The Refinement of Two Fundamental Tools in Information Theory

Raymond W. Yeung
Institute of Network Coding
The Chinese University of Hong Kong

Joint work with Siu Wai Ho and Sergio Verdu
Discontinuity of Shannon’s Information Measures

- Shannon’s information measures: $H(X)$, $H(X|Y)$, $I(X;Y)$ and $I(X;Y|Z)$.  

- They are described as continuous functions [Shannon 1948] [Csiszár & Körner 1981] [Cover & Thomas 1991] [McEliece 2002] [Yeung 2002].

- All Shannon's information measures are indeed discontinuous everywhere when random variables take values from countably infinite alphabets [Ho & Yeung 2005].

- e.g., $X$ can be any positive integer.
Discontinuity of Entropy

- Let $P_X = \{1, 0, 0, ...\}$ and
  
  $$P_{X_n} = \left\{1 - \frac{1}{\sqrt{\log n}}, \frac{1}{n\sqrt{\log n}}, \frac{1}{n\sqrt{\log n}}, ..., 0, 0, ...\right\}.$$

- As $n \to \infty$, we have
  
  $$\sum_i |P_X(i) - P_{X_n}(i)| = \frac{2}{\sqrt{\log n}} \to 0.$$

- However,
  
  $$\lim_{n \to \infty} H(X_n) = \infty.$$
**Theorem 1:** For any $c \geq 0$ and any $X$ taking values from a countably infinite alphabet with $H(X) < \infty$,

$$\exists P_{X_n} \text{ s.t. } V(P_X, P_{X_n}) = \sum_i |P_X(i) - P_{X_n}(i)| \to 0$$

but

$$H(X_n) \to H(X) + c$$
Theorem 2: For any $c \geq 0$ and any $X$ taking values from countably infinite alphabet with $H(X) < \infty$,

$$\exists P_{X^n} \text{ s.t. } D(P_X \| P_{X^n}) = \sum_i P_X(i) \log \frac{P_X(i)}{P_{X^n}(i)} \to 0$$

but

$$H(X_n) \to H(X) + c$$
Pinsker’s inequality

\[ D(p \parallel q) \geq \frac{1}{2 \ln 2} V^2(p, q) \]

- By Pinsker’s inequality, convergence w.r.t. \( D(\cdot \parallel \cdot) \) implies convergence w.r.t. \( V(\cdot; \cdot) \).
- Therefore, Theorem 2 implies Theorem 1.
Discontinuity of Entropy
Theorem 3: For any $X$, $Y$ and $Z$ taking values from countably infinite alphabet with $I(X; Y|Z) < \infty$,

$$\exists P_{XnYnZn} \text{ s.t. } \lim_{n \to \infty} D(P_{XYZ} \parallel P_{XnYnZn}) = 0$$

but

$$\lim_{n \to \infty} I(X_n; Y_n | Z_n) = \infty.$$
Discontinuity of Shannon’s Information Measures

Applications:
- channel coding theorem
- lossless/lossy source coding theorems, etc.

Typicality

Fano’s Inequality

Shannon’s Information Measures
To Find the Capacity of a Communication Channel

\[ \text{Capacity} \geq C_1 \quad \text{Typicality} \]

\[ \text{Capacity} \leq C_2 \quad \text{Fano’s Inequality} \]
On Countably Infinite Alphabet

Applications:
channel coding theorem
lossless/lossy source coding theorems, etc.

Typicality

Fano’s Inequality

Shannon’s Information Measures

discontinuous!
Typicality

- Weak typicality was first introduced by Shannon [1948] to establish the source coding theorem.
- Strong typicality was first used by Wolfowitz [1964] and then by Berger [1978]. It was further developed into the method of types by Csiszár and Körner [1981].
- Strong typicality possesses stronger properties compared with weak typicality.
- It can be used only for random variables with finite alphabet.
Consider an i.i.d. source \( \{X_k, k \geq 1\} \), where \( X_k \) taking values from a countable alphabet \( \mathcal{X} \).

