- Differences between \( E_6 \) and \( G_0 \)
  
  Focusing mostly on square loss and linear regression

- WC-loss bounds

- Kernel trick

- One-norm versus two-norm regularization
TARGET \(1000\)

\[x \in \{-1,1\}^n; k=1; n=4\]

- \(GD\)
- \(\mathcal{E}G^2\)

**TOTAL LOSS**

**TRIAL NUMBER**
TARGET \((1, 0, 0, \ldots, 0)_{100}\)

\(x \in \{-1, 1\}^n; k=1; n=100\)

TOTAL LOSS

EG^\frac{1}{2}

TRIAL NUMBER
TARGET \((1,0,0,0,\ldots,0,0,0)\)

\(x \in \{-1,1\}^n; k=1; n=1000\)

\[\text{EG}^2\]
$x \in \{-1,1\}^n; k=1; n=1000$
GD weights; $x \in \{-1, 1\}^n$; $k=1$; $n=100$
\[ E^{\dagger} \] weights; \( x \in \{-1,1\}^n \); \( k=1 \); \( n=100 \)
Large random vectors in \( \{-1, 1\}^n \) are approximately orthogonal:

\[ \mathbf{x} \cdot \mathbf{x}' \approx 0 \]

Hadamard matrices

\[ H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad H_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \]

\[ H_t = \begin{pmatrix} H_{t-1} & H_{t-1} \\ H_{t-1} & -H_{t-1} \end{pmatrix} \]

All rows are orthogonal!

Use rows as instances!
TARGET $(1, 0, 0, \ldots, 0)$

Hadamard, $n=128$, $k=1$

GD & LLS
$x \in \{-1, 1\}^n$; $n=20$; $k=3$
$x \in \{-1, 1\}^n$; $n=20$; $k=6$

$G_0$

$E_0^\perp$
$L_p$ NORMS

$\| \mathbf{x} \|_p = \left( |x_1|^p + |x_2|^p + \ldots + |x_n|^p \right)^{\frac{1}{p}}$

$\| \mathbf{x} \|_2 = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$

$\| \mathbf{x} \|_1 = |x_1| + |x_2| + \ldots + |x_n|$

$\| \mathbf{x} \|_\infty = \max (|x_1|, |x_2|, \ldots, |x_n|)$

$\| \mathbf{x} \|_\infty \leq \| \mathbf{x} \|_2 \leq \| \mathbf{x} \|_1$

$p$-NORM \& $q$-NORM ARE DUAL IF

$\frac{1}{p} + \frac{1}{q} = 1$

$2$-NORM IS SELF-DUAL

$\infty$-NORM \& 1-NORM ARE DUAL
INCOMPARABLE LOSS BOUNDS

EXAMPLE SEQUENCE:

$S = (\bar{x}_1, y_1), \ldots, (\bar{x}_t, y_t)$, where $y_t = \bar{u} \cdot \bar{x}_t$

$\text{L}_{GD}(S) \leq (\|\bar{u}\|_2 \max_t (\|\bar{x}_t\|_2))^2$

$\text{L}_{EG^\dagger}(S) \leq (\|\bar{u}\|_1 \max_t (\|\bar{x}_t\|_\infty))^2 \ln (2n)$

THE PRODUCTS OF THE TWO PAIRS OF DUAL NORMS ARE INCOMPARABLE

\[ \begin{array}{ccc}
\|\bar{u}\| & \|\bar{x}_t\| \\
1 & \text{GD} & 2 \\
2 & \text{EG}^\dagger & \infty
\end{array} \]

TARGET NORMS
GD'S IS SMALLER

INSTANCE NORMS
EG'S IS SMALLER

Experiments so far

$1 \approx 2$

Sparse

\( \bar{x}_t \in \{-1,1\}^n \)
GD HAS ADVANTAGE

1

2 = ∞

DENSE

\( \bar{u} = (1, 1, \ldots, 1) \)

