Junta Test

\[ f: \{0,1\}^n \to \{0,1\} \]

**def** \( f \) is a \( k \)-Junta if depends on \( \leq k \) vars

**Why interesting?**
- how many relevant features in data?
- interesting also for more general \( f \)
- can we test or estimate it?

**A first idea:**
find all relevant vars \(<--\) How do you do this?
if \( k > k \), then not \( k \)-junta

Finding any relevant var:
1. Find \( x, y \) s.t. \( f(x) \neq f(y) \)
2. \( S \leftarrow \{ i \mid x_i \neq y_i \} \) \(<--\) must contain relevant \( i \)

\[ X = x^{(0)} = x_1, \ldots, x_n \]
\[ X^{(1)} = y_1, x_2, \ldots, x_n \]
\[ x^{(a)} = y_1, y_2, x_3, \ldots, x_n \]
\[ y = x^{(n)} = y_1, \ldots, y_n \]

\[ f(x) = f(x^{(0)}) \]
\[ f(x^{(i)}) \] must be \( \geq 1 \) \( i \)
\[ f(x^{(a)}) \neq f(x^{(1)}) \]

\[ f(y) = f(x^{(n)}) \]
A Problem:

do you need $\Omega(n)$ queries?

No!
can find $i$ via binary search:
if $f(x^{(i)}) \neq f(x^{(i/2)})$
  recurse on $0, \ldots, n/2$
else
  \[ \bigcup_{i=0}^{n/2} \bigcup_{j=n/2}^{n} \] \( O(\log n) \) queries

Can you do better than $O(\log n)$?

Yes - you don't need to find $i$ ...

Idea: Solve a weaker problem with fewer queries:

- Partition vars into "groups"
- Find all groups containing relevant vars \( \in O(\log \#\text{groups}) \)
- if \( \#\text{groups} > k \) then not $k$-junta

Faster - but why does it work?

- if $k$-junta, only $k$ groups have relevant $i$
- if not $k$-junta, why \( \#\text{groups} > k \) groups have relevant $i$?
  Possibly groups have $>1$ relevant var?

  \[ \text{idea: Partition randomly} \]
Plan: Random partition into many groups:
- show why no two relevant vars in same group
- # groups poly in \( k + \frac{1}{\varepsilon} \)

Algorithm \( \text{given } k, \varepsilon \):
- randomly partition \( 1..n \) into \( s \) parts \( I_1, \ldots, I_s \)
  \( \text{poly}(k, \varepsilon) \)
- \( R \leftarrow \emptyset \) (set of relevant parts)
- repeat up to \( r = O(k/\varepsilon) \) times:
  - generate pair \( (x, y) \) randomly s.t. \( X_k = Y_j \)
    agree on "relevant parts" indices
  - if \( f(x) \neq f(y) \)
    use binary search to find relevant \( I_j \)
  - \( R \leftarrow R \cup I_j \)
  - if \( R \) has \( > k \) relevant parts, REJECT
Analysis

- If \( f \) is \( k \)-junta, always accepts \( \checkmark \) \( \text{1-sided error} \)
- If \( f \) \( \varepsilon \)-far ...

To show:

- Whp over choice of \( I_1, \ldots, I_s \), \( \forall R \) containing \( \leq k \) ports

\[
\Pr_{x,y} \left[ f(x) \neq f(y) \right] \text{ is high}
\]
\[
X_R = Y_R
\]

Some defs + notation:

**Notation:** \( X_s = \) ordered list \( (X_i : i \in S) \)

\[
X_s Y_s = Z = (Z_1, \ldots, Z_n) \text{ st. } Z_s = X_s \quad Z_s = Y_s
\]

**Def:** Given \( f \), Influence of \( s \subseteq [n] \) is

\[
\text{Inf}_f(s) = 2 \Pr_{x,y} \left[ f(x) \neq f(y) \right] \quad \text{st. } X_s = Y_s
\]

