Today

- Prove two lemmas from last time
- Finish first unit on PAC learning via PTF dag upper bounds
- Start second unit on uniform-dist PAC
- Fourier analysis of Boolean functions
\[ P_t(x) = (x-1)(x-2)^2(x-3)^2 \cdots (x-2^t)^2 \]

**Lemma:** \( 1 \leq x \leq 2^t \)  \( 4P_t(x) \leq -P_t(-x) \)

**Proof:**
\[
P_t(-x) = (x+1)(-x-2)^2(-x-3)^2 \cdots (-x-2^t)^2
\]

\[ P_t(x) \leq P_t(-x) \text{ term-wise.} \]

It suffices to show one \( k \) s.t.

\[
4\left(x-2^k\right)^2 \leq \left(-x-2^k\right)^2
\]

Pick \( k \) s.t. \( 2^{k-1} \leq x < 2^k \)

\[
4\left(x-2^k\right)^2 = 4(2^k-x)^2 \leq 4\left(2^{k-1}\right)^2 = 2^{2k} \leq (x+2^k)^2 = \left(-x-2^k\right)^2
\]
\[ S_t(x) = \frac{P_t(-x) - P_t(x)}{P_t(-x) + P_t(x)}. \]

Lemma:

1. For \( 1 \leq x \leq 2^t \), \( S_t(x) \in \left[ 1, \frac{5}{3} \right] \).
2. For \( -2^t \leq x \leq -1 \), \( S_t(x) \in \left[ -\frac{5}{3}, -1 \right] \).
3. For \( -2^t \leq x \leq -1 \), \( S_t(x) \in \left[ -\frac{5}{3}, -1 \right] \).

Proof:

Suffices to prove (1). Can check that \( -S_t(-x) = S_t(x) \).

So (1) \( \Rightarrow \) (1). Let \( 1 \leq x \leq 2^t \).

Case 1. \( P_t(x) = 0 \). \( S_t(y) = 1 \).
Case 2. $P_t(x) \neq 0$. Divide num + denom of $S_t(x)$ by $P_t(x)$.

$$S_t(x) = \frac{(P_t(-x)/-P_t(x)) + 1}{(P_t(-x)/-P_t(x)) - 1} = \frac{2}{1 + \frac{-P_t(-x)}{P_t(x)} - 1} \leq 1 + \frac{2}{3} = \frac{5}{3}.$$  

Previous lemma says $4 \leq \frac{-P_t(-x)}{P_t(x)}$.  

(by)
Consequences for learning:

- Intersection of two weight $W$ halfspaces has PTF degree $O(\log W)$. Can PAC learn in time $nO(\log W)$.

- Intersection of $k$ weight $W$ halfspaces?

  PTF degree $O(k \log k \log W)$.

  $f(x) = \text{sgn}(\sum_{w \in Z} w_i x_i - \theta)$. $w_i \in \mathbb{Z}$, $\sum_{i=1}^{W} |w_i| \leq W$. 
Let $g(y_1, \ldots, y_k)$ be any boolean function.
$h_1, \ldots, h_k$ be weight $W$ halfspaces.
Then $g(h_1, \ldots, h_k)$ has $\mathbf{P}$-time
degree $O(k^2 \log W)$.

For details: Klivans-O’Donnell-Severdio.
Intersection of $2$ poly($n$) weighted halfspaces has PTF degree $O(\log n)$. Can we do better?

Lower bounds:

- Minsky-Papert ’68: $w(1)$.
- O’Donnell–Servedio ’08: $w\left(\frac{\log n}{\log\log n}\right)$.
- Sherson ’08: $\Omega(\log n)$. 
Intersection of 2 a.r.b. halfspaces? (no restr. on weight).

Shestov: PTF degree $\Omega(n)$.

No way to get a non-trivial running time via PTF degree method for this class.

Open Q: Learn intersection of 2 a.r.b. halfspace in time $2^{o(n)}$. 
2nd unit: Uniform distribution learning.

