I Finishing Online Gradient Descent

Thm: Let $f_1, \dots, f_T$ be convex functions and

$$w_1 = 0$$

$$w_{t+1} = w_t - \eta \nabla f_t(w_t)$$

for some $\eta > 0$. Suppose

$$\|\nabla f_t\| \le G \quad \forall t.$$

Then $\forall w$

$$\sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w) \le \eta G^2 T + \frac{\|w\|_2^2}{\eta}$$

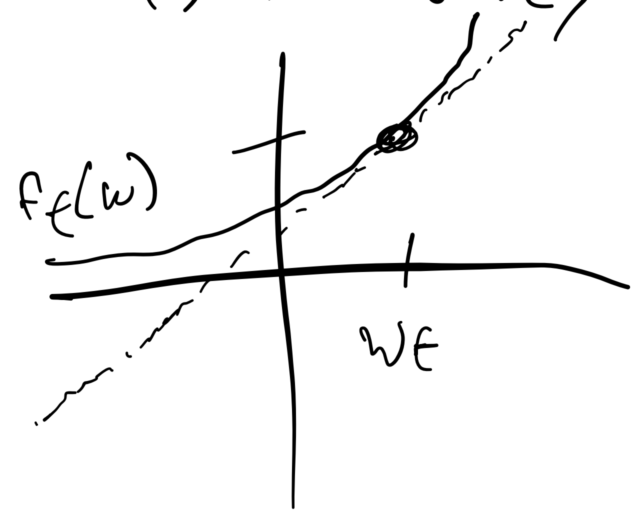
If $\eta = \frac{1}{\sqrt{T}}$ $\rightsquigarrow = (G^2 + \|w\|_2^2)\sqrt{T}$

''

Pf: Shown last time when f_t is linear.

Reduction to linear:

$$\forall w \quad f_t(w_t) + \langle \nabla f_t, w - w_t \rangle \le f_t(w)$$



$$\sum_{t=1}^{T} \underbrace{\left(f_t(w_t) - f_t(w)\right)}_{\text{regret at time } t} \le \sum_{t=1}^{T} \underbrace{\langle \nabla f, w_t - w \rangle}_{\text{regret for}}$$

$$\ell_t(w) = \langle \nabla f_t(w_t), w \rangle$$

Apply OGD to the sequence

$\ell_1 \ldots \ell_T$

$$y_t = \langle w^*, x_t \rangle + \xi_t$$

$$f_t(w) = \left(\langle w^*, x_t \rangle - \langle w, x_t \rangle\right)^2$$

Back to Contextual Bandit

$$\text{Decision Maker} \xleftarrow{\text{context } x_t} \text{Nature}$$

$$\text{Decision Maker} \xrightarrow[\text{(arm)}]{\text{decision } \pi_t} \text{Nature}$$

$$\text{Decision Maker} \xleftarrow{\text{reward } r_t \in \mathbb{R}} \text{Nature}$$

$\pi_t \in [K]$

K: # of arms

Expected Regret

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t - \sum_{t=1}^{T} \max_{\pi} f^*(x_t, \pi)\right]$$

Assumption:

independent mean zero noise

$$r_t = f^*(x_t, \pi_t) + \xi_t$$

for some $f^* \in \mathcal{F}$.

unknown

Known to us.

Example (Linear Contextual Bandit) $\theta \in \mathbb{R}^d$

$$\mathcal{F} = \{ f_\theta(x_t, \pi_t) = \langle \theta, \phi(x_t, \pi_t) \rangle \}$$

$$f^*(x_t, \pi_t) = \langle \theta^*, \phi(x_t, \pi_t) \rangle$$

Known feature map

$\gamma > 0$

SquareCB(γ) [Foster - Rakhlin '20]

Assumes access to online regression oracle for $\mathcal{F}$. (e.g. OGD for linear classes)

For $t=1$ to T:

1. Receive context x_t.
2. Use oracle $\hat{f}_t$ to forecast $\hat{r}_t(\pi) = \hat{f}_t(x_t, \pi) \quad \forall \pi \in [K]$

Let $\hat{\pi} = \operatorname{argmax}_{\pi} \hat{r}_t(\pi)$

$\sum_{\pi} p_t(\pi) = 1$

3. $p_t(\pi) = \frac{1}{\lambda + \gamma(\hat{f}_t(\hat{\pi}) - \hat{f}_t(\pi))}$ for $\pi \in [2, K]$
4. Sample $\pi \sim p_t$ and play it, observe reward r_t
5. Update regression oracle w/ (x_t, π_t, r_t).

Thm: