Chapter 2

Generalized Information Measures for Arbitrary System Statistics

Po-Ning Chen
Department of Communications Engineering
National Chiao-Tung University
Hsin Chu, Taiwan 30050
Shannon's entropy

- Entropy of a discrete random variable $X$:
  \[
  H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x) = E_X [-\log P_X(X)] \text{ nats}
  \]
  is a measure of the average amount of uncertainty in $X$.

- Entropy rate for a sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ is
  \[
  \lim_{n \to \infty} \frac{1}{n} H(X^n) = \lim_{n \to \infty} \frac{1}{n} E [-\log P_{X^n}(X^n)],
  \]
  assuming the limit exists.

- Operation meaning: Shannon’s coding theorems for stationary and ergodic system statistics.

- Question: Does these measures have the same operational significance for systems with non-stationary or time-varying nature. Answer: No.

- Solution: Require new entropy measure which can appropriately characterize the operational limits of arbitrary stochastic systems.
Arbitrary system statistics

• In general, there are two indices for observations: time index and space index.

• When a sequence of observations is denoted by

\[ X_1, X_2, \ldots, X_n, \ldots, \]

the subscript \( i \) of \( X_i \) can be treated as either a time index or a space index, but not both.

• Hence, when a sequence of observations are functions of both time and space, the notation of \( X_1, X_2, \ldots, X_n, \ldots, \) is by no means sufficient; and therefore, a new model for a time-varying multiple-sensor system, such as

\[ X_1^{(n)}, X_2^{(n)}, \ldots, X_t^{(n)}, \ldots, \]

where \( t \) is the time index and \( n \) is the space or position index (or vice versa), becomes significant.
Arbitrary system statistics

- For instance, periodic observations from infinite number of sensors.

\[
\begin{array}{ccccccc}
\text{time 1} & \text{time 2} & \text{time 3} & \cdots \\
X_1^{(1)} & X_2^{(1)} & X_3^{(1)} & \cdots \\
X_1^{(2)} & X_2^{(2)} & X_3^{(2)} & \cdots \\
X_1^{(3)} & X_2^{(3)} & X_3^{(3)} & \cdots \\
X_1^{(4)} & X_2^{(4)} & X_3^{(4)} & \cdots \\
\vdots & \vdots & \vdots & \vdots 
\end{array}
\]

- When block-wise (block in the sense of “time-block”) compression of such source (with block length \(n\)) is considered, same question as to the compression of i.i.d. source arises:

  *what is the minimum compression rate (bits per source sample) for which the error can be make arbitrarily small as the block length goes to infinity?*

- To answer the question, information theorists have to find a sequence of data compression codes for each block length \(n\) and investigate if the decompression error goes to zero as \(n\) approaches infinity.

- However, unlike those simple source models considered in Volume I, the arbitrary source for each block length \(n\) may exhibit distinct statistics at respective
Arbitrary system statistics

sample, i.e.,

\[ n = 1 : X_1^{(1)} \]
\[ n = 2 : X_1^{(2)}, X_2^{(2)} \]
\[ n = 3 : X_1^{(3)}, X_2^{(3)}, X_3^{(3)} \]
\[ n = 4 : X_1^{(4)}, X_2^{(4)}, X_3^{(4)}, X_4^{(4)} \]
\[ : \]

and the statistics of \( X_1^{(4)} \) could be different from \( X_1^{(1)} \), \( X_1^{(2)} \) and \( X_1^{(3)} \).

• Since it is the most general model for the above question, and the system statistics can be arbitrarily defined, it is therefore named arbitrary statistics system.

• In notation, the triangular array of random variables is often denoted by a boldface letter as

\[ \mathbf{X} \triangleq \{ X^n \}_{n=1}^{\infty} , \]

where

\[ X^n \triangleq \left( X_1^{(n)}, X_2^{(n)}, \ldots, X_n^{(n)} \right) ; \]

for convenience, the above statement is sometimes briefed as

\[ \mathbf{X} \triangleq \left\{ X^n = \left( X_1^{(n)}, X_2^{(n)}, \ldots, X_n^{(n)} \right) \right\}_{n=1}^{\infty} . \]
**Spectrum and Quantile**  

**Definition 2.1 (inf/sup-spectrum)** If \( \{A_n\}_{n=1}^{\infty} \) is a sequence of random variables, then its *inf-spectrum* \( \underline{u}(\cdot) \) and its *sup-spectrum* \( \overline{u}(\cdot) \) are defined by

\[
\underline{u}(\theta) \triangleq \lim_{n \to \infty} \inf \Pr\{A_n \leq \theta\},
\]

and

\[
\overline{u}(\theta) \triangleq \lim_{n \to \infty} \sup \Pr\{A_n \leq \theta\}.
\]

- \( \underline{u}(\cdot) \) and \( \overline{u}(\cdot) \) are respectively the liminf and the limsup of the cumulative distribution function (CDF) of \( A_n \).

