Online Learning Summer School
Copenhagen 2015
Lecture 2

Shai Shalev-Shwartz

School of CS and Engineering,
The Hebrew University of Jerusalem

Online Learning

Outline

Reminder: The Online Classification Game

For $t = 1, 2, \ldots$

- Environment presents a question x_t
- Learner predicts an answer $\hat{y}_t \in \{\pm 1\}$
- Environment reveals true label $y_t \in \{\pm 1\}$
- Learner pays $1[\hat{y}_t \neq y_t]$

Reminder: The Online Classification Game

For $t = 1, 2, \ldots$

- Environment presents a question x_t
- Learner predicts an answer $\hat{y}_t \in \{\pm 1\}$
- Environment reveals true label $y_t \in \{\pm 1\}$
- Learner pays $1[\hat{y}_t \neq y_t]$

- Realizability by $\mathcal{H}$ assumption: $\exists f \in \mathcal{H}$ s.t. $\forall t,\ y_t = f(x_t)$
- What if this assumption is wrong ? What should be a reasonable goal for the learner ?

Relaxing the prior knowledge

- Regret: the difference between the number of mistakes the learner made and the number of mistakes of the best $f \in \mathcal{H}$

$$\text{Regret}_T := \sum_{t=1}^{T} 1[\hat{y}_t \neq y_t] - \min_{f \in \mathcal{H}} \sum_{t=1}^{T} 1[f(x_t) \neq y_t]$$

Relaxing the prior knowledge

- Regret: the difference between the number of mistakes the learner made and the number of mistakes of the best $f \in \mathcal{H}$

$$\mathrm{Regret}_T := \sum_{t=1}^{T} 1[\hat{y}_t \neq y_t] - \min_{f \in \mathcal{H}} \sum_{t=1}^{T} 1[f(x_t) \neq y_t]$$

- Vanishing regret: Our modified goal is to have $\mathrm{Regret}_T = o(T)$. If this holds then

$$\frac{1}{T} \sum_{t=1}^{T} 1[\hat{y}_t \neq y_t] - \min_{f \in \mathcal{H}} \frac{1}{T} \sum_{t=1}^{T} 1[f(x_t) \neq y_t] \quad \rightarrow \quad 0$$

Relaxing the prior knowledge

- Regret: the difference between the number of mistakes the learner made and the number of mistakes of the best $f \in \mathcal{H}$

$$\mathrm{Regret}_T := \sum_{t=1}^{T} 1[\hat{y}_t \neq y_t] - \min_{f \in \mathcal{H}} \sum_{t=1}^{T} 1[f(x_t) \neq y_t]$$

- Vanishing regret: Our modified goal is to have $\mathrm{Regret}_T = o(T)$. If this holds then

$$\frac{1}{T}\sum_{t=1}^{T} 1[\hat{y}_t \neq y_t] - \min_{f \in \mathcal{H}} \frac{1}{T}\sum_{t=1}^{T} 1[f(x_t) \neq y_t] \quad \rightarrow \quad 0$$

- Is this a good goal ?

Is low regret a good goal ?

- Data dependent: yes, if there's $f \in \mathcal{H}$ that makes a "small" number of mistakes
- We'll later generalize $\mathcal{H}$ to be strategies (instead of fixed functions), and then the concept of regret becomes even stronger

Cover's impossibility Result

Can we have a vanishing regret for a finite $\mathcal{H}$?

Cover's impossibility Result

Can we have a vanishing regret for a finite $\mathcal{H}$?

- Take $\mathcal{H} = \{h_+, h_-\}$ where $h_+(x)$ is always 1 and $h_-(x)$ is always -1

Cover's impossibility Result

Can we have a vanishing regret for a finite $\mathcal{H}$?

- Take $\mathcal{H} = \{h_+, h_-\}$ where $h_+(x)$ is always 1 and $h_-(x)$ is always -1
- For every choice $\hat{y}_t$ the adversary will pick $y_t = -\hat{y}_t$

Cover's impossibility Result

Can we have a vanishing regret for a finite $\mathcal{H}$?

- Take $\mathcal{H} = \{h_+, h_-\}$ where $h_+(x)$ is always 1 and $h_-(x)$ is always -1
- For every choice $\hat{y}_t$ the adversary will pick $y_t = -\hat{y}_t$
- Claim: the regret is $\geq T/2$

Can we have a vanishing regret for a finite $\mathcal{H}$?

- Take $\mathcal{H} = \{h_+, h_-\}$ where $h_+(x)$ is always 1 and $h_-(x)$ is always -1
- For every choice $\hat{y}_t$ the adversary will pick $y_t = -\hat{y}_t$
- Claim: the regret is $\geq T/2$
- Proof: The learner makes T mistakes while $h_{\mathrm{MAJORITY}(y_1,\ldots,y_T)}$ makes at most $T/2$ mistakes.

Circumventing the impossibility result

Intuitively, we can:

- Make the adversary (slightly) weaker
- Make the regret (slightly) weaker

Circumventing the impossibility result

- Randomization: the learner calculates $p_t \in [0,1]$, and the loss is redefined to be $\mathbb{P}_{\hat{y}_t \sim p_t}[\hat{y}_t \neq y_t]$
- Multiplicative factor: redefine the regret to be

$$\sum_{t=1}^{T} 1[\hat{y}_t \neq y_t] - 2 \min_{f \in \mathcal{H}} \sum_{t=1}^{T} 1[f(x_t) \neq y_t]$$

Circumventing the impossibility result

- Randomization: the learner calculates $p_t \in [0, 1]$, and the loss is redefined to be $\mathbb{P}_{\hat{y}_t \sim p_t}[\hat{y}_t \neq y_t]$
- Multiplicative factor: redefine the regret to be

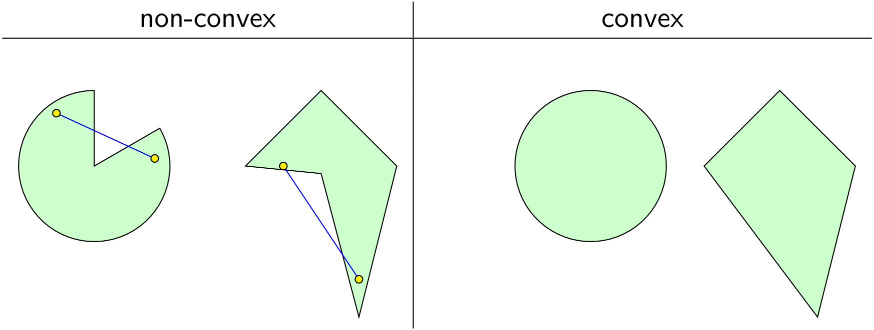
$$\sum_{t=1}^{T} 1[\hat{y}_t \neq y_t] - 2 \min_{f \in \mathcal{H}} \sum_{t=1}^{T} 1[f(x_t) \neq y_t]$$

As we'll show, both techniques rely on convexification

Outline

Definition (Convex Set)

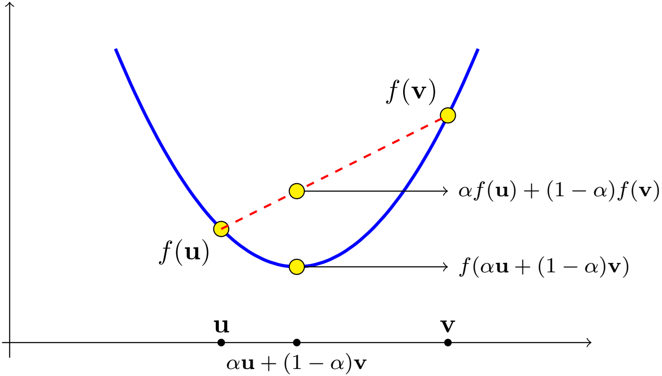
A set C in a vector space is convex if for any two vectors $\mathbf{u}, \mathbf{v}$ in C, the line segment between $\mathbf{u}$ and $\mathbf{v}$ is contained in C. That is, for any $\alpha \in [0, 1]$ we have that the convex combination $\alpha\mathbf{u} + (1 - \alpha)\mathbf{v}$ is in C.



