1. Definition: Bernoulli Distribution	2. Definition: Binary Entropy Function
3. Definition: Convergence In Probability	4. Definition: Entropy
5. Remark: Properties Of Entropy	6. Theorem: Convergence To Common Entropy In Probability
7. Corollary: Aep	8. Definition: Source.Memoryless
9. Definition: Fixed Rate Code	10. Definition: Code Rate
11. Definition: Code Error Probability	12. Theorem: Fixed Rate Coding Theorem
13. Definition: Hypothesis Test	14. Definition: Hypothesis Test Error Probability

15. Definition: Relative Entropy	16. Theorem: Steins Lemma
17. Theorem: Neyman Pearson Lemma	18. Proposition: Informatation Theoretic Form Of Neyman Pearson Decision Region
19. Theorem: Jensens Inequality	20. Theorem: Log Sum Inequality
21. Proposition: Bounds On Entropy And Relative Entropy	22. Definition: Joint Entropy
23. Definition: Conditional Entropy	24. Proposition: Joint Entropy Is Single Entropy Plus Conditional Entropy
25. Proposition: Entropy Chain Rule	26. Proposition: Conditioning Reduces Entropy

27. Definition: Conditional Independence	28. Note: X Y And Function Of Y Form Markov Chain
29. Corollary: Subadditivity Of Entropy	30. Remark: Expression For Conditional Entropy
31. Proposition: Entropy Data Processing	32. Proposition: Properties Of Conditional Entropy
33. Theorem: Fano Inequality	34. Theorem: Relative Entropy Data Processing
35. Definition: Total Variation Distance	36. Theorem: Pinskers Inequality
37. Theorem: Relative Entropy Is Convex	38. Corollary: Entropy Is Concave

39. Theorem: Binomial Converges To Poisson	40. Theorem: Relative Entropy Between Sum Of Bernoullis And Poisson Is Bounded
41. Corollary: Relative Entropy Between Sum Of Independent Bernoullis And Poisson Is Bounded	42. Corollary: Binomial Convergence To Poisson Follows From Relative Entropy Bound
43. Lemma: Binomial Maximum Entropy	44. Theorem: Poisson Maximum Entropy
45. Definition: Mutual Information	46. Proposition: Expressions For Mutual Information

47. Proposition: Bounds On Mutual Information And Conditional Mutual Information	48. Proposition: Mutual Information Chain Rule
49. Theorem: Mutual Information Data Processing	50. Definition: Synergy
51. Definition: Redundancy Orthogonality Synergisticy	52. Theorem: Condition For Correlated Observations Giving More Information Than Independent Ones
53. Definition: Sumset	54. Definition: Difference Set
55. Proposition: Sumset Bounds	56. Proposition: Ruzsa Triangle Inequality

57. Proposition: Sum Entropy Bounds	58. Lemma: Entropy Ruzsa Triangle Inequality Lemma
59. Theorem: Entropy Ruzsa Triangle Inequality	60. Definition: Entropy Increase
61. Proposition: Mutual Information Expression For Entropy Increase	62. Lemma: Triple Sum Reduces Entropy
63. Theorem: Doubling Difference Inequality	64. Definition: Entropy Rate
65. Definition: Source.Stationary	66. Lemma: Cesaro

67. Theorem: Entropy Rate Of Stationary Source	68. Theorem: Entropy Rate Is Conditional Entropy Given Infinite Past
69. Definition: Source.Ergodic	70. Theorem: Birkhoff
71. Theorem: Shannon Mcmillan Breiman	72. Definition: Type
73. Definition: N Types	74. Proposition: Upper Bound On Number Of N Types
75. Proposition: Prob Under Prod Dist Of String Of Type	76. Definition: Type Class

77. Lemma: Most Likely Type Class Under Prod Dist Is Type Class Of That Dist	78. Proposition: Bounds On Size Of Type Class
79. Corollary: Bounds On Probability Of Type Class	80. Theorem: Sanov
81. Definition: Log Mgf	82. Proposition: Chernoff Bound
83. Corollary: Minimising Distribution In Sanov Is Unique For Nonempty Closed Convex Events	84. Theorem: Pythagorean Identity
85. Lemma: Expectation Of Bounded Rvs Converges To Limit Of Convergence In Probability	86. Theorem: Gibbs Conditioning Principle

87. Theorem: Error Exponents For Fixed Rate Compression	88. Definition: Variable Rate Code
89. Definition: Variable Rate Code.Prefix Free	90. Theorem: Krafts Inequality
91. Theorem: Codes Distributions Correspondence	92. Theorem: Entropic Bounds On Expected Length Of Prefix Free Codes
93. Corollary: Entropy Rate Of Source Is Best Asymptotic Prefix Free Compression Rate	94. Definition: Shannon Code
95. Definition: Ideal Shannon Codelength	96. Theorem: Competitive Optimality Of Shannon Codes

97. Definition: Mle Code	98. Proposition: Mle Code Price Of Universality
99. Proposition: Mle Code Expected Price Of Universality	100. Definition: Counting Code
101. Proposition: Counting Code Description Length	102. Definition: Uniform Mixture
103. Definition: Mixture Code	104. Lemma: Closed Form Expression For Uniform Mixture Of Bernoullis
105. Proposition: Mixture Code Description Length	106. Definition: Predictive Code

107. Proposition: Predictive Code Description Length	108. Lemma: Exponential Upper Bound On Binomial Coefficient
109. Definition: Fisher Information	110. Proposition: Upper Bound On Price Of Universality Of Counting Mixture And Predictive Codes
111. Notation: Mdl Estimator	112. Definition: Mdl Code
113. Proposition: Mdl Code Description Length	114. Definition: Redundancy
115. Definition: Worst Case Maximal Redundancy	116. Definition: Minimax Maximal Redundancy

117. Definition: Worst Case Average Redundancy	118. Definition: Minimax Average Redundancy
119. Theorem: Normalised Maximum Likelihood Code	120. Definition: Gamma Function
121. Theorem: Shtarkov	122. Definition: Channel Capacity
123. Theorem: Redundancy Capacity Theorem	124. Definition: Standard Parameterisation Of Pmfs On Alphabet
125. Theorem: Rissanen	126. Corollary: Bounds On Minimax Maximal And Average Redundancies

1. Entropy

1.1. Introduction

Notation 1.1 Write $x_1^n := (x_1, ..., x_n) \in \{0,1\}^n$ for an length n bit string.

Notation 1.2 We use P to denote a probability mass function. Write P_1^n for the joint proability mass function of a sequence of n random variables $X_1^n = (X_1, ..., X_n)$.

Definition: Bernoulli Distribution

Definition 1.3 A random variable X has a **Bernoulli distribution**, $X \sim \text{Bern}(p)$, if for some fixed $p \in (0,1)$,

$$X = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1-p \end{cases}$$

i.e. the probability mass function (PMF) of X is $P : \{0,1\} \to \mathbb{R}$, $P(0) = 1-p$, $P(1) = p$.

Notation 1.4 Throughout, we take log to be the base-2 logarithm, $\log_2$.

Definition: Binary Entropy Function

Definition 1.5 The **binary entropy function** $h : (0, 1) \to [0, 1]$ is defined as

$$h(p) := -p \log p - (1 - p) \log(1 - p)$$

Example 1.6 Let $x_1^n \in \{0, 1\}^n$ be an n bit string which is the realisation of binary random variables (RVs) $X_1^n = (X_1, ..., X_n)$, where the X_i are independent and identically distributed (IID), with common distribution $X_i \sim \text{Bern}(p)$. Let $k = |\{i \in [n] : x_i = 1\}|$ be the number of ones in x_1^n. We have

$$\mathbb{P}(X_1^n = x_1^n) := P^n(x_1^n) = \prod_{i=1}^n P(x_i) = p^k(1-p)^{n-k}.$$

Now by the law of large numbers, the proportion of ones in a random x_1^n is $k/n \approx p$ with high probability for large n. Hence,

$$P^n(x_1^n) \approx p^{np}(1-p)^{n(1-p)} = 2^{-nh(p)}.$$

Note that this reveals an amazing fact: this approximation is independent of x_1^n, so any message we are likely to encounter has roughly the same probability $\approx 2^{-nh(p)}$ of occurring.

Remark 1.7 By the above example, we can split the set of all possible n-bit messages, $\{0,1\}^n$, into two parts: the set B_n of **typical** messages which are approximately uniformly distributed with probability $\approx 2^{-nh(p)}$ each, and the non-typical messages that occur with negligible probability. Since all but a very small amount of the probability is concentrated in B_n, we have $|B_n| \approx 2^{nh(p)}$.

Remark 1.8 Suppose an encoder and decoder both already know B_n and agree on an ordering of its elements: $B_n = \{x_1^n(1), ..., x_1^n(b)\}$, where $b = |B_n|$. Then instead of transmitting the actual message, the encoder can transmit its index $j \in [b]$, which can be described with

$$\lceil \log b \rceil = \lceil \log |B_n| \rceil \approx nh(p)$$

bits.

Remark 1.9

- The closer p is to $\frac{1}{2}$ (intuitively, the more random the messages are), the larger the entropy $h(p)$, and the larger the number of typical strings $|B_n|$.
- Assuing we ignore non-typical strings, which have vanishingly small probability for large n, the "compression rate" of the above method is $h(p)$, since we encode n-bit strings using $nh(p)$-bit strings. $h(p) < 1$ unless the message is uniformly distributed over all of $\{0, 1\}^n$.
- So the closer p is to 0 or 1 (intuitively, the less random the messages are), the smaller the entropy $h(p)$, so the greater the compression rate we can achieve.

1.2. Asymptotic equipartition property

Notation 1.10 We denote a finite alphabet by $A = \{a_1, ..., a_m\}$.

Notation 1.11 If $X_1, ..., X_n$ are IID RVs with values in A, with common distribution described by a PMF $P : A \to [0, 1]$ (i.e. $P(x) = \mathbb{P}(X_i = x)$ for all $x \in A$), then write $X \sim P$, and we say "X has distribution P on A".

Notation 1.12 For $i \leq j$, write X_i^j for the block of random variables $\left(X_i, ..., X_j\right)$, and similarly write x_i^j for the length $j-i+1$ string $\left(x_i, ..., x_j\right) \in A^{i-j+1}$.

Notation 1.13 For IID RVs $X_1, ..., X_n$ with each $X_i \sim P$, denote their joint PMF by $P^n : A^n \to [0,1]$:

$$P^n(x_1^n) = \mathbb{P}(X_1^n = x_1^n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) = \prod_{i=1}^{n} P(x_i),$$

and we say that "the RVs X_1^n have the product distribution P^n".

Definition: Convergence In Probability

Definition 1.14 A sequence of RVs $(Y_n)_{n\in\mathbb{N}}$ **converges in probability** to an RV Y if $\forall\varepsilon > 0$,

$$\mathbb{P}(|Y_n - Y| > \varepsilon) \to 0 \quad \text{as } n \to \infty.$$

Definition: Entropy

Definition 1.15 Let $X \sim P$ be a discrete RV on a countable alphabet A. The **entropy** of X is

$$H(X) = H(P) := -\sum_{x \in A} P(x) \log P(x) = \mathbb{E}[-\log P(X)].$$

Remark: Properties Of Entropy

Remark 1.16

- We use the convention $0 \log 0 = 0$ (this is natural due to continuity: $x \log x \to 0$ as $x \downarrow 0$, and also can be derived measure-theoretically).
- Entropy is technically a functional the probability distribution P and not of X, but we use the notation $H(X)$ as well as $H(P)$.
- $H(X)$ only depends on the probabilities $P(x)$, not on the values $x \in A$. Hence for any bijective $f : A \to A$, we have $H(f(X)) = H(X)$.
- All summands of $H(X)$ are non-negative, so the sum always exists and is in $[0, \infty]$, even if A is countable infinite.

- $H(X) = 0$ iff all summands are 0, i.e. if $P(x) \in \{0, 1\}$ for all $x \in A$, i.e. X is **deterministic** (constant, so equal to a fixed $x_0 \in A$ with probability 1).

Theorem: Convergence To Common Entropy In Probability

Theorem 1.17 Let $X = \{X_n : n \in \mathbb{N}\}$ be IID RVs with common distribution P on a finite alphabet A. Then

$$-\frac{1}{n} \log P^n(X_1^n) \longrightarrow H(X_1) \quad \text{in probability} \quad \text{as } n \to \infty$$

Proof (Hints). Straightforward. $\square$

Proof. We have

$$P^n(X_1^n) = \prod_{i=1}^{n} P(X_i)$$

$$\implies -\frac{1}{n}\log P^n(X_1^n) = -\frac{1}{n}\sum_{i=1}^{n}\log P(X_i) \to \mathbb{E}[-\log P(X_1)] \quad \text{in probability}$$

by the weak law of large numbers (WLLN) for the IID RVs $Y_i = -\log P(X_i)$. $\square$

Corollary: Aep

Corollary 1.18 (Asymptotic Equipartition Property (AEP)) Let $\{X_n : n \in \mathbb{N}\}$ be IID RVs on a finite alphabet A with common distribution P and common entropy $H = H(X_i)$. Then

- $(\Longrightarrow)$: for all $\varepsilon > 0$, the set of **typical strings** $B_n^*(\varepsilon) \subseteq A^n$ defined by

$$B_n^*(\varepsilon) := \left\{x_1^n \in A^n : 2^{-n(H+\varepsilon)} \leq P^n(x_1^n) \leq 2^{-n(H-\varepsilon)}\right\}$$

satisfies

$$|B_n^*(\varepsilon)| \leq 2^{n(H+\varepsilon)} \quad \forall n \in \mathbb{N}, \quad \text{and}$$
$$P^n(B_n^*(\varepsilon)) = \mathbb{P}(X_1^n \in B_n^*(\varepsilon)) \longrightarrow 1 \quad \text{as } n \to \infty$$

- ($\Longleftarrow$): for any sequence $(B_n)_{n\in\mathbb{N}}$ of subsets of A^n, if $\mathbb{P}(X_1^n \in B_n) \to 1$ as $n \to \infty$, then $\forall \varepsilon > 0$,

$$|B_n| \geq (1-\varepsilon)2^{n(H-\varepsilon)} \quad \text{eventually}$$

$$\text{i.e. } \exists N \in \mathbb{N} : \forall n \geq N, \quad |B_n| \geq (1-\varepsilon)2^{n(H-\varepsilon)}.$$

Proof (Hints).

- ($\Longrightarrow$): straightforward.
- ($\Longleftarrow$): show that $P^n(B_n \cap B_n^*(\varepsilon)) \to 1$ as $n \to \infty$.

$\square$

Proof.

- ($\Longrightarrow$):
  - Let $\varepsilon > 0$. By Theorem 1.17, we have

$$\mathbb{P}(X_1^n \notin B_n^*(\varepsilon)) = \mathbb{P}\left(\left|-\frac{1}{n}\log P^n(X_1^n) - H\right| > \varepsilon\right) \to 0 \quad \text{as } n \to \infty.$$

  - By definition of $B_n^*(\varepsilon)$,

$$1 \geq P^n(B_n^*(\varepsilon)) = \sum_{x_1^n \in B_n^*(\varepsilon)} P^n(x_1^n) \geq |B_n^*(\varepsilon)| 2^{-n(H+\varepsilon)}.$$

- ($\Longleftarrow$):

- We have $P^n(B_n \cap B_n^*(\varepsilon)) = P^n(B_n) + P^n(B_n^*(\varepsilon)) - P^n(B_n \cup B_n^*(\varepsilon)) \geq P^n(B_n) + P^n(B_n^*(\varepsilon)) - 1$, so $P^n(B_n \cap B_n^*(\varepsilon)) \to 1$.
- So $P^n(B_n \cap B_n^*(\varepsilon)) \geq 1 - \varepsilon$ eventually, and so

$$1 - \varepsilon \leq P^n(B_n \cap B_n^*(\varepsilon)) = \sum_{x_1^n \in B_n \cap B_n^*(\varepsilon)} P^n(x_1^n)$$
$$\leq |B_n \cap B_n^*(\varepsilon)| 2^{-n(H-\varepsilon)} \leq |B_n| 2^{-n(H-\varepsilon)}.$$

$\square$

Remark 1.19

- The $\Longrightarrow$ part of AEP states that a specific object (in this case, the $B_n^*(\varepsilon)$) can achieve a certain performance, while the $\Longleftarrow$ part states that no other object of this type can significantly perform better. This is common type of result in information theory.
- Theorem 1.17 gives a mathematical interpretation of entropy: the probability of a random string X_1^n generally decays exponentially with n ($P^n(X_1^n) \approx 2^{-nH}$ with high probability for large n). The AEP gives a more "operational interpretation": the smallest set of strings that can carry almost all the probability of P^n has size $\approx 2^{nH}$.

- The AEP tells us that higher entropy means more typical strings, and so the possible values of X_1^n are more unpredictable. So we consider "high entropy" RVs to be "more random" and "less predictable".

1.3. Fixed-rate lossless data compression

Definition: Source.Memoryless

Definition 1.20 A **memoryless source** $X = \{X_n : n \in \mathbb{N}\}$ is a sequence of IID RVs with a common PMF P on the same alphabet A.

Definition: Fixed Rate Code

Definition 1.21 A **fixed-rate lossless compression code** for a source X consists of a sequence of **codebooks** $\{B_n : n \in \mathbb{N}\}$, where each $B_n \subseteq A^n$ is a set of source strings of length n.

Assume the encoder and decoder share the codebooks, each of which is sorted. To send x_1^n, an encoder checks if $x_1^n \in B_n$; if so, they send the index of x_1^n in B_n, along with a flag bit 1, which requires $1 + \lceil \log |B_n| \rceil$ bits. Otherwise, they send x_1^n uncompressed, along with a flag bit 0 to indicate an "error", which requires $1 + \lceil \log |A^n| \rceil = 1 + \lceil n \log |A| \rceil$ bits.

Definition: Code Rate

Definition 1.22 For each $n \in \mathbb{N}$, the **rate** of a fixed-rate code $\{B_n : n \in \mathbb{N}\}$ for a source X is

$$R_n := \frac{1}{n}(1 + \lceil \log |B_n| \rceil) \approx \frac{1}{n} \log |B_n| \quad \text{bits/symbol.}$$

Definition: Code Error Probability

Definition 1.23 For each $n \in \mathbb{N}$, the **error probability** of a fixed-rate code $\{B_n : n \in \mathbb{N}\}$ for a source X is

$$P_e^{(n)} := \mathbb{P}(X_1^n \notin B_n).$$

Theorem: Fixed Rate Coding Theorem

Theorem 1.24 (Fixed-rate Coding Theorem) Let $X = \{X_n : n \in \mathbb{N}\}$ be a memoryless source with distribution P and entropy $H = H(X_i)$.

- ($\Longrightarrow$): $\forall \varepsilon > 0$, there is a fixed-rate code $\{B_n^*(\varepsilon) : n \in \mathbb{N}\}$ with vanishing error probability ($P_e^{(n)} \to 0$ as $n \to \infty$) and with rate

$$R_n \leq H + \varepsilon + \frac{2}{n} \quad \forall n \in \mathbb{N}.$$

- ($\Longleftarrow$): let $\{B_n : n \in \mathbb{N}\}$ be a fixed-rate with vanishing error probability. Then $\forall \varepsilon > 0$, its rate R_n satisfies

$$R_n > H - \varepsilon \quad \text{eventually}.$$

Proof (Hints). ($\Longrightarrow$): straightforward. ($\Longleftarrow$): explain why $0 < \varepsilon < 1/2$ WLOG. $\square$

Proof.

- ($\Longrightarrow$):
  - Let $B_n^*(\varepsilon)$ be the sets of typical strings defined in AEP (Asymptotic Equipartition Property (AEP)). Then $P_e^{(n)} = 1 - \mathbb{P}(X_1^n \in B_n^*) \to 0$ as $n \to \infty$ by AEP.
  - Also by AEP, $R_n = \frac{1}{n}(1 + \lceil \log|B_n^*| \rceil) \le \frac{1}{n}\log|B_n^*| + \frac{2}{n} \le H + \varepsilon + \frac{2}{n}$.
- ($\Longleftarrow$):
  - WLOG let $0 < \varepsilon < 1/2$. By AEP,

$$R_n \ge \frac{1}{n}\log|B_n^*| + \frac{1}{n} \ge \frac{1}{n}\log(1-\varepsilon) + H - \varepsilon + \frac{1}{n} = H - \varepsilon + \frac{1}{n}\log(2(1-\varepsilon)) > H$$

eventually.

$\square$

2. Relative entropy

Definition: Hypothesis Test

Definition 2.1 Suppose $x_1^n \in A^n$ are observations generated by IID RVs X_1^n and we want to decide whether $X_1^n \sim P^n$ or Q^n, for two distinct candidate PMFs P, Q on A. A **hypothesis test** is described by a **decision region** $B_n \subseteq A^n$ such that

- If $x_1^n \in B_n$, then we declare that $X_1^n \sim P^n$.
- Otherwise, if $x_1^n \notin B_n$, then we declare that $X_1^n \sim Q^n$.

