CSE599s, Spring 2014, Online Learning Lecture 5 - 04/15/2014

FTRL with Arbitrary Strongly Convex Regularization and Experts

Lecturer: Brendan McMahan and Ofer Dekel *Scribe: Christopher Lin*

1 Strong Convexity

A function f is $\sigma-$*strongly convex* with respect to a norm $||\cdot||$ iff for all $w \in W$ and $g \in \partial f(w)$, we have that for all u, $f(u) \geq f(w) + g(u-w) + \frac{\sigma}{2}||u-w||^2$.

1.1 Properties

- Let f_1, f_2 be functions that are σ_1-, σ_2- strongly convex. If for some $a, b \geq 0$, $f_3 = af_1(w) + bf_2(w)$, then f_3 is $a\sigma_1 + b\sigma_2$ strongly convex.
- Let $w^* = \text{argmin}_w f(w)$, where f is $\sigma-$strongly convex. Then $f(w) - f(w^*) \geq \frac{\sigma}{2}||w - w^*||^2$, by the fact that $0 \in \partial f(w^*)$ and the definition of strong convexity.

1.2 Examples

- $R(w) = \frac{\sigma}{2}||w||_2^2$ is strongly convex. It has a quadratic lower bound that is tight at every chosen point.
- $R(w) = \frac{\sigma}{2}||w||_2^2 + I_W(w)$ is strongly convex where I_W is the indicator function on a convex set W such that $I_W(w) = 0$ when $w \in W$ and ∞ otherwise.

2 Norms

A *norm* is a function $||\cdot||$ such that for all $w \in \mathbb{R}^n$, we have

- $||w|| \geq 0$
- $||w|| = 0$ iff $w = 0$
- For all w and for all $a \in \mathbb{R}$, $||aw|| = |a| \; ||w||$
- For all u, w, $||u + w|| \leq ||u|| + ||w||$.

2.1 Examples

- The *l2-norm* L2 is $||w||_2 = \sqrt{\sum_{i=1}^n w_i^2}$.
- The *l1-norm* L1 is $||w||_1 = \sum_i |w_i|$
- The L_∞ norm is $||w||_\infty = \max_i |w_i|$.

We have $||w||_1 \geq ||w||_2 \geq ||w||_\infty$. Let $B_p = \{w : ||w||_p \leq 1\}$. Then, $p = 1$ is the unit diamond, $p = 2$ is the unit circle, $p = \infty$ is the unit square.

2.2 Dual Norm

Given an arbitrary norm $||\cdot||$, the *dual norm* $||\cdot||_*$ is $||g||_* = \max_{w:||w||\leq 1} wg$. The dual norm is a norm, and the dual of $||\cdot||_*$ is $||\cdot||$.

Holder's inequality: Let $a, b \in \mathbb{R}^n$. Then, $a\cdot b \leq \max_{w:||w||\leq 1}(a\cdot||b||w) = ||b|| \max_{w:||w||\leq 1} a\cdot w = ||b||||a||_*$. (Let $w = b/||b||$.)

3 Analyzing FTRL with arbitrary strongly convex regularizations

FTRL selects $w_t = \text{argmin}_w \sum_{s=1}^{t-1} f_s(w) + R(w)$ where R is $\sigma-$strongly convex with respect to the norm. Define $F_t(w) = f_{1:t-1}(w) + R(w)$ for convenience. Using the definition and the fact that the sum of a convex and a strongly convex function is strongly convex, we have

$$F_t(w_{t+1}) - F_t(w_t) \geq \frac{\sigma}{2}||w_{t+1} - w_t||^2$$

and

$$F_{t+1}(w_t) - F_{t+1}(w_{t+1}) \geq \frac{\sigma}{2}||w_{t+1} - w_t||^2.$$

Summing the inequalities, we get $f_t(w_t) - f_t(w_{t+1}) \geq \sigma||w_{t+1} - w_t||^2$. Then, apply the definition of convexity to f_t to get $f_t(w_t) - f_t(w_{t+1}) \leq g_t(w_t - w_{t+1}) \leq ||g_t||_*||w_t - w_{t+1}||$ for some $g_t \in \partial f_t(w_t)$. Then, $||w_t - w_{t+1}|| \leq ||g_t||_*/\sigma$. Plugging this bound into the above bound, we get $f_t(w_t) - f_t(w_{t+1}) \leq ||w_t - w_{t+1}||||g_t||_* \leq \frac{1}{\sigma}||g_t||_*^2$.

