Scale-Free Algorithms for Online Linear Optimization

Francesco Orabona **Dávid Pál**

Yahoo Labs NYC

March 5, 2015

Large Scale Machine Learning Problems

Convex optimization problem

$$\underset{w}{\text{minimize}} \sum_{t=1}^{T} \ell(w, z_t)$$

where w is a vector of parameters and z_t is a data record.

A data record z_t could be:

- "`Hi, My name is Nastasjushka :)`" is a spam email.
- Coca-Cola ad on `www.nytimes.com` was not clicked on by David at 3:14:15pm

Loss function $\ell(w, z_t)$ is **convex** in w.

Methods of Solution

Data is huge

- T is between 10^6 and 10^{10}
- w has dimension between 10^6 and 10^9

First-order methods

$$w_{t+1} = w_t - \eta \nabla_w \ell(w_t, z_t)$$

- How to tune step size η?
- What is the **test** loss of the learned model?

Overview

Online Learning 101:

1. Online Convex Optimization (OCO)
2. Solving OCO implies low test error
3. Online Linear Optimization (OLO)
4. OLO solves OCO

Scale-Free algorithms for OLO:

1. Follow The Regularized Leader (FTRL)
2. Strongly convex regularizers
3. Scale-free variants of FTRL
4. Upper/Lower Bounds on Regret
5. Open Problem

OL 101: Online Convex Optimization (OCO)

For $t = 1, 2, \ldots$

- predict $w_t \in K$
- receive convex loss function $\ell_t : K \to \mathbb{R}$
- suffer loss $\ell_t(w_t)$

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

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

Goal: Design algorithms with sublinear Regret_T.

OL 101: Solving OCO implies low test error

We really want to solve a stochastic optimization problem

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

and D is unknown. We have only i.i.d. sample $z_1, z_2, \ldots, z_T$.

- Run an OCO algorithm on $\ell_t(\cdot) = \ell(\cdot, z_t)$.
- Take $\overline{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$
- It can be proved that

$$\mathbf{E}[\text{Risk}(\overline{w})] - \text{Risk}(w^*) \leq \frac{1}{T}\,\mathbf{E}[\text{Regret}_T(w^*)]$$

- High probability result:

$$\text{Risk}(\overline{w}) - \text{Risk}(w^*) \leq \frac{1}{T}\,\text{Regret}_T(w^*) + O\left(\sqrt{\log(1/\delta)/T}\right)$$

No regularization needed!

OL 101: Online Linear Optimization (OLO)

For $t = 1, 2, \ldots$

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

How well an algorithm is doing compared to u:

$$\text{Regret}_T(u) = \sum_{t=1}^{T} \langle g_t, w_t \rangle - \sum_{t=1}^{T} \langle g_t, u \rangle$$

Goal: Design algorithms with sublinear Regret_T.

OL 101: OLO solves OCO

- Feed OLO algorithm with $g_t = \nabla \ell_t(w_t)$
- It can be proved that

$$\text{Regret}^{(OCO)}(u) \le \text{Regret}^{(OLO)}(u)$$

Proof:

$$\begin{aligned}
\text{Regret}^{(OCO)}(u) &= \sum_{t=1}^{T} \ell_t(w_t) - \ell_t(u) \\
&\le \sum_{t=1}^{T} \langle \nabla \ell_t(w_t), w_t - u \rangle \\
&= \sum_{t=1}^{T} \langle g_t, w_t \rangle = \text{Regret}^{(OLO)}(u)
\end{aligned}$$

Linear functions are the hardest convex functions to minimize!

Overview

Online Learning 101:

1. Online Convex Optimization (OCO) ✓
2. Solving OCO implies low test error ✓
3. Online Linear Optimization (OLO) ✓
4. OLO solves OCO ✓

Scale-Free algorithms for OLO:

1. Follow The Regularized Leader (FTRL)
2. Strongly convex regularizers
3. Scale-free variants of FTRL
4. Upper/Lower Bounds on Regret
5. Open Problem

Follow The Regularized Leader (FTRL)

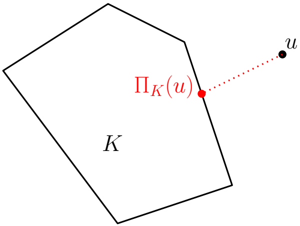
Let be $R : K \to \mathbb{R}$ be a convex and $\eta > 0$. FTRL chooses

