Notes on Adaptive Online Learning

Mingjie Qian QIANMINGJIE@GMAIL.COM

Editor: anonymous

Abstract

We will discuss adaptive online learning where the learning rate is scheduled in an adaptive manner. Specifically we will discuss adaptive Follow-The-Regularized-Leader (FTRL) and give regret bound for General FTRL and FTRL-Proximal algorithms. We also discuss adaptive FTRL with an additional regularization term. This chapter is to supplement McMahan (2014) where proofs of some claims are not provided.

1. Adaptive FTRL

The general template for adaptive FTRL is listed below.

$\mathbf{w}_1 \leftarrow \arg\min_{\mathbf{w} \in \Re^n} r_0(\mathbf{w})$

For $t \leftarrow 1, 2, \cdots$

Observe convex loss function $f_t(\mathbf{w}; \{\mathbf{x}_t, y_t\}) : \Re^n \to \Re \cup \{\infty\}$

Incur loss $f_t(\mathbf{w}_t; \{\mathbf{x}_t, y_t\})$

Choose incremental convex regularizer r_t based on $f_1, \cdots, f_t$

Update: $\mathbf{w}_{t+1} \leftarrow \arg\min_{\mathbf{w} \in \Re^n} \sum_{s=1}^{t} f_s(\mathbf{w}) + \sum_{s=0}^{t} r_s(\mathbf{w})$

EndFor

Some choices of the loss function $f_t(\mathbf{w}; \{\mathbf{x}_t, y_t\})$ are

- Square loss: $f_t(\mathbf{w}; \{\mathbf{x}_t, y_t\}) = \frac{1}{2}(y_t - \langle \mathbf{w}, \mathbf{x}_t \rangle)^2$.
- Hinge loss: $f_t(\mathbf{w}; \{\mathbf{x}_t, y_t\}) = \max\{0, 1 - y_t \langle \mathbf{w}, \mathbf{x}_t \rangle\}$.
- Logistic loss: $f_t(\mathbf{w}; \{\mathbf{x}_t, y_t\}) = \log\left(1 + \exp^{-y_t \langle \mathbf{w}, \mathbf{x}_t \rangle}\right)$.
- Cross entropy loss: $f_t(\mathbf{w}; \{\mathbf{x}_t, y_t\}) = -\log \mathrm{softmax}(\langle \mathbf{w}_{y_t}, \mathbf{x}_t \rangle)$.

Note that we consider a weight vector or vectors as model parameters without loss of generality because the bias term can be viewed as additional dimension in the weight vector space where all examples share a constant feature value (e.g. 1).

In practice, to reduce computation cost and storage of the loss function, we often consider the linearized loss function

$$\hat{f}_t(\mathbf{w}) = f_t(\mathbf{w}) + \langle \mathbf{g}_t, \mathbf{w} - \mathbf{w}_t \rangle,\ \mathbf{g}_t \in \partial f_t(\mathbf{w}_t).$$

It is well known that the regret bound w.r.t. f can be bounded by its linearized lower bound. From complexity we have

$$\hat{f}_t(\mathbf{w}_t) = f_t(\mathbf{w}_t), \ \hat{f}_t(\mathbf{u}) \leqslant f_t(\mathbf{u}),$$

therefore

$$\text{Regret}(\mathbf{u}; f) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \leqslant \sum_{t=1}^{T} \hat{f}_t(\mathbf{w}_t) - \hat{f}_t(\mathbf{u}) = \text{Regret}\left(\mathbf{u}; \hat{f}\right).$$

We can also drop the constant and only use the inner product of weight vector and example. By complexity of f,

$$f_t(\mathbf{u}) \geqslant f_t(\mathbf{w}_t) + \langle \mathbf{g}_t, \mathbf{u} - \mathbf{w}_t \rangle, \ \mathbf{g}_t \in \partial f_t(\mathbf{w}_t),$$

we have

$$f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \leqslant \langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{u} \rangle = \langle \mathbf{g}_t, \mathbf{w}_t \rangle - \langle \mathbf{g}_t, \mathbf{u} \rangle.$$

Define $g_t(\mathbf{w}) \triangleq \langle \mathbf{g}_t, \mathbf{w} \rangle$, we have

$$\text{Regret}(\mathbf{u}; f) = \sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \leqslant \sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{w}_t \rangle - \langle \mathbf{g}_t, \mathbf{u} \rangle = \text{Regret}(\mathbf{u}; g) = \text{Regret}\left(\mathbf{u}; \hat{f}\right).$$

Throughout the chapter we will use the notation $f_{1:t}(\mathbf{w}) \triangleq \sum_{s=1}^{t} f_s(\mathbf{w})$, and

$$\begin{aligned} h_0(\mathbf{w}) &= r_0(\mathbf{w}) \\ h_t(\mathbf{w}) &= f_t(\mathbf{w}) + r_t(\mathbf{w}) \quad \text{for } t = 1, 2, \ldots \end{aligned}$$

we see that the updating optimization problem for general FTRL is

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} h_{0:t}(\mathbf{w}).$$

In practice, f_t are convex and $r_t \geqslant 0$ are chosen so that $r_{0:t}$ is strongly convex for all t, e.g., $r_{0:t}(\mathbf{w}) = \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2$. Also we derive the dual norm of $\sigma\|\mathbf{x}\|$

$$(\sigma\|\mathbf{x}\|)_* = \sup_{\mathbf{y}:\sigma\|\mathbf{y}\| \leqslant 1} \langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{\sigma} \sup_{\|\sigma\mathbf{y}\| \leqslant 1} \langle \mathbf{x}, \sigma\mathbf{y} \rangle = \frac{1}{\sigma}\|\mathbf{x}\|_*.$$

We will also use the following inequality

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

We prove the following lemma for the rationale of lazy projection.

