CS292F Lecture 14 Online (Projected) Gradient Descent

Online Convex Optimization

$$\text{Regret}_T(A) = \sup_{\{f_1,\ldots,f_T\}} \sum_{t=1}^{T} f_t(x_t) - \min_{u\in K} \sum_{t=1}^{T} f_t(u)$$

$f_i \in \mathcal{F}$ ← Family of convex functions

K ← convex set

Setting: Player declares Alg A.

Adversary chooses $f_1, f_2, \ldots, f_T \in \mathcal{F}$

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

1. Player plays $x_t \in K$, using A
2. incur loss $f_t(x_t)$
3. receive feedback
   - $\nabla f_t(x_t)$ ← Full information Feedback
   - $f_t(x_t)$ ← Bandit Feedback
   - f_t as a function ← Full function access

Example 1: Learning from expert advice.

$$K = \{x \mid x \geq 0, \sum_{i=1}^{n} x[i] \leq 1\} =: \Delta_n$$

$$f_t(x) = \langle l_t, x\rangle = \sum_{i=1}^{n} x[i]\cdot l_t[i] = \mathbb{E}_{a\sim x}[l_t[a]]$$

Feedback model: full info: $\nabla f_t(x_t) = l_t$

$f_t(e_a) = l_t[a]$ ← Bandit feedback

Example 2: Online linear models

$x_t \in K \subseteq \mathbb{R}^n$, $\phi_t \in \Phi \subseteq \mathbb{R}^n$

$f_t(x) = (\langle x, \phi_t\rangle - y_t)^2$

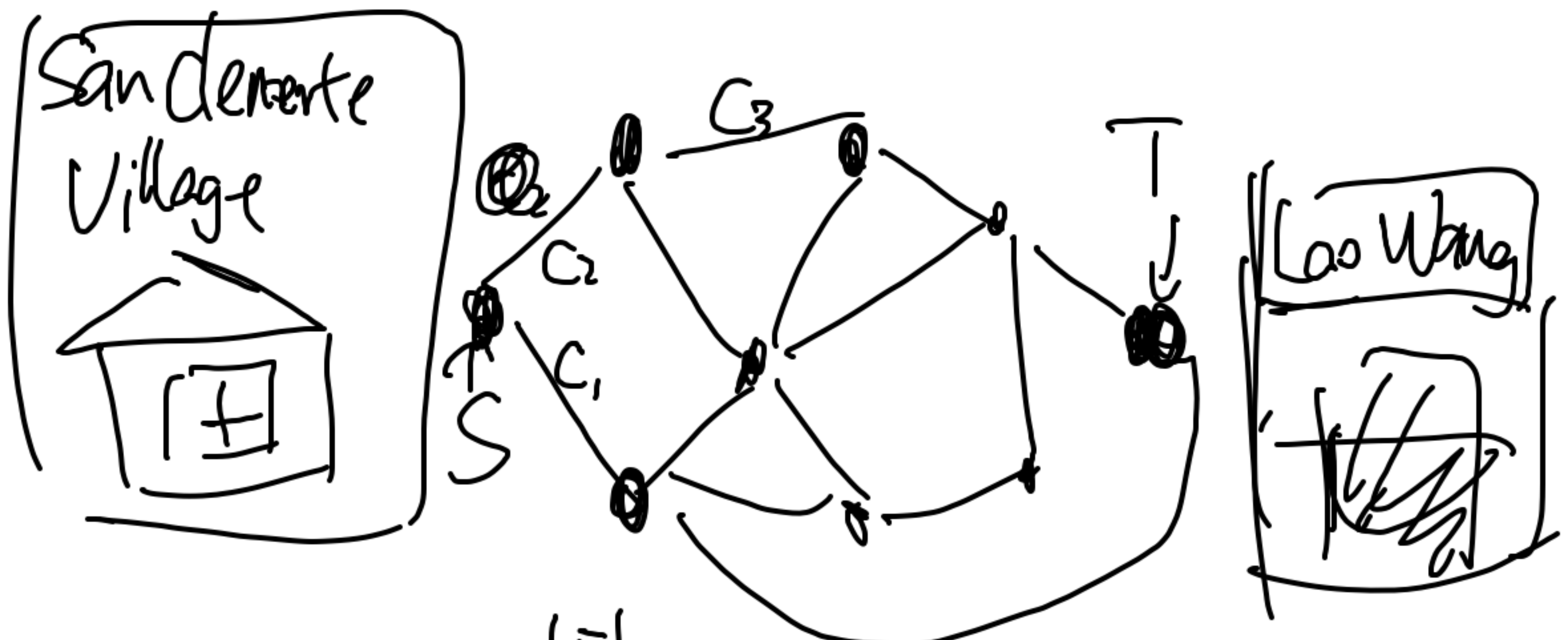
or $|\langle x,\phi_t\rangle - y_t|$, or $\text{huber}(|\langle x,\phi_t\rangle - y_t|)$,

or $\text{Logistic}(\langle x_t, \phi_t\rangle, y_t)$

(ϕ_t, y_t)

