CSE599s, Spring 2012, Online Learning Lecture 4 - 04/05/2012

Linearizing Convex Loss

Lecturer: Brendan McMahan, Ofer Dekel *Scribe: Hossein Safavi*

1 Follow the Leader Against Quadratics

In some cases FTL can be a very good algorithm. An example of such a case is a 1-D linear regression that can be generalized to the n-dimensional case.

- on each round, we chose $w_t \in [-1, 1]$
- adversary reveals $y_t \in [-1, 1]$
- we pay $f_t(w) = \frac{1}{2}(w - y_t)^2$

Applying FTL:

$$w_{t+1} = \underset{w \in [-1,1]}{\operatorname{argmin}} f_{1:t}(w) \tag{1}$$

$$f_t(w) = \frac{1}{2}(w - y_t)^2 = \frac{1}{2}w^2 - wy_t + \underbrace{\frac{1}{2}y_t^2}_{\text{constant}} \tag{2}$$

Without loss of generality we have:

$$f_t(w) = \frac{1}{2}w^2 - wy_t \tag{3}$$

$$f_{1:t}(w) = \underbrace{\frac{t}{2}w^2 - y_{1:t}w}_{\text{objective to minimize to chose } w_{t+1}} \tag{4}$$

We solve by differentiating (4) and setting to zero. Optimal solution is then:

$$w_{t+1} = \frac{y_{1:t}}{t} \tag{5}$$

Lemma 1. *f is a convex function $\Rightarrow \forall w, w_t \in \mathcal{W}$:*

$$f(w) \geq f(w_t) + \nabla f(w_t)(w - w_t) \tag{6}$$

$$\begin{aligned} f_t(w_t) - f_t(w_{t+1}) &\leq \nabla f_t(w_t)(w_t - w_{t+1}) && (7) \\ &\leq \underbrace{\|\nabla f_t(w_t)\|}_{\textit{bound on gradient}} \overbrace{\|w_t - w_{t+1}\|}^{\textit{bound on distance}} \textit{ (Cauchy-Schwarz)} && (8) \end{aligned}$$

Now, we compute the bound on the regret in the quadratic case.

$$\begin{aligned} w_t - w_{t+1} &= \frac{y_{1:t-1}}{t-1} - \left(\frac{y_{1:t-1}}{t-1} \cdot \frac{t-1}{t} + \frac{y_t}{t}\right) && (9) \\ &= \frac{y_{1:t-1}}{t(t-1)} - \frac{y_t}{t} \leq \frac{t-1}{t(t-1)} + \frac{1}{t} \leq \frac{2}{t} && (10) \end{aligned}$$

$$Regret \leq \sum_t f_t(w_t) - f_t(w_{t+1}) \leq \sum_t 2(\frac{2}{t}) \leq 4\sum_{t=1}^{T} \frac{1}{t} \leq 4(log(T) + 1) \tag{11}$$

Conclusion: On quadratic functions, the FTL algorithm is very good. Linear functions are in some sense hardest for online convex optimization.

2 Convexity

Definition 2. *A set $\mathcal{W} \subseteq \mathbb{R}^n$ is convex if $\forall w, w' \in \mathcal{W}$ $\forall \alpha \in [0, 1]$ it holds that:*

$$\alpha w + (1 - \alpha)w' \in \mathcal{W} \tag{12}$$

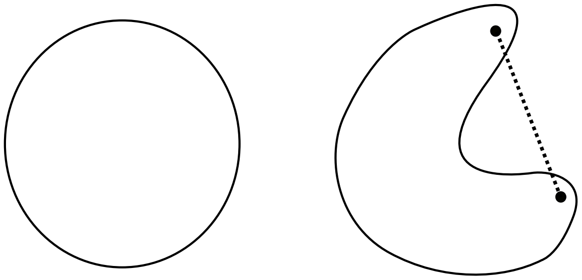

Figure 1: Example of a convex set(left) and a set that is not convex (right)

Definition 3. *A function $f : \mathcal{W} \to \mathbb{R}$ is convex if $\mathcal{W}$ is convex and if $\forall w, w' \in \mathcal{W}$ $\forall \alpha \in [0, 1]$ it holds that:*

$$\alpha f(w) + (1 - \alpha)f(w') \geq f(\alpha w + (1 - \alpha w') \tag{13}$$

Theorem 4. *A convex function has convex level sets.*

Note that the converse of the above theorem is not true. However, a function with convex level sets is *quasi convex*. An example of such a function is $f(w) = \sqrt{|w|}$.

Definition 5. *$g \in \mathbb{R}^n$ is a subgradient of the convex function $f : \mathcal{W} \to \mathbb{R}$ at the point $\hat{w} \in \mathcal{W}$ if $\forall w \in \mathcal{W}$:*

$$f(w) \geq \underbrace{f(\hat{w}) + (w - \hat{w}) \cdot g}_{first\ order\ tailor\ expansion\ of\ f\ at\ \hat{w}} \tag{14}$$

It should be noted that it is possible to have multiple subgradients at one point (figure 2).

Definition 6. *The subdifferential, denoted $\partial f(\hat{w})$, is defined as:*

$$\partial f(\hat{w}) = \{g : g\ is\ a\ subgradient\ of\ f\ at\ \hat{w}\} \tag{15}$$

The subdifferential itself is a convex set and can have 0, 1, or ∞ elements.

