Application of Information Theory, Lecture 6
Differential Entropy and Other Entropy Measures

Handout Mode

Iftach Haitner

Tel Aviv University.

December 05, 2019
Part I

Differential Entropy
Entropy of continues random variable

- Entropy of discrete random variable: $H(X) = - \sum_i p_i \log p_i$
- Also used when $X$ has infinite support (entropy might be infinite)
- Continues random variable is defined by its density function: $f : \mathbb{R} \mapsto \mathbb{R}^+$, for which $\int_{\mathbb{R}} f(x)dx = 1$.
- $F_X(x) := \Pr [X \leq x] = \int_{-\infty}^{x} f(x)dx$
- $E X = \int x \cdot f(x)dx$ and $V X = \int x^2 \cdot f(x)dx - (E X)^2$
- Examples: $X \sim [0, 1]$, $X \sim N(0, 1)$
- $H(X)$ must be infinite! it takes infinite number of bits to describe $X$
- The differential entropy of $X$ is defined by $h(X) = - \int f(x) \log f(x)dx$.
- We focus on cases where $h(X)$ is well defined.
- Since $h$ is a function of the density function, we sometimes write $h(f)$
- If not stated otherwise, we integrate over $\mathbb{R}$
Intuition for definition of $h$

Let $X^\Delta$ be rounding of $X$ for precision $\Delta$:

$$X^\Delta \sim (\ldots, p_{-2}, p_{-1}, p_0, p_1, p_2, \ldots),$$

where $p_i = \int_{i \cdot \Delta}^{(i+1) \cdot \Delta} f(x) dx = f(x_i) \cdot \Delta$

for some $x_i \in [i \cdot \Delta, (i + 1) \cdot \Delta]$ (?)

$$H(X^\Delta) = - \sum_{i=-\infty}^{\infty} p_i \log p_i$$

$$H(X^\Delta) = - \sum_{i=-\infty}^{\infty} f(x_i) \cdot \Delta \cdot \log(f(x_i) \cdot \Delta) = - \sum_{i=-\infty}^{\infty} f(x_i) \cdot \Delta \cdot (\log f(x_i) + \log \Delta)$$

$$= - \sum_{i=-\infty}^{\infty} f(x_i) \cdot \log f(x_i) \cdot \Delta - \left( \sum_{i=-\infty}^{\infty} f(x_i) \cdot \Delta \right) \log \Delta$$

$$\lim_{\Delta \to 0} H(X^\Delta) = h(X) - \lim_{\Delta \to 0} \log \Delta$$

Hence, $\lim_{\Delta \to 0}(H(X^\Delta) + \log \Delta) = h(x)$

Intuitively, $h(X)$ is the entropy of $X$ plus const $(\lim_{\Delta \to 0} - \log \Delta)$.

Note that $\lim_{\Delta \to 0} - \log \Delta = \infty$
Properties of the entropy function

\[ h(X) = -\int f(x) \log f(x) \, dx \]

- **Shift invariant:** \( h(f) = h(g) \) for \( g(x) = f(x + a) \)
- \( h(f) \) might be infinite
- For any discrete \( X \) exists \( f \) with \( h(f) = H(X) \):
  for \( X \sim (p_1, p_2, \ldots) \), let \( f_{\tilde{X}}(x) = p_i \) for all \( x \in [i, i + 1] \)
- \( h(X) \) might be negative
- Example: \( X \sim [0, a] \) – \( f(x) = \frac{1}{a} \) on \( [0, a] \)
  \[ -\int f(x) \log f(x) \, dx = -\log \frac{1}{a} = \log a \]. Negative for \( a < 1 \).
- \( h(X) \) should be interpreted as the uncertainty up to a certain constant
- Used for comparing two distributions
Common distribution (in nature)

- The uniform distribution: $X \sim [a, b]$
- Normal (Gaussian) distribution: (we focus on $E = 0$ and $V = 1$)
  $X \sim N(0, 1): f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}$
- Boltzmann (Gibbs) distribution:
  $X \in \{E_1, E_2, \ldots, E_m\}$, $\Pr [X = E_i] = C \cdot e^{-\beta E_i}$ for $\beta > 0$ (the distribution constant) and $C = 1/\sum_i e^{-\beta E_i}$.
  - Describes a (discrete) physical system that can take states $\{1, \ldots, m\}$ with energies $E_1, \ldots, E_m$.
  - Probability is inverse to the energy