Lemma 1 *The following two optimization problems are equivalent:*

$$\begin{cases} \mathbf{u}_{t+1} = \arg\min_{\mathbf{w}\in\Re^n} \mathbf{g}_{1:t} \cdot \mathbf{w} + \frac{1}{2\eta}\|\mathbf{w}\|_2^2 \\ \mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\chi} \|\mathbf{w} - \mathbf{u}_{t+1}\|_2^2 \end{cases} \Leftrightarrow \mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\chi} \mathbf{g}_{1:t} \cdot \mathbf{w} + \frac{1}{2\eta}\|\mathbf{w}\|^2 \tag{1}$$

Proof For the first two-stage optimization problem we have

$$\mathbf{u}_{t+1} = \arg\min_{\mathbf{w}\in\Re^n} \mathbf{g}_{1:t}\cdot\mathbf{w} + \frac{1}{2\eta}\|\mathbf{w}\|_2^2 \Leftrightarrow \mathbf{g}_{1:t} + \frac{1}{\eta}\mathbf{u}_{t+1} = \mathbf{0},$$

and

$$\begin{aligned}
&\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\chi}\|\mathbf{w}-\mathbf{u}_{t+1}\|_2^2 = \arg\min \frac{1}{2}\|\mathbf{w}-\mathbf{u}_{t+1}\|_2^2 + I_\chi(\mathbf{w})\\
&\Leftrightarrow -\left(\mathbf{w}_{t+1}-\mathbf{u}_{t+1}\right)\in\partial I_\chi(\mathbf{w}_{t+1})\\
&\Leftrightarrow -\left(\frac{1}{\eta}\mathbf{w}_{t+1}-\frac{1}{\eta}\mathbf{u}_{t+1}\right)\in\partial I_\chi(\mathbf{w}_{t+1})\\
&\Leftrightarrow -\left(\frac{1}{\eta}\mathbf{w}_{t+1}+\mathbf{g}_{1:t}\right)\in\partial I_\chi(\mathbf{w}_{t+1})
\end{aligned}$$

For the other optimization problem,

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\chi}\mathbf{g}_{1:t}\cdot\mathbf{w} + \frac{1}{2\eta}\|\mathbf{w}\|^2 = \arg\min \mathbf{g}_{1:t}\cdot\mathbf{w} + \frac{1}{2\eta}\|\mathbf{w}\|^2 + I_\chi(\mathbf{w})$$

is equivalent to

$$-\left(\mathbf{g}_{1:t}+\frac{1}{\eta}\mathbf{w}_{t+1}\right)\in\partial I_\chi(\mathbf{w}_{t+1}).$$

■

To understand $\partial I_\chi(\mathbf{w})$, since χ is a convex set, by complexity we have

$$I_\chi(\mathbf{w}) \geqslant I_\chi(\mathbf{w}_0) + \langle\mathbf{g},\mathbf{w}-\mathbf{w}_0\rangle,\ \mathbf{w}_0\in\chi,\ \forall\mathbf{w}\in\chi,$$

where $\mathbf{g}\in\partial I_\chi(\mathbf{w}_0)$. We therefore have

$$0\geqslant 0+\langle\mathbf{g},\mathbf{w}-\mathbf{w}_0\rangle \Leftrightarrow \langle\mathbf{g},\mathbf{w}-\mathbf{w}_0\rangle\leqslant 0.$$

$$\begin{aligned}
&\therefore\partial I_\chi(\mathbf{w}_0) = \{\mathbf{g}|\langle\mathbf{g},\mathbf{w}-\mathbf{w}_0\rangle\leqslant 0,\ \forall\mathbf{w}\in\chi\},\ \mathbf{w}_0\in\chi\\
&\therefore\partial I_\chi(\mathbf{w}_0) = \gamma\partial I_\chi(\mathbf{w}_0),\ \forall\gamma>0\\
&\mathbf{w}_0\in\mathrm{int}(\chi)\Rightarrow\exists\varepsilon>0,\mathbf{w}_0\pm\varepsilon\mathbf{w}_0\in\mathrm{int}(\chi)\Rightarrow\langle\mathbf{g},\pm\varepsilon\mathbf{w}_0\rangle\leqslant 0\Rightarrow\partial I_\chi(\mathbf{w}_0)=\mathbf{0}\\
&\mathbf{w}_0\in\partial\chi\Rightarrow\partial I_\chi(\mathbf{w}_0) = \{\mathbf{g}|\langle\mathbf{g},\mathbf{w}-\mathbf{w}_0\rangle\leqslant 0,\ \forall\mathbf{w}\in\chi\}.
\end{aligned}$$

2. Regret Bound for General FTRL and FTRL-Proximal

To compute regret bound for adaptive FTRL, the following three lemmas are very important.

