# Origin of Native FTRL-Proximal

Gyuseung Baek

January 15, 2019

# Native FTRL-Proximal

| DATA | FTRL-PROXIMAL | RDA | FOBOS |
|---|---|---|---|
| BOOKS | 0.874 (0.081) | **0.878** (**0.079**) | 0.877 (0.382) |
| DVD | 0.884 (0.078) | 0.886 (**0.075**) | **0.887** (0.354) |
| ELECTRONICS | 0.916 (0.114) | **0.919** (**0.113**) | 0.918 (0.399) |
| KITCHEN | 0.931 (**0.129**) | **0.934** (0.130) | 0.933 (0.414) |
| NEWS | 0.989 (**0.052**) | **0.991** (0.054) | 0.990 (0.194) |
| RCV1 | 0.991 (**0.319**) | **0.991** (0.360) | 0.991 (0.488) |
| WEB SEARCH ADS | **0.832** (**0.615**) | 0.831 (0.632) | 0.832 (0.849) |

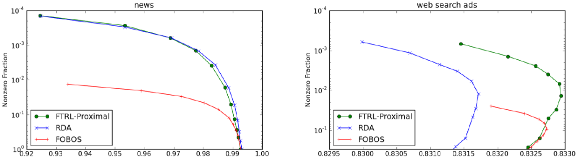



Figure: Table 2, Figure 1, 2 of H. B. McMahan, 2011

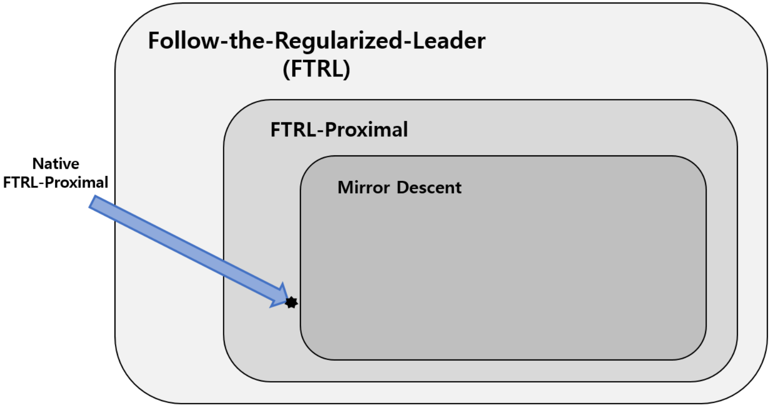
| | Num. Non-Zero's | AucLoss Detriment |
|---|---|---|
| FTRL-PROXIMAL | baseline | baseline |
| RDA | +3% | 0.6% |
| FOBOS | +38% | 0.0% |
| OGD-COUNT | +216% | 0.0% |

Figure: Table 1 of H. B. McMahan et al., 2013

# Native FTRL-Proximal

$$x_{t+1} = \underset{x}{\operatorname{argmin}}\, g_{1:t} \cdot x + t\lambda ||x||_1 + \frac{1}{2} \sum_{s=1}^{t} ||Q_s^{\frac{1}{2}} (x - x_s)||_2^2$$

# Introduction



# Online Convex Optimization(OCO)

- At each round $t \in \{1, 2, \cdots\}$, select a point $x_t \in \mathbb{R}^n$
- From convex loss function $f_t$, observe the $t$ time's loss $f_t(x_t)$

- Regret of the algorithm $\{x_t\}$ at the round $T$ at a given point $x*$

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

  If $\{f_t\}$ and $T$ are clear then we omit them.