Definition: Hypothesis Test Error Probability

Definition 2.2 The associated **error probabilities** for a hypothesis test are

$$e_1^{(n)} = e_1^{(n)}(B_n) := \mathbb{P}(\text{declare } P \mid \text{data} \sim Q) = Q^n(B_n)$$
$$e_2^{(n)} = e_2^{(n)}(B_n) := \mathbb{P}(\text{declare } Q \mid \text{data} \sim P) = P^n(B_n^c).$$

Definition: Relative Entropy

Definition 2.3 The **relative entropy** between PMFs P and Q on the same countable alphabet A is

$$D(P \parallel Q) := \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)} = \mathbb{E}\left[\log \frac{P(X)}{Q(X)}\right], \quad \text{where } X \sim P.$$

Remark 2.4

- We use the convention that $0 \log \frac{0}{0} = 0$ (this can be avoided by defining relative entropy measure-theoretically).
- $D(P \parallel Q)$ always exists and $D(P \parallel Q) \geq 0$ with equality iff $P = Q$.
- Relative entropy is not symmetric: $D(P \parallel Q) \neq D(Q \parallel P)$ in general, and does not satisfy the triangle inequality.
- Despite this, it is reasonable and natural to think of $D(P \parallel Q)$ as a statistical "distance" between P and Q.

Remark 2.5 Let $X \sim P$. We have, by WLLN,

$$\begin{aligned}\frac{1}{n}\log\left(\frac{P^n(X_1^n)}{Q^n(X_1^n)}\right) &= \frac{1}{n}\log\prod_{i=1}^{n}\frac{P(X_i)}{Q(X_i)}\\ &= \frac{1}{n}\sum_{i=1}^{n}\log\frac{P(X_i)}{Q(X_i)}\\ &\longrightarrow D(P \parallel Q) \text{ in probability} \quad \text{as } n \to \infty.\end{aligned}$$

So for large n, $\frac{P^n(X_1^n)}{Q^n(X_1^n)} \approx 2^{nD(P \parallel Q)}$ with high probability. Hence, the random string X_1^n is exponentially more likely under its true distribution P than under Q.

2.1. Asymptotically optimal hypothesis testing

Theorem: Steins Lemma

Theorem 2.6 (Stein's Lemma) Let P, Q be PMFs on a finite alphabet A, with $D = D(P \parallel Q) \in (0, \infty)$. Let $X = \{X_n : n \in \mathbb{N}\}$ be a memoryless source on A, with either each $X_i \sim P$ or each $X_i \sim Q$.

- $(\Longrightarrow)$: for all $\varepsilon > 0$, there is a hypothesis test with decision regions $\{B_n^*(\varepsilon) : n \in \mathbb{N}\}$ such that

$$\forall n \in \mathbb{N}, \quad e_1^{(n)}(B_n^*(\varepsilon)) \leq 2^{-n(D-\varepsilon)}$$

  and $e_2^{(n)} \to 0$ as $n \to \infty$.
- $(\Longleftarrow)$: for any hypothesis test with decision regions $\{B_n : n \in \mathbb{N}\}$ such that $e_2^{(n)}(B_n) \to 0$ as $n \to \infty$, we have $\forall \varepsilon > 0$,

$$e_1^{(n)}(B_n) \geq 2^{-n\left(D+\varepsilon+\frac{1}{n}\right)} \quad \text{eventually.}$$

Proof (Hints).

- ($\Longrightarrow$):
  - Let $B_n^*(\varepsilon) = \left\{ x_1^n \in A^n : 2^{n(D-\varepsilon)} \leq \frac{P^n(x_1^n)}{Q^n(x_1^n)} \leq 2^{n(D+\varepsilon)} \right\}$. The rest is straightforward (use above remark).
- ($\Longleftarrow$):
  - Show that $P^n(B_n^*(\varepsilon) \cap B_n) \to 1$ as $n \to \infty$, use that $\frac{1}{2} = 2^{-n(1/n)}$.

$\square$

Proof.

- $(\Longrightarrow)$:
  - Let $B_n^*(\varepsilon) = \left\{ x_1^n \in A^n : 2^{n(D-\varepsilon)} \le \frac{P^n(x_1^n)}{Q^n(x_1^n)} \le 2^{n(D+\varepsilon)} \right\}$.
  - Then the convergence in probability of $\frac{1}{n}\sum_{i=1}^n \log \frac{P(X_i)}{Q(X_i)}$ is equivalent to $\mathbb{P}(X_1^n \notin B_n^*) = P^n(B_n^*(\varepsilon)) = e_2^{(n)} \to 0$ as $n \to \infty$, when $X_1^n \sim P^n$.
  - Also, $1 \ge P^n(B_n^*) = \sum_{x_1^n \in B_n^*(\varepsilon)} Q^n(x_1^n) \frac{P^n(x_1^n)}{Q^n(x_1^n)} \ge 2^{n(D-\varepsilon)} \sum_{x_1^n \in B_n^*(\varepsilon)} Q^n(x_1^n) = 2^{n(D-\varepsilon)} Q^n(B_n^*(\varepsilon))$.
- $(\Longleftarrow)$:

- We havee $e_2^{(n)}(B_n^*(\varepsilon)) = P^n(B_n^*(\varepsilon)) \to 0$ as $n \to \infty$. Suppose $e_2^{(n)}(B_n) = P^n(B_n^c) \to 0$. Then $P^n(B_n \cap B_n^*(\varepsilon)) \to 1$. So eventually,

$$
\begin{aligned}
\frac{1}{2} \leq P^n(B_n \cap B_n^*(\varepsilon)) &= \sum_{x_1^n \in B_n \cap B_n^*(\varepsilon)} P^n(x_1^n) \frac{Q^n(x_1^n)}{Q^n(x_1^n)} \\
&\leq 2^{n(D+\varepsilon)} \sum_{x_1^n \in B_n} Q^n(x_1^n) \\
&= 2^{n(D+\varepsilon)} Q^n(B_n) = 2^{n(D+\varepsilon)} e_1^{(n)}(B_n)
\end{aligned}
$$

$\square$

Remark 2.7

- The decision regions B_n^* are asymptotically optimal in that, among all tests that have $e_2^{(n)} \to 0$, they achieve the asymptotically smallest possible $e_1^{(n)} \approx 2^{-nD}$. However, they are not the most optimal decision regions for finite n. For finite regions, the optimal regions are given by the Neyman-Pearson Lemma.
- Assuming $D \neq 0$ is a trivial assumption, as otherwise $P = Q$ on A, so any test would give the correct answer.
- Assuming $D < \infty$ is a reasonable assumption, as otherwise there is some $a \in A$ such that $P(a) > 0$ but $Q(a) = 0$. In that case, we check whether any such a appear in x_1^n or not.

- In Stein's Lemma, we assume one error vanishes at possibly an arbitrarily slow rate, while the other decays exponentially. This is a natural asymmetry in many applications, e.g. in diagnosing disease.
- Stein's Lemma shows why the relative entropy is a natural measure of "distance" between two distributions, as large D means a smaller error probability (one vanishes exponentially at rate D), so easier to tell apart the distributions from the data.

2.2. Relative entropy and optimal hypothesis testing

Theorem: Neyman Pearson Lemma

Theorem 2.8 (Neyman-Pearson Lemma) For a hypothesis test between P and Q based on n data samples, the **likelihood ratio decision regions**

$$B_{\mathrm{NP}} = \left\{ x_1^n \in A^n : \frac{P^n(x_1^n)}{Q^n(x_1^n)} \geq T \right\}, \quad \text{for some threshold } T > 0,$$

are optimal in that, for any decision region $B_n \subseteq A^n$, if $e_1^{(n)}(B_n) \leq e_1^{(n)}(B_{\mathrm{NP}})$, then $e_2^{(n)}(B_n) \geq e_2^{(n)}(B_{\mathrm{NP}})$, and vice versa.

Proof (Hints). Consider the inequality

$$\left(P^n(x_1^n) - TQ^n(x_1^n)\right)\Big(\mathbb{1}_{B_{\mathrm{NP}}}(x_1^n) - \mathbb{1}_{B_n}(x_1^n)\Big) \geq 0$$

(justify why this holds). $\square$

Proof.

- Consider the obvious inequality

$$\left(P^n(x_1^n) - TQ^n(x_1^n)\right)\Big(\mathbb{1}_{B_{\mathrm{NP}}}(x_1^n) - \mathbb{1}_{B_n}(x_1^n)\Big) \geq 0$$

- Then, summing over all x_1^n,

$$\begin{aligned}
0 &\leq P^n(B_{\mathrm{NP}}) - P^n(B_n) - TQ^n(B_{\mathrm{NP}}) + TQ^n(B_n) \\
&= 1 - e_2^{(n)}(B_{\mathrm{NP}}) - \Big(1 - e_2^{(n)}(B_n)\Big) - T\Big(e_1^{(n)}(B_{\mathrm{NP}}) - e_1^{(n)}(B_n)\Big) \\
&\Longrightarrow e_2^{(n)}(B_n) - e_2^{(n)}(B_{\mathrm{NP}}) \geq T\Big(e_1^{(n)}(B_{\mathrm{NP}}) - e_1^{(n)}(B_n)\Big)
\end{aligned}$$

$\square$

Remark 2.9 Neyman-Pearson says that if any decision region has an error as small as that of B_{NP}, then its other error must be larger than that of B_{NP}.

Notation 2.10 Let $\hat{P}_n$ denote the empirical distribution (or **type**) induced by x_1^n on A^n (the frequency with which $a \in A$ occurs in x_1^n):

$$\forall a \in A, \quad \hat{P}_n(a) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i = a\}}$$

Proposition: Informatation Theoretic Form Of Neyman Pearson Decision Region

Proposition 2.11 The Neyman-Pearson decision region B_{NP} can be expressed in information-theoretic form as

$$B_{\mathrm{NP}} = \left\{ x_1^n \in A^n : D\left(\hat{P}_n \parallel Q\right) \geq D\left(\hat{P}_n \parallel P\right) + T' \right\}$$

where $T' = \frac{1}{n} \log T$.

Proof (Hints). Rewrite the expression $\frac{1}{n}\log\frac{P^n(x_1^n)}{Q^n(x_1^n)}$. $\square$

Proof. We have

$$\begin{aligned}
\frac{1}{n}\log\frac{P^n(x_1^n)}{Q^n(x_1^n)} &= \frac{1}{n}\log\left(\prod_{i=1}^{n}\frac{P(x_i)}{Q(x_i)}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\log\frac{P(x_i)}{Q(x_i)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A}\mathbb{1}_{\{x_i=a\}}\log\frac{P(a)}{Q(a)} \\
&= \sum_{a\in A}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{x_i=a\}}\right)\log\frac{P(a)}{Q(a)}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{a \in A} \hat{P}_n(a) \log\left(\frac{P(a)}{Q(a)} \cdot \frac{\hat{P}_n(a)}{\hat{P}_n(a)}\right) \\
&= D\big(\hat{P}_n \parallel Q\big) - D\big(\hat{P}_n \parallel P\big).
\end{aligned}
$$

$\square$

Theorem: Jensens Inequality

Theorem 2.12 (Jensen's Inequality) Let I be an interval, $f : I \to \mathbb{R}$ be convex and X be an RV with values in I. Then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

Moreover, if f is strictly convex, then equality holds iff X is almost surely constant.

Proof. Omitted. $\square$

Theorem: Log Sum Inequality

Theorem 2.13 (Log-sum Inequality) Let $a_1, ..., a_n$, $b_1, ..., b_n$ be non-negative constants. Then

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality iff $\frac{a_i}{b_i} = c$ for all i, for some constant c. We use the convention that $0 \log 0 = 0 \log \frac{0}{0} = 0$.

Remark 2.14 This also holds for countably many a_i and b_i.

Proof (Hints). Use Jensen's inequality with X the RV such that $\mathbb{P}\left(X = \frac{a_i}{b_i}\right) = \frac{b_i}{\sum_{j=1}^n b_j}$ for all $i \in [n]$, and a suitable f. $\square$

Proof.

- Define
$$f(x) = \begin{cases} x \log x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}.$$
f is strictly convex.
- Let $A = \sum_i a_i$, $B = \sum_i b_i$. Let X be the RV with $\mathbb{P}\left(X = \frac{a_i}{b_i}\right) = \frac{b_i}{B}$ for all $i \in [n]$.
- Then $\mathbb{E}[f(X)] = \sum_i \frac{b_i}{B} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \frac{1}{B} \sum_i a_i \log \frac{a_i}{b_i}$.
- $f(\mathbb{E}[X]) = \mathbb{E}[X] \log \mathbb{E}[X] = \sum_i \frac{a_i}{b_i} \frac{b_i}{B} \log \sum_i \frac{a_i}{b_i} \frac{b_i}{B} = \frac{A}{B} \log \frac{A}{B}$.
- So by Jensen's inequality, $\frac{A}{B} \log \frac{A}{B} \leq \frac{1}{B} \sum_i a_i \log \frac{a_i}{b_i}$.

Proposition: Bounds On Entropy And Relative Entropy

Proposition 2.15

1. If P and Q are PMFs on the same finite alphabet A, then

$$D(P \parallel Q) \geq 0$$

with equality iff $P = Q$.

2. If $X \sim P$ on a finite alphabet A, then

$$0 \leq H(X) \leq \log|A|$$

with equality to 0 iff X is a constant, and equality to $\log|A|$ iff X is uniformly distributed on A.

Remark 2.16 This also holds for countably infinite A.

Proof (Hints).

1. Straightforward.
2. For $\leq \log|A|$, consider $D(P \parallel Q)$ where Q is the uniform distribution on A. ≥ 0 is straightforward.

$\square$

Proof.

- ▸ By the log-sum inequality,

$$D(P \parallel Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)} \geq \left(\sum_{x \in A} P(x) \right) \log \frac{\sum_{x \in A} P(x)}{\sum_{x \in A} Q(x)} = 0$$

  with equality if $\frac{P(x)}{Q(x)}$ is the same constant for all $x \in A$, i.e. $P = Q$.
- ▸ Let Q be the uniform distribution on A, so $H(Q) = \sum_{x \in A} \frac{1}{|A|} \log \frac{1}{1/|A|} = \log|A|$.
  ▸ Now $0 \leq D(P \parallel Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{1/|A|} = \log|A| - H(X)$ with equality iff $P = Q$, i.e. P is uniform.

- Each term in $-H(X)$ is ≤ 0, with equality iff each $P(x) \log P(x)$ is 0, i.e. $P(x) = 0$ or 1.

$\square$

Remark 2.17 If $X = \{X_n : n \in \mathbb{N}\}$ is a memoryless source with PMF P on A, then we have shown that it can be at best compressed to $\approx H(P)$ bits/symbol. This means that we can always achieve non-trivial compression, i.e. a description using $\approx H(P) < \log|A|$ bits/symbol, unless the source X is completely random (i.e. IID and uniformly distribute), in which case we cannot do better than simply describing each x_1^n uncompressed using $\frac{\lceil \log|A^n| \rceil}{n} \approx \log|A|$ bits/symbol.

3. Properties of entropy and relative entropy

3.1. Joint entropy and conditional entropy

Definition: Joint Entropy

Definition 3.1 Let X_1^n be an arbitrary finite collection of discrete RVs on corresponding alphabets $A_1, ..., A_n$. Note we can think of X_1^n itself a discrete RV on alphabet $A_1 \times \cdots \times A_n$. Let X_1^n have PMF P_n, then the **joint entropy** of X_1^n is

$$H(X_1^n) = H(P_n) = H(X_1, ..., X_n) := \mathbb{E}[-\log P_n(X_1^n)] = -\sum_{x_1^n \in A^n} P_n(x_1^n) \log P_n(x_1^n)$$

Example 3.2 Note that if X and Y are independent, then $P_{X,Y}(x,y) = P_X(x)P_Y(y)$, so

$$H(X,Y) = \mathbb{E}\big[-\log P_{X,Y}(X,Y)\big] = \mathbb{E}[-\log P_X(X) - \log P_Y(Y)] = H(X) + H(Y)$$

Example 3.3 Let X and Y have joint PMF given by

$Y \backslash X$	1	2	3	
0	1/10	1/5	1/4	11/20
1	1/5	1/20	1/5	9/20
	3/10	1/4	9/20	

Note that X and Y are not independent. We have

$$H(X) = -\frac{3}{10}\log\frac{3}{10} - \frac{1}{4}\log\frac{1}{4} - \frac{9}{20}\log\frac{9}{20} \approx 1.539,$$

$$H(Y) = -\frac{11}{20}\log\frac{11}{20} - \frac{9}{20}\log\frac{9}{20} \approx 0.993,$$

$$H(X, Y) = -\frac{1}{10}\log\frac{1}{10} - \cdots - \frac{1}{5}\log\frac{1}{5} \approx 2.441 < H(X) + H(Y).$$

In general, if X and Y are not independent, then $P_{XY}(x, y) = P_X(x)P_{Y \mid X}(y \mid x)$, so

$$H(X, Y) = \mathbb{E}[-\log P_{XY}(x, y)] = \mathbb{E}[-\log P_X(x)] + \mathbb{E}\left[-\log P_{Y \mid X}(y \mid x)\right].$$

Definition: Conditional Entropy

Definition 3.4 Let X and Y be discrete random variables with joint PMF $P_{X,Y}$, then the **conditional entropy** of Y given X is

$$H(Y \mid X) = \mathbb{E}\Big[-\log P_{Y \mid X}(Y \mid X)\Big] = -\sum_{x,y} P_{X,Y}(x,y) \log P_{Y \mid X}(y \mid x)$$

Note 3.5 $P_{Y \mid X}$ is a function of $(x, y) \in X$, and so for the expected value we multiply the log by the probability that $X = x$ and $Y = y$.

Proposition: Joint Entropy Is Single Entropy Plus Conditional Entropy

Proposition 3.6 For discrete RVs X and Y, we have

$$H(Y \mid X) = H(X, Y) - H(X).$$

Proof (Hints). Straightforward. $\square$

Proof. Note that $P_{Y \mid X}(y \mid x) = \mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(Y=y, X=x)}{\mathbb{P}(X=x)} = P_{X,Y}(x,y)P_X(x)$. Hence

$$\begin{aligned} H(X,Y) &= \mathbb{E}\Big[-\log P_{X,Y}(X,Y)\Big] \\ &= \mathbb{E}\Big[-\log P_X(X) - \log P_{Y \mid X}(Y \mid X)\Big] \\ &= \mathbb{E}[-\log P_X(X)] + \mathbb{E}\Big[-\log P_{Y \mid X}(Y \mid X)\Big]. \end{aligned}$$

$\square$

3.2. Properties of entropy, joint entropy and conditional entropy

Proposition: Entropy Chain Rule

Proposition 3.7 (Chain Rule for Entropy) Let X_1^n be a collection of discrete RVs. Then

$$H(X_1^n) = \sum_{i=1}^{n} H(X_i \mid X_1^{i-1}).$$

In particular, if the X_1^n are independent, then

$$H(X_1^n) = \sum_{i=1}^{n} H(X_i).$$

Proof (Hints). By induction. $\square$

Proof. We can write

$$\begin{aligned} P_{X_1^n}(x_1^n) &= P_{X_1}(x_1)P_{X_2|X_1}(x_2 \mid x_1)\cdots P_{X_n \mid X_1,\ldots,x_{n-1}}(x_n \mid x_1,\ldots,x_{n-1}) \\ &= \prod_{i=1}^{n} P_{X_i \mid X_1^{i-1}}(x_i \mid x_1^{i-1}). \end{aligned}$$

Then the result follows by inductively using the above proposition. $\square$

Proposition: Conditioning Reduces Entropy

Proposition 3.8 (Conditioning Reduces Entropy) For discrete RVs X and Y,

$$H(Y \mid X) \leq H(Y)$$

with equality iff X and Y are independent.

Proof (Hints). Express $H(Y) - H(Y \mid X)$ as a relative entropy. $\square$

Proof. We have

$$
\begin{aligned}
H(Y) - H(Y \mid X) &= \mathbb{E}[-\log P_Y(Y)] - \mathbb{E}\left[-\log P_{Y \mid X}(Y \mid X)\right] \\
&= \mathbb{E}\left[\log \frac{P_{Y \mid X}(Y \mid X)}{P_Y(Y)}\right] \\
&= \mathbb{E}\left[\log \frac{P_{Y \mid X}(Y \mid X) P_X(X)}{P_Y(Y) P_X(X)}\right] \\
&= \mathbb{E}\left[\log \frac{P_{X,Y}(X, Y)}{P_X(X) P_Y(Y)}\right]
\end{aligned}
$$

$$= D\big(P_{X,Y} \parallel P_X P_Y\big) \geq 0,$$

with equality iff $P_{X,Y} = P_X P_Y$, i.e. X and Y are independent. $\square$

Definition: Conditional Independence

Definition 3.9 Discrete RVs X and Z are **conditionally independent given Y** if:

- $P_{X,Z \mid Y}(x, z \mid y) = P_{X \mid Y}(x \mid y) P_{Z \mid Y}(z \mid y)$,
- or equivalently, $P_{X \mid Z,Y}(x \mid z, y) = P_{X \mid Y}(x \mid y)$,
- or equivalently, $P_{Z \mid X,Y}(z \mid x, y) = P_{Z \mid Y}(z \mid y)$.