Let \( P_X = P_{X_k} \) for all \( k \).

Assume \( H(P_X) < \infty \).

Let \( X = (X_1, X_2, \ldots, X_n) \)

For a sequence \( x = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n \),

\( N(x; x) \) is the number of occurrences of \( x \) in \( x \)

\( q(x; x) = n^{-1}N(x; x) \) and

\( Q_X = \{q(x; x)\} \) is the empirical distribution of \( x \)

e.g., \( x = (1, 3, 2, 1, 1) \).

\[
N(1; x) = 3, \quad N(2; x) = N(3; x) = 1
\]

\( Q_X = \{3/5, 1/5, 1/5\} \).
Weak Typicality

Definition (Weak typicality): For any $\varepsilon > 0$, the weakly typical set $W^n_{[X]_\varepsilon}$ with respect to $P_X$ is the set of sequences $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ such that

$$\left| -\frac{1}{n} \log P_X(x) - H(P_X) \right| \leq \varepsilon$$
Definition 1 (Weak typicality): For any $\varepsilon > 0$, the weakly typical set $W^n_{[x]\varepsilon}$ with respect to $P_X$ is the set of sequences $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ such that

$$| D(Q_X \| P_X) + H(Q_X) - H(P_X) | \leq \varepsilon$$

Note that

$$H(Q_X) = -\sum_x Q_X(x) \log Q_X(x)$$

while

$$\text{Empirical entropy} = -\sum_x Q_X(x) \log P_X(x)$$
Asymptotic Equipartition Property

- **Theorem 4 (Weak AEP):** For any \( \varepsilon > 0 \):
  1. If \( x \in W_{[X]_\varepsilon}^n \), then
     \[
     2^{-n(H(X)+\varepsilon)} \leq p(x) \leq 2^{-n(H(X)-\varepsilon)}
     \]
  2. For sufficiently large \( n \),
     \[
     \Pr\left\{ X \in W_{[X]_\varepsilon}^n \right\} > 1 - \varepsilon
     \]
  3. For sufficiently large \( n \),
     \[
     (1 - \varepsilon)2^{-n(H(X)-\varepsilon)} \leq \left| W_{[X]_\varepsilon}^n \right| \leq 2^n(H(X)+\varepsilon)
     \]
Strong Typicality

- Strong typicality has been defined in slightly different forms in the literature.

- **Definition 2 (Strong typicality):** For $|\mathcal{X}| < \infty$ and any $\delta > 0$, the strongly typical set $T_n^{[\mathcal{X}]\delta}$ with respect to $P_X$ is the set of sequences $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ such that

$$V(P_X, Q_X) = \sum_x |P_X(x) - q(x; \mathbf{x})| \leq \delta$$

the variational distance between the empirical distribution of the sequence $\mathbf{x}$ and $P_X$ is small.
Theorem 5 (Strong AEP): For a finite alphabet $\mathcal{X}$ and any $\delta > 0$:

1) If $x \in T^n_{[\mathcal{X}]\delta}$, then

$$2^{-n(H(X)+\delta)} \leq p(x) \leq 2^{-n(H(X)-\delta)}$$

2) For sufficiently large $n$,

$$\Pr\{X \in T^n_{[\mathcal{X}]\delta}\} > 1 - \delta$$

3) For sufficiently large $n$,

$$(1 - \delta)2^{n(H(X)-\gamma)} \leq |T^n_{[\mathcal{X}]\delta}| \leq 2^{n(H(X)+\gamma)}$$
Breakdown of Strong AEP

- If strong typicality is extended (in the natural way) to countably infinite alphabets, strong AEP no longer holds.
- Specifically, Property 2 holds but Properties 1 and 3 do not hold.
Typicality