- Goal : if our searching space is $\mathcal{X}$, then find the algorithm which minimizes the regret on the set $\mathcal{X}$:

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

# Basic Convex Optimization Definitions

- Assume $\mathcal{X}$ is convex set, $\psi : \mathcal{X} \to \mathbb{R} \bigcup \{\infty\}$ is convex function $\text{dom}\psi \equiv \{x : \psi(x) < \infty\}$
- $g$ is a subgradient of $\psi$ at $x$ if

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

  $\partial\psi(x)$ : Set of subgradients of $\psi$ at $x$
  ***Note*** If $x \in \text{int}(\text{dom}\psi)$, then $\partial\psi(x) \neq \emptyset$.

- Let $|| \cdot ||$ be a norm on $\mathcal{X}$. $\psi : \mathcal{X} \to \mathbb{R} \bigcup \{\infty\}$ is $\sigma$-*strongly* convex function w.r.t. a norm $|| \cdot ||$ if for all $x, y \in cX$,

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

# Basic Convex Optimization Definitions

- $\mathcal{X}^*$ is a dual space correspond to $\mathcal{X}$ if $\mathcal{X}^* = \{\phi : \mathcal{X} \to \mathbb{R} | \phi \text{ is linear.}\}$.
- For a norm $||\cdot||$, the dual norm $||\cdot||_*$ is a norm on $\mathcal{X}^*$. It is given by

$$||\phi||_* \equiv \sup_{x:||x||\leq 1} \phi(x)$$

- $\forall g \in \partial\psi(x), x \in \mathcal{X}, \psi :$ convex $, g(x) \equiv g \cdot x \in \mathcal{X}^*$.

  For convinience, let $||g(\cdot)||_* = ||g||_*$.

# Linearization

- Computing $Regret_T(\mathcal{X}, \{f_t\})$ is hard.
  In general, computing the upper bound of $Regret_T(\mathcal{X})$.
- Let $g_t$ be a subgradient of $f_t$ at $x_t$. Let $\bar{f}_t(x) = g_t \cdot x$. Then

$$Regret_T(\mathcal{X}, \{f_t\}) \leq Regret_T(\mathcal{X}, \{\bar{f}_t\})$$

  since $\forall x^* \in \mathcal{X}, f_t(x^*) - f_t(x_t) \geq g_t(x^* - x_t) = \bar{f}_t(x^*) - \bar{f}_t(x_t)$.
- Linearization help to compute closed form of $x_t$.

# Follow-the-Leader

- $x_{t+1} = \text{argmin}_{x \in \mathcal{X}} f_{1:t}(x)$

- Simplest online algorithm
- Similar to empirical risk minimization(ERM)
- Impractical.

# Follow-the-Regularized-Leader(FTRL)

- Add additional smoothing regularizer $r(x) \geq 0$

$$x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}}\, f_{1:t}(x) + r(x)$$

- Consider regularizer varies while round $T$ increases.
  Let $r_t(x) \geq 0 \quad \forall x \in \mathcal{X}$.
  Then we can consider the adaptive algorithm.

$$\begin{aligned} x_1 &= \underset{x \in \mathcal{X}}{\operatorname{argmin}}\, r_0(x) \\ x_{t+1} &= \underset{x \in \mathcal{X}}{\operatorname{argmin}}\, f_{1:t}(x) + r_{0:t}(x) \quad \text{for } t = 1, 2, \cdots \end{aligned}$$

# FTRL-Centered and FTRL-Proximal

- FTRL-Centered : Each $r_t$ is minimized at a fixed point,
  $x_1 = \text{argmin}_{x \in \mathcal{X}} r_0(x)$
  $r_{0:t}$ is also minimized by $x_1$.
  $r_{0:t}$ is called the *prox-function*.

- FTRL-Proximal : Each $r_t$ is minimized by $x_t$.
  $r_t$ is called incemental proximal regularizers.

