CS 2416 - Machine Learning Theory

Lecture #10: 17 November 2005

Lecturer: Toniann Pitassi

Scribe Notes by: Benjamin Bartlett

Boosting and Hard Core Sets

Hardness Amplification

Definition Given a boolean function $f : \{0,1\}^n \rightarrow \{0,1\}$ that is mildly inapproximable by size g circuits, construct from f a new function f' that is highly inapproximable by all size g' circuits.

This is crucial to derandomization attempts for BPP.

The most famous hardness amplification problem is Yao's XOR lemma, which states that if f is mildly inapproximable, then $f(x_1) \oplus f(x_2) \oplus \cdots \oplus f(x_k)$ is highly inapproximable.

Impagliazzo '95 reduces problem of amplifying hardness of f to problem of constructing a distribution $\mathcal{D}$ such that f is high inapproximable by size g' circuits for inputs chosen according to $\mathcal{D}$.

Definition Let f be a boolean function on $\{0,1\}^n$ and $\mathcal{D}$ a distribution on $\{0,1\}^n$. Let $0 < \epsilon < \frac{1}{2}$, and $n \leq g \leq \frac{2^n}{n}$. f is ϵ-hard for g under $\mathcal{D}$ if for any boolean circuit C of size at most g, $Pr_{\mathcal{D}}[f(x) = c(x)] \leq 1 - \epsilon$.

Definition A measure on $\{0,1\}^n$ is a function $M : \{0,1\}^n \rightarrow [0,1]$
The absolute size of M, $|M| = \sum_x M(x)$
The relative size of M, $\mu(M) = \frac{|M|}{2^n}$

Definition $\mathcal{L}_\infty(\xi) = \max_x |\xi(x)|$, where ξ is a real valued function.
$\log(\mathcal{L}_\infty(\mathcal{D})^{-1})$: max entropy of $\mathcal{D}$

Correspondence betwen measures and distributions
$M \Rightarrow \mathcal{D}_M : \mathcal{D}_M = \frac{M(x)}{|M|}$
$\mathcal{D} \Rightarrow M_{\mathcal{D}} : M_{\mathcal{D}}(x) = \frac{\mathcal{D}(x)}{\mathcal{L}_\infty(\mathcal{D})}$ (so some points have measure 1, and some 4)
$|M_{\mathcal{D}}| = \frac{1}{\mathcal{L}_\infty(\mathcal{D})}$
Large measure $\Leftrightarrow$ distributions which do not assign large weight to any point $\Leftrightarrow$ high minimum entropy

Definition f is γ-hardcore on M for size g of $Pr_{\mathcal{D}_M}[f(x) = c(x)] \leq \frac{1}{2} + \gamma$ for every circuit C of size at most g. For $S \subseteq \{0,1\}^n$, we say f is γ-hardcore on S for size g if f is γ-hardcore on M_S for g, $M_S(x) =$ the characteristic function of S.

Impagliazzo: new proof of XOR lemma

Proof sketch:

Theorem 1 (ϵ-hard $\Rightarrow$ γ-hardcore measure) *Let f be ϵ-hard for circuits of size g under U (the uniform distribution), and let $0 < \gamma < \frac{1}{2}$. Then $\exists M$, $\mu(M) \geq \epsilon$ such that f is γ-hard on M for size $g' = O(\epsilon^2\gamma^2)g$.*

Theorem 2 (γ-hardcore measure $\Rightarrow$ large hardcore set) *Let f be γ-hardcore on M for size g' ($g' \leq \frac{1}{2n}2^n\gamma^2\epsilon^2$), $\mu(M) \geq \epsilon$. Then $\exists S$, $|S| \geq \frac{\epsilon}{2}2^n$ such that f is 4γ-hardcore on S for g'.*

Theorem 3 *Let $\epsilon' = \frac{\epsilon}{2}$, and let $\gamma' = 4\gamma$. Then if f is γ'-hardcore for some set of $\epsilon' 2^n$ inputs of size g', then $\overline{f}(x_1, \cdots, x_n) = f(x_1) \oplus \cdots \oplus f(x_n)$ is $(\gamma + (1-\epsilon)^k)$-hardcore for size g.*

Intuition of Proof: if we ask several questions of a somewhat hard problem then with high probability one is a hard instance, and hence XOR is hard.

