A Survey of Algorithms and Analysis for Adaptive Online Learning

H. Brendan McMaha
(2017)

Presented by Jiin Seo

July 1, 2019

Outline

Outline

1. Introduction

Online Convex Optimization

- On each round $t \in \{1, 2, \ldots\}$, select a point $x_t \in \mathbb{R}^n$.
  For a convex loss function f_t,

$$\text{Regret}(x^*) := \text{Regret}_T(x^*, f_t) \equiv \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*)$$

- Main interest :

$$\text{Regret}(\mathcal{X}) \equiv \sup_{x^* \in \mathcal{X}} \text{Regret}(x^*)$$

- $\text{Regret}(x^*)$: frequently bounded by a function of $\|x^*\|$

Outline

2. FTRL Family

Follow-the-Leader(FTL)

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^n} f_{1:t}(x)$$

- f_t : strongly convex functions $\Rightarrow$ sublinear regret

Follow-the-Regularized-Leader(FTRL)

- $r_t(x) \geq 0$: smoothing regularizer (adaptive member)

$$x_1 \in \arg\min_{x \in \mathbb{R}^n} r_0(x)$$
$$x_{t+1} = \arg\min_{x \in \mathbb{R}^n} \left(f_{1:t}(x) + r_{0:t}(x) \right) \quad , \text{for } t = 1, 2, \ldots$$

- Flexible decisions
  1. exact $f_{1:t}(x)$ or efficient lower bounds $\overline{f}_{1:t}(x)$
  2. r_t are minimized at fixed stationary point x_1 or current x_t

2. FTRL Family

Linearization of convex ftn

- For convex f_t and subgradient $g_t \in \partial f_t(x_t)$,

$$f_t(x_t) - f_t(x^*) \leq g_t \cdot (x_t - x^*), \forall x^*$$

- Let $\overline{f}_t(x) = g_t \cdot x$: (Linearization)

$$\text{Regret}(x^*, f_t) \leq \text{Regret}\left(x^*, \overline{f}_t\right)$$

- Update

$$x_{t+1} = \arg\min_x \left(g_{1:t} \cdot x + r_{0:t}(x)\right)$$

2. FTRL Family

Regularization in FTRL

1. FTRL-Centered
   - Each r_t is minimized at a fixed point. ($x_1 = \arg\min_x r_0(x)$)
   - $r_{0:t}$: prox-function
2. FTRL-Proximal
   - Each r_t is minimized at x_t.
   - r_t : incremental proximal regularizers

2. FTRL Family

Notation and Definitions

- $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ an extended-value convex function, if

$$\psi(\theta x + (1-\theta)y) \leq \theta\psi(x) + (1-\theta)\psi(y), \quad \theta \in (0,1)$$
$$\text{dom}\,\psi \equiv \{x : \psi(x) < \infty\} \text{ is the convex set.}$$

- ψ is proper if

$$\exists x \in \mathbb{R}^n \text{ s.t. } \psi(x) < +\infty \quad \text{and } \forall x \in \mathbb{R}^n, \psi(x) > -\infty$$

- extended-value proper convex functions ="convex functions"
- $g(\in \partial\psi(x))$ is subgradient of ψ, if

$$\forall y \in \mathbb{R}^n, \psi(y) \geq \psi(x) + g \cdot (y - x)$$

3. A General Analysis Technique

Notation and Definitions

- Convex conjugate (or Fenchel conjugate) of $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$:

$$\psi^\star(g) \equiv \sup_x \left[g \cdot x - \psi(x) \right]$$

- Dual norm :

$$\|x\|_\star \equiv \sup_{y : \|y\| \leq 1} x \cdot y$$

- $\|\cdot\|_{(t)}$, $\|\cdot\|_{(t),\star}$: functions of the round t

2. FTRL Family

Notation and Definitions

- $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is σ -strongly convex w.r.t. a norm $\|\cdot\|$, if

$$\forall g \in \partial\psi(x), \quad \psi(y) \geq \psi(x) + g \cdot (y - x) + \frac{\sigma}{2}\|y - x\|^2, \forall x, y \in \mathbb{R}^n.$$

- A function ψ is σ -strongly-smooth w.r.t. a norm $\|\cdot\|$, if it is differentiable and for all x, y we have

$$\psi(y) \leq \psi(x) + \nabla\psi(x) \cdot (y - x) + \frac{\sigma}{2}\|y - x\|^2$$

Lemma 9 Let ψ be closed and convex. Then,
ψ is σ-strongly convex w.r.t. norm $\|\cdot\|$.
$\Leftrightarrow$ $\psi^\star$ is $\frac{1}{\sigma}$ -strongly smooth with respect to the dual norm $\|\cdot\|_\star$.

- fundamental duality between strongly convex and strongly smooth functions

2. FTRL Family

Setting1

- Algorithm :

$$x_1 \in \arg\min_{x \in \mathbb{R}^n} r_0(x)$$

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^n} \underbrace{\left(f_{1:t}(x) + r_{0:t}(x)\right)}_{h_{0:t}(x)} \quad , \text{for } t = 1, 2, \ldots$$

- $r_t(x) \geq 0, \quad f_t : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}, \quad f_t, r_t$: convex.
- $h_{0:t} = r_{0:t} + f_{1:t}, \quad \mathrm{dom}\left(h_{0:t}\right) \neq \emptyset, \quad \partial f_t\left(x_t\right) \neq \emptyset$

2. FTRL Family

Theorem 1. General FTRL Bound

- Setting1
- r_t are chosen s.t. $h_{0:t} + f_{t+1} = r_{0:t} + f_{1:t+1}$ is 1-strongly convex w.r.t. $\|\cdot\|_{(t)}$.
- Then, $\forall x^* \in \mathbb{R}^n$ and $\forall T > 0$,

$$\text{Regret}_T(x^*) \leq r_{0:T-1}(x^*) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|^2_{(t-1),\star}$$

