LECTURE 14 - First Order Algorithms for Online Convex Optimization

Contents

I Reminder: Online Convex Optimization
II Lower bounds on the regret
III Online gradient descent
  (1) Convex functions
  (2) Strongly convex functions
IV Online Mirror Descent
  (1) General statements
  (2) The EWA forecaster as a special case of Online Mirror Descend.
V Follow the regularized leader
  (1) Follow the leader
  (2) Things can go wrong without strong convexity
  (3) Follow the regularized leader.

I Reminder : Online Convex Optimization

We start by briefly reminding the reader about the general problem of Online Convex Optimization.

Suppose given a convex set $\mathcal{H} \subset \mathbb{R}^d$ as well as a set $\mathcal{F}$ of convex functions $f : \mathcal{H} \to \mathbb{R}$.

At each round $t \geq 1$, a player chooses a point $\theta_t \in \mathcal{H}$. Once the player commits to his choice, the environment (or an adversary) reveals a function $f_t \in \mathcal{F}$. The player then incurs loss $f_t(\theta_t)$ and moves on to the next round.

In this setting, the goal for the player is to minimize his regret

$$R_T = \sum_{t=1}^{T} f_t(\theta_t) - \min_{\theta \in \mathcal{H}} \sum_{t=1}^{T} f_t(\theta),$$

i.e. to design a strategy such that his cumulative loss is as good as the best fixed strategy in hindsight up to a small error term.

II Lower bounds on the regret

Before discussing specific algorithm, we briefly discuss lower bounds on the regret showing that there is a limit to what one can hope for. The following table

summarizes the limitations on the regret depending on different assumptions on the functions $f \in \mathcal{F}$

	Convex	β-exp concave	α-strongly convex
Lower bound on R_T	$\sqrt{T}$	$\frac{d}{\beta} \log T$	$\frac{1}{\alpha} \log T$

The above table displays "worst case" lower bounds on the regret and should be understood as follows.

<< There exist a convex set $\mathcal{H} \subset \mathbb{R}^d$ and a set $\mathcal{F}$ of convex (resp. β-exp concave, resp. α-strongly convex) functions $f : \mathcal{H} \to \mathbb{R}$ such that, for any OCO algorithm, we have

$$\sup_{f_1, \dots, f_T} R_T \geq c\sqrt{T} \quad \left(\text{resp.} \geq \frac{d}{\beta} \log T, \text{ resp.} \geq \frac{1}{\alpha} \log T \right)$$

where the sup is taken over all $f_1, \dots, f_T \in \mathcal{F}$. >>

The goal is then to devise specific algorithms that achieve these lower bounds up to constants. For more details in this direction, we refer the reader to the references listed at the end of Lecture 12.

II Online gradient descent

We start with online gradient descent. First introduced by Zinkevich (2003) in its online form, the algorithm operates along the same strategy as classical (offline) gradient descent for the optimization of a fixed convex function f: at each iteration the algorithm takes a step in the direction of minus the gradient of the previous loss function. While the next cost function may have nothing to do with the previous one, this strategy still yields sublinear regret. We study the algorithm in the setting of convex functions and in the setting of strongly convex functions.

Online Gradient Descent

Input : Convex set $\mathcal{H}$, T, $\theta_1 \in \mathcal{H}$, step sizes $\{\eta_t\}$

For $t = 1$ **to** T **do**

→ Play θ_t

→ Receive f_t and incur loss $f_t(\theta_t)$

→ Update

$$\theta_{t+1} = \Pi_{\mathcal{H}}(\theta_t - \eta_t \nabla_t)$$

where $\nabla_t \in \partial f_t(\theta_t)$

End for

---

① Convex function

We start with the case where elements of $\mathcal{F}$ are only assumed convex. We introduce

$$D = \operatorname{diam}_{\ell_2}(\Theta) = \max_{\theta, \theta' \in \Theta} \|\theta - \theta'\|_2$$

and

$$G = \sup \left\{ \|\nabla\|_2 : \nabla \in \partial f(\theta), f \in \mathcal{F}, \theta \in \Theta \right\}$$

which we suppose to be finite.

Theorem

Online gradient descent with step size $\eta_t = \frac{D}{G\sqrt{t}}$ guarantees, $\forall T \geq 1$,

$$R_T \leq \frac{3}{2} GD\sqrt{T}$$

Proof: For all $t \geq 1$ and all $\theta \in \Theta$, convexity of f_t implies that

$$f_t(\theta_t) - f_t(\theta) \leq \nabla_t^T (\theta_t - \theta).$$

Using the properties of projection and developping the ℓ_2-norm, we get

$$\|\theta_{t+1} - \theta\|_2^2 = \|\Pi_{\mathcal{B}(0)}(\theta_t - \eta_t \nabla_t) - \theta\|_2^2$$

$$\leq \|\theta_t - \eta_t \nabla_t - \theta\|_2^2$$

$$= \|\theta_t - \theta\|_2^2 - 2\eta_t \nabla_t^T(\theta_t - \theta) + \eta_t^2 \|\nabla_t\|_2^2$$

$$\Rightarrow \quad \nabla_t^T(\theta_t - \theta) \leq \frac{\|\theta_t - \theta\|_2^2 - \|\theta_{t+1} - \theta\|_2^2}{2\eta_t} + \frac{\eta_t}{2}\|\nabla_t\|_2^2$$

