Lecture 6

Information Theory

September 18, 2013

Asymptotic Equipartition Principle AEP

- Law of Large Numbers
  - Sample mean is close to expected value
- Asymptotic Equipartition Principle (AEP)
  - $-\log P(x_1, x_2, \ldots, x_n)/n$ is close to entropy H
- The Typical Set
  - Probability of each sequence close to 2^{-nH}
  - Size ($\sim 2^{nH}$) and total probability (~ 1)
- The Atypical Set
  - Unimportant and could be ignored

Typicality: Example

X = {a, b, c, d}, $\mathbf{p} = [0.5\ 0.25\ 0.125\ 0.125]$

$-\log \mathbf{p} = [1\ 2\ 3\ 3] \Rightarrow H(\mathbf{p}) = 1.75$ bits

Sample eight i.i.d. values

- typical $\Rightarrow$ correct proportions

  adbabaac $-\log p(\mathbf{x}) = 14 = 8 \times 1.75 = nH(\mathbf{x})$

- not typical $\Rightarrow \log p(\mathbf{x}) \neq nH(\mathbf{x})$

  dddddddd $-\log p(\mathbf{x}) = 24$

Convergence of Random Variables

- Convergence

$$X_n \underset{n\to\infty}{\to} Y \quad \Rightarrow \quad \forall \varepsilon > 0, \exists m \text{ such that } \forall n > m, |X_n - Y| < \varepsilon$$

Example: $X_n = \pm 2^{-n}, \quad Y = 0$

choose $m = -\log \varepsilon$

- Convergence in probability (weaker than convergence)

$$X_n \overset{\text{prob}}{\to} Y \quad \Rightarrow \quad \forall \varepsilon > 0, \quad P\left(|X_n - Y| > \varepsilon\right) \to 0$$

Example: $x_n \in \{0;1\}, \quad p = [1 - n^{-1}; n^{-1}]$

for any small ε, $p(|x_n| > \varepsilon) = n^{-1} \xrightarrow{n\to\infty} 0$

so $x_n \overset{\text{prob}}{\to} 0$ (but $x_n \not\to 0$)

Note: y can be a constant or another random variable

Law of Large Numbers

Given i.i.d. $\{X_i\}$, sample mean $s_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

- $E\, s_n = E\, X = \mu \qquad \mathrm{Var}\, s_n = n^{-1}\,\mathrm{Var}\, X = n^{-1}\sigma^2$

As n increases, $\mathrm{Var}\, s_n$ gets smaller and the values become clustered around the mean

LLN: $s_n \xrightarrow{\text{prob}} \mu$

$$\Leftrightarrow \quad \forall \varepsilon > 0, \quad P\left(|\, s_n - \mu \,| > \varepsilon\right) \underset{n\to\infty}{\to} 0$$

The expected value of a random variable is equal to the long-term average when sampling repeatedly.

Asymptotic Equipartition Principle

- **x** is the i.i.d. sequence $\{X_i\}$ for $1 \le i \le n$
  - Prob of a particular sequence is $p(\mathbf{x}) = \prod_{i=1}^{n} p(X_i)$
  - Average $E - \log p(\mathbf{x}) = n\, E - \log p(x_i) = nH(X)$
- AEP:

$$-\frac{1}{n}\log p(\mathbf{x}) \xrightarrow{\text{prob}} H(X)$$

- Proof:

