Lecture 15 Follow the Regularized Leader.

Recall OCO

Adversary chooses $f_1, f_2, \dots, f_T$

for $t = 1, 2, \dots, T$

1. Player plays $x_t \in K$
2. Incurs a loss of $f_t(x_t)$
3. Receive Feedback

$\nabla f_t(x_t)$ ← Full info.

$f_t(x_t)$ ← Bandits

f_t ← full function access

$$\boxed{\text{Regret}(A) = \sum_{t=1}^{T} f_t(x_t) - \inf_{u \in X} \sum_{t=1}^{T} f_t(u)}$$

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

Thm. $\eta_t = \frac{D}{G\sqrt{t}}$, $\sup_{x,y \in K} \|x-y\|_2 \le D$, $\|\nabla f_t(x)\|_2 \le G$

$\text{Regret}(OGD) \le \frac{3}{2} G D \sqrt{T} = O(\sqrt{T})$

(Strong Convexity) f_t obeys m-Strong Convexity

Thm. $\eta_t = \frac{m}{t}$, $\text{Regret} \le \frac{G^2}{m}\underbrace{(1 + \log(T))}$

$1 + \frac{1}{2} + \dots + \frac{1}{T} \le \log_2(1+T)$

Convergence of (Stochastic) Subgradient Method

$\text{Suboptimality} \le \frac{1}{T} \text{Regret}(OGD)$.

① $E[\tilde{g}_t \mid x_t] = g_t$

② $E[\|\tilde{g}_t - g_t\|^2 \mid x_t] \le G^2$

③ $\|g_t\|_2^2 \le G^2$

→ $P(\|\tilde{g}_t\|_2^2 \le G^2 + G^2) = 1$

Today: a different style of Alg for OCO

FTL: Follow the leader

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

at time t, play x_t.

$$x_{t+1} = \arg\min_x \sum_{t=1}^{t-1} \langle \nabla f_t, x \rangle$$

Example: $K = [-1, 1]$, $f_1 = \frac{1}{2}x$, $f_2 = -x$, $f_3 = x$, $f_4 = -x$, ...

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

FTL will play: $x_1, x_2, x_3, \dots, x_T$

$0, -1, 1, -1, 1, -1, 1, \dots$

$f_3(x_3) = 1$ $\quad f_3(x_3) = 1$

Regret, v.s. $u = 0$ $\quad 0 + 1 + 1 + 1 + \cdots = T - 1$

$f_2(u) = 0$ $\quad f_3(u) = 0$

Learning with expert advice

$f_t = -\langle x_t, r_t \rangle$, $x_t \in \text{Conv}(e_1, \dots, e_n) = \Delta_n$

$\|r_t\|_\infty \leq 1$ $\quad \mathcal{X} = \{e_1, \dots, e_n\}$

RWM, Hedge, Exponential Average

Initiate: $x_1(i) = 1 \quad \forall i \in [n]$

for $t = 1, 2, 3, \dots T$

1. play i with prob $\propto x_t(i)$

   if $K = \Delta_n$, play $\frac{x_t}{1^T x_t}$ ← normalization

2. incur loss of $-\langle x_t, r_t \rangle$ (expected)

3. update $x_{t+1}^{(i)} = x_t^{(i)} e^{\varepsilon r_t(i)}$ for all $i \in [n]$

$[n] = \{1, 2, \dots, n\}$

$$\log(x_{t+1}) = \log(x_t) + \varepsilon \vec{r}_t = \log(x_t) - \varepsilon \nabla f_t(x_t)$$

Exponentiated Gradient Alg.

FTL fails because of instability.

OGD is stable. $\|g_t\|_2 \le G$, also $\eta_t = O\left(\frac{1}{\sqrt{t}}\right) \to 0$

Recall Wedge Alg. Regret: Lecture 13 Intro to OCO / Learning from experts

Thm. Regret(RMW, Wedge)

$$\le C\cdot\left(\varepsilon\cdot m + \frac{\log n}{\varepsilon}\right)$$

\# of mistakes of best expert

$$\varepsilon \sim \frac{1}{\sqrt{m}} \Rightarrow C\cdot\sqrt{m\log n} \le C\sqrt{T\cdot\log n}$$

$$\varepsilon = \frac{1}{\sqrt{T}}$$

Idea: Regularization

FTL $\Rightarrow$ FTRL

Example 1.

