Information Theory, Part II

Manan Shah
manan.shah.777@gmail.com
The Harker School

February 23, 2017

This document contains lecture notes from Harker's Advanced Topics in Mathematics class in Information Theory II, taught by Dr. Anuradha Aiyer. This course is the second part of a two part offering that explores the basic concepts of Information Theory, as initially described by Claude Elwood Shannon at Bell Labs in 1948. These notes were taken using TeXShop and LaTeX2ϵ and will be updated for each class. The reader is advised to note any errata at the source control repository https://github.com/mananshah99/infotheory.

Contents

1 Unit 1: Gambling

We'll discuss the duality between the growth rate of investment (i.e. a horse race) and the entropy rate of the horse race and how the side information's financial value is tied to mutual information.

Definition 1 (Horse Race)**.** We have m horses in a race in which the ith horse wins with probability p_i. If horse i wins, the payoff is o_i for 1[1]. We'll assume that the gambler invests his wealth across all horses and doesn't hold on to any of his money. Specifically, b_i is the fraction of wealth invested in horse i where $b_i \geq 0$ and $\sum b_i = 1$. If horse i wins, the gambler wins $o_i b_i$; this case occurs with probability p_i. The wealth at the end of the race is a random variable which we will attempt to maximize.

1.1 Repeated Gambling

Define S_n as the total growth in the gambler's wealth after n races. We have

$$S_n = \prod_{j=1}^{n} S(X_j) \qquad \text{and} \qquad S(X_j) = b(X_j)o(X_j)$$

with X representing the horse that wins (this changes between races). Here, $S(X_i)$ represents the factor by which the gambler's wealth grows. We can define the doubling rate of a race W as

$$E(\log S(X)) = \sum_{k=1}^{m} p_k \log(b_k o_k) = W(b, p)$$

Theorem 1. Let race outcomes $X_1, X_2, X_3, \ldots, X_n$ be identically and independently distributed $\sim p(x)$. The wealth of a gambler using betting strategy b grows exponentially at the rate $W(b, p)$ such that $S_n = 2^{nW(b,p)}$.

Proof. Functions of independent random variables are also independent, so $\log S(X_1), \ldots \log S(X_n)$ are i.i.d. From our earlier definition of S_n we have

$$\frac{1}{n} \log S_n = \frac{1}{n} \sum \log S(X_i)$$

By the weak law of large numbers[2], this equates to $E(\log S(X)) = W(b, p)$. So we can conclude that $S_n = 2^{nW(b,p)}$ and the proof is complete. So if to maximize S_n, we'll need to maximize W. □

Definition 2. The optimum doubling rate over all choices of b_i is

$$W^*(p) = \max_b W(b, p) = \max_{b\colon b_i \geq 0, \sum b_i = 1} \sum_i p_i \log b_i o_i$$

We must formally maximize $W(b, p)$ such that $\sum b_i = 1$. To do this, we'll apply Lagrange optimization. We have

$$J(b) = \sum p_i \log b_i o_i + \lambda \sum b_i$$

[1] There are two ways to describe a bet: either a for 1 or b to 1. The first notation indicates an exchange that happens prior to the race, and the latter indicates and exchange that happens post-race (although in both cases the horses are picked before the race). More concretely, a for 1 indicates that if one places \$1 on a particular horse before the race, the payoff is \$a iff the horse wins and \$0 if the horse loses. b to 1 indicates that one would pay \$1 after the race if a particular horse loses and win \$b if the horse wins. The equivalency between these scenarios is $b = a - 1$.

[2] See `http://mathworld.wolfram.com/WeakLawofLargeNumbers.html` for more information.

Taking the partial with respect to b_i and setting it equal to 0,

$$\frac{\partial J}{\partial b_i} = \frac{p_i}{b_i} + \lambda$$

where $i \in \{1 \ldots m\}$. Solving for b_i as a function of p_i and λ, substituting the resulting value into the constraint $\sum b_i = 1$, and evaluating the differential expression with $\lambda = -1$ results in $b_i = p_i$. Technically, we'd have to take the second derivative to prove that this is a maximum; this verification is left to the reader.