Why are these distributions so common?
- What is common to these distributions?
Second law of thermodynamics

- The entropy of a closed physical system never decreases.
- If we wait enough time, the system tends to be in maximal entropy.
- If there are constraints, it tends to be in maximal entropy under this constraint.
- This suggests that distributions that are common in nature, are distributions of maximal entropy, under some constraints.
The normal distribution

- \( X \sim N(0, 1) \): 
  \[
  f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}
  \]

- Why is it so common?
- Answer: the central limit theorem (CLT):
  
  Let \( X_1, \ldots, X_n \) be iid with \( E X_i = 0 \) and \( V X_i = 1 \). Then
  \[
  \lim_{n \to \infty} \frac{\sum_i X_i}{\sqrt{n}} = N(0, 1).
  \]

- But why does it converge to \( N(0, 1) \)??
- CLT holds also in many other variants: not id, not fully independent, ...
- We know that \( E \frac{\sum_i X_i}{\sqrt{n}} = 0 \) and \( V \frac{\sum_i X_i}{\sqrt{n}} = 1 \), but it could have converge to any other distribution with these constraints.

- The reason is that \( N(0, 1) \) has the highest entropy among all distribution with \( E = 0 \) and \( V = 1 \).
- CLT and the normal distribution where known and studied way before Shannon, yet this striking property was not known until his theory.
The normal distribution, cont.

Theorem 1

\[ h(X) \leq h(N(0, 1)), \text{ for any } v \text{ X with } \sqrt{V} X = 1. \]

- Among the distributions of \( V = 1 \), the distribution \( N(0, 1) \) has maximal entropy.
- Generalizes to any variance:
  \[ h(X) \leq h(N(0, V(X))) = \frac{1}{2} \cdot \log(2\pi e) \cdot V(X) \]

Let \( g \) be a density function with \( \int g(x)x^2dx = 1 \), and let \( f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \).

We will show that

1. \[ -\int g(x) \log g(x)dx \leq -\int g(x) \log f(x)dx \]
2. \[ -\int g(x) \log f(x)dx = -\int f(x) \log f(x)dx \]
$- \int g(x) \log g(x) dx \leq - \int g(x) \log f(x) dx$

**Claim 2**

$- \int g(x) \log g(x) dx \leq - \int g(x) \log q(x) dx$ for any two density functions $q, g$.

**Proof:**

- **Jensen:** For any function $t$ and density function $\lambda$:
  
  $\int \lambda(x) \log t(x) \leq \log \int \lambda(x) t(x) dx$

- Assume for simplicity that $g(x) > 0$ for all $x$.

- By Jensen, $\int g(x) \log \frac{q(x)}{g(x)} \leq \log \int g(x) \frac{q(x)}{g(x)} dx = \log 1 = 0$

- Hence, $- \int g(x) \log g(x) \leq - \int g(x) \log q(x)$
\[- \int g(x) \log f(x) \, dx = - \int f(x) \log f(x) \, dx \]

**Claim 3**

Exists $c \in \mathbb{R}$ such that $- \int g(x) \log f(x) \, dx = c$ for any density function $g$ with $\int g(x)x^2 \, dx = 1$.

Hence, $- \int g(x) \log f(x) \, dx = - \int f(x) \log f(x) \, dx$.