Theorem 2. FTRL-Proximal Bound

- Setting1
- r_t are chosen s.t. $h_{0:t} = r_{0:t} + f_{1:t}$ is 1-strongly convex w.r.t. $\|\cdot\|_{(t)}$.
- r_t are proximal. (i.e. x_t is a minimizer of r_t)
- Then, $\forall x^* \in \mathbb{R}^n$ and $\forall T > 0$,

$$\text{Regret}_T(x^*) \leq r_{0:T}(x^*) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|^2_{(t),\star} \quad , g_t \in \partial f_t(x_t)$$

off-by-one difference

2. FTRL Family

Non-Adaptive Algorithms

- Setting1
- r_0 : fixed non-adaptive regularizer, 1-strongly convex w.r.t. $\|\cdot\|$
- $r_t(x) = 0 \quad , t \geq 1$
- Then, $\forall t,\ \|x\|_{(t),\star} = \|x\|_\star$ and so both thms provide the identical statement

$$\text{Regret}\,(x^*) \leq r_0\,(x^*) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|_\star^2$$

2. FTRL Family

Incorporating a Feasible Set

- $x \in \mathbb{R}^n \rightarrow x_t \in \mathcal{X}$: convex feasible set

$$I_{\mathcal{X}}(x) = \begin{cases} 0 & , x \in \mathcal{X} \\ \infty & , \text{otherwise} \end{cases}$$

- $r_0 \equiv r_0 + I_{\mathcal{X}}$

$$\arg\min_{x \in \mathbb{R}^n} f_{1:t}(x) + r_{0:t}(x) + I_{\mathcal{X}}(x) = \arg\min_{x \in \mathcal{X}} f_{1:t}(x) + r_{0:t}(x)$$

- $I_{\mathcal{X}}$ appears in regret bound.

$$(\text{Thm1}) : \ \text{Regret}_T\left(x^*\right) \leq I_{\mathcal{X}}\left(x^*\right) + r_{0:T-1}\left(x^*\right) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|^2_{(t-1),\star}$$

Outline

3. A General Analysis Technique

Inductive Lemmas

Lemma 5 (Strong FTRL Lemma).

- f_t : seq. of arbitrary (possibly non-convex) loss functions
- r_t : arbitrary non-nega regularization functions, s.t.
- $h_{0:t}(x) \equiv f_{1:t}(x) + r_{0:t}(x)$
- $x_{t+1} = \arg\min_x h_{0:t}(x)$: well defined

Then, the algorithm that selects these x_t achieves

$$\text{Regret}\left(x^*\right) \leq r_{0:T}\left(x^*\right) + \sum_{t=1}^{T}\left[h_{0:t}\left(x_t\right) - h_{0:t}\left(x_{t+1}\right) - r_t\left(x_t\right)\right]$$

(Standard FTRL Lemma).

$$\text{Regret}\left(x^*\right) \leq r_{0:T}\left(x^*\right) + \sum_{t=1}^{T}\left[f_t\left(x_t\right) - f_t\left(x_{t+1}\right)\right].$$

3. A General Analysis Technique

Proof of Lemma 5 (Strong FTRL Lemma).

- FTL algorithm against $h_1, \dots h_T$

$$\sum_{t=1}^{T} \left(f_t\left(x_t\right) + r_t\left(x_t\right)\right) - f_{1:T}\left(x^*\right) - r_{0:T}\left(x^*\right)$$

$$= \sum_{t=1}^{T} h_t\left(x_t\right) - h_{0:T}\left(x^*\right) = \sum_{t=1}^{T} \left(h_{0:t}\left(x_t\right) - h_{0:t-1}\left(x_t\right)\right) - h_{0:T}\left(x^*\right)$$

$$\leq \sum_{t=1}^{T} \left(h_{0:t}\left(x_t\right) - h_{0:t-1}\left(x_t\right)\right) - h_{0:T}\left(x_{T+1}\right) \qquad \left(x_{T+1} \text{ minimizes } h_{0:T}\right)$$

$$\leq \sum_{t=1}^{T} \left(h_{0:t}\left(x_t\right) - h_{0:t}\left(x_{t+1}\right)\right) \quad \left(, -h_0\left(x_1\right) = -r_0\left(x_1\right) \leq 0\right)$$

■ Proof of Lemma5

3. A General Analysis Technique

Inductive Lemmas

Corollary 6

- The same conditions as Lemma 5
- $f_t(x) = g_t \cdot x_t$: Linear loss

Then,

$$h_{0:t}(x_t) - h_{0:t}(x_{t+1}) - r_t(x_t) = r^\star_{0:t}(-g_{1:t}) - r^\star_{0:t-1}(-g_{1:t-1}) + g_t \cdot x_t$$

which implies

$$\text{Regret}(x^*) \leq r_{0:T}(x^*) + \sum_{t=1}^{T} \left[r^\star_{0:t}(-g_{1:t}) - r^\star_{0:t-1}(-g_{1:t-1}) + g_t \cdot x_t \right]$$

3. A General Analysis Technique

Proof of Corollary 6.

1. $h_{0:t}(x_{t+1}) \rightarrow$Fenchel conjugate:

$$\begin{aligned} r^{\star}_{0:t}(-g_{1:t}) &= \max_{x}\left[-g_{1:t}\cdot x - r_{0:t}(x)\right] && \text{: Def. of Fenchel conjugate} \\ &= -\min_{x}\left[g_{1:t}\cdot x + r_{0:t}(x)\right] \\ &= -h_{0:t}(x_{t+1}) \end{aligned}$$

2.

$$\begin{aligned} h_{0:t}(x_t) - r_t(x_t) &= g_{1:t}\cdot x_t + r_{0:t}(x_t) - r_t(x_t) \\ &= g_{1:t-1}\cdot x_t + r_{0:t-1}(x_t) + g_t\cdot x_t \\ &= h_{0:t-1}(x_t) + g_t\cdot x_t \\ &= -r^{\star}_{0:t-1}(-g_{1:t-1}) + g_t\cdot x_t && \text{(using 1 with } t\rightarrow t-1) \end{aligned}$$

■Proof of Coro6

3. A General Analysis Technique

Tools from Convex Analysis

Lemma 7.