**Definition 2.2 (quantile of inf/sup-spectrum)** For any \( 0 \leq \delta \leq 1 \), the *quantiles* \( U_\delta \) and \( \overline{U}_\delta \) of the sup-spectrum and the inf-spectrum are defined by

\[
U_\delta \triangleq \sup\{\theta : \overline{u}(\theta) \leq \delta\}
\]

and

\[
\overline{U}_\delta \triangleq \sup\{\theta : \underline{u}(\theta) \leq \delta\},
\]

respectively.

It follows from the above definitions that \( U_\delta \) and \( \overline{U}_\delta \) are right-continuous and non-decreasing in \( \delta \). Note that the supremum of an empty set is defined to be \(-\infty\).

- If \( \overline{u}(\cdot) \) (or \( \underline{u}(\cdot) \)) is strictly increasing, then the quantile is nothing but its inverse:

\[ U_\delta = \overline{u}^{-1}(\delta). \]
• \textit{liminf in probability} $U$ of $\{A_n\}_{n=1}^{\infty}$ is the largest extended real number such that for all $\xi > 0$,

$$\lim_{n \to \infty} \Pr[A_n \leq U - \xi] = 0.$$ 

• \textit{limsup in probability} $\bar{U}$ is defined as the smallest extended real number such that for all $\xi > 0$,

$$\lim_{n \to \infty} \Pr[A_n \geq \bar{U} + \xi] = 0.$$ 

$$U = \lim_{\delta \downarrow 0} U_{\delta} = U_0$$

and

$$\bar{U} = \lim_{\delta \uparrow 1} \bar{U}_{\delta} = \sup\{\theta : u(\theta) < 1\},$$

• Straightforwardly by their definitions,

$$U \leq U_{\delta} \leq \bar{U}_{\delta} \leq \bar{U} \quad \text{for} \ \delta \in [0, 1).$$

• $\bar{U}_1 = U_1 = \infty.$
The asymptotic CDFs of a sequence of random variables \( \{A_n\}_{n=1}^{\infty} \).

\( \bar{u}(\cdot) = \text{sup-spectrum of } A_n; \)

\( u(\cdot) = \text{inf-spectrum of } A_n; \)

\( U_{1-} = \lim_{\xi \uparrow 1} U_\xi. \)
Properties of quantile

Lemma 2.3 Assume:

- Two random sequences: \( \{A_n\}_{n=1}^{\infty} \) and \( \{B_n\}_{n=1}^{\infty} \);
- \( \bar{u}(\cdot) = \text{sup-spectrum of } \{A_n\}_{n=1}^{\infty}; \bar{U}_\delta = \text{quantile of } \bar{u}(\cdot) \);
- \( u(\cdot) = \text{inf-spectrum of } \{A_n\}_{n=1}^{\infty}; \bar{U}_\delta = \text{quantile of } u(\cdot) \);
- \( \bar{v}(\cdot) = \text{sup-spectrum of } \{B_n\}_{n=1}^{\infty}; \bar{V}_\delta = \text{quantile of } \bar{u}(\cdot) \);
- \( v(\cdot) = \text{inf-spectrum of } \{B_n\}_{n=1}^{\infty}; \bar{V}_\delta = \text{quantile of } u(\cdot) \);
- \( (\bar{u} + \bar{v})(\cdot) = \text{sup-spectrum of } \{A_n + B_n\}_{n=1}^{\infty}, \text{i.e.,} \)
  \[
  (\bar{u} + \bar{v})(\theta) \triangleq \limsup_{n \to \infty} \Pr\{A_n + B_n \leq \theta\};
  \]
  \( (\bar{U} + \bar{V})_\delta = \text{quantile of } (\bar{u} + \bar{v})(\cdot) \);
- \( (\bar{u} + \bar{v})(\cdot) = \text{inf-spectrum of } \{A_n + B_n\}_{n=1}^{\infty}, \text{i.e.,} \)
  \[
  (\bar{u} + \bar{v})(\theta) \triangleq \liminf_{n \to \infty} \Pr\{A_n + B_n \leq \theta\};
  \]
  \( (\bar{U} + \bar{V})_\delta = \text{quantile of } (\bar{u} + \bar{v})(\cdot) \).
Properties of quantile

Then the following statements hold.

1. $\bar{U}_\delta$ and $\bar{U}_\delta$ are both non-decreasing and right-continuous functions of $\delta$ for $\delta \in [0, 1]$.

2. $\lim_{\delta \downarrow 0} \bar{U}_\delta = \bar{U}_0$ and $\lim_{\delta \downarrow 0} \bar{U}_\delta = \bar{U}_0$.