Definition (Convex function)

Let C be a convex set. A function $f : C \to \mathbb{R}$ is convex if for every $\mathbf{u}, \mathbf{v} \in C$ and $\alpha \in [0, 1]$,

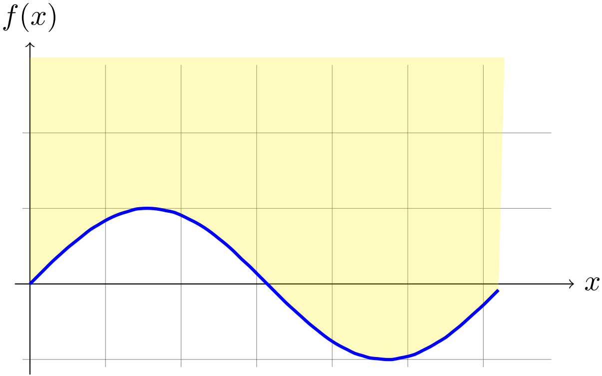
$$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \;\leq\; \alpha f(\mathbf{u}) + (1-\alpha) f(\mathbf{v}) \; .$$



Epigraph

A function f is convex if and only if its *epigraph* is a convex set:

$$\text{epigraph(f)} = \{(\mathbf{x}, \beta) : f(\mathbf{x}) \leq \beta\} \ .$$

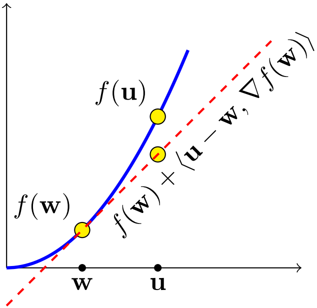


Relevant Property: tangents lie below f

If f is convex and differentiable, then

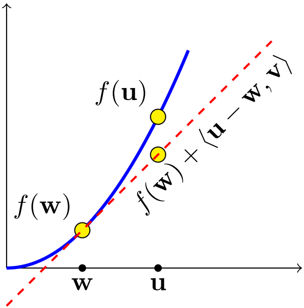
$$\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle$$

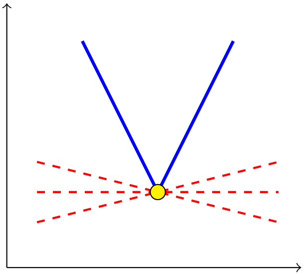
(recall, $\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \ldots, \frac{\partial f(\mathbf{w})}{\partial w_d} \right)$ is the gradient of f at $\mathbf{w}$)



Sub-gradients

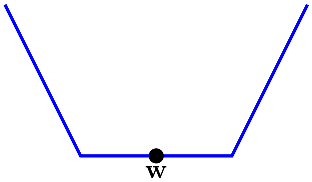
- $\mathbf{v}$ is sub-gradient of f at $\mathbf{w}$ if $\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{v}, \mathbf{u} - \mathbf{w} \rangle$
- The differential set, $\partial f(\mathbf{w})$, is the set of sub-gradients of f at $\mathbf{w}$
- Lemma: f is convex iff for every $\mathbf{w}$, $\partial f(\mathbf{w}) \neq \emptyset$





Tangents lie below f

f is "locally flat" around $\mathbf{w}$ (i.e. $\mathbf{0}$ is a sub-gradient) iff $\mathbf{w}$ is a global minimizer



Convex Calculus

Exercises:

- If $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable then f is convex iff f' is monotonically non-decreasing iff f'' is non-negative

Convex Calculus

Exercises:

- If $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable then f is convex iff f' is monotonically non-decreasing iff f'' is non-negative
- Composing convex function on linear function preserves convexity

Convex Calculus

Exercises:

- If $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable then f is convex iff f' is monotonically non-decreasing iff f'' is non-negative
- Composing convex function on linear function preserves convexity
- Max of convex functions is convex

Convex Calculus

Exercises:

- If $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable then f is convex iff f' is monotonically non-decreasing iff f'' is non-negative
- Composing convex function on linear function preserves convexity
- Max of convex functions is convex
- Positive sum of convex functions is convex

Online Convex Optimization

Game Board:

- $\mathcal{X}$: a set of contexts
- S: A convex set of vectors
- $\mathcal{F}$: A set of convex loss functions from S to $\mathbb{R}$

The Online Convex Optimization Game

For $t = 1, 2, \ldots, T$

- Environment presents a context $x_t \in \mathcal{X}$
- Learner predicts $\mathbf{w}_t \in S$
- Environment picks a loss function $f_t \in \mathcal{F}$
- Learner pays $f_t(\mathbf{w}_t)$

Outline

1. Regret
   - Relaxing the prior knowledge
   - Cover's impossibility Result
2. Online Convex Optimization
   - Convexity
   - Online Convex Optimization
   - Convexification
3. Follow The (Regularized) Leader
4. Online Gradient Descent and Online Mirror Descent
   - Linearization
   - Online Gradient Descent
   - Online Mirror Descent

Modeling Online Classification using OCO

Online Classification

For $t = 1, 2, \ldots, T$

- Environment presents a context $x_t \in \mathcal{X}$
- Learner predicts $\hat{y}_t \in \{\pm 1\}$
- Environment reveals $y_t \in \{\pm 1\}$
- Learner pays $1[\hat{y}_t \neq y_t]$

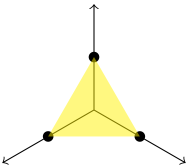
Online Convex Optimization

For $t = 1, 2, \ldots, T$

- Environment presents a context $x_t \in \mathcal{X}$
- Learner predicts $\mathbf{w}_t \in S$
- Environment picks a loss function $f_t \in \mathcal{F}$
- Learner pays $f_t(\mathbf{w}_t)$

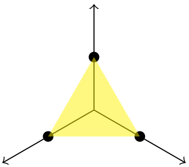
Convexification

- Suppose $\mathcal{H} = \{h_1, \ldots, h_d\}$, the state of the online convex optimizer will be a distribution over $\mathcal{H}$: $S = \{\mathbf{w} \in [0,1]^d \;:\; \|\mathbf{w}\|_1 = 1\}$



Convexification

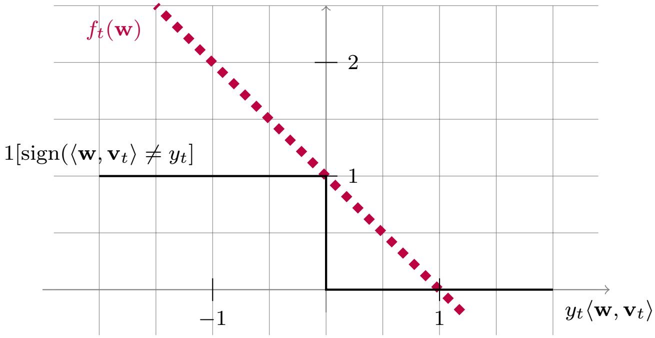
- Suppose $\mathcal{H} = \{h_1, \ldots, h_d\}$, the state of the online convex optimizer will be a distribution over $\mathcal{H}$: $S = \{\mathbf{w} \in [0,1]^d \ : \ \|\mathbf{w}\|_1 = 1\}$