- $\phi_1 : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$: convex function s.t. $\exists x_1 = \arg\min_x \phi_1(x)$
- ψ : convex function s.t. $\phi_2(x) = \phi_1(x) + \psi(x)$ is strongly convex w.r.t. norm $\|\cdot\|$
- $x_2 = \arg\min_x \phi_2(x)$.

Then,

$$\forall b \in \partial\psi\left(x_1\right), \quad \|x_1 - x_2\| \leq \|b\|_\star$$

and

$$\forall x', \quad \phi_2\left(x_1\right) - \phi_2\left(x'\right) \leq \frac{1}{2}\|b\|_\star^2$$

3. A General Analysis Technique

Proof of Lemma 7.

1. $\phi_2(x) = \phi_1(x) + \psi(x)$: 1-strongly convex w.r.t. $\|\cdot\|$
2. Construct ϕ_1' : 1-strongly convex w.r.t. $\|\cdot\|$, but minimizer is x_1.

$$\begin{aligned} \phi_1'(x) &\equiv \phi_1(x) + \psi(x) - b \cdot x, && \text{: 1-strongly convex w.r.t. } \|\cdot\| \\ 0 &\in \partial(\psi(x) - b \cdot x)|_{x=x_1} && \text{: } x_1 = \arg\min \phi_1'(x) \end{aligned}$$

$$\Rightarrow \quad \phi_2(x) = \phi_1'(x) + b \cdot x$$

3. Applying Lemma 16 to ϕ_1' and ϕ_2

$$\forall x', \quad \phi_2(x_1) - \phi_2(x') \leq \phi_2(x_1) - \phi_2(x_2) \leq \frac{1}{2}\|b\|_\star^2$$

and $\|x_1 - x_2\| \leq \|b\|_\star$

■ Proof of Lemma7

3. A General Analysis Technique

Tools from Convex Analysis

Lemma 16

- $\phi_1 : \mathbb{R}^n \to \mathbb{R}$: strongly convex w.r.t. norm $\|\cdot\|$
- $\phi_2(x) := \phi_1(x) + b \cdot x$ for $b \in \mathbb{R}^n$.
- $x_1 = \arg\min_x \phi_1(x), \quad x_2 = \arg\min_x \phi_2(x)$

Then,

$$\phi_2(x_1) - \phi_2(x_2) \leq \frac{1}{2}\|b\|_\star^2, \quad \text{and} \quad \|x_1 - x_2\| \leq \|b\|_\star$$

3. A General Analysis Technique

Proof of Lemma 16.

1. Fenchel conjugate:

$$
\begin{aligned}
-\phi_1^\star(0) &= -\max_x \left[0 \cdot x - \phi_1(x)\right] = \min_x \left[\phi_1(x)\right] = \phi_1\left(x_1\right) \\
-\phi_1^\star(-b) &= -\max_x \left[-b \cdot x - \phi_1(x)\right] = \min_x \left[b \cdot x + \phi_1(x)\right] \\
&= b \cdot x_2 + \phi_1\left(x_2\right)
\end{aligned}
$$

2. $\phi_1^\star$: strongly-smooth (by Lemma 9)

$$
\begin{aligned}
\phi_1^\star(-b) &\leq \phi_1^\star(0) + \nabla\phi_1^\star(0) \cdot (-b - 0) + \frac{1}{2}\| - b - 0\|^2 \\
&= \phi_1^\star(0) + x_1 \cdot (-b - 0) + \frac{1}{2}\|b\|_\star^2 \qquad (, x_1 = \nabla\phi_1^\star(0))
\end{aligned}
$$

Proof of Lemma 16.

3. 1st bound : Combining 1 and 2

$$\begin{aligned}\phi_2(x_1) - \phi_2(x_2) &= \phi_1(x_1) + b \cdot x_1 - \phi_1(x_2) - b \cdot x_2 \\ &= -\phi_1^\star(0) + b \cdot x_1 + \phi_1^\star(-b) \\ &\leq -\phi_1^\star(0) + b \cdot x_1 + \phi_1^\star(0) + x_1 \cdot (-b) + \frac{1}{2}\|b\|_\star^2 \\ &= \frac{1}{2}\|b\|_\star^2\end{aligned}$$

4. 2nd bound : using Lemma 15
$\nabla\phi_1^\star(0) = x_1$, and $\nabla\phi_1^\star(-b) = x_2$

$$\|x_1 - x_2\| = \|\nabla\phi_1^*(0) - \nabla\phi_1^\star(-b)\| \leq \|b\|_\star$$

■ Proof of Lemma16

3. A General Analysis Technique

Tools from Convex Analysis

Lemma 15.

- ψ : 1 -strongly convex w.r.t. $\|\cdot\|$
  ($\Leftrightarrow \psi^*$: 1 -strongly smooth w.r.t. $\|\cdot\|_\star$)

Then,

$$\left\|\nabla\psi^\star(z) - \nabla\psi^\star\left(z'\right)\right\| \leq \left\|z - z'\right\|_\star$$

and

$$\arg\min_x \left[g \cdot x + \psi(x)\right] = \nabla\psi^\star(-g)$$

3. A General Analysis Technique

Tools from Convex Analysis

Corollary 8.