We denote this by writing $X - Y - Z$ and we say that X, Y, Z form a Markov chain. Note that $X - Y - Z$ is equivalent to $Z - Y - X$, but not to $X - Z - Y$.

Note: X Y And Function Of Y Form Markov Chain

Note 3.10 For any function g on Y, we have $X - Y - g(Y)$.

Corollary: Subadditivity Of Entropy

Corollary 3.11 $H(X_1^n) \leq \sum_{i=1}^n H(X_i)$ with equality iff all X_1^n are independent.

Proof. Straightforward. $\square$

Proof. $H(X_1^n) = \sum_{i=1}^n H(X_i \mid X_1^{i-1}) \le \sum_{i=1}^n H(X_i)$ by the chain rule and conditioning reducing entropy. $\square$

Remark: Expression For Conditional Entropy

Remark 3.12 We can write

$$\begin{aligned}
H(Y \mid X) &= -\sum_{x,y} \left(P_{X,Y}(x,y)\right) \log P_{Y \mid X}(y \mid x) \\
&= \sum_{x} P_X(x) \left(-\sum_{y} P_{Y \mid X}(y \mid x) \log P_{Y \mid X}(y \mid x) \right) \\
&=: \sum_{x} P_X(x) H(Y \mid X = x)
\end{aligned}$$

Note $H(Y \mid X = x)$ is **not** a conditional entropy, and in particular, we do not always have $H(Y \mid X = x) \leq H(Y)$. Since $0 \leq H(Y \mid X =$

$x) \leq \log|A_Y|$, we have $0 \leq H(Y \mid X) \leq \log|A_Y|$ with equality to 0 iff Y is a function of X (i.e. $H(Y \mid X = x) = 0$ for all x).

Proposition: Entropy Data Processing

Proposition 3.13 (Data Processing Inequality for Entropy) Let X be discrete RV on alphabet A and f be function on A. Then

1. $H(f(X)|X) = 0$.
2. $H(f(X)) \leq H(X)$ with equality iff f is injective.

Proof (Hints). Use that $x \mapsto (x, f(x))$ is injective and the chain rule.
$\square$

Proof. We have already shown the "if" direction of 2. We have $H(X) = H(X, f(X)) = H(f(X)|X) + H(X)$, since $x \mapsto (x, f(x))$ is injective. Also, $H(X) = H(X, f(X)) = H(X \mid f(X)) + H(f(X)) \geq H(f(X))$. So $H(X) \geq H(f(X))$ with equality iff $H(X \mid f(X)) = 0$, i.e. X is a deterministic function of $f(X)$, i.e. f is invertible. $\square$

Proposition: Properties Of Conditional Entropy

Proposition 3.14 (Properties of Conditional Entropy) For discrete RVs X, Y, Z:

- Chain rule: $H(X, Z \mid Y) = H(X \mid Y) + H(Z \mid X, Y)$.
- Subadditivity: $H(X, Z \mid Y) \leq H(X \mid Y) + H(Z \mid Y)$ with equality iff X and Z are conditionally independent given Y.
- Conditioning reduces entropy: $H(X \mid Y, Z) \leq H(X \mid Y)$ with equality iff X and Z are conditionally independent given Y.

Proof. Exercise. $\square$

Theorem: Fano Inequality

Theorem 3.15 (Fano's Inequality) Let X and Y be RVs on respective alphabets A and B. Suppose we are interested in the RV X but only are allowed to observe the possibly correlated RV Y. Consider the estimate $\widehat{X} = f(Y)$, with probability of error $P_e := \mathbb{P}(\widehat{X} \neq X)$. Then

$$H(X \mid Y) \leq h(P_e) + P_e \log(|A| - 1),$$

where h is the binary entropy function.

Proof (Hints). Consider an "error" Bernoulli RV E which depends on X and Y. Use the chain rule in two directions on $H(X, E \mid Y)$. Merge these and split up into the cases when $E = 0$ and $E = 1$ (using) $\square$

Proof. Let E be the binary RV taking value 1 when there is an error (i.e. $\widehat{X} \neq X$), and taking value 0 otherwise. So $E \sim \mathrm{Bern}(P_e)$ and $H(E) = h(P_e)$. Then

$$H(X, E \mid Y) = H(X \mid Y) + H(E \mid X, Y) = H(X \mid Y)$$

since E is function of (X, Y). Using the chain rule in the other direction,

$$H(X, E \mid Y) = H(E \mid Y) + H(X \mid E, Y) \leq H(E) + E(X \mid E, Y).$$

Now

$$
\begin{aligned}
H(X \mid Y) - h(P_e) &\le H(X \mid E, Y) \\
&= P_e H(X \mid E = 1, Y) + (1 - P_e) H(X \mid E = 0, Y)
\end{aligned}
$$

When $E = 0$, given Y, we can determine $X = f(Y)$ as a function of Y, so $H(X \mid E = 0, Y) = 0$. When $E = 1$, given Y, we know X doesn't take value $f(Y)$, so there are $|A| - 1$ possible values that it takes, so $H(X \mid E = 1, Y) \le \log(|A| - 1)$. $\square$

3.3. Properties of relative entropy

Theorem: Relative Entropy Data Processing

Theorem 3.16 (Data Processing Inequality for Relative Entropy) Let $X \sim P_X$ and $X' \sim Q_X$ be RVs on the same alphabet A, and $f : A \to B$ be an arbitrary function. Let $P_{f(X)}$ and $Q_{f(X)}$ be the PMFs of $f(X)$ and $f(X')$ respectively. Then

$$D\big(P_{f(X)} \parallel Q_{f(X)}\big) \leq D(P_X \parallel Q_X).$$

Proof (Hints). Use that $P_{f(X)}(y) = \sum_{x \in f^{-1}(\{y\})} P_X(x)$. $\square$

Proof. For each $y \in B$, let $A_y = \{x \in A : f(x) = y\} = f^{-1}(\{y\})$. Then

$$\begin{aligned}
D\Big(P_{f(X)} \,\|\, Q_{f(X)}\Big) &= \sum_{y\in B} P_{f(X)}(y) \log \frac{P_{f(X)}(y)}{Q_{f(X)}(y)} \\
&= \sum_{y\in B} \left(\sum_{x\in A_y} P_X(x) \right) \log \frac{\sum_{x\in A_y} P_X(x)}{\sum_{x\in A_y} Q_X(x)} \\
&\leq \sum_{y\in B} \sum_{x\in A_y} P_X(x) \log \frac{P_X(x)}{Q_X(x)} \quad \text{by log-sum inequality}
\end{aligned}$$

$$= \sum_{x \in A} P_X(x) \log \frac{P_X(x)}{Q_X(x)} = D(P_X \parallel Q_X).$$

$\square$

Remark 3.17 The data processing inequality for relative entropy shows that we cannot make two distributions more "distinguishable" by first "processing" the data (by applying f).

Definition: Total Variation Distance

Definition 3.18 The **total variation distance** between PMFs P and Q on the same alphabet A is

$$\|P - Q\|_{\mathrm{TV}} = \sum_{x \in A} |P(x) - Q(x)|.$$

Remark 3.19 Let $B = \{x \in A : P(x) > Q(x)\}$, then

$$\begin{aligned}
\|P - Q\|_{\mathrm{TV}} &= \sum_{x \in A} |P(x) - Q(x)| \\
&= \sum_{x \in B} (P(x) - Q(x)) + \sum_{x \in B^c} (Q(x) - P(x)) \\
&= P(B) - Q(B) + Q(B^c) - P(B^c) \\
&= P(B) - Q(B) + (1 - Q(B)) + (1 - P(B)) \\
&= 2(P(B) - Q(B)).
\end{aligned}$$

Notation 3.20 Write

$$D_e(P \parallel Q) = (\ln 2) P(D \parallel Q) = \sum_{x \in A} P(x) \log_e \frac{P(x)}{Q(x)}$$

and more generally, write

$$D_c(P \parallel Q) = (\log_c 2) P(D \parallel Q) = \sum_{x \in A} P(x) \log_c \frac{P(x)}{Q(x)}.$$

Theorem: Pinskers Inequality

Theorem 3.21 (Pinsker's Inequality) Let P and Q be PMFs on the same alphabet A. Then

$$\|P - Q\|_{\mathrm{TV}}^2 \le (2\ln 2) D(P \parallel Q) = 2D_e(P \parallel Q).$$

Proof (Hints).

- First prove for case that P and Q are PMFs of $\text{Bern}(p)$ and $\text{Bern}(q)$ (explain why we can assume $q \leq p$ WLOG), by definining $\Delta(p,q) = 2D_e(P \parallel Q) - \|P-Q\|_{\text{TV}}^2$, and showing that $\frac{\partial \Delta(p,q)}{\partial q} \leq 0$.
- Then show for general PMFs by using data processing, where $f = \mathbb{1}_B$ for $B = \{x \in A : P(x) > Q(x)\}$.

$\square$

Proof. First, assume that P and Q are the PMFs of the distributions $\mathrm{Bern}(p)$ and $\mathrm{Bern}(q)$ for some $0 \le q \le p \le 1$ ($q \le p$ WLOG since we can simultaneously interchange both p with $1-p$ and q with $1-q$ if necessary). Let

$$\Delta(p,q) = (2\ln 2)D(P \parallel Q) - \|P-Q\|_{\mathrm{TV}}^2 = 2p\ln\frac{p}{q} + 2(1-p)\ln\frac{1-p}{1-q} - (2(p-q))^2$$

Since $\Delta(p,p)=0$ for all p, it suffices to show that $\frac{\partial\Delta(p,q)}{\partial q} \le 0$. Indeed,

$$\frac{\partial\Delta(p,q)}{\partial q} = 2\frac{p}{q} - 2\frac{1-p}{1-q} - 8(q-p) = 2(q-p)\left(\frac{1}{q(1-q)} - 4\right) \le 0$$

since $q(1-q) \leq \frac{1}{4}$ for all $q \in [0,1]$.

Now, assume P and Q are general PMFs and let $B = \{x \in A : P(x) > Q(x)\}$ and $f = \mathbb{1}_B$. Define the RVs $X \sim P$ and $X' \sim Q$, and let P_f and Q_f be the respective PMFs of the RVs $f(X)$ and $f(X')$. Note that $f(X) \sim \mathrm{Bern}(p)$, $f(X') \sim \mathrm{Bern}(q)$ where $p = P(B)$ and $q = Q(B)$. Then

$$\begin{aligned} 2D_e(P \parallel Q) &\geq 2D_e\left(P_f \parallel Q_f\right) && \text{by data-processing} \\ &\geq \left\|P_f - Q_f\right\|_{\mathrm{TV}}^2 && \text{by above} \\ &= (2(p-q))^2 \end{aligned}$$

$$
\begin{aligned}
&= (2(P(B) - Q(B)))^2 \\
&= \|P - Q\|_{\mathrm{TV}}^2.
\end{aligned}
$$

$\square$

Theorem: Relative Entropy Is Convex

Theorem 3.22 (Convexity of Relative Entropy) The relative entropy $D(P \parallel Q)$ is jointly convex in P, Q: for all PMFs P, P', Q, Q' on the same alphabet and for all $0 < \lambda < 1$,

$$D(\lambda P + (1 - \lambda)P' \parallel \lambda Q + (1 - \lambda)Q') \leq \lambda D(P \parallel Q) + (1 - \lambda)D(P' \parallel Q').$$

Proof. Exercise. $\square$

Corollary: Entropy Is Concave

Corollary 3.23 (Concavity of Entropy) The entropy of $H(P)$ is a concave function on all PMFs P on a finite alphabet.

Proof (Hints). Use convexity of relative entropy of P and a suitable distribution. $\square$

Proof. Let P be a PMF on finite alphabet A and U be the uniform PMF on A. Then by convexity of relative entropy, $D(P \parallel U) = \sum_{x \in A} p(x) \log \frac{P(x)}{1/|A|} = \log m - H(P)$ is convex in P, so $H(P)$ is concave in P. $\square$

4. Poisson approximation

4.1. Poisson approximation via entropy

Theorem: Binomial Converges To Poisson

Theorem 4.1 Let $X_1, ..., X_n$ be IID RVs with each $X_i \sim \text{Bern}(\lambda/n)$, let $S_n = X_1 + \cdots + X_n$. Then $P_{S_n} \to \text{Pois}(\lambda)$ in distribution as $n \to \infty$, i.e. $\forall k \in \mathbb{N}$,

$$\mathbb{P}(S_n = k) \to e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{as } n \to \infty$$

Remark 4.2 Using information theory, we can derive stronger and more general statements than the one above.

Theorem: Relative Entropy Between Sum Of Bernoullis And Poisson Is Bounded

Theorem 4.3 Let $X_1, ..., X_n$ be (not necessarily independent) RVs with each $X_i \sim \text{Bern}(p_i)$. Let $S_n = \sum_{i=1}^n X_i$ and $\lambda = \sum_{i=1}^n p_i = \mathbb{E}[S_n]$. Then

$$D_e\Big(P_{S_n} \,\|\, \text{Pois}(\lambda)\Big) \leq \sum_{i=1}^n p_i^2 + \sum_{i=1}^n H_e(X_i) - H_e(X_1^n).$$

Proof (Hints).

- Let $Z_i = \mathrm{Pois}(p_i)$ for each $i \in [n]$ be independent Poisson RVs so that $T_n = \sum_{i=1}^{n} Z_i \sim \mathrm{Pois}(\lambda)$.
- Use data processing inequality for relative entropy, and prove the fact that $D_e(\mathrm{Bern}(p) \,\|\, \mathrm{Pois}(p)) \leq p^2$ for all $p \in [0,1]$ (use that $1 - p \leq e^{-p}$).

$\square$

Proof. Let $Z_i = \mathrm{Pois}(p_i)$ for each $i \in [n]$ be independent Poisson RVs so that $T_n = \sum_{i=1}^n Z_i \sim \mathrm{Pois}(\lambda)$. Then

$$
\begin{aligned}
P_{S_n} \,\|\, \mathrm{Pois}(\lambda)\Big) &= D_e\Big(P_{S_n} \,\|\, P_{T_n}\Big) \\
&\le D_e\Big(P_{X_1^n} \,\|\, P_{Z_1^n}\Big) \quad \text{by data-processing with } f(x_1^n) = x_1 + \cdots + x \\
&= \mathbb{E}\left[\ln \frac{P_{X_1^n}(X_1^n)}{P_{Z_1^n}(X_1^n)}\right] \\
&= \mathbb{E}\left[\ln\left(\frac{P_{X_1^n}(x_1^n)}{\prod_{i=1}^n P_{Z_1^n}(X_i)} \cdot \frac{\prod_{i=1}^n P_{X_i}(X_i)}{\prod_{i=1}^n P_{X_i}(X_i)}\right)\right]
\end{aligned}
$$

$$= \mathbb{E}\left[\ln\left(\prod_{i=1}^{n}\frac{P_{X_i}(x_i)}{P_{Z_i}(x_i)}\right)\right] + \sum_{x_1^n\in A^n} P_{X_1^n}(x_1^n)\ln\frac{1}{\prod_{i=1}^n P_{X_i}(x_i)} - H_e($$

$$= \sum_{i=1}^{n} D_e\Big(P_{X_i} \parallel P_{Z_i}\Big) + \sum_{i=1}^{n} H_e(X_i) - H_e(X_1^n)$$

since for given $x_1 \in A$, $\sum_{x_2^n\in A^n} P_{X_1^n}(x_1^n) = P_{X_1}(x_1)$ (and similarly for each x_j, $j = 2, ..., n$). Now note that $D_e\Big(P_{X_i} \parallel P_{Z_i}\Big) = D_e(\mathrm{Bern}(p_i) \parallel \mathrm{Pois}(p_i))$, and for all $p \in (0, 1)$,

$$D_e(\mathrm{Bern}(p) \parallel \mathrm{Pois}(p)) = (1-p)\ln\frac{1-p}{e^{-p}} + p\ln\frac{p}{pe^{-p}}$$

$$
\begin{aligned}
&= (1-p)\ln(1-p) + (1-p)p + p^2 \\
&\leq (1-p)\ln(e^{-p}) + p \\
&= p^2
\end{aligned}
$$

since $1-p \leq e^{-p}$ for all $p \in [0,1]$. Similarly, if $p = 0$ or 1, then $D_e(\mathrm{Bern}(p) \,\|\, \mathrm{Pois}(p)) = 0 \leq p^2$. $\square$

Corollary: Relative Entropy Between Sum Of Independent Bernoullis And Poisson Is Bounded

Corollary 4.4 Let $X_1, ..., X_n$ be independent, with each $X_i \sim \text{Bern}(p_i)$. Then

$$D_e\Big(P_{S_n} \,\|\, \text{Pois}(\lambda)\Big) \leq \sum_{i=1}^{n} p_i^2$$

Corollary: Binomial Convergence To Poisson Follows From Relative Entropy Bound

Corollary 4.5 Theorem 4.1 follows directly from Theorem 4.3.

Proof (Hints). Use Pinsker's Inequality. $\square$

Proof. Let P_λ be the PMF of the $\text{Pois}(\lambda)$ distribution. Then by Pinsker's Inequality,

$$\left\|P_{S_n} - P_\lambda\right\|_{\mathrm{TV}}^2 \le 2D_e\left(P_{S_n} \,\|\, \mathrm{Pois}(\lambda)\right) \le 2\sum_{i=1}^{n} \frac{\lambda^2}{n^2} = 2\frac{\lambda^2}{n}.$$

So for each $k \in \mathbb{N}$, $\left|P_{S_n}(k) - P_\lambda(k)\right| \le \left\|P_{S_n} - P_\lambda\right\|_{\mathrm{TV}} \le \sqrt{\frac{2}{n}}\lambda \to 0$ as $n \to \infty$. $\square$

Remark 4.6 Theorem 4.3 is stronger than Theorem 4.1 in that it holds for all n rather than being asymptotic. It also provides an easily computable bound on the difference between P_{S_n} and $\mathrm{Pois}(\lambda)$, and does not assume the p_i are equal, or that the RVs $X_1, \ldots, X_n$ are independent.

Remark 4.7 It is known that for independent $X_1, ..., X_n$, $P_{S_n} \to \text{Pois}(\lambda)$ iff $\sum_{i=1}^{n} p_i^2 \to 0$. So the bound in Theorem 4.3 is the best possible.

4.2. What is the Poisson distribution?

Lemma: Binomial Maximum Entropy

Lemma 4.8 (Binomial Maximum Entropy) Let $B_n(\lambda)$ be set of distributions on $\mathbb{N}_0$ that arise from sums $\sum_{i=1}^{n} X_i$ where $X_i \sim \text{Bern}(p_i)$ are independent and $\sum_{i=1}^{n} p_i = \lambda$. For all $n \geq \lambda$,

$$H_e(\text{Bin}(n, \lambda/n)) = \sup\{H_e(P) : P \in B_n(\lambda)\}$$

Proof. Exercise. $\square$

Theorem: Poisson Maximum Entropy

Theorem 4.9 (Poisson Maximum Entropy) We have

$$
\begin{aligned}
&H_e(\mathrm{Pois}(\lambda)) \\
&= \sup\left\{ H_e(S_n) : S_n = \sum_{i=1}^{n} X_i, X_i \sim \mathrm{Bern}(p_i) \text{ independent} \wedge \sum_{i=1}^{n} p_i = \lambda, n \geq 1 \right\} \\
&= \sup_{n \in \mathbb{N}} \sup\{H_e(P) : P \in B_n(\lambda)\}.
\end{aligned}
$$

Proof. Let $H^* = \sup_{n\in\mathbb{N}} \sup\{H_e(P) : P \in B_n(\lambda)\}$. Note that $B_n(\lambda) \subseteq B_{n+1}(\lambda)$, hence $H^* = \lim_{n\to\infty} \sup\left\{H_{e(P)} : P \in B_n(\lambda)\right\} = \lim_{n\to\infty} H_e(\mathrm{Bin}(n, \lambda/n))$.