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

$\nabla^2 R = I$

$B_R(x\|y) = \frac{1}{2}\|x-y\|^2$

Example 2:

$R(x) = \langle x, \log x\rangle$

$x \in \Delta_n$

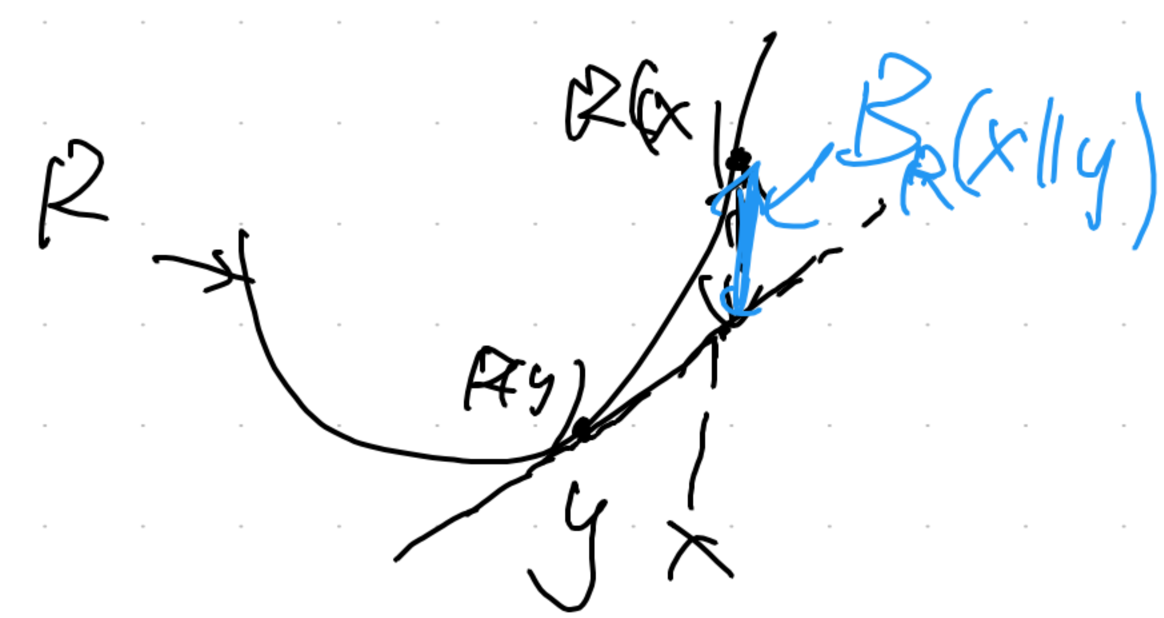
$B_R(x\|y)$

$= \sum_i x_i \log\frac{x_i}{y_i}$

$= KL\text{-}div(x\|y)$

Bregman Divergence: for a convex function R, twice differentiable

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



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

Taylor's thm

$$R(x) = R(y) + \langle\nabla R(y), x-y\rangle + \underbrace{\boxed{\frac{1}{2}(x-y)^T\nabla^2 R(z)(x-y)}}_{B_R(x\|y)}, \quad z\in[x,y]$$

Other Properties of Bregman divergence

1. $B_R(x,y) \geq 0$
2. $B_R(x,y)$ is convex in x
3. $B_{R_1+\lambda R_2}(x,y) = B_{R_1}(x\|y) + \lambda B_{R_2}(x\|y)$
4. $B_{R^*}(x^*\|y^*) = B_R(x,y)$ (Duality)

   for $x^* = \nabla R(y)$, $y^* = \nabla R(x)$
5. $y \sim D$, $\underset{D}{E}[y] = \underset{x}{\operatorname{argmin}} \underset{y \sim D}{E} B_R(x\|y)$

   $\forall$ convex function R

More properties, Xinhua Zhang notes on Bregman Divergence

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

1. let $x_1 = \underset{x \in K}{\operatorname{argmin}} R(x)$
2. for $t = 1, 2, 3, \ldots T$
   a. play x_t
   b. observe $\nabla_t := \nabla f_t(x_t)$
   c. update $x_{t+1} = \underset{x \in K}{\operatorname{argmin}} \left\{ \eta \sum_{s=1}^{t} \nabla_s^T x + R(x) \right\}$

Recall Dual Norm $\|x\|_*$

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

$$\langle x, y \rangle \leq \|x\| \|y\|_*$$

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

$$B_R(x\|y) =: \frac{1}{2} \|x-y\|_{x,y}^2$$

$$B_R(x_t\|x_{t+1}) = : \frac{1}{2} \|x_t - x_{t+1}\|_t^2 \in \nabla^2 R(z)$$

Example: $R = \frac{1}{2}\|x\|^2$ then this is Euclidean projection

$$x_{t+1} = \underset{x \in K}{\operatorname{argmin}} \eta \sum_{s=1}^{t} \nabla_s^T x + \frac{1}{2}\|x\|_2^2$$

$$= \underset{x \in K}{\operatorname{argmin}} \Big\| x - \underbrace{\eta \sum_{s=1}^{t} \nabla_s}_{\pi} \Big\|^2$$

(Lazy) OGD

Example: $R(x) = x^T \log x$

$= \sum x[i] \log(x[i])$

$$\boxed{\min_x \; \eta \sum_{s=1}^{t} \ell_t^T x + \sum_{i=1}^{n} x_i \log x_i \quad \text{s.t. } x \geq 0,\ \mathbb{1}^T x = 1}$$

$$L(x,u,v) = \eta \sum_{s=1}^{t} \ell_s^T x + \sum_{i=1}^{n} x_i \log x_i + u^T(-x) + v(\mathbb{1}^T x - 1)$$

$$\nabla_x L(x,u,v) = \eta \sum_{s=1}^{t} \ell_s + \log x + \vec{\mathbb{1}} + (-u) + v\vec{\mathbb{1}} = 0$$

$$\log x = \underbrace{-\eta \sum_{s=1}^{t} \ell_s}_{\uparrow} + \underset{\substack{\| \\ 0}}{u} \underbrace{- (1+v)\vec{\mathbb{1}}}_{}$$

0 by complementary slackness

ensure x is probability

Recovers Hedge Alg.

