MATH FOR PREDICTION GAMES

$\mathbb{R}$ are the real numbers

symbol for "belongs to" — "imaginary" number

EXAMPLES: $3, 4.8, \pi, -6 \in \mathbb{R}$ $\quad \sqrt{-1} \notin \mathbb{R}$

---

vectors (in n dimensions)

$$u = [u_1 \quad u_2 \quad \cdots \quad u_{n-1} \quad u_n]$$

$$u - w = [u_1 - w_1 \quad \cdots \quad u_n - w_n]$$

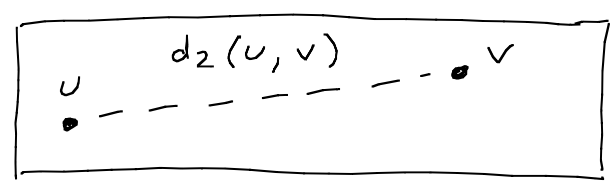
EXAMPLE: $[9 \quad 4.2 \quad 8 \quad 6 \quad 12] \in \mathbb{R}^5$ ← denotes $n = 5$ dimensions

---

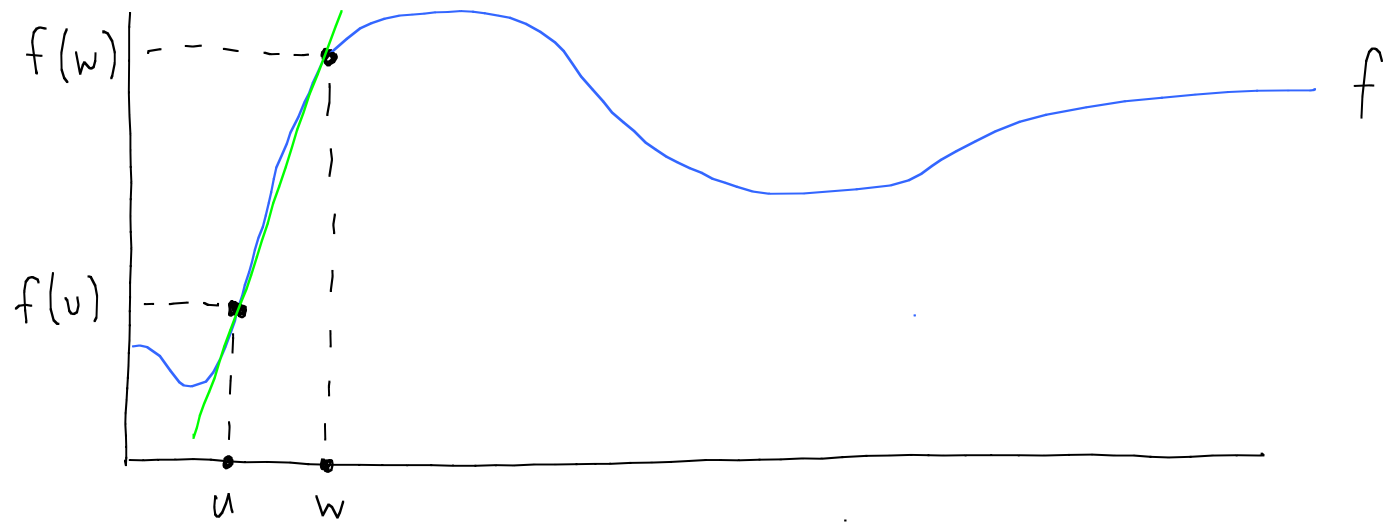
For $u, w \in \mathbb{R}^n$, the Euclidean distance between u and v is $d_2(u,v) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_n - v_n)^2}$

It generalizes the distance between $u, v \in \mathbb{R}$ $\quad |u - v|$

EXAMPLE: in the "Euclidean plane" $\mathbb{R}^2$



Let's measure how wiggly a function is.



A function f is L-Lipshitz if

$$\forall u, w \quad |f(u) - f(w)| \leq L \cdot d(u, w)$$

We'll see some concrete examples later.

