HW 2 $\|W_i^j\|_1 = B_i$ (Problem 4)

$F_1$ $F_2$ ..

$Rad(F_t) \le Rad(F_{t-1}) \cdot C$

Online (convex) optimization;

- decision set $\Omega$ (action space) often cvx

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

- learner picks $w_t \in \Omega$ ($w_1 f_1$ $w_2 f_2$ .. $w_{t-1} f_{t-1}$)
- environment picks loss fn $f_t : \Omega \to \mathbb{R}$. cvx
- learner suffers loss $f_t(w_t)$,

$w_t = \underset{w \in \Omega}{\operatorname{argmin}} \sum_{s=1}^{t-1} f_s(w)$ Follow-the-leader

$w_t \leftarrow$ update $(w_{t-1}, f_t)$

key performance measure:

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

$R_T(\Omega) = \max_{u \in \Omega} R_T(u)$.

Goal: want $R_T(u)$ to be $o(T)$

connection to statistical learning:

$S = (z_1 \cdots z_T) \overset{iid}{\sim} D$.

$\hat{w} = A(S)$ $\quad \Omega$ : hypothesis class

$\ell(w, z)$

goal: generate $\hat{w}$ s.t. $L_D(\hat{w}) = \mathbb{E}_{z \sim D} \ell(\hat{w}, z)$ small.

excess loss $L_D(\hat{w}) - L_D(w^*)$ $\left(= o(1)\right)$.

$$w^* = \underset{w \in \Omega}{\operatorname{argmin}} L_D(w)$$

Differences:

- online learning does not necessarily assume iid assumptions
- sometimes online learning algs are more computationally efficient ( b/c it uses fast update rule )

Connection: (online to batch conversion)

given online learning alg w/ good regret guarantees, use it to construct a stat. learning alg w/ good excess loss guarantees.

Alg: input $(z_1 \cdots z_T) \overset{iid}{\sim} D$. online learning alg $A$. w/ action space $\Omega$.

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

$A$ outputs $w_t$

$f_t(w) = \ell(w, z_t)$

$w_1 \cdots w_T$ $\qquad \hat{w} = \frac{1}{T} \sum_{t=1}^{T} w_T = \bar{w}$ $L_D(\hat{w})$

Regret guarantees of $A$ on $u$:

$$\frac{R_T(u)}{T} = \frac{1}{T}\sum_{t=1}^{T} \ell(w_t, z_t) - \frac{1}{T}\sum_{t=1}^{T} \ell(u, z_t) \quad = O(1)$$

$\ell(w_t, z_t)$: $\downarrow$ Hoeffding. $\quad$ $\ell(u, z_t)$: $\downarrow$ Hoeffding.

$\frac{1}{T}\sum_{t=1}^{T} L_D(w_t) \qquad L_D(u)$

Assuming $L_D(w)$ is cvx ( $\ell(w, z)$ is cvx $\forall z$ ).

$$L_D(\hat{w}) \le \frac{1}{T}\sum_{t=1}^{T} L_D(w_t) \le L_D(u) + \underbrace{\frac{R_T(u)}{T}}_{\le O(1)} + \text{concentration factors}$$

**Thm**: Assume $\ell(w, z) \in [0, B]$. is cvx in $w$, $\qquad O(1)$.

then w.p. $1-\delta$, $\forall$

$$L_D(\bar{w}) \le L_D(w^*) + \frac{R_T(w^*)}{T} + 2B\sqrt{\frac{2\ln(4/\delta)}{T}} \qquad \forall w^* \in \Omega$$

**Pf**: Hoeffding $\Rightarrow$ w.p. $1-\delta/2$:

$$\left| \frac{1}{T}\sum_{t=1}^{T} \ell(w^*, z_t) - L_D(w^*) \right| \le B\sqrt{\frac{2\ln(4/\delta)}{T}}.$$

$\ell(w_t, z_t)$ may not be indept from $\ell(w_{t+1}, z_{t+1})$

Azuma $\Rightarrow$ w.p. $1-\delta/2$:

$$\left| \frac{1}{T}\sum_{t=1}^{T} \ell(w_t, z_t) - \frac{1}{T}\sum_{t=1}^{T} L_D(w_t) \right| \le B\sqrt{\frac{2\ln(4/\delta)}{T}}.$$

Azuma's inequality: $X_1 \cdots X_T \in [-B, B]$. $(B > 0)$.