$\nabla f_t(x_t) = 2\phi_t\cdot(\langle x_t,\phi_t\rangle - y_t)$

Example 3. Online Shortest Path



$c \in \mathbb{R}^{|E|}$

$x_t \in \mathbb{R}^{|E|}$ $x_t \geq 0$ probability of taking a particular edge

$$\min \sum_t \langle c, x_t \rangle = \sum_t f_t(x_t)$$

$$\text{s.t.} \begin{cases} \sum_{\substack{i,j \\ i=S \\ (i,j)\in E}} x_t(i,j) = 1, \quad \sum_{\substack{i \\ j=T \\ (i,j)\in E}} x_t(i,j) = 1 \\ \sum_{\substack{i \\ (i,v)\in E}} x_t(i,v) = \sum_{\substack{i \\ (v,i)\in E}} x_t(v,i) \quad \forall v \in E \setminus \{S,T\} \end{cases}$$

K $\forall t$

Example 4. Portfolio selection

$r_t \in \mathbb{R}^n$ $r_t \geq 0$

stock i, $r_t[i]$ relative ratio of stock i

100 today 95 tomorrow

$r_t[\text{apple}] = \frac{95}{100}$

$x_t \in \Delta_n$

$\prod_{t=1}^{T} \langle r_t, x_t \rangle =$ cumulative relative gain

$\sum_{t=1}^{T} \log(r_t^T x_T)$

$f_t(x_t) = -\log(r_t^T x_t)$

$$\underbrace{-\sum_{t=1}^{T} \log(r_t^T x_T)}_{\text{log of your (cumulative return)}} + \underbrace{\max_{x \in \Delta_n} \sum_t \log(r_t^T x)}_{\text{Best fixed proportion (investment strategy in the hindsight)}}$$

Thomas Cover Universal Portfolio (1994)

OGD: consider subgradient feedback

Input: Convex $\mathcal{K}$, Horizon T, stepsizes $\{\eta_t\}_{t=1,\cdots,T}$

x_1 as an initialization

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

play x_t, observe $g_t \in \partial f_t(x_t)$

update: $x_{t+\frac{1}{2}} = x_t - \eta_t g_t$

$x_{t+1} = \Pi_{\mathcal{K}}(x_{t+\frac{1}{2}})$

Assumptions

G-Lipschitz: $f(x) - f(y) \leq G\|x-y\|_2 \quad \forall x, y$

or $\|\nabla f(x)\|_2 \leq G \quad \forall x$

f convex

$\mathcal{K}$: $\operatorname{diam}(\mathcal{K}) \leq D$

$\forall x, y \in \mathcal{K} \quad \|x-y\|_2 \leq D$

Theorem (OGD) $\eta_t = \dfrac{D}{G\sqrt{t}}$, $t = 1, 2, 3, \cdots, T$

$$\text{Regret}(u) := \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(u) \leq \frac{3}{2} GD\sqrt{T}$$

$\forall u \in \mathcal{K}$ $\qquad \simeq \sqrt{T}$

Proof: $\underbrace{f_t(x^*) \geq f_t(x_t) + \langle g_t, x^* - x_t\rangle}_{\text{By convexity of } f} \iff \underbrace{f_t(x_t) - f_t(x^*)}_{} \leq \langle g_t, x_t - x^*\rangle \leftarrow \times 2$

$$\|x_{t+1} - x^*\|_2^2 \underset{\text{By OGD alg}}{=} \|\Pi_{\mathcal{K}}(x_t - \eta_t g_t) - x^*\|_2^2 \underset{\text{non-expansiveness of } \Pi_{\mathcal{K}}}{\leq} \|x_t - \eta_t g_t - x^*\|^2 = \|x_t - x^*\|^2 + \eta_t^2\|g_t\|^2 - \underbrace{2\eta_t\langle g_t, x_t - x^*\rangle}_{\text{move this to the other side}}$$

$$2(f_t(x_t) - f_t(x^*)) \leq 2\langle g_t, x_t - x^*\rangle \leq \frac{\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2}{\eta_t} + \eta_t\|g_t\|^2$$

$$2(f_t(x_t) - f_t(x^*)) \le \frac{1}{\eta_t}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right) + \eta_t \|g_t\|^2$$