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

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

$$w_t = \Pi_K \left(-\eta \sum_{s=1}^{t-1} g_s \right)$$

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



FTRL = Gradient Descent with Lazy Projections

$$w_t = \Pi_K\left(-\eta \sum_{s=1}^{t-1} g_s\right)$$

$$x_t = x_{t-1} - \eta g_{t-1}$$
$$w_t = \Pi_K(x_{t-1})$$

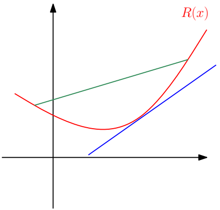
Strongly Convex Regularizers

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

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

Equivalently, for all $t \in [0,1]$ and all $x, y \in K$,

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



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 Bound for FTRL

Theorem

If $R(w) \geq 0$ and 1-strongly convex with respect to $\|\cdot\|$,

$$\text{Regret}_T(u) \leq \frac{1}{\eta} R(u) + \eta \sum_{t=1}^{T} \|g_t\|_*^2$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Optimal choice of η when K is bounded

$$\eta = \sqrt{\frac{\sup_{u \in K} R(u)}{\sum_{t=1}^{T} \|g_t\|_*^2}} \qquad \text{Regret}_T(u) \leq 2\sqrt{\sup_{u \in K} R(u) \sum_{t=1}^{T} \|g_t\|_*^2}$$

How do you choose η 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 OLO algorithm is **scale-free** if $w_1, w_2, \dots$ remains the same.

For a scale-free algorithm

$$\mathrm{Regret}_T(u) \to c\,\mathrm{Regret}_T(u)$$

and

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

Scale-Free FTRL

For FTRL

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

to be scale-free $1/\eta_t$ needs to be 1-homogeneous function of $g_1, g_2, \ldots, g_{t-1}$.

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

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

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

Bad Scale-Free Choices for η_t

For example,

$$\eta_t = \frac{1}{\sum_{s=1}^{t-1} \|g_s\|_*}$$
$$\eta_t = \frac{1}{\|g_{t-1}\|_* + 42\|g_{t-2}\|_*}$$
$$\eta_t = \frac{1}{\sqrt[t-1]{\prod_{s=1}^{t-1} \|g_s\|_*}}$$
$$\eta_t = \frac{1}{\langle g_{t-1}, w_{t-1}\rangle + 47\langle g_{t-2}, w_{t-2}\rangle}$$
$$\vdots$$

makes $1/\eta_t$ 1-homogeneous in $g_1, g_2, \ldots, g_{t-1}$.

Unfortunately, regret will be $\Omega(T)$ for all of these.

Two Good Scale-Free Choices of η_t

$$\eta_t = \frac{1}{\sqrt{\sum_{s=1}^{t-1} \|g_s\|_*^2}} \qquad \text{(SOLO FTRL)}$$

$$\eta_t = \frac{1}{\sum_{s=1}^{t-1} \frac{1}{\eta_s} D_{R^*}\left(-\eta_s \sum_{j=1}^{s} g_j, -\eta_s \sum_{j=1}^{s-1} g_j\right)} \qquad \text{(AdaFTRL)}$$

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

Regret of Scale-Free FTRL

Theorem

Suppose $R : K \to \mathbb{R}$ is non-negative and λ-strongly convex w.r.t. $\|\cdot\|$. K had diameter D w.r.t. to $\|\cdot\|$.
SOLO FTRL:

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

ADAFTRL:

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

Optimization of λ for Bounded K

- Choose $R(w) = \lambda f(w)$ where f is non-negative 1-strongly convex.
- Use $D \leq \sqrt{8 \sup_{u \in K} f(u)}$
- Optimize λ

For both algorithms, with optimal choices of λ,

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

Bits of the Proof: Homogeneous Inequalities

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

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

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

$$0 \le b_t \le \min\left\{ a_t, \frac{a_t^2}{\sum_{s=1}^{t-1} b_s} \right\}$$

implies that

$$\sum_{t=1}^{T} b_t \le 2 \sqrt{\sum_{t=1}^{T} a_t^2}$$

OLO Lower Bound

Theorem

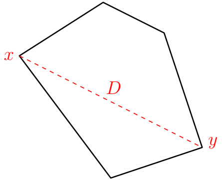
For any $a_1, a_2, \ldots, a_T$ and any OLO algorithm there exists $\ell_1, \ell_2, \ldots, \ell_T$ and $u \in K$ such that

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

Proof.

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

$$\|x - y\| = D \qquad \|\ell\|_* = 1$$
$$\operatorname*{argmin}_{x \in K} \langle \ell, x \rangle = x \qquad \operatorname*{argmax}_{x \in K} \langle \ell, y \rangle = y$$



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

□

Open Problem

Our regret bound is

$$\sqrt{\sup_{u \in K} f(u) \sum_{t=1}^{T} \|g_t\|_*^2}$$

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

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

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

Trivial lower bound: If diameter of K is D, then $\sup_{u \in K} f(u) \geq D^2/8$.

Questions?

Paper: http://arxiv.org/abs/1502.05744