# Chapter 7 Follow The Regularized Leader

**Algorithm 7.1** Follow-the-Regularized-Leader

**Require:** Closed and non-empty set $V \subseteq \mathbb{R}^d$, a sequence of regularizers $\psi_1, \dots, \psi_T : \mathbb{R}^d \to (-\infty, +\infty]$

1: **for** $t = 1$ **to** $T$ **do**
2: Output $\boldsymbol{x}_t \in \operatorname{argmin}_{\boldsymbol{x} \in V} \ \psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1} \ell_i(\boldsymbol{x})$
3: Receive $\ell_t : \mathbb{R}^d \to (-\infty, +\infty]$ and pay $\ell_t(\boldsymbol{x}_t)$
4: **end for**

## 7.2 FTRL Regret Bound using Strong Convexity

We consider a easy case to upper bound the regret for FTRL, where the losses plus regularizer are strongly convex

### 7.2.1 Properties of Strongly Convex Functions

We first give several properties of strongly convex functions.

$\mu$-strongly convex:

$$f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{\mu}{2}\|x - x_0\|^2 \le f(x), \ \forall x$$

**Lemma 7.5.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ $\mu$-strongly convex with respect to a norm $\|\cdot\|$. Then, for all $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$, $\boldsymbol{g} \in \partial f(\boldsymbol{y})$, and $\boldsymbol{g}' \in \partial f(\boldsymbol{x})$, we have*

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \le \langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{1}{2\mu}\|\boldsymbol{g} - \boldsymbol{g}'\|_*^2 .$$

$$f(x) \le f(y) + \langle g, x - y \rangle + \frac{1}{2\mu}\|g - g'\|_*^2$$

Proof: Define $\phi(z) = f(z) - \langle g, z \rangle$

$$\partial\phi(z) = \partial f(z) - g$$

Thus $\partial\phi(y) = \partial f(y) - g$. Thus $0 \in \partial\phi(y)$ ($g \in \partial f(y)$)

Based on Subgradient optimality condition

$$f(x^*) = \min_x f(x) \iff 0 \in \partial f(x^*)$$

Therefore, $y$ is the minimizer of $\phi(z)$ ①

What's more, $\partial \phi(x) = \partial f(x) - g$

And $g' \in \partial f(x)$

Thus, $g' - g \in \partial \phi(x)$ ②

Hence, we can write

$$\phi(y) = \min_{z \in \text{dom}\phi} \phi(z) \quad (\text{based on ①})$$

$$\geq \min_{z \in \text{dom}\phi} \left( \phi(x) + \langle g' - g, z - x \rangle + \frac{u}{2}\|z - x\|^2 \right) \quad (\text{strongly-convex})$$

$g' - g$: subgradient of $x$ (based on ②)

$$\geq \inf_{z \in R^d} \left( \phi(x) + \langle g' - g, z - x \rangle + \frac{u}{2}\|z - x\|^2 \right)$$

$$= \phi(x) - \frac{1}{2u}\|g' - g\|_*^2 \quad (\text{discussed below})$$

The last equality is because:

$f(x) = \frac{1}{2}\|x\|^2$ 's conjugate is $f^*(\theta) = \frac{1}{2}\|\theta\|_*^2$

Recall:

the dual norm $\|z\|_* = \sup_{\|x\| \le 1} z^T x$

the Fenchel conjugate $f^*(\theta) = \sup_{x \in R^d} \langle \theta, x \rangle - f(x)$

Thus,

$$\frac{1}{2}\|g' - g\|_*^2 = \sup_{z \in R^d} \langle g - g', z - x \rangle - \frac{1}{2}\|z - x\|^2$$

Multiply $-\frac{1}{u}$:

$$-\frac{1}{2u}\|g' - g\|_*^2 = \inf_{z \in R^d} \frac{1}{2u}\|z - x\|^2 + \frac{1}{u}\langle g' - g, z - x \rangle$$

$$= \inf_{z \in R^d} \langle g' - g, z - x \rangle + \frac{u}{2}\|z - x\|^2$$

This is where the last eqhality comes from.

**Corollary 7.6.** *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ $\mu$-strongly convex with respect to a norm $\|\cdot\|$. Let $\boldsymbol{x}^\star = \operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$. Then, for all $\boldsymbol{x} \in \operatorname{dom} f$, and $\boldsymbol{g} \in \partial f(\boldsymbol{x})$, we have*

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^\star) \le \frac{1}{2\mu}\|\boldsymbol{g}\|_*^2 .$$

In words, the above lemma says that an upper bound to the suboptimality gap is proportional to the squared norm of the subgradient.

