Future scribing: go to Piazza, edit link on overleaf repo.

HW3: errata page on Piazza.

Mid-project progress report due today on gradescope.

OCO for strongly cvx fns. / kernel methods.

Motivation: fast algs for regularized loss minimization

logistic / hinge.

$$\underbrace{\min_{w \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \left(\ell(w,(x_i,y_i)) + \frac{\lambda}{2} \|w\|_2^2 \right)}_{F_S(w)} . \quad \|x_i\| \le B.$$

$$w. \quad F_S(w) \le F_S(\hat{w}) + \varepsilon. \qquad \hat{w} = \arg\min_w F_S(w).$$

evaluating this would take $O(m)$.

- Idea 1: gradient descent

$$w_{t+1} \leftarrow w_t - \eta \nabla F_S(w_t)$$

OGD's guarantee

$$f_t \equiv F_S. \qquad F_S(\bar{w}_T) - F_S(\hat{w}) \le \frac{1}{\sqrt{T}}.$$

#iters

$$T = O\left(\frac{1}{\varepsilon^2}\right).$$

$$\text{running time} = O\left(\frac{m}{\varepsilon^2}\right).$$

- Idea 2: stochastic gradient descent.

define $\hat{D} = \text{unif}\left((x_i,y_i)_{i=1}^{m} \right)$,

use OGD on the regularized losses induced by random examples drawn from $\hat{D}$, and do online to batch conversion.

For $t = 1, 2, \cdots T$.

sample $i_t \sim \text{uniform}(\{1 \cdots m\})$.

$f_t(w) = \underline{\ell(w, (x_{i_t}, y_{i_t})) + \frac{\lambda}{2}\|w\|^2}$. λ-SC wrt $\|\cdot\|_2$

$w_{t+1} \leftarrow w_t - \eta \cdot g_t$. where $g_t \in \partial f_t(w_t)$.

$\{w_1 \cdots w_T\} \xrightarrow{?} \bar{w}_T$.

$$\mathbb{E}[F_S(\bar{w}_T)] - F_S(\hat{w}) \overset{?}{\leq} \frac{1}{\sqrt{T}}.$$

$$\mathbb{E}[L_D(\bar{w}_T)] - \underline{\min_w L_D(w)} \leq \underline{\mathbb{E}[R_T(w^*)]} \leq \frac{1}{\sqrt{T}}.$$

(see Remark 2 in Mar 23's lecture)

$T = O(\frac{1}{\varepsilon^2})$. running time $= O(\frac{1}{\varepsilon^2})$.

can we do even better?

Thm: Ω wx. $\{f_t\}_{t=1}^T$ are $\underline{\lambda\text{-SC}}$ wrt $\|\cdot\|_2$.

we run time varying step size OGD:

$$w_{t+1} \leftarrow w_t - \eta_t \cdot g_t. \quad g_t \in \partial f_t(w_t)$$

$\eta_t = \underline{\frac{1}{\lambda t}}$, and assume $\forall t$. $\|g_t\|_2 \leq L$.

then,

$$R_T(\Omega) \leq \frac{(\ln T + 1) L^2}{2\lambda}$$

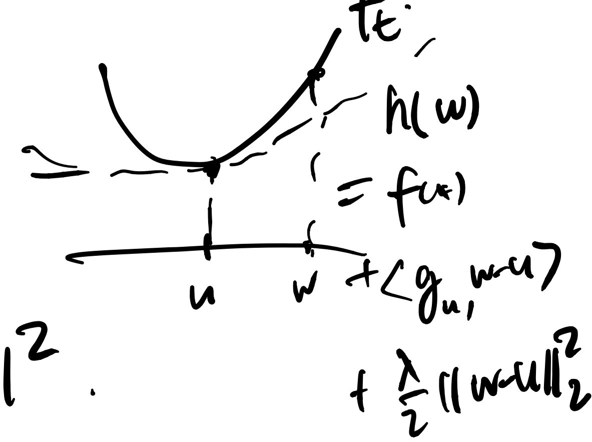
(much better than the basic $O(\sqrt{T})$ regret, by exploiting strong convexity).

This implies that running $\frac{1}{\lambda t}$-step-size SGD yields

$$F_S(\bar{w}_T) - F_S(\hat{w}) \le \tilde{O}\left(\frac{1}{T}\right).$$

$$T = O\left(\frac{1}{\varepsilon}\right) \ll O\left(\frac{1}{\varepsilon^2}\right).$$

<u>Pf</u>: ① by quadratic lower bound property of λ-sc fns



$u \in \Omega$

$$\Rightarrow f_t(w_t) - f_t(u) \le \langle g_t, w_t - u\rangle - \frac{\lambda}{2}\|w_t - u\|^2.$$

(improving over the linearization step by utilizing λ-sc).