# Regret bound of FTRL

- Consider the linearized case. The followings are taken from H. B. McMahan, 2017.
- (Setting 1) $r_t \geq 0$, $f_t, r_t$ : convex. $\text{dom}(r_{0:t} + f_{1:t}) \neq \emptyset$, $\partial f_t(x_t) \neq \emptyset$.
- (Thm 1 - Thm 1 of McMahan, 2017) General FTRL Bound
  *(Setting 1)* $+$
  $r_t$ are chosen s.t. $f_{1:t+1} + r_{0:t}$ is 1-strongly convex w.r.t. some norm $||\cdot||_{(t)}$.
  Then, for any $x^* \in \mathcal{X}$ and $T > 0$,

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

where $g_t \in \partial f_t(x_t)$.

# Regret bound of FTRL

- (Thm 2 - Thm 2 of McMahan, 2017) FTRL-Proximal Bound *(Setting 1)* +
  $r_t$ are chosen s.t. $f_{1:t} + r_{0:t}$ is 1-strongly convex w.r.t. some norm $||\cdot||_{(t)}$ and $r_t$ are proximal. Then, for any $x^* \in \mathcal{X}$ and $T > 0$,

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

  where $g_t \in \partial f_t(x_t)$.

- *off-by-one* difference
  In thm 1, $r_t$ affect $||g_{t+1}||_*$, whereas $r_t$ affect $||g_t||_*$ in thm 2.
  For this reason, FTRL-Proximal can choose $r_t$ adaptive to $g_t$.

# Regret bound of FTRL

- Compute regret bounds for some cases.
- Consider $L2$ regularizer for $r$.
- Show the importance of adaptivity.

# Non-adaptive case

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

$$x_1 = \operatorname*{argmin}_{x \in \mathcal{X}} r_0(x) = 0$$

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} g_{1:t} \cdot x + \frac{1}{2\eta}||x||_2^2 \quad \text{for } t = 1, 2, \cdots$$

$$= x_t - \eta g_t$$

  It is a online grandient descent with constant learning rate.

- By Thm 1,

$$Regret_T(x^*) \leq \frac{1}{2\eta}||x^*||_2^2 + \frac{1}{2}\sum_{t=1}^{T} \eta ||g_t||_2^2$$

- Suppose $||x^*||_2 \leq R, ||g_t||_2 \leq G$.
  If we want to minimize regret after exactly $T'$ round, we need to choose $\eta = \frac{R}{G\sqrt{T'}}$ and we have

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

  for $T = T'$. It does not work when $T \neq O(T')$.

# Dual Averaging

- $r_t(x) = \frac{\sigma_t}{2}||x||_2^2$ for $t \geq 0$. Let $\eta_t = 1/\sigma_{0:t}$. Then

$$x_1 = \underset{x \in \mathcal{X}}{\operatorname{argmin}}\, r_0(x) = 0$$
$$x_{t+1} = \frac{\eta_t}{\eta_{t-1}} x_t - \eta_t g_t \quad \text{for } t = 1, 2, \cdots$$

- By Thm 1,

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

- Suppose $||x^*||_2 \leq R, ||g_t||_2 \leq G$.
If we choose $\eta_t = \frac{R}{\sqrt{2}G\sqrt{t+1}}$, then we have

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

# FTRL-Proximal

- $r_0(x) = I_{\mathcal{X}}(x), r_t(x) = \frac{\sigma_t}{2}||x||_2^2$ for $t \geq 1$. Let $\eta_t = 1/\sigma_{0:t}$. Then

$$x_1 = \text{any } \bar{x} \in \mathcal{X}$$
$$x_{t+1} = x_t - \eta_t g_t \quad \text{for } t = 1, 2, \cdots$$

- By Thm 1,

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

- Suppose $\forall x \in \mathcal{X}, ||x||_2 \leq R, ||g_t||_2 \leq G$.
If we choose $\eta_t = \frac{\sqrt{2}R}{G\sqrt{t}}$, then we have

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

: twice bigger than Dual averaging.
*Reason*: $||r_t||_2 \leq 2R$ in FTRL-Proximal, whereas $||r_t||_2 \leq R$ in Dual Averaging.

