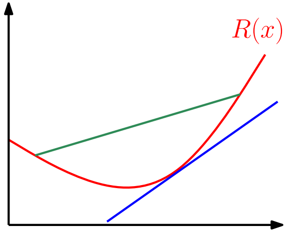
A convex function $R : K \to \mathbb{R}$ is λ**-strongly convex** w.r.t. $\|\cdot\|$ iff

$$\forall x, y \in K \quad \forall t \in [0,1]$$

$$R(tx + (1-t)y) \leq tR(x) + (1-t)R(y) - \frac{\lambda}{2} t(1-t)\|x-y\|^2$$

If R is differentiable, this is equivalent to

$$\forall x, y \in K \qquad R(y) \geq R(x) + \langle \nabla R(x), y - x \rangle + \frac{\lambda}{2}\|x-y\|^2$$



Follow The Regularized Leader (FTRL)

- $R : K \to \mathbb{R}$ non-negative 1-strongly convex w.r.t. $\|\cdot\|$.
- FTRL chooses

$$w_t = \operatorname*{argmin}_{w \in K} \left(\frac{1}{\eta_t} R(w) + \sum_{i=1}^{t-1} \langle \ell_i, w \rangle \right)$$

where $\eta_t > 0$ is a learning rate.

- Constant learning rate $\eta_1 = \eta_2 = \cdots = \eta_T = \sqrt{\frac{\sup_{v \in K} R(v)}{\sum_{t=1}^{T} \|\ell_t\|_*^2}}$ gives [Rakhlin '09; Shalev-Shwartz '11]

$$\text{Regret}_T(u) \le 2 \underbrace{\sqrt{\sup_{v \in K} R(v)}}_{C_{K,\|\cdot\|}} \sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2}$$

- How to choose η_t adaptively?

Scale-Free Property

Multiply loss vectors by $c > 0$:

$$\ell_1, \ell_2, \ell_3, \dots \to c\ell_1, c\ell_2, c\ell_3, \dots$$

An algorithm is **scale-free** if $w_1, w_2, w_3, \dots$ remains the same.

For a scale-free algorithm

$$\mathrm{Regret}_T(u) \to c\,\mathrm{Regret}_T(u) \qquad \sum_{t=1}^T \langle \ell_t, w_t \rangle \to c \sum_{t=1}^T \langle \ell_t, w_t \rangle$$

$$\sqrt{\sum_{t=1}^T \|\ell_t\|_*^2} \to c \sqrt{\sum_{t=1}^T \|\ell_t\|_*^2}$$

Scale-Free FTRL

For FTRL

$$w_t = \operatorname*{argmin}_{w \in K} \left(\frac{1}{\eta_t} R(w) + \sum_{i=1}^{t-1} \langle \ell_i, w \rangle \right)$$

to be scale-free $1/\eta_t$ needs to be **positive** 1**-homogeneous** function of $\ell_1, \ell_2, \ldots, \ell_{t-1}$.

That is, $(\ell_1, \ell_2, \ldots, \ell_{t-1}) \to (c\ell_1, c\ell_2, \ldots, c\ell_{t-1})$ causes

$$1/\eta_t \to c/\eta_t$$

$$w_t = \operatorname*{argmin}_{w \in K} \left(\frac{1}{\eta_t} R(w) + \sum_{i=1}^{t-1} \langle \ell_i, w \rangle \right)$$

$$\downarrow$$

$$w_t = \operatorname*{argmin}_{w \in K} \left(\frac{c}{\eta_t} R(w) + \sum_{i=1}^{t-1} \langle c\ell_i, w \rangle \right)$$

Two Good Scale-Free Choices of η_t

SOLO FTRL:

$$\frac{1}{\eta_t} = \sqrt{\sum_{i=1}^{t-1} \|\ell_i\|_*^2}$$

AdaFTRL:

$$\frac{1}{\eta_t} = \begin{cases} 0 & \text{if } t = 1 \\ \frac{1}{\eta_{t-1}} + \frac{1}{\eta_{t-1}} D_{R^*}\left(-\eta_{t-1} \sum_{i=1}^{t-1} \ell_i, -\eta_{t-1} \sum_{i=1}^{t-2} \ell_i\right) & \text{if } t \geq 2 \end{cases}$$

$D_{R^*}(\cdot, \cdot)$ is the Bregman divergence of Fenchel conjugate of R.

Regret of Scale-Free FTRL

Theorem

Let $R : K \to \mathbb{R}$ be non-negative and λ-strongly convex w.r.t. $\|\cdot\|$.
Suppose K has diameter D w.r.t. to $\|\cdot\|$.