- Given $x_t \in \mathcal{X}$ define $\mathbf{v}_t = (h_1(x_t), \ldots, h_d(x_t))$
- With $\mathbf{w}_t \in S$, the prediction will be:
  1. Majority: $\hat{y}_t = \mathrm{sign}(\langle \mathbf{w}_t, \mathbf{v}_t \rangle)$
  2. Random: $\mathbb{P}[\hat{y}_t = 1] = p_t := \frac{1+\langle \mathbf{w}_t, \mathbf{v}_t \rangle}{2}$

Convexification – The Loss Function

For the majority option: $f_t(\mathbf{w}) = 1 - y_t\langle \mathbf{w}, \mathbf{v}_t\rangle$



Regret for the majority option

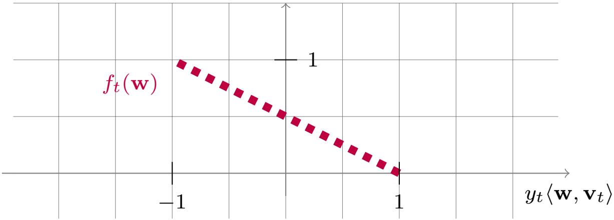
$$f_t(\mathbf{w}) = 1 - y_t\langle \mathbf{w}, \mathbf{v}_t\rangle$$

- Observe:
  1. $f_t(\mathbf{w}) \geq 1[y_t \neq \mathrm{sign}(\langle \mathbf{w}, \mathbf{v}_t\rangle)]$
  2. $\forall j,\ f_t(\mathbf{e}_j) = 2 \cdot 1[y_t \neq h_j(x_t)]$
- Therefore, vanishing regret of OCO guarantees:

$$\sum_{t=1}^{T} 1[\hat{y}_t \neq y_t] \ \leq \ 2 \ \cdot \ \min_{f \in \mathcal{H}} \sum_{t=1}^{T} 1[f(x_t) \neq y_t] \ + \ o(T)$$

Convexification – The Randomized Option

For the random option: $f_t(\mathbf{w}) = \frac{1}{2}\left(1 - y_t\langle \mathbf{w}, \mathbf{v}_t\rangle\right)\mathbb{P}_{\hat{y}_t \sim p_t}[\hat{y}_t \neq y_t]$



Regret for the Convexification by Randomization

Since singletons are in S, vanishing regret of OCO guarantees:

$$\sum_{t=1}^{T} \underset{\hat{y}_t \sim p_t}{\mathbb{P}} [\hat{y}_t \neq y_t] \ \leq \ \min_{f \in \mathcal{H}} \sum_{t=1}^{T} 1[f(x_t) \neq y_t] \ + \ o(T)$$

Outline

Follow The Leader

The most straightforward online learner:

Follow The Leader (FTL)

At each round, choose the $\mathbf{w}_t$ in S that minimizes the sum of previous loss functions:

$$\forall t, \quad \mathbf{w}_t = \operatorname*{argmin}_{\mathbf{w} \in S} \sum_{i=1}^{t-1} f_i(\mathbf{w}) \qquad \text{(break ties arbitrarily)}$$

Regret of FTL

Lemma

Let $\mathbf{w}_1, \mathbf{w}_2, \ldots$ *be the sequence of vectors produced by FTL. Then, for all* $\mathbf{u} \in S$ *we have*

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

Regret of FTL

Lemma

Let $\mathbf{w}_1, \mathbf{w}_2, \ldots$ *be the sequence of vectors produced by FTL. Then, for all* $\mathbf{u} \in S$ *we have*

$$\mathrm{Regret}_T(\mathbf{u}) = \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \le \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})) \ .$$

The lemma shows that for FTL: stability $\Rightarrow$ Low regret

Proof

Equivalent inequalities:

$$\forall \mathbf{u}, \quad \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \le \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}))$$

Proof

Equivalent inequalities:

$$\begin{aligned} & \forall \mathbf{u}, \quad \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \le \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})) \\ \iff & \forall \mathbf{u}, \quad \sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \le \sum_{t=1}^{T} f_t(\mathbf{u}) \end{aligned}$$

Proof

Equivalent inequalities:

$$
\begin{aligned}
& \forall \mathbf{u}, \quad \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \leq \sum_{t=1}^{T} (f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})) \\
\iff \quad & \forall \mathbf{u}, \quad \sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \leq \sum_{t=1}^{T} f_t(\mathbf{u}) \\
\iff \quad & \sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \leq \sum_{t=1}^{T} f_t(\mathbf{w}_{T+1})
\end{aligned}
$$

Proof (cont.)

Proof by induction on T:

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) = \left(\sum_{t=1}^{T-1} f_t(\mathbf{w}_{t+1})\right) + f_T(\mathbf{w}_{T+1})$$

Proof (cont.)

Proof by induction on T:

$$\begin{aligned}\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) &= \left(\sum_{t=1}^{T-1} f_t(\mathbf{w}_{t+1})\right) + f_T(\mathbf{w}_{T+1}) \\ &\le \left(\sum_{t=1}^{T-1} f_t(\mathbf{w}_{T})\right) + f_T(\mathbf{w}_{T+1}) \qquad \text{(inductive assumption)}\end{aligned}$$

Proof (cont.)

Proof by induction on T:

$$\begin{aligned}
\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) &= \left(\sum_{t=1}^{T-1} f_t(\mathbf{w}_{t+1})\right) + f_T(\mathbf{w}_{T+1}) \\
&\le \left(\sum_{t=1}^{T-1} f_t(\mathbf{w}_{T})\right) + f_T(\mathbf{w}_{T+1}) && \text{(inductive assumption)} \\
&\le \left(\sum_{t=1}^{T-1} f_t(\mathbf{w}_{T+1})\right) + f_T(\mathbf{w}_{T+1}) && \text{(by def. of } \mathbf{w}_T\text{)}
\end{aligned}$$

Proof (cont.)

Proof by induction on T:

$$\begin{aligned}
\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) &= \left(\sum_{t=1}^{T-1} f_t(\mathbf{w}_{t+1})\right) + f_T(\mathbf{w}_{T+1}) \\
&\le \left(\sum_{t=1}^{T-1} f_t(\mathbf{w}_{T})\right) + f_T(\mathbf{w}_{T+1}) \qquad \text{(inductive assumption)} \\
&\le \left(\sum_{t=1}^{T-1} f_t(\mathbf{w}_{T+1})\right) + f_T(\mathbf{w}_{T+1}) \qquad \text{(by def. of } \mathbf{w}_T\text{)} \\
&= \sum_{t=1}^{T} f_t(\mathbf{w}_{T+1})
\end{aligned}$$

□

Example: Success of FTL

Online Quadratic Optimisation:

- $S \subset \mathbb{R}^d$ is a convex set, and for every t, $f_t(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{z}_t\|^2$, for some $\mathbf{z}_t \in S$

Example: Success of FTL

Online Quadratic Optimisation:

- $S \subset \mathbb{R}^d$ is a convex set, and for every t, $f_t(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{z}_t\|^2$, for some $\mathbf{z}_t \in S$
- FTL rule: $\mathbf{w}_t = \mathrm{argmin}_{\mathbf{w} \in S} \sum_{i<t} \|\mathbf{w} - \mathbf{z}_i\|^2 = \frac{1}{t-1} \sum_{i=1}^{t-1} \mathbf{z}_i$