\( \bar{x}_\epsilon \) ARE UNIT VECTORS
$\text{norm}(x, 2) = 1$, $\text{norm}(\text{target}, 1) = 1$; $n = 20$, $k = 10$
$\text{norm}(x, 2) = 1, \text{norm}(\text{target}, 1) = 1; n = 20, k = 20$
SUMMARY

EG⁺ BETTER WHEN

INSTANCES ARE "DENSE" \((X₀ ≤ X₂)\)

AND BEST WHEN WEIGHT VECTOR IS "SPARSE"
\((U₁ ≈ U₂)\)

GD BETTER WHEN

INSTANCES ARE "SPARSE" \((X₂ ≈ X₀)\)

AND BEST WEIGHT VECTOR IS "DENSE"
\((U₂ ≪ U₁)\)

PRODUCTS OF DUAL NORMS

\[ EG⁺ = (U, X₀)^2 \]

\[ GD = (U₂ X₂)^2 \]
EXAMPLE

\[ \bar{u} = (1, 1, 1, 0, 0, 0, 0, 0) \]

\[ \bar{x}_t \in \{1, -1\}^N \]

\[ \| \bar{u} \|_2 = \sqrt{k} \]
\[ \| \bar{x}_t \|_2 = \sqrt{N} \]

\[ \| \bar{u} \|_1 = k \]
\[ \| \bar{x}_t \|_\infty = 1 \]

\[ \text{LGD}(S) \leq \lambda (\text{L} \bar{u}(S) + kN) \]

\[ \text{LEG}(S) \leq \lambda' (\text{L} \bar{u}(S) + k^2 \text{log}N) \]

LINEAR VERSUS LOGARITHMIC GROWTH IN N
WHEN E6 BEATS GD & VICA VERSA

ASSUME $L_4(S) = 0$

$x_4 \in \mathbb{Z}^{N}$, $u = e_i$

I

$\|x\|_2 = \sqrt{N}$
$\|u\|_2 = 1$
$\|x\|_2^2 \|u\|_2^2 = N$  GD

$\|x\|_\infty = 1$
$\|u\|_\infty = 1$
$\|x\|_\infty^2 \|u\|_\infty^2 \log N = \log(N)$  E6±

$x_4 = 0$
$u = \{ \pm 1 \}^N$

II

$\|x\|_2 = 1$
$\|u\|_2 = \sqrt{N}$
$\|x\|_2^2 \|u\|_2^2 = N$  GD

$\|x\|_\infty = 1$
$\|u\|_\infty = N$
$\|x\|_\infty^2 \|u\|_\infty^2 \log N = N^2 \log N$  E6±

$\frac{GD}{E6^\pm} = \frac{N}{\log N}$

$I$ $\frac{E6^\pm}{GD} = \frac{N^2 \log N}{N} = N \log N$

WARNING: THE BOUNDS MIGHT BE WEAK!
THE GENERAL BOUNDS

\( \forall \text{sequences } S \text{ with } \|x_t\|_2 \leq x_2 \)

\[
\log(...)(S) \leq \\
\inf_{\|\bar{u}\|_2 \leq u_2} \left( L\bar{u}(S) + a \sqrt{L\bar{u}(S)} \right) u_2 x_2 + b u_2^2 x_2^2
\]

\( \forall \text{sequences } S \text{ with } \|x_t\|_{\infty} \leq x_{\infty} \)

\[
\log(...)(S) \leq \\
\inf_{\|\bar{u}\|_1 \leq u_1} \left( L\bar{u}(S) + c \sqrt{L\bar{u}(S)} \right) u_1 x_{\infty} \sqrt{\log N} + d u_1^2 x_{\infty}^2 \log N
\]
EXPAND

\((x_1, \ldots, x_{20}) \quad (x_1, \ldots, x_{20}, \ldots, x_i x_j \ldots x_i x_j x_k \ldots)\)

20 vars, degree 3 \((p=1771), k=2\)