Theorem 2. $W^* = \sum p_i \log o_i - H(p)$

Proof.

$$\begin{aligned} W(b,p) &= \sum p_i \log b_i o_i \\ &= p_i \log \left(\frac{b_i}{p_i} \times p_i o_i \right) \\ &= \sum p_i \log o_i - H(p) - D(p||b) \end{aligned}$$

The last term, D, is known as relative entropy. It has some of the same properties of entropy, one of them being that $D \geq 0$. So, $W(b,p) \leq \sum p_i \log o_i - H(p)$ with equality when $p = b$. □

1.2 Kullback-Liebler Divergence

The function D, known as the relative entropy or Kullback-Liebler Divergence, is a measure of distance[3] between two distributions. If p and q are the two distributions, then $D(p||q)$ is a measure of inefficiency of assuming q when the true distribution is p. The average code length for distribution p is $H(p)$, but if we were to use the code for q to encode p, then $H(p) + D(p||q)$ bits.

Definition 3 (Kullback-Liebler Divergence)**.** The KL divergence $D(p||q)$ is expressed as

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_x \left[\log \frac{p(x)}{q(x)} \right]$$

where $0 \log 0/q = 0$ and $p \log p/0 = \infty$. We can then write $I(X;Y) = \sum \sum p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$ which is simplified to $D(p(x,y)||p(x)p(Y))$

Example 1. $p(0) = 1 - r, q(0) = 1 - s, p(1) = r, q(1) = s$

$$D(p||q) = (1-r) \log \frac{1-r}{1-s} + r \log \frac{r}{s}$$

$$D(q||p) = (1-s) \log \frac{1-s}{1-r} + s \log \frac{s}{r}$$

Example 2. Consider a case with two horses where horse 1 wins with probability p_1 and horse 2 wins with p_2. Assume even odds (2-for-1). (a) What is the optimal bet? (b) Doubling rate? (c) Resulting wealth?

(a) The optimal bet is according to the probabilities of the horses, (b) The doubling rate is $1 - H(p)$, and the resulting wealth (c) is $2^{n(1-H(p))}$

[3] This isn't technically a measure of distance as it doesn't satisfy the triangle inequality

We further have that $W(b,p) = \sum p \log \frac{p_i}{r_i} - \sum p \log \frac{p}{b} = D(p||r) - D(p||b)$ where $r_i = 1/o_i$. The doubling rate is the difference between the distance of the bookie's estimates from the truth. The gambler only makes money when b is closer than r. When the odds are m-for-1, we have

$$W^*(p) = D(p||1/m) = \log m - H(p)$$

and $W^*(p) + H(p) = \log m$.

Example 3. Three horses run a race. A gambler offers 3-for-1 odds on each horse. Fair odds under the assumption that all horses are equally likely to win. $p = (1/2, 1/4, 1/4)$. (a) Expected wealth, (b) b^*, (c) W^*

(a) We have that $W(b) = \sum p_i \log b_i o_i = \sum p_i \log 3b$ since $o_i = 1/3$ due to fair odds. Therefore, $W(b) = \sum p_i \log 3 + \sum p_i \log b_i = \log 3 + \sum p_i \log b_i$. (b) $b^* = p = (1/2, 1/4/, 1/4)$ and (c) $W^* = W(b^*) - \log 3 - 3/2$. Note that we can solve (c) with the identity discussed above.

1.3 The Value of Side Information

One measure is the increase in the doubling rate based on the information. We'll connect this increase with mutual information (as we connected W^* with KL divergence and entropy before). Define $X \in \{1, 2, \dots m\}$ as the horse betting space, $p(x)$ as the probabilities associated with $1 \to m$, $o(x)$ for 1 odds, and y as the side information. Furthermore, we have $\sum_x b(x|y)$ as the conditional betting depending on side information y and $b(x|y)$ as the proportion of wealth bet on horse x when y is observed. Based on these definitions, we have

$$W^*(X) = \max_{b(x)} \sum_x p(x) \log b(x) o(x)$$

and given our side information,

$$W^*(X|Y) = \max_{b(x|y)} \sum_{x,y} p(x,y) \log b(x|y) o(x)$$

so we have

$$\Delta W = W^*(X|Y) - W(X)$$

Theorem 3. The increase doubling rate ΔW due to side information Y for a horse race X is $\Delta W = I(X;Y)$.