# AdaGrad style update

- In previous FTRL-Proximal setting, if we choose

$$\eta_t = \frac{\sqrt{2}R}{\sqrt{\sum_{s=1}^{t} g_s^2}}$$

then we have

$$Regret_T(x^*) \leq 2\sqrt{2}R\sqrt{\sum_{t=1}^{T} g_t^2}$$

It would give better bound than previous results.

# AdaGrad Dual Averaging

- In Dual Averaging setting, it is necessary to choose $\eta_t$ as

$$\eta_t \simeq \frac{R}{G^2 + \sqrt{\sum_{s=1}^{t} g_s^2}}$$

  where $|g_t| \geq G$.

- Additional $G^2$ is due to the "off-by-one" difference.

# Additional regularization

- Consider additional regularization term $\alpha_t \Psi(x)$ on each round $t$ where $\Psi \geq 0$ is convex and $\alpha_t \geq 0$ for $t \geq 1$ are non-increasing in $t$. Further, assume $x_1 = \operatorname{argmin}_{x \in \mathcal{X}} \Psi(x)$ and w.l.o.g. $\Psi(x_1) = 0$.

- (Composite Objective FTRL)

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} g_{1:t} \cdot x + \alpha_{1:t} \Psi(x) + r_{0:t}(x).$$

- $\Psi$ can be not strongly convex, unlike $r$.

# Regret bound of FTRL for Composite Objectives

- Thm 3 (Thm 10 of McMahan, 2017)
  *(Setting 1)* + $r_t$ are chosen s.t.
  $f_{1:t} + \alpha_{1:t}\Psi + r_{0:t}$ is 1-strongly convex w.r.t. some norm $||\cdot||_{(t)}$
  and $r_t$ are proximal. Then, for any $x^* \in \mathcal{X}$ and $T > 0$,

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

# Bregman divergence

- For convex differentiable function $\phi$, the Bregman divergence $\mathcal{B}_\phi$is defined as:

$$\mathcal{B}_\phi(u, v) = \phi(u) - (\phi(v) + \nabla\phi(v) \cdot (u - v))$$

- If we take $\phi(u) = ||u||^2$, then $\mathcal{B}_\phi(u, v) = (u - v)^2$.

# Mirror Descent

- Composite-Objective Mirror Descent

$$\hat{x}_1 = \underset{x}{\operatorname{argmin}}\, r(x)$$
$$\hat{x}_{t+1} = \underset{x}{\operatorname{argmin}}\, g_t \cdot x + \alpha\Psi(x) + \mathcal{B}_r(x, \hat{x}_t) \quad \text{for } t = 1, 2, \cdots$$

- Adaptive Composite-Objective Mirror Descent

$$\hat{x}_1 = \underset{x}{\operatorname{argmin}}\, r_0(x)$$
$$\hat{x}_{t+1} = \underset{x}{\operatorname{argmin}}\, g_t \cdot x + \alpha_t\Psi(x) + \mathcal{B}_{r_{o:t}}(x, \hat{x}_t) \quad \text{for } t = 1, 2, \cdots$$

# Mirror Descent is an FTRL-Proximal Algorithm

- Define $r_t^{\mathcal{B}}$ as

$$r_0^{\mathcal{B}}(x) \equiv r_0(x)$$
$$r_t^{\mathcal{B}}(x) \equiv \mathcal{B}_{r_t}(x, x_t) \quad \text{for } t = 1, 2, \cdots$$

with this regularizer $r_t^{\mathcal{B}}$, define the FTRL-Proximal algorithm

$$x_1 = \underset{x}{\operatorname{argmin}}\, r_0^{\mathcal{B}}(x)$$
$$x_{t+1} = \underset{x}{\operatorname{argmin}}\, g_{1:t} \cdot x + g_{1:t-1}^{(\Psi)} \cdot x + \alpha_t \Psi(x) + r_{o:t}^{\mathcal{B}}(x) \quad \text{for } t = 1, 2, \cdots$$