Since $f_t(\theta_t) - f_t(\theta) \leq \nabla_t^T(\theta_t - \theta)$, we deduce that

$$\sum_{t=1}^{T} (f_t(\theta_t) - f_t(\theta))$$

$$\leq \sum_{t=1}^{T} \frac{\|\theta_t - \theta\|_2^2 - \|\theta_{t+1} - \theta\|_2^2}{2\eta_t} + \sum_{t=1}^{T} \frac{\eta_t}{2}\|\nabla_t\|_2^2.$$

We set $\frac{1}{\eta_0} = 0$ by convention

$$= \frac{1}{2}\sum_{t=1}^{T} \underbrace{\|\theta_t - \theta\|_2^2}_{\leq D^2} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) - \underbrace{\frac{\|\theta_{T+1} - \theta\|_2^2}{2\eta_T}}_{\geq 0} + \sum_{t=1}^{T} \frac{\eta_t}{2} \underbrace{\|\nabla_t\|_2^2}_{\leq G^2}$$

$$\leq \frac{D^2}{2}\sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{G^2}{2}\sum_{t=1}^{T} \eta_t$$

$$= \frac{D^2}{2\eta_T} + \frac{G^2}{2}\sum_{t=1}^{T} \eta_t.$$

Plugging $\eta_t = \frac{D}{G\sqrt{t}}$ in the above expression and using the upper bound $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$, we

we get that, $\forall \theta \in \Theta$,

$$\sum_{t=1}^{T} (f_t(\theta_t) - f_t(\theta)) \leq \frac{3}{2} GD\sqrt{T}$$

As a result

$$R_T = \max_{\theta \in \Theta} \sum_{t=1}^{T} (f_t(\theta_t) - f_t(\theta)) \leq \frac{3}{2} GD\sqrt{T}.$$

□

(2) Strongly convex functions

Theorem

Suppose that all functions $f \in \mathcal{F}$ are α-strongly convex. Then Online Gradient Descent with step size $\eta_t = \frac{1}{\alpha t}$ satisfies, for all $T \geq 1$,

$$R_T \leq \frac{G^2}{2\alpha}(1 + \log(T))$$

Proof. The proof is a slight modification of the previous one. For all $t \geq 1$ and all $\theta \in \Theta$, we get by strong convexity of f_t that

$$f_t(\theta_t) - f_t(\theta) \leq \nabla_t^T(\theta_t - \theta) - \frac{\alpha}{2}\|\theta_t - \theta\|_2^2.$$

As in the previous proof, we get that

$$\nabla_t^T(\theta_t - \theta) \leqslant \frac{\|\theta_t - \theta\|_2^2 - \|\theta_{t+1} - \theta\|_2^2}{2\eta_t} + \frac{\eta_t}{2}\|\nabla_t\|_2^2$$

As a result

We set $1/\eta_0 = 0$ as in the previous proof by convention

$$\sum_{t=1}^{T}(f_t(\theta_t) - f_t(\theta))$$

$$\leqslant \sum_{t=1}^{T} \frac{\|\theta_t - \theta\|_2^2 - \|\theta_{t+1} - \theta\|_2^2}{2\eta_t} - \frac{\alpha}{2}\|\theta_t - \theta\|_2^2 + \frac{\eta_t}{2}\|\nabla_t\|_2^2$$

$$= \frac{1}{2}\sum_{t=1}^{T}\|\theta_t - \theta\|_2^2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \alpha\right) - \frac{\|\theta_{T+1} - \theta\|_2^2}{\eta_T} + \frac{1}{2}\sum_{t=1}^{T}\eta_t\|\nabla_t\|_2^2$$

$\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \alpha = 0$ by the choice of $\eta_t = \frac{1}{\alpha t}$

$\frac{\|\theta_{T+1} - \theta\|_2^2}{\eta_T} \geqslant 0$

$\|\nabla_t\|_2^2 \leqslant G^2$

$$\leqslant \frac{G^2}{2}\sum_{t=1}^{T}\eta_t.$$

$$= \frac{G^2}{2\alpha}\sum_{t=1}^{T}\frac{1}{t}.$$

It remains to note that $\sum_{t=1}^{T}\frac{1}{t} \leqslant 1 + \log(T)$.

□

Remark: Unlike the case of convex functions, the previous analysis seems to hold even for a convex set $\mathcal{H}$ of unbounded diameter. The thing is that there are no α-strongly convex functions on an unbounded convex set with bounded gradients! In otherwords if $D = +\infty$ then $G = +\infty$ here!

