Lecture 13 Follow the Regularized Leader

Recall OCO:

Player declares Alg A
Adversary chooses $f_1 \cdots f_T$
for $t = 1, 2, \cdots, T$:
1. player plays $x_t \in K$
2. Incur a loss of $f_t(x_t)$
3. receive feedback $\begin{cases} \nabla f_t(x_t) \leftarrow \text{Full Information} \\ f_t(x_t) \leftarrow \text{Bandits} \\ f_t \leftarrow \text{Function access} \end{cases}$

$$\text{Regret}(A) = \sum_{t=1}^{T} f_t(x_t) - \inf_u \sum_{t=1}^{T} f_t(u)$$

Algorithm = Online Gradient Descent (Sub)

$$x_{t+1} = \Pi_K(x_t - \eta_t \nabla f_t(x_t))$$

Thm: $\eta_t = \frac{D}{G\sqrt{T}}$

$\|\nabla f_t(x)\|_2 \le G$, $\forall f_t, \forall x$ $\|x - y\|_2 \le D$ $\forall x, y \in K$

$\text{Regret}_T \le \frac{3}{2} GD\sqrt{T}$

if f_t is m-strongly convex

$\text{Regret} \le \frac{G^2}{m} \log_2(1+T)$

Today: a different kind of algorithm.

FTL: Follow the leader

$$x_{t+1} = \arg\min_x \sum_{\tau=1}^{t} f_\tau(x)$$

$K = [-1, 1]$ $f_1 = \frac{1}{2}x$, $f_2 = -x$, $f_3 = x, \cdots$

$f_4 = -x$, $f_5 = x \cdots$

$$\sum_{\tau=1}^{t} f_\tau(x) = \begin{cases} \frac{1}{2}x & \text{when } t \text{ is odd} \\ -\frac{1}{2}x & \text{when } t \text{ is even} \end{cases}$$

FTL will output $0, -1, 1, -1, 1, -1, 1 \cdots$

$$\text{Regret} \le 0 + 1 + 1 + 1 + 1 + \cdots + 1 - \sum_{t=1}^{T} f_t(u)$$

$$= T - 1$$

Learning with expert advice

$f_t = \langle x_t, r_t \rangle$ $\|r_t\|_\infty \le 1$

Randomized Multiplicative Weight (RMW) or Hedge Alg.

Initialize $\forall i \in [N]$, $X_1(i) = 1$
For $t = 1 \cdots T$:
1. Pick $i_t \sim \text{Prob}(i) = \frac{X_t(i)}{\sum X_t(i)}$
2. Incur loss $\ell_t(i_t)$
3. Update $X_{t+1}(i) = X_t(i) e^{-\varepsilon \ell_t(i)}$

Exponentiated Gradient (GD in log space)

$$\log(X_{t+1}) = \log(X_t) - \varepsilon r_t$$

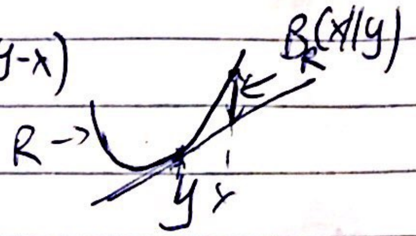
Stability is what we want (OGD is stable because learning rate is small)

Induce stability by ① Randomization Follow the Perturbed Leader (Kalai and Vempala 2005 (Hannan, 1957))

② By regularization: $R(x): K \to \mathbb{R}$

- twice differentiable
- Smooth and strongly convex on K

Bregman Divergence: $B_R(x||y) = R(x) - R(y) - \nabla R(y)^T(y-x)$



$D_R = \sqrt{\max_{x,y \in K} \{R(x) - R(y)\}}$

$B_R(x||y) = \frac{1}{2}\|x-y\|_z^2$ for some $z \in [x,y]$

Taylor's theorem (~~Mean~~ Mean value thm)

$R(x) = R(y) + \langle \nabla R(y), x-y \rangle + \underbrace{\frac{1}{2}(x-y)^T \nabla^2 R(z)(x-y)}_{\text{Bregman Divergence}}$

