COMP2610 / COMP6261 Information Theory
Lecture 10: Typicality and Asymptotic Equipartition Property

Quanling Deng

Computational Mathematics Group
School of Computing
College of Engineering & Computer Science
The Australian National University
Canberra, Australia

Australian National University

Last time

Markov's inequality

Chebyshev's inequality

Law of large numbers

Law of Large Numbers

Theorem

Let $X_1, \ldots, X_n$ be a sequence of iid random variables, with

$$\mathbb{E}[X_i] = \mu$$

and $\mathbb{V}[X_i] < \infty$. Define

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}.$$

Then, for any $\beta > 0$,

$$\lim_{n \to \infty} p(|\bar{X}_n - \mu| < \beta) = 1.$$

This is also called $\bar{X}_n \to \mu$ **in probability**.

Definition: For random variables $v_1, v_2, \ldots$, we say $v_n \to v$ **in probability** if for all $\beta > 0$ $\lim_{n \to \infty} P(|v_n - v| > \beta) = 0$.

β is fixed (not shrinking like $\frac{1}{n}$). Not max/min. Reduction in variability.

This time

- Ensembles and sequences
- Typical sets
- Asymptotic Equipartition Property (AEP)

Ensembles

Ensemble

An **ensemble** X is a triple $(x, \mathcal{A}_X, \mathcal{P}_X)$; x is a **random variable** taking **values** in $\mathcal{A}_X = \{a_1, a_2, \ldots, a_I\}$ with **probabilities** $\mathcal{P}_X = \{p_1, p_2, \ldots, p_I\}$.

We will call $\mathcal{A}_X$ the alphabet of the ensemble

Ensembles

Example: Bent Coin

Let X be an ensemble with outcomes $\mathtt{h}$ for *heads* with probability 0.9 and $\mathtt{t}$ for *tails* with probability 0.1.

- The outcome set is $\mathcal{A}_X = \{\mathtt{h}, \mathtt{t}\}$
- The probabilities are $\mathcal{P}_X = \{p_{\mathtt{h}} = 0.9, p_{\mathtt{t}} = 0.1\}$

Extended Ensembles

We can also consider blocks of outcomes, which will be useful to describe sequences:

Example (Coin Flips):

$$
\begin{aligned}
\texttt{hhhhthhththh} &\rightarrow \texttt{hh hh th ht ht hh} && (6 \times 2 \text{ outcome blocks}) \\
&\rightarrow \texttt{hhh hth hth thh} && (4 \times 3 \text{ outcome blocks}) \\
&\rightarrow \texttt{hhhh thht hthh} && (3 \times 4 \text{ outcome blocks})
\end{aligned}
$$

Extended Ensemble

Let X be a single ensemble. The **extended ensemble** of blocks of size N is denoted X^N. Outcomes from X^N are denoted $\mathbf{x} = (x_1, x_2, \ldots, x_N)$. The **probability** of $\mathbf{x}$ is defined to be $P(\mathbf{x}) = P(x_1)P(x_2)\ldots P(x_N)$.

Extended Ensembles

Example: Bent Coin

Let X be an ensemble with outcomes $\mathcal{A}_X = \{\mathtt{h}, \mathtt{t}\}$ with $p_\mathtt{h} = 0.9$ and $p_\mathtt{t} = 0.1$.

Consider X^4 – i.e., 4 flips of the coin.

$$\mathcal{A}_{X^4} = \{\mathtt{hhhh}, \mathtt{hhht}, \mathtt{hhth}, \ldots, \mathtt{tttt}\}$$

$$P(\mathtt{hhhh}) = (0.9)^4 \approx 0.6561$$
$$P(\mathtt{tttt}) = (0.1)^4 = 0.0001$$
$$P(\mathtt{hthh}) = 0.9 \cdot 0.1 \cdot 0.9 \cdot 0.9 = (0.9)^3(0.1) \approx 0.0729$$
$$P(\mathtt{htht}) = 0.9 \cdot 0.1 \cdot 0.9 \cdot 0.1 = (0.9)^2(0.1)^2 \approx 0.0081.$$

Extended Ensembles

Example: Bent Coin

Entropy of extended ensembles

We can view X^4 as comprising 4 independent random variables, based on the ensemble X

Entropy is additive for independent random variables

Thus,

$$H(X^4) = 4H(X) = 4.\,(-0.9\log_2 0.9 - 0.1\log_2 0.1) = 1.88\text{bits}.$$

More generally,

$$H(X^N) = NH(X).$$

Counting Types of Sequences

Criteria for dividing 2^N sequences into **types**

In the bent coin example,

$$\begin{aligned}(0.9)^2(0.1)^2 &= P(\mathtt{hhtt}) \\ &= P(\mathtt{htht}) \\ &= P(\mathtt{htth}) \\ &= P(\mathtt{thht}) \\ &= P(\mathtt{thth}) \\ &= P(\mathtt{tthh}).\end{aligned}$$

The order of outcomes in the sequence is irrelevant

Counting Types of Sequences

Let X be an ensemble with alphabet $\mathcal{A}_X = \{a_1, \ldots, a_I\}$.

