Scale-Free Algorithms for Online Linear Optimization

Francesco Orabona **Dávid Pál**

Yahoo Labs NYC

October 9, 2015

AI Seminar @ University of Alberta

Overview

1. Online Linear Optimization
2. Applications
3. Non-adaptive algorithms
4. Adaptive (i.e. **scale-free**) algorithms
5. Lower bounds, Recent developments, Open problems

Overview

1. Online Linear Optimization
2. Applications
3. Non-adaptive algorithms
4. Adaptive (i.e. **scale-free**) algorithms
5. Lower bounds, Recent developments, Open problems

Remember

$$\text{GD step size} = \frac{1}{\sqrt{\sum_{\text{past iterations}} \|\text{gradient}_t\|^2}}$$

Online Linear Optimization

For $t = 1, 2, \ldots$

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

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

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

Goal: Design algorithms with *sublinear* Regret_T.

Applications

1. Batch **convex** optimization
2. Stochastic optimization i.e. minimization of **test** error
3. Genuinely online/control problems

Regret bound implies results in all of these areas.

(Take Csaba's Online learning course!)

Application 1: Batch convex optimization

- We want to solve
$$\underset{w \in K}{\text{minimize}} f(w)$$
- Suppose $f : K \to \mathbb{R}$ is convex
- $w^* = \text{argmin}_{w \in K} f(w)$
- Feed online algorithm with $g_t = \nabla f(w_t)$
- $\widehat{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$ is approximately optimal:
$$f(\widehat{w}) \leq f(w^*) + \frac{\text{Regret}_T(w^*)}{T}$$

Application 2: Stochastic optimization

- We want to solve

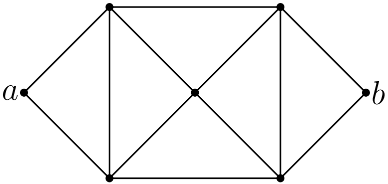
$$\underset{w \in K}{\text{minimize}} \ \mathrm{Risk}(w) \qquad \text{where} \qquad \mathrm{Risk}(w) = \mathop{\mathbf{E}}_{z \sim D}\left[\ell(w, z)\right]$$

- D is unknown; we have i.i.d. sample $z_1, z_2, \ldots, z_T$ from D
- $\ell(w, z)$ is convex in w
- $w^* = \mathrm{argmin}_{w \in K} \mathrm{Risk}(w)$
- Feed online algorithm with $g_t = \nabla \ell(w_t, z_t)$
- $\widehat{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$ is approximately optimal:

$$\mathbf{E}\left[\mathrm{Risk}(\widehat{w})\right] \leq \mathrm{Risk}(w^*) + \frac{\mathbf{E}\left[\mathrm{Regret}_T(w^*)\right]}{T}$$

Application 3: Online Shortest Path

- Given graph $G = (V, E)$ and source-sink pair a, b



- Algorithm chooses path p_t from a to b
- Receives loss of each edge: $\ell_t : E \to \mathbb{R}$
- Regret w.r.t. a path q

$$\text{Regret}_T(q) = \sum_{t=1}^{T} \ell_t(p_t) - \sum_{t=1}^{T} \ell_t(q)$$

- Vector $w \in K \subseteq \mathbb{R}^{|E|}$ is a unit flow from a to b

Typical Yahoo/Google applications

Stochastic optimization problem

$$\underset{w \in \mathbb{R}^d}{\text{minimize}} \underset{z \sim D}{\mathbf{E}} [\ell(w, z)]$$

where w is a vector of parameters and z is a data record

- i.i.d. sample $z_1, z_2, \ldots, z_T$ from D
- A data record z_t could be:
  - "`Hi, My name is Nastasjushka :)`" is a spam email.
  - Coca-Cola ad on `www.cbc.ca` was not clicked on by Csaba at 3:14:15pm
- Data is huge
  - T is between 10^6 and 10^{10}
  - w has dimension between 10^5 and 10^8

Follow The Regularized Leader (FTRL)