3. For $\delta \geq 0$, $\gamma \geq 0$, and $\delta + \gamma \leq 1$,

$$\bar{(U + V)}_{\delta + \gamma} \geq \bar{U}_\delta + \bar{V}_\gamma,$$  \hspace{1cm} (2.2.1)

and

$$\bar{(U + V)}_{\delta + \gamma} \geq \bar{U}_\delta + \bar{V}_\gamma.$$  \hspace{1cm} (2.2.2)

4. For $\delta \geq 0$, $\gamma \geq 0$, and $\delta + \gamma \leq 1$,

$$\bar{(U + V)}_{\delta} \leq \bar{U}_\delta + \bar{V}_{(1-\gamma)},$$  \hspace{1cm} (2.2.3)

and

$$\bar{(U + V)}_{\delta} \leq \bar{U}_\delta + \bar{V}_{(1-\gamma)}.$$  \hspace{1cm} (2.2.4)
Generalized information measures

In Definitions 2.1 and 2.2,

- $A_n = \text{normalized entropy density}$, i.e.,

$$\frac{1}{n} h_{X^n}(X^n) \triangleq - \frac{1}{n} \log P_{X^n}(X^n),$$

\[ \delta\text{-inf-entropy rate } H_\delta(X) = \text{quantile of sup-spectrum of } \frac{1}{n} h_{X^n}(X^n) \]

\[ \delta\text{-sup-entropy rate } \tilde{H}_\delta(X) = \text{quantile of inf-spectrum of } \frac{1}{n} h_{X^n}(X^n). \]

- $A_n = \text{normalized information density}$, i.e.,

$$\frac{1}{n} i_{X^nW^n}(X^n; Y^n) = \frac{1}{n} i_{X^n,Y^n}(X^n; Y^n) \triangleq - \frac{1}{n} \log \frac{P_{X^n,Y^n}(X^n, Y^n)}{P_{X^n}(X^n)P_{Y^n}(Y^n)},$$

\[ \delta\text{-inf-information-rate } I_\delta(X; Y) = \text{quantile of sup-spectrum of } \frac{1}{n} i_{X^nW^n}(X^n; Y^n) \]

\[ \delta\text{-sup-information-rate } \bar{I}_\delta(X; Y) = \text{quantile of inf-spectrum of } \frac{1}{n} i_{X^nW^n}(X^n; Y^n). \]
Generalized information measures

- $A_n = \text{normalized log-likelihood ratio}$, i.e.,

$$\frac{1}{n} d_{X^n}(X^n \parallel \hat{X}^n) \triangleq \frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\hat{X}^n}(X^n)}$$

$\delta$-inf-divergence rate $D_{\delta}(X \parallel \hat{X}) = \text{quantile of sup-spectrum of } \frac{1}{n} d_{X^n}(X^n \parallel \hat{X}^n)$

$\delta$-sup-divergence rate $\bar{D}_{\delta}(X \parallel \hat{X}) = \text{quantile of inf-spectrum of } \frac{1}{n} d_{X^n}(X^n \parallel \hat{X}^n)$.

- Notes:
  - The inf-entropy-rate $H(X)$ and the sup-entropy-rate $\bar{H}(X)$ are special cases of the $\delta$-inf/sup-entropy rate measures:

$$\underline{H}(X) = H_0(X), \quad \text{and} \quad \bar{H}(X) = \lim_{\delta \uparrow 1} \bar{H}_\delta(X).$$

  - Concept: If the random variable $(1/n)h(X^n)$ exhibits a limiting distribution, and suppose the limiting distribution of $(1/n)h_{X^n}(X^n)$ is positive over $(-2, 2)$; and zero, otherwise. Then $\bar{H}(X) = 2$ and $\underline{H}(X) = -2$. 
### Generalized information measures

<table>
<thead>
<tr>
<th>Entropy Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>system</strong></td>
</tr>
<tr>
<td>norm. entropy density</td>
</tr>
<tr>
<td>entropy sup-spectrum</td>
</tr>
<tr>
<td>entropy inf-spectrum</td>
</tr>
<tr>
<td>$\delta$-inf-entropy rate</td>
</tr>
<tr>
<td>$\delta$-sup-entropy rate</td>
</tr>
<tr>
<td>sup-entropy rate</td>
</tr>
<tr>
<td>inf-entropy rate</td>
</tr>
</tbody>
</table>

Generalized entropy measures where $\delta \in [0, 1]$. 
### Generalized information measures

#### Mutual Information Measures

| System | Arbitrary channel $P_W = P_{Y|X}$ with input $X$ and output $Y$ |
| --- | --- |
| Norm. information density | $i_{X^nW^n}(X^n; Y^n) \triangleq \frac{1}{n} \log \frac{P_{X^nY^n}(X^n, Y^n)}{P_{X^n}(X^n) \times P_{Y^n}(Y^n)}$ |
| Information sup-spectrum | $\tilde{i}(\theta) \triangleq \limsup_{n \to \infty} \Pr\left\{ \frac{1}{n} i_{X^nW^n}(X^n; Y^n) \leq \theta \right\}$ |
| Information inf-Spectrum | $\hat{i}(\theta) \triangleq \liminf_{n \to \infty} \Pr\left\{ \frac{1}{n} i_{X^nW^n}(X^n; Y^n) \leq \theta \right\}$ |
| $\delta$-inf-information rate | $I_\delta(X; Y) \triangleq \sup \{ \theta : \tilde{i}(\theta) \leq \delta \}$ |
| $\delta$-Sup-Information Rate | $\bar{I}_\delta(X; Y) \triangleq \sup \{ \theta : \hat{i}(\theta) \leq \delta \}$ |
| Sup-information rate | $\bar{I}(X; Y) \triangleq \lim_{\delta \uparrow 1} \bar{I}_\delta(X; Y)$ |
| Inf-information rate | $\underline{I}(X; Y) \triangleq \underline{I}_0(X; Y)$ |

Generalized mutual information measures where $\delta \in [0, 1]$. 