Proof. We have that $b^*(x|y) = p(x|y)$. Since $W^*(X|Y) = \max_{b(x|y)} E(\log S)$[4]. This equates to $\max_{b(x|y)} \sum p(x,y) \log[o(x) b(x|y)]$. So, we have that

$$W^*(X|Y) = \sum p(x,y) \log[p(x)p(x|y)] = \sum p(x) \log o(x) - H(X|Y)$$

Without side information $W^* = \sum p(x) \log o(x) - H(X)$, so we have $\Delta W = \sum p(x) \log o(x) - H(X|Y) - [\sum p(x) \log o(x) - H(X)]$. Finally, we have $\Delta W = H(X) - H(X|Y) = I(X;Y)$. □

Example 4. Given a three horse race $p = (1/2, 1/4, 1/4)$ with odds with respect to the false distribution $r_1, r_2, r_3 = (1/4, 1/4, 1/2)$ and $o_1, o_2, o_3 = (4, 4, 2)$[5]. Find (a) the entropy of the race and (b) (b_1, b_2, b_3) such that compounded wealth $\to \infty$.

The entropy of the race is easily calculated as $3/2$. It's intuitive that $b_i = o_i p_i$ so we have $(2, 1, 1/2)$, which we re-normalize to $(4/7, 2/7, 1/7)$. Our final $W = \sum p_i \log b_i o_i$.

[4] S was defined earlier as the aggregate wealth

[5] This is because $o_i = 1/r_i$ when determining the odds given the false distribution. "Fair odds" are defined such that $\sum 1/o_i = 1$

Example 5. Let the distribution be (p_1, p_2, p_3) with odds $o = (1, 1, 1)$ and wealth proportions $b = (b_1, b_2, b_3)$. $S_n \to 0$ exponentially. (a) Find the exponent, (b) b^*, and (c) What p causes $S_n \to 0$ at the fastest rate.

We always have that $b_i = \frac{p_i o_i}{\sum_i b_i}$, so we can write $b^* = p$. Furthermore, the exponent is simply the doubling rate $W = \sum p_i \log b_i o_i$, and the P that causes $S_n \to 0$ most quickly is the one that maximizes $H(p)$ or $p = (1/3, 1/3, 1/3)$.

Example 6. Now assume you have the most common form of side information, the past performance of the horses. If the results of each successive horse race are independent, then this additional information will not reveal anything new about the race. Suppose instead that the races are dependent. How might we calculate $W^*(X_k|X_{k-1}, X_{k-2}, ..., X_1)$?

At the end of of n races,

$$S_n = \prod_i S(X_i)$$

Therefore,

$$\begin{aligned}\frac{1}{n}\mathrm{E}\left[\log S_n\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\log S(X_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\log m - H(X_i|...)\right] \\ W^* &= \log m - \frac{H(X_1, X_2, ..., X_n)}{n}\end{aligned}$$

The term $\frac{H(X_1, X_2, ..., X_n)}{n}$ can be thought of as the average entropy across all the races.

2 Unit 2: Statistics

2.1 Introduction

Statistics considers two types of studies: observational and experimental. In an experimental study, treatments are assigned to subjects; this is not the case in observational studies. The first part of this topic was covered with traditional statistics worksheets involving the definition of p-values and statistical tests (t, z, etc.) A diagram that connects these concepts is as follows:

Information Theory	Statistics	Machine Learning
Source	Observational Studies	Unsupervised Learning
Channel	Experimental Studies	Supervised Learning

We'll start by modeling the source, which has a distribution Q and outputs the vector X. We will define a notion of X as too ridiculous to have come from Q. It is incorrect to define the ridiculousness criterion as "$\mathrm{Prob}_Q(X)$ is small" as looking at one event whose probability will almost always be small is insufficient.

Definition 4. X is "too ridiculous" to have originated from source Q if and only if probability Prob_Q(X and its entire subsequent tail) is sufficiently small[6]. That is to say, given X we can define

$$S_X = \{X' \mid \mathrm{Prob}_Q(X') \leq \mathrm{Prob}_Q(X)\}$$

[6] Define ϵ as in traditional proofs to quantify this

We then have that X is too ridiculous to have come from Q if the probability $\text{Prob}_Q(S_X)$ is sufficiently small.

Definition 5 (Confidence Interval)**.** Given X, identify all Qs it may have originated from. The confidence interval[7] is defined as

$$\{Q \mid X \text{ is not too ridiculous to have originated from } Q\}$$

Definition 6 (Hypothesis Test)**.** Given hypothesis Q, find all X that will allow for disproving Q. The rejection interval (or region) is defined as

$$\{X \mid X \text{ is too ridiculous to have come from } Q\}$$

2.2 Two Significant Theorems

We'll discuss two important theorems that define the notions of $\text{Prob}_Q(X)$ and $\text{Prob}_Q(S)$.