Example: Success of FTL

Online Quadratic Optimisation:

- $S \subset \mathbb{R}^d$ is a convex set, and for every t, $f_t(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{z}_t\|^2$, for some $\mathbf{z}_t \in S$
- FTL rule: $\mathbf{w}_t = \mathrm{argmin}_{\mathbf{w} \in S} \sum_{i<t} \|\mathbf{w} - \mathbf{z}_i\|^2 = \frac{1}{t-1} \sum_{i=1}^{t-1} \mathbf{z}_i$
- Stability term: by standard algebraic manipulations

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) = \left(\frac{1}{t} - \frac{1}{2t^2}\right) \|\mathbf{w}_t - \mathbf{z}_t\|^2 \leq \frac{\mathrm{diameter}(S)^2}{t}$$

Example: Success of FTL

Online Quadratic Optimisation:

- $S \subset \mathbb{R}^d$ is a convex set, and for every t, $f_t(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{z}_t\|^2$, for some $\mathbf{z}_t \in S$
- FTL rule: $\mathbf{w}_t = \mathrm{argmin}_{\mathbf{w} \in S} \sum_{i<t} \|\mathbf{w} - \mathbf{z}_i\|^2 = \frac{1}{t-1} \sum_{i=1}^{t-1} \mathbf{z}_i$
- Stability term: by standard algebraic manipulations

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) = \left(\frac{1}{t} - \frac{1}{2t^2}\right) \|\mathbf{w}_t - \mathbf{z}_t\|^2 \le \frac{\mathrm{diameter}(S)^2}{t}$$

- Since $\sum_{t=1}^{T}(1/t) \le \log(T) + 1$ we conclude:

$$\mathrm{Regret}_T(\mathbf{u}) \le \mathrm{diameter}(S)^2 \left(\log(T) + 1\right) = o(T)$$

Example: Failure of FTL

Online Linear Optimisation:

- $S = [-1, 1]$, and for every t, $f_t(\mathbf{w}) = wz_t$, where $z_t = 1$ if t is even and $z_t = -1$ if t is odd

Example: Failure of FTL

Online Linear Optimisation:

- $S = [-1, 1]$, and for every t, $f_t(\mathbf{w}) = wz_t$, where $z_t = 1$ if t is even and $z_t = -1$ if t is odd
- FTL rule: $\mathbf{w}_t = \mathrm{argmin}_{w \in S} \sum_{i<t} wz_i = \begin{cases} 1 & \text{for odd } t \\ -1 & \text{for even } t \end{cases}$

Example: Failure of FTL

Online Linear Optimisation:

- $S = [-1, 1]$, and for every t, $f_t(\mathbf{w}) = wz_t$, where $z_t = 1$ if t is even and $z_t = -1$ if t is odd
- FTL rule: $\mathbf{w}_t = \mathrm{argmin}_{w \in S} \sum_{i<t} wz_i = \begin{cases} 1 & \text{for odd } t \\ -1 & \text{for even } t \end{cases}$
- Regret is T:

$$\mathrm{Regret}_T(u = 0) = T - 0 = T$$

Example: Failure of FTL

Online Linear Optimisation:

- $S = [-1, 1]$, and for every t, $f_t(\mathbf{w}) = wz_t$, where $z_t = 1$ if t is even and $z_t = -1$ if t is odd
- FTL rule: $\mathbf{w}_t = \mathrm{argmin}_{w \in S} \sum_{i<t} wz_i = \begin{cases} 1 & \text{for odd } t \\ -1 & \text{for even } t \end{cases}$
- Regret is T:
$$\mathrm{Regret}_T(u = 0) = T - 0 = T$$
- Intuitively, FTL fails here because it is not stable

Follow The Regularized Leader (FoReL)

Follow The Regularized Leader (FoReL)

At each round, choose the $\mathbf{w}_t$ in S that minimizes the sum of previous loss functions plus regularization:

$$\forall t, \quad \mathbf{w}_t = \operatorname*{argmin}_{\mathbf{w} \in S} \sum_{i=1}^{t-1} f_i(\mathbf{w}) + R(\mathbf{w})$$

Analyzing FoReL

Lemma

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

Analyzing FoReL

Lemma

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

Proof.

Running FoReL on $f_1, \ldots, f_T$ is equivalent to running FTL on $f_0, f_1, \ldots, f_T$ where $f_0 = R$. □

Analyzing FoReL: Regularization as Stabilization

- We need to make sure that $f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})$ is small (on average)

Analyzing FoReL: Regularization as Stabilization

- We need to make sure that $f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})$ is small (on average)
- Lipschitzness: If f_t is L-Lipschitz (w.r.t. a norm $\|\cdot\|$) then:

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \;\leq\; L\,\|\mathbf{w}_t - \mathbf{w}_{t+1}\|$$

Analyzing FoReL: Regularization as Stabilization

- We need to make sure that $f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})$ is small (on average)
- Lipschitzness: If f_t is L-Lipschitz (w.r.t. a norm $\|\cdot\|$) then:

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \;\leq\; L \, \|\mathbf{w}_t - \mathbf{w}_{t+1}\|$$

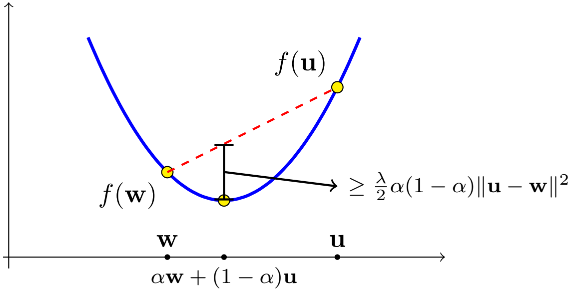
- So, it suffices that the regularizer will ensure $\|\mathbf{w}_t - \mathbf{w}_{t+1}\|$ is small

Strongly Convex Regularizers

Strongly convex function

A function f is λ-strongly convex if for all $\mathbf{w}$, $\mathbf{u}$ and $\alpha \in (0,1)$ we have

$$f(\alpha\mathbf{w} + (1-\alpha)\mathbf{u}) \le \alpha f(\mathbf{w}) + (1-\alpha) f(\mathbf{u}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2 .$$



Strongly Convex Regularizers

Lemma

Assume f is λ-strongly convex. Then:

1. *If g is convex, then $f + g$ is λ-s.c.*
2. *If f is λ-s.c. on S' and $S \subset S'$, then it is also λ-s.c. on S*
3. *If $\mathbf{u}$ is a minimizer of f, then, $\forall \mathbf{w}, \quad f(\mathbf{w}) - f(\mathbf{u}) \geq \frac{\lambda}{2}\|\mathbf{w} - \mathbf{u}\|^2$*

Strongly Convex Regularizers

Lemma

Assume f is λ-strongly convex. Then:

1. *If g is convex, then $f+g$ is λ-s.c.*
2. *If f is λ-s.c. on S' and $S \subset S'$, then it is also λ-s.c. on S*
3. *If $\mathbf{u}$ is a minimizer of f, then, $\forall \mathbf{w}$, $\quad f(\mathbf{w}) - f(\mathbf{u}) \geq \frac{\lambda}{2}\|\mathbf{w}-\mathbf{u}\|^2$*

Proof of (3):

Divide the definition of strong convexity by α and rearrange terms to get that

$$\frac{f(\mathbf{u}+\alpha(\mathbf{w}-\mathbf{u})) - f(\mathbf{u})}{\alpha} \leq f(\mathbf{w}) - f(\mathbf{u}) - \frac{\lambda}{2}(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2 .$$

Now take the limit $\alpha \to 0$.