Let $p(X = a_i) = p_i$.

For a sequence $\mathbf{x} = x_1, x_2, \ldots, x_N$, how to compute $p(\mathbf{x})$?

let n_i = # of times symbol a_i appears in $\mathbf{x}$ (symbol count)

Given the n_i's, we can compute the probability of seeing $\mathbf{x}$:

$$\begin{aligned} P(\mathbf{x}) &= P(x_1) \cdot P(x_2) \cdot \ldots \cdot P(x_N) \\ &= P(a_1)^{n_1} \cdot P(a_2)^{n_2} \cdot \ldots \cdot P(a_I)^{n_I} \\ &= p_1^{n_1} \cdot p_2^{n_2} \ldots p_I^{n_I} \end{aligned}$$

Sufficient statistics: $\{n_1, n_2, \ldots, n_I\}$. Use it as a criteria of partitioning.

Counting Types of Sequences

Sequence Types

Each unique choice of $(n_1, n_2, \ldots, n_I)$ gives a different type of sequence

- 4 heads, (3 heads, 1 tail), (2 heads, 2 tails), ...
- Sequences in each type are equiprobable.

For a given type of sequence how many sequences are there with these symbol counts?

$$\text{\# of sequences with } n_i \text{ copies of } a_i = \frac{N!}{n_1!n_2!\ldots n_I!}$$

$$\binom{N}{n_1}\binom{N-n_1}{n_2}\binom{N-n_1-n_2}{n_3}\cdots$$
$$= \frac{N!}{n_1!(N-n_1)!}\cdot\frac{(N-n_1)!}{n_2!(N-n_1-n_2)!}\cdot\frac{(N-n_1-n_2)!}{n_3!(N-n_1-n_2-n_3)!}\cdots$$

Counting Types of Sequences

Example

Probability of types

Let $\mathcal{A} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ with $P(\mathtt{a}) = 0.2$, $P(\mathtt{b}) = 0.3$, $P(\mathtt{c}) = 0.5$.

Each sequence of type $(n_{\mathtt{a}}, n_{\mathtt{b}}, n_{\mathtt{c}}) = (2, 1, 3)$ has length 6 and probability $(0.2)^2(0.3)^1(0.5)^3 = 0.0015$.

There are $\frac{6!}{2!1!3!} = 60$ such sequences.

The probability $\mathbf{x}$ is of type $(2, 1, 3)$ is $(0.0015) \cdot 60 = 0.09$.

Study probabilities at the level of types (most likely, average/typical)

Extended Ensembles

Example

With $p_{\mathrm{h}} = 0.75$, what are the probabilities for X^N?

$N = 2$

$\mathbf{x}$	$P(\mathbf{x})$
hh	0.5625
ht	0.1875
th	0.1875
tt	0.0625

$N = 3$

$\mathbf{x}$	$P(\mathbf{x})$
hhh	0.4219
hht	0.1406
hth	0.1406
thh	0.1406
htt	0.0469
tht	0.0469
tth	0.0469
ttt	0.0156

$N = 4$

$\mathbf{x}$	$P(\mathbf{x})$	$\mathbf{x}$	$P(\mathbf{x})$
hhhh	0.3164	thht	0.0352
hhht	0.1055	thth	0.0352
hhth	0.1055	tthh	0.0352
hthh	0.1055	httt	0.0117
thhh	0.1055	thtt	0.0117
htht	0.0352	ttht	0.0117
htth	0.0352	ttth	0.0117
hhtt	0.0352	tttt	0.0039

Observations

As N increases, there is an increasing spread of probabilities

The most likely single sequence will always be the all $\mathtt{h}$'s

However, for $N = 4$, the most likely sequence type is 3 $\mathtt{h}$'s and 1 $\mathtt{t}$

Not surprising because $3 = N \cdot p_{\mathtt{h}}$, pretty much average case.

Symbol Frequency in Long Sequences

To judge if a sequence is typical/average, a natural question to ask is:

How often does each symbol appear in a sequence $\mathbf{x}$ from X^N?