WORST-CASE
BOUND

TARGET = \(\frac{1}{2} x_2 x_{12} x_{13} + \frac{1}{2} x_2 x_{18} x_{20}\)
$E^+$ has good WC loss bound but needs to maintain one weight per base function in many cases. The GD update for expanded instances can be made efficient.
GD: \( w_t = \sum_{q} x_q \phi(x_q) \)

\( \hat{y}_t = \phi(x) \cdot w_t \)

\( = \phi(x) \cdot \sum_{q} x_q \phi(x_q) \)

\( = \sum_{q} \phi(x) \cdot \phi(x_q) \)

\( \underbrace{\text{k}(x, x_q)}_{\text{kernel}} \)

Only need efficient way to compute dot products via kernel function.

\( x \quad \phi(x) \)

\((x_1, x_2, x_3) \rightarrow (1, x_1, x_2, x_3, x_1 x_2, x_1 x_3, x_2 x_3, x_1 x_2 x_3)\)

8 features

\( \phi(x) \cdot \phi(z) = 1 + x_1 z_1 + x_2 z_2 + x_3 z_3 + \)

\( + x_1 x_2 z_1 z_2 + x_1 x_3 z_1 z_3 + x_2 x_3 z_2 z_3 + \)

\( + x_1 x_2 x_3 z_1 z_2 z_3 \)

\( = (1 + x_1 z_1) \cdot (1 + x_2 z_2) \cdot (1 + x_3 z_3) \)

8 terms: \( z > z \cdot z \)
\[ \phi(x) \]

\[ (x_1, x_2, \ldots, x_n) \]

\[ (1, x_i, x_i x_j, x_i x_j x_k, \ldots) \]

\[ \forall i \neq j \quad \forall i \neq k \]

\[ 2^m \text{ MONOMIALS} \]

\[ \phi(x) \cdot \phi(z) = \sum_{I \subseteq \{1, \ldots, n\}} \prod_{i \in I} x_i \cdot \prod_{i \notin I} z_i \]

\[ = \prod_{i=1}^{n} (1 + x_i z_i) \]

\[ O(2^n) \text{ TIME} \]

**EFFICIENCY:**

\[
\left( \sum_{q=1}^{2^t} \phi(q) \cdot \phi(x_t) \right) \cdot \phi(x_t) \]

\[ w_t \]

**DIMENSION** \( 2^n \)

**SEEMINGLY TIME** \( O(2^n) \)

\[
= \sum_{q=1}^{2^t} \phi(q) \cdot \phi(x_t) \]

\[
\text{TIME } O(1 + n) \]

\[
\text{TIME } O(t + n) \]
MORE EXAMPLES

\[(x_1, \ldots, x_n) \rightarrow (\sum_{i \leq j \leq n} x_i x_j)\]

WITH REPEATS

\[\phi(\bar{x})\]

\[\phi(\bar{x}) = \sum_{(i,j)=(1,1)}^{(n,n)} (x_i x_j)(z_i z_j)\]

\[= \left(\sum_{i=1}^{n} x_i z_i\right)^2 \left(\sum_{j=1}^{n} x_j z_j\right)\]

\[= \left(\sum_{i=1}^{n} x_i z_i\right)^2 \left(\sum_{j=1}^{n} x_j z_j\right)\]

\[= (x, z)^2\]
### Families of update algorithms

<table>
<thead>
<tr>
<th>Parameter divergence</th>
<th>Name of family</th>
<th>Update algs.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>expert algs Normalized Winnc “AdaBoost”</td>
</tr>
</tbody>
</table>
Families of update algorithms (cont)

\[ \sum_{i=1}^{n} w_i \ln \frac{w_i}{w_{t,i}} + w_{t,i} - w_i \]

Unnormalized Winnow
Exp. Grad. Alg.

\[ \sum_{i=1}^{n} w_i \ln \frac{w_i}{w_{t,i}} + (1 - w_i) \ln \frac{1-w_i}{1-w_{t,i}} \]