- $\phi_1 : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$: a convex function s.t. $\exists x_1 = \arg\min_x \phi_1(x)$
- ψ , Ψ : convex functions s.t. $\phi_2(x) = \phi_1(x) + \psi(x) + \Psi(x)$: strongly convex *w.r.t.* norm $\|\cdot\|$.
- $x_2 = \arg\min_x \phi_2(x)$.

Then, $\forall b \in \partial\psi(x_1)$ and $\forall x'$,

$$\phi_2(x_1) - \phi_2(x') \leq \frac{1}{2}\|b\|_\star^2 + \Psi(x_1) - \Psi(x_2)$$

3. A General Analysis Technique

Proof of Corollary 8.

1. $x_2' := \arg\min_x [\phi_1(x) + \psi(x)]$

$$\phi_1(x_1) + \psi(x_1) - \phi_1(x_2') - \psi(x_2') \le \frac{1}{2}\|b\|_\star^2 \quad \text{(by Lemma7)}$$
$$\phi_1(x_2') + \psi(x_2') \le \phi_1(x_2) + \psi(x_2) \quad \text{(by Def)}$$

2.

$$\begin{aligned}\forall x', \quad & \phi_2(x_1) - \phi_2(x') \\ &= \phi_1(x_1) + \psi(x_1) + \Psi(x_1) - \phi_1(x_2) - \psi(x_2) - \Psi(x_2) \\ &\le \phi_1(x_1) + \psi(x_1) + \Psi(x_1) - \phi_1(x_2') - \psi(x_2') - \Psi(x_2) \\ &\le \frac{1}{2}\|b\|_\star^2 + \Psi(x_1) - \Psi(x_2)\end{aligned}$$

■ Proof of Corollary 8

3. A General Analysis Technique

Theorem 1. General FTRL Bound

- Setting1
- r_t are chosen s.t. $h_{0:t} + f_{t+1} = r_{0:t} + f_{1:t+1}$ is 1-strongly convex w.r.t. $\|\cdot\|_{(t)}$.
- Then, $\forall x^* \in \mathbb{R}^n$ and $\forall T > 0$,

$$\text{Regret}_T\left(x^*\right) \leq r_{0:T-1}\left(x^*\right) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|_{(t-1),\star}^2$$

3. A General Analysis Technique

General FTRL including FTRL-Centered (Proof of Theorem 1)

1. Bound the stability terms : $h_{0:t}(x_t) - h_{0:t}(x_{t+1})$
   - For fixed t,

$$\begin{aligned} \phi_1(x) &:= h_{0:t-1}(x) && , x_t = \arg\min_x \phi_1(x) \\ \phi_2(x) &:= h_{0:t-1}(x) + f_t(x) && : \text{1-strongly-convex w.r.t. } \|\cdot\|_{(t-1)} \end{aligned}$$

   - applying Lemma 7 :

$$\phi_2(x_t) - \phi_2(x_{t+1}) \le \frac{1}{2}\|g_t\|^2_{(t-1),\star} \text{ for } g_t \in \partial f_t(x_t)$$

$$\begin{aligned} &h_{0:t}(x_t) - h_{0:t}(x_{t+1}) - r_t(x_t) \\ &\quad = \phi_2(x_t) + r_t(x_t) - \phi_2(x_{t+1}) - r_t(x_{t+1}) - r_t(x_t) \\ &\quad \le \frac{1}{2}\|g_t\|^2_{(t-1),\star} && , r_t(x) \ge 0 \end{aligned}$$

3. A General Analysis Technique

General FTRL including FTRL-Centered (Proof of Theorem 1)

2. Sum these up for all rounds. & Use Lemma 5.

$$\begin{aligned}
\text{Regret}\left(x^*\right) &\leq r_{0:T}\left(x^*\right) + \sum_{t=1}^{T}\left[h_{0:t}\left(x_t\right) - h_{0:t}\left(x_{t+1}\right) - r_t\left(x_t\right)\right] \\
&\leq r_{0:T}\left(x^*\right) + \frac{1}{2}\sum_{t=1}^{T}\left\|g_t\right\|_{(t-1),\star}^2 \\
&= r_{0:T-1}\left(x^*\right) + \frac{1}{2}\sum_{t=1}^{T}\left\|g_t\right\|_{(t-1),\star}^2 \qquad , r_T(x) := 0 \text{ (WLOG)}
\end{aligned}$$

■Proof of Thm1

3. A General Analysis Technique

Theorem 2. FTRL-Proximal Bound

- Setting1
- r_t are chosen s.t. $h_{0:t} = r_{0:t} + f_{1:t}$ is 1-strongly convex w.r.t. $\|\cdot\|_{(t)}$.
- r_t are proximal. (i.e. x_t is a minimizer of r_t)
- Then, $\forall x^* \in \mathbb{R}^n$ and $\forall T > 0$,

$$\text{Regret}_T(x^*) \leq r_{0:T}(x^*) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|^2_{(t),\star} \quad , g_t \in \partial f_t(x_t)$$

3. A General Analysis Technique

FTRL-Proximal (Proof of Theorem 2)

1. Bound the stability terms : $h_{0:t}(x_t) - h_{0:t}(x_{t+1})$
   - For fixed t,
     $\phi_1(x) := f_{1:t-1}(x) + r_{0:t}(x) = h_{0:t}(x) - f_t(x)$
     $x_t = \arg\min_x \phi_1(x)$, $\quad x_{t+1} = \arg\min_x [\phi_1(x) + f_t(x)]$
   - applying Lemma 7 and $\phi_2(x) = \phi_1(x) + f_t(x) = h_{0:t}(x)$

$$
\begin{aligned}
& h_{0:t}(x_t) - h_{0:t}(x_{t+1}) - r_t(x_t) \\
& \qquad \leq h_{0:t}(x_t) - h_{0:t}(x_{t+1}) \qquad , (r_t(x) \geq 0) \\
& \qquad = \phi_1(x_t) + f_t(x_t) - \phi_1(x_{t+1}) - f_t(x_{t+1}) \\
& \qquad \leq \frac{1}{2} \|g_t\|^2_{(t),\star} \qquad , (\text{Lemma 7})
\end{aligned}
$$