$$-\frac{1}{n}\log p(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i)$$

law of large numbers

$$\xrightarrow{\text{prob}} \quad E - \log p(x_i) = H(X)$$

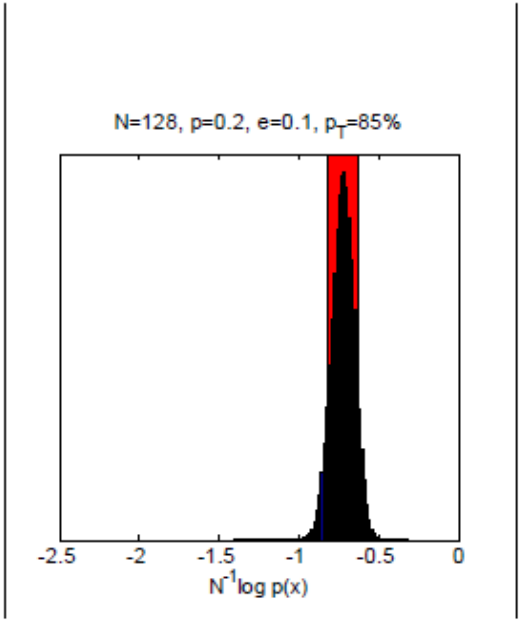
Typical Set

Typical set (for finite n)

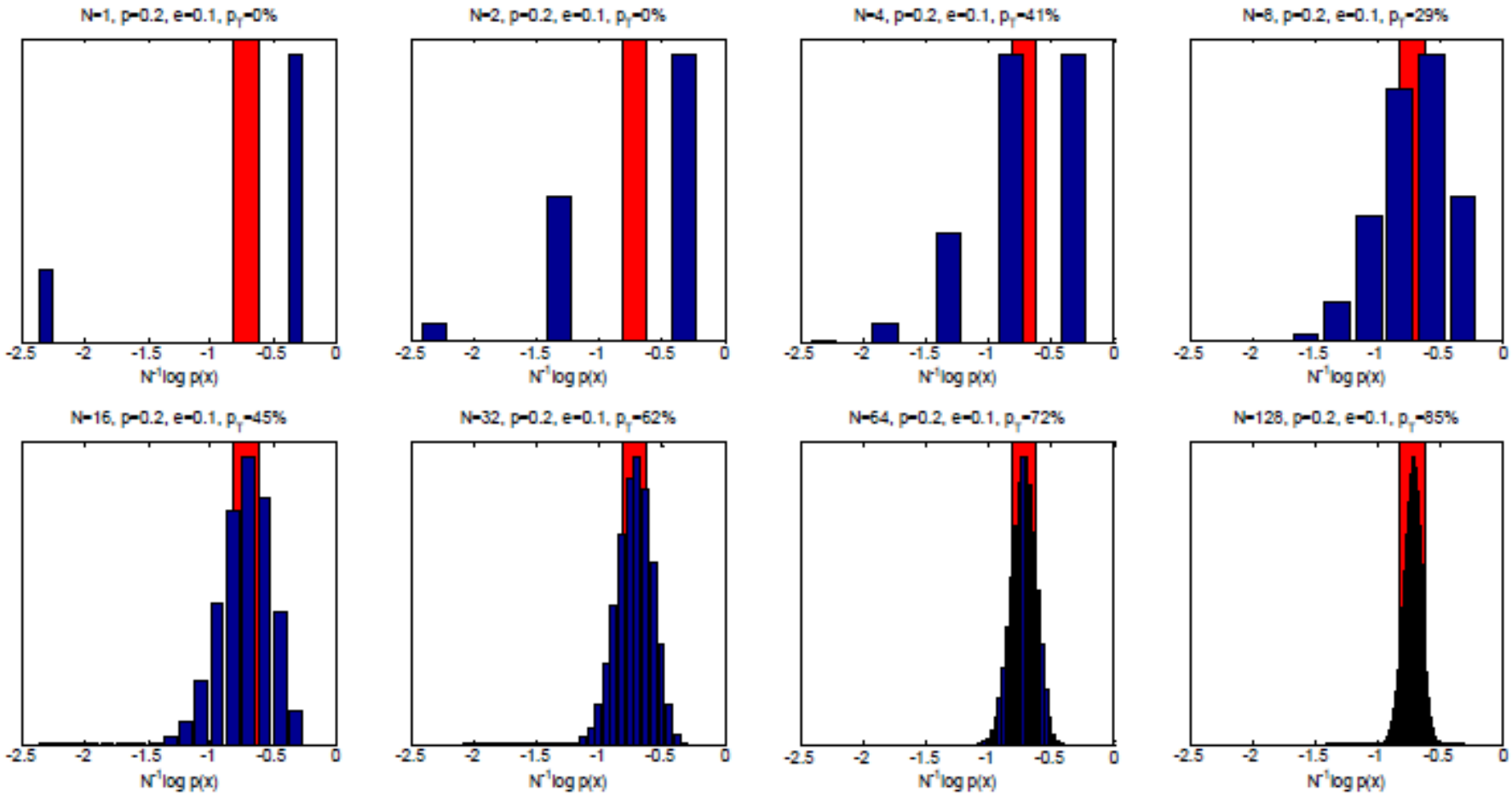
$$T_\varepsilon^{(n)} = \left\{ \mathbf{x} \in \mathrm{X}^n : \left| -n^{-1} \log p(\mathbf{x}) - H(X) \right| < \varepsilon \right\}$$