Recall: A is a PAC learning alg for concept class 

dist-free PAC \text{ if } \exists \text{ Uniform distributions } D_{\text{over } \{0,1\}^n} \text{,}

unif-dist PAC: \forall \varepsilon, \delta > 0. \exists A \text{ target class } \mathcal{C}.

if A is given access to \( \mathbb{E}_{x \sim D} \left( \mathbb{E}_{l \sim L(x)} \left( \mathbb{E}_{z \sim R} \left( \mathbb{1}_{A(z, x, l, z) \neq y} \right) \right) \right) \leq \varepsilon \).

\text{ s.t. } A \left( A(z, x, l, z) = y \right) \leq \delta.
Uniform over \(\{0,1\}^n\), each \(x \in \{0,1\}^n\) verifies \(\frac{1}{2^n}\) w.r.t. \(U\).

Let me write \(\text{error}(h, c) = \Pr_{x \sim U}[h(x) \neq c(x)]\)

\[
= \frac{1}{2^n} \# \{x \in \{0,1\}^n : h(x) \neq c(x)\}
\]

\[\text{error}(h, c) = \frac{1}{2^n}\]
Main approach in dist-free setting.
- Prove every $c \in \mathcal{C}$ has a low deg PTF.
  $p : \{0,1\}^n \rightarrow \mathbb{R}$ s.t. $\text{sgn}(p(x)) = c(x) \forall x \in \{0,1\}^n$.
- Use LP to find/approx $p$.
- The lower the deg of $p$, the faster the alg.

What if $\text{sgn}(p(x)) = c(x)$ for "most" $x \in \{0,1\}^n$?
Will not succeed for the dist. \( D \) that puts all its wt on those \( x \) s.t. \( \text{sgn}(p(x)) \neq c(x) \).

\[ \text{Pr}[h(x) \neq c(y)] \text{ huge! Not a worry for } U! \]

- Find poly \( \{0,1\}^n \to \mathbb{R} \) s.t. \( \mathbb{E}[c(p(x) - c(x))] \leq \epsilon \).

- Easy to show \( \text{error}(\text{sgn}(p), c) \) is small.

- Alg is fast if \( p \) is low deg, \( p \) is "small".

- Useful tool: Fourier analysis over \( \{0,1\}^n \).
Learning poly(n)-term DNFs.
PTF degree lower bound of $\Omega(n^{1/3})$.
Klivans-Servedio alg that runs in time $2^{O(n^{1/3} \log n)}$.
Essentially opt for PTF deg method.
High-level sketch of $n^{O(1/\log n)}$-time alg if distrib is promised to be uniform.
Lemma: Let $f$ be a $s$-term DNF. Let $g$ be $f$ with all “long” terms of width $\geq \log \frac{s}{\epsilon}$ removed. Then $\text{error}(f, g) \leq \epsilon$.

Proof: A term $T$ in $f$ of width $k$ is sat. with prob $\frac{1}{2^k}$ (using unif dist asump crucially here).

So a long term is sat. w.p. $\leq \frac{1}{\log (s/\epsilon)} \leq \frac{1}{s}$.

So throwing it out incurs $\leq \frac{1}{s}$ error. Take union bound over all terms.
In order to approx/learn f w.r.t. U, suffices to approx g w.r.t. U, and g is "nice": all terms in gate width $\leq \log(\frac{\delta}{\epsilon})$. (i.e. $g$ is a $\log(\frac{\delta}{\epsilon})$-DNF).

- View each term of width $\leq \log(\frac{\delta}{\epsilon})$ as a meta-var.
  There are $N = O(\log(\frac{\delta}{\epsilon}))$ many meta-vars.
- View $g$ as OR of $s$ of these $N$ meta-vars.

$\Rightarrow$ Run standard alg (Winnow from $\frac{1}{2} \delta$).

$(\text{learn polyln})$-term DNF in time $n^{O(\log \frac{1}{\epsilon})}$. 
Fourier analysis of Boolean Functions.

- $f : \{-1,1\}^n \rightarrow \{-1,1\}$ (instead of $\{0,1\}^n$).
- View $f : \{-1,1\}^n \rightarrow \{-1,1\}$ (discrete/robing forial)
  as a polynomial $P_f : \mathbb{R}^n \rightarrow \mathbb{R}$ (continuous).
- Use analytic tools to reason about $P_f$ instead of $f$.
- Actually, let's be more general and $f : \{1,1\}^n \rightarrow \mathbb{R}$.
View $f: \{-1,1\}^n \rightarrow \mathbb{R}$ as a $2^n$ dimensional vector in $\mathbb{R}^{2^n}$.