3. A General Analysis Technique

FTRL-Proximal (Proof of Theorem 2)

2. Sum up these for all rounds. & Use Lemma 5.

$$\begin{aligned}\text{Regret}\left(x^*\right) &\leq r_{0:T}\left(x^*\right) + \sum_{t=1}^{T}\left[h_{0:t}\left(x_t\right) - h_{0:t}\left(x_{t+1}\right) - r_t\left(x_t\right)\right] \\ &\leq r_{0:T}\left(x^*\right) + \frac{1}{2}\sum_{t=1}^{T}\left\|g_t\right\|_{(t),\star}^2\end{aligned}$$

■ Proof of Thm2

Outline

4. Additional Regularization Terms and Composite Objectives

generalized FTRL

- Ψ : non-negative convex function
- $\alpha_t \geq 0$ for $t \geq 1$: non-increasing in t
- $x_1 = \arg\min_x \Psi(x) = \arg\min_x r_0(x), \quad$ w.l.o.g. $\Psi(x_1) = 0$
- Objective function :

$$h_t(x) = g_t \cdot x + \underbrace{\alpha_t \Psi(x)}_{\text{: additional regularization}} + r_t(x)$$

- FTRL update

$$x_{t+1} = \arg\min_x h_{0:t}(x) = \arg\min g_{1:t} \cdot x + \alpha_{1:t}\Psi(x) + r_{0:t}(x)$$

4. Additional Regularization Terms and Composite Objectives

generalized FTRL

non-smooth regularizer Ψ

- L_2 penalty : $\Psi(x) = \|x\|_2^2$
- L_1 penalty : $\Psi(x) = \|x\|_1 \Rightarrow$ sparse solutions
- indicator function : $\Psi(x) = I_{\mathcal{X}}(x) \Rightarrow$ feasible solutions

4. Additional Regularization Terms and Composite Objectives

generalized FTRL

1. Ψ terms in the objective g_t
   - $f_t^{\Psi}(x) \equiv g_t \cdot x + \alpha_t \Psi(x)$
   - Regret

$$\text{Regret}\left(x^*, f^{\Psi}\right) \equiv \sum_{t=1}^{T} f_t^{\Psi}\left(x_t\right) - f_t^{\Psi}\left(x^*\right)$$

2. Ψ terms in the regularizer r_t
   - $f_t(x) = g_t \cdot x$
   - Regret

$$\text{Regret}\left(x^*, g_t\right) \equiv \sum_{t=1}^{T} g_t \cdot x_t - g_t \cdot x^*$$

4. Additional Regularization Terms and Composite Objectives

Theorem 10. FTRL-Proximal Bounds for Composite Objectives

- Ψ : a non-negative convex function,
  $x_1 = \arg\min_x \Psi(x), \quad \Psi(x_1) = 0$
- $\alpha_t \geq 0$ be a non-increasing seq. of constants
- Setting1
- $h_t(x) := g_t \cdot x + \alpha_t \Psi(x) + r_t(x)$
- r_t are choosen s.t. $h_{0:t}$: 1 -strongly-convex w.r. t $\|\cdot\|_{(t)}$
- r_t : proximal (x_t is a global minimizer of r_t)

1. Case1 : $f_t^{\Psi}(x) = g_t \cdot x + \alpha_t \Psi(x)$

$$\text{Regret}\left(x^*, f^{\Psi}\right) \leq r_{0:T}(x^*) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|_{(t),\star}^2$$

2. Case2 : $f_t(x) = g_t \cdot x$

$$\text{Regret}(x^*, g_t) \leq r_{0:T}(x^*) + \alpha_{1:T}\Psi(x^*) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|_{(t),\star}^2$$

4. Additional Regularization Terms and Composite Objectives

Proof of Theorem 10.

1. Bound the stability terms : $h_{0:t}(x_t) - h_{0:t}(x_{t+1})$

- For fixed t,

$$\begin{aligned} \phi_1(x) &:= h_{0:t-1}(x) + r_t(x) && , x_t = \arg\min_x \phi_1(x) \\ \phi_2(x) &= \phi_1(x) + g_t \cdot x + \alpha_t \Psi(x) = h_{0:t}(x) && , x_{t+1} = \arg\min_x \phi_2(x) \end{aligned}$$

- applying Corollary 8

$$\begin{aligned} & h_{0:t}(x_t) - h_{0:t}(x_{t+1}) - r_t(x_t) \\ & \quad \le h_{0:t}(x_t) - h_{0:t}(x_{t+1}) \qquad , (r_t(x) \ge 0) \\ & \quad = \phi_1(x_t) + f_t(x_t) + \alpha_t \Psi(x_t) - \phi_1(x_{t+1}) - f_t(x_{t+1}) - \alpha_t \Psi(x_{t+1}) \\ & \quad \le \frac{1}{2} \|g_t\|^2_{(t),\star} + \alpha_t \Psi(x_t) - \alpha_t \Psi(x_{t+1}) \qquad , (\text{Corollary 8}) \end{aligned}$$

4. Additional Regularization Terms and Composite Objectives

Proof of Theorem 10.