Let P_n and Q be respective PMFs of $\mathrm{Bin}(n, \lambda/n)$ and $\mathrm{Pois}(\lambda)$. Using that $k! \le k^k \le e^{k^2}$, we have

$$
\begin{aligned}
H_e(Q) &= \sum_{k=0}^{\infty} Q(k) \ln \frac{k!}{e^{-\lambda}\lambda^k} \\
&\le \sum_{k=0}^{\infty} Q(k)\left(\lambda - k \ln \lambda + k^2\right)
\end{aligned}
$$

$$= \lambda^2 + 2\lambda - \lambda \ln \lambda < \infty$$

since $\mathbb{E}[X] = \lambda$ and $\mathbb{E}[X^2] = \lambda + \lambda^2$ for $X \sim \mathrm{Pois}(\lambda)$. So $H_e(Q)$ is finite. The convergence is left as an exercise. $\square$

5. Mutual information

Definition: Mutual Information

Definition 5.1 The **mutual information** between discrete RVs X and Y is

$$I(X;Y) = H(X) - H(X|Y).$$

The **conditional mutual information** between X and Y given a discrete RV Z is

$$\begin{aligned} I(X;Y \mid Z) &= H(X \mid Z) - H(X \mid Y, Z) \\ &= H(X \mid Z) + H(Y \mid Z) - H(X, Y \mid Z) \\ &= H(Y \mid Z) - H(Y \mid X, Z). \end{aligned}$$

Proposition: Expressions For Mutual Information

Proposition 5.2 Let X and Y be discrete RVs with marginal PMFs P_X and P_Y respectively, and joint PMF $P_{X,Y}$, then the mutual information can be expressed as:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(Y \mid X) \\ &= D\big(P_{X,Y} \parallel P_X P_Y\big). \end{aligned}$$

Proof (Hints). Straightforward. $\square$

Proof. The first two lines are by the chain rule. For the third, we have

$$
\begin{aligned}
(X) + H(Y) - H(X,Y) &= \mathbb{E}[-\log P_X(X)] + \mathbb{E}[-\log P_Y(Y)] - \mathbb{E}\big[-\log P_{X,Y}(X,Y \\
&= \mathbb{E}\left[\log\left(\frac{P_{X,Y}(X,Y)}{P_X(X)P_Y(Y)}\right)\right] \\
&= D\big(P_{X,Y} \parallel P_X P_Y\big).
\end{aligned}
$$

$\square$

Remark 5.3

- $I(X;Y)$ is symmetric in X and Y.
- The sum of the information contain in X and Y separately minus the information contained in the pair indeed is the amount of mutual information shared by both.
- Considering Stein's Lemma, we can consider $I(X;Y)$ as a measure of how well data generated from $P_{X,Y}$ can be distinguished from independent pairs (X', Y') generated by the product distribution $P_X P_Y$, so is a measure of how far X and Y are from being independent.

Proposition: Bounds On Mutual Information And Conditional Mutual Information

Proposition 5.4

- $0 \leq I(X;Y) \leq H(X)$ with equality to 0 iff X and Y are independent.
- Similarly, $I(X;Z \mid Y) \geq 0$ with equality iff $X - Y - Z$, i.e. X and Z are conditionally independent given Y.

Proof. First is by Proposition 5.2 and non-negativity of conditional entropy, second is an exercise. ∎

Proposition: Mutual Information Chain Rule

Proposition 5.5 (Chain Rule for Mutual Information) For all discrete RVs $X_1, ..., X_n, Y$,

$$I(X_1^n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_1^{i-1}).$$

Proof (Hints). Straighforward. $\square$

Proof. By the chain rule for entropy,

$$
\begin{aligned}
I(X_1^n; Y) &= H(X_1^n) - H(X_1^n \mid Y) \\
&= \sum_{i=1}^{n} H(X_i \mid X_1^{i-1}) - \sum_{i=1}^{n} H(X_i \mid X_1^{i-1}, Y) \\
&= \sum_{i=1}^{n} \big(H(X_i \mid X_1^{i-1}) - H(X_i \mid X_1^{i-1}, Y)\big) \\
&= \sum_{i=1}^{n} I(X_i; Y \mid X_1^{i-1}).
\end{aligned}
$$

$\square$

Theorem: Mutual Information Data Processing

Theorem 5.6 (Data Processing Inequalities for Mutual Information) If $X - Y - Z$ (so X and Z are conditionally independent given Y), then

$$I(X;Z), I(X;Y \mid Z) \leq I(X;Y).$$

Proof (Hints). Use chain rule for mutual information twice on the same expression. $\square$

Proof. By the chain rule, we have

$$\begin{aligned} I(X; Y, Z) &= I(X; Y) + I(X; Z \mid Y) \\ &= I(X; Z) + I(X; Y \mid Z). \end{aligned}$$

Now $I(X; Z \mid Y) = 0$ by conditional independence, so $I(X; Y) = I(X; Z) + I(X; Y \mid Z)$. $\square$

Example 5.7 We always have $X - Y - f(Y)$, hence $I(X; f(Y)) \leq I(X;Y)$, so applying a function to Y cannot make X and Y "less independent".

5.1. Synergy and redundancy

Note 5.8 $I(X;Y_1,Y_2)$ can greater than, equal to, or less than $I(X;Y_1)+(X;Y_2)$.

Definition: Synergy

Definition 5.9 The **synergy** of Y_1, Y_2 about X is

$$\begin{aligned} S(X; Y_1, Y_2) &= I(X; Y_1, Y_2) - (I(X; Y_1) + I(X; Y_2)) \\ &= I(X; Y_2 \mid Y_1) - I(X, Y_2). \end{aligned}$$

So the synergy can be < 0, > 0 or $= 0$.

Definition: Redundancy Orthogonality Synergisticy

Definition 5.10 If $S(X;Y_1,Y_2)$ is:

- negative, then Y_1 and Y_2 contain **redundant** information about X;
- zero, then Y_1 and Y_2 are **orthogonal**;
- positive, then Y_1 and Y_2 are **synergistic**. Intuitively, knowing Y_1 already makes the information in Y_2 more valuable (in that it gives more information about X).

Theorem: Condition For Correlated Observations Giving More Information Than Independent Ones

Theorem 5.11 Let RVs Y_1, Y_2 be conditionally independent given X, each with distribution $P_{Y \mid X}$, and RVs Z_1, Z_2 be distributed according to $Q_{Z \mid Y}(\cdot \mid Y_1), Q_{Z \mid Y}(\cdot \mid Y_2)$ respectively. Let RV Y have distribution $P_{Y \mid X}$, and W_1, W_2 be conditionally independent given Y, distributed according to $Q_{Z \mid Y}(\cdot \mid Y)$.

If $S(X; W_1, W_2) > 0$, then $I(X; W_1, W_2) > I(X; Z_1, Z_2)$, for independent Z_1 and Z_2, i.e. correlated observations are better than independent ones.

Proof (Hints). Use data processing for mutual information. $\square$

Proof. As in Definition 5.9, we have $I(X;W_2 \mid W_1) > I(X;W_2)$. $I(X;W_2) = I(X;Z_2)$ since (X,W_2) has the same joint distribution as (X,Z_2). By the data processing inequality, we have $I(X;Z_2 \mid Z_1) = I(Z_2;X \mid Z_1) \leq I(Z_2;X) = I(X;Z_2)$, since Z_1 and Z_2 are conditionally independent given X. Hence $I(X;W_2 \mid W_1) > I(X;Z_2 \mid Z_1)$, so $I(X;W_2 \mid W_1) + I(X;W_1) > I(X;Z_2 \mid Z_1) + I(X;Z_1)$, and the result follows by the chain rule. $\square$

Example 5.12 Given two equally noisy channels of a signal X, we want to decide whether it is better (gives more information about X) for the channels to be independent (this corresponds with choosing the Y_1, Y_2, Z_1, Z_2) or correlated (this corresponds with choosing the Y, W_1, W_2).

The natural assumption that the conditionally independent observations Z_1, Z_2 would be "better" than W_1, W_2 (i.e. $I(X; Z_1, Z_2) \geq I(X; W_1, W_2)$) is **false**. We can show diagramatically as

$$
\begin{tikzcd}
& X \arrow[dl] \arrow[dr] & \\
\mathrm{P}(y \mid x) \arrow[d] & & P(y \mid x) \arrow[d] \\
Y_1 \arrow[d] & & Y_2 \arrow[d] \\
Q(z \mid y_1) \arrow[d] & & Q(z \mid y_2) \arrow[d] \\
Z_1 & & Z_2
\end{tikzcd}
\qquad
\begin{tikzcd}
& X \arrow[d] & \\
& \mathrm{P}(y \mid x) \arrow[d] & \\
& Y \arrow[dl] \arrow[dr] & \\
Q(z \mid y) \arrow[d] & & Q(z \mid y) \arrow[d] \\
W_1 & & W_2
\end{tikzcd}
$$

Example 5.13 For example, let $P_{Y \mid X}$ be the Z-channel: if $X = 0$, then $Y = 0$ with probability 1, and if $X = 1$, then $Y \sim \text{Bern}(1 - \delta)$ for some $\delta \in (0, 1)$. Let $Q_{Z \mid Y}$ be a binary symmetric channel: given Y taking values in $0, 1$, $Z = Y$ with probability $1 - \varepsilon$, and $Z = 1 - Y$ with probability ε for some $\varepsilon \in (0, 1)$. We can represent this as

$$
\begin{tikzcd}
0 \arrow[r, "1"] & 0 \\
1 \arrow[ur, "\delta"] \arrow[r, "1-\delta"] & 1
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
0 \arrow[r, "1-\varepsilon"] \arrow[dr, "\varepsilon"'] & 0 \\
1 \arrow[ur, "\varepsilon"] \arrow[r, "1-\varepsilon"] & 1
\end{tikzcd}
$$

If $X \sim \mathrm{Bern}(1/2)$, $\delta = 0.85$ and $\varepsilon = 0.1$, then $I(X; W_1, W_2) \approx 0.047 > I(X; Z_1, Z_2) \approx 0.039$. So the correlated observations W_1, W_2 are better than the independent observations Z_1, Z_2.

6. Entropy and additive combinatorics

6.1. Simple sumset entropy bounds

Definition: Sumset

Definition 6.1 For $A, B \subseteq \mathbb{Z}$ the **sumset** of A and B is

$$A + B := \{a + b : a \in A, b \in B\}.$$

Definition: Difference Set

Definition 6.2 For $A, B \subseteq \mathbb{Z}$ the **difference set** of A and B is

$$A - B := \{a - b : a \in A, b \in B\}.$$

Proposition: Sumset Bounds

Proposition 6.3 Let $A, B \subseteq \mathbb{Z}$ be finite. Then

$$\max\{|A|, |B|\} \leq |A + B| \leq |A||B|.$$

Proof (Hints). Trivial. $\square$

Proof. Trivial. $\square$

Proposition: Ruzsa Triangle Inequality

Proposition 6.4 (Ruzsa Triangle Inequality) Let $A, B, C \subseteq \mathbb{Z}$ be finite. Then

$$|A - C| \cdot |B| \leq (|A - B||B - C|).$$

Proof (Hints). Show that an appropriate function is injective. $\square$

Proof. Fix a presentation $y = a_y - c_y$ (where $a_y \in A, c_y \in C$) for each $y \in A - C$. Let

$$\begin{aligned} f : B \times (A - C) &\to (A - B) \times (B - C) \\ (b, y) &\mapsto \big(a_y - b, b - c_y\big). \end{aligned}$$

If $f(b, y) = f(b', y')$, then $a_{y'} - b' = a_y - b$ and $b' - c_{y'} = b - c_y$. So $a_y - a_{y'} = b - b' = c_y - c_{y'}$. So $y = a_y - c_y = a_{y'} - c_{y'} = y'$. Hence $a_y = a_{y'}$, and so $b = b'$. So f is injective, so $|B \times (A - C)| \leq |(A - B) \times (B - C)|$. $\square$

Remark 6.5 If X_1^n is a large collection of IID RVs with common PMF P on alphabet A, then the Asymptotic Equipartition Property (AEP) tells us that we can concentrate on the 2^{nH} typical strings. $2^{nH} = (2^H)^n$ is typically much smaller than all $|A|^n = \left(2^{\log|A|}\right)^n$ strings. We can think of $(2^H)^n$ as the effective support size of P^n, and can of 2^H as the effective support size of a single RV with entropy H.

Remark 6.6 We can use the above interpretation to obtain useful conjectures about bounds for the entropy of discrete RVs, from corresponding results on bounds on sumsets. We start with a sumset bound, then replace subsets of $\mathbb{Z}$ by independent RVs on $\mathbb{Z}$, and replace $\log|A|$ of each set A by the entropy of the corresponding RV.

Proposition: Sum Entropy Bounds

Proposition 6.7 Let X and Y are independent RVs on alphabet $\mathbb{Z}$, then

$$\max\{H(X), H(Y)\} \leq H(X+Y) \leq H(X) + H(Y).$$

Proof (Hints).

- For lower bound, show that $H(X) \leq H(X+Y)$ using data processing and similarly for $H(Y)$. The upper bound should follow directly from this calculation.

$\square$

Proof. For the lower bound,

$$\begin{aligned}
H(X)+H(Y) &= H(X,Y) && \text{by Chain Rule for Entropy} \\
&= H(Y, X+Y) && \text{by Data Processing} \\
&= H(X+Y)+H(Y \mid X+Y) && \text{by Chain Rule for Entropy} \\
&\leq H(X+Y)+H(Y) && \text{by Conditioning Reduces Entropy}
\end{aligned}$$

Note we have equality for data processing, since $(x,y) \mapsto (x, x+y)$ is injective. Hence $H(X+Y) \geq H(X)$, and the same argument shows that $H(X+Y) \geq H(Y)$.

For the upper bound, we have $H(X) + H(Y) = H(X+Y) + H(Y \mid X+Y) \geq H(X+Y)$ by non-negativity of conditional entropy. $\square$

Lemma: Entropy Ruzsa Triangle Inequality Lemma

Lemma 6.8 Let X, Y, Z be independent RVs on alphabet $\mathbb{Z}$. Then

$$H(X-Z)+H(Y) \leq H(X-Y, Y-Z).$$

Proof (Hints).

- Show that $I(X; X - Z) \leq I(X; (X - Y, Y - Z))$ using the Chain Rule for mutual information.
- Rewrite both sides of the above inequality in terms of entropies, using Data Processing.

$\square$

Proof. Since $X - Z = (X - Y) + (Y - Z)$, X and $X - Z$ are conditionally independent given $(X - Y, Y - Z)$ by Note 3.10. Thus by Data Processing for mutual information, we have $I(X; (X - Y, Y - Z)) \geq I(X; X - Z)$. Now

$$\begin{aligned} I(X; X - Z) &= H(X - Z) - H(X - Z \mid X) \\ &= H(X - Z) - H(Z \mid X) = H(X - Z) - H(Z) \end{aligned}$$

by Data Processing (since, given $X = x$, $x - z \mapsto z$ is injective), and independence of X and Z. Also,

$$I(X; (X - Y, Y - Z)) = H(X - Y, Y - Z) + H(X) - H(X, X - Y, Y - Z)$$

$$
\begin{aligned}
&= H(X-Y, Y-Z) + H(X) - H(X,Y,Z) \\
&= H(X-Y, Y-Z) - H(Y) - H(Z)
\end{aligned}
$$

by Data Processing (since $(x, x-y, y-z) \mapsto (x,y,z)$ is injective), and independence of X, Y and Z. $\square$

Theorem: Entropy Ruzsa Triangle Inequality

Theorem 6.9 (Ruzsa Triangle Inequality for Entropy) Let X, Y, Z be independent RVs on alphabet $\mathbb{Z}$. Then

$$H(X - Z) + H(Y) \leq H(X - Y) + H(Y - Z).$$

Proof (Hints). By above lemma. $\square$

Proof. By the above lemma, we have

$$\begin{aligned}
H(X - Z) + H(Y) &\leq H(X - Y, Y - Z) \\
&= H(X - Y) + H(Y - Z \mid X - Y) \quad \text{by Chain Rule for Entrop} \\
&\leq H(X - Y) + H(Y - Z).
\end{aligned}$$

by Conditioning Reduces Entropy. $\square$

6.2. The doubling-difference inequality for entropy

Definition: Entropy Increase

Definition 6.10 For IID RVs X_1, X_2 on alphabet $\mathbb{Z}$, the **entropy-increase** due to addition (Δ^+) or subtraction (Δ^-) is

$$\Delta^+ := H(X_1 + X_2) - H(X_1),$$
$$\Delta^- := H(X_1 - X_2) - H(X_1).$$

Proposition: Mutual Information Expression For Entropy Increase

Proposition 6.11 For IID X_1, X_2 on $\mathbb{Z}$, we have

$$\Delta^+ = I(X_1 + X_2; X_2),$$
$$\Delta^- = I(X_1 - X_2; X_2).$$

Proof (Hints). Straightforward. $\square$

Proof. We have

$$\begin{aligned} I(X_1 + X_2; X_2) &= H(X_1 + X_2) + H(X_2) - H(X_1 + X_2, X_2) \\ &= H(X_1 + X_2) + H(X_2) - H(X_1, X_2) \\ &= H(X_1 + X_2) + H(X_2) - H(X_1) - H(X_2) \end{aligned}$$

by Data Processing (since $(x_1 + x_2, x_2) \mapsto (x_1, x_2)$ is injective) and Chain Rule for Entropy. The proof is identical for Δ^-. □

Lemma: Triple Sum Reduces Entropy

Lemma 6.12 Let X, Y, Z be independent RVs on alphabet $\mathbb{Z}$. Then

$$H(X+Y+Z) + H(Y) \leq H(X+Y) + H(Y+Z).$$

Proof (Hints).

- Show that $I(X; X + Y + Z) \leq I(X + Y; X)$.
- Rewrite both sides in terms of entropies.

$\square$

Proof. Since $X - (X+Y, Z) - (X+Y+Z)$ form a Markov chain by Note 3.10, we have, by Data Processing and Chain Rule for mutual information,

$$\begin{aligned} I(X; X+Y+Z) &\leq I(X+Y, Z; X) = I(X+Y; X) + I(Z; X \mid X+Y). \\ &= I(X+Y; X) \end{aligned}$$

since Z is (conditionally) independent of X given $X+Y$. Now

$$\begin{aligned} I(X+Y; X) &= H(X+Y) + H(X) - H(X+Y, X) \\ &= H(X+Y) + H(X) - H(Y, X) \\ &= H(X+Y) + H(X) - H(Y) - H(X) \end{aligned}$$

$$= H(X+Y) - H(Y)$$

since $(y, x) \mapsto (x+y, x)$ is injective and X and Y are independent. Also,

$$\begin{aligned}
I(X+Y+Z; X) &= H(X+Y+Z) + H(X+Y+Z \mid X) \\
&= H(X+Y+Z) - H(Y+Z \mid X) \\
&= H(X+Y+Z) - H(Y+Z)
\end{aligned}$$

since, given $X = x$, $x+y+z \mapsto y+z$ is injective, and X and $Y+Z$ are independent. $\square$

Theorem: Doubling Difference Inequality

Theorem 6.13 (Doubling-difference Inequality) Let X_1 and X_2 be IID RVs on $\mathbb{Z}$. Then

$$\frac{1}{2} \leq \frac{\Delta^+}{\Delta^-} \leq 2.$$

Proof (Hints).

- For lower bound, use Ruzsa Triangle Inequality for appropriate RVs.
- For upper bound, use Lemma 6.12 and Proposition 6.7.

$\square$

Proof. For the lower bound, let $X, -Y, Z$ be IID with the same distribution as X_1. Then by the Ruzsa Triangle Inequality,

$$H(X_1 - X_2) + H(X_1) \leq H(X_1 + X_2) + H(X_1 + X_2).$$

So $2(H(X_1 + X_2) - H(X_1)) \geq H(X_1 - X_2) - H(X_1)$.

For the upper bound, let $X, -Y, Z$ be IID with the same distribution as X_1. Then by the above lemma and Proposition 6.7,

$$H(X_1 + X_2) + H(X_1) \leq H(X_1 - X_2) + H(X_1 - X_2)$$

so $H(X_1 + X_2) - H(X_1) \leq 2(H(X_1 - X_2) - H(X_1))$. $\square$

7. Entropy rate

Definition: Entropy Rate

Definition 7.1 For an arbitrary source $\boldsymbol{X} = \{X_n : n \in \mathbb{N}\}$, the **entropy rate** $H(\boldsymbol{X})$ of $\boldsymbol{X}$ is the limit of the average number of bits per symbol:

$$H(\boldsymbol{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1^n)$$

whenever the limit exists.