One entry for each $x \in \{-1,1\}^n$ stacked on top of each other (in some fixed order). If $f: \{-1,1\}^n \rightarrow \{-1,1\}$, this is just its truth table. Function $\Leftrightarrow$ vector.

Example:

$$\min(x_1, x_2) = x_1 \text{ if } x_1 \leq x_2 \quad \eta = 2$$

$$\eta = 2 \text{ o.w.} \quad 2^n = 4$$
Write $V$ to denote vector space ($2^n$-dim) of all functions $f: \{0,1\}^n \to \mathbb{R}$.

$(f + g)(x) = f(x) + g(x) = \text{add vectors point-wise}$

$\alpha \in \mathbb{R}$ $(\alpha f)(x) = \alpha \cdot f(x) = \text{scalar mult of vec.}$

- Basis? Inner Product? Norm?
One possible basis ("standard basis").

For each \( z \in \{-1,1\}^n \), \( \delta_z : \{-1,1\}^n \to \{0,1\} \) as

\[
\delta_z(x) = \begin{cases} 
1 & \text{if } x = z \\
0 & \text{otherwise}
\end{cases}
\]

Claim: that the set of 2\(^n\) functions \( \{ \delta_z \}_{z \in \{-1,1\}^n} \)
forms a basis: Let \( f \in V \) (i.e. \( f : \{-1,1\}^n \to \mathbb{R} \))

\[
f(x) = \sum_{z \in \{-1,1\}^n} f(z) \delta_z(x).
\]
Fourier Analysis: Different basis, express \( f \) as linear comb. of "parity functions". Why?

- Many interesting combinatorial prop of \( f \) can be "read off" \( f \)'s Fourier expansion.

- Particularly well-suited to reason about approx such as \( \mathbb{E}
\left[ (\rho(x) - f(x))^2 \right] \)
Def: Let \( S \subseteq \{1, 2, \ldots, n\} \). The parity corresponding to \( S \), "parity-on-\( S \)" is \( \text{PAR}_S : \{-1, 1\}^n \rightarrow \{-1, 1\} \)

\[
\text{PAR}_S(x) = \prod_{i \in S} x_i = \begin{cases} 1 & \text{even \# of } i \in S \text{ is even} \\ -1 & \text{odd \# of } i \in S \text{ is odd} \end{cases}
\]

Denote \( \text{PAR}_S \) as \( \chi_S \).

Adopt convention that \( \text{PAR}_\emptyset = \chi_\emptyset = 1 \).
**Def (inner prod)**. Let $f, g \in V$, $f, g : \{-1,1\}^n \to \mathbb{R}$, define the inner prod of $f$ and $g$ as

$$
\langle f, g \rangle = \mathbb{E}[f(x)g(x)] = \mathbb{E}[fg] = \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} f(x)g(x)
$$

Scaled version of dot prod btw $f$ and $g$ viewed as $2^n$-dim vecs.
Def: The inner prod induces a norm on $V$, let $f \in V$, the norm of $f$ (the "length of $f$ as a vec")

$$\|f\| := \sqrt{\langle f, f \rangle}.$$ 

Properties of $\{x_s : S \in \mathbb{S}\}$.

1. $\|x_s\| = 1$ for all $S$. 
2. $\langle x_s, x_s \rangle = \mathbb{E}\left[\varphi_s(x) \cdot \varphi_s(x) \right] = \mathbb{E}[1] = 1$. 

True for all $f \in \mathbb{R}^d$, $S \in \mathbb{S}^d$. 