where $g_t^{(\Psi)} \in \partial(\alpha_t \Psi)(x_{t+1})$ satisfies

$$g_{1:t} + g_{1:t}^{(\Psi)} + \nabla r_{0:t}^{\mathcal{B}}(x_{t+1}) = 0$$

**Then this FTRL-Proximal update is equal to the Adaptive Composite-Objective Mirror Descent update.**

# Native FTRL

- Mirror descent linearizes the past $\alpha_s \Psi(x)$ terms for $s < t$.
- Consider the non-linearized version, Native FTRL algorithm

$$x_1 = \underset{x}{\operatorname{argmin}}\, r_0^{\mathcal{B}}(x)$$

$$x_{t+1} = \underset{x}{\operatorname{argmin}}\, g_{1:t} \cdot x + \alpha_{1:t}\Psi(x) + r_{o:t}^{\mathcal{B}}(x) \quad \text{for } t = 1, 2, \cdots$$

- FTRL-Proximal and Mirror descent has same regret upper bound.
- There can be a substantial *practical differences* for some choices of $\Psi$.

$\Psi(x) = ||x||_1$

- FTRL Proximal give sparser solutions than Mirror descent
- Example) one dimension $x$. $r = ||\cdot||_2^2$, $\alpha_t = \lambda$ for all $t$.

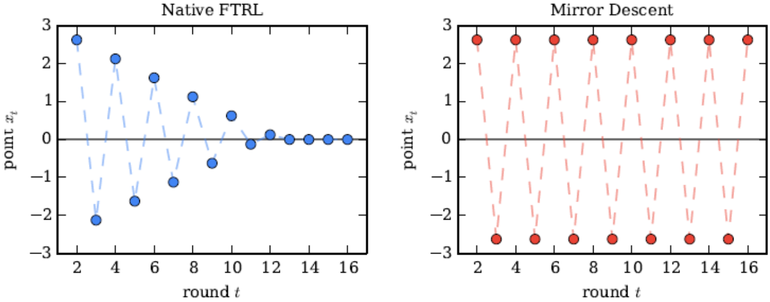



Figure: Fig 4 of H. B. McMahan, 2017

# Lazy and Greedy projection

- FTRL-Proximal : Lazy-projection
- Mirror Descent : Greedy-projection

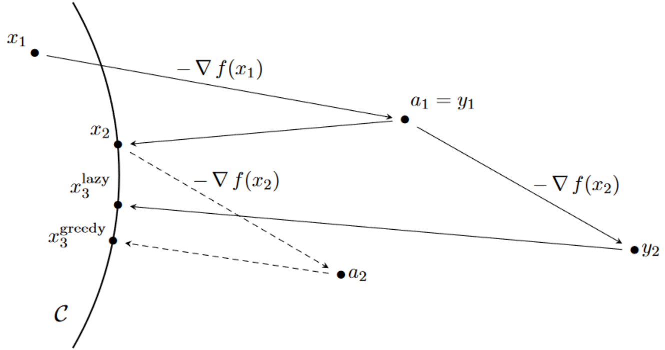



Figure: Fig 1 of J. Kwon & P. Mertikopoulos, 2014

# References I


- McMahan, H. Brendan, and Matthew Streeter. "Adaptive bound optimization for online convex optimization." arXiv preprint arXiv:1002.4908 (2010).
- McMahan, H. Brendan. "Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization." (2011).
- McMahan, H. Brendan, et al. "Ad click prediction: a view from the trenches." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, (2013).
- Kwon, Joon, and Panayotis Mertikopoulos. "A continuous-time approach to online optimization." arXiv preprint arXiv:1401.6956 (2014).
- McMahan, H. Brendan. "A survey of algorithms and analysis for adaptive online learning." The Journal of Machine Learning Research 18.1 (2017): 3117-3166.