- Factor of 2 to match standard single var def
- Equivalent to \( 2 \Pr_{x,y} \left[ f(x) \neq f(y, X_s) \right] \)

Useful Proposition

\[
\text{Inf}_f(s) = \sum_{T : s \cap T \neq \emptyset} \hat{f}(T)^2 = \sum_{T \subseteq [n]} \hat{f}(T)^2 - \sum_{T \subseteq S} \hat{f}(T)^2
\]
Proof of useful prop:

\[ \inf_t (s) = \frac{2 \Pr_{x,y} [ f(x) \neq f(y) ]}{\sum_s x_s = y_s} \]

\[ = \sum_{x,y} \left[ \frac{1 - f(x)f(y)}{x} \right] \]

\[ = 1 - \sum_{x,y} f(x)f(y) \]

\[ = \sum_{T \in \Omega} \hat{f}(T)^2 \]

\[ = \sum_{u \in \Lambda} \hat{f}(u) \hat{f}(T) \frac{1}{T} \frac{1}{T} \]

\[ = \sum_{T \leq S} \hat{f}(T)^2 \]

if \( \Lambda \neq \emptyset \)

or \( \Omega \neq \emptyset \)

then 0 since \( x,y \) indep. in \( S \)

so \( U \leq S + T \leq S \)

+ if \( U \neq T \) will be 0 by pairing.
Useful Corollary  
**monotonicity, subadditivity**

\[
\inf_f(s) \leq \inf_f(s \cup s') = \inf_f(s) + \inf_f(s')
\]

**why?**

\[
\inf_f(s) = \sum_{T : s \cap T \neq \emptyset} \hat{f}(T)^2
\]

\[
\inf_f(s \cup s') = \sum_{T : (s \cup s') \cap T \neq \emptyset} \hat{f}(T)^2
\]

\[
\inf_f(s) + \inf_f(s') = \sum_{T : s \cap T \neq \emptyset} \hat{f}(T)^2 + \sum_{T : s' \cap T \neq \emptyset} \hat{f}(T)^2
\]

*this contains all terms in above*  

*contains all sets (some may be added twice)*
Important Lemma
(can think of as a warmup)

Lemma \( f \) \( \varepsilon \)-far from \( k \)-junta \( \implies \)

\( \forall J \text{ st. } |J| \leq k \quad \inf \left( \left( \mathbb{R} \setminus J \right) \right) \geq 2\varepsilon \)

Pf.
Fix \( J \) st. \( |J| \leq k \)

Define \( h \) st. \( h(x) = \text{sign} \left( \mathbb{E}_x \left[ f \left( x_J, Z \right) \right] \right) \)

\( \equiv \text{maj}_{z} f(x_J, Z) \)

Note: 1) \( h \) depends only on \( J \)
2) \( h \) is best \( \varepsilon \)-junta from on \( J \)
Need to redo lemma for fixing any k parts rather than just k vars

Main Lemma
\[ f \text{ } \varepsilon \text{-far from } k \text{-junta} \]
\[ \exists \text{ random partition with } s = \text{poly}(k) \text{poly}(\varepsilon) \text{ parts} \]
\[ \text{With prob } \geq 5/6 \]
\[ f \text{ is } \frac{\varepsilon}{2} \text{-far from } k \text{-part junta wrt. } \mathcal{I} \]

Def. \[ f \text{ is } k \text{-part junta wrt } \mathcal{I} \text{ if relevant}
coords in } \leq k \text{ parts of } \mathcal{I}
\[ f \text{ is } \frac{\varepsilon}{2} \text{-far from } k \text{-part junta wrt } \mathcal{I}
\]
\[ \text{if } \forall J \text{ st. } J \text{ is union of } k \text{ parts in } \mathcal{I}
\[ \inf_{f} (|\mathcal{I} \setminus J|) \geq \frac{\varepsilon}{2} \]

How do we use this?