Example:

- x_i Bernoulli with $p(x_i = 1) = p$
- *e.g.* $p([0\ 1\ 1\ 0\ 0\ 0]) = p^2(1-p)^4$
- For $p = 0.2$, $H(X) = 0.72$ bits
- Red bar shows $T_{0.1}^{(n)}$



Typical Set Frames



Typical Set Properties

1. Individual prob: $\mathbf{x} \in T_\varepsilon^{(n)} \Rightarrow \log p(\mathbf{x}) = -nH(X) \pm n\varepsilon$
2. Total prob: $p(\mathbf{x} \in T_\varepsilon^{(n)}) > 1-\varepsilon \text{ for } n > N_\varepsilon$
3. Size: $(1-\varepsilon)2^{n(H(X)-\varepsilon)} \overset{n>N_\varepsilon}{<} \left|T_\varepsilon^{(n)}\right| \le 2^{n(H(X)+\varepsilon)}$

Proof 2: $-n^{-1}\log p(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} -\log p(x_i) \overset{\text{prob}}{\to} E -\log p(x_i) = H(X)$

Hence $\forall \varepsilon > 0\ \exists N_\varepsilon \text{ s.t. } \forall n > N_\varepsilon \quad p(\left|-n^{-1}\log p(\mathbf{x}) - H(X)\right| > \varepsilon) < \varepsilon$

Proof 3a: $\text{f.l.e.}\, n, \quad 1-\varepsilon < p(\mathbf{x} \in T_\varepsilon^{(n)}) \le \sum_{\mathbf{x}\in T_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)} = 2^{-n(H(X)-\varepsilon)}\left|T_\varepsilon^{(n)}\right|$

Proof 3b: $1 = \sum_{\mathbf{x}} p(\mathbf{x}) \ge \sum_{\mathbf{x}\in T_\varepsilon^{(n)}} p(\mathbf{x}) \ge \sum_{\mathbf{x}\in T_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)} = 2^{-n(H(X)+\varepsilon)}\left|T_\varepsilon^{(n)}\right|$

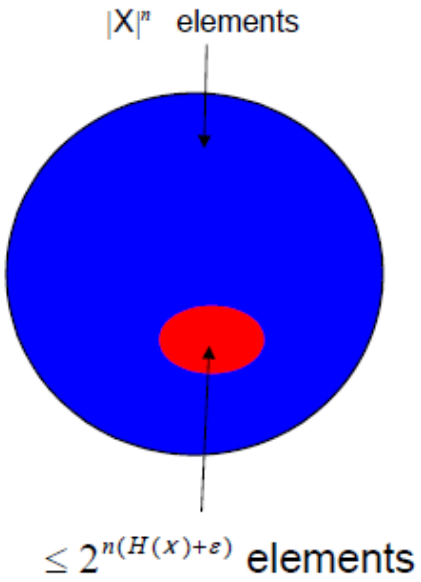
Consequence of AEP

- for any ε and for $n > N_\varepsilon$
  "Almost all events are almost equally surprising"
- $p(\mathbf{x} \in T_\varepsilon^{(n)}) > 1-\varepsilon$ and $\log p(\mathbf{x}) = -nH(X) \pm n\varepsilon$