Proof of Theorem 3
Assume for the sake of contradiction that C is a circuit of size at most g that achieves an advantage of more than $\gamma' + (1-\epsilon')^k$ for computing XOR of k independent copies of f.

Let A_l = C's advantage given that exactly l instances are from hardcore set.

$Adv = LHS = A_0 Pr(O) + A_1 Pr(1) + \cdots + A_k Pr(k) \geq \gamma' + (1-\epsilon')$

$Pr(0) = (1-\epsilon')^k$ (no instance from hardcore set)

If all $A_1, \cdots, A_k$ are $< \gamma'$ then $LHS < (1-\epsilon')^k + \gamma'$, a contradiction

Therefor: $\exists\, l \geq 1$ such that $A_l \geq \gamma'$
Let C' be a circuit constructed probabilistically as follows:

Pick $a_1, \cdots, a_l$ uniformly from hardcore

Pick $b_1, \cdots, b_{k-l}$ uniformly from non-hardcore

$C' = \tilde{C}(\pi(a_1, \cdots, a_{l-1}, b_1, \cdots, b_{k-l}, x))$ where π is a random permutation, and

$$\tilde{C} = \begin{cases} C & : \; f(a_1) \oplus \cdots \oplus f(a_{l-1}) \oplus f(b_1) \oplus \cdots \oplus f(b_{k-l}) \oplus f(x) = 0 \\ \overline{C} & : \; otherwise \end{cases}$$

Average advantage over a_i, b_j, π, x from hardcore set is $A_l \geq \gamma'$
So $\exists C'$ with fixed a_i, b_j, π achieving at least average, size(C') $\approx$ size(C) $\Rightarrow$ contradiction to fact that S is hardcore.

Proof of Theorem 2
Let C be a size $\leq g$ circuit.
Consider the following randomized construction of a set $S \subseteq \{0,1\}^n$:

For each $x \in \{0,1\}^n$, $x \in S$ with prob $M(x)$.

Let M_S be the characteristic function of S.

1. For each $x \in 0,1^n$, the expected value of $M_S(x)$ is $M(x)$. Thus, by linearity of expectation,

$$E\left[\sum_{x \in \{0,1\}^n} M_S(x)C(x)f(x)\right] = \sum M(x)C(x)f(x) \leq 2\gamma|M|$$

Note that for each x, $M_S(x)C(x)f(x)$ is a random var in $[-1,1]$.

Hoeffding sizes:

$$\begin{aligned} Pr\left[\frac{1}{2^n}\sum M_S(x)C(x)f(x) \geq 4\delta\mu(M)\right] &\leq exp\left(\frac{-2\cdot 2^n(2\gamma\mu(M))^2}{4}\right) \\ &\leq exp\left(\frac{-2^n\epsilon^2}{2}\right) \end{aligned}$$

where $\mu(M) = \frac{|M|}{2^n}$, and $\mu(M) \geq \epsilon$.
The number of boolean circuits of size $\leq g$ is $\leq ((n+5)g^2)^g$.
By union bound,

$$Pr\left[\exists C,\ |C| \leq g \textit{ such that } \sum_{x\in\{0,1\}^n} M_s(x)C(x)f(x) \geq 4\gamma|\mu|\right] < \frac{1}{10}$$

2. $E[\ |S|\ ] = |M|$ by Hoeffding.

$$\begin{aligned} Pr\left[\frac{|s|}{2^n} < \frac{\mu(M)}{2}\right] &\leq exp\left(-2\cdot 2^n\left(\frac{\mu(M)}{2}\right)^2\right) \\ &\leq exp\left(\frac{-2^n\epsilon^2}{2}\right) \end{aligned}$$

for $\epsilon \geq \frac{1}{2^{\frac{n}{2}}}$, $Pr\left[\ |S| \geq \frac{|M|}{2}\right] > \frac{1}{10}$

By 1 and 2:
$\exists S$, $|S| \geq \frac{|M|}{2} \geq \frac{\epsilon}{2}2^n$ such that for every circuit C with at most g gates

$$\sum_{x\in\{0,1\}^n} M_S(x)C(x)f(x) \leq 4\gamma|M| \leq 8\gamma|S| = 8\gamma|M_S|$$

Fact $Pr_{D_M}[C(x) = f(x)] = \frac{1}{2} + \rho$ iff $\sum_{x\in\{1,0\}^n} M(x)C(x)f(x) = 2\rho|M|$ because

$$\begin{aligned} \frac{\sum M(x)}{|M|}C(x)f(x) &= \textit{probability correct} - \textit{probability incorrect} \\ &= \frac{1}{2} + \rho - (1 - \frac{1}{2} - \rho) \\ &= 2\rho \end{aligned}$$

Therefor: $Pr_{x\in S}[C(x) = f(x)] \leq \frac{1}{2} + 4\gamma$ for every C $\Rightarrow$ f is 4γ hardcore on S.