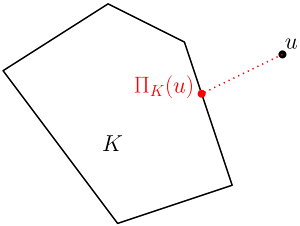
Let be $R : K \to \mathbb{R}$ be a convex and $\eta_t > 0$. FTRL chooses

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

For example with $R(w) = \frac{1}{2}\|w\|_2^2$

$$w_t = \Pi_K \left(-\eta_t \sum_{i=1}^{t-1} g_i \right)$$

where $\Pi_K(u)$ is the projection of u to K.



FTRL $\approx$ Gradient Descent

Suppose $K = \mathbb{R}^d$ and $R(w) = \frac{1}{2}\|w\|_2^2$.

FTRL:

$$w_t = -\eta_t \sum_{i=1}^{t-1} g_i$$

Gradient Descent:

$$w_t = -\sum_{i=1}^{t-1} \eta_i g_i$$

Strong Convexity

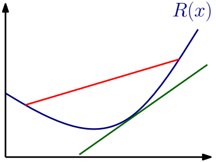
A convex function $R : K \to \mathbb{R}$ is λ**-strongly convex** w.r.t. $\|\cdot\|$ iff

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

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

If R is differentiable, this is equivalent to

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



For example,

- $R(w) = \frac{1}{2}\|w\|_2^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$
- $R(w) = \sum_{i=1}^{d} w_i \ln w_i$ is 1-strongly convex w.r.t. $\|\cdot\|_1$ on

$$K = \left\{ w \in \mathbb{R}^d \ : \ w \geq 0, \sum_{i=1}^{d} w_i = 1 \right\}$$

Regret of FTRL for Bounded K

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

Let $K \subseteq \mathbb{R}^d$ *be convex bounded.*
Let $R : K \to \mathbb{R}$ *be non-negative,* 1*-strongly convex w.r.t.* $\|\cdot\|$.
FTRL *with* $\eta_1 = \eta_2 \cdots = \eta_T = \sqrt{\frac{\sup_{v \in K} R(v)}{\sum_{t=1}^T \|g_t\|_*^2}}$ *satisfies*

$$\mathrm{Regret}_T(u) \leq 2 \sqrt{\sup_{v \in K} R(v) \sum_{t=1}^T \|g_t\|_*^2} \,.$$

Corollary

If $\|g_t\|_* \leq B$ *then* $\mathrm{Regret}_T(u) \leq 2B\sqrt{T \sup_{v \in K} R(v)}$.

Algorithm needs to know T, B, $\sum_{t=1}^T \|g_t\|_*^2$ in advance.

Adaptive algorithm?

Is there an algorithm such that

$$\text{Regret}_T(u) \le 100 \sqrt{\sup_{v \in K} R(v) \sum_{t=1}^{T} \|g_t\|_*^2}$$

for any T and any sequence $g_1, g_2, \ldots, g_T$ without knowing T, B, or $\sum_{t=1}^{T} \|g_t\|_*^2$ in advance?

Scale-Free Property

Multiply loss vectors by $c > 0$:

$$g_1, g_2, \dots \to c g_1, c g_2, \dots$$

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

For a scale-free algorithm

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

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

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, $(g_1, g_2, \ldots, g_{t-1}) \to (cg_1, cg_2, \ldots, cg_{t-1})$ causes

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

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

$$\downarrow$$

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

Two Good Scale-Free Choices of η_t

SOLO FTRL:

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

ADAFTRL:

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

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

$$\begin{aligned} D_{R^*}(u,v) &= R^*(u) - R^*(v) - \langle u - v, \nabla R^*(v) \rangle \\ R^*(u) &= \sup_{v \in K} \langle u, v \rangle - R(v) \end{aligned}$$

Regret of Scale-Free FTRL

Theorem (Orabona & P. ’15)

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*:*

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

AdaFTRL*:*

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

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 λ depends only on $\sup_{v \in K} f(v)$.