Strongly Convex Regularizers Yield Stability

Lemma

If R is σ-strongly convex (w.r.t. $\|\cdot\|$) and f_t is L_t-Lipschitz (w.r.t. the same $\|\cdot\|$), then

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1}) \;\le\; L_t \|\mathbf{w}_t - \mathbf{w}_{t+1}\| \;\le\; \frac{L_t^2}{\sigma}$$

Proof

- Define $F_t(\mathbf{w}) = \sum_{i=1}^{t-1} f_i(\mathbf{w}) + R(\mathbf{w})$ and note that $\mathbf{w}_t = \mathrm{argmin}_{\mathbf{w}\in S} F_t(\mathbf{w})$.

Proof

- Define $F_t(\mathbf{w}) = \sum_{i=1}^{t-1} f_i(\mathbf{w}) + R(\mathbf{w})$ and note that $\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w} \in S} F_t(\mathbf{w})$.
- F_t is σ-s.c., hence

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

Proof

- Define $F_t(\mathbf{w}) = \sum_{i=1}^{t-1} f_i(\mathbf{w}) + R(\mathbf{w})$ and note that $\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w} \in S} F_t(\mathbf{w})$.
- F_t is σ-s.c., hence

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

- Repeating the same argument for F_{t+1} and its minimizer $\mathbf{w}_{t+1}$ we get

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

Proof

- Define $F_t(\mathbf{w}) = \sum_{i=1}^{t-1} f_i(\mathbf{w}) + R(\mathbf{w})$ and note that $\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w}\in S} F_t(\mathbf{w})$.
- F_t is σ-s.c., hence

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

- Repeating the same argument for F_{t+1} and its minimizer $\mathbf{w}_{t+1}$ we get

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

- Summing up, and rearranging,

$$\sigma\,\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \;\leq\; f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})$$

Proof

- Define $F_t(\mathbf{w}) = \sum_{i=1}^{t-1} f_i(\mathbf{w}) + R(\mathbf{w})$ and note that $\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w} \in S} F_t(\mathbf{w})$.
- F_t is σ-s.c., hence

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

- Repeating the same argument for F_{t+1} and its minimizer $\mathbf{w}_{t+1}$ we get

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

- Summing up, and rearranging,

$$\sigma \, \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 \;\leq\; f_t(\mathbf{w}_t) - f_t(\mathbf{w}_{t+1})$$

- Proof follows by combining with Lipschitzness and rearranging

Back to Online Classification

Recall:

- $\mathcal{H} = \{h_1, \ldots, h_d\}$, $\mathbf{v}_t = (h_1(x_t), \ldots, (h_d(x_t))$,
  $S = \{\mathbf{w} \in [0,1]^d : \|\mathbf{w}\|_1 = 1\}$
- Randomization: $f_t(\mathbf{w}) = \frac{1}{2}\left(1 - y_t \langle \mathbf{w}, \mathbf{v}_t \rangle\right)$
- Majority: $f_t(\mathbf{w}) = 1 - y_t \langle \mathbf{w}, \mathbf{v}_t \rangle$

Back to Online Classification

Recall:

- $\mathcal{H} = \{h_1, \ldots, h_d\}$, $\mathbf{v}_t = (h_1(x_t), \ldots, (h_d(x_t))$,
  $S = \{\mathbf{w} \in [0,1]^d : \|\mathbf{w}\|_1 = 1\}$
- Randomization: $f_t(\mathbf{w}) = \frac{1}{2}(1 - y_t\langle \mathbf{w}, \mathbf{v}_t\rangle)$
- Majority: $f_t(\mathbf{w}) = 1 - y_t\langle \mathbf{w}, \mathbf{v}_t\rangle$

Definition: Dual Norm Given a norm $\|\cdot\|$, its dual norm is defined as

$$\|\mathbf{v}\|_* = \max_{\mathbf{z}:\|\mathbf{z}\|\le 1} \langle \mathbf{z}, \mathbf{v}\rangle$$

Back to Online Classification

Recall:

- $\mathcal{H} = \{h_1, \ldots, h_d\}$, $\mathbf{v}_t = (h_1(x_t), \ldots, (h_d(x_t))$,
  $S = \{\mathbf{w} \in [0,1]^d : \|\mathbf{w}\|_1 = 1\}$
- Randomization: $f_t(\mathbf{w}) = \frac{1}{2}\left(1 - y_t\langle \mathbf{w}, \mathbf{v}_t\rangle\right)$
- Majority: $f_t(\mathbf{w}) = 1 - y_t\langle \mathbf{w}, \mathbf{v}_t\rangle$

Definition: Dual Norm Given a norm $\|\cdot\|$, its dual norm is defined as

$$\|\mathbf{v}\|_\star = \max_{\mathbf{z}:\|\mathbf{z}\|\le 1} \langle \mathbf{z}, \mathbf{v}\rangle$$

Lipschitzness: For an affine function, $f_t(\mathbf{w}) = \langle \mathbf{w}, \mathbf{z}\rangle + b$, we have

$$|f_t(\mathbf{w}) - f_t(\mathbf{w}')| = |\langle \mathbf{w} - \mathbf{w}', \mathbf{z}\rangle| \le \|\mathbf{w} - \mathbf{w}'\| \, \|\mathbf{z}\|_\star$$

Back to Online Classification

Recall:

- $\mathcal{H} = \{h_1, \ldots, h_d\}$, $\mathbf{v}_t = (h_1(x_t), \ldots, (h_d(x_t))$,
  $S = \{\mathbf{w} \in [0,1]^d : \|\mathbf{w}\|_1 = 1\}$
- Randomization: $f_t(\mathbf{w}) = \frac{1}{2}\left(1 - y_t \langle \mathbf{w}, \mathbf{v}_t \rangle\right)$
- Majority: $f_t(\mathbf{w}) = 1 - y_t \langle \mathbf{w}, \mathbf{v}_t \rangle$

Definition: Dual Norm Given a norm $\|\cdot\|$, its dual norm is defined as

$$\|\mathbf{v}\|_* = \max_{\mathbf{z}:\|\mathbf{z}\| \le 1} \langle \mathbf{z}, \mathbf{v} \rangle$$

Lipschitzness: For an affine function, $f_t(\mathbf{w}) = \langle \mathbf{w}, \mathbf{z} \rangle + b$, we have

$$|f_t(\mathbf{w}) - f_t(\mathbf{w}')| = |\langle \mathbf{w} - \mathbf{w}', \mathbf{z} \rangle| \le \|\mathbf{w} - \mathbf{w}'\| \, \|\mathbf{z}\|_*$$

How to choose $R(\mathbf{w})$?