**Proof:**

\[
- \int g(x) \log f(x) \, dx = - \int g(x) \log \left( \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \right) \, dx \\
= - \int g(x) \left( \log \frac{1}{\sqrt{2\pi}} - \frac{x^2}{2} \cdot \log e \right) \, dx \\
= - \log \frac{1}{\sqrt{2\pi}} \int g(x) \, dx + \frac{\log e}{2} \int g(x)x^2 \, dx \\
= - \log \frac{1}{\sqrt{2\pi}} + \frac{\log e}{2}.
\]
The Boltzmann distribution

- States \( \{1, \ldots, m\} \), energies \( E_1, \ldots, E_m \).
- \( \text{Pr} [X = E_i] = C \cdot e^{-\beta E_i} \) for \( \beta > 0 \) and \( C = 1 / \sum_i e^{-\beta E_i} \).
- We will denote it by \( \sim B(\beta, E_1, \ldots, E_m) \).
- Like the exponential distribution (i.e., \( f(x) = \lambda e^{-\lambda x} \)), but discrete.
  - Describes a (discrete) physical system that can take states \( \{1, \ldots, m\} \) with energies \( E_1, \ldots, E_m \).
  - Probability is inverse to energy.

**Theorem 4**

Let \( X \sim B(\beta, E_1, \ldots, E_m) \). Then \( H(Y) \leq H(X) \) for any rv \( Y \) over \( \{E_1, \ldots, E_m\} \), with \( EY = EX \).

- The Boltzmann distribution is maximal among all distributions of the same energy.
Proving Theorem 4

- **X** ~ $B(\beta, E_1, \ldots, E_m)$ and **E Y = E X**
- Let **X** ~ $(p_1, \ldots, p_m)$ and **Y** ~ $(q_1, \ldots, q_m)$ over $\{E_1, \ldots, E_m\}$.
- $H(Y) \leq \sum_i q_i \log p_i$ (Q3 in Handout 1)
- Let $C = 1/\sum_i e^{-\beta E_i}$.

Then
\[
\sum_i q_i \log p_i = \sum_i q_i \log (C \cdot e^{-\beta E_i})
= \sum_i q_i \log C - \sum_i q_i \beta E_i \cdot \log e
= \log C - \beta \cdot \log e \cdot \sum_i q_i E_i
= \log C - \beta \cdot \log e \cdot E_X
\]

- Hence, $\sum_i q_i \log p_i = \sum_i p_i \log p_i$. ☐
The uniform distribution

- $X \sim [a, b]$.
- $E X = \frac{1}{2} (a + b)$ and $V X = \frac{1}{12} (b - a)^2$
- What come to mind when saying “$X$ takes values in $[0, 1]$”.

Theorem 5

$h(X) \leq h(\sim [a, b])$, for any RV with $\text{Supp}(X) \subseteq [a, b]$.

Proof: HW
Using diff. entropy to bound discrete entropy

**Proposition 6**

Let \( X \sim (p_1, p_2, \ldots) \), then \( H(X) \leq \log \frac{2\pi e}{2} \cdot (V(X) + \frac{1}{12}) \)

We assume wlg. that \( p_i = \Pr[X = i] \).

- Let \( U \sim [0, 1] \), let \( \tilde{X} = X + U \) and let \( f_{\tilde{X}} \) be the density function of \( \tilde{X} \).

\[
H(X) = -\sum_{i=1}^{\infty} p_i \log p_i \\
= -\sum_{i=1}^{\infty} \left( \int_{i}^{i+1} f_{\tilde{X}}(x) \, dx \right) \cdot \log p_i = -\sum_{i=1}^{\infty} \int_{i}^{i+1} f_{\tilde{X}}(x) \log p_i \, dx \\
= -\sum_{i=1}^{\infty} \int_{i}^{i+1} f_{\tilde{X}}(x) \log f_{\tilde{X}}(x) \, dx \quad \text{(} f_{\tilde{X}}(x) = p_i \text{ for all } x \in [i, i + 1] \text{)} \\
= -\int_{1}^{\infty} f_{\tilde{X}}(x) \log f_{\tilde{X}}(x) \, dx \\
= h(\tilde{X})
\]
Using diff. entropy to bound discrete entropy, cont.

Hence,

\[ H(X) = h(\tilde{X}) \]
\[ \leq \frac{1}{2} \log(2\pi e) V(\tilde{X}) \]
\[ = \frac{1}{2} \log(2\pi e) (V(X) + V(U)) \]
\[ = \frac{\log 2\pi e}{2} \cdot \left( \left( \sum_{i=1}^{\infty} p_i \cdot i^2 - \left( \sum_{i=1}^{\infty} p_i \cdot i \right)^2 \right) + \frac{1}{12} \right) \]

How good is this bound?

