Online Learning and Online Convex Optimization

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

Outline

Online Convex Optimization (OCO)

Algorithm

Input: A convex set S

For $t = 1, 2, \ldots$

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

Regret Definition

Regret of the Algorithm:

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

Regret relative to a set of vectors U:

$$\mathrm{Regret}_T(U) = \max_{u \in U} \mathrm{Regret}_T(u). \tag{2}$$

Follow-the-Leader Algorithm

FTL Strategy

At round t, select:

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

- 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 $u \in S$:

$$Regret_T(\mathrm{u}) = \sum_{t=1}^{T} (f_t(\mathrm{w}_t) - f_t(\mathrm{u})) \leq \sum_{t=1}^{T} (f_t(\mathrm{w}_t) - f_t(\mathrm{w}_{t+1})) \,.$$

proof

Step 1: Equivalent to

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

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

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

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

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

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

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

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

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

Lemma 2.3: Follow-the-Regularized-Leader

Lemma 2.3. Let $\mathrm{w}_1, \mathrm{w}_2, \ldots$ be the sequence of vectors produced by FoReL. Then, for all $\mathrm{u} \in S$ we have:

$$\sum_{t=1}^{T}(f_t(\mathrm{w}_t) - f_t(\mathrm{u})) \leq R(\mathrm{u}) - R(\mathrm{w}_1) + \sum_{t=1}^{T}(f_t(\mathrm{w}_t) - f_t(\mathrm{w}_{t+1})).$$

Proof of Lemma 2.3

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

$$\sum_{t=0}^{T} (f_t(\mathrm{w}_t) - f_t(\mathrm{u})) \leq \sum_{t=0}^{T} (f_t(\mathrm{w}_t) - f_t(\mathrm{w}_{t+1})).$$

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

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 $w_{t+1} = w_t - \eta z_t$ Then, for all u we have

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

Choice of η and Final Bound for linear functions

Tunings:

- Define the set $U = \{\mathrm{u} : \|\mathrm{u}\| \leq B\}$.
- Assume that
$$\frac{1}{T} \sum_{t=1}^{T} \|\mathrm{z}_t\|_2^2 \leq L^2.$$
- Set $\eta = \frac{B}{L\sqrt{2T}}$.

Conclusion:

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

From linear functions to Online Gradient Descent

Example (OGD from FTRL)

Consider the OCO setup where the functions $f_1, f_2, \ldots$ are differentiable. Let η be the learning rate.

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

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

Regret bound on OGD: From FTRL theorem:

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

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

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

Bounding the norm of u

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

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

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 η)

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

$$BL\sqrt{2T}$$

Definition 2.4: Strong Convexity

Strong Convexity

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

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

Lemma 2.8: Strong Convexity implication

Lemma 2.8

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

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

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

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

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 w, x,

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

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

Example 2.4: Euclidean Regularization

The function

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

is 1-strongly-convex with respect to the ℓ_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

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

is $\frac{1}{B}$-strongly-convex with respect to the ℓ_1 norm over the set

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

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

$\frac{\partial^2}{\partial w[i]^2} w[i] \log w[i] = \frac{1}{w[i]}$

$$\langle \nabla^2 R(\mathrm{w})\mathrm{x}, \mathrm{x} \rangle = \sum_i \frac{x[i]^2}{w[i]}$$

$$= \frac{1}{\|\mathrm{w}\|_1} \left(\sum_i w[i] \right) \left(\sum_i \frac{x[i]^2}{w[i]} \right)$$

$$\geq \frac{1}{\|\mathrm{w}\|_1} \left(\sum_i \sqrt{w[i]} \frac{x[i]}{\sqrt{w[i]}} \right)^2 = \frac{\|\mathrm{x}\|_1^2}{\|\mathrm{w}\|_1},$$

where the inequality follows from Cauchy–Schwarz inequality.

Single Step of FTRL with Strong Convexity

Let

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

be a σ-strongly-convex function over S with respect to a norm $\|\cdot\|$. Let $\mathrm{w}_1, \mathrm{w}_2, \ldots$ be the predictions of the FoReL algorithm. Then, for all t, if f_t is L_t-Lipschitz with respect to $\|\cdot\|$, then:

$$f_t(\mathrm{w}_t) - f_t(\mathrm{w}_{t+1}) \leq L_t \|\mathrm{w}_t - \mathrm{w}_{t+1}\| \leq \frac{L_t^2}{\sigma}.$$

Proof (Single Step of FTRL with Strong Convexity)

For all t let

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

and note that the FoReL rule is

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

Note also that F_t is σ-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:

$$F_t(\mathrm{w}_{t+1}) \geq F_t(\mathrm{w}_t) + \frac{\sigma}{2}\|\mathrm{w}_t - \mathrm{w}_{t+1}\|^2.$$

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:

$$F_{t+1}(\mathrm{w}_t) \geq F_{t+1}(\mathrm{w}_{t+1}) + \frac{\sigma}{2}\|\mathrm{w}_t - \mathrm{w}_{t+1}\|^2.$$

Taking the difference between the last two inequalities and rearranging, we obtain:

$$\sigma\|\mathrm{w}_t - \mathrm{w}_{t+1}\|^2 \leq f_t(\mathrm{w}_t) - f_t(\mathrm{w}_{t+1}). \quad (2.7)$$

Final Steps (Single Step of FTRL with Strong Convexity)

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

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

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

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

Together with the above, we conclude our proof. ■

Main theorem regarding σ-strongly convex regularization functions

Let $f_1, \ldots, f_T$ 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 σ-strongly-convex with respect to the same norm. Then, for all $\mathrm{u} \in S$,

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

Corollary for l_2 regularization

Let $f_1, \ldots, f_T$ 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

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

Then, for all u,

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

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.

Experts using l_2 regularization (1)

S be a convex set and consider running FoReL with the regularization function:

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

Where S us the d dimensional simplex.
Then, for all $\mathrm{u} \in S$,

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

Experts using l_2 regularization (2)

If

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

Setting

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

then,

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

Entropic Regularization

Let $f_1, \ldots, f_T$ 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

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

and with the set

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

Then,

$$\mathrm{Regret}_T(S) \leq \frac{B\log(d)}{\eta} + \eta BTL^2.$$

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

$$\mathrm{Regret}_T(S) \leq BL\sqrt{2\log(d)T}.$$

Entropic regularization for Experts

The Entropic regularization is strongly convex with respect to the ℓ_1 norm, and therefore the Lipschitzness requirement of the loss functions is also with respect to the ℓ_1-norm.
For linear functions,

$$f_t(\mathrm{w}) = \langle \mathrm{w}, \mathrm{x_t} \rangle,$$

we have by Hölder's inequality that,

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

Therefore, the Lipschitz parameter grows with the ℓ_∞ norm of $\mathrm{x_t}$ rather than the ℓ_2 norm of $\mathrm{x_t}$.
expert advice: $B = 1$ and $L = 1$), we obtain the regret bound of

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

Comparison between regularizations

- entropic regularization vs. ℓ_2 regularization.
- $\log d$ vs $\sqrt{d}$
- L: $\|x_t\|_\infty \geq \|x_t\|_2$ Liphsitz condition carries heavier penalty with entropic regularization.
- B: $\|u\|_1 \leq \|u\|_2$ Comparator length carries heavier penalty with l_2 norm.