$\forall t$, $\mathbb{E}[X_t \mid X_1 \cdots X_{t-1}] \overset{\text{ind}}{=} \mathbb{E}[X_t] = 0$. ( $X_1 \cdots X_T$ is a martingale difference sequence), w.p. $1-\delta$:

$$\left| \sum_{t=1}^{T} X_t \right| \le B\sqrt{2T\ln\frac{2}{\delta}}.$$

Azuma with $X_t = \ell(w_t, z_t) - L_D(w_t) \in [-B, B]$

$\mathbb{E}[X_t \mid$ all observations up to $t-1$. and $w_t] = 0$.

$\Rightarrow \mathbb{E}[X_t \mid X_1 \cdots X_{t-1}] = 0$.

using union bound & algebra ✗.

Ex: $c_1 \cdots c_T \overset{iid}{\sim} U(\pm 1)$.

$X_t$ depend on $c_1 \cdots c_{t-1}$. $\underline{\text{sign}(X_t)}$ $X_t \in [-B, B]$

$|X_t|$.

profit at $t$: $c_t \cdot X_t = Z_t$

$\sum_{t=1}^{T} Z_t \in [\pm B \cdot \sqrt{T}]$. .

pf of Azuma's ineq:

$\mathbb{E}[\sum_{t=1}^{T} X_t] = 0$ why?

M+g difference seq. defn.

$\mathbb{E}[\sum_{t=1}^{T} X_t] = \mathbb{E}_{X_1 \cdots X_{T-1}} \mathbb{E}_{X_T}[\sum_{t=1}^{T} X_t] = \mathbb{E}[\sum_{t=1}^{T+1} X_t] = 0$.

$\mathbb{E}[\ \mid X_1 \cdots X_{T-1}] \cdot X_T \mid X_1 \cdots X_{T-1}$ supported on $[-B, B]$.

verify $\sum_{t=1}^{T} X_t$ is $T \cdot B^2$-SG

$\forall \lambda$, $\mathbb{E}[e^{\lambda \sum_{t=1}^{T} X_t}] = \mathbb{E}_{X_1 \cdots X_{T-1}} \mathbb{E}_{X_T}[e^{\lambda \sum_{t=1}^{T+1} X_t} \cdot e^{\lambda X_T}]$

$X$ supported on $[a, b]$. $\mathbb{E} e^{\lambda Z} \le e^{\frac{\lambda^2}{2} \cdot \frac{(b-a)^2}{4}}$.

$$\le \mathbb{E}\left[ e^{\lambda \sum_{t=1}^{T+1} X_t} \cdot e^{\frac{\lambda^2}{2} B^2} \right]$$

$$\le \cdots$$

$$\le e^{\frac{\lambda^2}{2} \cdot T \cdot B^2}$$

SG $\Rightarrow$ an exponential tail property of SG r.v.'s. ☒

Algorithms for online learning / online optimization.

First algorithm: online gradient descent.

Motivation: $\ell(w, (x,y)) = \ln\left(1 + e^{-y\langle w, x\rangle}\right)$.

given $(x_i, y_i) \overset{iid}{\sim} D$.

can we use online learning + online to batch conversion to develop algs that can output $\hat{w}$ with $L_D(\hat{w})$ small ?

- gradient descent.
- convex fn.

convex set

**Recall**: fn $f$ is **cvx** in domain $\Omega$, if $\forall x, y \in \Omega$,

$$f(\alpha x + (1-\alpha) y) \le \alpha f(x) + (1-\alpha) f(y).$$

$\forall x \in \Omega$. $f(x) \in \mathbb{R}$.