IV Online Mirror Descent

(1) General Statements.

With an analysis similar to the one performed in Lectures 9 & 10, it is immediate to show that the updates of Online Gradient Descent can be formulated as

$$\theta_{t+1} \in \underset{\theta \in \mathcal{H}}{\operatorname{argmin}} \left\{ \theta^T \nabla_t + \frac{\|\theta - \theta_t\|_2^2}{2\eta_t} \right\}$$

where $\nabla_t \in \partial f_t(\theta_t)$ (the only $\neq$ being that the function f_t changes at each time step). According to same philosophy as in the context of classical (offline) Mirror Descent, one may want to adapt Online GD to different settings where the use of the Euclidean norm doesn't necessarily reflect the "geometry" of the problem. Online Mirror Descent is the straightforward adaptation of offline Mirror Descent to the the online framework.

Let us recall some notation allready introduced in Lectures 9 and 10.

We suppose that $\mathcal{H}$ is a closed convex and compact set $\subset \mathbb{R}^d$. We introduce a Mirror map

$$\phi : \mathcal{D} \longrightarrow \mathbb{R}$$

where $\mathcal{D}$ is an open convex set. We further suppose that $\mathcal{H} \cap \mathcal{D} \neq \emptyset$ and that $\mathcal{H} \subset \overline{\mathcal{D}}$. Finally we suppose that

- ϕ is strictly convex and differentiable on $\mathcal{D}$
- $\nabla\phi : \mathcal{D} \to \mathbb{R}^d$ is surjective
- $\lim_{x \to \partial\mathcal{D}} \|\nabla\phi(x)\| = +\infty$ (for some and therefore any norm $\|.\|$ on $\mathbb{R}^d$)

The Bregman divergence associated to ϕ is defined, $\forall \theta, \theta' \in \mathcal{D}$, by

$$D_\phi(\theta' \| \theta) = \phi(\theta') - \phi(\theta) - \nabla\phi(\theta)^T(\theta' - \theta).$$

Online Mirror Descent

Inputs: Convex set $\mathcal{H}$, T, step sizes $\{\eta_t\}$.
Mirror map $\phi : \mathcal{D} \to \mathbb{R}$

Initialization: $\theta_1 \in \underset{\theta \in \mathcal{H} \cap \mathcal{D}}{\operatorname{argmin}} \ \phi(\theta)$

For $t = 1$ **to** T **do**

- Play θ_t
- Receive f_t and incur loss $f_t(\theta_t)$
- Update

$$\theta_{t+1} \in \underset{\theta \in \mathcal{H} \cap \mathcal{D}}{\operatorname{argmin}} \left\{ \theta^T \nabla_t + \frac{D_\phi(\theta \| \theta_t)}{\eta_t} \right\}$$

where $\nabla_t \in \partial f_t(\theta_t)$.

End for

As for the offline setting, the analysis relies on the fact that the updates of the Online MD algorithm can be equivalently written as :

$$\begin{cases} \nabla\phi(\theta'_{t+1}) = \nabla\phi(\theta_t) - \eta_t \nabla_t \\ \text{and} \\ \theta_{t+1} \in \underset{\theta \in \mathcal{H} \cap \mathcal{D}}{\operatorname{argmin}} \; D_\phi(\theta \| \theta'_{t+1}) \end{cases}$$

The equivalence of the two formulations have been established in Lectures 9 & 10. We now move to the analysis of Online MD.

Theorem

Suppose ϕ is ρ-strongly convex on $\mathcal{H} \cap \mathcal{D}$ with respect to some norm $\|.\|$ on $\mathbb{R}^d$. Suppose that there exists $G > 0$ such that $\forall f \in \mathcal{F}, \forall \theta \in \mathcal{H}, \forall \nabla \in \partial f(\theta) : \|\nabla\|_* \leq G$ where $\|.\|_*$ is the dual norm of $\|.\|$. Suppose all functions $f \in \mathcal{F}$ are convex. Then $\forall T \geq 1$ Online Mirror Descent with $\eta_t = \eta = \frac{\Delta_\phi}{G}\sqrt{\frac{2\rho}{T}}$ satisfies,

$$R_T \leq G\Delta_\phi\sqrt{\frac{2T}{\rho}}$$

where $\Delta_\phi = \max_{\theta, \theta'} \sqrt{\phi(\theta) - \phi(\theta')}$

Proof: We only sketch the proof since most of the work has been done in Lectures 9 & 10. First note that, as usual, convexity of f_t implies that, $\forall\, \theta \in \mathcal{H} \cap \mathcal{D}$,

$$f_t(\theta_t) - f_t(\theta) \leq \nabla_t^T (\theta_t - \theta).$$

The exact same analysis as in Lectures 9 & 10 reveals that (the reader can check that the fact that $\nabla_t \in \partial f_t(\theta_t)$ rather than $\nabla_t \in \partial f(\theta_t)$ for some fixed f makes no difference in the analysis!)