Let \( X \sim (\frac{1}{2}, \frac{1}{2}) \). Hence, \( V[X] = \frac{1}{4} \) and \( H(X) = 1 \).

Proposition 6 grantees that \( H(X) \leq \frac{\log 2\pi e}{2} \left( \frac{1}{4} + \frac{1}{12} \right) \sim 1.255 \)
Part II

Statistical Distance
Statistical distance

Let \( p = (p_1, \ldots, p_m) \) and \( q = (q_1, \ldots, q_m) \) be distributions over \([m]\).

Their statistical distance (also known as variation distance) is defined by

\[
SD(p, q) := \frac{1}{2} \sum_{i \in [m]} |p_i - q_i|
\]

This is simply the \( L_1 \) norm between the distribution vectors.

We will soon see another “distance” measures for distributions next lecture.

For \( X \sim p \) and \( Y \sim q \), let \( SD(X, Y) = SD(p, q) \).

Claim (HW): \( SD(p, q) = \max_{S \subseteq [m]} \left( \sum_{i \in S} p_i - \sum_{i \in S} q_i \right) \)

Hence, \( SD(p, q) = \max_D \left( \Pr_{X \sim p} [D(X) = 1] - \Pr_{X \sim q} [D(X) = 1] \right) \)

Interpretation

Claim (data processing): \( SD(f(X), f(Y)) \leq SD(X, Y) \) for any function \( f \).
Distance from the uniform distribution

- Let $X$ be rv over $[m]$
- $H(X) \leq \log m$
- $H(X) = \log m \iff X$ is uniform over $[m]$

**Theorem 7 (next lecture)**

Let $X$ rv over $[m]$. Assume $H(X) \geq \log m - \varepsilon$, then

$$\text{SD}(X, \sim [m]) \leq \sqrt{\varepsilon \cdot \frac{\ln 2}{2}} = O(\sqrt{\varepsilon})$$
Part III

Other Entropy Measures
Other entropy measures

Let $X \sim p$ be a random variable over $\mathcal{X}$.

- Recall that Shannon entropy of $X$ is
  \[ H(X) = \sum_{x \in \mathcal{X}} -p(x) \cdot \log p(x) = E_X [-\log p(X)] = E_X [H_X(X)] \]

- Max entropy of $X$ is $H_0(X) = \log |\text{Supp}(X)|$

- Min entropy of $X$ is $H_\infty(X) = \min_{x \in \mathcal{X}} \{-\log p(x)\} = -\log \max_{x \in \mathcal{X}} \{p(x)\}$

- Collision probability of $X$ is $\text{CP}(X) = \sum_{x \in \mathcal{X}} p(x)^2 = \|p\|_2^2$
  Probability of collision when drawing two independent samples from $X$

- Collision entropy/Renyi entropy of $X$ is $H_2(X) = -\log \text{CP}(X)$

- For $\alpha \neq 1 \in \mathbb{N}$ — $H_\alpha = \frac{1}{1-\alpha} \log \left( \sum_{i=1}^n p_i^\alpha \right) = \frac{\alpha}{1-\alpha} \log(\|p\|_\alpha)$

- $H_\infty(X) \leq H_2(X) \leq H(X) \leq H_0(X)$ (Jensen)
  Equality iff $X$ is uniform over $\mathcal{X}$

- For instance, $\text{CP}(X) \leq \sum_{x} p(x) \max_{x'} p(x') = \max_{x'} p(x')$. Hence, $H_2(X) \geq -\log \max_{x'} p(x') = H_\infty(X)$.

- Claim: $H_2(X) \leq 2 \cdot H_\infty(X)$

- Proof: $\text{CP}(X) \geq (\max_{x'} p(x'))^2$. Hence, $-\log \text{CP}(X) \leq -2 H_\infty(X)$
Other entropy measures, cont

- No simple chain rule.
- Let $X = \perp$ with probability $\frac{1}{2}$ and uniform over $\{0, 1\}^n$ otherwise, and let $Y$ be indicator for $X = \perp$.
- $H_\infty(X|Y = 1) = 0$ and $H_\infty(X|Y = 0) = n$. But $H_\infty(X) = 1$. 
Section 1

Shannon to min entropy
Shannon to Min entropy

Given rv $X \sim p$, let $X^n$ denote $n$ independent copies of $X$, and let $p^n(x_1 \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$.