② Recall that in OGD analysis: ~~the same~~

$$\langle g_t, w_t - u\rangle \le \frac{\|u - w_t\|_2^2 - \|u - w_{t+1}\|_2^2}{2\eta_t} + \frac{\eta_t}{2}\|g_t\|_2^2$$

Combining ① ②. summing over $t = 1, 2, \dots T$.

$$\sum_{t=1}^{T} f_t(w_t) - f_t(u)$$

$$\le \sum_{t=1}^{T} \frac{\|u-w_t\|_2^2 - \|u-w_{t+1}\|_2^2}{2\eta_t} - \sum_{t=1}^{T} \frac{\lambda}{2}\|u-w_t\|_2^2$$

$\eta_t = \frac{1}{\lambda t}$

$$+ \sum_{t=1}^{T} \eta_t \cdot \|g_t\|_2^2 .$$

$$\frac{\lambda\|u-w_1\|_2^2}{2\cancel{\eta_1}} - \frac{\lambda\|u-w_2\|_2^2}{2\cancel{\eta_1}} + \frac{\|u-w_2\|_2^2}{2\eta_2} - \frac{\|u-w_3\|_2^2}{2\eta_2} +$$

$$\cdots + \frac{\|u-w_T\|^2}{2\eta_T} - \underbrace{\frac{\|u-w_{T+1}\|^2}{2\eta_T}}_{} \qquad \le 0$$

$T-1$.

Coefficient of $\|u-w_1\|_2^2$: $\frac{1}{2\eta_1} - \frac{\lambda}{2} = 0$.

$\|u-w_2\|_2^2$: $-\frac{1}{2\eta_1} + \frac{1}{2\eta_2} - \frac{\lambda}{2} = 0$

$\|u-w_T\|_2^2$ $-\frac{1}{2\eta_{T-1}} + \frac{1}{2\eta_T} - \frac{\lambda}{2} = 0$

$$R_T(u) \le \sum_{t=1}^{T} \frac{\eta_t}{2}\|g_t\|_2^2 .$$

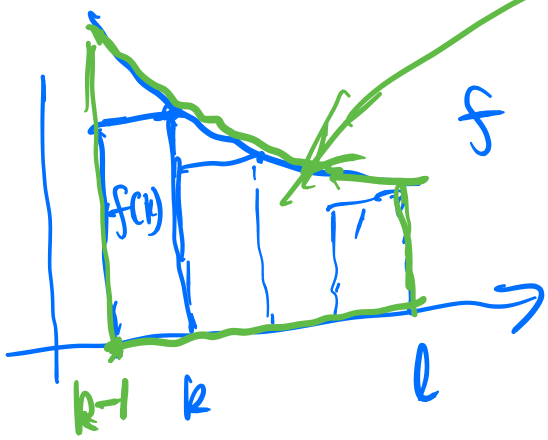
$\le \int_1^T \frac{1}{x}dx = \ln T$

$$= \frac{L^2}{2\lambda}\left(\sum_{t=1}^{T} \frac{1}{t}\right) \qquad \sum_{t=1}^{T} \frac{1}{t} = 1 + \sum_{t=2}^{T} \frac{1}{t} \le 1 + \ln T. \quad k=2,\ \ell = T.$$

$f(x) = \frac{1}{x}$

Fact: f decreasing, then

$$\sum_{i=k}^{\ell} f(i) \le \int_{k-1}^{\ell} f(x)\,dx.$$



Instantiating this result to unconstrained regularized loss minimization:

For $t = 1, 2, \cdots T$.

sample $i_t \sim$ uniform$(\{1 \cdots m\})$. β-Lip

$\Omega = \mathbb{R}^d$. $f_t(w) = \boxed{\ell(w, (x_{i_t}, y_{i_t}))} + \frac{\lambda}{2}\|w\|^2$. λ-SC wrt $\|\cdot\|_2$

$\ell_t(w)$

$w_{t+1} \leftarrow w_t - \frac{1}{\lambda t} g_t$. where $g_t \in \partial f_t(w_t)$.

(Fact: if f, g cvx. $h = f + g$.

on w, $a \in \partial f(w)$.