Lemma 2 (Strong FTRL Lemma) *Let f_t be a sequence of arbitrary loss functions, and $r_t\geqslant 0$ such that $\mathbf{w}_{t+1}=\arg\min_{\mathbf{w}}h_{0:t}(\mathbf{w})$ is well defined, where $h_{0:t}(\mathbf{w})\triangleq f_{1:t}(\mathbf{w})+r_{0:t}(\mathbf{w})$. Then we have*

$$\textit{Regret}(\mathbf{u})\leqslant r_{0:T}(\mathbf{u})+\sum_{t=1}^{T}h_{0:t}(\mathbf{w}_t)-h_{0:t}(\mathbf{w}_{t+1})-r_t(\mathbf{w}_t).$$

Proof

$$\begin{aligned}
\sum_{t=1}^{T} h_t(\mathbf{w}_t) - h_{0:T}(\mathbf{u}) &= \sum_{t=1}^{T} h_{0:t}(\mathbf{w}_t) - h_{0:t-1}(\mathbf{w}_t) - h_{0:T}(\mathbf{u}) \\
&\leqslant \sum_{t=1}^{T} h_{0:t}(\mathbf{w}_t) - h_{0:t-1}(\mathbf{w}_t) - h_{0:T}(\mathbf{w}_{T+1}) \\
&= \sum_{t=1}^{T} h_{0:t}(\mathbf{w}_t) - \sum_{t=1}^{T} h_{0:t-1}(\mathbf{w}_t) - h_{0:T}(\mathbf{w}_{T+1}) \\
&= \sum_{t=1}^{T} h_{0:t}(\mathbf{w}_t) - \sum_{t=0}^{T-1} h_{0:t}(\mathbf{w}_{t+1}) - h_{0:T}(\mathbf{w}_{T+1}) \\
&= \sum_{t=1}^{T} h_{0:t}(\mathbf{w}_t) - \sum_{t=1}^{T} h_{0:t}(\mathbf{w}_{t+1}) - h_0(\mathbf{w}_1) \\
&= \sum_{t=1}^{T} h_{0:t}(\mathbf{w}_t) - \sum_{t=1}^{T} h_{0:t}(\mathbf{w}_{t+1}) - r_0(\mathbf{w}_1) \\
&\leqslant \sum_{t=1}^{T} h_{0:t}(\mathbf{w}_t) - h_{0:t}(\mathbf{w}_{t+1}).
\end{aligned}$$

$$\therefore \sum_{t=1}^{T} f_t(\mathbf{w}_t) + r_t(\mathbf{w}_t) - f_{1:T}(\mathbf{u}) - r_{0:T}(\mathbf{u}) \leqslant \sum_{t=1}^{T} h_{0:t}(\mathbf{w}_t) - h_{0:t}(\mathbf{w}_{t+1}).$$

By rearranging we have

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_{1:T}(\mathbf{u}) \leqslant r_{0:T}(\mathbf{u}) + \sum_{t=1}^{T} h_{0:t}(\mathbf{w}_t) - h_{0:t}(\mathbf{w}_{t+1}) - r_t(\mathbf{w}_t).$$

■

Lemma 3 *Let $\phi_1 : \Re^n \to \Re \cup \{\infty\}$ be a convex function such that $\mathbf{w}_1 = \arg\min_{\mathbf{w}} \phi_1(\mathbf{w})$ exists. Let ψ be a convex function such that $\phi_2(\mathbf{w}) = \phi_1(\mathbf{w}) + \psi(\mathbf{w})$ is strongly convex w.r.t. norm $\|\cdot\|$. Let $\mathbf{w}_2 = \arg\min_{\mathbf{w}} \phi_2(\mathbf{w})$. Then for any $\mathbf{b} \in \partial\psi(\mathbf{w}_1)$, we have*

$$\|\mathbf{w}_1 - \mathbf{w}_2\| \leqslant \|\mathbf{b}\|_*,$$

and for any $\mathbf{w}'$,

$$\phi_2(\mathbf{w}_1) - \phi_2(\mathbf{w}') \leqslant \frac{1}{2}\|\mathbf{b}\|_*^2.$$

Lemma 4 *Let $\phi_1 : \Re^n \to \Re \cup \{\infty\}$ be a convex function such that $\mathbf{w}_1 = \arg\min_{\mathbf{w}} \phi_1(\mathbf{w})$ exists. Let ψ and Ψ be a convex functions such that $\phi_2(\mathbf{w}) = \phi_1(\mathbf{w}) + \psi(\mathbf{w}) + \Psi(\mathbf{w})$ is strongly convex w.r.t. $\|\cdot\|$. Let $\mathbf{w}_2 = \arg\min_{\mathbf{w}} \phi_2(\mathbf{w})$. Then for any $\mathbf{b} \in \partial\psi(\mathbf{w}_1)$ and any $\mathbf{w}'$, we have*

$$\phi_2(\mathbf{w}_1) - \phi_2(\mathbf{w}') \leqslant \frac{1}{2}\|\mathbf{b}\|_*^2 + \Psi(\mathbf{w}_1) - \Psi(\mathbf{w}_2).$$

Theorem 5 (General FTRL Bound including FTRL-Centered) *Suppose the r_t are chosen such that $h_{0:t} + f_{t+1} = r_{0:t} + f_{1:t+1}$ is 1-strongly-convex w.r.t. some norm $\|\cdot\|_{(t)}$. Then, choosing any $\mathbf{g}_t \in \partial f_t(\mathbf{w}_t)$ on each round, for any $\mathbf{u} \in \Re^n$ and for any $T > 0$,*

$$\textit{Regret}_T(\mathbf{u}) \leqslant r_{0:T-1}(\mathbf{u}) + \frac{1}{2}\sum_{t=1}^{T} \|\mathbf{g}_t\|^2_{(t-1),*}.$$