$$\nabla_t^T (\theta_t - \theta) \leq \frac{D_\phi(\theta \| \theta_t) - D_\phi(\theta \| \theta_{t+1})}{\eta} + \frac{\eta}{2\rho} \|\nabla_t\|_*^2.$$

As a result, we get

$$\begin{aligned}
&\sum_{t=1}^{T} (f_t(\theta_t) - f_t(\theta)) \\
&\leq \sum_{t=1}^{T} \frac{D_\phi(\theta \| \theta_t) - D_\phi(\theta \| \theta_{t+1})}{\eta} + \frac{\eta}{2\rho} \sum_{t=1}^{T} \underbrace{\|\nabla_t\|_*^2}_{\leq G^2} \\
&= \frac{D_\phi(\theta \| \theta_1) - \underbrace{D_\phi(\theta \| \theta_{T+1})}_{\geq 0}}{\eta} + \frac{\eta T}{2\rho} G^2 \\
&\leq \frac{D_\phi(\theta \| \theta_1)}{\eta} + \frac{\eta T}{2\rho} G^2.
\end{aligned}$$

As in Lectures 9 & 10, we have by choice of θ_1 that

$$D_\phi(\theta \| \theta_1) \leq \phi(\theta) - \phi(\theta_1) \leq \Delta_\phi^2.$$

Plugging the value of $\eta = \frac{\Delta_\phi}{G}\sqrt{\frac{2\rho}{T}}$, we get

$$\sum_{t=1}^{T} f_t(\theta_t) - f_t(\theta) \leq \frac{\Delta_\phi^2}{\eta} + \frac{\eta T}{2\rho} G^2 \leq \Delta_\phi G \sqrt{\frac{2T}{\rho}}.$$

Since this holds for all $\theta \in \mathcal{H} \cap \mathcal{D}$ and since $\mathcal{H} \subset \overline{\mathcal{D}}$ letting $\theta \to \theta^*$ which minimzes the sum on the left hand side leads to the claim. □

(2) The EWA forecaster as a special case of online MD.

Recall from previous lectures the pbl of prediction with expert advice (with N experts). In lecture 12, we've argued that (see lecture 12, III, (2), paragraphs 2.2 & 2.3) if one plans to compute his decision $\hat{a}_t$ at time t as a convex combination of the expert advice $\{a_{i,t}\}_{i=1}^N$, i.e.

$$\hat{a}_t = \sum_{i=1}^{N} p_{i,t}\, a_{i,t} = p_t^T a_t$$

where $p_t = (p_{1,t}, \dots, p_{N,t})^T \in \Delta_N$ and $a_t = (a_{1,t}, \dots, a_{N,t})^T$ then for any strategy for choosing $p_t \in \Delta_N$ at each time step, we have that

$$R_T^{exp} := \sum_{t=1}^{T} \ell(\hat{a}_t, y_t) - \min_{1 \le i \le N} \sum_{t=1}^{T} \ell(a_{i,t}, y_t)$$

$$\le \sum_{t=1}^{T} p_t^T \ell_t - \min_{p \in \Delta_N} \sum_{t=1}^{T} p^T \ell_t =: R_T^{oco}$$

where $\ell_t := (\ell_{1,t}, \dots, \ell_{N,t})^T \in [a,b]^N$ is the vector of losses $\ell_{i,t} := \ell(a_{i,t}, y_t)$ of the experts at time t.

From this point of view, choosing p_t in order to minimize this upper bound can be understood as an OCO problem where

- $\mathcal{H} = \Delta_N$: the probability simplex in $\mathbb{R}^N$
- $\mathcal{F} = \{ p \in \Delta_N \mapsto p^T \ell : \ell \in [a,b]^N \}$

In the setting of Δ_N, an interesting mirror map is the negative entropy:

$$\phi : \mathcal{D} = (0, +\infty)^N \to \mathbb{R}$$

$$\phi(x) = \sum_{i=1}^{N} x(i) \log x(i).$$

We have already seen that ϕ is 1-strongly convex with respect to the ℓ_1-norm. In addition, if $\ell \in [a,b]^N$ and $p \in \Delta_N$, the function $f(p) = p^T \ell$ obviously satisfies

$$\| \nabla f \|_\infty = \| \ell \|_\infty \le \max\{|a|, |b|\}$$

Which suits perfectly the assumptions needed for Online MD with $\|.\| = \|.\|_1$ ($\|.\|_* = \|.\|_\infty$). In addition, in this very specific setting since

* $\nabla\phi(x) = (1+\log x(1), \dots, 1+\log x(N))^T$
* $f_t(p) = p^T \ell_t \Rightarrow \nabla_t = \ell_t$
* $\underset{p \in \Delta_N \cap (0,+\infty)^N}{\operatorname{argmin}} D_\phi(p \| x) = \dfrac{x}{\|x\|_1} = \dfrac{x}{\sum_{i=1}^N x(i)}$

One easily sees that the Online MD updates can be written as (see Lecture 9 & 10 for more details)

$$p_{i,t+1} = \frac{p_{i,t}\, e^{-\eta_t \ell_{i,t}}}{\sum_{j=1}^N p_{j,t}\, e^{-\eta_t \ell_{j,t}}}$$

(Using the more natural notation p_t instead of θ_t in the context of the simplex).