# 7.2.2 An Explicit Regret Bound

Now, we state a Lemma quantifying the intuition on the "stability" of the predictions.

**Lemma 7.7.** *With the notation and assumptions of Lemma 7.1, assume that $F_t$ is proper and $\lambda_t$-strongly convex w.r.t. $\|\cdot\|$, and $\ell_t$ proper and convex. Also, assume that $\partial \ell_t(\boldsymbol{x}_t)$ is non-empty. Then, we have*

$$F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_{t+1}) + \ell_t(\boldsymbol{x}_t) \leq \frac{\|\boldsymbol{g}_t\|_*^2}{2\lambda_t} + \psi_t(\boldsymbol{x}_{t+1}) - \psi_{t+1}(\boldsymbol{x}_{t+1}),$$

*for all $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$.*

show the stability

**Proof:** First we review Lemma 7.1.

**Lemma 7.1.** *Let $V \subseteq \mathbb{R}^d$ be closed and non-empty. Denote by $F_t(\boldsymbol{x}) = \psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1} \ell_i(\boldsymbol{x})$. Assume that $\operatorname{argmin}_{\boldsymbol{x}\in V} F_t(\boldsymbol{x})$ is not empty and set $\boldsymbol{x}_t \in \operatorname{argmin}_{\boldsymbol{x}\in V} F_t(\boldsymbol{x})$. Then, for any $\boldsymbol{u}$, we have*

$$\sum_{t=1}^{T}(\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) = \psi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x}\in V} \psi_1(\boldsymbol{x}) + \sum_{t=1}^{T}[F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_{t+1}) + \ell_t(\boldsymbol{x}_t)] + F_{T+1}(\boldsymbol{x}_{T+1}) - F_{T+1}(\boldsymbol{u}) .$$

This providing an equality for the regret, which is suprising.

Then we prove Lemma 7.7.

$$F_t(x_t) - F_{t+1}(x_{t+1}) + l_t(x_t)$$

$$= (F_t(x_t) + l_t(x_t)) - (F_t(x_{t+1}) + l_t(x_{t+1})) + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1})$$

because of:

$$F_{t+1}(x) = \psi_{t+1}(x) + \sum_{i=1}^{t} l_i(x) = F_t(x) - \psi_t(x) + \psi_{t+1}(x) + l_t(x)$$

$$\left(F_t(x_t)+l_t(x_t)\right)-\left(F_t(x_{t+1})+l_t(x_{t+1})\right)+\psi_t(x_{t+1})-\psi_{t+1}(x_{t+1})$$

$$\le \left(F_t(x_t)+l_t(x_t)\right)-\left(F_t(x_t^*)+l_t(x_t^*)\right)+\psi_t(x_{t+1})-\psi_{t+1}(x_{t+1})$$

$$\le \frac{1}{\sum\lambda_t}\|g_t'\|_*^2+\psi_t(x_{t+1})-\psi_{t+1}(x_{t+1})$$ (By corollary 7.6)

**Corollary 7.6.** *Let $f:\mathbb{R}^d\to(-\infty,+\infty]$ $\mu$-strongly convex with respect to a norm $\|\cdot\|$. Let $\boldsymbol{x}^\star=\operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$. Then, for all $\boldsymbol{x}\in\operatorname{dom} f$, and $\boldsymbol{g}\in\partial f(\boldsymbol{x})$, we have*

$$f(\boldsymbol{x})-f(\boldsymbol{x}^\star)\le\frac{1}{2\mu}\|\boldsymbol{g}\|_*^2\,.$$

where $g_t'\in\partial\left(F_t(x_t)+l_t(x_t)\right)$ but we want find $g_t\in\partial l_t(x_t)$

What's more, $x_t=\operatorname{argmin}_{x\in V} F_t(x)$ (in our setting)

By Theorem 6.10:

**Theorem 6.10.** *Let $f:\mathbb{R}^d\to(-\infty,+\infty]$ proper. Then $\boldsymbol{x}^\star\in\operatorname{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$ iff $\mathbf{0}\in\partial f(\boldsymbol{x}^\star)$.*