Intuitively, in a sequence of length N, let a_i appear for n_i times.
Then **in expectation**

$$n_i \approx N \cdot P(a_i)$$

Note $p_i = P(a_i)$, and

$$P(\mathbf{x}) = P(a_1)^{n_1} P(a_2)^{n_2} \ldots P(a_I)^{n_I} \approx p_1^{Np_1} p_2^{Np_2} \ldots p_I^{Np_I}$$

So the *information content* $-\log_2 P(\mathbf{x})$ of that sequence is approximately

$$-p_1 N \log_2 p_1 - \ldots - p_I N \log_2 p_I = -N \sum_{i=1}^{I} p_i \log_2 p_i = NH(X)$$

Typical Sets

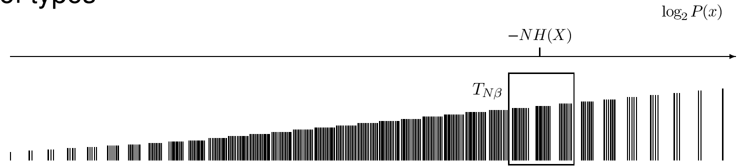
We want to consider elements $\mathbf{x}$ that have $-\log_2 P(\mathbf{x})$ "close" to $NH(X)$

Typical Set

For "closeness" $\beta > 0$ the typical set $T_{N\beta}$ for X^N is

$$T_{N\beta} \stackrel{\text{def}}{=} \{\mathbf{x} : |-\log_2 P(\mathbf{x}) - NH(X)| < N\beta\}$$
$$= \left\{\mathbf{x} : \left|-\frac{1}{N}\log_2 P(\mathbf{x}) - H(X)\right| < \beta\right\}$$

Union of types



What when $\beta = 0$ (and replace $<$ by $\leq$)?
Criterion based on information content. Other criterion (KL divergence)?

Typical Sets

The name "typical" is used since $\mathbf{x} \in T_{N\beta}$ will have roughly $p_1 N$ occurrences of symbol a_1, $p_2 N$ of a_2, . . ., $p_K N$ of a_K.

$\mathbf{x}$	$\log_2(P(\mathbf{x}))$
...1...................1.....1....1.1.......1........1...........1.....................1.......11...	−50.1
......................1.....1.....1.......1....1.........1.....................................1....	−37.3
........1....1..1...1....11..1.1.........11.........................1...1.1..1...1................1.	−65.9
1.1...1................1.......................11.1..1............................1.....1..1.11.....	−56.4
...11...........1...1.....1.1......1..........1....1...1.....1............1.........................	−53.2
..............1......1.........1.1.......1..........1............1...1......................1.......	−43.7
.....1........1.......1...1............1............1...........1......1..11........................	−46.8
.....1..1..1...............111...................1...............1.........1.1...1...1.............1	−56.4
.........1..........1.....1......1..........1....1..1...	−37.3
......1........................1..............1.....1..1.1.1..1...................................1.	−43.7
1.......................1..........1...1...................1....1....1........1..11..1.1...1........	−56.4
...........11.1.........1................1......1.....................1.............................	−37.3
.1..........1...1.1.............1.......11...........1.1...1..............1.............11..........	−56.4
......1...1..1.....1..11.1.1.1...1.....................1............1.............1..1..............	−59.5
............11.1......1....1..1............................1.......1..............1.......1.........	−46.8
....	−15.2
1111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111	−332.1

Randomly drawn sequences for $P(1) = 0.1$. Note: $H(X) \approx 0.47$

Typical Sets

Properties

Typical sequences are nearly equiprobable: Every $\mathbf{x} \in T_{N\beta}$ has

$$2^{-N(H(X)+\beta)} \leq P(\mathbf{x}) \leq 2^{-N(H(X)-\beta)}.$$

Variation is small when β is small

Number of sequences in the typical set: For any N, β,

$$|T_{N\beta}| \leq 2^{N(H(X)+\beta)}.$$

Typical Sets

Proof of Cardinality Bound

For every $\mathbf{x} \in T_{N\beta}$,

$$p(\mathbf{x}) \geq 2^{-N(H(X)-\beta)}.$$

Thus,

$$\begin{aligned} 1 &= \sum_{\mathbf{x}} p(\mathbf{x}) \\ &\geq \sum_{\mathbf{x} \in T_{N\beta}} p(\mathbf{x}) \\ &\geq \sum_{\mathbf{x} \in T_{N\beta}} 2^{-N(H(X)-\beta)} \\ &= 2^{-N(H(X)-\beta)} \cdot |T_{N\beta}|. \end{aligned}$$

