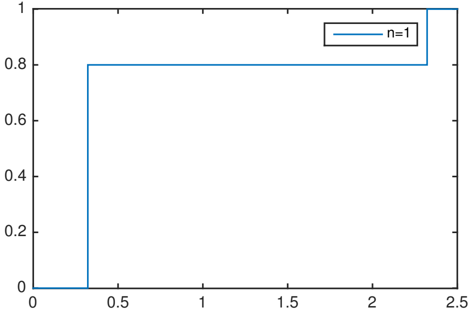
EEE 551 Information Theory (Spring 2022)

Chapter 3: Asymptotic Equipartition Property (AEP)

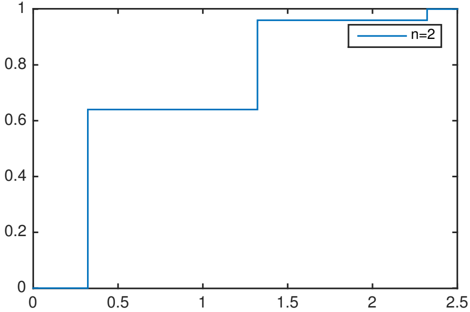
AEP Roughly

- Consider an **independent and identically distributed** (i.i.d.) sequence of random variables
$$(X_1, X_2, \ldots, X_{n-1}, X_n) = X^n$$
also called a **discrete memoryless source** (DMS)
- How does $p(X^n)$ behave for large n?
- For example, let $X_i \sim \text{Bern}(0.2)$
- Plot the CDF of $-\frac{1}{n}\log p(X^n)$:



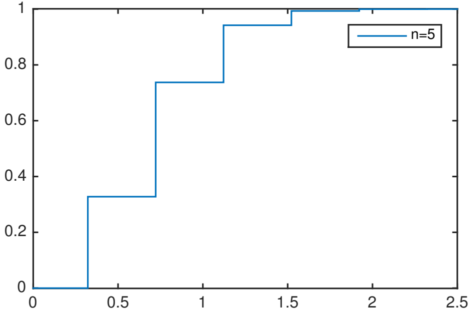
AEP Roughly

- Consider an **independent and identically distributed** (i.i.d.) sequence of random variables
$$(X_1, X_2, \ldots, X_{n-1}, X_n) = X^n$$
also called a **discrete memoryless source** (DMS)
- How does $p(X^n)$ behave for large n?
- For example, let $X_i \sim \text{Bern}(0.2)$
- Plot the CDF of $-\dfrac{1}{n}\log p(X^n)$:



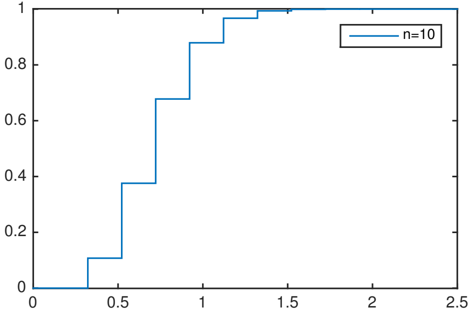
AEP Roughly

- Consider an **independent and identically distributed** (i.i.d.) sequence of random variables
$$(X_1, X_2, \dots, X_{n-1}, X_n) = X^n$$
also called a **discrete memoryless source** (DMS)
- How does $p(X^n)$ behave for large n?
- For example, let $X_i \sim \text{Bern}(0.2)$
- Plot the CDF of $-\frac{1}{n} \log p(X^n)$:



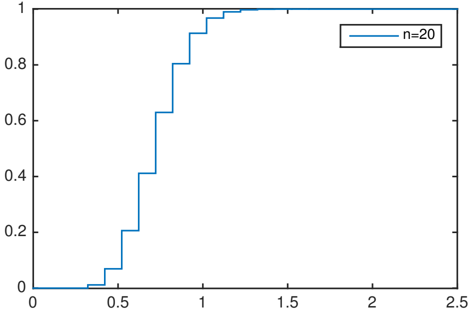
AEP Roughly

- Consider an **independent and identically distributed** (i.i.d.) sequence of random variables
$$(X_1, X_2, \ldots, X_{n-1}, X_n) = X^n$$
also called a **discrete memoryless source** (DMS)
- How does $p(X^n)$ behave for large n?
- For example, let $X_i \sim \text{Bern}(0.2)$
- Plot the CDF of $-\frac{1}{n}\log p(X^n)$:



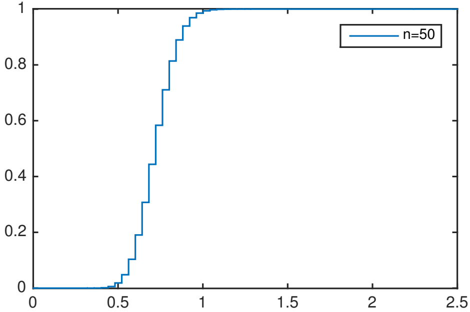
AEP Roughly

- Consider an **independent and identically distributed** (i.i.d.) sequence of random variables
$$(X_1, X_2, \ldots, X_{n-1}, X_n) = X^n$$
also called a **discrete memoryless source** (DMS)
- How does $p(X^n)$ behave for large n?
- For example, let $X_i \sim \text{Bern}(0.2)$
- Plot the CDF of $-\frac{1}{n} \log p(X^n)$:



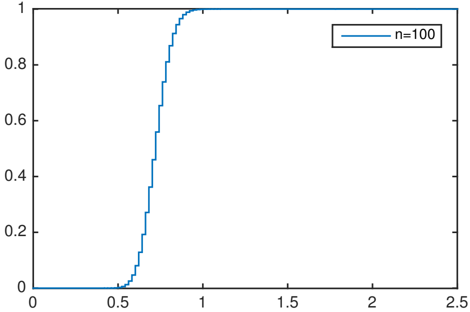
AEP Roughly

- Consider an **independent and identically distributed** (i.i.d.) sequence of random variables
$$(X_1, X_2, \ldots, X_{n-1}, X_n) = X^n$$
also called a **discrete memoryless source** (DMS)
- How does $p(X^n)$ behave for large n?
- For example, let $X_i \sim \text{Bern}(0.2)$
- Plot the CDF of $-\frac{1}{n}\log p(X^n)$:



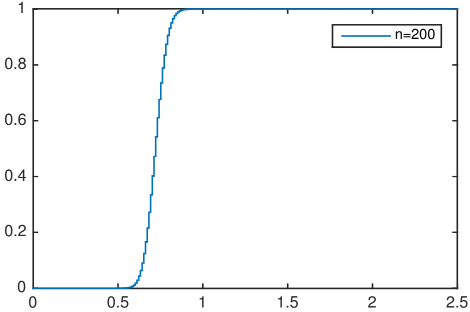
AEP Roughly

- Consider an **independent and identically distributed** (i.i.d.) sequence of random variables
$$(X_1, X_2, \ldots, X_{n-1}, X_n) = X^n$$
also called a **discrete memoryless source** (DMS)
- How does $p(X^n)$ behave for large n?
- For example, let $X_i \sim \text{Bern}(0.2)$
- Plot the CDF of $-\dfrac{1}{n} \log p(X^n)$:



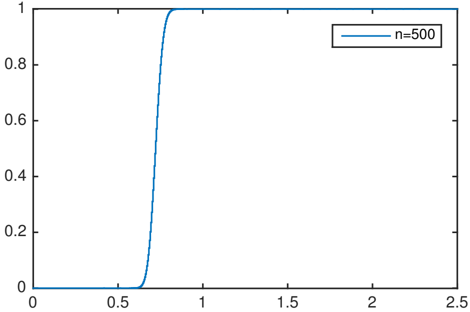
AEP Roughly

- Consider an **independent and identically distributed** (i.i.d.) sequence of random variables
$$(X_1, X_2, \ldots, X_{n-1}, X_n) = X^n$$
also called a **discrete memoryless source** (DMS)
- How does $p(X^n)$ behave for large n?
- For example, let $X_i \sim$ Bern(0.2)
- Plot the CDF of $-\frac{1}{n}\log p(X^n)$:



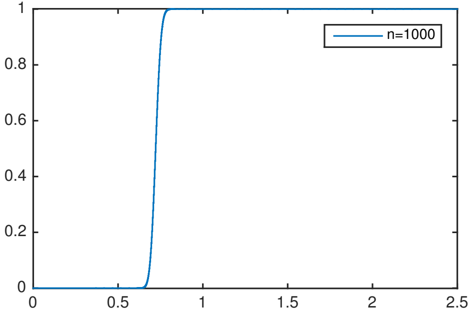
AEP Roughly

- Consider an **independent and identically distributed** (i.i.d.) sequence of random variables
$$(X_1, X_2, \ldots, X_{n-1}, X_n) = X^n$$
also called a **discrete memoryless source** (DMS)
- How does $p(X^n)$ behave for large n?
- For example, let $X_i \sim \text{Bern}(0.2)$
- Plot the CDF of $-\frac{1}{n} \log p(X^n)$:



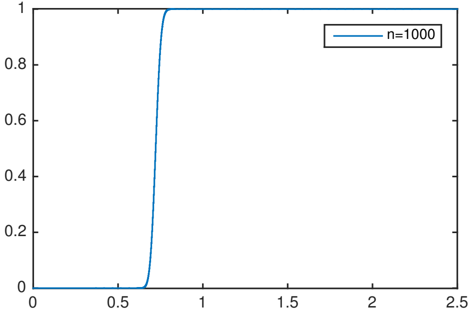
AEP Roughly

- Consider an **independent and identically distributed** (i.i.d.) sequence of random variables
$$(X_1, X_2, \ldots, X_{n-1}, X_n) = X^n$$
also called a **discrete memoryless source** (DMS)
- How does $p(X^n)$ behave for large n?
- For example, let $X_i \sim \text{Bern}(0.2)$
- Plot the CDF of $-\frac{1}{n} \log p(X^n)$:



AEP Roughly

- Consider an **independent and identically distributed** (i.i.d.) sequence of random variables
$$(X_1, X_2, \dots, X_{n-1}, X_n) = X^n$$
also called a **discrete memoryless source** (DMS)
- How does $p(X^n)$ behave for large n?
- For example, let $X_i \sim \text{Bern}(0.2)$
- Plot the CDF of $-\frac{1}{n}\log p(X^n)$:



- Concentrates around $H(X) = 0.7219$!

AEP Roughly, ctd

- Let $X^n \overset{\text{iid}}{\sim} \text{Bern}(p)$
- Given $x^n = (x_1, x_2, \ldots, x_n)$,

$$p(x^n) = \prod_{i=1}^{n} p(x_i) = \left(\prod_{i:x_i=1} p \right) \left(\prod_{i:x_i=0} (1-p) \right) = p^{\sum_i x_i} (1-p)^{\sum_i (1-x_i)}$$

- By the law of large numbers, $\sum_i X_i \approx np$ and $\sum_i (1 - X_i) \approx n(1-p)$

$$\begin{aligned} p(X^n) &\approx p^{np}(1-p)^{n(1-p)} \\ &= \left(2^{\log p}\right)^{np} \left(2^{\log(1-p)}\right)^{n(1-p)} \\ &= 2^{np \log p} 2^{n(1-p)\log(1-p)} \\ &= 2^{n(p \log p + (1-p)\log(1-p))} \\ &= 2^{-nH(X)} \end{aligned}$$

i.e. $-\frac{1}{n} \log p(X^n) \approx H(X)$

- "Almost all events are almost equally likely"

Weak Law of Large Numbers

A sequence of random variables $Z_1, Z_2, Z_3, \ldots$ converges to c **in probability** if, for all $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\{|Z_n - c| > \epsilon\} = 0.$$

Theorem (Weak Law of Large Numbers)

Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. random variables with expectation μ, and finite variance. Then $\frac{1}{n}\sum_{i=1}^{n} Y_i \to \mu$ in probability.

The proof is based on two **concentration inequalities:**

1. Markov's inequality: If $P(X < 0) = 0$, then for any $a > 0$, $P(X > a) \leq \dfrac{\mathbb{E}(X)}{a}$

Proof:

- Let $I(x > a) = \begin{cases} 1 & x > a \\ 0 & x \leq a \end{cases}$
- $I(x > a) \leq \dfrac{x}{a}$ for all $x \geq 0$
- $P(X > a) = \mathbb{E}\big(I(X > a)\big) \leq \mathbb{E}\left(\dfrac{X}{a}\right) = \dfrac{\mathbb{E}(X)}{a}$

Weak Law of Large Numbers, ctd

2. Chebyshev's inequality: For any random variable X, for any $a > 0$,

$$P(|X - \mathbb{E}(X)| > a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

Proof: Apply Markov's inequality to $Y = (X - \mathbb{E}(X))^2$:

$$P(|X - \mathbb{E}(X)| > a) = P\big((X - \mathbb{E}(X))^2 > a^2\big) = P(Y > a^2)$$
$$\leq \frac{\mathbb{E}(Y)}{a^2} = \frac{\mathrm{Var}(X)}{a^2}$$