*Proof.* We have that

$$\boldsymbol{x}^\star\in\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})\Leftrightarrow\forall\boldsymbol{y}\in\mathbb{R}^d,\ f(\boldsymbol{y})\ge f(\boldsymbol{x}^\star)=f(\boldsymbol{x}^\star)+\langle\mathbf{0},\boldsymbol{y}-\boldsymbol{x}^\star\rangle\Leftrightarrow\mathbf{0}\in\partial f(\boldsymbol{x}^\star)\,.$$ □

Thus. $0\in\partial\left(F_t(x_t)+i_V(x_t)\right)$

By Theorem 2-19, we can select $g_t'\in\partial l_t(x_t)$

**Theorem 2.19** ([Bauschke and Combettes, 2011, Theorem 18.5]). *Let $(f_i)_{i\in I}$ be a finite set of convex functions from $\mathbb{R}^d$ to $(-\infty,+\infty]$ and suppose $\boldsymbol{x}\in\bigcap_{i\in I}\operatorname{dom} f_i$ and $f_i$ continuous at $\boldsymbol{x}$. Set $F=\max_{i\in I} f_i$ and let $A(\boldsymbol{x})=\{i\in I|f_i(\boldsymbol{x})=F(\boldsymbol{x})\}$ the set of the active functions. Then*

$$\partial F(\boldsymbol{x})=\operatorname{conv}\bigcup_{i\in A(\boldsymbol{x})}\partial f_i(\boldsymbol{x})\,.$$

function composition's subgradient

Let's see some immediate applications of FTRL.

**Corollary 7.8.** *Let $\ell_t$ a sequence of convex loss functions. Let $\psi : V \to \mathbb{R}$ a $\mu$-strongly convex function w.r.t. $\|\cdot\|$. Set the sequence of regularizers as $\psi_t(\boldsymbol{x}) = \frac{1}{\eta_{t-1}}(\psi(\boldsymbol{x}) - \min_{\boldsymbol{z}} \psi(\boldsymbol{z}))$, where $\eta_{t+1} \leq \eta_t$, $t = 1, \dots, T$. Then, FTRL guarantees*

$$\sum_{t=1}^{T} \ell(\boldsymbol{x}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \leq \frac{\psi(\boldsymbol{u}) - \min_{\boldsymbol{x} \in V} \psi(\boldsymbol{x})}{\eta_{T-1}} + \frac{1}{2\mu} \sum_{t=1}^{T} \eta_{t-1} \|\boldsymbol{g}_t\|_\star^2,$$

*for all $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$. Moreover, if the functions $\ell_t$ are $L$-Lipschitz, setting $\eta_{t-1} = \frac{\alpha\sqrt{\mu}}{L\sqrt{t}}$ we get*

$$\sum_{t=1}^{T} \ell(\boldsymbol{x}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \leq \left( \frac{\psi(\boldsymbol{u}) - \min_{\boldsymbol{x}} \psi(\boldsymbol{x})}{\alpha} + \alpha \right) \frac{L\sqrt{T}}{\sqrt{\mu}} .$$

Proof: This corollary is immediate from Lemma 7.1, 7.7

**Lemma 7.1.** *Let $V \subseteq \mathbb{R}^d$ be closed and non-empty. Denote by $F_t(\boldsymbol{x}) = \psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1} \ell_i(\boldsymbol{x})$. Assume that $\operatorname{argmin}_{\boldsymbol{x} \in V} F_t(\boldsymbol{x})$ is not empty and set $\boldsymbol{x}_t \in \operatorname{argmin}_{\boldsymbol{x} \in V} F_t(\boldsymbol{x})$. Then, for any $\boldsymbol{u}$, we have*

$$\sum_{t=1}^{T} (\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{u})) = \psi_{T+1}(\boldsymbol{u}) - \min_{\boldsymbol{x} \in V} \psi_1(\boldsymbol{x}) + \sum_{t=1}^{T} [F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_{t+1}) + \ell_t(\boldsymbol{x}_t)] + F_{T+1}(\boldsymbol{x}_{T+1}) - F_{T+1}(\boldsymbol{u}) .$$

$$\leq \psi_{T+1}(u) - \min_{x \in V} \psi_1(x) + F_{T+1}(x_{T+1}) - F_{T+1}(u) \leq 0 \text{ by}$$

set $\psi_T = \psi_{T+1}$

$$+ \sum_{t=1}^{T} \left[ \frac{\|g_t\|_*^2}{2\lambda_t} + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}) \right]$$

(by Lemma 7.7)