Euclidean Regularization

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

Euclidean Regularization

- $R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|_2^2$
- Exercise: show that R is $(1/\eta)$-strongly convex w.r.t. $\|\cdot\|_2$

Euclidean Regularization

- $R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|_2^2$
- Exercise: show that R is $(1/\eta)$-strongly convex w.r.t. $\|\cdot\|_2$
- For every $\mathbf{u} \in S$ we have $R(\mathbf{u}) - R(\mathbf{w}_1) \leq 1/(2\eta)$

Euclidean Regularization

- $R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|_2^2$
- Exercise: show that R is $(1/\eta)$-strongly convex w.r.t. $\|\cdot\|_2$
- For every $\mathbf{u} \in S$ we have $R(\mathbf{u}) - R(\mathbf{w}_1) \leq 1/(2\eta)$
- For every t, $\|\mathbf{v}_t\|_2 = d$

Corollary

FoReL with Euclidean regularization yields the regret bound $\frac{1}{2\eta} + \eta d T$. In particular, for $\eta = (2dT)^{-1/2}$ we obtain the regret bound $\sqrt{2dT} = o(T)$

Euclidean Regularization — The resulting algorithm

Define $\mathbf{z}_t = a\,\eta\,\sum_{i=1}^{t} y_i \mathbf{v}_i$ (where a is 1 or $1/2$). Then,

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in S} \frac{1}{2\eta}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{t} f_t(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}\in S} \frac{1}{2}\|\mathbf{w}-\mathbf{z}_t\|_2^2$$

Euclidean Regularization — The resulting algorithm

Define $\mathbf{z}_t = a\,\eta \sum_{i=1}^{t} y_i \mathbf{v}_i$ (where a is 1 or $1/2$). Then,

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in S} \frac{1}{2\eta}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{t} f_t(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}\in S} \frac{1}{2}\|\mathbf{w}-\mathbf{z}_t\|_2^2$$

Exercise: Show that the solution has the form: $w_{t+1,i} = [z_{t,i} - \theta]_+$

Euclidean Regularization — The resulting algorithm

Define $\mathbf{z}_t = a\,\eta\sum_{i=1}^{t} y_i \mathbf{v}_i$ (where a is 1 or $1/2$). Then,

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in S} \frac{1}{2\eta}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{t} f_t(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}\in S} \frac{1}{2}\|\mathbf{w}-\mathbf{z}_t\|_2^2$$

Exercise: Show that the solution has the form: $w_{t+1,i} = [z_{t,i} - \theta]_+$

Interpretation: Each hypothesis in $\mathcal{H}$ gets an initial score of #correct − #wrong. We subtract θ from all scores and clamp at zero.

Entropic Regularization

- $R(\mathbf{w}) = \frac{1}{\eta} \sum_i w_i \log(w_i)$

Entropic Regularization

- $R(\mathbf{w}) = \frac{1}{\eta} \sum_i w_i \log(w_i)$
- Exercise: show that R is $(1/\eta)$-strongly convex w.r.t. $\|\cdot\|_1$

Entropic Regularization

- $R(\mathbf{w}) = \frac{1}{\eta} \sum_i w_i \log(w_i)$
- Exercise: show that R is $(1/\eta)$-strongly convex w.r.t. $\|\cdot\|_1$
- For every $\mathbf{u} \in S$ we have $R(\mathbf{u}) - R(\mathbf{w}_1) \le \log(d)/\eta$

Entropic Regularization

- $R(\mathbf{w}) = \frac{1}{\eta} \sum_i w_i \log(w_i)$
- Exercise: show that R is $(1/\eta)$-strongly convex w.r.t. $\|\cdot\|_1$
- For every $\mathbf{u} \in S$ we have $R(\mathbf{u}) - R(\mathbf{w}_1) \leq \log(d)/\eta$
- For every t, $\|\mathbf{v}_t\|_\infty = 1$

Corollary

FoReL with Entropic regularization yields the regret bound $\frac{\log(d)}{\eta} + \eta T$. In particular, for $\eta = (\log(d)/T)^{1/2}$ we obtain the regret bound $2\sqrt{\log(d)\, T}$

Entropic Regularization — The resulting algorithm

Define $\mathbf{z}_t = a\,\eta\,\sum_{i=1}^{t} y_i \mathbf{v}_i$ (where a is 1 or $1/2$). Then,

$$\begin{aligned}\mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w}\in S} \frac{1}{\eta}\sum_i w_i \log(w_i) + \sum_{i=1}^{t} f_t(\mathbf{w}) \\ &= \operatorname*{argmin}_{\mathbf{w}\in S} \sum_i w_i \log(w_i) - \langle \mathbf{w}, \mathbf{z}_t\rangle\end{aligned}$$

Entropic Regularization — The resulting algorithm

Define $\mathbf{z}_t = a\,\eta \sum_{i=1}^t y_i \mathbf{v}_i$ (where a is 1 or $1/2$). Then,

$$\begin{aligned}\mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w}\in S} \frac{1}{\eta}\sum_i w_i \log(w_i) + \sum_{i=1}^t f_t(\mathbf{w}) \\ &= \operatorname*{argmin}_{\mathbf{w}\in S} \sum_i w_i \log(w_i) - \langle \mathbf{w}, \mathbf{z}_t \rangle\end{aligned}$$

Exercise: Show that the solution is:

$$w_{t+1,i} = \frac{e^{z_{t,i}}}{\sum_j e^{z_{t,j}}}$$

Entropic Regularization — The resulting algorithm

Define $\mathbf{z}_t = a\,\eta\,\sum_{i=1}^t y_i \mathbf{v}_i$ (where a is 1 or $1/2$). Then,

$$\begin{aligned}\mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w}\in S} \frac{1}{\eta}\sum_i w_i \log(w_i) + \sum_{i=1}^t f_t(\mathbf{w}) \\ &= \operatorname*{argmin}_{\mathbf{w}\in S} \sum_i w_i \log(w_i) - \langle \mathbf{w}, \mathbf{z}_t\rangle\end{aligned}$$

Exercise: Show that the solution is:

$$w_{t+1,i} = \frac{e^{z_{t,i}}}{\sum_j e^{z_{t,j}}}$$

Interpretation: Each hypothesis in $\mathcal{H}$ gets a score of $\exp(\eta(\#\text{correct} - \#\text{wrong}))$. Then, we normalize the scores.

Outline

Linearization

- Recall that if f_t is convex then there is $\mathbf{v}_t \in \partial f_t(\mathbf{w}_t)$ such that

$$\forall \mathbf{u} \in S, \quad f_t(\mathbf{u}) \geq f_t(\mathbf{w}_t) + \langle \mathbf{v}_t, \mathbf{u} - \mathbf{w}_t \rangle$$

Linearization

- Recall that if f_t is convex then there is $\mathbf{v}_t \in \partial f_t(\mathbf{w}_t)$ such that

$$\forall \mathbf{u} \in S, \quad f_t(\mathbf{u}) \geq f_t(\mathbf{w}_t) + \langle \mathbf{v}_t, \mathbf{u} - \mathbf{w}_t \rangle$$

- Rearranging, we obtain

$$\forall \mathbf{u} \in S, \quad f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \ \leq \ \langle \mathbf{v}_t, \mathbf{w}_t - \mathbf{u} \rangle$$

Linearization

- Recall that if f_t is convex then there is $\mathbf{v}_t \in \partial f_t(\mathbf{w}_t)$ such that

$$\forall \mathbf{u} \in S, \quad f_t(\mathbf{u}) \geq f_t(\mathbf{w}_t) + \langle \mathbf{v}_t, \mathbf{u} - \mathbf{w}_t \rangle$$

- Rearranging, we obtain

$$\forall \mathbf{u} \in S, \quad f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \;\leq\; \langle \mathbf{v}_t, \mathbf{w}_t - \mathbf{u} \rangle$$

- This implies that we can define $\tilde{f}_t(\mathbf{w}) = \langle \mathbf{v}_t, \mathbf{w} \rangle$ and regret w.r.t. $\tilde{f}_t$ upper bounds the regret w.r.t. f_t

Linearization

- Recall that if f_t is convex then there is $\mathbf{v}_t \in \partial f_t(\mathbf{w}_t)$ such that

$$\forall \mathbf{u} \in S, \quad f_t(\mathbf{u}) \geq f_t(\mathbf{w}_t) + \langle \mathbf{v}_t, \mathbf{u} - \mathbf{w}_t \rangle$$