## Generalized information measures

<table>
<thead>
<tr>
<th>Divergence Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>system</td>
</tr>
<tr>
<td>arbitrary sources $X$ and $\hat{X}$</td>
</tr>
<tr>
<td>$A_n$ : norm. log-likelihood ratio</td>
</tr>
<tr>
<td>$\frac{1}{n}d_{Xn}(X^n</td>
</tr>
<tr>
<td>divergence sup-spectrum</td>
</tr>
<tr>
<td>$\bar{d}(\theta) \triangleq \limsup_{n \to \infty} \Pr \left{ \frac{1}{n}d_{Xn}(X^n</td>
</tr>
<tr>
<td>divergence inf-spectrum</td>
</tr>
<tr>
<td>$d(\theta) \triangleq \liminf_{n \to \infty} \Pr \left{ \frac{1}{n}d_{Xn}(X^n</td>
</tr>
<tr>
<td>$\delta$-inf-divergence rate</td>
</tr>
<tr>
<td>$D_{\delta}(X</td>
</tr>
<tr>
<td>$\delta$-sup-divergence rate</td>
</tr>
<tr>
<td>$\tilde{D}_{\delta}(X</td>
</tr>
<tr>
<td>sup-divergence rate</td>
</tr>
<tr>
<td>$\bar{D}(X</td>
</tr>
<tr>
<td>inf-divergence rate</td>
</tr>
<tr>
<td>$D(X</td>
</tr>
</tbody>
</table>

Generalized divergence measures where $\delta \in [0, 1]$. 

Properties of generalized information measures

- Example of basic property for Shannon’s entropy: \( I(X;Y) = H(Y) - H(Y|X) \).

  - By taking \( \delta = 0 \) and letting \( \gamma \downarrow 0 \) in
    \[
    (U + V)_{\delta + \gamma} \geq U_\delta + V_\gamma \text{ for } \delta \geq 0, \gamma \geq 0, \text{ and } \delta + \gamma \leq 1
    \]
    and
    \[
    (U + V)_{\delta} \leq U_{\delta + \gamma} + \bar{V}_{(1-\gamma)} \text{ for } \delta \geq 0, \gamma \geq 0, \text{ and } \delta + \gamma \leq 1,
    \]
we obtain
    \[
    (U + V) \geq U_0 + \lim_{\gamma \downarrow 0} V_\gamma \geq U + V
    \]
and
    \[
    (U + V) \leq \lim_{\gamma \downarrow 0} U_\gamma + \lim_{\gamma \downarrow 0} \bar{V}_{(1-\gamma)} = U + \bar{V}.
    \]
  - Meaning: The liminf in probability of a sequence of random variables \( A_n + B_n \) is upper bounded by the liminf in probability of \( A_n \) plus the limsup in probability of \( B_n \); and is lower bounded by the sum of the liminfs in probability of \( A_n \) and \( B_n \).
  - This fact is used in the paper of Verdú and Han to show that
    \[
    I(X;Y) + H(Y|X) \leq H(Y) \leq I(X;Y) + \bar{H}(Y|X),
    \]
or equivalently,

\[ H(Y) - \bar{H}(Y|X) \leq I(X;Y) \leq H(Y) - \bar{H}(Y|X). \]
Lemma 2.4 For finite alphabet $\mathcal{X}$, the following statements hold.

1. $\bar{H}_\delta(\mathbf{X}) \geq 0$ for $\delta \in [0, 1]$. (This property also applies to $H_\delta(\mathbf{X}), I_\delta(\mathbf{X}; \mathbf{Y}), \bar{I}_\delta(\mathbf{X}; \mathbf{Y}), \bar{D}_\delta(\mathbf{X} \parallel \mathbf{X}),$ and $D_\delta(\mathbf{X} \parallel \mathbf{X}).$)

2. $\bar{I}_\delta(\mathbf{X}; \mathbf{Y}) = I_\delta(\mathbf{Y}; \mathbf{X})$ and $\bar{I}_\delta(\mathbf{X}; \mathbf{Y}) = I_\delta(\mathbf{Y}; \mathbf{X})$ for $\delta \in [0, 1]$.

