CSE599s, Spring 2014, Online Learning Lecture 4 - 04/10/2014

Convexity, Online Gradient Descent

Lecturer: Brendan McMahan *Scribe: Marco Tulio Ribeiro*

1 Review - Definitions

Here are some definitions and nomenclature that may be used interchangeably:

- $w_t \in W \to$ model, feasible point, strategy, point, play.
- $u \in W \to$ comparator.
- Regret $= \sum_t f_t(w_t) - \min_{u \in W} \sum_t f_t(u)$.
- Regret$(u) = \sum_t f_t(w_t) - f_t(u)$.
  If we bound this regret $\forall u \in W$, we've bounded the first definition of regret.

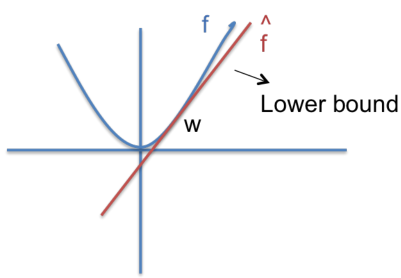
2 Convexity

2.1 Definition

Lemma 1. *For $w \subseteq \mathbb{R}^n$, $f : W \to \mathbb{R}$ is convex iff $\forall w \in W$, $\exists g \in \mathbb{R}^n$ s.t. $\forall u \in W$,*

$$f(u) \geq \underbrace{f(w) + g.(u - w)}_{\hat{f}, \textit{linear approximation to } f} . \tag{1}$$

Note that $\hat{f}$ is always a lower bound for f if f is convex, as illustrated in the figure below.



Definitions:

- g is a subgradient of f at w if inequality (1) holds.
- $\partial f(w) \to$ subdifferential of f at w, or set of subgradients.

Facts about subgradients:

- If f is differential, $\partial f(w) = \{\nabla f(w)\}$.
- $0 \in \partial f(w) \iff w \in \underset{w}{\operatorname{argmin}} f(w)$.

- For $a \in \partial f(w), b \in \partial h(w)$ and $\phi = cf(w) + dh(w)$,
  $ca + db \in \partial\phi(w)$.

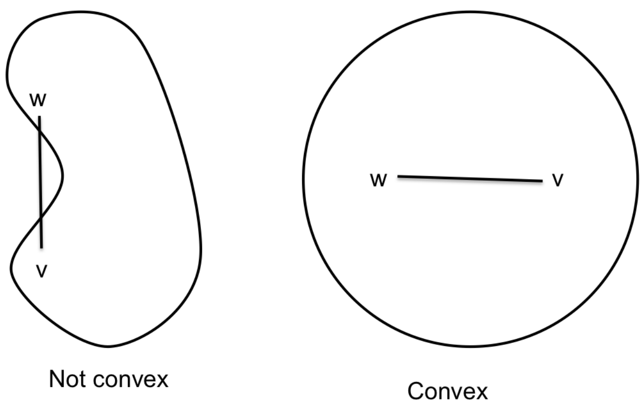
If we have a function f such that $f : W \to \mathbb{R}$, and we want to extend its tomain to be R^n (so that we can feed it to an optimization algorithm for example), it is useful to define $\hat{f}$ such that

$$\hat{f} : \mathbb{R}^n \to \{\mathbb{R}, +\infty\}$$
$$\hat{f}(w) = \begin{cases} f(w) & w \in W, \\ +\infty & otherwise. \end{cases}$$

Note that for all practical purposes (optimization), $\hat{f}$ maintains the properties of f. Particularly, if f is convex, $\hat{f}$ is also convex.

2.2 Convex set

Lemma 2. *$W \in \mathbb{R}^n$ is convex if $\forall w, v \in W$ and $\forall \alpha \in [0,1]$, $\alpha v + (1-\alpha)w \in W$*



Examples of convex sets:

- $W = \{w \mid \|w\| \leq R\} \to$ norm ball.
- $W = \{w \mid Aw \leq b\}$.