- Rearranging, we obtain

$$\forall \mathbf{u} \in S, \quad f_t(\mathbf{w}_t) - f_t(\mathbf{u}) \ \leq \ \langle \mathbf{v}_t, \mathbf{w}_t - \mathbf{u} \rangle$$

- This implies that we can define $\tilde{f}_t(\mathbf{w}) = \langle \mathbf{v}_t, \mathbf{w} \rangle$ and regret w.r.t. $\tilde{f}_t$ upper bounds the regret w.r.t. f_t
- The definition of $\tilde{f}_t$ depends on $\mathbf{w}_t$, but this doesn't matter because regret is a worst-case guarantee

Online Gradient Descent

- Suppose $S = \mathbb{R}^d$

Online Gradient Descent

- Suppose $S = \mathbb{R}^d$
- Apply the linearization trick, so $\tilde{f}_t(\mathbf{w}) = \langle \mathbf{v}_t, \mathbf{w} \rangle$ for $\mathbf{v}_t \in \partial f_t(\mathbf{w}_t)$

Online Gradient Descent

- Suppose $S = \mathbb{R}^d$
- Apply the linearization trick, so $\tilde{f}_t(\mathbf{w}) = \langle \mathbf{v}_t, \mathbf{w} \rangle$ for $\mathbf{v}_t \in \partial f_t(\mathbf{w}_t)$
- Apply FoReL with Euclidean regularization on the sequence, we have,

$$\mathbf{w}_{t+1} = -\eta \sum_{i=1}^{t} \mathbf{v}_i = \mathbf{w}_t - \eta \mathbf{v}_i$$

Analysis of Online Gradient Descent

- Since $\tilde{f}_t$ is $\|\mathbf{v}_t\|_2$-Lipschitz, we have,

$$\forall \mathbf{u} \in \mathbb{R}^d, \quad \mathrm{Regret}_T(\mathbf{u}) \;\leq\; \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta\, T \mathop{\mathbb{E}}_{t} \|\mathbf{v}_t\|_2^2$$

Analysis of Online Gradient Descent

- Since $\tilde{f}_t$ is $\|\mathbf{v}_t\|_2$-Lipschitz, we have,

$$\forall \mathbf{u} \in \mathbb{R}^d, \quad \mathrm{Regret}_T(\mathbf{u}) \;\le\; \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta\, T \operatorname*{\mathbb{E}}_t \|\mathbf{v}_t\|_2^2$$

- This doesn't yield a regret bound w.r.t. the entire $S = \mathbb{R}^d$. But, we can have a regret w.r.t. a bounded $U \subset S$

Analysis of Online Gradient Descent

- Since $\tilde{f}_t$ is $\|\mathbf{v}_t\|_2$-Lipschitz, we have,

$$\forall \mathbf{u} \in \mathbb{R}^d, \quad \mathrm{Regret}_T(\mathbf{u}) \ \leq \ \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \eta\, T \mathop{\mathbb{E}}_t \|\mathbf{v}_t\|_2^2$$

- This doesn't yield a regret bound w.r.t. the entire $S = \mathbb{R}^d$. But, we can have a regret w.r.t. a bounded $U \subset S$
- Specifically, assuming $\mathbb{E}_t \|\mathbf{v}_t\|_2^2 \leq L^2$, and setting $\eta = \mathrm{Radius}(U)/(L\sqrt{2T})$, we obtain

$$\forall \mathbf{u} \in U, \quad \mathrm{Regret}_T(\mathbf{u}) \ \leq \ \mathrm{Radius}(U)\, L\, \sqrt{2T}$$

Online Mirror Descent

Online Mirror Descent

- **parameter:** a regularization function $R : S \to \mathbb{R}$
- **initialize:** $\mathbf{z}_1 = \mathbf{0}$
- **for** $t = 1, 2, \ldots$
  - predict $\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w} \in S} R(\mathbf{w}) - \langle \mathbf{w}, \mathbf{z}_t \rangle$
  - update $\mathbf{z}_{t+1} = \mathbf{z}_t - \mathbf{v}_t$ where $\mathbf{v}_t \in \partial f_t(\mathbf{w}_t)$

Online Mirror Descent is FoReL $+$ the linearization trick
(so we don't need to analyze Online Mirror Descent)

Example I: Online Gradient Descent

- $S = \mathbb{R}^d$, $R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|_2^2$
- $\mathbf{w}_t = \mathrm{argmin}_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2\eta}\|\mathbf{w}\|_2^2 - \langle \mathbf{w}, \mathbf{z}_t \rangle = \eta \mathbf{z}_t = \mathbf{w}_{t-1} - \eta \mathbf{v}_{t-1}$

Example II: Online Gradient Descent with Lazy Projections

- $S \subset \mathbb{R}^d$, $R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|_2^2$
- $\mathbf{w}_t$ is the projection of $\eta \mathbf{z}_t$ on S:

$$\mathbf{w}_t = \operatorname*{argmin}_{\mathbf{w} \in S} \frac{1}{2\eta}\|\mathbf{w}\|_2^2 - \langle \mathbf{w}, \mathbf{z}_t \rangle = \operatorname*{argmin}_{\mathbf{w} \in S} \frac{1}{2}\|\mathbf{w} - \eta \mathbf{z}_t\|_2^2$$

Example III: Normalized Exponentiated Gradient Descent

- $S \subset \{\mathbf{w} \in [0,1]^d : \|\mathbf{w}\|_1 = 1\}$, $R(\mathbf{w}) = \frac{1}{\eta}\sum_i w_i \log(w_i)$
- Exercise: show that $\mathbf{w}_1 = (1/d, \ldots, 1/d)$ and for $t \geq 1$:

$$\forall i, \quad w_{t+1,i} \; = \; \frac{w_{t,i} e^{-\eta v_{t,i}}}{\sum_j w_{t,j} e^{-\eta v_{t,j}}}$$

Perceptron

- Recall the class of Halfspaces: $\mathcal{H} = \{\mathbf{x} \mapsto \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}$

Perceptron

- Recall the class of Halfspaces: $\mathcal{H} = \{\mathbf{x} \mapsto \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}$
- Realizability: Suppose that there exists $\mathbf{u}$ s.t. $y_t \langle \mathbf{u}, \mathbf{x}_t \rangle \geq 1$ for all t

Perceptron

- Recall the class of Halfspaces: $\mathcal{H} = \{\mathbf{x} \mapsto \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}$
- Realizability: Suppose that there exists $\mathbf{u}$ s.t. $y_t \langle \mathbf{u}, \mathbf{x}_t \rangle \geq 1$ for all t
- The Perceptron starts with $\mathbf{w}_1 = 0$ and update

$$\mathbf{w}_{t+1} = \mathbf{w}_t + 1[y_t \langle \mathbf{w}_t, \mathbf{x}_t \rangle \leq 0] \; y_t \; \mathbf{x}_t$$

Perceptron

- Recall the class of Halfspaces: $\mathcal{H} = \{\mathbf{x} \mapsto \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}$
- Realizability: Suppose that there exists $\mathbf{u}$ s.t. $y_t \langle \mathbf{u}, \mathbf{x}_t \rangle \geq 1$ for all t
- The Perceptron starts with $\mathbf{w}_1 = 0$ and update

$$\mathbf{w}_{t+1} = \mathbf{w}_t + 1[y_t \langle \mathbf{w}_t, \mathbf{x}_t \rangle \leq 0] \; y_t \; \mathbf{x}_t$$