Example 7.2 If $\boldsymbol{X}$ is memoryless (so a sequence of IID RVs) with common entropy $H = H(X_i)$, then the entropy rate is

$$H(\boldsymbol{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1^n) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i) = H.$$

Example 7.3 Let $\boldsymbol{X} = \{X_n : n \in \mathbb{N}\}$ be an irreducible, aperiodic Markov chain on a finite alphabet A with transition matrix Q, where

$$Q_{ab} = \mathbb{P}(X_{n+1} = b \mid X_n = a), \quad \forall a, b \in A$$

Let $X_1 \sim P_{X_1}$ be the initial distribution and π be the unique stationary distribution ($\mathbb{P}(X_n = x) \to \pi(x)$ as $n \to \infty$). $\boldsymbol{X}$ has a unique invariant distribution π to which it converges:

$$\forall x \in A, \quad \mathbb{P}(X_n = x) \to \pi(x) \quad \text{as } n \to \infty$$

and hence also

$$\mathbb{P}(X_{n-1} = x, X_n = y) = \mathbb{P}(X_n = x)Q_{xy} \to \pi(x)Q_{xy}.$$

Then by the Chain Rule for Entropy and conditional independence,

$$\begin{aligned} H(X_1^n) &= \sum_{i=1}^{n} H(X_i \mid X_1^{i-1}) \\ &= H(X_1) + \sum_{i=2}^{n} H(X_i \mid X_{i-1}) \\ &= H(X_1) - H(X_{n+1} \mid X_n) + \sum_{i=1}^{n} H(X_{i+1} \mid X_i). \end{aligned}$$

By the convergence theorem for Markov chains, we have $P_{X_n} \to \pi$ as $n \to \infty$. $H(X \mid Y)$ is a continuous function of the joint distribution $P_{X,Y}$, so $H(X_n \mid X_{n-1}) \to H\left(\overline{X_1} \mid \overline{X_0}\right)$ as $n \to \infty$, where $\overline{X_0} \sim \pi$ and $\mathbb{P}\left(\overline{X_1} = b \mid \overline{X_1} = a\right) = Q_{ab}$. We have

$$\frac{1}{n}H(X_1^n) = \frac{1}{n}(H(X_1) - H(X_{n+1} \mid X_n)) + \frac{1}{n}\sum_{i=1}^{n} H(X_{i+1} \mid X_i)$$

The first term tends to 0 since the numerator is bounded, and the summands in the second term tend to $H\left(\overline{X_1} \mid \overline{X_0}\right)$. So the entropy rate exists and is equal to $H(\boldsymbol{X}) = H\left(\overline{X_1} \mid \overline{X_0}\right)$.

Definition: Source.Stationary

Definition 7.4 A source $\boldsymbol{X}$ is **stationary** if for any block length $n \in \mathbb{N}$, the distribution of X_{k+1}^{k+n} is independent of k.

Remark 7.5 If $\boldsymbol{X} = \{X_n : n \in \mathbb{N}\}$ is one-sided stationary process, then by Kolmogorov's extension theorem, $\boldsymbol{X}$ admits a unique two-sided extension to $\boldsymbol{X} = \{X_n : n \in \mathbb{Z}\}$.

Lemma: Cesaro

Lemma 7.6 (Cesàro) Let (a_n) be a sequence. The $\boldsymbol{n}$**-th Cesaro mean** is defined as

$$\sigma_n = \frac{1}{n}\sum_{k=1}^{n} a_k.$$

If (a_n) has limit L, then

$$\lim_{n\to\infty} \sigma_n = \lim_{n\to\infty} a_n = L.$$

Theorem: Entropy Rate Of Stationary Source

Theorem 7.7 If $\boldsymbol{X} = \{X_n : n \in \mathbb{N}\}$ is a stationary process on finite alphabet A, then its entropy rate exists and is equal to

$$H(\boldsymbol{X}) = \lim_{n\to\infty} H\big(X_n \mid X_1^{n-1}\big).$$

Proof (Hints). Show that the sequence $\{H(X_n) \mid X_1^{n-1} : n \in \mathbb{N}\}$ is non-increasing and use the Cesàro Lemma. $\square$

Proof. The sequence $\{H(X_n) \mid X_1^{n-1} : n \in \mathbb{N}\}$ is non-negative by non-negativity of conditional entropy, and is non-increasing, since

$$\begin{aligned}
H(X_{n+1} \mid X_1^n) &\leq H(X_{n+1} \mid X_2^n) && \text{by Conditioning Reduces Entropy} \\
&= H\big(X_2^{n+1}\big) - H(X_2^n) && \text{by Chain Rule for Entropy} \\
&= H(X_1^n) - H\big(X_1^{n-1}\big) && \text{by stationarity} \\
&= H\big(X_{n-1} \mid X_1^{n-2}\big) && \text{by Chain Rule for Entropy.}
\end{aligned}$$

Hence the limit $\lim_{n\to\infty} H\big(X_n \mid X_1^{n-1}\big)$ exists, and so by the Cèsaro Lemma, the averages converge to the same limit. But by the Chain Rule for Entropy, the averages are

$$\frac{1}{n}\sum_{i=1}^{n} H(X_i \mid X_1^{i-1}) = \frac{1}{n}H(X_1^n).$$

$\square$

Theorem: Entropy Rate Is Conditional Entropy Given Infinite Past

Theorem 7.8 For a stationary process $\boldsymbol{X} = \{X_n : n \in \mathbb{Z}\}$ on a finite alphabet A,

$$H(\boldsymbol{X}) = \lim_{n\to\infty} H\big(X_0 \mid X_{-n}^{-1}\big) = H\big(X_0 \mid X_{-\infty}^{-1}\big).$$

Proof (Hints). Non-examinable. $\square$

Proof. By Martingale convergence, we have that

$$P(x_0 \mid X_{-n}^{-1}) \to P(x_0 \mid X_{-\infty}^{-1}) \quad \text{almost surely} \quad \text{as } n \to \infty,$$

where $P(\cdot \mid x_{-n}^{-1})$ is the conditional distribution of X_0 given $X_{-n}^{-1} = x_{-n}^{-1}$, and $P(\cdot \mid x_{-\infty}^{-1})$ is the conditional distribution of X_0 given $X_{-\infty}^{-1} = x_{-\infty}^{-1}$. Now, we can take expectations to obtain that, by the bounded convergence theorem (since $p \mapsto p \log p$ is continuous and bounded for $p \in [0, 1]$),

$$H(X_0 \mid X_{-n}^{-1}) = \mathbb{E}\left[-\sum_{x_0 \in A} P(x_0 \mid X_{-n}^{-1}) \log P(x_0 \mid X_{-n}^{-1})\right]$$

$$
\begin{aligned}
&\to \mathbb{E}\left[-\sum_{x_0 \in A} P(x_0 \mid X_{-\infty}^{-1}) \log P(x_0 \mid X_{-\infty}^{-1})\right] \\
&=: H(X_0 \mid X_{-\infty}^{-1}) \quad \text{almost surely} \quad \text{as } n \to \infty.
\end{aligned}
$$

Finally, $H(X_0 \mid X_{-n}^{-1}) = H(X_{n+1} \mid X_1^n)$ by stationarity, so we are done by Theorem 7.7. $\square$

Definition: Source.Ergodic

Definition 7.9 Let $\boldsymbol{X} = \{X_n : n \in \mathbb{Z}\}$ be a stationary source on finite alphabet A, and define the (left) **shift** operator $T : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ on sequences $A^{\mathbb{Z}}$ by

$$(Tx)_n = x_{n+1} \quad \forall n \in \mathbb{Z}.$$

$\boldsymbol{X}$ is **ergodic** if all shift invariant events are trivial, i.e. for any measurable $B \subseteq A^{\mathbb{Z}}$, we have

$$T^{-1}B = B \Longrightarrow \mathbb{P}(X_{-\infty}^{\infty} \in B) = 0 \text{ or } 1.$$

Intuitively, an ergodic process is one which satisfies the general form of the strong law of large numbers.

It turns out that ergodicity is equivalent to the validity of the following:

Theorem: Birkhoff

Theorem 7.10 (Birkhoff's Ergodic Theorem) Let $\boldsymbol{X} = \{X_n : n \in \mathbb{Z}\}$ be a stationary ergodic source on alphabet A. Then for any measurable function $f : A^{\mathbb{Z}} \to \mathbb{R}$ such that

$$\mathbb{E}[|f(X_{-\infty}^{\infty})|] < \infty,$$

we have

$$\frac{1}{n}\sum_{i=1}^{n} f(T^i X_{-\infty}^{\infty}) \to \mathbb{E}[f(X_{-\infty}^{\infty})] \quad \text{almost surely} \quad \text{as } n \to \infty$$

Proof (Hints). Beyond the scope of this course. $\square$

Proof. Omitted. $\square$

Remark 7.11 The strong law of large numbers follows instantly from Birkhoff by setting $f(x_{-\infty}^{\infty}) = x_1$.

Example 7.12 Every IID source is ergodic.

Theorem: Shannon Mcmillan Breiman

Theorem 7.13 (Shannon-McMillan-Breiman) Let $\boldsymbol{X} = \{X_n : n \in \mathbb{N}\}$ be a stationary ergodic source on alphahbet A with entropy rate $H = H(\boldsymbol{X})$, then

$$-\frac{1}{n} \log P_n(X_1^n) \to H \text{ almost surely as } n \to \infty$$

where P_n is the PMF of X_1^n.

Proof (Hints). Non-examinable. $\square$

Proof. Idea: by Chain Rule for Entropy, we have

$$-\frac{1}{n}\log P_n(X_1^n) = -\frac{1}{n}\log\prod_{i=1}^{n} P(X_i \mid X_1^{i-1}) = \frac{1}{n}\sum_{i=1}^{n}[-\log P(X_i \mid X_1^{i-1})]$$

but we cannot directly apply the ergodic theorem to this, since $-\log P(X_i \mid X_1^{i-1})$ is not of the form $f(T^i x_{-\infty}^{\infty})$. Instead, note that by Birkhoff's Ergodic Theorem and Theorem 7.8,

$$\begin{aligned}
-\frac{1}{n}\log P(X_1^n \mid X_{-\infty}^0) &= \frac{1}{n}\sum_{i=1}^{n}[-\log P(X_i \mid X_{-\infty}^{i-1})] \\
&\to \mathbb{E}[-\log P(X_0 \mid X_{-\infty}^{-1})]
\end{aligned}$$

$$=: H(X_0 \mid X_{-\infty}^{-1}) = H \text{ almost surely} \quad \text{as } n \to \infty.$$

Also, by Birkhoff's Ergodic Theorem, for each fixed $k \geq 1$,

$$\frac{1}{n}\sum_{i=1}^{n}(-\log P(X_i \mid X_{i-k}^{i-1})) \to \mathbb{E}[-\log P(X_0 \mid X_{-k}^{-1})]$$

$$=: H(X_0 \mid X_{-k}^{-1}) \text{ almost surely} \quad \text{as } n \to \infty.$$

We have

$$\mathbb{P}\left(-\frac{1}{n}\log P(X_1^n \mid X_{-\infty}^0) - \left(-\frac{1}{n}\log P_n(X_1^n)\right) > \varepsilon\right) = \mathbb{P}\left(\frac{1}{n}\log\frac{P_n(X_1^n)}{P(X_1^n \mid X_{-\infty}^0)} >\right.$$

$$\mathbb{P}\left(\frac{P_n(X_1^n)}{P(X_1^n \mid X_{-\infty}^0)} > 2^{n\varepsilon}\right)$$

$$2^{-n\varepsilon}\mathbb{E}\left[\frac{P_n(X_1^n)}{P(X_1^n \mid X_{-\infty}^0)}\right] \quad \text{by markov's inequality}$$

$$2^{-n\varepsilon}\mathbb{E}\left[\mathbb{E}\left[\frac{P_n(X_1^n)}{P(X_1^n \mid X_{-\infty}^0)} \mid X_{-\infty}^0\right]\right]$$

$$2^{-n\varepsilon}\mathbb{E}\left[\sum_{\substack{x_1^n \\ P(x_1^n \mid X_{-\infty}^0)>0}} P(x_1^n \mid X_{-\infty}^0)\frac{P_n(x_1^n)}{P(x_1^n \mid X_{-\infty}^0)}\right]$$

$2^{-n\varepsilon}$

which is summable, so by Borel-Cantelli,

$$\liminf_{n\to\infty} -\frac{1}{n} \log P\big(X_1^n \mid X_{-\infty}^0\big) \leq \liminf_{n\to\infty} -\frac{1}{n} \log P_n(X_1^n) \text{ almost surely.}$$

For each fixed k, consider the sequence of PMFs $Q_n^{(k)}(x_1^n) = P_k\big(x_1^k\big) \prod_{i=k+1}^n P\big(x_i \mid X_{i-k}^{i-1}\big)$ for $x_1^n \in A^n$. Then

$$-\frac{1}{n} \log Q_n^{(k)}(X_1^n) - \left[-\frac{1}{n} \sum_{i=1}^n \log P\big(x_i \mid x_{i-k}^{i-1}\big)\right]$$

$$= -\frac{1}{n}\left[\log P_k\big(x_1^k\big) - \sum_{i=1}^{k} \log P\big(X_i \mid X_{i-k}^{i-1}\big)\right]$$

$$\to 0 \text{ almost surely as } n \to \infty$$

So suffices to show that $\limsup_{n\to\infty} -\frac{1}{n}\log P_n(X_1^n) \leq \limsup_{n\to\infty} -\frac{1}{n}\log Q_n^{(k)}(X_1^n)$ almost surely. So again, let $\varepsilon > 0$ be arbitrary, then

$$\mathbb{P}\bigg(-\frac{1}{n}\log P_n(X_1^n) - \Big(-\frac{1}{n}\log Q_n^{(k)}(X_1^n)\Big) > \varepsilon\bigg)$$

$$= \mathbb{P}\left(\frac{Q_n^{(k)}(X_1^n)}{P_n(X_1^n)} > 2^{n\varepsilon} \right) \leq 2^{-n\varepsilon} \mathbb{E}\left[\frac{Q_n^{(k)}(X_1^n)}{P_n(X_1^n)} \right] \text{ by Markov's inequality}$$

$$\leq 2^{-n\varepsilon} \sum_{x_1^n \in A^n} P_n(x_1^n) \frac{Q_n^{(k)}(x_1^n)}{P_n(x_1^n)} = 2^{-n\varepsilon}$$

which is summable, so by Borel-Cantelli and the fact that $\varepsilon > 0$ was arbitrary, we have

$$\limsup_{n\to\infty} -\frac{1}{n} \log P_n(X_1^n) \leq \limsup_{n\to\infty} -\frac{1}{n} \sum_{i=1}^{n} \log P\big(X_i \mid X_{i-k}^{i-1}\big).$$

8. Types and large deviations

8.1. The method of types

Definition: Type

Definition 8.1 Let A be a finite alphabet and $x_1^n \in A^n$. The **type** of x_1^n is its empirical distribution $\widehat{P}_n = \widehat{P}_{x_1^n}$:

$$\widehat{P}_n(a) = \widehat{P}_{x_1^n}(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i = a\}}.$$

Notation 8.2 For a finite alphabet $A = \{a_1, ..., a_m\}$, let $\mathcal{P}$ denote the set of all PMFs on A:

$$\mathcal{P} = \left\{ P \in [0,1]^m : \sum_{a \in A} P(a) = 1 \right\}.$$

Note that $\mathcal{P}$ is an m-simplex.

Definition: N Types

Notation 8.3 We write $\mathcal{P}_n$ for the set of all ***n*-types**:

$$\mathcal{P}_n = \{P \in \mathcal{P} : nP(a) \in \mathbb{Z} \ \forall a \in A\}.$$

Note that $\mathcal{P}_n$ is finite.

Proposition: Upper Bound On Number Of N Types

Proposition 8.4 We have $|\mathcal{P}_n| \leq (n+1)^m$.

Proof (Hints). Straightforward. $\square$

Proof. Each $P \in \mathcal{P}_n$ is of the form $(k_1/n, ..., k_m/n)$. There are at most $(n+1)$ choices $(0, ..., n)$ for each k_i. $\square$

Proposition: Prob Under Prod Dist Of String Of Type

Proposition 8.5 Let $x_1^n \in A^n$ have type $\hat{P}_n$. Then for any PMF Q,

$$Q^n(x_1^n) = 2^{-n(H(\hat{P}_n)+D(\hat{P}_n \,\|\, Q))}.$$

In particular, if $Q = \hat{P}_n$, then $Q^n(x_1^n) = 2^{-nH(Q)}$.

Proof (Hints). Rewrite $\log Q^n(x_1^n)$. $\square$

Proof. We have

$$\begin{aligned}
\log Q^n(x_1^n) &= \sum_{i=1}^{n} \log Q(x_i) \\
&= \sum_{i=1}^{n} \sum_{a \in A} \mathbb{1}_{\{x_i = a\}} \log Q(a) \\
&= n \sum_{a \in A} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i = a\}} \log Q(a) \\
&= n \sum_{a \in A} \hat{P}_n(a) \log Q(a) = -\sum_{a \in A} \hat{P}_n(a) \log \left(\frac{\hat{P}_n(a)}{Q(a)} \frac{1}{\hat{P}_n(a)} \right)
\end{aligned}$$

$$
\begin{aligned}
&= -n\left(\sum_{a\in A}\hat{P}_n(a)\log\frac{\hat{P}_n(a)}{Q(a)} + \sum_{a\in A}\hat{P}_n(a)\log\frac{1}{\hat{P}_n(a)}\right) \\
&= -n(D(\hat{P}_n \parallel Q) + H(\hat{P}_n))
\end{aligned}
$$

$\square$

Definition: Type Class

Definition 8.6 Given a n-type P, its **type class** is

$$T(P) := \left\{ x_1^n \in A^n : \hat{P}_{x_1^n} = P \right\}.$$

Note that $A^n = \coprod_{P \in \mathcal{P}_n} T(P)$: since A^n has size $|A|^n$ exponential in n, and the union is over $|\mathcal{P}_n| \leq (n+1)^m$ (polynomial in n) elements, at least one type class must contain exponentially many strings.

$T(P)$ consists of all possible arrangements of $nP(a_1)$ a_1's, ..., $nP(a_m)$ a_m's, so

$$|T(P)| = \frac{n!}{\prod_{j=1}^{m} \big(nP(a_j)\big)!}.$$

Lemma: Most Likely Type Class Under Prod Dist Is Type Class Of That Dist

Lemma 8.7 Let $P \in \mathcal{P}_n$. Then

$$P^n(T(P)) = \max\{P^n(T(Q)) : Q \in \mathcal{P}_n\}.$$

i.e. the most likely type class under P^n is $T(P)$.

Proof (Hints).

- For $Q \in \mathcal{P}_n$, find an expression for $P^n(x_1^n)$ which should be independent of x_1^n, for each case $x_1^n \in T(P)$ and $x_1^n \in T(Q)$.
- Show that $\frac{P^n(T(P))}{P^n(T(Q))} \geq 1$, using the fact that $k!/\ell! \geq \ell^{k-\ell}$ (why?).

$\square$

Proof. Let $Q \in \mathcal{P}_n$ be arbitrary. Then

$$\begin{aligned}
\frac{P^n(T(P))}{P^n(T(Q))} &= \frac{|T(P)| \cdot \prod_{i=1}^m P(a_i)^{nP(a_i)}}{|T(Q)| \cdot \prod_{i=1}^m P(a_i)^{nQ(a_i)}} \\
&= \frac{n!}{\prod_{i=1}^m (nP(a_i))!} \cdot \frac{\prod_{i=1}^m (nQ(a_i))!}{n!} \cdot \prod_{i=1}^m P(a_i)^{n(P(a_i)-Q(a_i))} \\
&= \prod_{i=1}^m P(a_i)^{n(P(a_i)-Q(a_i))} \cdot \prod_{i=1}^m \frac{(nQ(a_i))!}{(nP(a_i))!}.
\end{aligned}$$

Now since $k!/\ell! \geq \ell^{k-\ell}$ (to show this, consider $k \geq \ell$ and $k < \ell$ cases separately), this is

$$
\begin{aligned}
&\geq \prod_{i=1}^{m} P(a_i)^{n(P(a_i)-Q(a_i))} \cdot \prod_{i=1}^{m} (n(P(a_i)))^{n(Q(a_i)-P(a_i))} \\
&= \prod_{i=1}^{m} n^{n(Q(a_i)-P(a_i))} \\
&= n^{n\sum_{i=1}^{m}(Q(a_i)-P(a_i))} = 1
\end{aligned}
$$

since probabilities sum to 1. $\square$

Proposition: Bounds On Size Of Type Class

Proposition 8.8 Let $|A| = m$. For any n-type $P \in \mathcal{P}_n$,

$$(n+1)^{-m} 2^{nH(P)} \leq |T(P)| \leq 2^{H(P)}.$$

Proof (Hints). Straightforward. ∎

Proof. By Proposition 8.5, we have $1 \geq P^n(T(P)) = |T(P)|2^{-nH(P)}$. For the lower bound,

$$\begin{aligned}
1 &= \sum_{x_1^n \in A^n} P^n(x_1^n) \\
&= \sum_{Q \in \mathcal{P}_n} P^n(T(Q)) \\
&\leq |\mathcal{P}_n| P^n(T(P)) && \text{by Lemma 8.7} \\
&\leq (n+1)^m |T(P)| 2^{-nH(P)}.
\end{aligned}$$

$\square$

Corollary: Bounds On Probability Of Type Class

Corollary 8.9 For any n-type $P \in \mathcal{P}_n$ and any PMF Q on A,

$$(n+1)^{-m} 2^{-nD(P \parallel Q)} \le Q^n(T(P)) \le 2^{-nD(P \parallel Q)}.$$

Proof (Hints). Straightforward. $\square$

Proof. Let $x_1^n \in T(P)$ be arbitrary. Then by Proposition 8.5,

$$Q^n(T(P)) = |T(P)|Q^n(x_1^n) = |T(P)|2^{-n(H(P)+D(P \parallel Q))}.$$

So we are done by Proposition 8.8. $\square$

8.2. Sanov's theorem

Theorem: Sanov

Theorem 8.10 (Sanov) Let X_1^n be IID with common PMF Q which has full support on alphabet A (i.e. $Q(a) > 0$ for all $a \in A$) with $|A| = m$. Let $\hat{P}_n$ be the empirical distribution of X_1^n. For all $E \subseteq \mathcal{P}$,

$$\mathbb{P}(\hat{P}_n \in E) \leq (n+1)^m 2^{-nD^*}.$$

where $D^* = \inf\{D(P \parallel Q) : P \in E\}$. Also, if $E = \overline{\mathrm{int}(E)}$ is equal to the closure of its interior, then

$$\lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}(\hat{P}_n \in E) = D^* = D(P^* \parallel Q),$$

where $P^* \in E$.