2. Sum up these for all rounds. & Use Lemma 5.

$$\begin{aligned}
&\sum_{t=1}^{T} \alpha_t \Psi(x_t) - \alpha_t \Psi(x_{t+1}) \\
&\quad = \alpha_1 \Psi(x_1) - \alpha_T \Psi(x_{T+1}) + \sum_{t=2}^{T} [\alpha_t \Psi(x_t) - \alpha_{t-1} \Psi(x_t)] \\
&\quad \leq 0 \qquad (\Psi(x) \geq 0, \alpha_t \leq \alpha_{t-1}, \text{ and } \Psi(x_1) = 0)
\end{aligned}$$

$$\begin{aligned}
&\sum_{t=1}^{T} [h_{0:t}(x_t) - h_{0:t}(x_{t+1}) - r_t(x_t)] \\
&\quad \leq \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{(t),\star}^2 + \sum_{t=1}^{T} [\alpha_t \Psi(x_t) - \alpha_t \Psi(x_{t+1})] \\
&\quad \leq \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_{(t),\star}^2
\end{aligned}$$

4. Additional Regularization Terms and Composite Objectives

Proof of Theorem 10.

3. Case1 : $f_t(x) \leftarrow g_t \cdot x + \alpha_t \Psi(x)$ and $r_t \leftarrow r_t$

$$\text{Regret}\left(x^*, f^{\Psi}\right) \leq r_{0:T}\left(x^*\right) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|^2_{(t),*} \qquad \text{(Lemma 5)}$$

4. Case2 : $f_t(x) \leftarrow g_t \cdot x$ and $r_t(x) \leftarrow \alpha_t \Psi(x) + r_t(x)$

$$\text{Regret}\left(x^*, g_t\right) \leq r_{0:T}\left(x^*\right) + \alpha_{1:T}\Psi\left(x^*\right) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|^2_{(t),*} \qquad \text{(Lemma 5)}$$

■ Proof of Thm10

Outline

5. Application to Specific Algorithms

Lemma 3.

- Setting1.
- $r_{0:t}(x) \geq r_{0:t-1}(x)$,
- $r^{\star}_{0:t}(x) \leq r^{\star}_{0:t-1}(x)$, where $r^{\star}_{0:t}$ is the convex-conjugate of $r_{0:t}$.
- If r_t is σ_t -strongly convex w.r.t. a norm $\|\cdot\|$ for $\sigma_t \geq 0$,
- Then,

$$r_{0:t} : \begin{cases} \sigma_{0:t}\text{-strongly convex w.r.t. } \|\cdot\| \\ \text{1-strongly-convex w.r.t } .\|x\|_{(t)} = \sqrt{\sigma_{0:t}}\|x\| \quad , \|x\|_{(t),\star} = \frac{1}{\sqrt{\sigma_{0:t}}}\|x\| \end{cases}$$

Learning rate schedule (η_t)

$$\eta_t = \frac{1}{\sigma_{0:t}} \Leftrightarrow \sigma_t = \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \quad , \textit{where } \sigma_0 = \frac{1}{\eta_0}$$

(Assumption) : $f_t(x) = g_t \cdot x$

5. Application to Specific Algorithms

Non-adaptive algorithm

- OGD with constant η

$$(\text{OGD}) : x_{t+1} = x_t - \eta g_t \quad , \eta > 0 \Leftrightarrow x_{t+1} = -\eta g_{1:t}$$

- FTRL with $r_0(x) = \frac{1}{2\eta}\|x\|_2^2$ and $r_t(x) = 0$ for $t \geq 1$

$$x_{t+1} = \arg\min_x g_{1:t} \cdot x + \frac{1}{2\eta}\|x\|_2^2 \Leftrightarrow x_{t+1} = -\eta g_{1:t}$$

- Regret

$$\begin{aligned}\text{Regret}_T(x^*) &\leq r_{0:T-1}(x^*) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|_{(t-1),*}^2 \\ &= \frac{1}{2\eta}\|x^*\|_2^2 + \frac{1}{2}\sum_{t=1}^{T} \eta \|g_t\|_2^2 \quad \text{,using } \|x\|_{(t),\star} = \sqrt{\eta}\|x\|_2\end{aligned}$$

5. Application to Specific Algorithms

Non-adaptive algorithm

- (Assumption) $\|x^*\|_2 \leq R, \|g_t\|_2 \leq G$
  we want to minimize regret after T' rounds.
- Choosing $\eta = \frac{R}{G\sqrt{T'}}$ minimize bound of $\text{Regret}_T(x^*)$

$$\begin{aligned}\text{Regret}_T(x^*) &\leq \frac{RG}{2}\sqrt{T'} + \frac{RG}{2}\frac{T}{\sqrt{T}} \\ &= RG\sqrt{T} \quad (\text{ ,when } T = T')\end{aligned}$$

- bound : $\mathcal{O}(\sqrt{T})$ when $T = \mathcal{O}(T')$

5. Application to Specific Algorithms

Non-adaptive algorithm

- OGD with Constant η and Feasible Set $\mathcal{X}$
- (Assumption) $\|x^*\|_2 \leq R$
- $r_0(x) = \frac{1}{2\eta}\|x\|_2^2 + I_{\mathcal{X}}(x)$

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^n} g_{1:t} \cdot x + \frac{1}{2\eta}\|x\|_2^2 + I_{\mathcal{X}}(x) = \arg\min_{x \in \mathcal{X}} g_{1:t} \cdot x + \frac{1}{2\eta}\|x\|_2^2$$

- equiv. to 2-step update (Lazy projection)

$$\Leftrightarrow \begin{cases} u_{t+1} = \arg\min_{x \in \mathbb{R}^n} g_{1:t} \cdot x + \frac{1}{2\eta}\|x\|^2 \\ x_{t+1} = \Pi_{\mathcal{X}}(u_{t+1}) \end{cases} \quad , \Pi_{\mathcal{X}}(u) \equiv \arg\min_{x \in \mathcal{X}} \|x - u\|_2$$

Appendix C.2.