Theorem FTRL, with a σ-strongly convex R, arbitrary convex f_t. Then for all $u \in \mathbb{R}^n$,

$$\text{Regret}(u) \leq R(u) + \frac{1}{\sigma}\sum_{t=1}^{T}||g_t||_*^2.$$

4 Recap

Optimization is when you optimize one single function. Statisical Machine Learning is when the functions are sampled from a distribution. Online Learning is when the functions are totally arbitrary, which is much more awesome.

5 Online Learning With Expert Advice

We play the following game. For $t = 1, \ldots, T$:

- Receive input from d experts
- Choose one expert and follow his advice
  - Specifically, choose a distribution $p_t \in \Delta_d = \{p \in \mathbb{R}^d, p_i \geq 0, \sum_{i=1}^d p_i = 1\}$.
  - draw I_t (the index of the expert you listen to) from p_t.
- Observe the loss of each expert $\ell_{t,1}, \ldots, \ell_{t,d} \in [0,1]^d$
- Incur loss ℓ_{t,I_t}.

The *cumulative expected loss* is $E[\sum_{t=1}^{T} \ell_{t,I_t}]$. The oblivious adversary defines all the losses ahead of time. The regret is a comparison to the best fixed expert in hindsight, defined to be $E[\sum_{t=1}^{T} \ell_{t,I_t}] - \min_{i\in\{1,\ldots,d\}} \sum_{t=1}^{T} \ell_{t,i}$. Note that we compare to the performance of a single expert, rather than to an arbitrary convex combination of experts. In game theory, the former is called a pure strategy, while the latter is called a mixed strategy. In this case, the best pure strategy is just as good as the best mixed strategy, and there is no advantage to taking combinations of experts.

We see that experts is a special case of online convex optimization:

- $E[\ell_{t,I_t}] = \sum_{i=1}^{d} p_{t,i}\ell_{t,i} = p_t\ell_t$. So we let $f_t(p) = p \cdot \ell_t$ (linear loss functions).
- Regret $\le \sum_{t=1}^{T} p_t \cdot \ell_t - \min_{p\in\Delta_d} \sum_{t=1}^{T} p \cdot \ell_t$.

Previously proved theorem: Let $f_1, \ldots, f_T$ be the convex functions, and we play $w_1, \ldots, w_T \in \mathbb{R}^n$ generated with FTRL with $R(w)$ ($\min R(w) = 0$), and $g_1, \ldots, g_T$ are subgradients $g_t \in \partial f_t(w_t)$. For any norm $||\cdot||$, let σ, G be constants such that for all t, $||g_t||_* \le G$ and R is σ-strongly convex with respect to $||\cdot||$, then the Regret$(u) \le R(u) + TG^2/\sigma$.

5.1 First attempt to solve experts

Run FTRL with $R(p) = \frac{1}{2\eta}||p||_2^2 + I_{\Delta_d}(p)$. Apply theorem with $||\cdot||_2$:

- The gradient is just the vector of losses, $g_t = \ell_t$. So $G = \max_{t=1,\ldots,T} ||\ell_t||_2 \le \sqrt{d}$. This bound G is very slack - we are bounding a quarter of the unit square with a unit circle.
- $R(p)$ is $\frac{1}{\eta}$-strongly convex with respect to $||\cdot||_2$ because $\frac{1}{2}||w||_2^2$ is 1-strongly convex with respect to $||w||_2^2$.

Then, we have that Regret$(q) \le \frac{1}{2\eta}||q||_2^2 + I_{\Delta_d}(q) + Td\eta = \frac{1}{2\eta} + Td\eta$. Since the best η is $\frac{1}{\sqrt{2dT}}$, we get a regret upper bounded by $\sqrt{2dT}$.