A function f is linear if it can be written as

$$f(x) = \sum_{i=1}^{n} \alpha_i x_i = \langle \alpha, w \rangle$$

the slope of f

where $w = [x_1 \; x_2 \; \dots \; x_{n-1} \; x_n]$ and $\alpha = [\alpha_1 \; \alpha_2 \; \dots \; \alpha_{n-1} \; \alpha_n]$

EXAMPLE: $f(x) = 3x \qquad \alpha = 3$

$$f([x_1 \; x_2 \; x_3]) = \langle [5 \; 2 \; 7], [x_1 \; x_2 \; x_3] \rangle$$

A function f is affine if it is a "translated" linear function:

$$f(u) = \overset{f(0)}{b_0} + \langle \alpha, u \rangle$$

or more generally:

$$f(w) = b_k + \langle \alpha, w - k \rangle$$

$$f(k) = b_0 + \langle \alpha, k \rangle$$

EXAMPLE: $f(w) = 3 + 5w$

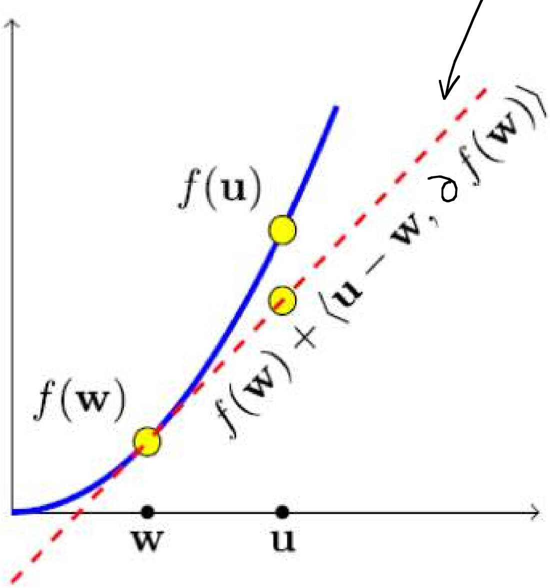
$f(7) = 3 + 5(7) = 38$

$k = 2 \qquad f(2) = 3 + 5(2) = 13$

$f(w) = 13 + 5(w - 2)$

$f(7) = 13 + 5(7 - 2) = 38$

The affine function tangent to f at w approximates f around w.



The differential "operator" ∂ takes a function f as input and returns another function ∂f (the derivative) as output.

$\partial f(w)$ is defined as the slope of the affine function tangent to f at w.

Yeah, this is a circular definition. You don't need to know how to differentiate (i.e. evaluate ∂). Just understand the diagram.

EXAMPLE: $f(u) = u^2$ $\quad \partial f(w) = 2w$ (take my word for it!)

The affine function tangent to f at $w = 3$:

$b_3 = f(3) = 3^2 = 9 \qquad \partial f(3) = 2(3) = 6$

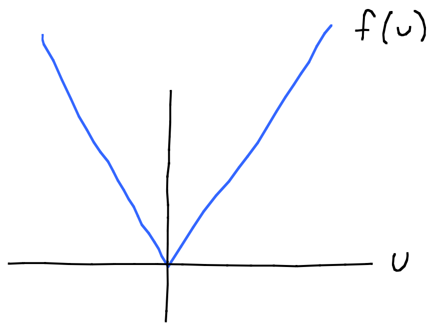
$g(u) = b_w + \langle u - w, \partial f(w) \rangle = 9 + \langle u - 3, 6 \rangle = 9 + 6(u-3)$

CHECK: $g(3) = 9 = f(3)$

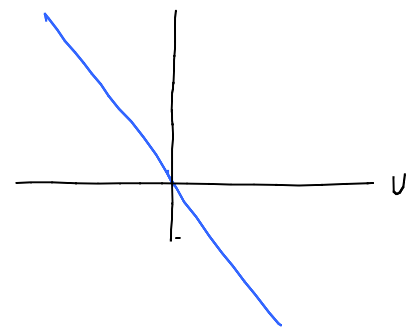
We will henceforth assume that all functions f are differentiable, which means $\partial f(w)$ is "well-defined". I'm not going to give a proper definition of this, because it doesn't matter to us. But most of the subsequent material actually applies to non-differentiable functions as well.