The reader recognizes here that the prediction

$$\hat{a}_t = \sum_{i=1}^N p_{i,t}\, a_{i,t}$$

constructed with the above $p_{i,t}$'s is the EWA forecaster!

V Follow the regularized leader

We now turn to another algorithm known as "Follow the Leader" (FTL) and its variant known as "Follow the Regularized Leader" (FTRL). The FTL algorithm consists in the following natural strategy

---

Follow The Leader (FTL)

Input: Convex set $\mathcal{H}$, $\theta_1 \in \mathcal{H}$, $T \geq 1$

For $t = 1$ **to** T **do**

- Play θ_t
- Receive f_t and incur loss $f_t(\theta_t)$
- Update

$$\theta_{t+1} \in \operatorname*{argmin}_{\theta \in \mathcal{H}} \sum_{s=1}^{t} f_s(\theta)$$

End for

The FTL algorithm seem natural for OCO since the ultimate goal is to compete against the best fixed strategy in hindsight i.e.

$$\theta_T^* \in \operatorname*{argmin}_{\theta \in \mathcal{H}} \sum_{t=1}^{T} f_t(\theta).$$

The first paragraph below shows that the strategy works great for strongly convex losses. However we show that using the FTL strategy

may go really wrong for simply convex losses. This observation motivates the introduction of a regularized version of the FTL algorithm : The FTRL algorithm.

① Follow the Leader

We start with a technical lemma.

Lemma

If $(\theta_t)_{t\geq 1}$ is the sequence produced by the FTL algorithm, then $\forall T \geq 1$ and $\forall \theta \in \mathcal{H}$,

$$\sum_{t=1}^{T} f_t(\theta_{t+1}) \leq \sum_{t=1}^{T} f_t(\theta)$$

Proof : We'll prove the result by induction on T. For $T=1$, the inequality is true by definition of

$$\theta_2 \in \operatorname*{argmin}_{\theta \in \mathcal{H}} f_1(\theta).$$

Now suppose the result is true for some $T \geq 1$. Then we have by the induction hypothesis that

$$\begin{aligned}
\sum_{t=1}^{T+1} f_t(\theta_{t+1}) &= \sum_{t=1}^{T} f_t(\theta_{t+1}) + f_{T+1}(\theta_{T+2}) \\
&\leq \sum_{t=1}^{T} f_t(\theta_{T+2}) + f_{T+1}(\theta_{T+2}) \\
&= \sum_{t=1}^{T+1} f_t(\theta_{T+2})
\end{aligned}$$

By by definition of θ_{T+2}, we know that

$$\sum_{t=1}^{T+1} f_t(\theta_{T+2}) \leq \sum_{t=1}^{T+1} f_t(\theta), \quad \forall \theta \in \mathcal{U}.$$

Combining the two inequalities above implies the result. □

Theorem

Suppose all functions $f \in \mathcal{F}$ are α-strongly convex and L-Lipschitz. Then the FTL algorithm satisfies, $\forall T \geq 1$,

$$R_T \leq \frac{2L^2}{\alpha}(1 + \log T).$$

Proof: Denote $\psi_t(\theta) = \sum_{s=1}^{t} f_s(\theta)$.

By assumption, ψ_t is $t\alpha$-strongly convex. Since by definition

$$\theta_{t+1} \in \operatorname*{argmin}_{\theta \in \mathcal{U}} \psi_t(\theta)$$

we deduce that $\forall \theta \in \mathcal{U}$ (see next lemma)

$$\frac{t\alpha}{2} \|\theta_{t+1} - \theta\|_2^2 \leq \psi_t(\theta) - \psi_t(\theta_{t+1}).$$

$*_1$

Next, lets prove by induction that $\forall\, t \geq 1$

$$\Psi_t(\theta_t) - \Psi_t(\theta_{t+1}) \leq f_t(\theta_t) - f_t(\theta_{t+1}). \quad *_2$$