Russel's Algorithm (IHA) for Theorem 1

Input: $0 < \epsilon < 1$, $0 < \gamma < \frac{1}{2}$, f
Output: circuit h such that $Pr_u]h(x) = f(x)] > 1 - \epsilon$

$i \leftarrow 0$
$M_0(x) = 1$
until $\mu(M_i) \leq \epsilon$ do:

> Let C_i be a size $\leq g'$ circuit such that $Pr_{D_{M_i}}[C_i(x) = f(x)] \geq \frac{1}{2} + \gamma$
>
> Set $R_{C_i} = \begin{cases} 1 & : \ f(x) = C_i(x) \\ -1 & : \ otherwise \end{cases}$
>
> $N_i(x) = \sum_{j=0}^{i} R_{C_j}(x) =$ margin by which $maj(C_0, \cdots, C_i)$ correctly predicts $f(x)$
>
> Set $M_{i+1} = \begin{cases} 1 & : \ N_i(x) = 0 \\ 1 - \epsilon\gamma N_i(x) & : \ 0 \leq N_i(x) \leq \frac{1}{\epsilon\gamma} \\ 0 & : \ N_i(x) > \frac{1}{\epsilon\gamma} \end{cases}$
>
> (The above sets the weight to 0 if the margin is large, 1 if the margin is negative, and somewhere in between if the margin is positive but small)
>
> $i = i + 1$

return $h = majority(C_0, C_1, \cdots, C_{i-1})$

Claim After at most $i_0 = O\left(\frac{1}{\epsilon^2\gamma^2}\right)$ cycles through the loop, $\mu(M_i) < \epsilon$.

Proof Assume for sake of contradiction that $\forall$ measure M with $\mu(M) \geq \epsilon$, $\exists C_M$ of size at most g' such that $Pr_{D_M}[f(x) = C_N(x)] > \frac{1}{2} + \gamma$.
Run IHA. By claim, when we exit loop, h has size $\leq g'i_0 + O(i_0) \leq g$ gates and $Pr_{\mathcal{U}}[f(x) = h(x) \geq 1 - \mu(M_i) < 1 - \epsilon$ contradicting the assumption that f is ϵ-hard for size g circuits under $\mathcal{U}$.

Canonical boosting in general for hardcore set constructions:

A canonical booster is a boosting algorithm with the following structure:

- At stage 0, algorithm starts with $\mathcal{D}_0 = \mathcal{D}$ and uses a weak learner to generate an approximator h_0 for f under $\mathcal{D}$ ($\frac{1}{2} - \gamma$ approximator)
- stage i
  - constructs distribution $\mathcal{D}_i$, favoring points where previous hypotheses $h_0, \cdots, h_{i-1}$ do poorly.
  - simulates example oracle $EX(f, \mathcal{D}_i)$ and lets the weak learner access simulated example oracle to produce hypothesis h_i, an approximator for f under D_i
- At end, combines $h_0, \cdots, h_{i-1}$ to obtain the final hypothesis h such that if each h_i is a $(\frac{1}{2} - \gamma)$ approximator for f under $\mathcal{D}_i$ then h is an ϵ-approximator for f under $\mathcal{D}$

2 important properties of boosting for hardcore set construction:
number of iterations and smoothness

Definition A distribution $\mathcal{D}$ is d-smooth if $\mathcal{L}_\infty(2^n\mathcal{D}) \leq d$

Fact If $\mathcal{D}$ is d-smooth, then $\mu(M_\mathcal{D}) \geq \frac{1}{2}$

Definition Let B be a canonical boosting algorithm which takes as input ϵ, δ, γ, $EX(f,\mathcal{D})$ and a $(\frac{1}{2}-\gamma)$-approximator learning algorithm, WL, for f.