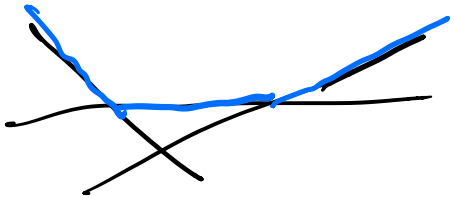
**Ex**: 1. $f(w) = \langle a, w\rangle + b$ cvx.

2. $f(w) = \|w\|_2$.

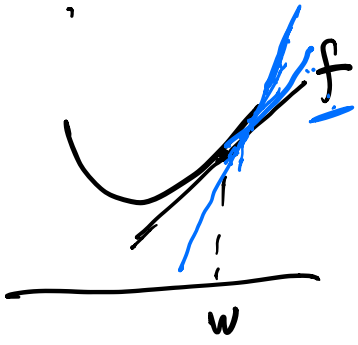
3. $f(w) = \frac{1}{2}\|w\|_2^2$ (1-sc $\Rightarrow$ cvx).

Basic properties of cvx fns:

1. $f, g$ cvx. $\alpha, \beta \geq 0$. $\Rightarrow$ $\alpha f + \beta g$ cvx.
2. $f$ cvx $\Rightarrow$ $g(x) = f(Ax + b)$ cvx.
3. $f_1 \cdots f_n$ cvx $\Rightarrow$ $g(x) = \max_{i=1} f_i(x)$ cvx.



Subgradients:




$f(u) \geq f(w) + \langle \nabla f(w), u - w \rangle$

for all $x$ in the interior of $\Omega$

$f$ is cvx on $\Omega$, then, for almost all pts $x$ in $\Omega$,

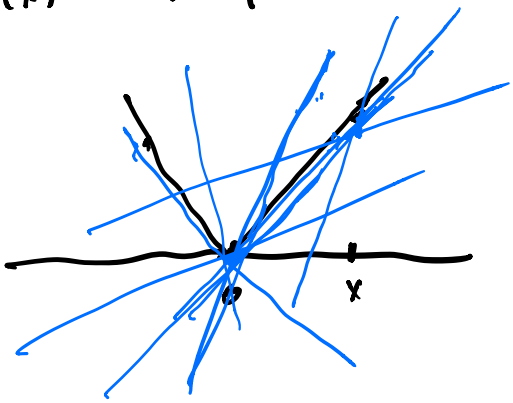
subgradient of $f$ on $x$

$\partial f(x) = \{ g : \forall y \in \Omega, f(y) \geq f(x) + \langle g, y - x \rangle \} \neq \phi$.

when $f$ is differentiable at $x$. $\partial f(x) = \{ \nabla f(x) \}$

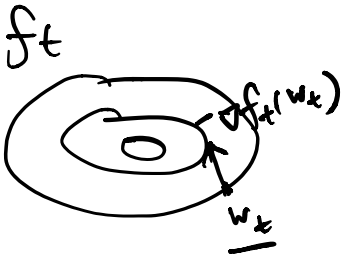
Ex: $f(x) = |x|$

$$\partial f(x) = \begin{cases} \{1\}, & x > 0 \\ \{-1\}, & x < 0 \\ [-1, 1], & x = 0 \end{cases}$$




Fact: $x^* = \underset{x \in \Omega}{\operatorname{argmin}} f(x) \Leftrightarrow 0 \in \partial f(x)$.

Online (sub) gradient descent algorithm:




Initialize $w_1 \in \Omega$.

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

- choose $w_t$.
- Receive $f_t$. suffer loss $f_t(w_t)$
- set $g_t \in \partial f_t(w_t)$
- update: $w'_{t+1} \leftarrow w_t - \eta g_t$.

$\Omega$

$w_{t+1} \leftarrow \Pi_\Omega(w'_{t+1})$

$= \operatorname{argmin}_{w \in \Omega} \| w - w'_{t+1} \|_2$.