$\mathcal{X}^n$ finite alphabet

Weak Typicality:

$$|D(Q_X \parallel P_X) + H(Q_X) - H(P_X)| \leq \varepsilon$$

Strong Typicality:

$$V(P_X, Q_X) \leq \delta$$
Unified Typicality

$\mathcal{X}^n$ countably infinite alphabet

**Weak Typicality:**

$$|D(Q_X \parallel P_X) + H(Q_X) - H(P_X)| \leq \varepsilon$$

**Strong Typicality:**

$$V(P_X, Q_X) \leq \delta$$

$\exists x$ s.t. $D(Q_X \parallel P_X)$ is small but $|H(Q_X) - H(P_X)|$ is large
Unified Typicality

$\mathcal{X}^n$ countably infinite alphabet

Weak Typicality:

$$|D(Q_X \parallel P_X) + H(Q_X) - H(P_X)| \leq \varepsilon$$

Strong Typicality:

$$V(P_X, Q_X) \leq \delta$$

Unified Typicality:

$$D(Q_X \parallel P_X) + |H(Q_X) - H(P_X)| \leq \eta.$$
Unified Typicality

- **Definition 3 (Unified typicality):** For any $\eta > 0$, the unified typical set $U^n_{[X] \eta}$ with respect to $P_X$ is the set of sequences $x = (x_1, x_2, ..., x_n) \in X^n$ such that

  $$D(Q_X \parallel P_X) + |H(Q_X) - H(P_X)| \leq \eta$$

- **Weak Typicality:** $|D(Q_X \parallel P_X) + H(Q_X) - H(P_X)| \leq \varepsilon$

- **Strong Typicality:** $V(P_X, Q_X) \leq \delta$

- Each typicality corresponds to a “distance measure”

- Entropy is continuous w.r.t. the distance measure induced by unified typicality
Theorem 6 (Unified AEP): For any $\eta > 0$:

1) If $x \in U_{[X]n}$, then

$$2^{-n(H(X)+\eta)} \leq p(x) \leq 2^{-n(H(X)-\eta)}$$

2) For sufficiently large $n$,

$$\Pr\{X \in U_{[X]n}\} > 1 - \eta$$

3) For sufficiently large $n$,

$$(1 - \eta)2^{n(H(X)+\mu)} \leq |U_{[X]n}| \leq 2^{n(H(X)+\mu)}$$
Unified Typicality

- **Theorem 7**: For any $x \in \mathcal{X}^n$,
  
  if $x \in U^n_{[X] \eta}$, then $x \in W^n_{[X] \varepsilon}$ and $x \in T^n_{[X] \delta}$,

  where $\varepsilon = \eta$ and $\delta = \sqrt{\eta \cdot 2 \ln 2}$. 
Unified Jointly Typicality

- Consider a bivariate information source \( \{(X_k, Y_k), k \geq 1\} \) where \((X_k, Y_k)\) are i.i.d. with generic distribution \( P_{XY} \).

- We use \((X, Y)\) to denote the pair of generic random variables.

- Let \((X, Y) = ((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))\).

- For the pair of sequence \((x, y)\), the empirical distribution is \( Q_{XY} = \{q(x,y; x,y)\} \) where \( q(x,y; x,y) = n^{-1}N(x,y; x,y) \).
Definition 4 (Unified jointly typicality): For any $\eta > 0$, the unified typical set $U^n_{[XY]\eta}$ with respect to $P_{XY}$ is the set of sequences $(x, y) \in \mathcal{X}^n \times \mathcal{Y}^n$ such that

$$D(Q_{XY} \parallel P_{XY}) + |H(Q_{XY}) - H(P_{XY})|$$
$$+ |H(Q_X) - H(P_X)| + |H(Q_Y) - H(P_Y)| \leq \eta.$$

This definition cannot be simplified.
Conditional AEP

- **Definition 5:** For any \( x \in U^n_{[X]\eta} \), the conditional typical set of \( Y \) is defined as

\[
U^n_{[Y|X]\eta}(x) = \left\{ y \in U^n_{[Y]\eta} : (x, y) \in U^n_{[XY]\eta} \right\}
\]

