Lower bounds for disjointness and $L_\infty$ estimation

Piotr Indyk

(most slides by Ravi Kumar and T. Jayram)
Recap

• The **Hellinger distance** $h$ between two distributions $P$ and $Q$ is such that:

$$h^2(P, Q) = 1/2 \sum_\omega (P(\omega)^{1/2} - Q(\omega)^{1/2})^2$$

$$= 1 - \sum_\omega (P(\omega)Q(\omega))^{1/2}$$

  – $h$ is the Euclidean distance between $P^{1/2}$ and $Q^{1/2}$
  – Its value lies in the interval $[0, 1]$
In the last episode

\[ I(U, V : P(U, V) | M) \]
\[ = \frac{1}{2} \left( I(Z : P(0, Z)) + I(Z : P(Z, 0)) \right) \]
\[ Z \in \mathbb{R} \{0, 1\} \]
\[ \geq \frac{1}{2} \left( h^2(P_{00}, P_{01}) + h^2(P_{00}, P_{10}) \right) \]
\[ \geq \frac{1}{4} \left( h(P_{00}, P_{01}) + h(P_{00}, P_{10}) \right)^2 \quad \text{[Cauchy-Schwarz]} \]
\[ \geq \frac{1}{4} h^2(P_{01}, P_{10}) \quad \text{[Triangle ineq.]} \]
A point to ponder

- \( \text{I}(U, V : P(U, V) | M) \geq \frac{1}{4} h^2(P_{01}, P_{10}) \)
- If \( P \) computes AND correctly, why should \( P_{01} \) be far from \( P_{10} \)

AND is 0 on both these inputs

The large distance is between \( P_{11} \) and \( P_{00}, P_{01}, P_{10} \)
Rectangular property of d-cc

A deterministic communication protocol partitions the input matrix into monochromatic rectangles

Alice and Bob send one bit in each round
**Theorem:** If $P$ is a deterministic communication protocol, then the set of inputs with same transcript is a combinatorial rectangle

\[
P_{ab} = \tau = P_{cd} \quad \Rightarrow \quad P_{ad} = \tau = P_{cb}
\]
**Fundamental theorem of r-cc**

**Theorem:** If $P$ is a randomized communication protocol and $T$ is the set of all transcripts, then there exist $p: T \times X \rightarrow [0, 1]$, $q: T \times Y \rightarrow [0, 1]$ such that for all $x, y, \tau$ we have

$$\Pr[P_{xy} = \tau] = p(\tau, x) \cdot q(\tau, y),$$

**Proof:**
- Consider extended input = input + private random coins
- Apply the rectangular property
Cut-and-paste lemma

Lemma: $h^2(P_{ab}, P_{cd}) = h^2(P_{ad}, P_{cb})$

Proof:

1 - $h^2(P_{ab}, P_{cd})$

$= \sum_\tau (\text{Pr}[P_{ab} = \tau] \text{Pr}[P_{cd} = \tau])^{1/2}$

$= \sum_\tau (p(\tau, a) q(\tau, b) p(\tau, c) q(\tau, d))^{1/2}$

$= \sum_\tau (\text{Pr}[P_{ad} = \tau] \text{Pr}[P_{cb} = \tau])^{1/2}$

$= 1 - h^2(P_{ad}, P_{cb})$
LB for IC\(_v\)(AND | M), contd.

- So, we got
  \[ I(U, V : P(U, V) | M) \geq \frac{1}{4} h^2(P_{01}, P_{10}) = \frac{1}{4} h^2(P_{00}, P_{11}) \]

- How to relate \( h(P_{00}, P_{11}) \) to the error of \( P \)?
- Via the **total variation distance**
  \[ V(P, Q) = \frac{1}{2} \sum_\omega |P(\omega) - Q(\omega)| \]

- We know that \( V(P_{00}, P_{11}) \geq 1 - 2\delta \)
  
  (since 00 is a NO instance and 11 is a YES instance)
Variational distance

• **IT Fact 2**: We have
  \[ V(P, Q) \leq h(P, Q) \left(2 - h^2(P, Q)\right)^{1/2} \]

• **Intuition:**
  – Consider the extreme case where \( V(P, Q) = 1 \)
  – Then if \( P(\omega) > 0 \) then \( Q(\omega) = 0 \), and vice versa
  – Therefore, \( h(P, Q) \) has the maximal possible value = 1

• **Corollary**: \( h^2(P_{00}, P_{11}) \geq 1 - 2\sqrt{\delta} \)
Putting all together

• Lower bound for $I_{\nu}(\text{AND} | M)$
  \[ I(U, V : P(U, V) | M) \geq \frac{1}{4} h^2(P_{01}, P_{10}) \]
  \[ = \frac{1}{4} h^2(P_{00}, P_{11}) \]
  \[ = \frac{1}{4} (1 - 2\sqrt{\delta}) \]