Binary
Exp. Grad. Alg.

any Bregman divergence

Members of different families exhibit different behavior

Good loss functions are often also Bregman Divergence

LATER
DEEP QUESTION?
CAN EC BE KERNELIZED

WHY IMPORTANT?
- EC EXPLOITS DENSE Instances / SPARSE TARGET
- KERNEL TRICK MANIPULATES FEATURES EFFICIENTLY

YES: EC CAN BE KERNELIZED IN MATRIX CASE LATER CLASS
RELATIVE LOSS BOUNDS FOR ONLINE
IN THE CASE OF LINEAR REGRESSION

Loop

\[ x_t, \]
\[ \hat{y}_t = w_t \cdot x_t, \]
\[ y_t \]
\[ (y_t - \hat{y}_t)^2, \]
\[ w_{t+1} = w_t - \eta \cdot 2(w_t x_t - y_t)x_t \]

PROOF OUTLINE

\[ \|u - w_t\|^2 - \|u - w_{t+1}\|^2 \geq \alpha (w_t x_t - y_t)^2 - \beta (u x_t - y_t)^2 \]

Progress towards comparator

Loss of all \( \bar{u} \)

\[ 2^{t+1} \geq \alpha \text{ TOTAL LOSS} - \beta \text{ TOTAL LOSS} \]

of \( G_0 \)

of \( u \)

\[ \|u - w_{t+1}\|^2 - \|u - w_t\|^2 \geq \alpha L_D(s) - \beta L_u(s) \]

\[ \leq L_D(s) \leq \frac{\beta}{\alpha} L_u(s) = \frac{\|u - w_t\|^2}{\alpha} \]

\( \alpha, \beta, \beta \) will depend on \( \eta \)
> #relative loss bound for GD
> progr:=(u-w)^2/2-(u-w+2*eta*(w*x-y)*x)^2/2;

\[
progr := \frac{1}{2} (u - w)^2 - \frac{1}{2} (u - w + 2 \eta (w x - y) x)^2
\]

> p1:=factor(progr);

\[
p1 := -2 \eta x (-w x + y) (\eta x y - \eta x^2 w - u + w)
\]

> p2:=-2*eta^2*(w*x-y)^2*x^2-2*eta*(u*x-w*x)*(w*x-y);

\[
p2 := -2 \eta^2 (w x - y)^2 x^2 - 2 \eta (u x - w x) (w x - y)
\]

> factor(progr-p2);

0

> #ew:=wx-y; eu=ux-y
> p3:=-2*eta^2*ew^2*x^2-2*eta*(eu-ew)*ew;

\[
p3 := -2 \eta^2 ew^2 x^2 - 2 \eta (eu - ew) ew
\]

> factor(progr-sub(p3));

0

> #to show that \( f \leq 0 \)
> f:=-p3+a*ew^2-b*eu^2;

\[
f := 2 \eta^2 ew^2 x^2 + 2 \eta (eu - ew) ew + a ew^2 - b eu^2
\]

> #replace eu by worst-case eu
> diff(f,eu$2); #f(eu) cupped downward

-2 b

> soleu:=solve(diff(f,eu)=0,eu);

\[
soleu := \frac{\eta \text{ew}}{b}
\]

> g:=factor(subs(eu=soleu,f));
\[ g := \frac{\text{ew}^2 \left( 2 \eta^2 x^2 b + \eta^2 - 2 \eta b + a b \right)}{b} \]

> diff(g,eta$2);#g(eta) cupped upward
\[ \frac{\text{ew}^2 \left( 4 x^2 b + 2 \right)}{b} \]

> soleta:=solve(diff(g,eta)=0,eta);
\[ \text{soleta} := \frac{b}{2 x^2 b + 1} \]

> h:=factor(subs(eta=soleta,g));
\[ h := \frac{(2 a x^2 b + a - b) \text{ew}^2}{2 x^2 b + 1} \]