Thm. \[ \exists \text{ alg } \mathcal{I} \text{ using } O(\varepsilon + k \log k) \text{ queries st.}
\[ \text{if } f \text{ is } k \text{-junta, } \mathcal{I} \text{ outputs PASS}
\[ \text{if } f \text{ is } \varepsilon \text{-far from } k \text{-junta, } \Pr[\mathcal{I} \text{ outputs FAIL}] \geq 2/3 \]
Proof of Thm

if \( f \) is \( k \)-junta, then \( \leq K \) relevant parts \( \Rightarrow \) \( \Gamma \) accepts

if \( f \) is \( \varepsilon \)-far from \( k \)-junta

- then main lemma implies w/ prob \( \geq 5/6 \) over choice of \( \gamma \),
  \( f \) is \( \varepsilon/2 \)-far from \( k \)-part juntas wrt \( \gamma \)

- so until \( \geq k+1 \) parts found

\[ \inf [ |n| \setminus \text{found parts}] \geq \varepsilon/2 \]

if only \( \leq k \) part found,

\[ \Pr [ \text{pick } x_2 y \text{ s.t. } f(x) \neq f(y_3 x_3)] \geq \varepsilon/2 \]

\[ E \text{ tries to find } \gamma \ldots \gamma \text{ } J \leq \frac{2}{\varepsilon} \]

\[ E \text{ tries to find } k+1 \text{ parts with relevant vars } \leq \frac{\rho(K+1)}{\varepsilon} \]

Markov's \( \Rightarrow \) \( \Pr [ \text{need } \geq 12 \frac{(K+1)}{\varepsilon} \text{ rounds to find } \geq k+1 \text{ relevant parts}] \leq \frac{1}{6} \)

\[ \therefore \Pr [ \Gamma \text{ passes}] \leq \frac{1}{6} + \frac{1}{6} < \frac{1}{3} \]
It remains to prove the lemma:

**Restricted Lemma** for most \( \mathcal{O} \) of \( \mathcal{J} \), union of \( k \)-parts

\[
\inf \frac{1}{s} \sum_{s} f(s)^2 = \sum_{s \in \mathcal{J}} f(s)^2 = \frac{3}{2} \geq \epsilon/2
\]

**Proof.** Some useful quantities:

\[
\inf \frac{1}{s} \sum_{s} f(s)^2 = \sum_{s \in 2k} f(s)^2 = \sum_{s \notin \emptyset} f(s)^2
\]

**Notation:** \( \inf (i) = \inf (\epsilon i^2) \)

Define coods with "large low order influence"

\[
H = \{ i \mid \inf \frac{1}{s} \sum_{s} f(s)^2 = \epsilon i \}
\]

Use \( \theta = \epsilon^2 \cdot \log \frac{k}{\epsilon} \cdot \frac{1}{\log k} \)

Ideas: break subsets \( S \) of \( J \) into three groups

- Bound each group separately
- Small sets with only coods that have large low order influence
- Other small sets
- Large sets
\[ \ln f_p ([n] \setminus J) = \sum_{s \in [n]} \hat{f}(s)^2 - \sum_{s \in A} \hat{f}(s)^2 - \sum_{s \in B} \hat{f}(s)^2 - \sum_{s \in C} \hat{f}(s)^2 \]

1. Show \( \geq \varepsilon \)
2. Show \( \leq \varepsilon /4 \)
3. Show \( \leq \varepsilon /4 \)

\[ \text{total} \geq \varepsilon - \varepsilon /4 - \varepsilon /4 \geq \varepsilon /2 \]

Start with 0:

\[ \sum_{s \in [n]} \hat{f}(s)^2 - \sum_{s \in A} \hat{f}(s)^2 = \sum_{s \in [n]} \hat{f}(s)^2 - \text{1st} \leq \text{2nd} \]

\[ = \ln f_p ([n] \setminus J \setminus H) \]