3. For $0 \leq \delta < 1, 0 \leq \gamma < 1$ and $\delta + \gamma \leq 1,$

$$I_\delta(\mathbf{X}; \mathbf{Y}) \leq H_{\delta+\gamma}(\mathbf{Y}) - H_{\gamma}(\mathbf{Y} | \mathbf{X}), \quad (2.4.1)$$

$$\bar{I}_\delta(\mathbf{X}; \mathbf{Y}) \leq \bar{H}_{\delta+\gamma}(\mathbf{Y}) - \bar{H}_{\gamma}(\mathbf{Y} | \mathbf{X}), \quad (2.4.2)$$

$$\bar{I}_\gamma(\mathbf{X}; \mathbf{Y}) \leq \bar{H}_{\delta+\gamma}(\mathbf{Y}) - \bar{H}_{\delta}(\mathbf{Y} | \mathbf{X}), \quad (2.4.3)$$

$$\bar{I}_{\delta+\gamma}(\mathbf{X}; \mathbf{Y}) \geq \bar{H}_\delta(\mathbf{Y}) - \bar{H}_{(1-\gamma)}(\mathbf{Y} | \mathbf{X}), \quad (2.4.4)$$

and

$$\bar{I}_{\delta+\gamma}(\mathbf{X}; \mathbf{Y}) \geq \bar{H}_\delta(\mathbf{Y}) - \bar{H}_{(1-\gamma)}(\mathbf{Y} | \mathbf{X}). \quad (2.4.5)$$

(Note that the case of $(\delta, \gamma) = (1, 0)$ holds for (2.4.1) and (2.4.2), and the case of $(\delta, \gamma) = (0, 1)$ holds for (2.4.3), (2.4.4) and (2.4.5).)

4. $0 \leq H_\delta(\mathbf{X}) \leq \bar{H}_\delta(\mathbf{X}) \leq \log |\mathcal{X}|$ for $\delta \in [0, 1)$, where each $X_i^{(n)}$ takes values in $\mathcal{X}$ for $i = 1, \ldots, n$ and $n = 1, 2, \ldots$.

5. $I_\delta(\mathbf{X}, \mathbf{Y}; \mathbf{Z}) \geq I_\delta(\mathbf{X}; \mathbf{Z})$ for $\delta \in [0, 1]$. 

Properties of generalized information measures

Property 1:

\[
\begin{align*}
\Pr \left\{ \frac{1}{n} \log P_{X^n}(X^n) < 0 \right\} &= 0, \\
\Pr \left\{ \frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\hat{X}^n}(X^n)} < -\nu \right\} &= P_{X^n} \left\{ x^n \in \mathcal{X}^n : \frac{1}{n} \log \frac{P_{X^n}(x^n)}{P_{\hat{X}^n}(x^n)} < -\nu \right\} \\
&= \sum \left\{ x^n \in \mathcal{X}^n : P_{X^n}(x^n) < P_{\hat{X}^n}(x^n) e^{-n\nu} \right\} P_{X^n}(x^n) \\
&\leq \sum \left\{ x^n \in \mathcal{X}^n : P_{X^n}(x^n) < P_{\hat{X}^n}(x^n) e^{n\nu} \right\} P_{\hat{X}^n}(x^n) e^{-n\nu} \\
&\leq e^{-n\nu} \cdot \sum \left\{ x^n \in \mathcal{X}^n : P_{X^n}(x^n) < P_{\hat{X}^n}(x^n) e^{n\nu} \right\} P_{\hat{X}^n}(x^n) \\
&\leq e^{-\nu n},
\end{align*}
\]

and, by following the same procedure as (2.4.6),

\[
\begin{align*}
\Pr \left\{ \frac{1}{n} \log \frac{P_{X^n,Y^n}(X^n,Y^n)}{P_{X^n}(X^n)P_{Y^n}(Y^n)} < -\nu \right\} &\leq e^{-\nu n}.
\end{align*}
\]
Properties of generalized information measures

Property 2: An immediate consequence of the definition.

Property 3: Follow from the facts that

$$\frac{1}{n} h_{Y^n}(Y^n) = \frac{1}{n} i_{X^n,Y^n}(X^n; Y^n) + \frac{1}{n} h_{X^n,Y^n}(Y^n|X^n),$$

where

$$\frac{1}{n} h_{X^n,Y^n}(Y^n|X^n) \triangleq -\frac{1}{n} \log P_{Y^n|X^n}(Y^n|X^n).$$

Property 4: $\bar{H}_\delta(\cdot)$ is non-decreasing in $\delta$, $\bar{H}_\delta(X) \leq \bar{H}(X)$, and $\bar{H}(X) \leq \log |\mathcal{X}|$. The last inequality can be proved as follows.

$$\Pr \left\{ \frac{1}{n} h_{X^n}(X^n) \leq \log |\mathcal{X}| + \nu \right\}$$

$$= 1 - P_{X^n} \left\{ x^n \in \mathcal{X}^n : \frac{1}{n} \log \frac{P_{X^n}(X^n)}{1/|\mathcal{X}|^n} < -\nu \right\}$$

$$\geq 1 - e^{-n\nu},$$

where the last step can be obtained by using the same procedure as (2.4.6). Therefore, $h(\log |\mathcal{X}| + \nu) = 1$ for any $\nu > 0$, which indicates $\bar{H}(X) \leq \log |\mathcal{X}|$. 
Properties of generalized information measures

Property 5:

\[
\frac{1}{n} i_{X^n, Y^n, Z^n}(X^n, Y^n; Z^n) = \frac{1}{n} i_{X^n, Z^n}(X^n; Y^n; Z^n) + \frac{1}{n} i_{X^n, Y^n, Z^n}(Y^n; Z^n|X^n).
\]
Properties of generalized information measures