Indeed, this inequality is obvious for $t=1$ since $\Psi_1 = f_1$. Supposing it is true for some $t \geq 1$, observe that

$$\Psi_{t+1}(\theta_{t+1}) - \Psi_{t+1}(\theta_{t+2}) = \underbrace{\Psi_t(\theta_{t+1}) - \Psi_t(\theta_{t+2})}_{\leq 0 \text{ by definition of } \theta_{t+1}} + f_{t+1}(\theta_{t+1}) - f_{t+1}(\theta_{t+2})$$

$$\leq f_{t+1}(\theta_{t+1}) - f_{t+1}(\theta_{t+2}),$$

which proves $*_2$. Now combining $*_1$ & $*_2$, we deduce that, $\forall\, t \geq 1$

$$\frac{t\alpha}{2} \|\theta_{t+1} - \theta_t\|_2^2 \leq f_t(\theta_t) - f_t(\theta_{t+1}) \quad *_3.$$

Since each f_t is L-Lipschitz, we also deduce that $\forall\, t \geq 1$

$$f_t(\theta_t) - f_t(\theta_{t+1}) \leq L \|\theta_{t+1} - \theta_t\|_2 \quad *_4$$

Combining $*_3$ & $*_4$, it follows that

$$\frac{t\alpha}{2} \|\theta_{t+1} - \theta_t\|_2^2 \leq L \|\theta_{t+1} - \theta_t\|_2$$

$$\Rightarrow \quad \|\theta_{t+1} - \theta_t\|_2 \leq \frac{2L}{\alpha t} \quad *_5$$

Plugging $*_5$ back in $*_4$ implies that, $\forall t \geq 1$,

$$f_t(\theta_t) - f_t(\theta_{t+1}) \leq \frac{2L^2}{\alpha t} \qquad *_6$$

Now applying the previous lemma, we get that, $\forall \theta \in \mathcal{H}$

$$\sum_{t=1}^{T} f_t(\theta_t) - f_t(\theta) \leq \sum_{t=1}^{T} f_t(\theta_t) - f_t(\theta_{t+1})$$

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

According to $*_6$

$$\leq \frac{2L^2}{\alpha} (1 + \log(T)).$$

Since this holds for all $\theta \in \mathcal{H}$, we get that

$$R_T = \max_{\theta \in \mathcal{H}} \sum_{t=1}^{T} f_t(\theta_t) - f_t(\theta)$$

$$\leq \frac{2L^2}{\alpha} (1 + \log(T)). \qquad \square$$

(2) Things can go wrong without strong convexity.

In this paragraph we show that the intuitive FTL algorithm (which we've shown works very well in the context of strongly convex losses) can behave very poorly

in the context of convex losses.

Indeed, consider the situation where $\Theta = [-1,1]$, $F = \{ f(\theta) = \theta z \;,\; z \in [-1,1] \}$ and the sequence $f_t(\theta) = \theta z_t$ provided by the adversary is defined by

$$z_t = \begin{cases} -1/2 & \text{for } t = 1 \\ +1 & \text{for } t \geq 2 \text{ even} \\ -1 & \text{for } t \geq 2 \text{ odd} \end{cases}$$

which implies that

$$\sum_{t=1}^{T} z_t = \frac{(-1)^T}{2}$$

This implies that the FTL algorithm computes (besides the arbitrary initial point $\theta_1 \in [-1,1]$) $\forall\, t \geq 2$

$$\begin{aligned} \theta_t &= \operatorname*{argmin}_{\theta \in [-1,1]} \sum_{s=1}^{t-1} \theta z_s \\ &= \operatorname*{argmin}_{\theta \in [-1,1]} \theta \frac{(-1)^{t-1}}{2} \\ &= (-1)^t \\ &= z_t \end{aligned}$$

Clearly this implies that the regret of the FTL algorithm in this setting is

$$R_T = \sum_{t=1}^{T} \theta_t z_t - \min_{-1 \le \theta \le 1} \sum_{t=1}^{T} \theta z_t = \frac{1-\theta_1}{2} + (T-1) \ge T-1.$$

$$= \frac{-\theta_1}{2} + (T-1) \qquad = -\frac{1}{2}$$

(3) Follow the Regularized Leader (FTRL)

A small modification of the FTL algorithm is able to automatically fix the issue of the previous paragraph and works very well in the context of convex functions. We mention two variants of the algorithm.

Follow the Regularized Leader (FTRL) Variant 1

Input: $\mathcal{H}$, $T \ge 1$, Mirror Map $\phi : \mathcal{D} \to \mathbb{R}$

Initialize:

$$\theta_1 \in \operatorname*{argmin}_{\theta \in \mathcal{H} \cap \mathcal{D}} \phi(\theta)$$

For $t=1$ **to** T **do**

→ Play θ_t

→ Receive f_t and incur loss $f_t(\theta_t)$

→ Update

$$\theta_{t+1} \in \operatorname*{argmin}_{\theta \in \mathcal{H} \cap \mathcal{D}} \left\{ \sum_{s=1}^{t} f_s(\theta) + \frac{\phi(\theta)}{\eta_s} \right\}$$