Proof To apply Lemma 3, take $\phi_1(\mathbf{w}) = h_{0:t-1}(\mathbf{w})$ and $\phi_2(\mathbf{w}) = h_{0:t-1}(\mathbf{w}) + f_t(\mathbf{w}) = h_{0:t}(\mathbf{w}) - r_t(\mathbf{w})$ so $\mathbf{w}_t = \arg\min_{\mathbf{w}} \phi_1(\mathbf{w})$. By assumption ϕ_2 is 1-strongly-convex w.r.t. $\|\cdot\|_{(t-1)}$. Applying Lemma 3 to ϕ_2 we have $\phi_2(\mathbf{w}_t) - \phi_2(\mathbf{w}_{t+1}) \leqslant \frac{1}{2}\|\mathbf{g}_t\|^2_{(t-1),*}$ for $\mathbf{g}_t \in \partial f_t(\mathbf{w}_t)$, and so

$$\begin{aligned} h_{0:t}(\mathbf{w}_t) - h_{0:t}(\mathbf{w}_{t+1}) - r_t(\mathbf{w}_t) &= \phi_2(\mathbf{w}_t) - \phi_2(\mathbf{w}_{t+1}) - r_t(\mathbf{w}_{t+1}) \\ &\leqslant \frac{1}{2}\|\mathbf{g}_t\|^2_{(t-1),*} - r_t(\mathbf{w}_{t+1}) \\ &\leqslant \frac{1}{2}\|\mathbf{g}_t\|^2_{(t-1),*}. \end{aligned}$$

Further, since r_T does not influence any of the points $\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_T$ selected by the algorithm, we can take $r_T(\mathbf{w}) = 0$ with loss of generality, and hence replace $r_{0:T}(\mathbf{u})$ with $r_{0:T-1}(\mathbf{u})$ in the final round. ∎

Theorem 6 (FTRL-Proximal Bound) *Suppose the r_t are chosen such that $h_{0:t} = r_{0:t} + f_{1:t}$ is 1-strongly-convex w.r.t. some norm $\|\cdot\|_{(t)}$ and further the r_t are proximal, that is $\mathbf{w}_t$ is a minimizer of r_t. Then, choosing any $\mathbf{g}_t \in \partial f_t(\mathbf{w}_t)$ on each round, for any $\mathbf{u} \in \Re^n$ and for any $T > 0$,*

$$\textit{Regret}_T(\mathbf{u}) \leqslant r_{0:T}(\mathbf{u}) + \frac{1}{2}\sum_{t=1}^{T} \|\mathbf{g}_t\|^2_{(t),*}.$$

Proof Take $\phi_1(\mathbf{w}) = f_{1:t-1}(\mathbf{w}) + r_{0:t}(\mathbf{w}) = h_{0:t}(\mathbf{w}) - f_t(\mathbf{w})$ and $\phi_2(\mathbf{w}) = h_{0:t}(\mathbf{w}) = \phi_1(\mathbf{w}) + f_t(\mathbf{w})$, since r_t is proximal, we have $\mathbf{w}_t = \arg\min_{\mathbf{w}} \phi_1(\mathbf{w})$, and $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \phi_1(\mathbf{w}) + f_t(\mathbf{w})$. Since ϕ_2 is 1-strongly-convex w.r.t. $\|\cdot\|_{(t)}$, by applying Lemma 3 we have

$$\phi_2(\mathbf{w}_t) - \phi_2(\mathbf{w}_{t+1}) \leqslant \frac{1}{2}\|\mathbf{g}_t\|^2_{(t),*} \text{ for } \mathbf{g}_t \in \partial f_t(\mathbf{w}_t),$$

therefore

$$\begin{aligned} h_{0:t}(\mathbf{w}_t) - h_{0:t}(\mathbf{w}_{t+1}) - r_t(\mathbf{w}_t) &= \phi_2(\mathbf{w}_t) - \phi_2(\mathbf{w}_{t+1}) - r_t(\mathbf{w}_{t+1}) \\ &\leqslant \frac{1}{2}\|\mathbf{g}_t\|^2_{(t),*} - r_t(\mathbf{w}_{t+1}) \\ &\leqslant \frac{1}{2}\|\mathbf{g}_t\|^2_{(t),*}. \end{aligned}$$

∎

3. Additional Regularization Terms and Composite Objectives

In this section, we consider generalized FTRL algorithms where an additional regularization term $\alpha_t \Psi(\mathbf{w})$ is added on each round, where Ψ is a non-negative convex function and the weights $\alpha_t > 0$ for $t \geq 1$ are non-increasing in t. We further assume Ψ and r_0 are both minimized at $\mathbf{w}_1$ and $\Psi(\mathbf{w}_1) = 0$. We generalize our definition of h_t to

$$\begin{aligned} h_0(\mathbf{w}) &= r_0(\mathbf{w}) \\ h_t(\mathbf{w}) &= f_t(\mathbf{w}) + \alpha_t \Psi(\mathbf{w}) + r_t(\mathbf{w}), \end{aligned}$$

so the FTRL update is

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} h_{0:t}(\mathbf{w}) = \arg\min_{\mathbf{w}} f_{1:t}(\mathbf{w}) + \alpha_{1:t} \Psi(\mathbf{w}) + r_{0:t}(\mathbf{w}). \tag{2}$$

Note that we use the linearization of the loss function f_t here. The regret w.r.t. f_t is bounded by that of the FTRL update for linearized loss function.