2.2.1 Prob$_Q(X)$

Assume that we are given a source Q that produces observed data outputs $X_1 \dots X_N$ (all abbreviated as the vector X) and that the values are identically and independently distributed. We will attempt to define the value $\text{Prob}_Q(X)$. If we were to histogram the vector X (with the y axis representing the frequency of occurrences of ξ in X and the x axis representing the discrete values $\xi_1 \dots \xi_k)^8$, each value ξ_i would have an associated frequency N_i. Call this histogram P_X with $N_1 + N_2 + \dots + N_k = N$. We then have

$$\begin{aligned}
\text{Prob}_Q(X) &= Q(X_1) \times Q(X_2) \times \dots \times Q(X_N) \\
&= Q(\xi_1)^{N_1} \times Q(\xi_2)^{N_2} \times \dots \times Q(\xi_k)^{N_k} \\
&= 2^{-[N_1 \log \frac{1}{Q(\xi_1)} + N_2 \log \frac{1}{Q(\xi_2)} + \dots + N_k \log \frac{1}{Q(\xi_k)}]} \\
&= 2^{-N[P_X(\xi_1) \log \frac{1}{Q(\xi_1)} + \dots + P_X(\xi_k) \log \frac{1}{Q(\xi_k)}]}
\end{aligned}$$

where $Q(X_i)$ is the probability of seeing X_i in the distribution of Q. We can multiply each log term by $P_X(\xi_i)$ on the numerator and denominator to express the value as a function of the KL divergence and entropy. We therefore have the following result[9].

Theorem 4. $\text{Prob}_Q(X) = 2^{-N[D(P_X||Q)+H(P_X)]}$

2.2.2 Prob$_Q(S)$

We can begin by writing

$$\begin{aligned}
\text{Prob}_Q(S) &= \sum_{X \in S} \text{Prob}_Q(X) \\
&= \sum_{X \in S} 2^{-N[D(P_X||Q)+H(P_X)]}
\end{aligned}$$

[7] It's better to write this as a confidence region as opposed to a confidence interval if we're working in spaces of higher dimensionality than $\mathbb{R}^1$

[8] X comprises of discrete values that are represented by ξ

[9] We've only proved the result for a discrete i.i.d distribution, but it can be shown to be applicable to continuous distributions (with differential entropy). This result cannot, however, be extended to non-i.i.d distributions because of the first step

With Q representing the true distribution, we have a space of histograms $\{P_X \, \forall \, X \in S\}$. We want to identify the distribution P^* that is "closest" to Q. By the Pythagorean inequality (which we'll prove later), we can write the above expression as

$$\text{Prob}_Q(S) \leq \sum_{X \in S} 2^{-N[D(P_X||P^*)+D(P^*||Q)+H(P_X)]} = 2^{-ND(P^*||Q)} \sum_{X \in S} 2^{-N[D(P_X||P^*)+H(P_X)]}$$

which can be written as

$$2^{-ND(P^*||Q)} \sum_{X \in S} \text{Prob}_{P^*}(X) = 2^{-ND(P^*||Q)} \text{Prob}_{P^*}(S)$$

Since the probability term is less than or equal to one, we've therefore bounded $\text{Prob}_Q(S)$.

Theorem 5 (Sanov's Theorem)**.** $\text{Prob}_Q(S) \leq 2^{-ND(P^*||Q)}$

2.2.3 Proof of the Pythagorean Inequality

Theorem 6 (Pythagorean Inequality)**.** $D(P||Q) \geq D(P||P^*) + D(P^*||Q)$ where Q is the true distribution and P is the estimated distribution. P^* is the closest distribution in terms of KL divergence.

Proof. Consider any P in the convex vector space of histograms S. By the definition of a convex vector space,

$$P_\lambda = \lambda P + (1 - \lambda) P^*$$

If we let $\lambda = 0$, $P_\lambda = P^*$, but since we know that P^* is defined to be the minimum of all P in set S (according to the condition $D(P^*||Q) = \min_{P \in S} D(P||Q)$), we know that it is also the minimum of $D(P_\lambda||Q)$ along the path $P^* \to P$. So, $D_\lambda = D(P_\lambda||Q) = \sum P_\lambda \log P_\lambda / Q$. Therefore, $dD_\lambda / d\lambda$ as a function of λ is nonnegative at $\lambda = 0$ because as $\lambda \to 0$, the divergence decreases. Specifically,

$$\frac{dD_\lambda}{d\lambda} = \sum_{\text{bins}} \left[(P - P^*) + (P - P^*) \log \frac{P_\lambda}{Q} \right]$$