Lemma 8

Let $X \sim p$ and let $\varepsilon > 0$. Then $\Pr \left[ -\log p^n(X^n) \leq n \cdot (H(X) - \varepsilon) \right] < 2 \cdot e^{-2\varepsilon^2 n}$.

▶ Compare to $-\log p^n(x) \leq n \cdot H_\infty(X)$, for any $x \in \text{Supp}(X^n)$

Corollary 9

$\exists$ rv $W$ that is $(2 \cdot e^{-2\varepsilon^2 n})$-close to $X^n$, and $H_\infty(W) \geq n(H(X) - \varepsilon)$.

Proof: $W = X^n$ if $X^n \in A_{n,\varepsilon} := \{x \in \text{Supp}(X^n) : 2^{-n(H(X)+\varepsilon)} \leq p^n(x) \leq 2^{-n(H(X)-\varepsilon)}\}$, and “well spread” outside $\text{Supp}(X^n)$ otherwise.
Shannon to min entropy, proof

\[ p^n(x_1 \ldots, x_n) = \prod_{i=1}^{n} p(x_i). \]

**Lemma 10 (Restatment of Lemma 8)**

Let \( X \sim p \) and let \( \varepsilon > 0 \). Then \( \Pr \left[ -\log p^n(X^n) \leq n \cdot (H(X) - \varepsilon) \right] < 2 \cdot e^{-2\varepsilon^2 n}. \)

**Proof:** (quantitative) AEP.

**Proposition 11 (Hoeffding’s inequality)**

Let \( Z_1, \ldots, Z_n \) be iids over \([0, 1]\) with expectation \( \mu \). Then,

\[ \Pr \left[ \left| \frac{1}{n} \sum_{i=1}^{n} Z_i - \mu \right| \geq \varepsilon \right] \leq 2 \cdot e^{-2\varepsilon^2 n} \text{ for every } \varepsilon > 0. \]

- \( A_{n,\varepsilon} := \{ x \in \text{Supp}(X^n) : 2^{-n(H(X)+\varepsilon)} \leq p^n(x) \leq 2^{-n(H(X)-\varepsilon)} \} \)
- \( -\log p^n(x) \geq n \cdot (H(X) - \varepsilon) \) for any \( x \in A_{n,\varepsilon} \)
- Taking \( Z_i = -\log p(X_i) \) and \( \mu = H(X_1) \), it follows that

\[ \Pr \left[ X^n \notin A_{n,\varepsilon} \right] = \Pr \left[ \left| n\mu - \sum_{i} Z_i \right| \geq \varepsilon n \right] = \Pr \left[ \left| \mu - \frac{1}{n} \sum_{i} Z_i \right| \geq \varepsilon \right] \leq 2 \cdot e^{-2\varepsilon^2 n}. \]
Shannon to Min entropy, conditional version

**Lemma 12**

Let \((X, Y) \sim p\) let \(\varepsilon > 0\). Then

\[
\Pr_{(x^n, y^n) \leftarrow (X, Y)^n} \left[ -\log p^n_{X|Y}(x^n|y^n) \leq n \cdot (H(X|Y) - \varepsilon) \right] < 2 \cdot e^{-2\varepsilon^2 n}.
\]

**Proof:** same proof, letting \(Z_i = \log p_{X|Y}(X_i|Y_i)\)

**Corollary 13**

\exists rv \(W\) over \(X^n \times Y^n\) such that

1. \(\text{SD}(W, (X, Y)^n) \leq 2 \cdot e^{-2\varepsilon^2 n}\).
2. \(\text{SD}(W_{Y^n}, Y^n) = 0\), and
3. \(H_\infty(W_{X^n}|W_{Y^n} = y) \geq n \cdot (H(X|Y) - \varepsilon)\), for any \(y \in \text{Supp}(Y^n)\)

**Proof:** ?
Section 2

Renyi-entropy to Uniform Distribution
Extraction

Goal: given a random variable over \( \{0, 1\}^n \), with \( k \) bits of “entropy”, extract close to \( k \) uniform bits.