Proof (Hints).

- For the inequality, use that $\mathbb{P}(\hat{P}_n \in E) = \mathbb{P}(\hat{P}_n \in E \cap \mathcal{P}_n) = \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P))$. Explain why D^* is finite.
- For the equality, use the above inequality, and explain why there is a sequence $\{P_n : n \in \mathbb{N}\}$ with each $P_n \in \mathcal{P}_n$ and $P_n \to P^*$ where $D(P^* \parallel Q) = D^*$ (why does P^* exist?)

$\square$

Proof. Since Q has full support, for any $P \in \mathcal{P}$, we have $D(P \parallel Q) \leq -\sum_{a \in A} \log Q(a) < \infty$, so D^* is finite. For the upper bound,

$$\begin{aligned}
\mathbb{P}(\hat{P}_n \in E) &= \mathbb{P}(\hat{P}_n \in E \cap \mathcal{P}_n) \\
&= \sum_{P \in E \cap \mathcal{P}_n} \mathbb{P}(\hat{P}_n = P) \\
&= \sum_{P \in E \cap \mathcal{P}_n} \mathbb{P}(X_1^n \in T(P)) \\
&= \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \\
&\leq |E \cap \mathcal{P}_n| \max\{Q^n(T(P)) : P \in E \cap \mathcal{P}_n\}
\end{aligned}$$

$$
\begin{aligned}
&\leq |E \cap \mathcal{P}_n| \max\left\{2^{-nD(P \,\|\, Q)} : P \in E \cap \mathcal{P}_n\right\} \quad \text{by Corollary 8.9} \\
&= |E \cap \mathcal{P}_n| \cdot 2^{-n \min\{D(P \,\|\, Q) : P \in E \cap \mathcal{P}_n\}} \\
&\leq (n+1)^m \cdot 2^{-nD^*}.
\end{aligned}
$$

So $\liminf_{n\to\infty} -\frac{1}{n} \log Q^n(\hat{P}_n \in E) \geq D^*$.

For the lower bound, since E is compact and $D(P \,\|\, Q)$ is continuous in P, the infimum D^* is attained by some P^*. (Note that since $\mathcal{P}$ itself is compact, there is always a minimising P^* but this is not necessarily in E). Also, note that $\bigcup_{n\in\mathbb{N}} \mathcal{P}_n$ is dense in $\mathcal{P}$, so we can find a sequence

$\{P_n : n \in \mathbb{N}\} \subseteq E$ such that each $P_n \in \mathcal{P}_n$ and $P_n \to P^*$ (as a vector). Now for each $n \in \mathbb{N}$,

$$\mathbb{P}(\hat{P}_n \in E) \geq \mathbb{P}(\hat{P}_n = P_n) = Q^n(T(P_n)) \geq (n+1)^{-m} 2^{-nD(P_n \,\|\, Q)}$$

by Corollary 8.9. We have $D(P_n \,\|\, Q) \to D(P^* \,\|\, Q)$ as $n \to \infty$ since $D(P \,\|\, Q)$ is continuous in P. So $\limsup_{n\to\infty} -\frac{1}{n} \log \mathbb{P}(\hat{P}_n \in E) \leq D(P^* \,\|\, Q) = D^*$. $\square$

Definition: Log Mgf

Definition 8.11 For a random variable Y, the **log-moment generating function** of Y is $\Lambda : \mathbb{R} \to \mathbb{R}$ defined by

$$\Lambda(\lambda) := \ln \mathbb{E}\big[e^{\lambda Y}\big].$$

Notation 8.12 Write $\Lambda^*(x) = \sup\{\lambda x - \Lambda(\lambda) : \lambda > 0\}$.

Proposition: Chernoff Bound

Proposition 8.13 (Chernoff Bound) Let X_1^n be IID RVs and $f : A \to \mathbb{R}$ have mean $\mu = \mathbb{E}[f(X_1)]$. Denote the empirical averages by $S_n := \frac{1}{n}\sum_{i=1}^{n} f(X_i)$. Then for all $\varepsilon > 0$,

$$\mathbb{P}(S_n \geq \mu + \varepsilon) \leq e^{-n\Lambda^*(\mu+\varepsilon)},$$

where Λ is the log-moment generating function of the $f(X_i)$.

Proof (Hints). Use Markov's inequality. $\square$

Proof. By Markov's inequality, for all $\lambda > 0$,

$$\mathbb{P}(S_n \geq \mu + \varepsilon) = \mathbb{P}\big(e^{n\lambda S_n} \geq e^{n\lambda(\mu+\varepsilon)}\big) \leq e^{-n\lambda(\mu+\varepsilon)}\mathbb{E}\big[e^{\lambda n S_n}\big].$$

Now since the X_i are independent,

$$\mathbb{E}\big[e^{\lambda n S_n}\big] = \mathbb{E}\big[e^{\lambda \sum_{i=1}^n f(X_i)}\big] = \mathbb{E}\left[\prod_{i=1}^n e^{\lambda f(X_i)}\right] = \prod_{i=1}^n \mathbb{E}\big[e^{\lambda f(X_i)}\big] = e^{n\Lambda(\lambda)}.$$

Hence,

$$\mathbb{P}(S_n \geq \mu + \varepsilon) \leq e^{-n\lambda(\mu+\varepsilon)} e^{n\Lambda(\lambda)} = e^{-n(\lambda(\mu+\varepsilon)-\Lambda(\lambda))},$$

and this holds for all $\lambda > 0$, so taking the infimum over λ gives the result. $\square$

Example 8.14 Let X_1^n be IID with common PMF Q on finite alphabet A, let $f : A \to \mathbb{R}$ with mean $\mu = \mathbb{E}_{X\sim Q}[f(X)]$. Denote the empirical averages by $S_n := \frac{1}{n}\sum_{i=1}^n f(X_i)$. Let $\varepsilon > 0$. By WLLN, $\mathbb{P}(S_n \geq \mu + \varepsilon) \to 0$ as $n \to \infty$. We want to estimate how small this probability is as a function of n. Typically, the way we bound $\mathbb{P}(S_n \geq \mu + \varepsilon)$ is by the Chernoff Bound. Alternatively, we have

$$S_n = \frac{1}{n}\sum_{i=1}^n f(X_i) = \frac{1}{n}\sum_{i=1}^n \sum_{a\in A} \mathbb{1}_{\{X_i=a\}} f(a) = \sum_{a\in A} \hat{P}_n(a) f(a) = \mathbb{E}_{X\sim\hat{P}_n}[f(X)].$$

Let B be the event $B = \{S_n \geq \mu + \varepsilon\}$, then we can write B as $\{\hat{P}_n \in E\}$ where $E = \{P \in \mathcal{P} : \mathbb{E}_{X\sim P}[f(X)] \geq \mu + \varepsilon\}$. But Sanov says that

$\mathbb{P}(S_n \geq \mu + \varepsilon) = \mathbb{P}(\hat{P}_n \in E) \leq (n+1)^m e^{-nD_e(P^* \,\|\, Q)}$ and in fact it tells us that $D_e(P^* \,\|\, Q) = \inf\{D_e(P \,\|\, Q) : P \in E\}$ is asymptotically the "correct" exponent.

Proposition 8.15 Let X_1^n be IID RVs with common PMF Q on alphabet A and $f : A \to \mathbb{R}$ have mean $\mu = \mathbb{E}[f(X_1)]$. Let P^* be the minimiser in Sanov for the event $E = \{P \in \mathcal{P} : \mathbb{E}_{X\sim P}[f(X)] \geq \mu + \varepsilon\}$. Then

$$\forall \varepsilon > 0, \quad \Lambda^*(\mu + \varepsilon) = D_e(P^* \parallel Q),$$

where Λ is the log-moment generating function of the $f(X_i)$.

Proof (Hints).

- $\leq$: show that $S_n = \mathbb{E}_{X \sim \hat{P}_n}[f(X)]$, then use the Chernoff Bound and Sanov.
- $\geq$: for each $\lambda \geq 0$, define a PMF on A by

$$P_\lambda(a) = \frac{e^{\lambda f(a)}}{\mathbb{E}\left[e^{\lambda f(X_1)}\right]} Q(a).$$

- Show that $\Lambda'(\lambda) = \mathbb{E}_{Y \sim P_\lambda}[f(Y)]$ and $\Lambda''(\lambda) \geq 0$.
- Deduce that there exists $\lambda^* > 0$ such that $\Lambda'(\lambda^*) = \mu + \varepsilon$, then use the definition of P^* and expressing a relative entropy as an appropriate expectation to conclude the result.

Proof. ($\leq$): Let $\varepsilon > 0$. We have

$$S_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A} \mathbb{1}_{\{X_i=a\}} f(a) = \sum_{a\in A} \hat{P}_n(a) f(a) = \mathbb{E}_{X\sim\hat{P}_n}[f(X)].$$

So we have $\mathbb{P}(\hat{P}_n \in E) = \mathbb{P}(S_n \geq \mu + \varepsilon)$, hence

$$\begin{aligned}
\Lambda^*(\mu+\varepsilon) &\leq \liminf_{n\to\infty} -\frac{1}{n}\mathbb{P}(S_n \geq \mu+\varepsilon) && \text{by the Chernoff Bound} \\
&\leq \lim_{n\to\infty} -\frac{1}{n}\ln \mathbb{P}(\hat{P}_n \in E) \\
&= D_e(P^* \parallel Q) && \text{by Sanov}.
\end{aligned}$$

$(\geq)$: For each $\lambda \geq 0$, define the PMF P_λ on A by

$$P_\lambda(a) = \frac{e^{\lambda f(a)}}{\mathbb{E}\left[e^{\lambda f(X_1)}\right]} Q(a).$$

Then

$$\Lambda'(\lambda) = \frac{\mathbb{E}\left[f(X_1)e^{\lambda f(X_1)}\right]}{\mathbb{E}\left[e^{\lambda f(X_1)}\right]} = \frac{1}{\mathbb{E}\left[e^{\lambda f(X_1)}\right]} \sum_{a \in A} Q(a) f(a) e^{\lambda f(a)} = \mathbb{E}_{Y \sim P_\lambda}[f(Y)]$$

and also, a straightforward calculation shows that

$$\Lambda''(\lambda) = \mathrm{Var}_{Y \sim P_\lambda}(f(Y)) \geq 0.$$

Hence, $\Lambda'(\lambda)$ is increasing from $\Lambda'(0) = \mu$ to $\lim_{\lambda\to\infty} \Lambda'(\lambda) =: f^*$, so there exists $\lambda^* > 0$ such that $\Lambda'(\lambda^*) = \mu + \varepsilon$, hence $\mathbb{E}_{Y \in P_{\lambda^*}}[f(Y)] = \mu + \varepsilon$, so $P_{\lambda^*} \in E$. Thus,

$$\begin{aligned}
D_e(P^* \parallel Q) &\leq D_e(P_{\lambda^*} \parallel Q) \\
&= \mathbb{E}_{Y \sim P_{\lambda^*}}\left[\log \frac{P_{\lambda^*}(Y)}{Q(Y)}\right] \\
&= \mathbb{E}_{Y \sim P_{\lambda^*}}\left[\log \frac{e^{\lambda^* f(Y)}}{\mathbb{E}\left[e^{\lambda^* f(X_1)}\right]}\right] \\
&= \lambda^* \mathbb{E}_{Y \sim P_{\lambda^*}}[f(Y)] - \Lambda(\lambda^*)
\end{aligned}$$

$$= \lambda^*(\mu + \varepsilon) - \Lambda(\lambda^*) \leq \Lambda^*(\mu + \varepsilon).$$

$\square$

Corollary: Minimising Distribution In Sanov Is Unique For Nonempty Closed Convex Events

Corollary 8.16 Let X_1^n be IID RVs with common PMF Q on alphabet A. The minimiser P^* in Sanov for the event $E = \{P \in \mathcal{P} : \mathbb{E}_{X \sim P}[f(X)] \geq \mu + \varepsilon\}$ is unique and is given by

$$P^*(a) = P_{\lambda^*}(a) = \frac{e^{\lambda^* f(a)}}{\mathbb{E}\left[e^{\lambda^* f(X_1)}\right]} Q(a).$$

where $\lambda^* > 0$ satisfies $\mathbb{E}_{Y \sim P_{\lambda^*}}[f(Y)] = \mu + \varepsilon$.

Proof (Hints). Existence: by above proposition. Uniqueness: use a property of $D(P \parallel Q)$ and the fact that E is non-empty, convex and closed. $\square$

Proof. $D(P \parallel Q)$ is strictly convex in P for fixed Q and E is non-empty, convex and closed, so the minimising P^* is unique. The existence is by the proof of the above proposition. $\square$

Theorem: Pythagorean Identity

Theorem 8.17 (Pythagorean Identity) Let $E \subseteq \mathcal{P}$ be closed and convex. Let $Q \notin E$ have full support on A, and let P^* achieve the minimum in Sanov's theorem. Then

$$\forall P \in E, \quad D(P \parallel Q) \geq D(P \parallel P^*) + D(P^* \parallel Q).$$

Proof (Hints).

- For $P \in E$, define $\overline{P}_\lambda = \lambda P + (1-\lambda)P^*$ for $\lambda \in [0,1]$. Show that $D(\overline{P}_\lambda \parallel Q) \geq D(\overline{P}_0 \parallel Q)$ for all $\lambda \in [0,1]$.
- Show the derivative of $D_e(\overline{P}_\lambda \parallel Q)$ at $\lambda = 0$ is $D_e(P \parallel Q) - D_e(P \parallel P^*) - D_e(P^* \parallel Q)$.

$\square$

Proof. Let $P \in E$. Define the mixture $\overline{P}_\lambda = \lambda P + (1-\lambda)P^*$ for $0 \le \lambda \le 1$. Since E is convex, $\overline{P}_\lambda \in E$ for all $\lambda \in [0,1]$, and by definition of P^*, $D(\overline{P}_\lambda \parallel Q) \ge D(P^* \parallel Q) = D(\overline{P}_0 \parallel Q)$ for all $\lambda \in [0,1]$. So we have

$$
\begin{aligned}
0 &\le \frac{\mathrm{d}}{\mathrm{d}\lambda} D_e(\overline{P}_\lambda \parallel Q)\bigg|_{\lambda=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}\lambda} \sum_{a\in A} \overline{P}_\lambda(a) \ln \frac{\overline{P}_\lambda(a)}{Q(a)}\bigg|_{\lambda=0} \\
&= \sum_{a\in A} (P(a) - P^*(a)) \ln \frac{\overline{P}_\lambda(a)}{Q(a)}\bigg|_{\lambda=0} + \sum_{a\in A} (P(a) - P^*(a))
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{a \in A} P(a) \ln \frac{P^*(a)P(a)}{Q(a)P(a)} - \sum_{a \in A} P^*(a) \ln \frac{P^*(a)}{Q(a)} \\
&= D_e(P \parallel Q) - D_e(P \parallel P^*) - D_e(P^* \parallel Q).
\end{aligned}
$$

$\square$

Remark 8.18

- The Pythagorean Identity is an L^2-style bound: the minimiser P^* can be viewed as the "orthogonal projection" of Q onto E.
- The Pythagorean Identity provides a quantatitive version of the uniqueness statement in Corollary 8.16: if $D(P \parallel Q) = D(P^* \parallel Q)$, then $P = P^*$; additionally, if $D(P \parallel Q) \leq D(P^* \parallel Q) + \delta$ (i.e. $D(P \parallel Q)$ is close to $D(P^* \parallel Q)$), then $D(P \parallel P^*) \leq \delta$ (i.e. P is close to P^*).

8.3. The Gibbs conditioning principle

Lemma: Expectation Of Bounded Rvs Converges To Limit Of Convergence In Probability

Lemma 8.19 Let $\{Z_n : n \in \mathbb{N}\}$ be a bounded sequence of RVs which converges to $z \in \mathbb{R}$ in probability. Then

$$\mathbb{E}[Z_n] \to z \quad \text{as } n \to \infty.$$

Proof (Hints). Use Jensen's inequality, then split the expectation into two terms, one bounded above by ε, the other $\to 0$, to show that $|\mathbb{E}[Z_n] - c| \to 0$. $\square$

Proof. Let $\varepsilon > 0$. Since the Z_n are bounded, we have $|Z_n| \leq M$ for all $n \in \mathbb{N}$, for some constant M. By Jensen's Inequality,

$$|\mathbb{E}[Z_n] - z| \leq \mathbb{E}[|Z_n - z|] = \mathbb{E}\Big[|Z_n - z| \cdot \mathbb{1}_{\{|Z_n - z| \leq \varepsilon\}}\Big] + \mathbb{E}\Big[|Z_n - z| \cdot \mathbb{1}_{\{|Z_n - z| > \varepsilon\}}\Big].$$

The first term is bounded above by ε. The second term is bounded above by

$$(M + |z|) \cdot \mathbb{E}\Big[\mathbb{1}_{\{|Z_n - z| > \varepsilon\}}\Big] = (M + |z|) \cdot \mathbb{P}(|Z_n - z| > \varepsilon) \to 0 \quad \text{as } n \to \infty.$$

Thus, $\limsup_{n\to\infty} |\mathbb{E}[Z_n] - c| \leq \varepsilon$, and $\varepsilon > 0$ was arbitrary. $\square$

Theorem: Gibbs Conditioning Principle

Theorem 8.20 (Gibbs' Conditioning Principle) Let X_1^n be IID with common PMF Q which has full support on A. Let $\hat{P}_n$ be the empirical distribution of X_1^n. If $E \subseteq \mathcal{P}$ is closed, convex, has non-empty interior, and $Q \notin E$, then

$$\forall a \in A, \quad \mathbb{E}[\hat{P}_n(a) \mid \hat{P}_n \in E] = \mathbb{P}(X_1 = a \mid \hat{P}_n \in E) \to P^*(a) \quad \text{as} \quad n \to \infty,$$

where P^* is the unique minimiser in Sanov for the event E.

Proof (Hints).

- Showing the equality is straightforward.
- Define $B(Q,\delta) := \{P \in \mathcal{P} : D(P \parallel Q) \leq D(P^* \parallel Q) + \delta\}$, $C = B(Q, 2\delta) \cap E$ and $D = E \setminus C$.
- Show that $\mathbb{P}(\hat{P}_n \in D \mid \hat{P}_n \in E) \leq (n+1)^{2m} 2^{-n\delta}$ by using the density of $\{\mathcal{P}_n : n \in \mathbb{N}\}$ in $\mathcal{P}$ to reason that some $P_n \in B(Q,\delta) \cap E \cap \mathcal{P}_n$ eventually exists.
- Use the Pythagorean Identity and Pinsker's Inequality to show that $\mathbb{P}(\left|\hat{P}_n(a) - P^*(a)\right| > \varepsilon \mid \hat{P}_n \in E) \to 0$.