- **Theorem 8:** For \( x \in U^n_{[X]\eta} \), if

\[
\left| U^n_{[Y|X]\eta} \right| \geq 1,
\]

then

\[
2^n(H(Y|X) - \nu) \leq \left| U^n_{[Y|X]\eta} \right| \leq 2^n(H(Y|X) + \nu)
\]

where \( \nu \to 0 \) as \( \eta \to 0 \) and then \( n \to \infty \).
Applications

- Rate-distortion theory
  - A version of rate-distortion theorem was proved by strong typicality [Cover & Thomas 1991][Yeung 2008]
  - It can be easily generalized to countably infinite alphabet

- Multi-source network coding
  - The achievable information rate region in multisource network coding problem was proved by strong typicality [Yeung 2008]
  - It can be easily generalized to countably infinite alphabet
Fano's Inequality

Fano's inequality: For discrete random variables $X$ and $Y$ taking values on the same alphabet $\mathcal{X} = \{1, 2, \ldots\}$, let

$$\varepsilon = P[X \neq Y] = 1 - \sum_{w \in \mathcal{X}} P_{XY}(w, w)$$

Then

$$H(X \mid Y) \leq \varepsilon \log(\| \mathcal{X} \| - 1) + h(\varepsilon),$$

where

$$h(x) = x \log \frac{1}{x} + (1 - x) \log \frac{1}{1 - x}$$

for $0 < x < 1$ and $h(0) = h(1) = 0$. 
Motivation 1

\[ H(X \mid Y) \leq \varepsilon \log(|X| - 1) + h(\varepsilon) \]

- This upper bound on \( H(X \mid Y) \) is not tight.
- For fixed \( \varepsilon \) and \(|X|\), can always find \( X \) such that
  \[ H(X \mid Y) \leq H(X) < \varepsilon \log(|X| - 1) + h(\varepsilon) \]
- Then we can ask, for fixed \( P_X \) and \( \varepsilon \), what is
  \[ \max_{P_{Y \mid X} : P[X \neq Y] = \varepsilon} H(X \mid Y) < \varepsilon \log(|X| - 1) + h(\varepsilon) \]
Motivation 2

- If $\mathcal{X}$ is countably infinite, Fano’s inequality no longer gives an upper bound on $H(X|Y)$.
- It is possible that $H(X|Y) \not\rightarrow 0$ as $\epsilon \rightarrow 0$ which can be explained by the discontinuity of entropy.
- Then $H(X_n|Y_n) = H(X_n) \rightarrow \infty$ but $\epsilon_n = \frac{1}{\sqrt{\log n}} \rightarrow 0$
- Under what conditions $\epsilon \rightarrow 0 \Rightarrow H(X|Y) \rightarrow 0$ for countably infinite alphabets?
Tight Upper Bound on $H(X|Y)$

- Theorem 9: Suppose $\varepsilon = P[X \neq Y] \leq 1 - P_X(1)$, then

$$H(X | Y) \leq \varepsilon \, H(Q(P_X, \varepsilon)) + h(\varepsilon)$$

where the right side is the tight bound dependent on $\varepsilon$ and $P_X$. (This is the simplest of the 3 cases.)

Let $\Phi_X(\varepsilon) = \varepsilon \, H(Q(P_X, \varepsilon)) + h(\varepsilon)$
Generalizing Fano’s Inequality

- Fano's inequality [Fano 1952] gives an upper bound on the conditional entropy $H(X|Y)$ in terms of the error probability $\varepsilon = \Pr\{X \neq Y\}$.
- e.g. $P_X = [0.4, 0.4, 0.1, 0.1]$
Generalizing Fano’s Inequality

- e.g., $X$ is a Poisson random variable with mean equal to 10.

- Fano's inequality no longer gives an upper bound on $H(X|Y)$. 

$H(X|Y)$ vs. $\epsilon$
Generalizing Fano’s Inequality

- e.g. $X$ is a Poisson random variable with mean equal to 10.
- Fano's inequality no longer gives an upper bound on $H(X|Y)$.