Theorem 7. *The function $f : \mathcal{W} \to \mathbb{R}$ is convex iff $\forall w$ in the relative interior of $\mathcal{W}$ the subdifferential is non-empty.*

In other words, for convex functions, there is at least 1 subgradient at every point.

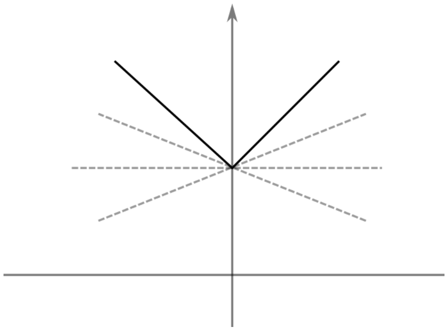

Figure 2: 3 subgradients shown for a point on the function

3 Playing Against Linear Functions

Now, recall the problems of FTRL:

1. Finite Horizon: The algorithm itself needs to know T. Solved using the doubling trick.
2. Focused on class of linear cost functions. Since cost is highest with linear functions, convex functions will do better.
3. The feasible set is $\mathcal{W} = \mathbb{R}^n$.

The question arises, what if the feasible set is constrained?

3.1 Linearizing Convex Functions

Linearizing a convex loss makes online learning more difficult. For any w^*,$g_t \in \partial f_t(w_t)$ specifically:

$$\begin{aligned} w^* &= \underset{w \in \mathcal{W}}{\text{argmin}} f_{1:t}(w^*) && (16) \\ Regret(T) &= f_{1:t}(w_t) - f_{1:t}(w^*) && (17) \\ &= \sum_{t=1}^{T} [f_t(w_t) - f_t(w^*)] \leq \sum_{t=1}^{T} [w_t \cdot g_t - w^* \cdot g_t] && (18) \end{aligned}$$

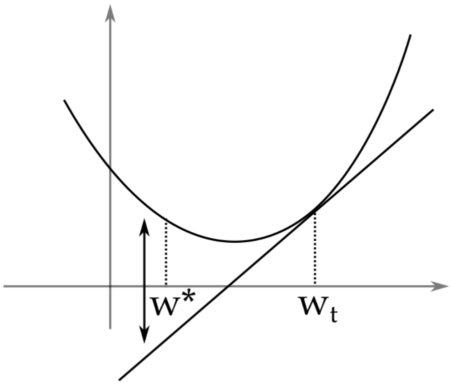

Figure 3: the cost to point w^* is lowered significantly due to linearization

As illustrated in figure 3, using the linear approximation, the loss of all other actions is decreased thereby increasing the regret for the player. Therefore, linearized regret is an upper bound for convex regret. Using this fact, by transforming the convex online learning problem to a linear online learning problem, we can have bounded regret. While in the linear case we cannot do better than $\sqrt{T}$ regret, in the case of a strongly convex function, we can do better.

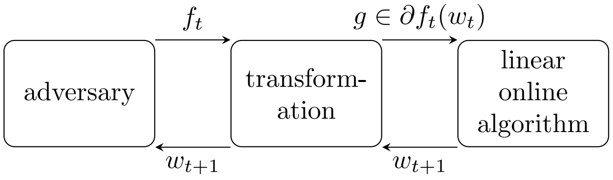

Figure 4: transforming online convex to online linear

Conclusion: Follow the Regularized Leader using the regularization $r(w) = \frac{\sigma}{2}\|w\|^2$ and the doubling trick enjoys an infinite horizon regret of $O(\sqrt{T})$ against an adversary that uses convex functions with bounded subgradients.

Definition 8. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is G-Lipschitz if $\forall w, w' \in \mathbb{R}^n$:*

$$|f(w) - f(w')| \leq G\|w - w'\| \tag{19}$$

Theorem 9. *For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, the function is G-Lipschitz iff $\forall w \in \mathbb{R}^n$ $\forall g \in \partial f_t(w_t)$, $|g| < G$.*

Proof. ($\Leftarrow$)
Choose $w \in \mathcal{W}$, $g \in \partial f_t(w_t)$.

$$G \underbrace{\|w - w - \frac{g}{\|g\|}\|}_{1} \geq f(w + \frac{g}{\|g\|}) - f(w) \tag{20}$$

using the definition of the subgradient:

$$f(w + \frac{g}{\|g\|}) - f(w) \geq (w + \frac{g}{\|g\|} - w)g = \frac{g \cdot g}{\|g\|} = \|g\| \tag{21}$$

($\Rightarrow$)
Choose $w' \in \mathcal{W}$, $g \in \partial f_t(w_t)$

$$f(w') - f(w) \leq (w' - w) \cdot g \leq \underbrace{\|w' - w\| \cdot \|g\|}_{\text{Cauchy-Schwarz}} \leq \|w' - w\|G \tag{22}$$

□

Conclusion: We see that FTRL using the regularization $r(w) = \frac{\sigma}{2}\|w\|^2$ and the doubling trick enjoys an infinite horizon regret of $O(\sqrt{T})$ against an adversary that uses convex *Lipschitz* functions.