- How big is this?
- Goal "reduce" to "important lemma" by showing if \( J \) & therefore \( J \setminus H \) are small

**Lemma 1:** \( \sum_{i \in [n]} \inf f_{\text{peak}} (i) \leq 2k \)

**pf.** \( \sum_{i \in [n]} \inf f_{\text{peak}} (i) \geq \sum_{T \in \text{all} T} \hat{f}(T)^2 \geq 2k \hat{f}(T)^2 \)

- Each \( T \) appears \( \leq 2k \) times

\[ \leq 2k \]

\[ \]
Corollary \(|H| = |\exists i | \inf_{f} (i) \geq 3| \leq 2k / \theta \)

why? Markovs \# 

if not, \(\sum \inf_{f} (i) \geq 2k \cdot \theta \) \(\Rightarrow\)

Now let's use it:

1. let \( H = \exists i | \inf_{f} (i) \geq 3 \)

2. Then Corr 1 \(\Rightarrow |H| \leq 2k / \theta \)

3. pick \( S \geq \frac{72k^2}{\theta^2} \)

\[ Pr[each \ partition \ of \ I \ has \ at \ most \ one \ member \ of \ H] \geq \left[ 1 - Pr[\exists \ partition \ that \ gets \ \geq 2 \ members \ of \ H] \right] \]

\[ \geq 1 - \left( \frac{|H|}{2} \right) \left( \frac{1}{5} \right)^2 \]

\[ \geq 1 - \left( \frac{2k}{\theta} \right)^2 \cdot \frac{1}{5} \]

\[ \geq \frac{17}{18} \]

\[ |J \cap H| \leq k \]

\[ \therefore \ |J \cap H| \leq k \]

4. since \( f \) is \( \epsilon \)-far from \( k \)-junta

Important lemma \( \Rightarrow \inf (|J| \setminus J \cap H) \geq \epsilon \)

so \( \epsilon \geq \epsilon \)
lets do \( \textcircled{3} \) next? (need to show that it is small compared to \( \textcircled{1} \))

\[
\sum_{S \in \mathcal{B}} \hat{f}(S)^2 : \text{big sets unlikely to be in } \leq k \text{ sets of partition}
\]

**Def.** set \( S \) is \( k \)-covered by \( \mathcal{I} \) if

\[
S \subseteq_k \mathcal{I}
\]

\[
\exists i_1 \ldots i_k \text{ s.t. } S \subseteq I_{i_1} \cup \ldots \cup I_{i_k}
\]

**Thm.** \# partitions \( S > \frac{c \cdot k}{\varepsilon} \)

with prob \( \geq \frac{17}{18} \) over choice of \( \mathcal{I} \) into \( s \) partitions

\[
\sum_{S \subseteq_k \mathcal{I} \\text{ s.t. } |S| > 2k} \hat{f}(S)^2 \leq \frac{\varepsilon}{4}
\]

**Why?**

\[
|S| > 2k
\]

\[
\Pr \left[ \text{all els of } S \text{ sent to } \leq k \text{ parts in } \mathcal{I} \right] \leq \left( \frac{k}{s} \right)^{2k+1} \leq \left( \frac{c}{k} \right)^{2k+1} = e^{k \left( \frac{c}{k} \right)^{k+1}} \leq \frac{\varepsilon}{72} \quad \text{for } \varepsilon = 72e
\]

\[
S \subseteq_k \mathcal{I}
\]

\[
|S| > 2k
\]

\[
|S| > 2k
\]

\[
\sum_{S \subseteq_k \mathcal{I} \\text{ s.t. } |S| > 2k} \hat{f}(S)^2 \cdot \Pr[|S| > 2k] \leq \frac{\varepsilon}{72}
\]

**Thm follows from Markov's inequality**
Finally, let's do (c):

$$\sum_{s \in \mathcal{L}} \hat{f}(s)^2 : \text{ will show small low-order influence coords have little combined influence.}$$