5. Application to Specific Algorithms

Dual Averaging

- adaptive FTRL-Centered algorithm with linearized loss function
- bound : $\mathcal{O}(\sqrt{T}), \forall T$
- $r_t(x) = \frac{\sigma_t}{2}\|x\|_2^2, \quad \sigma_t \geq 0$
- $r_{0:t}$: 1 -strongly-convex w.r.t. the norm $\|x\|_{(t)} = \sqrt{\sigma_{0:t}}\|x\|_2$
- dual norm : $\|x\|_{(t),\star} = \frac{1}{\sqrt{\sigma_{0:t}}}\|x\|_2 = \sqrt{\eta_t}\|x\|_2$

$$\forall T, \text{ Regret }_T(x^*) \leq r_{0:T-1}(x^*) + \frac{1}{2}\sum_{t=1}^{T}\|g_t\|_{(t-1),\star}^2$$

$$= \frac{1}{2\eta_{T-1}}\|x^*\|_2^2 + \frac{1}{2}\sum_{t=1}^{T}\eta_{t-1}\|g_t\|_2^2$$

5. Application to Specific Algorithms

Dual Averaging

- (Assumption) $\|x^*\|_2 \leq R, \|g_t\|_2 \leq G$
- Choosing $\eta_t = \frac{R}{\sqrt{2}G\sqrt{t+1}}$ and using $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$

$$\forall T,\ \text{Regret}_T(x^*) \leq \frac{\sqrt{2}}{2}\left(R + \frac{\|x^*\|_2^2}{R}\right) G\sqrt{T}$$

- When $\|x^*\| \leq R$, we have Regret $\leq \sqrt{2}RG\sqrt{T}$
- RDA(Regularized Dual Averaging algorithm)
  linearized f_t + non-smooth regularization (adding $\lambda_1\|x\|_1 + \lambda_2\|x\|_2^2$ to r_0)

5. Application to Specific Algorithms

FTRL-Proximal

- $\mathcal{X} \subseteq \{x | \|x\|_2 \leq R\}$
- $r_0(x) = I_{\mathcal{X}}(x), \quad r_t(x) = \frac{\sigma_t}{2} \|x - x_t\|_2^2, t > 1$
- $r_{0:t}(x^*) \leq \frac{\sigma_{1:t}}{2}(2R)^2, \forall x^* \in \mathcal{X}$, since each $x_t \in \mathcal{X}$
- $r_{0:t}$ is 1 -strongly-convex w.r.t. the norm $\|x\|_{(t)} = \sqrt{\sigma_{1:t}}\|x\|_2$
- dual norm : $\|x\|_{(t),*} = \frac{1}{\sqrt{\sigma_{1it}}}\|x\|_2$

$$\forall T, \ \text{Regret}_T(x^*) \leq r_{0:T}(x^*) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|_{(t),\star}^2$$

$$= \frac{1}{2\eta_T}(2R)^2 + \frac{1}{2}\sum_{t=1}^{T} \eta_t \|g_t\|^2, \text{where } \eta_t = \frac{1}{\sigma_{1:t}}$$

- (Assumption) $\|x^*\| \leq R$, $\quad \|g_t\|_2 \leq G$
- Choosing $\eta_t = \frac{\sqrt{2}R}{G\sqrt{t}}$.

$$(x^*) \leq 2\sqrt{2}RG\sqrt{T}$$

- One would hope x_t is closer to x^* than 0

5. Application to Specific Algorithms

FTRL-Proximal with Diagonal Matrix Learning Rates

- 1-dimensional + AdaGrad FTRL-Proximal algorithm : adaptive to seq. of g_t
- Let $r_0 = I_{\mathcal{X}}$ with $\mathcal{X} = [-R, R]$ and $\eta_t = \frac{\sqrt{2}R}{\sqrt{\sum_{s=1}^{t} g_s^2}}$.

$$\forall x^* \in \mathcal{X}, \quad \text{Regret}\,(x^*) \leq \frac{1}{2\eta_T}(2R)^2 + \frac{1}{2}\sum_{t=1}^{T} \eta_t \left\|g_t\right\|^2$$

$$\leq 2\sqrt{2}R\sqrt{\sum_{t=1}^{T} g_t^2}$$

- Fully adaptive bound : do not need to know T and G
- R is chosen in advance.

5. Application to Specific Algorithms

FTRL-Proximal with Diagonal Matrix Learning Rates

- n-dimensional AdaGrad-style algorithm
- $\mathcal{X} \subseteq [-R_\infty, R_\infty]^n$
- per-coordinate basis ; For coordinate i, $\eta_{t,i} = \frac{\sqrt{2}R_\infty}{\sqrt{\sum_{s=1}^{t} g_{s,i}^2}}$
- $r_0 = I_{\mathcal{X}}, \quad r_t(x) = \frac{1}{2}\left\| Q_t^{\frac{1}{2}}(x - x_t)\right\|_2^2, t \geq 1$
  $Q_t = \text{diag}(\sigma_{t,i})$: P.S.D. , $\quad \sigma_{t,i} = \eta_{t,i}^{-1} - \eta_{t-1,i}^{-1}$
- $r_{0:t}$ is 1 -strongly-convex w.r.t. the norm $\|x\|_{(t)} = \left\|(Q_{1:t})^{\frac{1}{2}} x\right\|_2$
- dual norm : $\|g\|_{(t),*} = \left\|(Q_{1:t})^{-\frac{1}{2}} g\right\|_2$

$$\forall x^* \in \mathcal{X}, \quad \text{Regret}_T(x^*) \leq r_{0:T}(x^*) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|_{(t),\star}^2$$

$$= r_{0:T}(x^*) + \frac{1}{2}\sum_{t=1}^{T} \left\|(Q_{1:t})^{-\frac{1}{2}} g_t\right\|_2$$