$b \in \partial g(w)$

$a+b \in \partial h(w)$

calculating g_t:

— $v_t \in \partial l_t(w_t)$

— $g_t = v_t + \lambda w_t$.

updating w_t:

$$w_{t+1} \leftarrow w_t - \frac{1}{\lambda t}(\lambda w_t + v_t)$$

$$= \left(1-\frac{1}{t}\right) w_t - \frac{1}{\lambda t} v_t .$$

observation:

$$\underbrace{t \cdot w_{t+1}}_{A_{t+1}} = \underbrace{(t-1) w_t}_{A_t} - \frac{1}{\lambda} v_t .$$

$$\Rightarrow \quad A_{t+1} = \sum_{s=1}^{t} -\frac{1}{\lambda} v_s .$$

$\alpha_t \phi(x_{i_t})$.

$$\Rightarrow \quad w_{t+1} = -\frac{1}{\lambda t} \sum_{s=1}^{t} v_s$$

$$\underset{\underset{\mathbb{R}^d}{\|}}{R_T(\Omega)} \le O\left(\frac{\ln T}{T}\right)$$

suppose additionally, ℓ_t's are B-Lip. wrt $\|\cdot\|_2$

f_t's are B-Lip.

then. $\|v_t\|_2 \le B \quad \forall t$

$\Rightarrow \|w_t\|_2 \le \frac{B}{\lambda}$.

$\Rightarrow g_t = \underbrace{\lambda w_t}_{\le B} + \underbrace{v_t}_{\le B}$ satisfies $\|g_t\|_2 \le 2B$.

L

Thm

$\Rightarrow \forall u \in \Omega = \mathbb{R}^d$, $\quad R_T(u) \le \frac{4B^2(\ln T+1)}{2\lambda}$.

online to batch conversion

$\Rightarrow \mathbb{E}\, F_S(\bar{w}_T) - F_S(\hat{w}) \le \frac{2B^2(\ln T+1)}{\lambda T}$.

$\Rightarrow$ setting $T = \tilde{O}\left(\frac{B^2}{\lambda \varepsilon}\right)$ ensures

$$\mathbb{E}\, F_S(\bar{w}_T) - F_S(\hat{w}) \le \varepsilon.$$

kernel methods: a brief introduction

suppose all examples are transformed w/ a nonlinear map: $\phi: \mathbb{R}^d \to \mathbb{R}^N$. N is very large.

goal: find w that approx minimizes

$$\frac{1}{m}\sum_{i=1}^{m} f(y_i \langle \underset{\in \mathbb{R}^N}{w}, \phi(x_i)\rangle) + \frac{\lambda}{2}\|w\|_2^2 .$$

$\ell(w, \phi(x_i), y_i)$

silver lining: assume that $\langle \phi(x), \phi(z)\rangle = k(x,z)$ (kernel fn by ϕ) can be evaluated efficiently.

$$k(x,z) = (1+\langle x,z\rangle)^{\ell}. \quad \text{(polynomial kernel)}$$

the ϕ induced by k will have $N = O(d^{\ell})$.

- can we develop efficient training & test algs w/ running time independent from N?

- key idea:

  keep track of coefficient of w_t. $\alpha_t \in \mathbb{R}^m$

  maintain invariant that $w_t = \sum_{i=1}^{m} \alpha_t(i)\phi(x_i)$.

$(m \ll N)$.

$$w = \sum_{i=1}^{m} \alpha(i) \phi(x_i).$$

prediction:

$$\langle w, \phi(x) \rangle = \sum_{i=1}^{m} \alpha_i \langle \phi(x_i), \phi(x) \rangle$$

$$= \sum_{i=1}^{m} \alpha_i k(x_i, x)$$. can be done efficiently

For $t = 1, 2, \cdots T$.

sample $i_t \sim$ uniform $(\{1 \cdots m\})$.

$\Omega = \mathbb{R}^d$.

$$f_t(w) = \underbrace{\ell(w, (x_{i_t}, y_{i_t}))}_{\ell_t(w)} + \frac{\lambda}{2}\|w\|^2.$$

λ-SC wrt $\|\cdot\|_2$

calculating g_t:

— $v_t \in \partial \ell_t(w_t)$

— $g_t = v_t + \lambda w_t$.

updating w_t:

$$w_{t+1} \leftarrow w_t - \frac{1}{\lambda t}(\lambda w_t + v_t)$$

$$= (1 - \frac{1}{t}) w_t - \frac{1}{\lambda t} v_t.$$

Q: can we modify the alg so that instead of keep w_t's, we keep α_t's.