Lemma 2.5 (data processing lemma) Fix $\delta \in [0, 1]$. Suppose that for every $n$, $X_1^n$ and $X_3^n$ are conditionally independent given $X_2^n$. Then

$$I_\delta(X_1; X_3) \leq I_\delta(X_1; X_2).$$

Proof: By property 5 of Lemma 2.4, we get

$$I_\delta(X_1; X_3) \leq I_\delta(X_1; X_2, X_3) = I_\delta(X_1; X_2),$$

where the equality holds because

$$\frac{1}{n} \log \frac{P_{X_1^n, X_2^n, X_3^n}(x_1^n, x_2^n, x_3^n)}{P_{X_1^n}(x_1^n)P_{X_2^n, X_3^n}(x_2^n, x_3^n)} = \frac{1}{n} \log \frac{P_{X_1^n, X_2^n}(x_1^n, x_2^n)}{P_{X_1^n}(x_1^n)P_{X_2^n}(x_2^n)}.$$

Lemma 2.6 (optimality of independent inputs) Fix $\delta \in [0, 1)$. Consider a finite-alphabet channel with $P_{W^n}(y^n|x^n) = P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^n P_{Y_i|X_i}(y_i|x_i)$ for all $n$. For any input $X$ and its corresponding output $Y$,

$$I_\delta(X; Y) \leq I_\delta(\bar{X}; \bar{Y}) = I(\bar{X}; \bar{Y}),$$

where $\bar{Y}$ is the output due to $\bar{X}$, which is an independent process with the same first order statistics as $X$, i.e., $P_{\bar{X}^n}(x^n) = \prod_{i=1}^n P_{X_i}(x_i)$. 
System setting:

- Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and
  \[
  Y_i^{(n)} = X_i^{(n)} \oplus Z_i^{(n)}
  \]
  where the arbitrary noise $Z$ is independent of the channel input $X$.

- Assume that $X$ is an i.i.d. random process with equal-probably marginal distribution.

- Then the resultant channel output $Y$ is also an i.i.d. random process with equal-probably marginal distribution.
Computation of general information measures

Derivations:

\[ \bar{i}(\theta) \triangleq \limsup_{n \to \infty} \Pr \left\{ \frac{1}{n} \log \frac{P_{Y^n|X^n}(Y^n|X^n)}{P_{Y^n}(Y^n)} \leq \theta \right\} \]

\[ = \limsup_{n \to \infty} \Pr \left\{ \frac{1}{n} \log P_{Z^n}(Z^n) - \frac{1}{n} \log P_{Y^n}(Y^n) \leq \theta \right\} \]

\[ = \limsup_{n \to \infty} \Pr \left\{ \frac{1}{n} \log P_{Z^n}(Z^n) \leq \theta - \log(2) \right\} \]

\[ = \limsup_{n \to \infty} \Pr \left\{ -\frac{1}{n} \log P_{Z^n}(Z^n) \geq \log(2) - \theta \right\} \]

\[ = 1 - \liminf_{n \to \infty} \Pr \left\{ -\frac{1}{n} \log P_{Z^n}(Z^n) < \log(2) - \theta \right\} . \]
Hence, for \( \varepsilon \in (0, 1) \),

\[
I_\varepsilon(X; Y) = \sup \{ \theta : \bar{i}(\theta) \leq \varepsilon \}
= \sup \left\{ \theta : 1 - \liminf_{n \to \infty} \Pr \left\{ -\frac{1}{n} \log P_{Z^n}(Z^n) < \log(2) - \theta \right\} \leq \varepsilon \right\}
= \sup \left\{ \theta : \liminf_{n \to \infty} \Pr \left\{ -\frac{1}{n} \log P_{Z^n}(Z^n) < \log(2) - \theta \right\} \geq 1 - \varepsilon \right\}
= \sup \left\{ \log(2) - \beta : \liminf_{n \to \infty} \Pr \left\{ -\frac{1}{n} \log P_{Z^n}(Z^n) < \beta \right\} \geq 1 - \varepsilon \right\}
= \log(2) + \sup \left\{ -\beta : \liminf_{n \to \infty} \Pr \left\{ -\frac{1}{n} \log P_{Z^n}(Z^n) < \beta \right\} \geq 1 - \varepsilon \right\}
= \log(2) - \inf \left\{ \beta : \liminf_{n \to \infty} \Pr \left\{ -\frac{1}{n} \log P_{Z^n}(Z^n) < \beta \right\} \geq 1 - \varepsilon \right\}
= \log(2) - \sup \left\{ \beta : \liminf_{n \to \infty} \Pr \left\{ -\frac{1}{n} \log P_{Z^n}(Z^n) < \beta \right\} < 1 - \varepsilon \right\}
\leq \log(2) - \sup \left\{ \beta : \liminf_{n \to \infty} \Pr \left\{ -\frac{1}{n} \log P_{Z^n}(Z^n) \leq \beta \right\} < 1 - \varepsilon \right\}
= \log(2) - \lim_{\delta \uparrow (1-\varepsilon)} \bar{H}_\delta(Z).
Computation of general information measures