Feb 9-12:33 PM
\[
X_S \cdot X_T = X_{SDT} \quad \text{where SDT is the symmetric diff of the sets } S, T \subseteq \{1, \ldots, n\}.
\]

\[
\prod_{i \in S} x_i \prod_{j \notin T} x_j = \prod_{i \in SDT} x_i
\]

\[
x_S^2 = 1
\]

\[
\mathbb{E}[X_\emptyset] = 1 \quad (X_\emptyset \equiv 1 \text{ by convention})
\]

\[
\mathbb{E}[X_S] = 0 \text{ if } S \neq \emptyset \quad \text{exactly half of } x \in S + R^n
\]

\[
\mathbb{E}[X_S] = 1 \text{ if } S = \emptyset \quad \text{half have } X_S(x) = 1, \text{ half have } X_S(x) = -1
\]
Orthonormality. For any basis functions $x_s, x_t$:

$$\langle x_s, x_t \rangle = \begin{cases} 1 & \text{if } S = T \\ 0 & \text{otherwise} \end{cases}$$

**Proof:**

$$\langle x_s, x_t \rangle = |\text{E}[x_s x_t]| = |\text{E}[x_{S \Delta T}]|$$

- If $S = T$, then $S \Delta T = \emptyset$ so $|\text{E}[x_{S \Delta T}]| = 1$.
- If $S \neq T$, then $S \Delta T \neq \emptyset$ so $|\text{E}[x_{S \Delta T}]| = 0$. \(\square\)
Thm. Let $f: \{-1, 1\}^n \to \mathbb{R}$. Then $f$ has a unique representation as a linear combo of

$$\mathcal{F} \subseteq \mathbb{F}_2^n$$

$$f(x) = \sum_{\mathcal{F} \subseteq \mathbb{F}_2^n} \hat{f}(\mathcal{S}) X_{\mathcal{S}}(x).$$

We call this the Fourier expansion of $f$, and $\hat{f}(\mathcal{S})$'s the Fourier coefficients of $f$. 
What is $\hat{f}(s)$?

Claim: $\hat{f}(s) = \mathbb{E}(f \cdot X_s) = \langle f, X_s \rangle$.

Proof: $\langle f, X_s \rangle = \left\langle \sum_{T \in [m]} \hat{f}(T)X_T, X_s \right\rangle$

$= \sum_{T \in [m]} \hat{f}(T)\mathbb{E}[X_T \big| X_s]$

$= \sum_{T \in [m]} f(T)\mathbb{E}[X_{s \cap T}] = f(s)$.\[\]
What about $\hat{f}(S)$ for $f: \{-1,1\}^n \rightarrow \{-1,1\}$?

$\hat{f}(S) = \mathbb{E}[f \cdot X_S]$

$= (1) \cdot \text{Pr}[f(x) = X_S(x)]$
$+ (-1) \cdot \text{Pr}[f(x) = -X_S(x)]$

$= 2 \cdot \text{Pr}[f(x) = X_S(x)] - 1.$

$\hat{f}(S)$ is a measure of the correlation btw $f$ and $X_S$.

"How much $f$ behaves like $X_S"
Examples: \( \text{AND}(x_1, \ldots, x_n) = 1 \) if all \( x_i = -1 \)
\(-1 \) if any \( x_i = 1 \).

First consider \( \{\pm 1\}^n \):

\[ \text{AND}'(x_1, \ldots, x_n) = 1 \) if all \( x_i = -1 \),
\( 0 \) if any \( x_i = 1 \).

\( \text{AND} = 2 \text{AND}' - 1. \)

\[ \text{AND}'(x_1, \ldots, x_n) = \left( \frac{1 - x_1}{2} \right) \left( \frac{1 - x_2}{2} \right) \ldots \left( \frac{1 - x_n}{2} \right) \]

\[ = \sum_{S \subseteq \{1, \ldots, n\}} (-1)^{|S|} x_S \]
Decision tree

\[ f(x) = \sum \mathbf{1}_{p(x)} f(p) \]

paths \( p \) in \( T \)

\[ \mathbf{1}_{p(x)} = \begin{cases} 1 & \text{if } x \text{ follows } p \\ 0 & \text{otherwise} \end{cases} \]

value of leaf at end of path \( p \)

\[ f(p) = 1 \]

\[ 1_{p(x)} = \left( \frac{1-x_1}{2} \right) \left( \frac{1+x_2}{2} \right) \left( \frac{1-x_3}{2} \right) \]