Coding consequence

- $\mathbf{x} \in T_\varepsilon^{(n)}$: '0' + at most $1+n(H+\varepsilon)$ bits
- $\mathbf{x} \notin T_\varepsilon^{(n)}$: '1' + at most $1+n\log|\mathrm{X}|$ bits
- L = Average code length

$$\le p(\mathbf{x} \in T_\varepsilon^{(n)})[2+n(H+\varepsilon)]$$
$$+p(\mathbf{x} \notin T_\varepsilon^{(n)})[2+n\log|\mathrm{X}|]$$
$$\le n(H+\varepsilon)+\varepsilon\left(n\log|\mathrm{X}|\right)+2\varepsilon+2$$
$$= n\left(H+\varepsilon+\varepsilon\log|\mathrm{X}|+2(\varepsilon+2)n^{-1}\right) = n(H+\varepsilon')$$



Data Compression and Source Coding

For any choice of $\varepsilon > 0$, we can, by choosing block size, n, large enough, do the following:

- make a lossless code using only $H(X)+\varepsilon$ bits per symbol on average:

$$\frac{L}{n} \leq H + \varepsilon$$

- The coding is one-to-one and decodable
  - However impractical due to exponential complexity
- Typical sequences have short descriptions of length $\approx nH$
  - Another proof of source coding theorem (Shannon's original proof)
- However, encoding/decoding complexity is exponential in n

Smallest High Probability Set

$T_\varepsilon^{(n)}$ is a small subset of X^n containing most of the probability mass. Can you get even smaller ?

For any $0 < \varepsilon < 1$, choose $N_0 = -\varepsilon^{-1}\log \varepsilon$, then for any $n > \max(N_0, N_\varepsilon)$ and any subset $S^{(n)}$ satisfying $\left|S^{(n)}\right| < 2^{n(H(X)-2\varepsilon)}$

$$
\begin{aligned}
p\left(\mathbf{x} \in S^{(n)}\right) &= p\left(\mathbf{x} \in S^{(n)} \cap T_\varepsilon^{(n)}\right) + p\left(\mathbf{x} \in S^{(n)} \cap \overline{T_\varepsilon^{(n)}}\right) \\
&< \left|S^{(n)}\right| \max_{\mathbf{x} \in T_\varepsilon^{(n)}} p(\mathbf{x}) + p\left(\mathbf{x} \in \overline{T_\varepsilon^{(n)}}\right) \\
&< 2^{n(H-2\varepsilon)} 2^{-n(H-\varepsilon)} + \varepsilon \qquad \text{for } n > N_\varepsilon \\
&= 2^{-n\varepsilon} + \varepsilon < 2\varepsilon \qquad \text{for } n > N_0, \quad 2^{-n\varepsilon} < 2^{\log \varepsilon} = \varepsilon
\end{aligned}
$$

Summary

- Typical Set
  - Individual Prob $\mathbf{x} \in T_\varepsilon^{(n)} \Rightarrow \log p(\mathbf{x}) = -nH(x) \pm n\varepsilon$
  - Total Prob $p(\mathbf{x} \in T_\varepsilon^{(n)}) > 1-\varepsilon \text{ for } n > N_\varepsilon$
  - Size $(1-\varepsilon)2^{n(H(x)-\varepsilon)} \overset{n>N_\varepsilon}{<} \left|T_\varepsilon^{(n)}\right| \leq 2^{n(H(x)+\varepsilon)}$
- No other high probability set can be much smaller than $T_\varepsilon^{(n)}$
- Asymptotic Equipartition Principle
  - Almost all event sequences are equally surprising
- Can be used to prove source coding theorem