Also, for $\varepsilon \in (0, 1)$,

$$I_\varepsilon(X;Y) \geq \sup \left\{ \theta : \limsup_{n \to \infty} \Pr \left[ \frac{1}{n} \log \frac{P_{X^n,Y^n}(X^n,Y^n)}{P_{X^n}(X^n)P_{Y^n}(Y^n)} < \theta \right] < \varepsilon \right\}$$

$$= \log(2) - \sup \left\{ \beta : \liminf_{n \to \infty} \Pr \left\{ -\frac{1}{n} \log P_{Z^n}(Z^n) \leq \beta \right\} \leq 1 - \varepsilon \right\}$$

$$= \log(2) - \bar{H}_{(1-\varepsilon)}(Z),$$

Therefore,

$$\log(2) - \bar{H}_{(1-\varepsilon)}(Z) \leq I_\varepsilon(X;Y) \leq \log(2) - \lim_{\gamma \uparrow (1-\varepsilon)} \bar{H}_{\gamma}(Z) \quad \text{for } \varepsilon \in (0, 1).$$

By taking $\varepsilon \downarrow 0$, we obtain

$$I(X;Y) = I_0(X;Y) = \log(2) - \bar{H}(Z).$$

Based on this result, we can now compute $I_\varepsilon(X;Y)$ for some specific examples.
Example 2.7

\[ Z = \begin{cases} 
\text{all-zero sequence with probability } \beta; \\
\text{Bernoulli (with parameter } p\text{) with probability } 1 - \beta.
\end{cases} \]

Then

\[
\frac{1}{n} h_{Z^n}(Z^n) \rightarrow \begin{cases} 
0, & \text{with probability } \beta; \\
h_b(p), & \text{with probability } 1 - \beta,
\end{cases}
\]

where \( h_b(p) \triangleq -p \log p - (1 - p) \log(1 - p) \).
Therefore,

\[ I_\varepsilon(X; Y) = \begin{cases} 
1 - h_b(p), & \text{if } 0 < \varepsilon < 1 - \beta; \\
1, & \text{if } 1 - \beta \leq \varepsilon < 1. 
\end{cases} \]
Computation of general information measures

**Example 2.8** $Z$ = non-stationary binary independent sequence with

$$\Pr \{ Z_i^{(n)} = 0 \} = 1 - \Pr \{ Z_i^{(n)} = 1 \} = p_i,$$

then by the fact that

$$\Var \left[ - \log P_{Z_i^{(n)}}(Z_i^{(n)}) \right] \leq E \left[ \left( \log P_{Z_i^{(n)}}(Z_i^{(n)}) \right)^2 \right]$$

$$\leq \sup_{0 < p_i < 1} [p_i (\log p_i)^2 + (1 - p_i) (\log (1 - p_i))^2]$$

$$< \log(2),$$

we have (by Chebyshev’s inequality)

$$\Pr \left\{ \left| - \frac{1}{n} \log P_{Z_i^{(n)}}(Z_i^{(n)}) - \frac{1}{n} \sum_{i=1}^{n} H \left( Z_i^{(n)} \right) \right| > \gamma \right\} \rightarrow 0,$$

for any $\gamma > 0.$
Computation of general information measures

The possible limiting spectrum of \((1/n)h_{Z^n}(Z^n)\).

The possible limiting spectrums of \((1/n)i_{X^n,Y^n}(X^n;Y^n)\).
Computation of general information measures

Therefore, $\bar{H}_{(1-\varepsilon)}(\mathbf{Z})$ is equal to

$$
\bar{H}_{(1-\varepsilon)}(\mathbf{Z}) = \begin{cases} 
\bar{H}(\mathbf{Z}) = \limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} H(Z_{i}^{(n)}) & \text{for } \varepsilon \in (0, 1]; \\
= \limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} h_b(p_i) & \text{for } \varepsilon = 0.
\end{cases}
$$

Consequently,

$$
I_\varepsilon(\mathbf{X}; \mathbf{Y}) = \begin{cases} 
1 - \bar{H}(\mathbf{Z}) = 1 - \limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} h_b(p_i) & \text{for } \varepsilon \in [0, 1), \\
\infty, & \text{for } \varepsilon = 1.
\end{cases}
$$
Rényi’s information measures

In this section, we will introduce alternative generalizations of information measures. They are respectively named

- Rényi’s entropy
- Rényi’s mutual information
- Rényi’s divergence

Definition 2.9 (Rényi’s entropy) For \( \alpha > 0 \), the Rényi’s entropy of order \( \alpha \) is defined by:

\[
H(X; \alpha) \triangleq \begin{cases} 
\frac{1}{1 - \alpha} \log \left( \sum_{x \in \mathcal{X}} [P_X(x)]^\alpha \right), & \text{for } \alpha \neq 1; \\
\lim_{\alpha \to 1} H(X; \alpha) = - \sum_{x \in \mathcal{X}} P_X(x) \log P_X(x), & \text{for } \alpha = 1.
\end{cases}
\]