- Let \( U_t \sim \{0, 1\}^t \).
- Deterministic extractors:
  \[ \text{Ext}: \{0, 1\}^n \mapsto \{0, 1\}^m \], such that for any \( X \) over \( \{0, 1\}^n \) with \( H_\infty(X) \geq k \) it holds that \( \text{SD}(\text{Ext}(X), U_m) \leq \varepsilon \).
- Impossible to achieve even for \( k = n - 1 \) and \( m = 1! \)
- Seeded extractors:
  \[ \text{Ext}: \{0, 1\}^n \times \{0, 1\}^d \mapsto \{0, 1\}^m \], such that for any \( X \) over \( \{0, 1\}^n \) with \( H_\infty(X) \geq k \) it holds that \( \text{SD}((\text{Ext}(X, U_d), U_d), (U_m, U_d)) \leq \varepsilon \).
- Typically, we would like \( d \) to be us small as possible.
- Very useful concept
Pairwise independent hashing

**Definition 14 (pairwise independent function family)**

A function family \( G = \{ g: \mathcal{D} \rightarrow \mathcal{R} \} \) is pairwise independent, if \( \forall x \neq x' \in \mathcal{D} \) and \( y, y' \in \mathcal{R} \), it holds that \( \Pr_{g \leftarrow G} [g(x) = y \land g(x') = y'] = \left( \frac{1}{|\mathcal{R}|} \right)^2 \).

- Example: for \( \mathcal{D} = \{0, 1\}^n \) and \( \mathcal{R} = \{0, 1\}^m \) let \( G = \{(A, b) \in \{0, 1\}^{m \times n} \times \{0, 1\}^m \} \) with \((A, b)(x) = A \times x + b\). (additions are over \( \mathbb{F}(2) \), e.g., \( \oplus \))

- 2-universal families: \( \Pr_{g \leftarrow G} [g(x) = g(x')] = \frac{1}{|\mathcal{R}|} \).

- Example for universal family that is not pairwise independent?

- Many-wise independent
Leftover hash lemma

Lemma 15 (leftover hash lemma)
Let $X$ be a rv over $\{0, 1\}^n$ with $H_2(X) \geq k$, let $G = \{g: \{0, 1\}^n \mapsto \{0, 1\}^m\}$ be 2-universal, and $G \leftarrow G$. Then $SD((G, G(X)), (G, \sim \{0, 1\}^m)) \leq \frac{1}{2} \cdot 2^{(m-k)/2}$.

Lemma 16
Let $p$ be dist. over $\mathcal{U}$ with $CP(p) \leq \frac{1+\delta}{|\mathcal{U}|}$, then $SD(p, \sim \mathcal{U}) \leq \frac{\sqrt{\delta}}{2}$.

Proof: Let $q$ be the uniform distribution over $\mathcal{U}$.

- $\|p - q\|^2_2 = \sum_{u \in \mathcal{U}} (p(u) - q(u))^2 = \|p\|^2_2 + \|q\|^2_2 - 2\langle p, q \rangle = CP(p) - \frac{1}{|\mathcal{U}|} \leq \frac{\delta}{|\mathcal{U}|}$
- Chebyshev Sum Inequality: $(\sum_{i=1}^n a_i)^2 \leq n \cdot \sum_{i=1}^n a_i^2$
- Hence, $\|p - q\|^2_1 \leq |\mathcal{U}| \cdot \|p - q\|^2_2$
- Thus, $SD(p, q) = \frac{1}{2} \|p - q\|_1 \leq \frac{\sqrt{\delta}}{2}$. □

To deuce the proof of Lemma 15, we notice that $CP(G, G(X)) \leq \frac{1}{|G|} \cdot (2^{-k} + 2^{-m}) = \frac{1+2^{m-k}}{|G| \cdot 2^m} = \frac{1+2^{m-k}}{|G \times \{0,1\}^m|}$. 