> sola:=solve(h=0,a);
\[ \text{sola} := \frac{b}{2 x^2 b + 1} \]

> #a and b as functions of eta
> sola:=eta;
\[ \text{sola} := \eta \]

> solb:=solve(soleta=eta,b);
\[ \text{solb} := \frac{\eta}{1 - 2 \eta x^2} \]

> bound:= Lu*solb/sola + (ED^2/2) / sola;
\[ \text{bound} := \frac{Lu}{1 - 2 \eta x^2} + \frac{1}{2} \frac{\eta \text{ED}^2}{\eta} \]

> factor(diff(bound,eta$2));
\[ (-8 \text{Lu} x^4 \eta^3 - \text{ED}^2 + 6 \text{ED}^2 \eta x^2 - 12 \text{ED}^2 \eta^2 x^4 + 8 \text{ED}^2 \eta^3 x^6) \]
\[
\frac{1}{(-1 + 2\eta x^2)^3}\eta^3
\]

> Soleta := solve(diff(bound,eta)=0,eta);

\[
Soleta := \frac{1}{2} \left( -4 ED^2 x^2 + 4 ED x \sqrt{Lu} \right) \cdot \frac{1}{2} \left( -4 ED^2 x^2 - 4 ED x \sqrt{Lu} \right)
\]

> SOLETA := ED/(2*x)/(sqrt(Lu)+ED*x);

\[
SOLETA := \frac{1}{2} \frac{ED}{x \left( ED x + \sqrt{Lu} \right)}
\]

> factor(Soleta[1]-SOLETA);

0

> bound2 := factor(subs(eta=factor(SOLETA),bound));

\[
bound2 := \frac{\left( ED x + \sqrt{Lu} \right) \left( Lu + ED x \sqrt{Lu} \right)}{\sqrt{Lu}}
\]

> finbound := (sqrt(Lu)+ED*x)^2; expand("");

\[
finbound := \left( ED x + \sqrt{Lu} \right)^2
\]

\[
ED^2 x^2 + 2 ED x \sqrt{Lu} + Lu
\]

> factor(subs(eta=SOLETA,diff(bound,eta$2)));#thus min. at Soleta

\[
8 x^3 \left( ED x + \sqrt{Lu} \right)^3 \left( Lu ED x + Lu^{3/2} \right) \frac{Lu^{3/2}}{ED}
\]

>
\[ E_6 \quad w_{t+1, i} = w_{t, i} e^{-2\eta (u_{t+1} - u_t) x_{t, i}} \]

\[ \text{NORMAL.} \]

\[ \text{DIVERGENCE: } \Delta (u, w_t) = \sum_i u_i \ln \frac{u_i}{w_{t, i}} \]

\[ \Delta (u, w_t) - \Delta (u, w_{t+1}) = \sum_i u_i \ln \frac{w_{t+1, i}}{w_{t, i}} \]

\[ = \sum_i u_i \ln \frac{e^{-\eta \nabla_i}}{\sum_k e^{-\eta \nabla_k}} \]

\[ = -2\eta (w_{t+1} - w_t) u_t - e^{\eta \nabla_t} \sum_i w_{t, i} e^{-\eta \nabla_i} \]

\[ \text{QUADRATIC APPROX.} \]

\[ \text{MAPLE} \]

If \[ \|x\|_{\infty} \leq X \infty \]

\[ \inf \{ u_n(t) : \|u_n(t)\|_{\infty} \leq K \} \]

Then with properly tuned \( \eta(u, t, X, \infty) \)

\[ L_2(t) \leq L_2(t) + c_1 \sqrt{k} ||u, x_{\infty}|| \sqrt{\ln n} \]

\[ + c_2 \frac{u, x_{\infty}}{(u, x_{\infty})^2} \ln n \]

\[ \text{DUAL NORM} \]

\[ E_6 \text{ can also be generalized to the matrix case.} \]