

# Online Learning and Online Convex Optimization

**Chapter 2 in Shai Shalev Shwartz / Online Learning and  
Online convex Optimization**

# Outline

## Online Convex Optimization

### Follow The Leader

- Quadratic Optimization

- Failure of Follow the Leader

### Follow The Regularized Leader

- FTRL for linear functions

- Online Gradient Descent

- Doubling Trick

- Strong Convexity

- General Theorem regarding FTRL with Strong Convexity

- Applications to expert advice

# Online Convex Optimization (OCO)

## Algorithm

**Input:** A convex set  $S$

**For**  $t = 1, 2, \dots$

- ▶ Predict a vector  $w_t \in S$
- ▶ Receive a convex loss function  $f_t : S \rightarrow \mathbb{R}$
- ▶ Suffer loss  $f_t(w_t)$

## Regret Definition

**Regret of the Algorithm:**

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

**Regret relative to a set of vectors  $U$ :**

$$\text{Regret}_T(U) = \max_{u \in U} \text{Regret}_T(u). \quad (2)$$

# Follow-the-Leader Algorithm

## FTL Strategy

At round  $t$ , select:

$$w_t = \operatorname{argmin}_{w \in S} \sum_{i=1}^{t-1} f_i(w)$$

- ▶ Natural approach: Choose best performer on past data
- ▶ Simple but can be unstable
- ▶ Requires solving optimization problem each round

## FTL Regret Analysis

Theorem (Lemma 2.1)

For any  $u \in S$ :

$$\text{Regret}_T(u) = \sum_{t=1}^T (f_t(w_t) - f_t(u)) \leq \sum_{t=1}^T (f_t(w_t) - f_t(w_{t+1})).$$

**proof**

**Step 1:** Equivalent to

$$\sum_{t=1}^T f_t(w_{t+1}) \leq \sum_{t=1}^T f_t(u)$$

**Step 2:** By induction on  $T$ :

- ▶ **Base case:**  $T = 1$  trivial as  $f_1(w_1) - f_1(u) \leq 0$
- ▶ **Inductive step:** Assume holds for  $T - 1$ , then

$$\begin{aligned}
 & \sum_{t=1}^T [f_t(w_t) - f_t(u)] \\
 &= \underbrace{\sum_{t=1}^{T-1} [f_t(w_t) - f_t(u)]}_{\leq \sum_{t=1}^{T-1} [f_t(w_t) - f_t(w_{t+1})]} + [f_T(w_T) - f_T(u)] \\
 &\leq \sum_{t=1}^T [f_t(w_t) - f_t(w_{t+1})]
 \end{aligned}$$

using  $w_{T+1} = \operatorname{argmin}_w \sum_{t=1}^T f_t(w)$

## FTL for Quadratic Optimization

For  $f_t(w) = \frac{1}{2} \|w - z_t\|_2^2$ :

- ▶ FTL update:  $w_t = \frac{1}{t-1} \sum_{i=1}^{t-1} z_i$
- ▶ Regret bound:  $O(\log T)$

Regret Calculation for quadratic optimization.

$$\begin{aligned} \text{Regret}_T(u) &\leq \sum_{t=1}^T \frac{1}{t} \|w_t - z_t\|^2 \\ &\leq \sum_{t=1}^T \frac{(2L)^2}{t} = 4L^2(\log T + 1) \end{aligned}$$

where  $L = \max_t \|z_t\|$





## Failure of follow the leader

$$f_t(w) = w \cdot z_t:$$



$$z_t = \begin{cases} -0.5 & \text{if } t = 1 \\ 1 & \text{if } t \text{ is even} \\ -1 & \text{if } t > 1 \text{ and } t \text{ is odd} \end{cases}$$

- ▶  $w_t = -1, 1, -1, 1, \dots$
- ▶ Cumulative loss is  $T$ .
- ▶ Cumulative loss of 0 is 0
- ▶ Regret is  $T$ .
- ▶ **Reason:** prediction is unstable
- ▶ We need to regularize.
- ▶  $R(W)$  penalizes vectors which are large.

## Follow-the-Regularized-Leader (FTRL)

$$\forall t, \quad w_t = \arg \min_{w \in S} \sum_{i=1}^{t-1} f_i(w) + R(w)$$

- ▶ For bad case above:  $w_t = 0, 0, 0, 0, \dots$
- ▶ Each step requires solving a minimization problem.

