Information Theory, Part II

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

January 26, 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. In Part 2 of the course, we explore other applications of Information Theory to the disciplines of Gambling, Statistics, Physics, Computer science, Economics and Philosophy. 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

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

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 $i \in \{1 \dots m\}$. $\sum b_i = 1$ 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, we have that $W(b,p) \leq \sum p_i \log o_i - H(p)$ 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. 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 $W(b,p) = \sum p \log \frac{p_i}{r_i} - \sum p \log \frac{p}{b} = D(p||r) - D(p||b)$. 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 $\Delta W = H(X) - H(X|Y) = I(X;Y)$. $\square$

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 $p = (1/3, 1/3, 1/3)$.

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)