Lemma \[ s = \frac{c \cdot k^2}{\varepsilon} \]

$$\Theta \leq \frac{c^e \varepsilon^2}{k^3 \log (18s)}$$

with prob $\geq \frac{17}{18}$

$$\ln_{f}^{\varepsilon_{2k}} (I_i \setminus \mathcal{H}) \leq \frac{\varepsilon}{414}, \quad \forall i \in [s]$$

Why is this useful?

$$\sum_{s \in \mathcal{L}} \hat{f}(s)^2 \leq \ln_{f}^{\varepsilon_{2k}} (J \setminus \mathcal{H}) \leq \sum_{I \subseteq J} \hat{f}(I)^2 \leq K \cdot \frac{\varepsilon}{4k} = \frac{\varepsilon}{4}$$

Proof of Lemma

Given $i \in [s]$ and $j \in [n]$, define $x_j \in \{0, \varepsilon_{2k} \}$ if $j \in I_i \setminus \mathcal{H}$ only count if in partition $I_i$ and not heavy, i.e., sum light guys in $I_i$. 

Proof continues with further details...
\[ \ln f^{e^{2k}}(I_x \setminus H) \leq \sum_{j \in I_x \setminus H} \ln f^{e^{2k}}(j) \quad \text{subadditivity} \]

\[ = \sum_{j \in \{1, \ldots, n\}} X_j \quad \text{def of } X_j \]

Show why \( \sum_{j} X_j \) is small:

\[ E \left[ \sum_{j \in \{1, \ldots, n\}} X_j \right] \leq \sum_{j \in \{1, \ldots, n\} \setminus H} \ln f^{e^{2k}}(j) \cdot E(1)_{j \in I_x} \]

\[ = \frac{1}{3} \sum \ln f^{e^{2k}}(j) \]

\[ \leq \frac{2k}{3} \quad \text{Lemma 1} \]

\[ \leq \frac{\varepsilon}{8k} \quad \text{choice of } S = \frac{16 \cdot k^2}{\varepsilon} \]

Hoeffding's \( \Phi \):

\[ \Pr \left[ s - E[s] \geq n \varepsilon \right] \leq e^{-\frac{2n^2 \varepsilon^2}{(2\varepsilon)^2}} \quad \text{for } X_i \in [a_i, b_i] \]

\[ \Pr \left[ \sum_{i \in I} X_i - E[\sum_{i \in I} X_i] \geq n \varepsilon \right] \leq e^{-2 \varepsilon^2 / E(\ln f^{e^{2k}}(i))^2} \]

What is \( \sum_{i \in \{1, \ldots, n\} \setminus H} \ln f^{e^{2k}}(i) \) \?

\[ \leq \max_{i \in \{1, \ldots, n\} \setminus H} \ln f^{e^{2k}}(i) \quad \sum_{i \in \{1, \ldots, n\} \setminus H} \ln f^{e^{2k}}(i) \]

\[ \leq \Theta \quad \text{since } H \setminus \text{removed} \]

\[ \leq 2k \]

So \( \Pr \left[ \ln f^{e^{2k}}(I_x \setminus H) > \varepsilon / 4k \right] \leq \Pr \left[ \exists X \geq \frac{\varepsilon}{8k} + \frac{\varepsilon}{9k} \right] \leq e^{-\frac{\varepsilon^2}{8k} / 6 \cdot 2k} \quad \log k \)

Union bnd \( \Rightarrow \) Thm
Combining $\Theta, \Theta, \Theta$

with prob \( \geq 1 - \frac{3}{18} = \frac{5}{6} \)

\( \forall J \) union \( \leq k \) parts in \( J \)

\[ \inf_{\Theta} \left[ \left( n \setminus J \right) \right] \geq \frac{3}{4} - \frac{3}{4} - \frac{3}{4} = \frac{3}{2} \]