5. Application to Specific Algorithms

FTRL-Proximal with Diagonal Matrix Learning Rates

- Bound of diagonal FTRL-Proximal is at least as good than bound of FTRL proximal.
- Less useful because of $\mathcal{O}\left(n^2\right)$ requirements
- Update simplifies to

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left(g_{1:t} - a_{1:t}\right) \cdot x + \sum_{i=1}^{n} \frac{1}{2\eta_{t,i}} x_i^2 \quad \text{where} \quad a_t = \sigma_t x_t$$

5. Application to Specific Algorithms

$$\sum_{t=1}^{T} 1/\sqrt{t} \leq 2\sqrt{T}$$

Lemma 4. For any non-negative real numbers $a_1, a_2, \ldots, a_n$

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \leq 2\sqrt{\sum_{i=1}^{n} a_i}$$

Proof)

- $n = 1$: Clear!
- Assume the lemma holds for some $n-1$,

$$\begin{aligned}
\sum_{i=1}^{n} \frac{x_i}{\sqrt{\sum_{j=1}^{i} x_j}} &\leq 2\sqrt{\sum_{i=1}^{n-1} x_i + \frac{x_n}{\sqrt{\sum_{i=1}^{n} x_i}}} \quad , (Z \equiv \sum_{i=1}^{n} x_i \text{ and } x \equiv x_n) \\
&= 2\sqrt{Z - x} + \frac{x}{\sqrt{Z}} \quad (\text{ RHS is maximized at } x = 0\ ) \\
&\leq 2\sqrt{Z}
\end{aligned}$$

- using derivative of RHS w.r.t. $x = \frac{-1}{\sqrt{Z-x}} + \frac{1}{\sqrt{Z}} < 0$ for $x > 0$

5. Application to Specific Algorithms

AdaGrad Dual Averaging

- off-by-one difference
- per-coordinate learning rates $\eta_t \approx \frac{R}{\sqrt{G^2+\sum_{s=1}^{t} g_s^2}}, \quad |g_t| \leq G$
- almost fully adaptive

5. Application to Specific Algorithms

AdaGrad Dual Averaging with the Entropic Regularizer ???

- probability distribution selection problems
  $x_t \in \Delta_n$ with $\Delta_n = \{x | \sum_{i=1}^n x_i = 1 \text{ and } x_i \geq 0\}$
- (assumption) $\|g_t\|_\infty \leq G_\infty$
- Using entropic regularizer

$$h(x) = I_\Delta(x) + \log n + \sum_{i=1}^n x_i \log x_i$$

- adaptive regularization schedule:

$$r_{0:t}(x) = \frac{1}{\eta_t} h(x) \qquad \text{where} \qquad \eta_t = \frac{\sqrt{\log n}}{\sqrt{G_\infty^2 + \sum_{s=1}^t \|g_s\|_\infty^2}}, \quad t \geq 0$$

- $x_1 = (1/n, \ldots, 1/n)$,where $h(x) = 0$
- h is maximized at the corners of Δ_n

5. Application to Specific Algorithms

AdaGrad Dual Averaging with the Entropic Regularizer

- h is 1 -strongly-convex w.r.t. L_1
- $r_{0:t}$ is 1 -strongly-convex w.r.t. $\|x\|_{(t)} = \frac{1}{\sqrt{\eta t}}\|x\|_1$, and $\|g\|^2_{(t),\star} = \eta_t \|g\|^2_\infty$

$$
\begin{aligned}
\text{Regret}\,(x^*) &\leq r_{0:T-1}\,(x^*) + \frac{1}{2}\sum_{t=1}^{T} \|g_t\|^2_{(t-1),\star} \\
&\leq \frac{\log n}{\eta_{T-1}} + \frac{1}{2}\sum_{t=1}^{T} \eta_{t-1}\,\|g_t\|^2_\infty \quad \text{since } \forall t, \|g_t\|_\infty \leq G_\infty \\
&\leq \frac{\log n}{\eta_{T-1}} + \frac{\sqrt{\log n}}{2}\sum_{t=1}^{T} \frac{\|g_t\|^2_\infty}{\sqrt{\sum_{s=1}^{t}\|g_s\|^2_\infty}} \qquad \text{Lemma 4 and } \|g_T\|_\infty \\
&\leq 2G_\infty\sqrt{T\log n}
\end{aligned}
$$

5. Application to Specific Algorithms

Strongly Convex Functions

- Suppose f_t is 1 -strongly-convex w.r.t. a norm $\|\cdot\|$, and $r_t(x) \equiv 0\ \forall t$. :Exact FTL
- (Assumption) $\|g_t\| \leq G$
- $\|x\|_{(t)} = \sqrt{t}\|x\|$
- $h_{0:t}(x)$ is 1 -strongly-convex w.r.t. $\|\cdot\|_{(t)}$ (by Lemma 3)
- using $\sum_{t=1}^{T} \frac{1}{t} \leq 1 + \log T$

$$\text{Regret}\,(x^*) \leq \frac{1}{2}\sum_{t=1}^{T} \|g_t\|_{(t),\star}^2 = \frac{1}{2}\sum_{t=1}^{T} \frac{1}{t}\|g_t\|^2 \leq \frac{G^2}{2}(1+\log T)$$

- $\bar{f}_t$, f_t : 1 -strongly-convex w.r.t. $\|\cdot\|$

$$f_t(x) \geq \underbrace{f_t(x_t) + g_t \cdot (x - x_t) + \frac{1}{2}\|x - x_t\|^2}_{\equiv \bar{f}_t}$$

5. Application to Specific Algorithms

Strongly Convex Functions

- FTL on $\overline{f}_t \Leftrightarrow$ OGD
- $x_{t+1} = \arg\min_x \overline{f}_{1:t}(x) \Leftrightarrow x_{t+1} = x_t - \frac{1}{t} g_t$