- Theorem: The Perceptron makes at most $\|\mathbf{u}\|_2^2 \max_t \|\mathbf{x}_t\|^2$ mistakes

Proof

- Define: $f_t(\mathbf{w}) = 1[y_t \langle \mathbf{w}_t, \mathbf{x}_t \rangle \le 0] \; (1 - y_t \langle \mathbf{w}, \mathbf{x}_t \rangle)$

Proof

- Define: $f_t(\mathbf{w}) = 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\ (1 - y_t\langle \mathbf{w}, \mathbf{x}_t\rangle)$
- Exercise: Show that for Online Gradient Descent we have

$$\mathbf{w}_{t+1} = \eta \sum_{i \le t} 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0] y_i \mathbf{x}_i = \mathbf{w}_t + \eta 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\, y_t \mathbf{x}_t$$

Proof

- Define: $f_t(\mathbf{w}) = 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\ (1 - y_t\langle \mathbf{w}, \mathbf{x}_t\rangle)$
- Exercise: Show that for Online Gradient Descent we have

$$\mathbf{w}_{t+1} = \eta \sum_{i \le t} 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0] y_i \mathbf{x}_i = \mathbf{w}_t + \eta 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\, y_t \mathbf{x}_t$$

- Exercise: show that the algorithm makes the same number of mistake no matter what the value of η is

Proof

- Define: $f_t(\mathbf{w}) = 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\;(1 - y_t\langle \mathbf{w}, \mathbf{x}_t\rangle)$
- Exercise: Show that for Online Gradient Descent we have

$$\mathbf{w}_{t+1} = \eta \sum_{i \le t} 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0] y_i \mathbf{x}_i = \mathbf{w}_t + \eta 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\, y_t \mathbf{x}_t$$

- Exercise: show that the algorithm makes the same number of mistake no matter what the value of η is
- Let $\mathbf{v}_t = \nabla f_t(\mathbf{w}_t)$, denote $M = |\{t \in [T] : y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0\}|$, and $R = \max_t \|\mathbf{x}_t\|_2$.

Proof

- Define: $f_t(\mathbf{w}) = 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\ (1 - y_t\langle \mathbf{w}, \mathbf{x}_t\rangle)$
- Exercise: Show that for Online Gradient Descent we have

$$\mathbf{w}_{t+1} = \eta \sum_{i \le t} 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0] y_i \mathbf{x}_i = \mathbf{w}_t + \eta 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\, y_t \mathbf{x}_t$$

- Exercise: show that the algorithm makes the same number of mistake no matter what the value of η is
- Let $\mathbf{v}_t = \nabla f_t(\mathbf{w}_t)$, denote $M = |\{t \in [T] : y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0\}|$, and $R = \max_t \|\mathbf{x}_t\|_2$.
- Exercise: conclude the proof by relying on the following regret bound for Online Gradient Descent (we proved a slightly worse regret bound, but this one is also true):

$$\mathrm{Regret}_T(\mathbf{u}) \ \le \ \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \frac{1}{2}\,\eta\, T \mathop{\mathbb{E}}_{t} \|\mathbf{v}_t\|_2^2$$

Proof

- Define: $f_t(\mathbf{w}) = 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\ (1 - y_t\langle \mathbf{w}, \mathbf{x}_t\rangle)$
- Exercise: Show that for Online Gradient Descent we have

$$\mathbf{w}_{t+1} = \eta \sum_{i\le t} 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0] y_i \mathbf{x}_i = \mathbf{w}_t + \eta 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\, y_t \mathbf{x}_t$$

- Exercise: show that the algorithm makes the same number of mistake no matter what the value of η is
- Let $\mathbf{v}_t = \nabla f_t(\mathbf{w}_t)$, denote $M = |\{t \in [T] : y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0\}|$, and $R = \max_t \|\mathbf{x}_t\|_2$.
- Exercise: conclude the proof by relying on the following regret bound for Online Gradient Descent (we proved a slightly worse regret bound, but this one is also true):

$$\mathrm{Regret}_T(\mathbf{u}) \ \le\ \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \frac{1}{2}\,\eta\, T \mathop{\mathbb{E}}_{t} \|\mathbf{v}_t\|_2^2$$

- Hints:

Proof

- Define: $f_t(\mathbf{w}) = 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\ (1 - y_t\langle \mathbf{w}, \mathbf{x}_t\rangle)$
- Exercise: Show that for Online Gradient Descent we have

$$\mathbf{w}_{t+1} = \eta \sum_{i \le t} 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0] y_i \mathbf{x}_i = \mathbf{w}_t + \eta 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\, y_t \mathbf{x}_t$$

- Exercise: show that the algorithm makes the same number of mistake no matter what the value of η is
- Let $\mathbf{v}_t = \nabla f_t(\mathbf{w}_t)$, denote $M = |\{t \in [T] : y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0\}|$, and $R = \max_t \|\mathbf{x}_t\|_2$.
- Exercise: conclude the proof by relying on the following regret bound for Online Gradient Descent (we proved a slightly worse regret bound, but this one is also true):

$$\mathrm{Regret}_T(\mathbf{u}) \ \le\ \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \frac{1}{2}\,\eta\, T \mathop{\mathbb{E}}_t \|\mathbf{v}_t\|_2^2$$

- Hints:
  - Show that $T\,\mathbb{E}_t \|\mathbf{v}_t\|_2^2 \ \le\ M\,R^2$

Proof

- Define: $f_t(\mathbf{w}) = 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\ (1 - y_t\langle \mathbf{w}, \mathbf{x}_t\rangle)$
- Exercise: Show that for Online Gradient Descent we have

$$\mathbf{w}_{t+1} = \eta \sum_{i \le t} 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0] y_i \mathbf{x}_i = \mathbf{w}_t + \eta 1[y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0]\, y_t \mathbf{x}_t$$

- Exercise: show that the algorithm makes the same number of mistake no matter what the value of η is
- Let $\mathbf{v}_t = \nabla f_t(\mathbf{w}_t)$, denote $M = |\{t \in [T] : y_t\langle \mathbf{w}_t, \mathbf{x}_t\rangle \le 0\}|$, and $R = \max_t \|\mathbf{x}_t\|_2$.
- Exercise: conclude the proof by relying on the following regret bound for Online Gradient Descent (we proved a slightly worse regret bound, but this one is also true):

$$\mathrm{Regret}_T(\mathbf{u}) \ \le\ \frac{1}{2\eta}\|\mathbf{u}\|_2^2 + \frac{1}{2}\,\eta\, T \mathop{\mathbb{E}}_t \|\mathbf{v}_t\|_2^2$$

- Hints:
  - Show that $T\,\mathbb{E}_t \|\mathbf{v}_t\|_2^2 \ \le\ M\, R^2$
  - Observe that $\mathrm{Regret}_T(\mathbf{u}) \ge M$

Exercises

- Solve the exercises on Page 16
- Solve the exercise on Page 40
- Solve the exercise on Page 41
- Solve the exercise on Page 42. Hint:
  - Show that a function is λ-strongly convex iff for every $\mathbf{w}$ we have

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

  - Prove that if R is twice differentiable then a sufficient condition for strong convexity is that $\langle \nabla^2 R(\mathbf{w}) \mathbf{x}, \mathbf{x} \rangle \geq \lambda \|\mathbf{x}\|^2$
- Solve the exercise on Page 43
- Solve the exercise on Page 51
- Solve the exercises on Page 53