$\square$

Proof. The conditional distribution of each X_i given $\hat{P}_n \in E$ is the same, so

$$\mathbb{E}[\hat{P}_n(a) \mid \hat{P}_n \in E] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(X_i = a \mid \hat{P}_n \in E) = \mathbb{P}(X_1 = a \mid \hat{P}_n \in E).$$

Define the relative entropy neighbourhoods

$$B(Q,\delta) := \{P \in \mathcal{P} : D(P \parallel Q) \leq D(P^* \parallel Q) + \delta\},$$

and write $C = B(Q, 2\delta) \cap E$ and $D = E \setminus C$.

Then

$$\mathbb{P}(\hat{P}_n \in D \mid \hat{P}_n \in E) = \frac{\mathbb{P}(\hat{P}_n \in D)}{\mathbb{P}(\hat{P}_n \in E)}.$$

By Sanov,

$$\mathbb{P}(\hat{P}_n \in D) \le (n+1)^m 2^{-n \inf\{D(P \parallel Q) : P \in D\}} \le (n+1)^m 2^{-n(D(P^* \parallel Q) + 2\delta)}$$

and for the denominator, since $\{\mathcal{P}_n : n \in \mathbb{N}\}$ is dense in $\mathcal{P}$, $\mathcal{P}_n$ eventually intersects every open set in $\mathcal{P}$, so eventually $B(Q, \delta) \cap E \cap \mathcal{P}_n$ is non-empty (since E has non-empty interior). So we can eventually find $P_n \in \mathcal{P}_n \cap E \cap B(Q, \delta)$. By Corollary 8.9,

$$
\begin{aligned}
\mathbb{P}(\hat{P}_n \in E) &\geq \mathbb{P}(\hat{P}_n \in B(Q,\delta) \cap E) \\
&\geq \mathbb{P}(\hat{P}_n = P_n) = Q^n(T(P_n)) \\
&\geq (n+1)^{-m} 2^{-nD(P_n \,\|\, Q)} \\
&\geq (n+1)^{-m} 2^{-n(D(P^* \,\|\, Q)+\delta)},
\end{aligned}
$$

since $P_n \in B(Q,\delta)$. Combining these, we obtain

$$
\mathbb{P}(\hat{P}_n \in D \mid \hat{P}_n \in E) \leq (n+1)^{2m} 2^{-n\delta} \to 0 \quad \text{as } n \to \infty.
$$

For $P \in C$, by the Pythagorean Identity,

$$D(P^* \parallel Q) \geq D(P \parallel Q) \geq D(P \parallel P^*) + D(P^* \parallel Q),$$

thus $D(P \parallel P^*) \leq 2\delta$. So

$$\mathbb{P}\Big(D\big(\hat{P}_n \parallel P^*\big) \leq 2\delta \mid \hat{P}_n \in E\Big) \geq \mathbb{P}\Big(\hat{P}_n \in C \mid \hat{P}_n \in E\Big) \to 1 \quad \text{as } n \to \infty.$$

Hence by Pinsker's Inequality, since $\delta > 0$ was arbitrary,

$$\mathbb{P}\left(\left\|\hat{P}_n - P^*\right\|_{\mathrm{TV}} > \varepsilon \,\middle|\, \hat{P}_n \in E\right) \to 0 \text{ as } n \to \infty$$

for all $\varepsilon > 0$. Thus also, $\mathbb{P}(\left|\hat{P}_n(a) - P^*(a)\right| > \varepsilon \mid \hat{P}_n \in E) \to 0$. So, conditional on $\hat{P}_n \in E$, $\hat{P}_n \to P^*$ in probability as $n \to \infty$. Therefore, since $(\widehat{P}_n(a))$ is a bounded sequence, we also have $\mathbb{E}\Big[\widehat{P}_n(a) \mid \widehat{P}_n \in E\Big] \to P^*(a)$ as $n \to \infty$ by Lemma 8.19. $\square$

Example 8.21 Suppose a fair die is rolled 1000 times, and the observed average of the rolls is at least 5. What proportion of the rolls was a 6?

Let X_1^{1000} be IID RVs with uniform distribution Q on $A = \{1, 2, 3, 4, 5, 6\}$. Let $f(x) = x$, $\mu = \mathbb{E}[X_1^{1000}] = 3.5$, let $E = \{P \in \mathcal{P} : \mathbb{E}_{X \sim P}[X] \geq 5\}$. By Corollary 8.16, the minimiser P^* is unique and is given by

$$P^*(a) = \frac{e^{\lambda^* a}}{\sum_{k=1}^{6} e^{\lambda^* k}}, \quad \forall a \in A,$$

where $\lambda^* > 0$ is such that $\mathbb{E}_{Y \sim P_{\lambda^*}}[Y] = 5$. We can directly compute $\lambda^* \approx 0.63$ and so

$$P^* \approx (0.021, 0.039, 0.14, 0.25, 0.48)$$

So by the Gibbs' Conditioning Principle, we expect that about 48% of the rolls were 6.

8.4. Error probability in fixed-rate data compression

Theorem: Error Exponents For Fixed Rate Compression

Theorem 8.22 Let $\boldsymbol{X} = \{X_n : n \in \mathbb{N}\}$ be a memoryless source with entropy $H = H(X_1)$ and with PMF Q which has full support on finite alphabet A. For any rate R with $H < R < \log|A|$,

- $\Longrightarrow$: There is a fixed-rate code $\{B_n^* : n \in \mathbb{N}\}$ with asymptotic rate no more than R bits/symbol:

$$\limsup_{n\to\infty} \frac{1}{n}(1 + \lceil \log|B_n^*| \rceil) = \limsup_{n\to\infty} \frac{1}{n} \log|B_n^*| \leq R,$$

  and with probability of error $P_e^{(n)}$ that decays to zero exponentially fast:

$$\limsup_{n\to\infty} \frac{1}{n} \log P_e^{(n)} \le -D^*,$$

where

$$D^* = \inf\{D(P \parallel Q) : H(P) \ge R\}.$$

- $\Longleftarrow$: for any fixed-rate code $\{B_n : n \in \mathbb{N}\}$ with asymptotic rate no more than R bits/symbol:

$$\limsup_{n\to\infty} \frac{1}{n}(1 + \lceil \log|B_n| \rceil) = \limsup_{n\to\infty} \frac{1}{n} \log|B_n| \le R,$$

then its probability of error $P_e^{(n)}$ cannot decay faster than exponentially with exponent D^*:

$$\liminf_{n\to\infty} \frac{1}{n} \log P_e^{(n)} \geq -D^*.$$

Proof (Hints).

- $\Longrightarrow$: let B_n^* be the codebook which is a union over the set of type classes $T(P)$ such that $H(P) < R$.
- $\Longleftarrow$: explain why there is $\delta > 0$ such that $\inf\{D(P \parallel Q) : H(P) \geq R + \delta\} \leq D^* + \varepsilon$.
- Explain why, for all n large enough, there is $P_n \in \mathcal{P}_n$ such that $H(P_n) \geq R + \delta/2$ and $D(P_n \parallel Q) \leq D^* + 2\varepsilon$.
- Show that $|B_n|/|T(P_n)| \to 0$ as $n \to \infty$, and hence that $P_e^{(n)} \geq \frac{1}{2}(n+1)^{-m} 2^{-n(D^*+2\varepsilon)}$ eventually.

$\square$

Proof. $\Longrightarrow$: define the codebook

$$B_n^* = \bigcup_{\substack{P \in \mathcal{P}_n \\ H(P) < R}} T(P).$$

Then by Proposition 8.4 and Proposition 8.8,

$$|B_n^*| = \sum_{\substack{P \in \mathcal{P}_n \\ H(P) < R}} |T(P)| \leq \sum_{\substack{P \in \mathcal{P}_n \\ H(P) < R}} 2^{nH(P)} \leq (n+1)^m 2^{nR},$$

and so $\limsup_{n\to\infty} \frac{1}{n} \log|B_n^*| \leq R$. For the probability of error,

$$P_e^{(n)} = \mathbb{P}(X_1^n \notin B_n^*) = Q^n\left(\bigcup_{\substack{P\in\mathcal{P}_n\\ H(P)\geq R}} T(P)\right) \leq \sum_{\substack{P\in\mathcal{P}_n\\ H(P)\geq R}} Q^n(T(P)) \leq (n+1)^m 2^{-nD^*}.$$

$\Longleftarrow$: let $\varepsilon > 0$ be arbitrary. By continuity, there is a $\delta > 0$ such that

$$\inf\{D(P \parallel Q) : H(P) \geq R + \delta\} \leq D^* + \varepsilon.$$

Since the n-types $\{\mathcal{P}_n : n \in \mathbb{N}\}$ are dense in $\mathcal{P}$, for all n large enough, we can find $P_n \in \mathcal{P}_n$ such that $H(P_n) \geq R + \delta/2$ and $D(P_n \parallel Q) \leq D^* + 2\varepsilon$. Also, by assumption, there is a sequence (r_n) such that $\frac{1}{n}\log|B_n| \leq R + r_n$ and $r_n \to 0$. Now

$$\frac{|B_n|}{|T(P_n)|} \leq \frac{2^{n(R+r_n)}}{(n+1)^{-m} 2^{nH(P_n)}} = (n+1)^m 2^{n(R-H(P_n)+r_n)}$$
$$\leq (n+1)^m 2^{n(r_n-\delta/2)} \to 0 \quad \text{as } n \to \infty.$$

So $|B_n|/|T(P_n)| \leq 1/2$ eventually. Then, for an arbitrary string $x_1^n \in T(P_n)$, we have

$$P_e^{(n)} = \mathbb{P}(X_1^n \in B_n^c) \geq \mathbb{P}(X_1^n \in T(P_n) \cap B_n^c)$$
$$= |T(P_n) \cap B_n^c| Q^n(x_1^n) = \frac{|T(P_n) \cap B_n^c|}{|T(P_n)|} Q^n(T(P_n))$$

$$
\begin{aligned}
&\geq \left(1 - \frac{|T(P_n) \cap B_n|}{|T(P_n)|}\right)(n+1)^{-m} 2^{-nD(P_n \,\|\, Q)} \\
&\geq \left(1 - \frac{|B_n|}{|T(P_n)|}\right)(n+1)^{-m} 2^{-nD(P_n \,\|\, Q)} \\
&\geq \frac{1}{2}(n+1)^{-m} 2^{-n(D^* + 2\varepsilon)} \quad \text{eventually}
\end{aligned}
$$

Thus,

$$
\liminf_{n \to \infty} \frac{1}{n} \log P_e^{(n)} \geq -(D^* + 2\varepsilon),
$$

and since $\varepsilon > 0$ was arbitrary, we are done. $\square$

Remark 8.23

- Theorem 8.22 gives the rate at which the error probabilities $P_e^{(n)}$ of the codes in the Fixed-rate Coding Theorem decay.
- Note that the code B_n^* is **universal**: it achieves the optimal error probability at rate R *simultaneously* for all memoryless sources with entropy $H < R$.
- The Fixed-rate Coding Theorem says that $P_e^{(n)}$ cannot tend to zero if $R < H$. In fact, it is possible to show a “strong converse” of the Fixed-rate Coding Theorem, which says that in this case, $P_e^{(n)} \to 1$ exponentially fast.

9. Variable-rate lossless data compression

Notation 9.1 Let $\{0,1\}^*$ denote the set of all binary strings of finite length.

Definition: Variable Rate Code

Definition 9.2 A **variable-rate lossless compression code** of block length n on a finite alphabet A is an injective map $C_n : A^n \to \{0,1\}^*$ which maps source strings to **codewords**. C_n is also known as the **encoder**.

Each C_n has an associated **length function** $L_n : A^n \to \mathbb{N}$, defined as

$$L_n(x_1^n) = \text{length of } C_n(x_1^n).$$

Definition: Variable Rate Code.Prefix Free

Definition 9.3 A code C_n is **prefix-free** if for all $x_1^n \neq y_1^n \in \{0,1\}^n$, the codeword $C_n(x_1^n)$ is not a prefix (an initial segment) of $C_n(y_1^n)$.

Example 9.4

x	$C(x)$
a	00
b	01
c	10
d	11

x	$C(x)$
a	0
b	10
c	110
d	111

x	$C(x)$
a	0
b	00
c	110
d	111

x	$C(x)$
a	0
b	1
c	00
d	11

The first two codes are prefix-free, the last two are not.

Remark 9.5 An advantage of prefix-free codes is that once a full codeword is received, it is guaranteed to be that codeword and not the start of another.

Theorem: Krafts Inequality

Theorem 9.6 (Kraft's Inequality)

- ($\Longrightarrow$): for any function $L_n : A^n \to \mathbb{N}$ satisfying **Kraft's inequality**:

$$\sum_{x_1^n \in A^n} 2^{-L_n(x_1^n)} \leq 1,$$

  there is a prefix-free code C_n on A^n with length function L_n.
- ($\Longleftarrow$): the length function of any prefix-free code satisfies Kraft's inequality.

Proof (Hints). For both directions, consider the complete binary tree of depth $\max\{L_n(x_1^n) : x_1^n \in A^n\}$. For $\Longleftarrow$, consider the number of descendants of each codeword in terms of its depth. $\square$

Proof. $\Longleftarrow$: let C_n be a prefix-free code with length function L_n. Let $L^* = \max\{L_n(x_1^n) : x_1^n \in A^n\}$ and consider the complete binary tree of depth L^*. If we mark all the codewords on the tree, then the prefix-free property implies that no codeword is a descendant of any other codeword. Each codeword $C_n(x_1^n)$ has $2^{L^*-L_n(x_1^n)}$ descendants (possibly including itself) at depth L^*. The prefix-free property also implies that these descendants are disjoint for different codewords. Since the total number of leaves at depth L^* is 2^{L^*}, we have

$$2^{L^*} \geq \sum_{x_1^n \in A^n} 2^{L^*-L_n(x_1^n)}.$$

$\Longrightarrow$: given a length function L_n satisfying Kraft's inequality, consider the complete binary tree of depth $L^* = \max\{L_n(x_1^n) : x_1^n \in A^n\}$. Then, ordering the $x_1^n \in A^n$ in the order of increasing $L_n(x_1^n)$, assign to each x_1^n (in order) any available node (i.e. any node that is not a prefix or descandant of any codewords already assigned) at depth $L_n(x_1^n)$. Kraft's inequality guarantees that there will always be such a node. $\square$

Remark 9.7 Kraft's Inequality informally says "not all codelengths for prefix-free codes can be short".

9.1. The codes-distributions correspondence

Theorem: Codes Distributions Correspondence

Theorem 9.8 (Codes-distributions Correspondence)

- $\Longrightarrow$: for any PMF Q_n on A^n, there is a prefix-free code C_n^* with length function L_n^* such that

$$\forall x_1^n \in A^n, \quad L_n^*(x_1^n) < -\log Q_n(x_1^n) + 1$$

- $\Longleftarrow$: for any prefix-free code C_n with length function L_n, there is a PMF Q_n on A^n such that

$$\forall x_1^n \in A^n, \quad -\log Q_n(x_1^n) \leq L_n(x_1^n).$$

Proof (Hints).

- $\Longrightarrow$: straightforward.
- $\Longleftarrow$: consider Kraft's Inequality to define a suitable Q_n.

$\square$

Proof. $\Longrightarrow$: Let $L_n^*(x_1^n) = \lceil -\log Q_n(x_1^n) \rceil < -\log Q_n(x_1^n) + 1$. L_n^* satisfies Kraft's inequality:

$$\sum_{x_1^n \in A^n} 2^{-L_n(x_1^n)} = \sum_{x_1^n \in A^n} 2^{-\lceil -\log Q_n(x_1^n) \rceil} \leq \sum_{x_1^n \in A^n} 2^{\log Q_n(x_1^n)} = \sum_{x_1^n \in A^n} Q_n(x_1^n) = 1.$$

So we are done by the first part of Kraft's Inequality.

$\Longleftarrow$: define the PMF Q_n on A^n by

$$Q_n(x_1^n) = \frac{2^{-L_n(x_1^n)}}{\sum_{y_1^n \in A^n} 2^{-L_n(y_1^n)}}.$$

Then

$$-\log Q_n(x_1^n) = L_n(x_1^n) + \log\left(\sum_{y_1^n \in A^n} 2^{-L_n(y_1^n)}\right) \leq L_n(x_1^n).$$

since L_n satisfies Kraft's inequality (i.e. $\sum_{y_1^n \in A^n} 2^{-L_n(y_1^n)} \leq 1$). $\square$

Remark 9.9

- Codes-distributions Correspondence says that the performance of any prefix-free can be dominated by a code with length function $L_n(x_1^n) \approx -\log Q_n(x_1^n)$ for some PMF Q_n on A^n, and that for any distribution Q_n such a code exists. So finding a good code is equivalent to finding a good distribution. This assumes nothing about the distribution of the source X_1^n or the block length n.

Theorem: Entropic Bounds On Expected Length Of Prefix Free Codes

Theorem 9.10 Let X_1^n have PMF P_n on A^n.

$\Longrightarrow$: there is a prefix-free code C_n^* with length function L_n^* that achieves an expected description length of

$$\mathbb{E}[L_n^*(X_1^n)] < H(X_1^n) + 1.$$

$\Longleftarrow$: for any prefix-free code C_n with length function L_n on A^n,

$$\mathbb{E}[L_n(X_1^n)] \geq H(X_1^n).$$

Proof (Hints). Straightforward. $\square$

Proof. $\Longrightarrow$: let C_n^* be the code with length function $L_n^*(x_1^n) = \lceil -\log P_n(x_1^n) \rceil$ as in the Codes-distributions Correspondence. Then

$$\mathbb{E}[L_n^*(X_1^n)] < \mathbb{E}[-\log P_n(X_1^n) + 1] = H(X_1^n) + 1.$$

$\Longleftarrow$: let Q_n be as in the Codes-distributions Correspondence. Then

$$\begin{aligned}
\mathbb{E}[L_n(X_1^n)] &\geq \mathbb{E}[-\log Q_n(X_1^n)] \\
&= \mathbb{E}\left[\log\left(\frac{1}{P_n(X_1^n)} \cdot \frac{P_n(X_1^n)}{Q_n(X_1^n)}\right)\right] \\
&= \mathbb{E}[-\log P_n(X_1^n)] + \mathbb{E}\left[\log \frac{P_n(X_1^n)}{Q_n(X_1^n)}\right]
\end{aligned}$$

$$= H(X_1^n) + D(P_n \parallel Q_n) \geq H(X_1^n).$$

$\square$

Corollary: Entropy Rate Of Source Is Best Asymptotic Prefix Free Compression Rate

Corollary 9.11 Let $\boldsymbol{X} = \{X_n : n \in \mathbb{N}\}$ be a stationary source with entropy rate $H = H(\boldsymbol{X})$. Then H is the best asymptotically achievable compression rate among all variable-rate prefix-free codes:

$$\lim_{n\to\infty} \inf_{(C_n, L_n) \text{ prefix-free}} \frac{1}{n}\mathbb{E}[L_n(X_1^n)] = H.$$

Proof (Hints). Straightforward. $\square$

Proof. By Theorem 9.10,

$$\frac{1}{n}H(X_1^n) \leq \inf_{(C_n,L_n)\text{ prefix-free}} \frac{1}{n}\mathbb{E}[L_n(X_1^n)] < \frac{1}{n}(H(X_1^n)+1).$$

$\square$

9.2. Shannon codes and their properties

Definition: Shannon Code

Definition 9.12 A **Shannon code** for a distribution Q_n on A^n is a code with length function

$$L_n(x_1^n) := \lceil -\log Q_n(x_1^n) \rceil.$$

Note this is the code used in the proof of the Codes-distributions Correspondence.

Remark 9.13

- Shannon codes do not always achieve the optimal (minimal) expected description length $\mathbb{E}[L_n(X_1^n)]$, which is achieved instead by the Huffman code. However, the difference between the expected description lengths of these codes is less than one bit by Theorem 9.10.
- Shannon codes give shorter descriptions to likely messages and longer descriptions to unlikely messages.

Definition: Ideal Shannon Codelength

Definition 9.14 We call the $L_n(x_1^n) = -\log Q_n(x_1^n)$ for $x_1^n \in A^n$ the **ideal Shannon codelengths**.

Theorem: Competitive Optimality Of Shannon Codes

Theorem 9.15 (Competitive Optimality of Shannon Codes) Let P_n be a distribution on A^n and $X_1^n \sim P_n$. For any other PMF Q_n on A^n,

$$\mathbb{P}(-\log Q_n(X_1^n) \leq -\log P_n(X_1^n) - K) \leq 2^{-K}.$$

Proof (Hints). Use Markov's inequality. $\square$

Proof. By Markov's inequality, we have

$$\begin{aligned}
\mathbb{P}(-\log Q_n(X_1^n) \leq -\log P_n(X_1^n) - K) &= \mathbb{P}\left(\frac{Q_n(X_1^n)}{P_n(X_1^n)} \geq 2^K\right) \\
&\leq 2^{-K}\mathbb{E}\left[\frac{Q_n(X_1^n)}{P_n(X_1^n)}\right] \\
&= 2^{-K} \sum_{x_1^n \in A^n} P_n(x_1^n) \cdot \frac{Q_n(x_1^n)}{P_n(x_1^n)} \\
&= 2^{-K}.
\end{aligned}$$

□

10. Universal data compression

In this chapter, assume that we want to compress a message $x_1^n \in \{0,1\}^n$ where each x_i is produced by an unknown distribution $P = P_{\theta^*}$ which belongs to the parametric family $\{P_\theta \sim \mathrm{Bern}(\theta) : \theta \in (0,1)\}$. We also assume codelengths can be non-integral for simplicity, since the actual codelength differs by at most one bit.