With optimal choices of λ,

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

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

Proof Techniques

Lemma

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

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

Lemma

Let $a_1, a_2, \ldots, a_T$ be non-negative. 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 (Orabona & P. '15)

For any $a_1, a_2, \ldots, a_T \geq 0$ and any algorithm there exists $g_1, g_2, \ldots, g_T \in \mathbb{R}^d$ and $u \in K$ such that

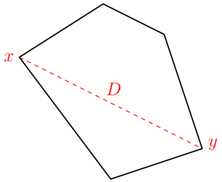
1. $\|g_1\|_* = a_1,\ \|g_2\|_* = a_2,\ \ldots,\ \|g_T\|_* = a_T$
2. $\text{Regret}_T(u) \geq \frac{D}{\sqrt{8}} \sqrt{\sum_{t=1}^{T} \|g_t\|_*^2}$

Proof sketch.

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

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

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



- Set $g_t = \pm a_t g$ where signs are i.i.d. random

$\square$

Open Problem: Bounded K

- Lower vs. upper bound

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

  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

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

Suboptimality of SOLO for Unbounded K

- SOLO for λ-strongly convex R,

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

- SOLO for $R(u) = \|u\|_2^2$, which is 2-strongly convex

$$\text{Regret}_T(u) \leq \|u\|_2^2 \sqrt{\sum_{t=1}^{T} \|g_t\|_*^2 + 3.125 \sqrt{T} \max_{t=1,2,\ldots,T} \|g_t\|_*}$$

- Take $\|u\|_2 \leq D$. SOLO with $K = \{u \ : \ \|u\|_2 \leq D\}$:

$$\text{Regret}_T(u) \leq 13.3 D \sqrt{\sum_{t=1}^{T} \|g_t\|_*^2}$$

What is the right bound for $K = \mathbb{R}^d$?

$$\mathrm{Regret}_T(u) \leq O\left(\|u\|_2 \sqrt{\sum_{t=1}^{T} \|g_t\|_*^2}\right)$$

vs.

$$\mathrm{Regret}_T(u) \leq O\left(\|u\|_2^2 \sqrt{\sum_{t=1}^{T} \|g_t\|_*^2}\right)$$

Upper Bound for $K = \mathbb{R}^d$ (Unpublished)

Theorem

If $\|g_t\|_2 \leq 1$*, the algorithm*

$$w_t = -\frac{\sum_{i=1}^{t-1} g_i}{2(t+1)} \left(\sqrt{t} - \sum_{i=1}^{t-1} \langle g_i, w_i \rangle \right)$$

has regret

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

Similar results [McMahan & Streeter '12; Orabona '13, '14; McMahan & Abernethy '13]

Lower Bound for $K = \mathbb{R}^1$ (Unpublished)

Theorem

*For any algorithm there exists a sequence $g_1, g_2, \ldots, g_T \in \mathbb{R}^1$ such that $|g_1| = |g_2| = \cdots = |g_T| = 1$ and **one** of the following holds:*

1. *For $u = \log T$, $\mathrm{Regret}_T(u) \geq \Omega\left(|u| \sqrt{T \log |u|}\right)$.*
2. $\mathrm{Regret}_T(0) \geq \Omega\left(\sqrt{T \log \log T}\right)$.

This rules out $O(|u| \sqrt{T})$ upper bound.

Open Problems: Unbounded K

- Is there an adaptive algorithm for $K = \mathbb{R}^d$ and 2-norm such that

$$\|u\|_2 \sqrt{T} \max_{t=1,2,\ldots,T} \|g_t\|_2 \cdot \mathrm{poly}(\log T, \log \|u\|_2)$$

  for any sequence $g_1, g_2, \ldots, g_T$?
- What about norms other than 2-norm?
- What about unbounded $K \neq \mathbb{R}^d$?

Questions?

Scale-Free Algorithms for Online Optimization, ALT 2015

http://arxiv.org/abs/1502.05744