## Lemma 2.3: Follow-the-Regularized-Leader

**Lemma 2.3.** Let  $w_1, w_2, \dots$  be the sequence of vectors produced by FoReL. Then, for all  $u \in S$  we have:

$$\sum_{t=1}^T (f_t(w_t) - f_t(u)) \leq R(u) - R(w_1) + \sum_{t=1}^T (f_t(w_t) - f_t(w_{t+1})).$$

## Proof of Lemma 2.3

*Proof.* Observe that running FoReL on  $f_1, \dots, f_T$  is equivalent to running FTL on  $f_0, f_1, \dots, f_T$  where  $f_0 = R$ . Using Lemma 2.1, we obtain:

$$\sum_{t=0}^T (f_t(w_t) - f_t(u)) \leq \sum_{t=0}^T (f_t(w_t) - f_t(w_{t+1})).$$

Rearranging the above and using  $f_0 = R$ , we conclude our proof.



## FTRL Regret Bound for linear functions

**For linear**  $f_t(w) = \langle w, z_t \rangle$  and  $R(w) = \frac{1}{2\eta} \|w\|_2^2$

**Update rule**  $w_{t+1} = w_t - \eta z_t$

Theorem (Theorem 2.4)

$$\text{Regret}_T(U) \leq \frac{B^2}{2\eta} + \eta T L^2$$

Then, for all  $u$  we have

$$\text{Regret}_T(u) \leq \frac{1}{2\eta} \|u\|_2^2 + \eta \sum_{t=1}^T \|z_t\|_2^2.$$

## Choice of $\eta$ and Final Bound for linear functions

### Tunings:

- ▶ Define the set  $U = \{u : \|u\| \leq B\}$ .
- ▶ Assume that

$$\frac{1}{T} \sum_{t=1}^T \|z_t\|_2^2 \leq L^2.$$

- ▶ Set  $\eta = \frac{B}{L\sqrt{2T}}$ .

### Conclusion:

$$\text{Regret}_T(U) \leq BL\sqrt{2T}.$$

## From linear functions to Online Gradient Descent

### Example (OGD from FTRL)

Consider the OCO setup where the functions  $f_1, f_2, \dots$  are differentiable.

Let  $\eta$  be the learning rate.

$$w_{t+1} = w_t - \eta z_t, \quad z_t = \nabla f_t(w_t)$$

Identical to FTRL with regularization:  $R(w) = \frac{1}{2\eta} \|w\|_2^2$

**Regret bound on OGD:** From FTRL theorem:

$$\begin{aligned} \text{Regret} &\leq \frac{\|u\|^2}{2\eta} + \eta \sum_{t=1}^T \|z_t\|^2 \\ &\leq \frac{B^2}{2\eta} + \eta TL^2 \end{aligned}$$

## Regret Bound for OGD

If we further assume that each  $f_t$  is  $L_t$ -Lipschitz with respect to  $\|\cdot\|_2$ , and let  $L$  be such that

$$\frac{1}{T} \sum_{t=1}^T L_t^2 \leq L^2.$$

Then, for all  $u$ , the regret of OGD satisfies

$$\text{Regret}_T(u) \leq \frac{1}{2\eta} \|u\|_2^2 + \eta T L^2.$$



## Bounding the norm of $\mathbf{u}$

In particular, if  $U = \{\mathbf{u} : \|\mathbf{u}\|_2 \leq B\}$  and  $\eta = \frac{B}{L\sqrt{2T}}$  then

$$\text{Regret}_T(U) \leq BL\sqrt{2T}.$$

## Practical Considerations

### Doubling Trick

- ▶ Removes need to know time horizon  $T$
- ▶ Divide time into epochs  $2^m, 2^{m+1} - 1$
- ▶ Regret increases by constant factor:

$$\sum_{m=0}^{\log T} \sqrt{2^m} = O(\sqrt{T})$$

### Example (Optimal $\eta$ )

Setting  $\eta = \frac{B}{L} \sqrt{\frac{2}{T}}$  gives:

$$BL\sqrt{2T}$$

## Definition 2.4: Strong Convexity

### Strong Convexity

A function  $f : S \rightarrow \mathbb{R}$  is  $\sigma$ -strongly convex over  $S$  with respect to a norm  $\|\cdot\|$  if for any  $w \in S$  we have:

$$\forall z \in \partial f(w), \quad \forall u \in S, \quad f(u) \geq f(w) + \langle z, u - w \rangle + \frac{\sigma}{2} \|u - w\|^2.$$