Note that in this case, $\theta_{\mathrm{MLE}} = k/n$ where k is the number of 1s in x_1^n. So the maximum likelihood distribution for x_1^n amsong all P_θ is its type $\hat{P}_n$, and by Proposition 8.5, for all $\theta \in \Theta$,

$$-\log P_{\theta_{\mathrm{MLE}}}^n(x_1^n) = nH\left(\hat{P}_n\right) \leq -\log P_\theta^n(x_1^n).$$

Definition: Mle Code

Definition 10.1 The **MLE code** first describes $\widehat{\theta}_{\mathrm{MLE}}$ to the decoder, then describes x_1^n using the Shannon code for $P^n_{\hat{\theta}_{\mathrm{MLE}}}$.

Proposition: Mle Code Price Of Universality

Proposition 10.2 The description length of the MLE code is

$$nH(\widehat{P}_n) + \log(n+1).$$

In particular, the price of universality of the MLE code is $\log n$ bits.

Proof (Hints). Trivial. $\square$

Proof. $\theta_{\text{MLE}} = k/n$ where k is the number of 1s in x_1^n, so $k \in \{0, ..., n\}$. So k can be described using $\log(n+1)$ bits. x_1^n is described using $-\log P^n_{\theta_{\text{MLE}}}(x_1^n) = nH\left(\hat{P}_n\right)$ bits. $\square$

Proposition: Mle Code Expected Price Of Universality

Proposition 10.3 The expected description length of the MLE code is bounded above by

$$nH(P^n_{\theta^*}) + \log(n+1).$$

In particular, the price of universality in expectation of the MLE code is $\log n$ bits.

Proof (Hints). Straightforward. $\square$

Proof. The expected description length is

$$\begin{aligned}\log(n+1) + \mathbb{E}\Big[-\log P^n_{\theta_{\text{MLE}}}(X_1^n)\Big] &\le \log(n+1) + \mathbb{E}[-\log P^n_{\theta^*}(X_1^n)] \\ &= \log(n+1) + nH(P_{\theta^*}).\end{aligned}$$

$\square$

Definition: Counting Code

Definition 10.4 The **counting code** first describes $\theta_{\text{MLE}} = k/n$ to the decoder, then describes the index of x_1^n in the ordered list of $\binom{n}{k}$ binary strings containing k 1s.

Proposition: Counting Code Description Length

Proposition 10.5 The description length of the counting code is

$$\log(n+1) + \log\binom{n}{k}.$$

Proof (Hints). Trivial. $\square$

Proof. Trivial. $\square$

Definition: Uniform Mixture

Definition 10.6 Given a parametric family of distributions $\{P_\theta : \theta \in [0,1]\}$, the **uniform mixture** of $\{P_\theta^n : \theta \in [0,1]\}$ is the PMF Q_n on A^n defined by

$$Q_n(x_1^n) = \int_0^1 P_\theta^n(x_1^n)\,\mathrm{d}\theta.$$

Definition: Mixture Code

Definition 10.7 The **mixture code** is the Shannon code for the uniform mixture Q_n of the P_θ^n.

Lemma: Closed Form Expression For Uniform Mixture Of Bernoullis

Lemma 10.8 For all $k, \ell \in \mathbb{N}_0$,

$$\int_0^1 \theta^k (1-\theta)^\ell \, \mathrm{d}\theta = \frac{k!\ell!}{(k+\ell+1)!}.$$

Proof. Exercise. $\square$

Proposition: Mixture Code Description Length

Proposition 10.9 The description length of the mixture code is

$$\log(n+1) + \log\binom{n}{k}.$$

Proof (Hints). Straightforward. $\square$

Proof. The uniform mixture is

$$Q_n(x_1^n) = \int_0^1 \theta^k (1-\theta)^{n-k} \, \mathrm{d}\theta,$$

where k is the number of 1s in x_1^n. By the above lemma with $\ell = n - k$, the description length is

$$-\log Q_n(x_1^n) = -\log \frac{k!(n-k)!}{(n+1)!} = \log(n+1) + \log \binom{n}{k}.$$

$\square$

Definition: Predictive Code

Definition 10.10 The **predictive code** describes the message x_1^n in steps instead of describing it all at once: having already communicated x_1^i, the encoder and decoder calculate the estimate

$$\hat{\theta}_i = \frac{k_i + 1}{i + 2},$$

where k_i is the number of 1s in x_1^i. Since $\hat{\theta}_i$ is known to the decoder, the encoder then describes x_{i+1} using $-\log P_{\hat{\theta}_i}(x_{i+1})$ bits. This is repeated for each $i = 1, ..., n-1$.

Proposition: Predictive Code Description Length

Proposition 10.11 The description length of the predictive code is

$$\log(n+1) + \log\binom{n}{k},$$

where k is the number of 1s in x_1^n.

Proof (Hints). Straightforward. $\square$

Proof. We have $k_0 = 0$ so $\hat{\theta}_0 = 1/2$. The description length is

$$\begin{aligned}
\sum_{i=1}^{n} -\log P_{\hat{\theta}_{i-1}}(x_i) &= \sum_{i=1}^{n} -\log\left(\hat{\theta}_{i-1}^{x_i}\left(1-\hat{\theta}_{i-1}\right)^{1-x_i}\right) \\
&= -\sum_{i=1}^{n}\left(x_i \log \hat{\theta}_{i-1} + (1-x_i)\log\left(1-\hat{\theta}_{i-1}\right)\right) \\
&= -\sum_{i=1}^{n}\left(x_i \log \frac{k_{i-1}+1}{i+1} + (1-x_i)\log \frac{i-k_{i-1}}{i+1}\right) \\
&= -\sum_{i:x_i=1} \log(k_{i-1}) - \sum_{i:x_i=0} \log(i-k_{i-1}) + \sum_{i=1}^{n} \log(i+1)
\end{aligned}$$

$$
\begin{aligned}
&= -\log(k_n!) - \log((n-k_n)!) + \log((n+1)!) \\
&= \log(n+1) + \log\binom{n}{k}.
\end{aligned}
$$

$\square$

Lemma: Exponential Upper Bound On Binomial Coefficient

Lemma 10.12 Let $n \in \mathbb{N}$, $0 \leq k \leq n$ and $p = k/n$. Then

$$\binom{n}{k} \leq \frac{1}{\sqrt{2\pi np(1-p)}} \cdot 2^{nH(\mathrm{Bern}(p))}.$$

Proof. Exercise. $\square$

Definition: Fisher Information

Definition 10.13 The **Fisher information** for a parametric family of PMFS $\{P_\theta : \theta \in \Theta\}$ is defined as

$$J(\theta) := \mathbb{E}_{X \sim P_\theta}\left[\frac{\frac{\partial}{\partial\theta} P_\theta(X)}{\left(P_\theta(X)\right)^2}\right].$$

Proposition: Upper Bound On Price Of Universality Of Counting Mixture And Predictive Codes

Proposition 10.14 The description length of the counting, mixture and predictive codes is bounded above by

$$nH(\hat{P}_n) + \frac{1}{2}\log\left(n\frac{J(\theta_{\mathrm{MLE}})}{2\pi}\right) + 1.$$

In particular, the price of universality of the counting, mixture and predictive codes is $\frac{1}{2}\log n$ bits.

Proof (Hints). Straightforward. $\square$

Proof. The description length of all three codes is $\log(n+1) + \log\binom{n}{k}$ by Proposition 10.5, Proposition 10.9 and Proposition 10.5. By Lemma 10.12, we have

$$\log\binom{n}{k} \leq nH\left(\hat{P}_n\right) - \frac{1}{2}\log(2\pi n\theta_{\mathrm{MLE}}(1-\theta_{\mathrm{MLE}})) = nH\left(\hat{P}_n\right) + \frac{1}{2}\log\left(\frac{J(\theta_{\mathrm{MLE}})}{2\pi n}\right),$$

where $J(\cdot)$ is the Fisher information of the family of Bernoulli PMFs. This concludes the result. $\square$

Notation: Mdl Estimator

Notation 10.15 Partitioning the interval $[0, 1]$ into $\sqrt{n}$ intervals of length $1/\sqrt{n}$, let θ_{MDL} denote the index of the interval that θ_{MLE} belongs to.

Definition: Mdl Code

Definition 10.16 The **MDL (minimum description length) code** first describes θ_{MDL} to the decoder, then describes x_1^n using the Shannon code for $P_{\theta_{\mathrm{MDL}}}$.

Remark 10.17 Note that we can write the MLE as

$$\theta_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \theta^* + \frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \theta^*)\right),$$

where θ^* is the true underlying parameter. The term in the brackets has mean $\mu = 0$ and variance $\sigma^2 = \theta^*(1 - \theta^*)$. So by the central limit theorem,

$$\theta_{\text{MLE}} \approx \theta^* + \frac{1}{\sqrt{n}}Z, \quad Z \sim N(\mu, \sigma^2).$$

Hence, θ_{MLE} has fluctuations of order $O(1/\sqrt{n})$. This suggests the MLE code strategy of describing it with $O(1/n)$ accuracy is too fine-grained, and the MDL code strategy of describing it with $O(1/\sqrt{n})$ accuracy is more appropriate.

Proposition: Mdl Code Description Length

Proposition 10.18 The description length of the MDL code is

$$nH\left(\hat{P}_n\right) + \frac{1}{2}\log n + O(1).$$

In particular, the price of universality of the MDL code is $\frac{1}{2}\log n$ bits.

Proof (Hints). Use that $D\Big(P_{\theta_{\mathrm{MLE}}} \parallel P_{\theta_{\mathrm{MDL}}}\Big) = O\Big((\theta_{\mathrm{MLE}} - \theta_{\mathrm{MDL}})^2\Big)$ (since $D(P \parallel Q)$ is locally quadratic in $(P - Q)$). $\square$

Proof. By Proposition 8.5, we have

$$-\log P^n_{\theta_{\mathrm{MDL}}}(x_1^n) = nD\Big(P_{\theta_{\mathrm{MLE}}} \parallel P_{\theta_{\mathrm{MDL}}}\Big) + nH\Big(\hat{P}_n\Big).$$

Since $D(P \parallel Q)$ is locally quadratic in $(P - Q)$, the Taylor expansion gives

$$D\Big(P_{\theta_{\mathrm{MLE}}} \parallel P_{\theta_{\mathrm{MDL}}}\Big) = O\Big((\theta_{\mathrm{MLE}} - \theta_{\mathrm{MLE}})^2\Big).$$

Now by construction, $|\theta_{\mathrm{MLE}} - \theta_{\mathrm{MDL}}| = O(1/\sqrt{n})$. Thus,

$$nD\Big(P_{\theta_{\mathrm{MLE}}} \parallel P_{\theta_{\mathrm{MDL}}}\Big) = O(1),$$

which concluses the result. $\square$

11. Redundancy and the price of universality

11.1. Redundancy

Definition: Redundancy

Definition 11.1 Suppose $x_1^n \in A^n$ is generated by a memoryless source with PMF P on a finite alphabet A, with $|A| = m$. The **redundancy** on x_1^n of a code with length function L_n is the difference between $L_n(x_1^n)$ and the target compression of $-\log P^n(x_1^n)$ bits (the ideal Shannon codelength with respect to P^n), so is given by

$$L_n(x_1^n) - (-\log P^n(x_1^n)).$$

If we use the Shannon code with respect to an arbitrary PMF Q_n on A^n, the redundancy is

$$\rho_n(x_1^n; P, Q_n) = -\log Q_n(x_1^n) - (-\log P^n(x_1^n)) = \log \frac{P^n(x_1^n)}{Q_n(x_1^n)}.$$

Remark 11.2 Note that by the Codes-distributions Correspondence, we can restrict our attention to (ideal) Shannon codes (assuming that we ignore integer codelength constraints).

Definition: Worst Case Maximal Redundancy

Definition 11.3 The **worst-case maximal redundancy** of the Shannon code with respect to Q_n is its largest redundancy over all strings and all source distributions:

$$\sup_{P \in \mathcal{P}} \max_{x_1^n \in A^n} \log \frac{P^n(x_1^n)}{Q_n(x_1^n)}.$$

Definition: Minimax Maximal Redundancy

Definition 11.4 The **minimax maximal redundancy** ρ_n^* over the class of all IID source distributions on A^n is the shortest possible worst-case maximal redundancy:

$$\rho_n^* = \inf_{Q_n} \sup_{P \in \mathcal{P}} \max_{x_1^n \in A^n} \log \frac{P^n(x_1^n)}{Q_n(x_1^n)}.$$

Definition: Worst Case Average Redundancy

Definition 11.5 The **worst-case average redundancy** of the Shannon code with respect to Q_n is its largest average redundancy over all source distributions:

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{X_1^n \sim P^n}\left[\log \frac{P^n(X_1^n)}{Q_n(X_1^n)}\right] = \sup_{P \in \mathcal{P}} D(P^n \parallel Q_n).$$

Definition: Minimax Average Redundancy

Definition 11.6 The **minimax average redundancy** over the class of all IID source distributions on A^n is the shortest possible worst-case average redundancy

$$\overline{\rho}_n = \inf_{Q_n} \sup_{P \in \mathcal{P}} D(P^n \parallel Q_n).$$

11.2. Shtarkov's upper bound

Theorem: Normalised Maximum Likelihood Code

Theorem 11.7 (Normalised Maximum Likelihood Code) Let $\{P_\theta : \theta \in \Theta\}$ be a parametric family of distributions on a finite alphabet B. Denote the minimax maximal redundancy over $\{P_\theta : \theta \in \Theta\}$ by

$$\rho^*(\Theta) := \inf_Q \sup_{\theta \in \Theta} \max_{x \in B} \log \frac{P_\theta(x)}{Q(x)}.$$

Then $\rho^*(\Theta) = \log Z$, where

$$Z = \sum_{x \in B} \sup_{\theta \in \Theta} P_\theta(x).$$

Proof (Hints).

- For $\leq$, consider a suitable distribution Q^* which is defined using Z.
- For $\geq$, use that for every Q, $Q(x) \leq Q^*(x)$ for at least one x.

$\square$

Proof. Define the distribution Q^* on B by $Q^*(x) = \frac{1}{Z} \sup_{\theta \in \Theta} P_\theta(x)$. We have

$$\begin{aligned}
\rho^*(\Theta) &\le \sup_{\theta \in \Theta} \max_{x \in B} \log \frac{P_\theta(x)}{Q^*(x)} \\
&= \max_{x \in B} \sup_{\theta \in \Theta} \log \frac{P_\theta(x)}{Q^*(x)} \\
&= \max_{x \in B} \log \frac{\sup_{\theta \in \Theta} P_\theta(x)}{Q^*(x)} = \max_{x \in B} \log Z = \log Z.
\end{aligned}$$

For the lower bound, note that for every Q, $Q(x) \le Q^*(x)$ for at least one x, say x^*. Therefore,

$$\sup_{\theta\in\Theta}\max_{x\in B}\log\frac{P_\theta(x)}{Q(x)} \geq \sup_{\theta\in\Theta}\log\frac{P_\theta(x^*)}{Q(x^*)} \geq \sup_{\theta\in\Theta}\log\frac{P_\theta(x^*)}{Q^*(x^*)} = \log\frac{\sup_{\theta\in\Theta}P_\theta(x^*)}{Q^*(x^*)} = \log Z$$

Taking the infimum over all Q gives that $\rho^*(\Theta) \geq \log Z$ which concludes the result. $\square$

Definition: Gamma Function

Definition 11.8 The **Gamma function** is defined as

$$\Gamma(z) := \int_0^\infty x^{z-1} e^{-x} \, \mathrm{d}x.$$

Note that for all $n \in \mathbb{N}_0$, $\Gamma(n+1) = n!$.

Theorem: Shtarkov

Theorem 11.9 (Shtarkov) The minimax maximal redundancy over the class of all memoryless sources on A with $|A| = m$ satisfies, for all $n \in \mathbb{N}$,

$$\rho_n^* \leq \frac{m-1}{2} \log\left(\frac{n}{2}\right) + \log \frac{\Gamma(1/2)}{\Gamma(m/2)} + \frac{C'}{\sqrt{n}}$$

for a constant C depending only on m.

Proof Sketch. By Normalised Maximum Likelihood Code applied to the parametric family of all IID distributions P^n on A^n, we have

$$\rho_n^* = \log\left(\sum_{x_1^n \in A^n} \sup_P P^n(x_1^n) \right).$$

By Proposition 8.5, the MLE in this family is the empirical distribution $\hat{P}_n = \hat{P}_{x_1^n}$, so

$$\rho_n^* = \log\left(\sum_{x_1^n \in A^n} \hat{P}^n_{x_1^n}(x_1^n) \right).$$

Evaluating this (after some length calculations) gives the result. $\square$

11.3. Rissanen’s lower bound

Definition: Channel Capacity

Definition 11.10 Let $\{W(y \mid x) : x \in A, y \in B\}$ be a family of conditional PMFs $W(\cdot \mid x)$, describing the distribution of the output y of a discrete **channel** with input x. The **capacity** of the channel is

$$C = \sup I(X;Y),$$

where the supremum is over all jointly distribution RVs (X, Y), where X has an arbitrary distribution and the distribution of Y given X is $\mathbb{P}(Y = y \mid X = x) = W(y \mid x)$.

Theorem: Redundancy Capacity Theorem

Theorem 11.11 (Redundancy-capacity Theorem) Let $\{P_\theta : \theta \in \Theta\}$ be a "nice" parametric family of distributions on a finite alphabet B. Denote the minimax average redundancy over $\{P_\theta : \theta \in \Theta\}$ by

$$\overline{\rho}(\Theta) := \inf_Q \sup_{\theta \in \Theta} D(P_\theta \parallel Q).$$

Then $\overline{\rho}(\Theta)$ is equal to the capacity of the channel with input θ and output $X \sim P_\theta$:

$$\overline{\rho}(\Theta) = \max_\pi I(T; X),$$

where the maximum is over all probability distributions π on Θ, $T \sim \pi$ and $X \mid T = \theta \sim P_\theta$ (so the pair of RVs (T, X) has joint distribution $\pi(\theta)P_\theta(x)$).

Proof. Omitted (non-examinable). $\square$

Definition: Standard Parameterisation Of Pmfs On Alphabet

Definition 11.12 The standard parameterisation of the set of PMFS on $A = \{a_1, ..., a_m\}$ is $\{P_\theta : \theta \in \Theta\}$, where $\Theta = \left\{\theta \in [0,1]^{m-1} : \sum_{i=1}^{m-1} \theta_i \leq 1\right\}$ and

$$P_\theta(a_i) = \begin{cases} \theta_i & \text{if } i \neq m \\ 1 - \sum_{j=1}^{m-1} \theta_j & \text{if } i = m \end{cases}.$$

Theorem: Rissanen

Theorem 11.13 (Rissanen) Let Θ parametrise the set of PMFs on A, where $|A| = m$. Let $\{Q_n : n \in \mathbb{N}\}$ be an arbitrary sequence of distributions on A^n. Then for all $\varepsilon > 0$, there exists a constant C and a subset $\Theta_0 \subseteq \Theta$ of volume less than ε such that for all $\theta \notin \Theta_0$,

$$D(P_\theta^n \parallel Q_n) \geq \frac{m-1}{2} \log n - C \quad \text{eventually.}$$

In particular, $\overline{\rho}_n \geq \frac{m-1}{2} \log n - C'$ eventually for some constant C'.

Proof. Non-examinable. ∎

Corollary: Bounds On Minimax Maximal And Average Redundancies

Corollary 11.14 We have (eventually)

$$\frac{m-1}{2}\log n - C' \leq \overline{\rho}_n \leq \rho_n^* \leq \frac{m-1}{2}\log n + C$$

for some constants C, C'.

Remark 11.15 The above bound has a probabilistic interpretation: there exists a sequence of distributions $\{Q_n : n \in \mathbb{N}\}$ which are "uniformly close" to all product distributions:

$$-\log Q_n(x_1^n) \approx -\log P^n(x_1^n) + \frac{m-1}{2}\log n,$$

for all $P \in \mathcal{P}$ and $x_1^n \in A^n$. Moreover, the error term $\frac{m-1}{2}\log n$ is the best possible (up to addition of constants).