EXAMPLE OF NON-DIFFERENTIABILITY (OPTIONAL):

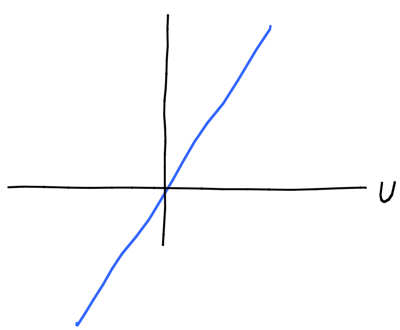
$f(u) = |u|$



For $u < 0$, tangents look like this:



For $u > 0$, tangents look like this:



What about $u = 0$?

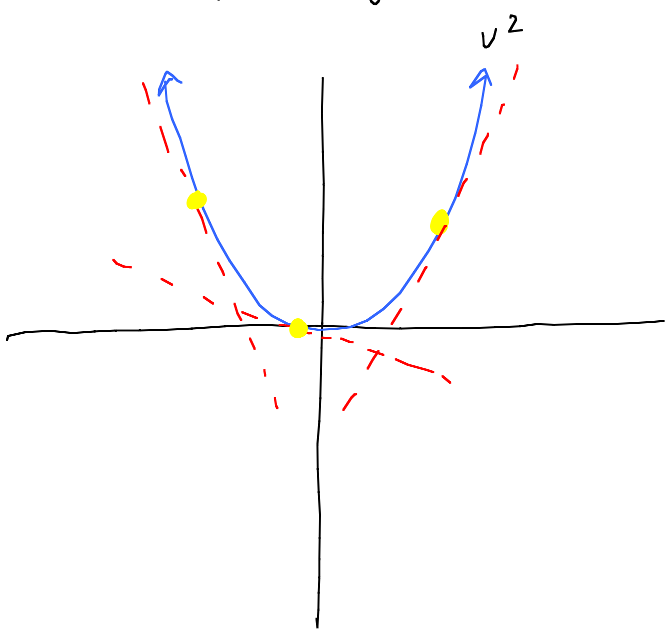
Let's not speak of such matters again.

A function f is convex if for all w, the affine function tangent to f at w lower-bounds f.

$\forall w, \forall u : f(u) \geq f(w) + \langle u - w, \partial f(w) \rangle$ SEE THE DIAGRAM ABOVE.

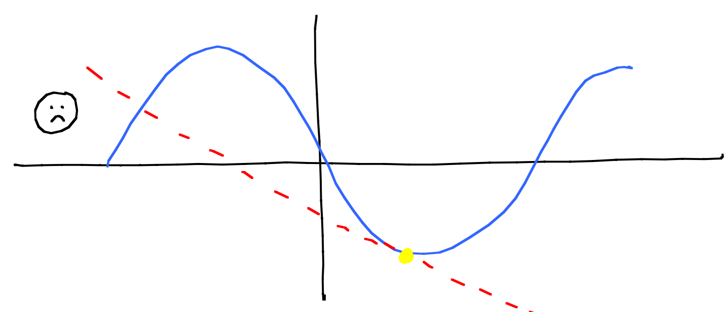
EXAMPLE: $f(u) = u^2$ is convex.

'Proof' by diagram:



Proof by contradiction:

$$u^2 < w^2 + \langle u - w, 2w \rangle$$
$$u^2 < w^2 + 2wu - 2w^2$$
$$u^2 < 2wu - w^2$$
$$u^2 - 2wu + w^2 < 0$$
$$(u+w)^2 < 0$$ which is impossible.

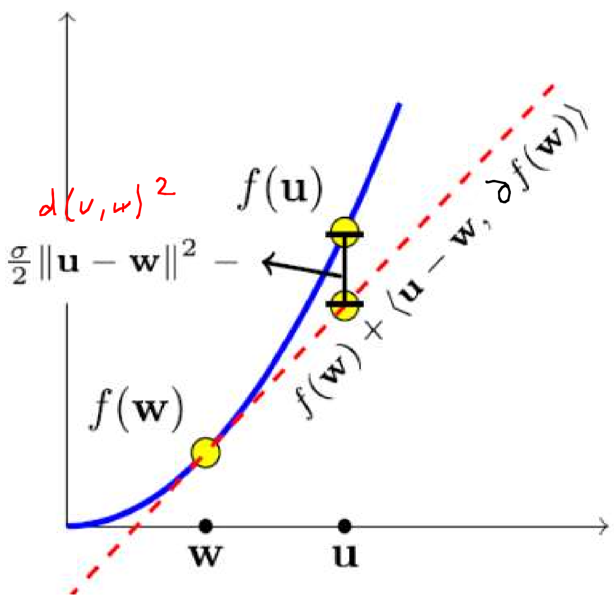


ANTI-EXAMPLE: $f(u) = \sin(u)$ is not convex.

A function f is σ-strongly convex (with respect to distance d) the above happens with a "gap":

$\sigma > 0$ is some number

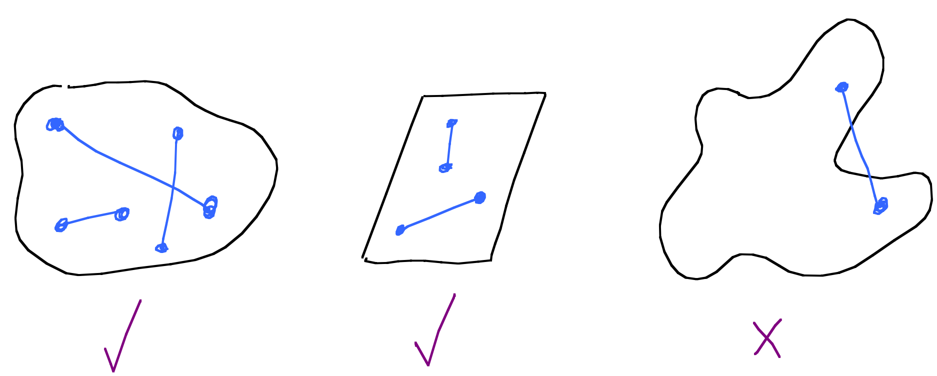
$$\forall w, \forall u : f(u) \geq f(w) + \langle u - w, \partial f(w) \rangle + \boxed{\frac{\sigma}{2} d(u, w)^2}$$



EXAMPLE: $f(u)$ is strongly convex, too. Proof omitted.

ANTI-EXAMPLE: linear functions are convex but not strongly convex.

A set S is convex if if $\forall$ endpoints $u, w \in S$, the line segment $\overline{uw}$ is fully contained in S.



PLAYING PREDICTION GAMES

reductions

upper bounds

induction

lower bounds

Movie critics [online linear regression]

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

- each critic i gives rating x_i^t
- you estimate the movie to be $s^t = \langle w^t, x^t \rangle$
- you see the movie and rate it y^t
- you penalize yourself $(y^t - s^t)^2 = f^t(w^t)$

Linearize: f^t convex, so

$$\forall w \quad f^t(w) \geq f^t(w^t) + \langle w - w^t, \partial f^t(w^t) \rangle$$

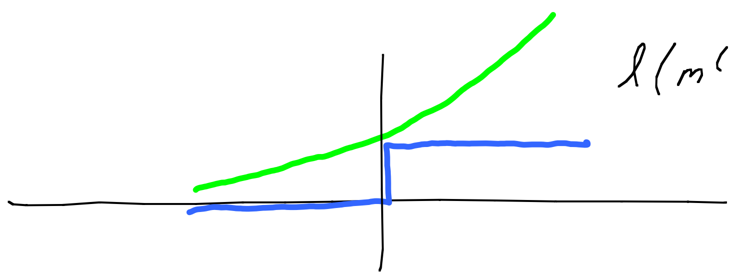
$$f^t(w^t) - f^t(w) \leq \langle w^t - w, \partial f^t(w^t) \rangle$$

$$\underbrace{\sum_t \left[f^t(w^t) - f^t(w) \right]}_{\substack{\min_w \text{ for regret against} \\ \text{convex } f^t}} \leq \sum_t \Big[\langle w^t, \underbrace{\partial f^t(w^t)}_{c^t} \rangle - \langle w, \underbrace{\partial f^t(w^t)}_{c^t} \rangle \Big]$$

... against affine functions tangent to f^t at w^t

Pro sports prediction. [online classification]