Sum $t = 1 \dots T$

$$2 \cdot \underline{\text{Regret}} := 2\left(\sum_t f_t(x_t) - \sum_t f_t(x^*)\right) \le \sum_{t=1}^T \frac{1}{\eta_t}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right) + \left(\sum_t \eta_t\right)\cdot G^2$$

$$= \left(\sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\cdot \|x_t - x^*\|^2\right) \underbrace{- \frac{1}{\eta_T}\|x_{T+1} - x^*\|^2}_{\le 0} + \left(\sum_t \eta_t\right) G^2$$

$$\boxed{\begin{gathered}\eta_0 = +\infty \\ \eta_t = \frac{D}{G\sqrt{t}}\end{gathered}}$$

$$\le D^2 \sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \left(\sum_t \eta_t\right) G^2$$

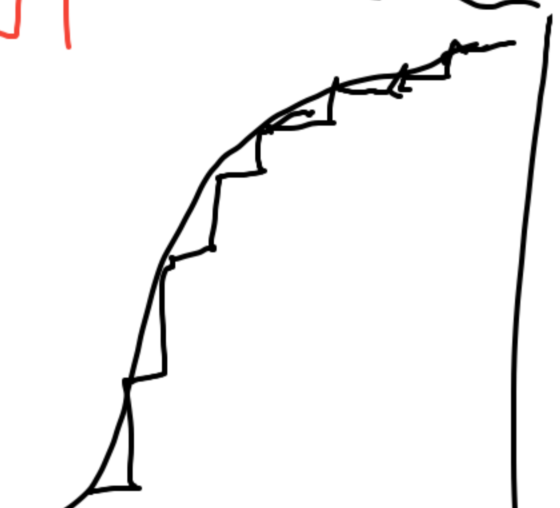
$$= D^2 \frac{1}{\eta_T} + G^2 \sum_{t=1}^T \eta_t$$

$$= \frac{D^{\cancel{2}} G\sqrt{T}}{\cancel{D}} + \frac{G^{\cancel{2}} D}{\cancel{G}} \sum_{t=1}^T \frac{1}{\sqrt{t}} \le \underline{DG\sqrt{T} + 2DG\sqrt{T}}$$

$$1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \dots + \frac{1}{\sqrt{T}} \le ?\ 2\sqrt{T}$$

$$\ge T \cdot \left(\frac{1}{\sqrt{T}}\right) = \sqrt{T}$$

$$\sum_{t=1}^T \frac{1}{\sqrt{t}} \le \int_0^T \frac{1}{\sqrt{x}}\,dx = \left[2\sqrt{x}\right]_0^T = 2\sqrt{T}$$



$$\underline{\text{Regret} \le 1.5\, DG\sqrt{T}}$$

$\eta_t = \frac{D}{G\sqrt{T}}$ $\quad 2\,\text{Regret} \le DG\sqrt{T} + DG\sqrt{T}$

Constant improvement from $1.5 \to 1$.

Strongly Convex case: f_t is m-strongly convex. f_t is G Lipschitz.

D doesn't have to be bounded

$\boxed{\text{Thm (OGD in strongly convex case). Regret} \le \frac{G^2}{2m}(1+\log T)}$

choosing $\eta_t = \frac{1}{mt}$, assume f_t is m-strongly convex.

improved from $O(\sqrt{T})$

Proof: By ~~convexity~~ m-strong convexity of f_t,

$$2\left(f_t(x_t) - f_t(x^*)\right) \le 2g_t^T(x_t - x^*) \underbrace{- m\|x^* - x_t\|^2}_{\text{new!}} \quad (1)$$

$$\|x_{t+1} - x^*\|^2 = \|\Pi_K(x_t - \eta_t g_t) - x^*\|^2 \le \|x_t - \eta_t g_t - x^*\|^2 = \|x_t - x^*\|^2 + \eta_t^2\|g_t\|^2 - 2\eta_t g_t^T(x_t - x^*)$$

$$2g_t^T(x_t - x^*) \le \frac{\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2}{\eta_t} + \eta_t G^2 \quad (2)$$

Combine (1) and (2), and then sum up over $t = 1, 2, \dots, T$

$$2\left(\sum_t f_t(x_t) - f_t(x^*)\right) \le \underbrace{\sum_t \frac{\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2}{\eta_t} - m\|x^* - x_t\|^2}_{} + G^2\sum_{t=1}^T \eta_t$$

$$= \sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - m\right)\|x^* - x_t\|^2 \underbrace{- \frac{1}{\eta_T}\|x_{T+1} - x^*\|^2}_{\le 0} + G^2\sum_{t=1}^T \eta_t$$

$$= \sum_{t=1}^T \underbrace{(mt - m(t-1) - m)}_{=0}\|x^* - x_t\|^2 + \frac{G^2}{m}\left(1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{T}\right) \le \frac{G^2}{m}(1 + \log T) \quad \square$$

Recall stochastic subgradient method

$$x_{t+1} = x_t - \eta_t \tilde{g}_t$$

1. $\mathbb{E}[\tilde{g}_t \mid x_t] = g_t \in \partial f(x_t)$
2. $\mathbb{E}\left[\|\tilde{g}_t - \mathbb{E}[\tilde{g}_t \mid x_t]\|^2 \mid x_t\right] \le \sigma^2$
3. f is convex, Lipschitz G