## Lemma 2.8: Strong Convexity implication

### Lemma 2.8

Let  $S$  be a nonempty convex set. Let  $f : S \rightarrow \mathbb{R}$  be a  $\sigma$ -strongly convex function over  $S$  with respect to a norm  $\|\cdot\|$ . Let:

$$w = \arg \min_{v \in S} f(v).$$

Then, for all  $u \in S$ , we have:

$$f(u) - f(w) \geq \frac{\sigma}{2} \|u - w\|^2.$$

## Strong Convexity Condition

If  $R$  is twice differentiable, then it is easy to verify that a sufficient condition for strong convexity of  $R$  is that for all  $\mathbf{w}, \mathbf{x}$ ,

$$\langle \nabla^2 R(\mathbf{w}) \mathbf{x}, \mathbf{x} \rangle \geq \sigma \|\mathbf{x}\|^2,$$

where  $\nabla^2 R(\mathbf{w})$  is the Hessian matrix of  $R$  at  $\mathbf{w}$ , namely, the matrix of second-order partial derivatives of  $R$  at  $\mathbf{w}$  [39, Lemma 14].

## Example 2.4: Euclidean Regularization

The function

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

is 1-strongly-convex with respect to the  $\ell_2$  norm over  $\mathbb{R}^d$ . To see this, simply note that the Hessian of  $R$  at any  $w$  is the identity matrix.

## Example 2.5: Entropic Regularization

The function

$$R(\mathbf{w}) = \sum_{i=1}^d w[i] \log(w[i])$$

is  $\frac{1}{B}$ -strongly-convex with respect to the  $\ell_1$  norm over the set

$$S = \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w} > 0 \wedge \|\mathbf{w}\|_1 \leq B\}.$$

In particular,  $R$  is 1-strongly-convex over the probability simplex, which is the set of positive vectors whose elements sum to 1.

## Proof of strong convexity for Entropic Regularization

To see this, note that

$$\begin{aligned}\langle \nabla^2 R(w) x, x \rangle &= \sum_i \frac{x[i]^2}{w[i]} \\&= \frac{1}{\|w\|_1} \left( \sum_i w[i] \right) \left( \sum_i \frac{x[i]^2}{w[i]} \right) \\&\geq \frac{1}{\|w\|_1} \left( \sum_i \sqrt{w[i]} \frac{x[i]}{\sqrt{w[i]}} \right)^2 = \frac{\|x\|_1^2}{\|w\|_1},\end{aligned}$$

where the inequality follows from Cauchy–Schwarz inequality.



## Single Step of FTRL with Strong Convexity

Let

$$R : S \rightarrow \mathbb{R}$$

be a  $\sigma$ -strongly-convex function over  $S$  with respect to a norm  $\|\cdot\|$ . Let  $w_1, w_2, \dots$  be the predictions of the FoReL algorithm. Then, for all  $t$ , if  $f_t$  is  $L_t$ -Lipschitz with respect to  $\|\cdot\|$ , then:

$$f_t(w_t) - f_t(w_{t+1}) \leq L_t \|w_t - w_{t+1}\| \leq \frac{L_t^2}{\sigma}.$$

## Proof (Single Step of FTRL with Strong Convexity)

For all  $t$  let

$$F_t(w) = \sum_{i=1}^{t-1} f_i(w) + R(w)$$

and note that the FoReL rule is

$$w_t = \arg \min_{w \in S} F_t(w).$$

Note also that  $F_t$  is  $\sigma$ -strongly-convex since the addition of a convex function to a strongly convex function keeps the strong convexity property. Therefore, Lemma 2.8 implies that:

$$F_t(w_{t+1}) \geq F_t(w_t) + \frac{\sigma}{2} \|w_t - w_{t+1}\|^2.$$

## Continuing the Proof (Single Step of FTRL with Strong Convexity)

Repeating the same argument for  $F_{t+1}$  and its minimizer  $w_{t+1}$ , we get:

$$F_{t+1}(w_t) \geq F_{t+1}(w_{t+1}) + \frac{\sigma}{2} \|w_t - w_{t+1}\|^2.$$

Summing the above two inequalities and rearranging, we obtain:

$$\sigma \|w_t - w_{t+1}\|^2 \leq f_t(w_t) - f_t(w_{t+1}). \quad (2.7)$$

## Final Steps (Single Step of FTRL with Strong Convexity)

Next, using the Lipschitzness of  $f_t$ , we get that:

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \leq L_t \|\mathbf{w}_t - \mathbf{w}_{t+1}\|.$$

Combining with Equation (2.7) and rearranging, we get:

$$\|\mathbf{w}_t - \mathbf{w}_{t+1}\| \leq L/\sigma.$$

Together with the above, we conclude our proof. ■

## Combining the steps

Let  $f_1, \dots, f_T$  be a sequence of convex functions such that  $f_t$  is  $L_t$ -Lipschitz with respect to some norm  $\|\cdot\|$ . Let  $L$  be such that

$$\frac{1}{T} \sum_{t=1}^T L_t^2 \leq L^2.$$

Assume that FoReL is run on the sequence with a regularization function which is  $\sigma$ -strongly-convex with respect to the same norm. Then, for all  $u \in S$ ,

$$\text{Regret}_T(u) \leq R(u) - \min_{v \in S} R(v) + \frac{TL^2}{\sigma}.$$

## Corollary 2.12

Let  $f_1, \dots, f_T$  be a sequence of convex functions such that  $f_t$  is  $L_t$ -Lipschitz with respect to  $\|\cdot\|_2$ . Let  $L$  be such that

$$\frac{1}{T} \sum_{t=1}^T L_t^2 \leq L^2.$$

Assume that FoReL is run on the sequence with the regularization function

$$R(w) = \frac{1}{2\eta} \|w\|_2^2.$$

Then, for all  $u$ ,

$$\text{Regret}_T(u) \leq \frac{1}{2\eta} \|u\|_2^2 + \eta T L^2.$$

## Applications to expert advice

- ▶ Distribution  $w_t$
- ▶ Action Losses:  $x_t \in [0, 1]^d$
- ▶ Algorithm Loss:  $\langle x_t, w_t \rangle$
- ▶ We want to bound regret.
- ▶ we will compare  $l_2$  regularization with Entropic Regularization.

## $l_2$ regularization

Assume that the conditions of Corollary 2.12 hold. Let  $S$  be a convex set and consider running FoReL with the regularization function:

$$R(w) = \begin{cases} \frac{1}{2\eta} \|w\|_2^2 & \text{if } w \in S \\ \infty & \text{if } w \notin S \end{cases}$$

Then, for all  $u \in S$ ,

$$\text{Regret}_T(u) \leq \frac{1}{2\eta} \|u\|_2^2 + \eta TL^2.$$



## $l_2$ regularization for Experts

In particular, if

$$B \geq \max_{u \in S} \|u\|_2$$

and

$$B = 1; \quad L = \sqrt{d}; \quad \eta = \frac{B}{L\sqrt{2T}} = \frac{1}{\sqrt{2dT}}$$

then,

$$\text{Regret}_T(S) \leq \sqrt{2dT}.$$

## Entropic Regularization

Let  $f_1, \dots, f_T$  be a sequence of convex functions such that  $f_t$  is  $L_t$ -Lipschitz with respect to  $\|\cdot\|_1$ . Let  $L$  be such that  $\frac{1}{T} \sum_{t=1}^T L_t^2 \leq L^2$ . Assume that FoReL is run on the sequence with the regularization function

$$R(w) = \frac{1}{\eta} \sum_i w[i] \log(w[i])$$

and with the set

$$S = \{w : \|w\|_1 = B \wedge w > 0\} \subset \mathbb{R}^d.$$

Then,

$$\text{Regret}_T(S) \leq \frac{B \log(d)}{\eta} + \eta B T L^2.$$

In particular, setting  $\eta = \frac{\sqrt{\log d}}{L\sqrt{2T}}$  yields

$$\text{Regret}_T(S) \leq B L \sqrt{2 \log(d) T}.$$

## Entropic regularization for Experts

The Entropic regularization is strongly convex with respect to the  $\ell_1$  norm, and therefore the Lipschitzness requirement of the loss functions is also with respect to the  $\ell_1$ -norm.

For linear functions,

$$f_t(w) = \langle w, x_t \rangle,$$

we have by Hölder's inequality that,

$$|f_t(w) - f_t(u)| = |\langle w - u, x_t \rangle| \leq \|w - u\|_1 \|x_t\|_\infty.$$

Therefore, the Lipschitz parameter grows with the  $\ell_\infty$  norm of  $x_t$  rather than the  $\ell_2$  norm of  $x_t$ .

expert advice:  $B = 1$  and  $L = 1$ ), we obtain the regret bound of

$$\sqrt{2 \log(d) T}.$$

## Comparison between regularizations

- ▶ entropic regularization vs.  $\ell_2$  regularization.
- ▶  $\log d$  vs  $\sqrt{d}$
- ▶  $L$ :  $\|x_t\|_\infty \geq \|x_t\|_2$  Lipschitz condition is more stringent with entropic
- ▶  $B$ :  $\|u\|_1 \leq \|u\|_2$  Comparator length condition is more stringent with  $\ell_2$