Follow the Regularized Leader (FTRL) Variant 2

Input: $\mathcal{H}$, $T \geq 1$, Mirror Map $\phi : \mathcal{D} \to \mathbb{R}$

Initialize:

$$\theta_1 \in \operatorname*{argmin}_{\theta \in \mathcal{H} \cap \mathcal{D}} \phi(\theta)$$

For $t = 1$ **to** T **do**

- Play θ_t
- Receive f_t and incur loss $f_t(\theta_t)$
- Update

$$\theta_{t+1} \in \operatorname*{argmin}_{\theta \in \mathcal{H} \cap \mathcal{D}} \left\{ \sum_{s=1}^{t} \theta^T \nabla_s + \frac{\phi(\theta)}{\eta_s} \right\}$$

for $\nabla_s \in \partial f_s(\theta_s)$

Variant 1 is a more obvious modification of the FTL algorithm. Variant 2 is however more computationally friendly version and corresponds to replacing $f_s(\theta)$ in Variant 1 by its 1st order approximation around θ_s, i.e., $f_s(\theta_s) + \nabla_s^T(\theta - \theta_s)$. Indeed, we easily observe that

$$\operatorname*{argmin}_{\theta} \left\{ \sum_{s=1}^{t} (f_s(\theta_s) + \nabla_s^T(\theta - \theta_s)) + \frac{\phi(\theta)}{\eta_t} \right\}$$

$$= \operatorname*{argmin}_{\theta} \left\{ \sum_{s=1}^{t} \theta^T \nabla_s + \frac{\phi(\theta)}{\eta_t} \right\}.$$

While the analysis of variant 1 can be easily

done along the same lines, we analyse Version 2 below.

Lemma

If $(\theta_t)_{t\geq 1}$ are the iterations of the FTRL (variant 2) algorithm, then $\forall\, T \geq 1$, $\forall\, \theta \in \mathcal{U} \cap \mathcal{D}$

$$\sum_{t=1}^{T} \nabla_t^T \theta_{t+1} + \frac{\phi(\theta_1)}{\eta_T} \leq \sum_{t=1}^{T} \nabla_t^T \theta + \frac{\phi(\theta)}{\eta_T}$$

Proof: As in a similar lemma, for the FTL Algorithm, we proceed by induction on T. For $T = 1$, the result is true since

$$\nabla_1^T \theta_2 + \frac{\phi(\theta_1)}{\eta_1} \leq \nabla_1^T \theta_2 + \frac{\phi(\theta_2)}{\eta_1}$$

(since $\theta_1 \in \operatorname{argmin}_{\theta \in \mathcal{U} \cap \mathcal{D}} \phi(\theta)$)

$$\leq \nabla_1^T \theta + \frac{\phi(\theta)}{\eta_1}$$

(by definition of θ_2).

Now supposing the result true for some $T \geq 1$, we have (next page):

$$\sum_{t=1}^{T+1} \nabla_t^T \theta_{t+1} + \frac{\phi(\theta_1)}{\eta_{T+1}}$$

$$= \nabla_{T+1}^T \theta_{T+2} + \phi(\theta_1)\left(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_T}\right) + \sum_{t=1}^{T} \nabla_t^T \theta_{t+1} + \frac{\phi(\theta_1)}{\eta_T}$$

$$\leq \nabla_{T+1}^T \theta_{T+2} + \phi(\theta_1)\left(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_T}\right) + \sum_{t=1}^{T} \nabla_t^T \theta_{T+2} + \frac{\phi(\theta_{T+2})}{\eta_T}$$

(induction hypothesis)

$$\leq \nabla_{T+1}^T \theta_{T+2} + \phi(\theta_{T+2})\left(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_T}\right) + \sum_{t=1}^{T} \nabla_t^T \theta_{T+2} + \frac{\phi(\theta_{T+2})}{\eta_T}$$

$\left(\theta_1 \in \operatorname{argmin}_{\theta \in \Theta \cap \mathcal{D}} \phi(\theta) \;\&\; \frac{1}{\eta_{T+1}} \geq \frac{1}{\eta_T}\right)$

$$= \sum_{t=1}^{T+1} \nabla_t^T \theta_{T+2} + \frac{\phi(\theta_{T+2})}{\eta_{T+1}}$$

$$\leq \sum_{t=1}^{T+1} \nabla_t^T \theta + \frac{\phi(\theta)}{\eta_{T+1}}, \quad \forall\, \theta \in \Theta \cap \mathcal{D}$$

(By definition of θ_{T+2}).

This concludes the proof. □

Theorem

Suppose that all functions $f \in \mathcal{F}$ are convex, that ϕ is ρ-strongly cvx (with respect to some norm $\|\cdot\|$) on $\Theta \cap \mathcal{D}$ and that $G = \sup \{ \|\nabla\|_* : \nabla \in \partial f(\theta), f \in \mathcal{F}, \theta \in \Theta \} < +\infty$.