Thus

$$|T_{N\beta}| \leq 2^{N(H(X)+\beta)}$$

Typical Sets

Most Likely Sequence

The most likely sequence may not belong to the typical set

e.g. with $p_{\mathrm{h}} = 0.75$, we have

$$-\frac{1}{4}\log_2 P(\mathtt{hhhh}) = 0.4150$$

whereas $H(X) = 0.8113$

The most likely single sequence $\rightarrow \mathtt{hhhh}$

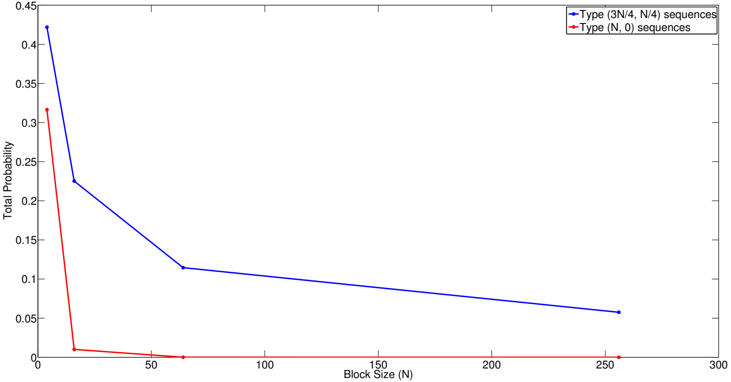
The most likely single sequence type $\rightarrow \{\mathtt{hhht}, \mathtt{hthh}, \ldots\}$

Typical Sets

Most Likely Sequence

Probability of most likely sequence decays like $(p_{\mathrm{h}})^N$ ($p_{\mathrm{h}} = 0.75$)

Sequences with $N \cdot p_{\mathrm{h}}$ heads contain much more total probability mass



Blue curve corresponds to typical set with $\beta = 0$. What if $\beta > 0$?

Asymptotic Equipartition Property

$\underbrace{\text{Asymptotic}}_{\text{Eventually}}$ $\underbrace{\text{Equipartition}}_{\text{Equally Divided}}$ Property

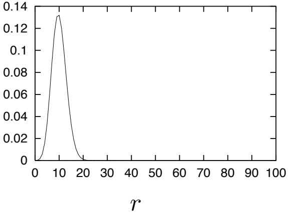
Informally

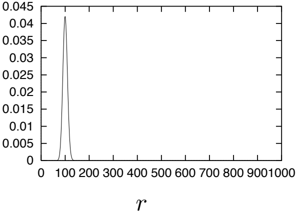
Asymptotic Equipartition Property (Informal)

As $N \to \infty$, $\log_2 P(x_1, \ldots, x_N)$ is close to $-NH(X)$ with high probability.

For large block sizes "almost all sequences are typical" (i.e., in $T_{N\beta}$)

$$n(r)P(\mathbf{x}) = \binom{N}{r} p_1^r (1 - p_1)^{N-r}$$





Probability sequence $\mathbf{x}$ has r heads for $N = 100$ (left) and $N = 1000$ (right). Here $P(X = \text{head}) = 0.1$.

Asymptotic Equipartition Property

If $x_1, x_2, \ldots$ are i.i.d. with distribution P then, **in probability**,

$$-\frac{1}{N}\log_2 P(x_1, \ldots, x_N) \to H(X).$$

In precise language:

$$(\forall \beta > 0) \lim_{N\to\infty} p\left(\left|-\frac{1}{N}\log_2 P(x_1, \ldots, x_N) - H(X)\right| < \beta\right) = 1.$$

Exactly the probability of $\mathbf{x} \in T_{N\beta}$.

Recall definition: for random variables $v_1, v_2, \ldots$, we say $v_N \to v$ **in probability** if for all $\beta > 0$ $\lim_{N\to\infty} P(|v_N - v| > \beta) = 0$

Here v_N corresponds to $-\frac{1}{N}\log_2 P(x_1, \ldots, x_N)$.

Asymptotic Equipartition Property

Comments

Why is it surprising/significant?

For an ensemble with binary outcomes, and low entropy,

$$|T_{N\beta}| \leq 2^{NH(X)+\beta} \ll 2^N$$

i.e. the typical set is a small fraction of all possible sequences

AEP says that for N sufficiently large, we are virtually guaranteed to draw a sequence from this small set

Significance in information theory

Asymptotic Equipartition Property

Proof

Since $x_1, \ldots, x_N$ are independent,

$$-\frac{1}{N}\log p(x_1, \ldots, x_N) = -\frac{1}{n}\log\prod_{n=1}^{N} p(x_n)$$
$$= -\frac{1}{N}\sum_{n=1}^{N}\log p(x_n).$$

Let $Y = -\log p(X)$ and $y_n = -\log p(x_n)$. Then, $y_n \sim Y$, and

$$\mathbb{E}[Y] = H(X).$$

But then by the law of large numbers,

$$(\forall\beta > 0)\ \lim_{N\to\infty} p\left(\left|\frac{1}{N}\sum_{n=1}^{N} y_n - H(X)\right| > \beta\right) = 0.$$

Summary & Conclusions

- Ensembles and sequences
- Typical sets
- Asymptotic Equipartition Property (AEP)

Next: Source Coding.

Acknowledgement

These slides were originally developed by Professor Robert C. Williamson.