Theorem 7 (FTRL-Proximal Bounds for Additional Regularization Terms) *Let Ψ be a non-negative convex function minimized at $\mathbf{w}_1$ with $\Psi(\mathbf{w}_1)$. Let $\alpha_t \geq 0$ be a non-increasing sequence of constants. Define h_t as in Eq. (2). Suppose the r_t are chosen such that $h_{0:t}$ is 1-strongly-convex w.r.t. some norm $\|\cdot\|_{(t)}$, and further r_t are proximal. We have*

$$\textit{Regret}(\mathbf{u}, f) \leqslant \textit{Regret}(\mathbf{u}, \mathbf{g}_t) \leqslant r_{0:T}(\mathbf{u}) + \alpha_{1:t} \Psi(\mathbf{u}) + \frac{1}{2} \sum_{t=1}^{T} \|\mathbf{g}_t\|_{(t),*}^2.$$

Proof Take $\phi_1(\mathbf{w}) = f_{1:t-1}(\mathbf{w}) + r_{0:t}(\mathbf{w}) = h_{0:t-1}(\mathbf{w}) + r_t(\mathbf{w})$ and $\phi_2(\mathbf{w}) = h_{0:t}(\mathbf{w}) = \phi_1(\mathbf{w}) + f_t(\mathbf{w}) + \alpha_t \Psi(\mathbf{w})$, since r_t is proximal, we have $\mathbf{w}_t = \arg\min_{\mathbf{w}} \phi_1(\mathbf{w})$, and $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \phi_1(\mathbf{w}) + f_t(\mathbf{w})$. Since ϕ_2 is 1-strongly-convex w.r.t. $\|\cdot\|_{(t)}$, by applying Lemma 4 we have

$$\phi_2(\mathbf{w}_t) - \phi_2(\mathbf{w}_{t+1}) \leqslant \frac{1}{2} \|\mathbf{g}_t\|_{(t),*}^2 + \alpha_t \Psi(\mathbf{w}_t) - \alpha_t \Psi(\mathbf{w}_{t+1}) \text{ for } \mathbf{g}_t \in \partial f_t(\mathbf{w}_t),$$

therefore

$$\begin{aligned} h_{0:t}(\mathbf{w}_t) - h_{0:t}(\mathbf{w}_{t+1}) - r_t(\mathbf{w}_t) &= \phi_2(\mathbf{w}_t) - \phi_2(\mathbf{w}_{t+1}) - r_t(\mathbf{w}_{t+1}) \\ &\leqslant \phi_2(\mathbf{w}_t) - \phi_2(\mathbf{w}_{t+1}) \\ &\leqslant \frac{1}{2} \|\mathbf{g}_t\|_{(t),*}^2 + \alpha_t \Psi(\mathbf{w}_t) - \alpha_t \Psi(\mathbf{w}_{t+1}). \end{aligned}$$

Considering only the Ψ terms, we have

$$\sum_{t=1}^{T} \alpha_t \Psi(\mathbf{w}_t) - \alpha_t \Psi(\mathbf{w}_{t+1}) = \alpha_1 \Psi(\mathbf{w}_1) - \alpha_T \Psi(\mathbf{w}_{T+1}) + \sum_{t=2}^{T} \alpha_t \Psi(\mathbf{w}_t) - \alpha_{t-1} \Psi(\mathbf{w}_t) \leqslant 0.$$

Thus

$$\sum_{t=1}^{T} h_{0:t}(\mathbf{w}_t) - h_{0:t}(\mathbf{w}_{t+1}) - r_t(\mathbf{w}_t) \leqslant \frac{1}{2} \sum_{t=1}^{T} \|\mathbf{g}_t\|_{(t),*}^2.$$

By applying the strong FTRL lemma, we have

$$\begin{aligned}
\mathrm{Regret}\left(\mathbf{u}, f_t\right) - \sum_{t=1}^{T} \alpha_t \Psi\left(\mathbf{u}\right) &\leqslant \mathrm{Regret}\left(\mathbf{u}, f_t\right) + \sum_{t=1}^{T} \alpha_t \Psi\left(\mathbf{w}_t\right) - \alpha_t \Psi\left(\mathbf{u}\right) \\
&= \mathrm{Regret}\left(\mathbf{u}, f_t + \alpha_t \Psi\right) \\
&\leqslant r_{0:T}\left(\mathbf{u}\right) + \sum_{t=1}^{T} h_{0:t}\left(\mathbf{w}_t\right) - h_{0:t}\left(\mathbf{w}_{t+1}\right) - r_t\left(\mathbf{w}_t\right) \\
&\leqslant r_{0:T}\left(\mathbf{u}\right) + \frac{1}{2} \sum_{t=1}^{T} \left\|\mathbf{g}_t\right\|_{(t),*}^2 .
\end{aligned}$$

By rearranging, we have

$$\mathrm{Regret}\left(\mathbf{u}, f_t\right) \leqslant r_{0:T}\left(\mathbf{u}\right) + \alpha_{1:T} \Psi\left(\mathbf{u}\right) + \frac{1}{2} \sum_{t=1}^{T} \left\|\mathbf{g}_t\right\|_{(t),*}^2 .$$