$\leq 0$ by assumption

**Lemma 7.7.** *With the notation and assumptions of Lemma 7.1 assume that $F_t$ is proper and $\lambda_t$-strongly convex w.r.t. $\|\cdot\|$, and $\ell_t$ proper and convex. Also, assume that $\partial \ell_t(\boldsymbol{x}_t)$ is non-empty. Then, we have*

$$F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_{t+1}) + \ell_t(\boldsymbol{x}_t) \leq \frac{\|\boldsymbol{g}_t\|_*^2}{2\lambda_t} + \psi_t(\boldsymbol{x}_{t+1}) - \psi_{t+1}(\boldsymbol{x}_{t+1}),$$

*for all $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$.*

$\leq$ the right-hand side.

# 7.3 FTRL with Linearized Losses

In linear case, the same bound with OMD

We consider the case in which the losses are linear,

i.e., $l_t(x) = \langle g_t, x \rangle$, $t = 1, \dots, T$, we have

$$x_{t+1} \in \operatorname*{argmin}_{x \in V} \psi_{t+1}(x) + \sum_{i=1}^{t} \langle g_i, x \rangle$$
$$= \operatorname*{argmax}_{x \in V} \langle -\sum_{i=1}^{t} g_i, x \rangle - \psi_{t+1}(x)$$

Denote by $\psi_{V,t}(x) = \psi_t(x) + i_V(x)$

Now, we assume $\psi_{V,t}$ to be proper, convex and closed, using the Theorem 5.5, we have that $x_{t+1} \in \partial \psi^*_{V,t+1}(-\sum_{i=1}^{t} g_i)$

**Theorem 5.5** ([Rockafellar, 1970, Corollary 23.5.1 and Theorem 23.5]). *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be convex, proper, and closed. Then*

1. $\boldsymbol{x} \in \partial f^\star(\boldsymbol{\theta})$ *iff* $\boldsymbol{\theta} \in \partial f(\boldsymbol{x})$.
2. $\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle - f(\boldsymbol{x})$ *achieves its supremum in* $\boldsymbol{x}$ *at* $\boldsymbol{x} = \boldsymbol{x}^\star$ *iff* $\boldsymbol{x}^\star \in \partial f^\star(\boldsymbol{\theta})$.

Moreover, if $\psi_{V,t+1}$ is strongly convex, we know that $\psi^*_{V,t+1}$ is differentiable and we get

$$x_{t+1} = \nabla \psi^*_{V,t+1}(-\sum_{i=1}^{t} g_i)$$

$f(\theta) = \sup_x (\theta x - f(x))$

In turn, the update can be written in the following way

$$\theta_{t+1} = \theta_t - g_t$$

$$x_{t+1} = \nabla \psi^*_{V,t+1}(\theta_{t+1})$$

As shown in Figure below.

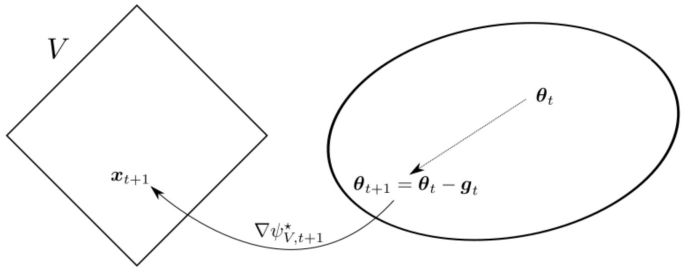


Figure 7.1: Dual mapping for FTRL with linear losses.

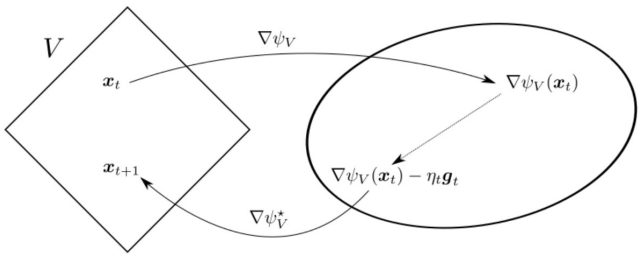


Figure 6.4: OMD update in terms of duality mappings.