Then the FTRL (variant 2) algorithm with

$$\eta_t = \frac{\Delta_\phi}{2G} \sqrt{\frac{\rho}{t}}$$

satisfies $\forall T \geq 1$

$$R_T \leq 4 \Delta_\phi G \sqrt{\frac{T}{\rho}}$$

where $\Delta_\phi = \sqrt{\sup_{\theta \in \Theta \cap \mathcal{D}} \phi(\theta) - \phi(\theta_1)}$.

Proof: Denote $\psi_t(\theta) = \sum_{s=1}^{t} \theta^T \nabla_s + \frac{\phi(\theta)}{\eta_t}$.

By assumption on ϕ, ψ_t is $\frac{\rho}{\eta_t}$-strongly convex on $\Theta \cap \mathcal{D}$.

Since θ_{t+1} minimizes ψ_t over $\Theta \cap \mathcal{D}$, we deduce that, $\forall \theta \in \Theta \cap \mathcal{D}$,

$$\frac{\rho}{2\eta_t} \|\theta_{t+1} - \theta\|^2 \leq \psi_t(\theta) - \psi_t(\theta_{t+1}). \quad (*_1)$$

On the other hand, we can easily show that, $\forall t \geq 1$,

$$\psi_t(\theta_t) - \psi_t(\theta_{t+1}) \leq \nabla_t^T (\theta_t - \theta_{t+1}). \quad (*_2)$$

[Indeed, for $t=1$, the inequality is clearly true since

$$\psi_1(\theta_1) - \psi_1(\theta_2) = \nabla_1^T(\theta_1 - \theta_2) + \frac{\phi(\theta_1)}{\eta_1} - \frac{\phi(\theta_2)}{\eta_1}$$

$$\leq \nabla_1^T(\theta_1 - \theta_2)$$

since $\phi(\theta_1) \leq \phi(\theta)$, $\forall \theta \in \mathcal{U} \cap \mathcal{D}$.
And for $t \geq 2$

$$\psi_t(\theta_t) - \psi_t(\theta_{t+1}) = \psi_{t-1}(\theta_t) - \psi_{t-1}(\theta_{t+1}) + \nabla_t^T(\theta_t - \theta_{t+1})$$

$$\leq \nabla_t^T(\theta_t - \theta_{t+1}),$$

Since $\theta_t \in \underset{\theta \in \mathcal{U} \cap \mathcal{D}}{\operatorname{argmin}} \psi_{t-1}(\theta)$ by definition of the FTRL algo.]

In addition, by assumption we get that

$$\nabla_t^T(\theta_t - \theta_{t+1}) \leq \|\nabla_t\|_* \|\theta_t - \theta_{t+1}\|$$

$$\leq G \|\theta_t - \theta_{t+1}\| \quad *_3$$

Combining $*_1$, $*_2$ & $*_3$ we deduce that

$$\nabla_t^T(\theta_t - \theta_{t+1}) \leq \frac{2G^2\eta_t}{\rho}. \quad *_4$$

Now by the previous lemma, we get that $\forall\ \theta \in \mathcal{U} \cap \mathcal{D}$

$$\sum_{t=1}^{T} \nabla_t^T (\theta_t - \theta) \leq \frac{\phi(\theta) - \phi(\theta_1)}{\eta_T} + \sum_{t=1}^{T} \nabla_t^T (\theta_t - \theta_{t+1})$$

$$\leq \frac{\phi(\theta) - \phi(\theta_1)}{\eta_T} + \frac{2G^2}{\rho} \sum_{t=1}^{T} \eta_t$$

where the last inequality follows from $\star_4$.
By convexity of each f_t, we deduce that

$$f_t(\theta_t) - f_t(\theta) \leq \nabla_t^T (\theta_t - \theta)$$

$$\Rightarrow \quad \sum_{t=1}^{T} (f_t(\theta_t) - f_t(\theta)) \leq \frac{\phi(\theta) - \phi(\theta_1)}{\eta_T} + \frac{2G^2}{\rho} \sum_{t=1}^{T} \eta_t.$$

$$\leq \frac{\Delta^2 \phi}{\eta_T} + \frac{2G^2}{\rho} \sum_{t=1}^{T} \eta_t.$$

Letting $\theta \to \theta_T^* \in \mathcal{U}$ such that

$$\theta_T^* \in \operatorname{argmin}_{\theta \in \mathcal{U}} \sum_{t=1}^{T} f_t(\theta),$$

(OK because $\theta \in \mathcal{U} \cap \mathcal{D}$ and $\mathcal{U} \subset \overline{\mathcal{D}}$)

we deduce finally that

$$R_T \leq \frac{\Delta^2 \phi}{\eta_T} + \frac{2G^2}{\rho} \sum_{t=1}^{T} \eta_t.$$

Plugging the value of $\eta_t = \frac{\Delta \phi}{2G} \sqrt{\frac{\rho}{t}}$ concludes. □