■

For general FTRL including FTRL-Centered algorithms, Theorem 5 immediately gives a regret bound if we add $\alpha_t \Psi$ to r_t on each round:

$$\mathrm{Regret}\left(\mathbf{u}, f_t\right) \leqslant r_{0:T-1}\left(\mathbf{u}\right) + \alpha_{1:T-1} \Psi\left(\mathbf{u}\right) + \frac{1}{2} \sum_{t=1}^{T} \left\|\mathbf{g}_t\right\|_{(t-1),*}^2 .$$

4. Regularized Dual Averaging

The regularized dual averaging (RDA) method is shown in the following algorithm:

Regularized Dual Averaging (RDA):

Input:

$h(\mathbf{w})$ is 1-strongly-convex w.r.t. $\|\cdot\|$

$\{\beta_t\}$ is a nonnegative and nondecreasing sequence

$\Psi(\mathbf{w})$ is convex and $\min_{\mathbf{w}} \Psi(\mathbf{w}) = 0$

$\mathbf{w}_1 \leftarrow \arg\min_{\mathbf{w}} h(\mathbf{w}) \in \operatorname{Argmin}_{\mathbf{w}} \Psi(\mathbf{w})$

$\mathbf{z}_0 \leftarrow \mathbf{0}$

For $t \leftarrow 1, 2, \ldots$

Observe a loss function f_t, compute $\mathbf{g}_t \in \partial f_t(\mathbf{w}_t)$

Update the average subgradient $\mathbf{z}_t$:

$$\mathbf{z}_t \leftarrow \frac{t-1}{t}\mathbf{z}_{t-1} + \frac{1}{t}\mathbf{g}_t$$

Compute the next iterate $\mathbf{w}_{t+1}$:

$$\mathbf{w}_{t+1} \leftarrow \arg\min_{\mathbf{w}} \langle \mathbf{z}_t, \mathbf{w}\rangle + \Psi(\mathbf{w}) + \frac{\beta_t}{t} h(\mathbf{w})$$

EndFor

We now show that RDA belongs to the general adaptive FTRL family. we first discuss the relation between the average subgradient $\mathbf{z}_t$ and the accumulated subgradient $\mathbf{g}_{1:t}$.

$$\begin{aligned}
\mathbf{z}_t = \frac{t-1}{t}\mathbf{z}_{t-1} + \frac{1}{t}\mathbf{g}_t &\Leftrightarrow t\mathbf{z}_t = (1-t)\,\mathbf{z}_{t-1} + \mathbf{g}_t \\
&\Leftrightarrow t\mathbf{z}_t = \sum\nolimits_{s=1}^{t} \mathbf{g}_s = \mathbf{g}_{1:t} \\
&\Leftrightarrow \mathbf{z}_t = \frac{1}{t}\mathbf{g}_{1:t}.
\end{aligned}$$

If we define

$$r_{0:t}(\mathbf{w}) = \beta_t h(\mathbf{w}) \text{ which is 1-strongly-convex w.r.t. } \sqrt{\beta_t}\,\|\cdot\|,$$

and

$$\alpha_t = 1 \text{ which is non-increasing in } t,$$

we have

$$\begin{aligned}
\mathbf{w}_{t+1} &= \arg\min_{\mathbf{w}} \langle \mathbf{z}_t, \mathbf{w}\rangle + \Psi(\mathbf{w}) + \frac{\beta_t}{t} h(\mathbf{w}) \\
&= \arg\min_{\mathbf{w}} \langle t\mathbf{z}_t, \mathbf{w}\rangle + t\Psi(\mathbf{w}) + \beta_t h(\mathbf{w}) \\
&= \arg\min_{\mathbf{w}} \langle \mathbf{g}_{1:t}, \mathbf{w}\rangle + \alpha_{1:t}\Psi(\mathbf{w}) + r_{0:t}(\mathbf{w}).
\end{aligned}$$

Therefore a regret bound can be derived for RDA following the regret bound for general FTRL with additional regularization terms. It's apparent that dual averaging is RDA with $\Psi(\mathbf{w}) \equiv 0$.

5. Special Cases

Now we set $r_0(\mathbf{w}) = I_\chi(\mathbf{w}),\ \chi = \{\mathbf{w} | \|\mathbf{w}\|_2 \leqslant R\}$, and we will show some special adaptive online learning methods.

The adaptive online gradient descent (OGD-Adaptive) has the following update:

$$\bar{\mathbf{w}}_{t+1} = \arg\min_{\mathbf{w}} \left(\mathbf{g}_t \cdot \mathbf{w} + \frac{1}{2\eta_t} \|\mathbf{w} - \bar{\mathbf{w}}_t\|_2^2 \right).$$

For FTRL-Proximal, we set $r_t(\mathbf{w}) = \frac{\sigma_t}{2} \|\mathbf{w} - \mathbf{w}_t\|_2^2$, giving

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \left(\mathbf{g}_{1:t} \cdot \mathbf{w} + \frac{1}{2} \sum_{s=1}^{t} \sigma_s \|\mathbf{w} - \mathbf{w}_s\|_2^2 \right).$$

Lemma 8 *Adaptive online gradient descent is equivalent to FTRL-Proximal.*

Proof For OGD-Adaptive, we have

$$\begin{aligned}
&\bar{\mathbf{w}}_{t+1} = \arg\min_{\mathbf{w}} \left(\mathbf{g}_t \cdot \mathbf{w} + \frac{1}{2\eta_t} \|\mathbf{w} - \bar{\mathbf{w}}_t\|_2^2 \right) \\
\Leftrightarrow\ & \mathbf{g}_t + \frac{1}{\eta_t} (\bar{\mathbf{w}}_{t+1} - \bar{\mathbf{w}}_t) = \mathbf{0} \\
\Leftrightarrow\ & \bar{\mathbf{w}}_{t+1} = \bar{\mathbf{w}}_t - \eta_t \mathbf{g}_t.
\end{aligned}$$