At $\lambda = 0$, $P_\lambda = P^*$. We also know that $\sum P(X) = \sum P^*(X) = 1$ by the property of a pdf. We can then write

$$\left. \frac{dD_\lambda}{d\lambda} \right|_{\lambda=0} = \sum [P(X) - P^*(X)] \log \frac{P^*}{Q}$$

This simplifies to $\sum P \log \frac{P^*}{Q} - \sum P^* \log \frac{P^*}{Q} \geq 0$. The first term can be written as $\sum P \log \frac{P^*}{P} \frac{P}{Q} - \sum P^* \log \frac{P^*}{Q}$ which completes the proof (as the initial point $\lambda = 0 \geq 0$ so the remainder of the function is). □

2.3 Connections to Statistics

From the definition of KL divergence, we can write $D[\text{Bernoulli}(p_1)||\text{Bernoulli}(p_2)]$ as

$$p_1 \log \frac{p_1}{p_2} + (1 - p_1) \log \frac{1 - p_1}{1 - p_2}$$

which is equivalent to[10]

$$\frac{1}{\ln 2} \left[p_1 \ln \frac{p_1}{p_2} + (1 - p_1) \ln \frac{1 - p_1}{1 - p_2} \right]$$

[10] Converting log to ln

A second order approximation of $\ln(1/1-\epsilon) \approx \epsilon + \epsilon^2/2$ allows us to write

$$\frac{1}{\ln 2}\left[p_1\left(\left(1-\frac{p_2}{p_1}\right)+\frac{1}{2}\left(1-\frac{p_1}{p_2}\right)^2\right)+(1-p_1)\left(\left(1-\frac{1-p_2}{1-p_1}\right)-\frac{1}{2}\left(1-\frac{1-p_2}{1-p_1}\right)^2\right)\right]$$

which simplifies to

$$\left(-\frac{1}{2\ln 2}\right)\frac{(p_1-p_2)^2}{p_1(1-p_1)}$$

What we've just obtained is an approximation to the KL divergence between two binomials. Note for future reference that

$$D(f||q) = \int_x f \log\frac{f}{q}dx$$

2.3.1 Connection 1: Observation Studies for Proportion

Consider a binomial distribution with probability p originating from source Q (that outputs values $x_1, x_2, \ldots, x_N$). We've defined X is too ridiculous to have come from Q as $2^{-ND(P^*||Q)} < \epsilon$ (by extension of Sanov's theorem). But since P^* (our best guess of the distribution) is equivalent to Binomial($\hat{p}$)[11], we can write

$$2^{-N\left[\frac{1}{2\ln 2}\frac{(\hat{p}-p)^2}{\hat{p}(1-\hat{p})}\right]} < \epsilon$$

Which is equivalent to

$$\frac{1}{2\ln 2}\frac{(\hat{p}-p)^2}{(\hat{p}(1-\hat{p}))/N} > \log\frac{1}{\epsilon}$$

We can rewrite this as

$$\frac{(\hat{p}-p)^2}{(\hat{p}(1-\hat{p}))/N} > \ln\frac{1}{\epsilon^2}$$

Taking the square root of both sides, we have

$$\frac{|\hat{p}-p|}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}} > \underbrace{\sqrt{\ln\frac{1}{\epsilon^2}}}_{\text{call this constant } C}$$

We can now express this in a more familiar form. In particular, given $X = \{P \mid P$ lies in the interval $\hat{p} \pm C\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$ $\}$, then we have the hypothesis test that $\hat{p}$ is "too ridiculous" to have come from X if it lies in the rejection region $\{ \frac{|\hat{p}-p|}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}} > C \}$.

2.3.2 Connection 2

[11] That is, we make our best guess given the observed probability distribution

Appendix A—Quotes

- “I have a problem. It’s called gambling.” (Dr. Aiyer)
- “What’s E? Entropy?” (David Zhu)
- “So is it the strong law of weak numbers?” (Jerry Chen)
- “It’s like a half life... but it’s a double life” (Steven Cao)
- “Isn’t this just Lagrange?”
  10 minutes later..
  “Wait, how do you do Lagrange again?” (Swapnil Garg)
- “I’m just amazed that you manage to learn something” (Dr. Aiyer)
- (Looking at Σ) That’s a backwards ξ! (Steven)
- “What’s a binomial distribution?” (Dr. Aiyer)
  “$a + bx$” (Shaya)
- “An approximation of $\ln(1 - x)$ is $1 + x + x^2 + x^3 + \ldots$” (Misha Ivkov)
- “Where does ϵ go to get a haircut?”
  “The epSALON” (Shaya)