SOLO FTRL *satisfies*

$$\text{Regret}_T(u) \le \left(R(u) + \frac{2.75}{\lambda}\right) \sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2} + 3.5 \min\left\{D, \frac{\sqrt{T-1}}{\lambda}\right\} \max_{1 \le t \le T} \|\ell_t\|_* .$$

AdaFTRL *satisfies*

$$\text{Regret}_T(u) \le 2 \max\left\{D, \frac{1}{\sqrt{\lambda}}\right\} (1 + R(u)) \sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2} .$$

Optimization of λ for Bounded K

- Choose $R(w) = \lambda \cdot f(w)$ where f is non-negative 1-strongly convex.
- Use $D \leq \sqrt{8 \sup_{v \in K} f(v)}$
- Optimize λ. Optimal choice depends only on $\sup_{v \in K} f(v)$.

With optimal choices of λ,

AdaFTRL: $$\text{Regret}_T(u) \leq 5.3 \sqrt{\sup_{v \in K} f(v) \sum_{t=1}^{T} \|\ell_t\|_*^2}$$

SOLO FTRL: $$\text{Regret}_T(u) \leq 13.3 \sqrt{\sup_{v \in K} f(v) \sum_{t=1}^{T} \|\ell_t\|_*^2}$$

Our Proof Techniques

Lemma

For non-negative numbers $C, a_1, a_2, \ldots, a_T$,

$$\sum_{t=1}^{T} \min\left\{ \frac{a_t^2}{\sqrt{\sum_{s=1}^{t-1} a_s^2}},\ C a_t \right\} \leq 3.5 \sqrt{\sum_{t=1}^{T} a_t^2} \ + \ 3.5 C \max_{1 \leq t \leq T} a_t$$

Lemma

For non-negative numbers $a_1, a_2, \ldots, a_T$ the recurrence

$$0 \leq b_t \leq \min\left\{ a_t, \frac{a_t^2}{\sum_{s=1}^{t-1} b_s} \right\} \quad \textit{implies that} \quad \sum_{t=1}^{T} b_t \leq 2 \sqrt{\sum_{t=1}^{T} a_t^2}$$

Lower Bound for Bounded K

Theorem

For any $a_1, a_2, \ldots, a_T$ and any algorithm there exists $\ell_1, \ell_2, \ldots, \ell_T$ and $u \in K$ such that

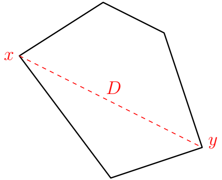
- $\|\ell_1\|_* = a_1$, $\|\ell_2\|_* = a_2$, $\ldots$, $\|\ell_T\|_* = a_T$
- $\text{Regret}_T(u) \geq \frac{D}{\sqrt{8}} \sqrt{\sum_{t=1}^T \|\ell_t\|_*^2}$

Proof.

- Choose $\ell \in \mathbb{R}^d$ and $x, y \in K$ such that

$$\|x - y\| = D \qquad \|\ell\|_* = 1$$

$$\operatorname*{argmin}_{w \in K} \langle \ell, w \rangle = x \qquad \operatorname*{argmax}_{w \in K} \langle \ell, w \rangle = y$$



- Set $\ell_t = \pm a_t \ell$ where signs are i.i.d. random

$\square$

Open Problem: Bounded K

- Lower vs. upper bound

$$\frac{D}{\sqrt{8}} \sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2} \qquad \text{vs.} \qquad 5.3 \sqrt{\sup_{u \in K} f(u) \sum_{t=1}^{T} \|\ell_t\|_*^2}$$

  where $f : K \to \mathbb{R}$ is 1-strongly convex w.r.t. $\|\cdot\|$.

- Upper bound is (almost) tight. [Srebro, Sridharan, Tewari '11]

- Open problem: [Kwon & Mertikopoulos '14]

  Given a convex set K and a norm $\|\cdot\|$, construct non-negative 1-strongly convex $f : K \to \mathbb{R}$ that minimizes

$$\sup_{u \in K} f(u) \ .$$

Open Problems: Unbounded K

- For λ-strongly convex R, SOLO FTRL:

$$\text{Regret}_T(u) \leq R(u) \sqrt{\sum_{t=1}^{T} \|\ell_t\|_*^2} + 6.25 \frac{\sqrt{T}}{\lambda} \max_{1 \leq t \leq T} \|\ell_t\|_*$$

- For 2-norm, $K = \mathbb{R}^d$, assuming $\|\ell_t\|_2 \leq 1$,
  PiSTOL: [Orabona '13, '14; McMahan & Orabona '13]

$$\text{Regret}(u) \leq O\left(\|u\|_2 \sqrt{T \log(T\|u\|_2)} \right) .$$

- Open problem 1:
  Algorithm for $K = \mathbb{R}^d$ that adapts to $\|\ell_t\|_2$ and has regret

$$\|u\|_2 \sqrt{T} \max_{1 \leq t \leq T} \|\ell_t\|_2 \cdot \text{poly}(\log T, \log \|u\|_2)$$

- Open problem 2:
  What about other norms and unbounded $K \neq \mathbb{R}^d$?

Questions?