ε-suboptimality, $O\left(\frac{1}{\varepsilon^2}\right)$ iterations

$$\mathbb{E}\left[f(x_k^{(best)}) - f^*\right] \le O\left(\frac{1}{\sqrt{k}}\right)$$

From Lecture on stochastic subgradient

$$\mathbb{E}\left[f(\bar{x}_k) - f^*\right] \le O\left(\frac{1}{\sqrt{k}}\right)$$

We did not prove the result of strong convexity

$$\tilde{O}\left(\frac{1}{mk}\right)$$

$$\tilde{O}\left(\frac{1}{m\varepsilon}\right) \text{ iteration complexity}$$

Analysis: Jensen's inequality

$$f(\bar{x}_T) - f(x^*) \le \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*)$$

$$\le \frac{1}{T}\sum_{t=1}^{T} \langle g_t, x_t - x^* \rangle - \frac{m}{2}\|x_t - x^*\|^2$$

convexity of f

$$-\frac{m}{2}(x_t - x^*)$$

$$= \frac{1}{T}\sum_{t=1}^{T} \langle \mathbb{E}[\tilde{g}_t \mid x_t], x_t - x^* \rangle$$

$$= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\langle \tilde{g}_t, x_t - x^* \rangle \mid x_t\right]$$

take full expectation on both sides.

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \le \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\langle \tilde{g}_t, x_t\rangle - \sum \langle \tilde{g}_t, x^*\rangle\right]$$

$$= \frac{1}{T}\mathbb{E}\Big[\underbrace{\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*)}_{\text{Regret}}\Big] \le \text{Regret(OGD)}$$

$$f_t(x) = \langle \tilde{g}_t, x \rangle$$

x^* is the minimizer of f, Not $\min \sum_t f_t(x^*)$

~~Strongly convex $\frac{(G^2+\sigma^2)\log T}{2mT}$~~

Convex $\frac{\text{OGD}\sqrt{T}}{T} = O\left(\frac{1}{\sqrt{T}}\right)$

$z_1 \cdots z_n \sim D$, find $h \in \mathcal{H}$ s.t. (Batch learning setting)

$h \to h^*$ $\quad \underbrace{\mathbb{E}_{z \sim D}[\ell(h,z)] - \mathbb{E}_{z \sim D}[\ell(h^*,z)]}_{\text{Excess Risk}} \to 0$

$\leq O(\sqrt{\frac{1}{n}})$

Online-to-Batch Conversion

(more in HW4)

You can bound excess risk of $\bar{x}_{1:T}$ by the regret.

$\boxed{\text{Corrections}}$ to the last part of the lecture on reducing Stochastic Subgradient method convergence to OGD.

1. The provided proof was only for general convex function.

2. We actually need $P(\|\check{g}_t\|_2 \le G) = 1$ for all t.

This is stronger than $E(\|\check{g}_t\|^2 | X_t] \le G^2 + G^2$.

← there is a way of getting this by parameterizing the regret by $\|\check{g}_t\|^2$...

for strongly convex problems, we need to tighten the arguments a bit and actually apply strong convexity.

$$f(\bar{x}) - f(x^*) \overset{\text{Jensen's}}{\le} \frac{1}{T}\sum_{t=1}^{T}\left(f(x_t) - f(x^*)\right) \overset{m\text{-strong convexity}}{\le} \frac{1}{T}\sum_{t=1}^{T} \langle g_t, x_t - x^* \rangle \underbrace{- \frac{m}{2}\|x_t - x^*\|^2}_{\text{new!}}$$

$$= \frac{1}{T}\sum_{t=1}^{T} \langle E[\check{g}_t | X_t], x_t - x^* \rangle - \frac{m}{2}\|x_t - x^*\|^2$$

Take expectation on both sides,

$$E f\bar{x} - f x^* \le \frac{1}{T} E\left[\sum_{t=1}^{T} \langle \check{g}_t, x_t \rangle - \left(\langle \check{g}_t, x^* \rangle + \frac{m}{2}\|x_t - x^*\|^2\right)\right]$$

$$\underbrace{\text{Define } f_t(x) = \langle \check{g}_t, x \rangle + \frac{m}{2}\|x - x_t\|^2}_{\text{new!}}$$

$$= \frac{1}{T} E\left[\sum_{t=1}^{T} f_t(x_t) - f_t(x^*)\right] \le \frac{1}{T}\,\text{Regret} = \frac{1}{T}\frac{G^2 \log T}{2m}$$

check f_t is m-strongly convex

$\nabla f_t(x_t) = \check{g}_t$, $\|\check{g}_t\|_2 \le G$.