- information $x^{(t)}$ • number of injured players • home game
- $z^{(t)} = \langle w^{(t)}, x^{(t)} \rangle$
- make prediction $p^{(t)} = \mathbb{1}(z^{(t)} \geq 0)$
- outcome $y^{(t)} \in \{-1, 1\}$
- suffer $\mathbb{1}(p^{(t)} \neq y^{(t)}) = \mathbb{1}(m^{(t)} \geq 0) = \ell(m^{(t)})$ where $m^{(t)} = z^{(t)} y^{(t)}$



$$\ell(m^{(t)}) \leq \bar{\ell}(m^{(t)}) = \begin{cases} \ln\left(1 + e^{-m^{(t)}}\right) & \text{if } m^{(t)} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_t \ell(m^{(t)}) \leq \sum_t \bar{\ell}(m^{(t)})$$

pay attention to

0/1 vs -1/+1

T-round Online linear optimization [portfolio management]

- pick $w^{(t)} \in S$ [s.t. $\sum_i w_i^{(t)} = 1$ and $w_i^{(t)} \geq 0$]
- observe $c^{(t)}$ [to be the (percentage) change down]
- lose $\langle w^{(t)}, c^{(t)} \rangle$ [dollars]

Regret $$R(T) = \sum_{t=1}^{T} \langle w^{(t)}, c^{(t)} \rangle - \min_{w \in S} \sum_{t=1}^{T} \langle w, c^{(t)} \rangle$$

No regret: $R(T)/T \to 0$ as $T \to \infty$

Rock - paper - scissors [prediction with expert advice]

- choose $a \in \{\text{rock, paper, scissors}\}$ w.p. $\{w_1^t, w_2^t / w_3^t\}$
- observe outcome c^t
- suffer c_a^t

randomization → online adaptive = oblivious.

minimize $\sum_{t=1}^{T} \mathbb{E}_a \left[c_a^{(t)} \right] = \sum_{t=1}^{T} \langle w^{(t)}, c^{(t)} \rangle$

Online affine optimization

An affine function in n dimensions is a linear function in $n+1$ dimensions.

Lemma 2.1. 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

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

Proof. Subtracting $\sum_t f_t(\mathbf{w}_t)$ from both sides of the inequality and rearranging, the desired inequality can be rewritten as

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

We prove this inequality by induction. The base case of $T = 1$ follows directly from the definition of $\mathbf{w}_{t+1}$. Assume the inequality holds for $T - 1$, then for all $\mathbf{u} \in S$ we have

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

Adding $f_T(\mathbf{w}_{T+1})$ to both sides we get

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

The above holds for all $\mathbf{u}$ and in particular for $\mathbf{u} = \mathbf{w}_{T+1}$. Thus,

$$\sum_{t=1}^{T} f_t(\mathbf{w}_{t+1}) \leq \sum_{t=1}^{T} f_t(\mathbf{w}_{T+1}) = \min_{\mathbf{u} \in S} \sum_{t=1}^{T} f_t(\mathbf{u}),$$

where the last equation follows from the definition of $\mathbf{w}_{T+1}$. This concludes our inductive argument. □

Follow - the - leader

$$w^t = \operatorname*{argmin}_{w \in S} \sum_{i=1}^{t-1} \langle w^i, c^i \rangle$$

Motivates Lipshitz parameter

If f_t is L-Lipschitz with respect to a norm $\|\cdot\|$ then

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

Therefore, we need to ensure that $\|\mathbf{w}_t - \mathbf{w}_{t+1}\|$ is small.

Example 2.2 (Failure of FTL). Let $S = [-1,1] \subset \mathbb{R}$ and consider the sequence of linear functions such that $f_t(w) = z_t w$ where

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

Then, the predictions of FTL will be to set $w_t = 1$ for t odd and $w_t = -1$ for t even. The cumulative loss of the FTL algorithm will therefore be T while the cumulative loss of the fixed solution $u = 0 \in S$ is 0. Thus, the regret of FTL is T !

Balance responsiveness with stability

FoReL

$$w^t = \arg\min_w F^t(w)$$

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

Still responsive

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

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

Proof. 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

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

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

Prev thm needs to be proved for convex functions not cost vectors.

lemma shows that if the regularization function $R(\mathbf{w})$ is strongly convex with respect to the same norm, then $\mathbf{w}_t$ will be close to $\mathbf{w}_{t+1}$.