For FTRL-Proximal, we have

$$\begin{aligned}
&\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \left(\mathbf{g}_{1:t} \cdot \mathbf{w} + \frac{1}{2} \sum_{s=1}^{t} \sigma_s \|\mathbf{w} - \mathbf{w}_s\|_2^2 \right) \\
\Leftrightarrow\ & \begin{cases} \mathbf{g}_{1:t} + \sum_{s=1}^{t} \sigma_s (\mathbf{w}_{t+1} - \mathbf{w}_s) = \mathbf{0} \\ \mathbf{g}_{1:t-1} + \sum_{s=1}^{t-1} \sigma_s (\mathbf{w}_t - \mathbf{w}_s) = \mathbf{0} \end{cases} \\
\Leftrightarrow\ & \mathbf{g}_t + \sigma_t (\mathbf{w}_{t+1} - \mathbf{w}_t) + \sum_{s=1}^{t-1} \sigma_s (\mathbf{w}_{t+1} - \mathbf{w}_t) = \mathbf{0} \\
\Leftrightarrow\ & \mathbf{g}_t + \sum_{s=1}^{t} \sigma_s (\mathbf{w}_{t+1} - \mathbf{w}_t) = \mathbf{0} \\
\Leftrightarrow\ & \mathbf{w}_{t+1} - \mathbf{w}_t = -\frac{1}{\sum_{s=1}^{t} \sigma_s} \mathbf{g}_t = -\frac{1}{\sigma_{1:t}} \mathbf{g}_t.
\end{aligned}$$

$$\therefore \eta_t^{\text{FTRL-Proximal}} = \frac{1}{\sigma_{1:t}} = \eta_t^{\text{OGD-Adaptive}}.$$

■

We see that

$$r_t(\mathbf{w}) = \frac{\sigma_t}{2} \|\mathbf{w} - \mathbf{w}_t\|_2^2 = \frac{\sigma_{1:t} - \sigma_{1:t-1}}{2} \|\mathbf{w} - \mathbf{w}_t\|_2^2 = \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}}\right) \|\mathbf{w} - \mathbf{w}_t\|_2^2,$$

so

$$\sigma_t = \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}.$$

If we set $\beta_t = \sigma_{1:t}$, $h(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$, or equivalently $r_t(\mathbf{w}) = \frac{1}{2}\sigma_t \|\mathbf{w}\|_2^2$, we have dual averaging update

$$\begin{aligned}
\mathbf{w}_{t+1} &= \arg\min_{\mathbf{w}} \left(\mathbf{g}_{1:t} \cdot \mathbf{w} + \frac{1}{2}\sum_{s=1}^{t} \sigma_s \|\mathbf{w}\|_2^2\right) \\
&= \arg\min_{\mathbf{w}} \left(\mathbf{g}_{1:t} \cdot \mathbf{w} + \frac{1}{2}\sigma_{1:t} \|\mathbf{w}\|_2^2\right) \\
&= -\eta_t \mathbf{g}_{1:t} \\
&= -\eta_t \mathbf{g}_{1:t-1} - \eta_t \mathbf{g}_t \\
&= \frac{\eta_t}{\eta_{t-1}}\left(-\eta_{t-1}\mathbf{g}_{1:t-1}\right) - \eta_t \mathbf{g}_t \\
&= \frac{\eta_t}{\eta_{t-1}}\mathbf{w}_t - \eta_t \mathbf{g}_t.
\end{aligned}$$

We now discuss AdaGrad FTRL-Proximal algorithm. For a one-dimensional problem, we use $r_0 = I_\chi$ with $\chi = [-R, R]$ and $r_t(w) = \frac{1}{2}\sigma_t \|w - w_t\|_2^2$, the learning rate schedule is

$$\eta_t = \frac{\sqrt{2}R}{\sqrt{\sum_{s=1}^{t} \|g_s\|_2^2}}.$$

We thus have $r_{0:t}(w) = \frac{1}{2}\sum_{s=1}^{t} \sigma_s \|w - w_s\|_2^2$ which implies that $h_{0:t}(w) = g_{1:t} \cdot w + r_{0:t}(w)$ is 1-strongly-convex with $\sqrt{\sigma_{1:s}}\|\cdot\|_2$. Therefore $\|g_t\|_{(t),*}^2 = \frac{1}{\sigma_{1:t}}\|g_t\|_2^2 = \eta_t \|g_t\|_2^2$. Now we derive its regret bound.

$$\begin{aligned}
\text{Regret}\left(\mathbf{u}\right) &\leqslant r_{0:T}\left(\mathbf{u}\right) + \frac{1}{2}\sum_{t=1}^{T}\|g_t\|_{(t),*}^2 \\
&\leqslant \frac{1}{2}\sum_{t=1}^{T}\sigma_t(2R)^2 + \frac{\sqrt{2}R}{2}\sum_{t=1}^{T}\frac{\|g_t\|_2^2}{\sqrt{\sum_{s=1}^{t}\|g_s\|_2^2}} \\
&\leqslant \frac{2R^2}{\eta_T} + \sqrt{2}R\sqrt{\sum_{t=1}^{T}\|g_t\|_2^2} \\
&= \frac{2R^2\sqrt{\sum\limits_{t=1}^{T}\|g_t\|_2^2}}{\sqrt{2}R} + \sqrt{2}R\sqrt{\sum_{t=1}^{T}\|g_t\|_2^2} \\
&= 2\sqrt{2}R\sqrt{\sum_{t=1}^{T}\|g_t\|_2^2}
\end{aligned}$$