Compare to Online Mirror Descent in a similar way:

$$\theta_{t+1} = \nabla\psi(x_t) - \eta_t g_t$$

$$x_{t+1} = \nabla\psi^*_V(\theta_{t+1})$$

Differences:

- In OMD, the state is kept in $x_t$, so we need to transform it into a dual variable before making the update and then back to the primal variable.
- In FTRL with linear losses, the state is kept directly in the dual space, updated and then transformed in the primal variable. The primal variable is only used to predict, but not directly in the update.
- In OMD, the samples are weighted by the learning rates that is typically decreasing $\downarrow \eta_t$
- In FTRL with linear losses, all the subgradients have the same weight, but the regularizer is typically increasing over time.

regularizer

**Algorithm 6.1** Online Mirror Descent

**Require:** Non-empty closed convex $V \subseteq X \subseteq \mathbb{R}^d$, $\psi : X \to \mathbb{R}$ strictly convex and continuously differentiable on $\operatorname{int} X$, $\boldsymbol{x}_1 \in V$ such that $\psi$ is differentiable in $\boldsymbol{x}_1$, $\eta_1, \cdots, \eta_T > 0$

1: **for** $t = 1$ **to** $T$ **do**
2: Output $\boldsymbol{x}_t$
3: Receive $\ell_t : \mathbb{R}^d \to (-\infty, +\infty]$ and pay $\ell_t(\boldsymbol{x}_t)$
4: Set $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$
5: $\boldsymbol{x}_{t+1} = \operatorname{argmin}_{\boldsymbol{x} \in V} \langle \boldsymbol{g}_t, \boldsymbol{x} \rangle + \frac{1}{\eta_t} B_\psi(\boldsymbol{x}; \boldsymbol{x}_t)$
6: **end for**

**Algorithm 7.2** Follow-the-Regularized-Leader on Linearized Losses

FRTL

**Require:** Convex, closed, and non-empty set $V \subseteq \mathbb{R}^d$, a sequence of regularizers $\psi_1, \ldots, \psi_T : \mathbb{R}^d \to (-\infty, +\infty]$

1: **for** $t = 1$ **to** $T$ **do**
2: Output $\boldsymbol{x}_t \in \operatorname{argmin}_{\boldsymbol{x} \in V} \psi_t(\boldsymbol{x}) + \sum_{i=1}^{t-1} \langle \boldsymbol{g}_i, \boldsymbol{x} \rangle$
3: Receive $\ell_t : \mathbb{R}^d \to (-\infty, +\infty]$ and pay $\ell_t(\boldsymbol{x}_t)$
4: Set $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$
5: **end for**

## 7.3.1 FTRL with Linearized Losses Can Be Equivalent to OMD

In certain cases, OMD and FTRL can be equivalent.

For example, consider that $V = X = \text{dom}\,\psi$

The output of OMD is:

$$x_{t+1} = \arg\min_{x} \langle \eta g_t, x \rangle + B_\psi(x; x_t)$$

Recall Bregman divergence:

$$B_\psi(x; y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle$$

We assume that $x_{t+1} \in \text{int dom}\,\psi$ for $t = 1, \cdots, T$

This implies that $\eta g_t + \nabla\psi(x_{t+1}) - \nabla\psi(x_t) = 0$

That is: $\nabla\psi(x_{t+1}) = \nabla\psi(x_t) - \eta g_t$

Assume $x_1 = \min_{x \in V} \psi(x)$, we have

? 

$$\nabla\psi(x_{t+1}) = -\eta \sum_{i=1}^{t} g_i$$

On the other hand, consider FTRL, with linearized losses with regularizers $\psi_t = \frac{1}{\eta}\psi$. then

$$x_{t+1} = \operatorname*{argmin}_x \frac{1}{\eta}\psi(x) + \sum_{i=1}^{t} \langle g_i, x \rangle$$

$$= \operatorname*{argmin}_x \psi(x) + \eta \sum_{i=1}^{t} \langle g_i, x \rangle$$

Assume $x_{t+1} \in \text{int dom}\psi$.

this implies that $\nabla\psi(x_{t+1}) = -\eta \sum_{i=1}^{t} g_i$

Thus the predictions of FTRL and OMD are the same.

This equivalence immediately gives us some intuition on the role of $\psi$ in both algorithm: The same function is inducing the Bregman divergence, that is our similarity measure, and is the regularizer in FTRL. Moreover, the inverse of the growth rate of the regularizers in FTRL takes the role of the learning rate in OMD.

**Example 7.10.** *Consider $\psi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$ and $V = \mathbb{R}^d$, then it satisfies the conditions above to have the predictions of OMD equal to the ones of FTRL.*