R(w) ensures stability

Lemma 2.10. Let $R : S \to \mathbb{R}$ be a σ-strongly-convex function over S with respect to a norm $\|\cdot\|$. Let $\mathbf{w}_1, \mathbf{w}_2, \ldots$ be the predictions of the FoReL algorithm. Then, for all t, if f_t is L_t-Lipschitz with respect to $\|\cdot\|$ then

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

Proof. For all t let $F_t(\mathbf{w}) = \sum_{i=1}^{t-1} f_i(\mathbf{w}) + R(\mathbf{w})$ and note that the FoReL rule is $\mathbf{w}_t = \operatorname{argmin}_{\mathbf{w} \in S} F_t(\mathbf{w})$. 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:

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

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 $\mathbf{w} = \operatorname{argmin}_{\mathbf{v} \in S} f(\mathbf{v})$. Then, for all $\mathbf{u} \in S$

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

Proof. To give intuition, assume first that f is differentiable and $\mathbf{w}$ is in the interior of S. Then, $\nabla f(\mathbf{w}) = \mathbf{0}$ and therefore, by the definition of strong convexity we have

$$\forall \mathbf{u} \in S,\ \ f(\mathbf{u}) - f(\mathbf{w}) \ge \langle \nabla f(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle + \frac{\sigma}{2}\|\mathbf{u} - \mathbf{w}\|^2 = \frac{\sigma}{2}\|\mathbf{u} - \mathbf{w}\|^2,$$

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

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

Summing the above two inequalities and rearranging we obtain

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

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 that $\|\mathbf{w}_t - \mathbf{w}_{t+1}\| \leq L/\sigma$ and together with the above we conclude our proof. □

Combining the above Lemma with Lemma 2.3 we obtain

Theorem 2.11. 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 $\mathbf{u} \in S$,

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

Corollary 2.12. 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 $R(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|_2^2$. Then, for all $\mathbf{u}$,

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

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

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

1-strongly convex
η wrt to l_2

$$w^{t+1} = -\eta \sum_{i=1}^{t} c^i$$
$$= w^t - \eta c^t$$

gradient descent

Corollary 2.14. 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 $R(\mathbf{w}) = \frac{1}{\eta}\sum_i w[i]\log(w[i])$ and with the set $S = \{\mathbf{w} : \|\mathbf{w}\|_1 = B \wedge \mathbf{w} > 0\} \subset \mathbb{R}^d$. Then,

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

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

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

$\frac{1}{\eta B}$ - strongly convex
wrt to l_1

$$w_i^{t+1} = \frac{w_i^t e^{-\eta c_i^t}}{\sum_j w_j^t e^{-\eta c_j^t}}$$

normalized exponentiated gradient

Consider an algorithm that enjoys a regret bound of the form $\alpha\sqrt{T}$, but its parameters require the knowledge of T. The doubling trick, described below, enables us to convert such an algorithm into an algorithm that does not need to know the time horizon. The idea is to divide the time into periods of increasing size and run the original algorithm on each period.

```
The Doubling Trick
```

input: algorithm A whose parameters depend on the time horizon
for $m = 0, 1, 2, \ldots$
  run A on the 2^m rounds $t = 2^m, \ldots, 2^{m+1} - 1$

The regret of A on each period of 2^m rounds is at most $\alpha\sqrt{2^m}$. Therefore, the total regret is at most

$$\begin{aligned}\sum_{m=1}^{\lceil \log_2(T) \rceil} \alpha\sqrt{2^m} &= \alpha \sum_{m=1}^{\lceil \log_2(T) \rceil} (\sqrt{2})^m \\ &= \alpha \frac{1 - \sqrt{2}^{\lceil \log_2(T) \rceil + 1}}{1 - \sqrt{2}} \\ &\le \alpha \frac{1 - \sqrt{2T}}{1 - \sqrt{2}} \\ &\le \frac{\sqrt{2}}{\sqrt{2} - 1} \alpha\sqrt{T}.\end{aligned}$$

That is, we obtain that the regret is worse by a constant multiplicative factor.