Proof of weak law of large numbers:

- Let $Z_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$
- $\mathbb{E}(Z_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(Y_i) = \mu$
- $\mathrm{Var}(Z_n) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(Y_i) = \frac{\mathrm{Var}(Y_1)}{n}$
- Apply Chebyshev's inequality:

  $$P(|Z_n - \mu| > \epsilon) \leq \frac{\mathrm{Var}(Z_n)}{\epsilon^2} = \frac{\mathrm{Var}(Y_1)}{\epsilon^2 n} \to 0$$

  Thus $Z_n \to \mu$ in probability

AEP Theorem

If $X_1, X_2, \ldots, X_n$ are drawn i.i.d. from $p(x)$, then

$$-\frac{1}{n} \log p(X^n) \to H(X) \text{ in probability}$$

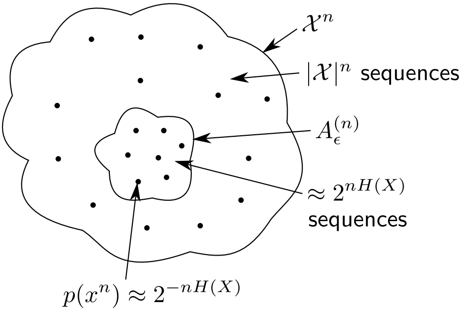
Proof:

$$\begin{aligned}
-\frac{1}{n} \log p(X^n) &= -\frac{1}{n} \log \prod_{i=1}^{n} p(X_i) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i) \\
&\to -\mathbb{E} \log p(X) && \text{by WLLN} \\
&= H(X)
\end{aligned}$$

The Typical Set

Given a distribution $p(x)$, the **typical set** $A_\epsilon^{(n)}$ is the set of $(x_1, x_2, \ldots, x_n) = x^n \in \mathcal{X}^n$ such that

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



Typical Set Properties

- $x^n \in A_\epsilon^{(n)}$ iff $\left| -\frac{1}{n} \log p(x^n) - H(X) \right| \leq \epsilon$
- $\lim_{n\to\infty} \Pr\{A_\epsilon^{(n)}\} = 1$
- $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$
- For n sufficiently large, $|A_\epsilon^{(n)}| \geq (1-\epsilon) 2^{n(H(X)-\epsilon)}$

Proofs of Typical Set Properties

- $x^n \in A_\epsilon^{(n)}$ iff $\left| -\frac{1}{n} \log p(x^n) - H(X) \right| \leq \epsilon$

  Proof: From the definition of the typical set:

$$2^{-n(H(X)+\epsilon)} \leq p(x^n) \leq 2^{-n(H(X)-\epsilon)}$$
$$-n(H(X)+\epsilon) \leq \log p(x^n) \leq -n(H(X)-\epsilon)$$
$$H(X)+\epsilon \geq -\frac{1}{n} \log p(x^n) \geq H(X)-\epsilon$$

- $\lim_{n\to\infty} \Pr\{A_\epsilon^{(n)}\} = 1$

  Proof: From the AEP theorem,

$$\Pr\left\{ \left| -\frac{1}{n} \log p(X^n) - H(X) \right| > \epsilon \right\} \to 0.$$

  Thus $\Pr\{A_\epsilon^{(n)}\} \to 1$ as $n \to \infty$

- $|A_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}$

  Proof:
$$
\begin{aligned}
1 &\ge \Pr\{A_\epsilon^{(n)}\} \\
&= \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \\
&\ge \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \\
&= |A_\epsilon^{(n)}| 2^{-n(H(X)+\epsilon)}
\end{aligned}
$$

- For n sufficiently large, $|A_\epsilon^{(n)}| \ge (1-\epsilon) 2^{n(H(X)-\epsilon)}$

  Proof: Since $\Pr\{A_\epsilon^{(n)}\} \to 1$, for sufficiently large n
$$
\begin{aligned}
1-\epsilon &\le \Pr\{A_\epsilon^{(n)}\} \\
&= \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \\
&\le \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} \\
&= |A_\epsilon^{(n)}| 2^{-n(H(X)-\epsilon)}
\end{aligned}
$$