• Combining with direct sum theorem
  \[ R(\text{DISJ}) \geq I_{\mu}(\text{DISJ} | D) \]
  \[ \geq n \cdot I_{\nu}(\text{AND} | M) \]
  \[ \geq (n/4) (1 - 2 \sqrt{\delta}) \]

QED
Better lower bound for $L_\infty$ estimation

• $L_\infty$ gap problem:
  – Alice has $x \in \{0\ldots m\}^n$
  – Bob has $y \in \{0\ldots m\}^n$
  – YES: if $||x-y||_\infty \leq 1$
  – NO: if $||x-y||_\infty \geq m$

• The “small problem”:
  – $\text{DIST}(u,v)$
  – YES: if $|u-v| \leq 1$
  – NO: if $|u-v| \geq m$

• We have $\text{GapL}_\infty(x,y) = \vee_i \text{DIST}(x_i,y_i)$
IC steps

1. Define the distribution
2. Apply the direct-sum theorem
3. Show that information complexity of DIST is $\Omega(1/m^2)$

**Theorem.** [Saks, Sun’02] [Bar-Yossef, Jayram, Kumar, Sivakumar’02]

The c.c. of $L_\infty$ gap problem is $\Omega(n/m^2)$
1. Distribution

- Distribution $\nu$ for DIST:
  - $(U,V,D,S) \sim \nu$
  - $U \perp V \mid D,S$
  - $(U,V)$ is a NO instance, i.e., $|U - V| \leq 1$

- Definition of the distribution:
  - $D \in R \{\text{-,|}\}$
  - $S \in R \{0,\ldots,m-1\}$
  - If $D = \text{-}$, then $U = S$, $V \in R \{S,S+1\}$
  - If $D = \text{|}$, then
    - $U \in R \{S,S+1\}$, $V = S+1$

- Distribution $\mu$ for instances of $\text{GapL}_{\infty}$ is defined to be $n$ independent copies of $\nu$
2. Direct-Sum theorem

\[ \text{IC}_\mu(\text{GapL}_\infty) \geq n \text{IC}_\nu(\text{DIST}) \]

\( \mu \) produces NO instances for \( \text{GapL}_\infty \)

Conditional independence
3. Information Cost of $P$

$$I(U, V : P | D, S)$$

$$= (1/2m) * \sum_s I(U, V : P | D = - , S = s) + I(U, V : P | D = | , S = s)$$

$$= (1/2m) * \sum_s I(V : P(s, V) | D = - , S = s) + I(U : P(U, s+1) | D = | , S = s)$$
Information Cost of $P$

$$I(U,V : P(U,V) | D,S)$$

$$= (1/2m) \sum_s I(V : P(s,V) | D = -, S = s) + I(U : P(U,s+1) | D = |, S = s)$$

$$\geq (1/2m) \sum_s h^2(P_{s,s}, P_{s,s+1}) + h^2(P_{s,s+1}, P_{s+1,s+1})$$

$$\geq (1/2m^2) h^2(P_{00}, P_{mm})$$
**Z-lemma**

**Lemma:** \( h^2(P_{xy}, P_{uv}) \geq \frac{1}{2} \left[ h^2(P_{xy}, P_{xv}) + h^2(P_{uy}, P_{uv}) \right] \)

**Proof:**

\[
\frac{1}{2}[ (1- h^2(P_{xy}, P_{xv})) + (1-h^2(P_{uy}, P_{uv})] \\
= \frac{1}{2} \sum_{\tau} (p(\tau, x) q(\tau, y) p(\tau, x) q(\tau, v))^{1/2} + \\
(p(\tau, u) q(\tau, y) p(\tau, u) q(\tau, v))^{1/2} \\
= \sum_{\tau} [p(\tau, x) + p(\tau, u)]/2 * (q(\tau, y) q(\tau, v))^{1/2} \\
\geq \sum_{\tau} [p(\tau, x) p(\tau, u) q(\tau, y) q(\tau, v))^{1/2} \\
= 1-h^2(P_{xy}, P_{uv})
\]
Finishing it up

• By Z-Lemma:

\[ h^2(P(0,0), P(m,m)) \geq \frac{1}{2} (h^2(P(0,0), P(0,m)) + h^2(P(m,0), P(m,m)) \]

• The latter two quantities are both \( \Omega(1) \)

• Therefore

\[ I(U,V : P(U,V) | D,S) \geq (1/2m^2)*h^2(P_{00}, P_{mm}) \]
\[ = \Omega(1/m^2) \]

QED