Thm (FTRL): $\text{Regret} \leq 2\eta \sum_{t=1}^{T} \| \underset{\substack{\| \\ \nabla_t}}{\nabla f_t(x_t)} \|_t^* + \frac{R(u) - R(x_1)}{\eta}$

Upper bound:

$\text{Regret} \leq 2 D_R G_R \sqrt{2T}$

$\|\cdot\|_t = \|\cdot\|_{\nabla^2 R(z)}$

$\|\cdot\|_t^* = \|\cdot\|_{(\nabla^2 R(z))^{-1}}$ $\quad z \in [x_t, x_{t+1}]$

Not changing $O(\sqrt{T})$

Adapting to the geometry of the problem

Be the Leader Alg. (Fictitious Alg.)

Play $x_{t+1} = \arg\min_x \sum_{s=1}^{t} \langle \nabla_t, x \rangle = \arg\min_x \sum_{s=1}^{t} f_t(x)$ at time t.

in linear losses

$$\boxed{\text{Lemma: (Regret of BTL)} \quad \text{Regret} \leq \sum_{t=1}^{T} \nabla_t^T x_t - \sum_{t=1}^{T} \nabla_t^T x_{t+1} + \frac{1}{\eta} D_R^2}$$

linearized vers. of the sum of loss of A

of BTL

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

Stability $\Rightarrow$ Bound

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

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

BTL is better than the best u in the hindsight

$$\sum_{t=1}^{T} g_t(x_t) - g_t(u) = \sum_{t=1}^{T} g_t(x_t) - g_t(x_{t+1}) + g_t(x_{t+1}) - g_t(u)$$

$$= \sum_{t=1}^{T} g_t(x_t) - g_t(x_{t+1}) + \underbrace{\sum_{t=0}^{T} g_t(x_{t+1}) - g_t(u)}_{\leq 0 \text{ Claim}} \underbrace{- g_0(x_1) + g_0(u)}$$

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

$x_1 = \arg\min_{x \in K} R(x)$

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

Proof of the claim, $\sum_{t=0}^{T} g_t(u) \geq \sum_{t=1}^{T} g_t(x_{t+1})$. By induction

① when $T = 0$, $g_0(u) \geq g_0(x_1)$ ✓

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

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

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

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

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

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

□

Proof of the Theorem:

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

By Taylor's Theorem

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

First order optimality condition of x_{t+1}

$$\geq \phi_t(x_{t+1}) + 0 + B_{\phi_t}(x_t \| x_{t+1})$$

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

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

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

$x_t = \arg\min_{x \in K} \phi_{t-1}(x) \rightarrow$

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

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

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

Holder $\eta_t \langle \nabla_t, x_t - x_{t+1} \rangle$

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

by $(*)$

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

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

$$\text{Regret} \leq 2\eta \sum (\|\nabla_t\|_t^*)^2 + \frac{1}{\eta} D_R^2$$

Choose η appropriately

$$\asymp D_R G_R \sqrt{T}$$

17

$\sqrt{\log n\, T}$: OGD: $\frac{3}{2} GD\sqrt{T} \le \frac{3}{2}\sqrt{nT}$ (exercise to verify this)

FTRL with Entropy Regularizer: $G_R D_R \sqrt{T}$. $R(x) = x^T \log x$

$$\|\nabla_t\|_t^* = \|\vec{1} + \log x\|_t^* = \sqrt{(1+\log x)^T \nabla^{-2} R(z) (1+\log x)}$$

$$= \sqrt{\sum_i (1+\log x_i) \cdot z_i} \le 1 \quad G_R$$

$z_t(i) \in [x_t, x_{t+1}]$ $\log x_i \le 0$

$$\nabla^2 R(z) = \begin{pmatrix} \frac{1}{z_1} & & 0 \\ & \frac{1}{z_2} & \\ 0 & & \ddots \; \frac{1}{z_n} \end{pmatrix}$$

$$\frac{\partial}{\partial z_i} (\vec{1} + \log z) = \begin{pmatrix} 0 \\ \frac{1}{z(i)} \\ 0 \end{pmatrix}$$

$$D_R = \sqrt{\max_{x,y \in \mathcal{K}} (R(x) - R(y))} \le \sqrt{\log n}$$ (Exercise 5.7.5 OCO)