For a d-dimensional problem, we only need to apply the above technique on a per-coordinate basis, namely we set $\chi = [-R, R]^n$ and use the learning rate

$$\eta_{t,i} = \frac{\sqrt{2}R}{\sqrt{\sum_{s=1}^{t}\|\mathbf{g}_s[i]\|_2^2}}$$

for coordinate i. We thus have

$$r_t\left(\mathbf{w}\right) = \frac{1}{2}\left\|\mathbf{Q}_t^{\frac{1}{2}}\left(\mathbf{w} - \mathbf{w}_t\right)\right\|_2^2 = \frac{1}{2}\sum_{i=1}^{d}\sigma_{t,i}(\mathbf{w}[i] - \mathbf{w}_t[i])^2,$$

and

$$r_{0:t}\left(\mathbf{w}\right) = \frac{1}{2}\sum_{i=1}^{d}\sigma_{0:t,i}(\mathbf{w}[i] - \mathbf{w}_t[i])^2 = \frac{1}{2}\sum_{i=1}^{d}\frac{1}{\eta_{t,i}}(\mathbf{w}[i] - \mathbf{w}_t[i])^2.$$

Define $\mathbf{Q}_t = diag\left(\sigma_t\right),\ \sigma_t[i] = \frac{1}{\eta_{t,i}} - \frac{1}{\eta_{t-1,i}}$, thus $r_{0:t}$ is 1-strongly-convex w.r.t. $\left\|\mathbf{Q}_{1:t}^{\frac{1}{2}}\mathbf{w}\right\|_2$ whose dual norm can be derived to be

$$\left(\left\|\mathbf{Q}_{1:t}^{\frac{1}{2}}\mathbf{x}\right\|_2\right)_* = \sup_{\mathbf{y}:\left\|\mathbf{Q}_{1:t}^{1/2}\mathbf{y}\right\|_2\leqslant 1}\langle\mathbf{x},\mathbf{y}\rangle = \sup_{\left\|\mathbf{Q}_t^{1/2}\mathbf{y}\right\|_2\leqslant 1}\langle\mathbf{Q}_{1:t}^{-\frac{1}{2}}\mathbf{x}, \mathbf{Q}_{1:t}^{\frac{1}{2}}\mathbf{y}\rangle = \left\|\mathbf{Q}_{1:t}^{-\frac{1}{2}}\mathbf{x}\right\|_2.$$

To derive the regret bound, we first determine the upper bound for the regularization term,

$$
\begin{aligned}
r_{0:T}(\mathbf{u}) &= \frac{1}{2}\sum_{i=1}^{d}\frac{1}{\eta_{T,i}}(\mathbf{u}[i]-\mathbf{w}_T[i])^2 \\
&\leqslant \frac{1}{2}\sum_{i=1}^{d}\frac{1}{\eta_{T,i}}4R^2 \\
&= 2R^2\sum_{i=1}^{d}\frac{1}{\sqrt{2}R}\sqrt{\sum_{t=1}^{T}\mathbf{g}_t[i]^2} \\
&= \sqrt{2}R\sum_{i=1}^{d}\sqrt{\sum_{t=1}^{T}\mathbf{g}_t[i]^2}.
\end{aligned}
$$

We then determine the upper bound for the stability term,

$$
\begin{aligned}
\frac{1}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|_{(t),*}^2 &= \frac{1}{2}\sum_{t=1}^{T}\left\|\mathbf{Q}_{1:t}^{-\frac{1}{2}}\mathbf{g}_t\right\|_2^2 \\
&= \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{d}\left(\sigma_{1:t}^{-1/2}[i]\cdot\mathbf{g}_t[i]\right)^2 \\
&= \frac{1}{2}\sum_{i=1}^{d}\sum_{t=1}^{T}\eta_{t,i}\mathbf{g}_t[i]^2 \\
&= \frac{1}{2}\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\sqrt{2}R}{\sqrt{\sum_{s=1}^{t}\|\mathbf{g}_s[i]\|_2^2}}\mathbf{g}_t[i]^2 \\
&\leqslant \sqrt{2}R\sum_{i=1}^{d}\sqrt{\sum_{t=1}^{T}\mathbf{g}_t[i]^2}
\end{aligned}
$$

Therefore the regret bound for d-dimensional AdaGrad FTRL-Proximal is

$$
\text{Regret}(\mathbf{u}) \leqslant r_{0:T}(\mathbf{u}) + \frac{1}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|_{(t),*}^2 \leqslant 2\sqrt{2}R\sum_{i=1}^{d}\sqrt{\sum_{t=1}^{T}\mathbf{g}_t[i]^2}.
$$

AdaGrad can also be applied to dual averaging, but due to the ”off-by-one” difference in the bound, we use learning rate

$$
\eta_{t,i} = \frac{R}{\sqrt{G_i^2+\sum_{s=1}^{t}\mathbf{g}_s[i]^2}},\ \mathbf{g}_s[i] \leqslant G_i.
$$

References

H Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *arXiv preprint arXiv:1403.3465*, 2014.