Definition 2.10 (Rényi’s divergence) For \( \alpha > 0 \), the Rényi’s divergence of order \( \alpha \) is defined by:

\[
D(X \parallel \hat{X}; \alpha) \triangleq \begin{cases} 
\frac{1}{\alpha - 1} \log \left( \sum_{x \in \mathcal{X}} \left[ P_X^\alpha(x) P_{\hat{X}}^{1-\alpha}(x) \right] \right), & \text{for } \alpha \neq 1; \\
\lim_{\alpha \to 1} D(X \parallel \hat{X}; \alpha) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{P_X(x)}{P_{\hat{X}}(x)}, & \text{for } \alpha = 1.
\end{cases}
\]
Rényi’s information measures

There are two possible Rényi’s extensions for mutual information. One is based on the observation of

\[
I(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{P_Y(y)}
\]

\[
= \min_{P_{\hat{Y}}} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{P_{\hat{Y}}(y)}.
\]

The other extension is a direct generalization of

\[
I(X; Y) = D(P_{X,Y} \| P_X \times P_Y) = \min_{P_{\hat{Y}}} D(P_{X,Y} \| P_X \times P_{\hat{Y}})
\]

to Rényi’s divergence.
Rényi’s information measures

**Definition 2.11 (type-I Rényi’s mutual information)** For $\alpha > 0$, the type-I Rényi’s mutual information of order $\alpha$ is defined by:

$$I(X; Y; \alpha) \triangleq \begin{cases} \min_{P_Y} \frac{1}{\alpha - 1} \sum_{x \in \mathcal{X}} P_X(x) \log \left( \sum_{y \in \mathcal{Y}} \left[ P_{Y|X}^\alpha(y|x) P_{Y}^{1-\alpha}(y) \right] \right), & \text{if } \alpha \neq 1; \\
lim_{\alpha \to 1} I(X; Y; \alpha) = I(X; Y), & \text{if } \alpha = 1,
\end{cases}$$

where the minimization is taken over all $P_Y$ under fixed $P_X$ and $P_{Y|X}$.

**Definition 2.12 (type-II Rényi’s mutual information)** For $\alpha > 0$, the type-II Rényi’s mutual information of order $\alpha$ is defined by:

$$J(X; Y; \alpha) \triangleq \min_{P_Y} D(P_{X,Y} \parallel P_X \times P_{Y}; \alpha)$$

$$= \begin{cases} \frac{\alpha}{\alpha - 1} \log \left[ \sum_{y \in \mathcal{Y}} \left( \sum_{x \in \mathcal{X}} P_X(x) P_{Y|X}^\alpha(y|x) \right)^{1/\alpha} \right], & \text{for } \alpha \neq 1; \\
lim_{\alpha \to 1} J(X; Y; \alpha) = I(X; Y), & \text{for } \alpha = 1.
\end{cases}$$
Rényi’s information measures

Lemma 2.13 For finite alphabet $\mathcal{X}$, the following statements hold.

1. $0 \leq H(X; \alpha) \leq \log |\mathcal{X}|$; the first equality holds if, and only if, $X$ is deterministic, and the second equality holds if, and only if, $X$ is uniformly distributed over $\mathcal{X}$.

2. $H(X; \alpha)$ is strictly decreasing in $\alpha$ unless $X$ is uniformly distributed over its support $\{x \in \mathcal{X} : P_X(x) > 0\}$.

3. $\lim_{\alpha \downarrow 0} H(X; \alpha) = \log |\{x \in \mathcal{X} : P_X(x) > 0\}|$.

4. $\lim_{\alpha \to \infty} H(X; \alpha) = -\log \max_{x \in \mathcal{X}} P_X(x)$.

5. $D(X \| \hat{X}; \alpha) \geq 0$ with equality holds if, and only if, $P_X = P_{\hat{X}}$.

6. $D(X \| \hat{X}; \alpha) = \infty$ if, and only if, either

$$\{x \in \mathcal{X} : P_X(x) > 0 \text{ and } P_{\hat{X}}(x) > 0\} = \emptyset$$

or

$$\{x \in \mathcal{X} : P_X(x) > 0\} \not\subset \{x \in \mathcal{X} : P_{\hat{X}}(x) > 0\} \text{ for } \alpha \geq 1.$$

7. $\lim_{\alpha \downarrow 0} D(X \| \hat{X}; \alpha) = -\log P_{\hat{X}}\{x \in \mathcal{X} : P_X(x) > 0\}$.
Rényi’s information measures

8. If $P_X(x) > 0$ implies $P_{\hat{X}}(x) > 0$, then

$$\lim_{\alpha \to \infty} D(X\|\hat{X}; \alpha) = \max_{\{x \in \mathcal{X} : P_{\hat{X}}(x) > 0\}} \log \frac{P_X(x)}{P_{\hat{X}}(x)}.$$ 

9. $I(X; Y; \alpha) \geq J(X; Y; \alpha)$ for $0 < \alpha < 1$, and $I(X; Y; \alpha) \leq J(X; Y; \alpha)$ for $\alpha > 1$.

**Lemma 2.14 (data processing lemma for type-I Rényi’s mutual information)** Fix $\alpha > 0$. If $X \to Y \to Z$, then $I(X; Y; \alpha) \geq I(X; Z; \alpha)$. 