$H(X|Y)$

[Ho & Verdú 2008]
Joint Source-Channel Coding

$(S_1, S_2, \ldots S_k) \rightarrow \text{Encoder} \rightarrow (X_1, X_2, \ldots X_n) \rightarrow \text{Channel} \rightarrow (Y_1, Y_2, \ldots Y_n) \rightarrow (\hat{S}_1, \hat{S}_2, \ldots \hat{S}_k)$

$k$-to-$n$ joint source-channel code
Error Probabilities

- The average symbol error probability is defined as
  \[ \lambda_k = \frac{1}{k} \sum_{i=1}^{k} P[S_i \neq \hat{S}_i] \]

- The block error probability is defined as
  \[ \mu_k = P[(S_1, S_2, \ldots, S_k) \neq (\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_k)] \]
Theorem 10: For any discrete memoryless source and general channel, the rate of a $k$-to-$n$ joint source-channel code with symbol error probability $\lambda_k$ satisfies

$$k \leq \frac{\sup_{X^n} n^{-1} I(X^n; Y^n)}{n} \leq \frac{k^{-1} H(S_k^k) - \Phi_{S^*}(\lambda_k)}{n}$$

where $S^*$ is constructed from $\{S_1, S_2, ..., S_k\}$ according to

$$P_{S^*}(1) = \min_j P_{S_j}(1),$$

$$P_{S^*}(a) = \min_j \sum_{i=1}^a P_{S_j}(i) - \sum_{i=1}^{a-1} P_{S^*}(i) \quad a \geq 2.$$
Theorem 11: For any general discrete source and general channel, the block error probability $\mu_k$ of a $k$-to-$n$ joint source-channel code is lower bounded by

$$\Phi^{-1}_{S^k} \left(H(S^k) - \sup_{X^n} I(X^n; Y^n)\right) \leq \mu_k$$
Information Theoretic Security

- Weak secrecy \( \lim_{n \to \infty} n^{-1} I(X^n; Y^n) = 0 \) has been considered in [Csiszár & Körner 78, Broadcast channel] and some seminal papers.

- [Wyner 75, Wiretap channel I] only stated that “a large value of the equivocation implies a large value of \( P_{ew} \)”, where the equivocation refers to \( n^{-1} H(X^n | Y^k) \) and \( P_{ew} \) means \( \mu_n \).

- It is important to clarify what exactly weak secrecy implies.
Weak Secrecy

- E.g., $P_X = (0.4, 0.4, 0.1, 0.1)$. 

\[ \varepsilon = P[X \neq Y] \]

- [Ho & Verdú 2008]
- [Fano 1952]
Weak Secrecy

Theorem 12: For any discrete stationary memoryless source (i.i.d. source) with distribution $P_X$, if

$$\lim_{n \to \infty} n^{-1} I(X^n; Y^n) = 0,$$

Then

$$\lim_{n \to \infty} \lambda_n = \lambda_{\text{max}} \quad \text{and} \quad \lim_{n \to \infty} \mu_n = 1.$$ 

Remark:

- Weak Secrecy together with the stationary source assumption is insufficient to show the maximum error probability.

- The proof is based on the tight upper bound on $H(X|Y)$ in terms of error probability.
Summary

Applications:
- Channel coding theorem
- Lossless/lossy source coding theorems

Typicality
- Weak Typicality
- Strong Typicality

Fano’s Inequality

Shannon’s Information Measures
On Countably Infinite Alphabet

Applications:
- channel coding theorem
- lossless/lossy source coding theorem

Typicality

Weak Typicality

Shannon’s Information Measures

discontinuous!
Unified Typicality

Applications:
- channel coding theorem
- MSNC/lossy SC theorems

Typicality

Unified Typicality

Shannon’s Information Measures
Generalized Fano’s Inequality

Applications:
- results on JSCC, IT security
- MSNC/lossy SC theorems

Typicality

Unified Typicality

Shannon’s Information Measures

Generalized Fano’s Inequality
Perhaps...

A lot of fundamental research in information theory are still waiting for us to investigate.
References

Q & A