Example1 $R(x) = \frac{1}{2}\|x\|^2$

$\nabla^2 R(x) = I \quad B_R(x||y) = \|x-y\|^2$

Dual norm $\|\cdot\|_*$

$\|y\|_* = \max_{\|x\| \le 1} \langle y, x \rangle$

Hölder: $\langle x, y \rangle \le \|x\| \|y\|_*$

$\|x\|_y := \|x\|_{\nabla^2 R(y)} = x^T \nabla^2 R(y) x$

$\|x\|_y^* = x^T \nabla^{-2} R(y) x$

$B_R(x_t || x_{t+1}) = \frac{1}{2}\|x_t - x_{t+1}\|_t^2$

Example2: $R(x) = \langle x, \log x \rangle$ for $0 < x \le 1$, $\sum x_i = 1$

$B_R(x||y) = \sum_i x_i \log\left(\frac{x_i}{y_i}\right) = KL(x||y)$

Generalized projection

$\tilde{x} = \arg\min_x B_R(x||y)$

By convexity: $T f(x^*) \ge \sum_{t=1}^{T} f(x_t) + \sum_{t=1}^{T} \langle x^* - x_t, \nabla_t \rangle$

$\Leftrightarrow$ Regret $\le \sum_{t=1} \nabla_t^T(x_t - x^*) \le \sum_t \nabla_t^T(x_t - u)$

(* linear loss is the worst loss in OCO)

Alg. FTRL. Input parameter $\eta > 0$, R, K

1. let $x_1 = \arg\min_{x \in K} R(x)$
2. For $t = 1, 2, 3, \cdots T$:
   a. Predict x_t.
   b. observe $\nabla_t = \nabla f_t(x)$
   c. update $x_{t+1} = \arg\min_{x \in K} \left\{ \eta \sum_{s=1}^{t} \nabla_s^T x + R(x) \right\}$

Scanned with CamScanner

~~When $R(x) = \frac{1}{2}\|x\|^2$~~ or $\arg\min_{x \in K} \left\{ \eta \sum_{t=1}^{T} f_t(x_t) + R(x) \right\}$

Example $R = \frac{1}{2}\|x\|^2$. this is Euclidean Projection to K

$R = x^T \log x$. this is Multiplication update

$$\min \ \eta \sum_{t=1}^{T} l^T x + \sum x_i \log x(i)$$
s.t. $x \geq 0$
$\sum x = 1$

$$L(x, u, v) = \eta \sum_{t=1}^{T} l_t^T x + \sum x_i \log x_i + U^T(-x) + V^T(\sum x - 1)$$

$$\boxed{\text{Thm: } Regret \leq 2\eta \sum_{t=1}^{T} \|\nabla_t\|_t^{*2} + \frac{R(u) - R(x_1)}{\eta}}$$
u can be x^*, or anything else

i.e. $Regret \leq 2 D_R G_R \sqrt{2T}$

* Same Regret but different Geometry

Be the leader: Play $x_{t+1} = \arg\min_x \sum_{t=1}^{t} f_t(x_t) = \arg\min_x \sum_{t=1}^{t} \langle \nabla_t, x_t \rangle$ instead of x_t

Lemma (Regret) $\forall u \in K$

$$Regret \leq \sum_{t=1}^{T} \nabla_t^T x_t - \sum_{t=1}^{T} \nabla_t^T x_{t+1} + \frac{1}{\eta} D_R^2$$

Proof: $g_0(x) = \frac{1}{\eta} R(x)$, $g_t(x) = \nabla_t^T x$. $Regret = \sum_{t=1}^{T} \left(g_t(x_t) - g_t(u)\right)$

Claim: $\sum_{t=0}^{T} g_t(u) \geq \sum_{t=0}^{T} g_t(x_{t+1})$

$$\sum_{t=1}^{T} g_t(x_t) - g_t(u) = \sum_{t=1}^{T} g_t(x_t) - g_t(x_{t+1}) + g_t(x_{t+1}) - g_t(u)$$
$$= \sum_{t=1}^{T} (g_t(x_t) - g_t(x_{t+1})) + \sum_{t=0}^{T} (g_t(x_{t+1}) - g_t(u)) - g_0(x_1) + g_0(u)$$
$$\leq \sum_{t=1}^{T} (g_t(x_t) - g_t(x_{t+1})) + \frac{1}{\eta} D_R^2$$

Prof of $\sum_{t=0}^{T} g_t(u) \geq \sum_{t=0}^{T} g_t(x_{t+1})$ by induction

① When $T=0$ $g_0(u) \geq g_0(x_1)$ but $x_1 = \arg\min_x R(x)$

② if for $T=T'$ $\sum_{t=0}^{T'} g_t(u) \geq \sum_{t=0}^{T'} g_t(x_{t+1})$

Then Consider $x_{T'+2} = \arg\min_x \sum_{t=0}^{T'+1} g_t(x)$

$$\sum_{t=0}^{T'+1} g_t(x_{t+1}) = \sum_{t=0}^{T'} g_t(x_{t+1}) + g_{T'+1}(x_{T'+2})$$

$$\leq \sum_{t=0}^{T'} g_t(V) + g_{T'+1}(x_{T'+2})$$

Take $V = T'+2$

$$\Rightarrow = \sum_{t=0}^{T'+2} g_t(x_{T'+2})$$

$$\leq \sum_{t=1}^{T'+2} g_t(u)$$

Prof of Thm:

$$\phi_t(x) := \eta \sum_{t=1}^{t} \nabla_t^T x + R(x)$$

Taylor's thm at x_{t+1}

$$\phi_t(x_t) = \phi_t(x_{t+1}) + \underbrace{(x_t - x_{t+1})^T \nabla\phi_t(x_{t+1})}_{\geq 0} + B_{\phi_t}(x_t \| x_{t+1})$$

$$\geq \phi_t(x_{t+1}) + B_{\phi_t}(x_t \| x_{t+1})$$

First order optimality condition

$$= \phi_t(x_{t+1}) + B_R(x_t \| x_{t+1})$$

thus $B_R(x_t \| x_{t+1}) \leq \phi_t(x_t) - \phi_t(x_{t+1})$

$$= (\phi_{t-1}(x_t) - \phi_{t-1}(x_{t+1})) + \eta \nabla_t^T (x_t - x_{t+1})$$

$x_t = \arg\min_x \phi_{t-1}(x)$ $\leq \eta_t \nabla_t^T (x_t - x_{t+1})$

Recall $B_R(x_t \| x_{t+1}) = \frac{1}{2} \|x_t - x_{t+1}\|_t^2$

$$\nabla_t^T (x_t - x_{t+1}) \leq \|\nabla_t\|_t^* \|x_t - x_{t+1}\|_t = \|\nabla_t\|_t^* \sqrt{2 B_R(x_t \| x_{t+1})} \leq \|\nabla_t\|_t^* \sqrt{2\eta \nabla_t^T (x_t - x_{t+1})}$$

$$\therefore \nabla_t^T (x_t - x_{t+1}) \leq 2\eta \|\nabla_t\|_t^{*2}$$

□

Revisit the Example of Learning from expert advice

OGD: $\frac{3}{2} GD\sqrt{T} \le \frac{3}{2}\sqrt{N}\sqrt{T}$

FTRL with Entropic Regularization $G_R D_R \sqrt{T}$

$$\nabla^2 R(x) = \begin{pmatrix} \frac{1}{x_{(1)}} & & \\ & \frac{1}{x_{(i)}} & \\ & & \frac{1}{x_{(n)}} \end{pmatrix}$$

$$G_R = \|\nabla_t\|_t^* = \|1+\log x\|_t^* = \sqrt{(1+\log x)^T \nabla^2 R(x)(1+\log x)} = \frac{(?)}{\sqrt{\sum_i (1+\log x_{(i)}) x_{(i)}}}$$

$$D_R = \sqrt{\max_{x,y \in K} R(x) - R(y)} \le \sqrt{\log n} \quad \text{(exercise 5.7.5 OCO Hazan)}$$

≤ 1

Back track to the Be the Leader Based Stability bound:

$$\sum_{t=1}^{T} \nabla_t^T (x_t - x_{t+1}) \le \sum_{t=1}^{T} \|\nabla_t\|_\infty \|x_t - x_{t+1}\|_1 \le \sum_{t=1}^{T} \|x_t - x_{t+1}\|_1 \le \eta T$$

why?

Proof:

$$x_t = \arg\min \sum_{t=1}^{t-1} \nabla_t^T x + \lambda R(x) =: F(x)$$

$$x_{t+1} = \arg\min \sum_{t=1}^{t} \nabla_t^T x + \lambda R(x) =: \tilde{F}(x)$$

By the Strong convexity of R in ℓ_1

$$F(x_{t+1}) \ge F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{\lambda}{2}\|x_{t+1} - x_t\|_1^2$$

$$\tilde{F}(x_t) \ge \tilde{F}(x_{t+1}) + \langle \nabla \tilde{F}(x_{t+1}), x_t - x_{t+1} \rangle + \frac{\lambda}{2}\|x_t - x_{t+1}\|_1^2$$

$$\|x_{t+1} - x_t\|_1^2 \le \frac{1}{\lambda}\left(-\nabla_t^T x_{t+1} + \nabla_t^T x_t\right) \le \frac{1}{\lambda}\|\nabla_t\|_\infty \|x_{t+1} - x_t\|_1$$

$$\Rightarrow \|x_{t+1} - x_t\|_1 \le \frac{1}{\lambda} \Rightarrow \sum_{t=1}^{T} \|x_t - x_{t+1}\|_1 \le \frac{T}{\lambda} = \eta T$$