Lets say we want to minimization over a certain parameter $w \in W$, where W is a convex set. We may want to set the objective function to $\min\limits_{w \in \mathbb{R}^n}$, in order to use optimization algorithms that operate over $\mathbb{R}^n$. For this purpose, we define the following indicator function:

$$I_w(w) = \begin{cases} 0 & w \in W, \\ +\infty & otherwise. \end{cases} \tag{2}$$

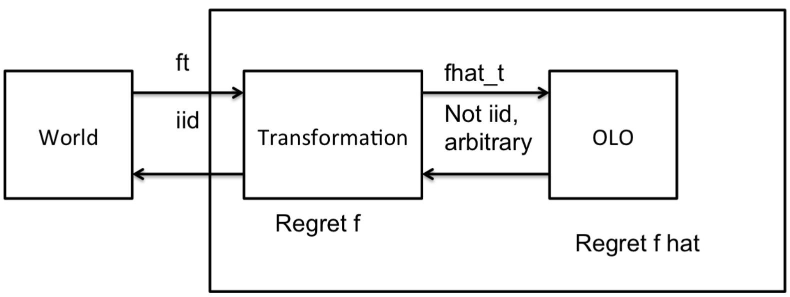
Note now that the following two optimization objectives are equivalent:

$$\min_{w \in W} f(w) = \min_{w \in \mathbb{R}^n} f(w) + I_w(w). \tag{3}$$

3 Algorithm

For $t = 1, ..., T$:

- algorithm selects w_t.
- adversary chooses f_t.
- suffer loss $f_t(w_t)$.
- Transformation:
  $\boxed{\hat{f}_t(w) = f_t(w_t) + g_t(w - w_t)}$, for $g_t \in \partial f_t(w_t)$.
- Give $\hat{f}_t$ to an algorithm for Online Linear Optimization (OLO), such as FTRL.
- w_{t+1} = output of OLO.



For this algorithm to work, we need two things:

1. $\hat{f}_t(w_t) = f_t(w_t)$ (True by definition).
2. $\forall u, \hat{f}_t(u) \leq f_t(u)$ (See convexity definition).

Note that from 2., we can get the following bound by just plugging in inequalities:

$$\underbrace{\sum_{t=1}^{T} f_t(w_t) - f_t(u)}_{\text{Regret}(u;f)} \leq \underbrace{\sum_{t=1}^{T} \hat{f}_t(w_t) - \hat{f}_t(u)}_{\text{Regret}(u;\hat{f})} \leq O(\sqrt{T}). \tag{4}$$

This bound means that the regret of the original convex function f is bounded by the regret on the modified linear function $\hat{f}$, if the same w_t is played for both at each turn. Since we have bounded the regret of linear functions with FTRL before ($O(\sqrt{T})$), this bound now holds for any convex function when we apply this algorithm.

4 Online Gradient Descent

Noting that we can write $\hat{f}_t(w) = g_t.w$ for $g_t \in \partial f_t(w_t)$, and considering the bound given on Equation 4, and the previous formulation of FTRL, we get the following algorithm:

$w_1 = 0$.
for $t \in T$:

- Observe f_t.
- Find $g_t \in \partial f_t(w_t)$.
- $w_{t+1} = w_t - \eta g_t$.

Note that this is what we would use in practice, and it is the same as applying the algorithm defined in Section 3, using FTRL as the OLO algorithm.

Corollary 3. *If all* $\|g_t\|_2 \leq G$*,* $\text{Regret}(u; f_1 ... fT) \leq \frac{1}{2\eta}||u||_2^2 + \eta T G^2$.

This regret bound comes straight from the previous analysis for FTRL on linear functions.

5 Strong convexity

Lemma 4. *A function* $f : W \to \mathbb{R}$ *is* σ *(for* $\sigma > 0$*) strongly convex w.r.t. norm* $\|\cdot\|$ *if* $\forall w \in W; \forall g \in \partial f(w); \forall u \in W$*:*

$$f(u) \geq f(w) + g.(u - w) + \frac{\sigma}{2} \|u - w\|^2 . \tag{5}$$