1. B is a $K(\epsilon,\gamma)$ stage boosting algorithm if:
   $\forall$ example oracles $EX(f,\mathcal{D})$, and $(\frac{1}{2}-\gamma)$-approximator learners WL for f, B simulates $k = K(\epsilon,\gamma)$ distribnutions $\mathcal{D}_0,\cdots,\mathcal{D}_{k-1}$ for WL and uses WL to generate k hypotheses $h_0,\cdots,h_{k-1}$

2. B is $d(\epsilon,\gamma)$ smooth if:
   For all f and $(\frac{1}{2}-\gamma)$-approximator learners WL, when B is given $EX(f,\mathcal{U})$ and WL with probability $1-\gamma$:
   (a) The simulated distributions $\mathcal{D}_0,\cdots,\mathcal{D}_{k-1}$ are all $d(\epsilon,\gamma)$ smooth
   (b) The hypothesis h output by B satisfies $Pr_\mathcal{U}[f(x)=h(x)] \geq 1-\epsilon$

Theorem 4 (Klivans, Servedio) *Let B be a $K(\epsilon,\gamma)$ stage, $d(\epsilon,\gamma)$ smooth boosting algorithm which outsputs as its final hypothesis a circuit of size r over inputs $h_0,...,h_{k-1}$. Let f be ϵ-hard for circuits of size g under (U) and let $0<\gamma<\frac{1}{2}$. Then $\exists$ a measure M on $\{0,1\}^n$ such that $\mu(M) \geq \frac{1}{d(\epsilon,\gamma)}$ such that f is γ-hardcore on M for size $g' = \frac{g-\gamma}{K(\epsilon,\gamma)}$*

Proof Assume for sake of contradiction that $\forall$ measures M with $\mu(M) \geq \frac{1}{d(\epsilon,\gamma)}$ $\exists C_M$ of size $\leq g'$ such that $Pr_{D_M}[f(x)=C_M(x)] \geq \frac{1}{2}+\gamma$

This implies that $\forall$ $d(\epsilon,\gamma)$ smooth distributions $\mathcal{D}$, $\exists$ circuit C_D of size $\leq g'$ such that $Pr_D[f(x)=C_D(x)] \geq \frac{1}{2}+\gamma$

Run B on inputs ϵ, δ, γ, and $EX(f,\mathcal{D})$
Since B is $d(\epsilon,\gamma)$ smooth, with probability $1-\delta$:

1. each $\mathcal{D}_i$ which B simulates is $d(\epsilon,\gamma)$ smooth
2. the final hypothesis h is an ϵ-approximator to f under $\mathcal{U}$

By (1), $\exists$ C_i of size at most g' which is a $(\frac{1}{2}-\gamma)$-approximator for f under $\mathcal{D}_i$.
Give B this circuit when it calls WL on $\mathcal{D}_i$.

By (2), h is an ϵ-approximator to f under $\mathcal{U}$. But since B is a $K(\epsilon,\gamma)$ stage, this h is a circuit of size $r+g'K(\epsilon,\gamma) \leq g$, contradicting our assumption that f is ϵ-hard for size g circuits under $\mathcal{U}$.

Results:

Given f such that no circuit of size g computes f correctly on more than $(1-\epsilon)2^n$ inputs then $\forall\ \gamma < \frac{1}{2}$, $\exists\ S$, $|S| = \epsilon 2^n$ such that no circuit of size g' can correctly computer f on more than $(\frac{1}{2}+\gamma)$ fraction of inputs in S.

Author	$\lvert S\rvert$	$\lvert g'\rvert$
Impagliazzo ('95)	$\omega(\epsilon)$	$O(\gamma^2\epsilon^2)g$
Msan (Imp. '95)	$\omega(\epsilon)$	$O(\gamma^2(\log(\frac{1}{\gamma\epsilon}))^{-1})g$
Klivers, Suredio	$\omega(\epsilon)$	$O(\gamma^2(log(\frac{1}{\epsilon}))^{-1})g$

In opposite direction, can use Impagliazzo's construction as a boosting procedure. This gives a (small) improvement to learning DNF under $\mathcal{U}$ with membership queries.

<u>Jackson</u> '97 $\tilde{O}(\frac{ns^{10}}{\epsilon^2}) \longrightarrow$ <u>Bshonty, Jackson, Tamon</u> '99 $\tilde{O}(\frac{ns^6}{\epsilon^4})$

<u>Klivans, Suredio</u> $\tilde{O}(\frac{ns^6}{\epsilon^2})$

n = number of variables, s = number of terms, ϵ = accuracy