Towards Shannon's Entropy Theorem

Contents

Chapter 1

Kraft–McMillan inequality

In coding theory, the **Kraft–McMillan inequality** gives a necessary and sufficient condition for the existence of a prefix code[1] (in Kraft's version) or a uniquely decodable code (in McMillan's version) for a given set of codeword lengths. Its applications to prefix codes and trees often find use in computer science and information theory.

Kraft's inequality was published in Kraft (1949). However, Kraft's paper discusses only prefix codes, and attributes the analysis leading to the inequality to Raymond Redheffer. The result was independently discovered in McMillan (1956). McMillan proves the result for the general case of uniquely decodable codes, and attributes the version for prefix codes to a spoken observation in 1955 by Joseph Leo Doob.

1.1 Applications and intuitions

Kraft's inequality limits the lengths of codewords in a prefix code: if one takes an exponential of the length of each valid codeword, the resulting set of values must look like a probability mass function, that is, it must have total measure less than or equal to one. Kraft's inequality can be thought of in terms of a constrained budget to be spent on codewords, with shorter codewords being more expensive. Among the useful properties following from the inequality are the following statements:

- If Kraft's inequality holds with strict inequality, the code has some redundancy.
- If Kraft's inequality holds with equality, the code in question is a complete code.
- If Kraft's inequality does not hold, the code is not uniquely decodable.
- For every uniquely decodable code, there exists a prefix code with the same length distribution.

1.2 Formal statement

Let each source symbol from the alphabet

$S = \{ s_1, s_2, \ldots, s_n \}$

be encoded into a uniquely decodable code over an alphabet of size r with codeword lengths

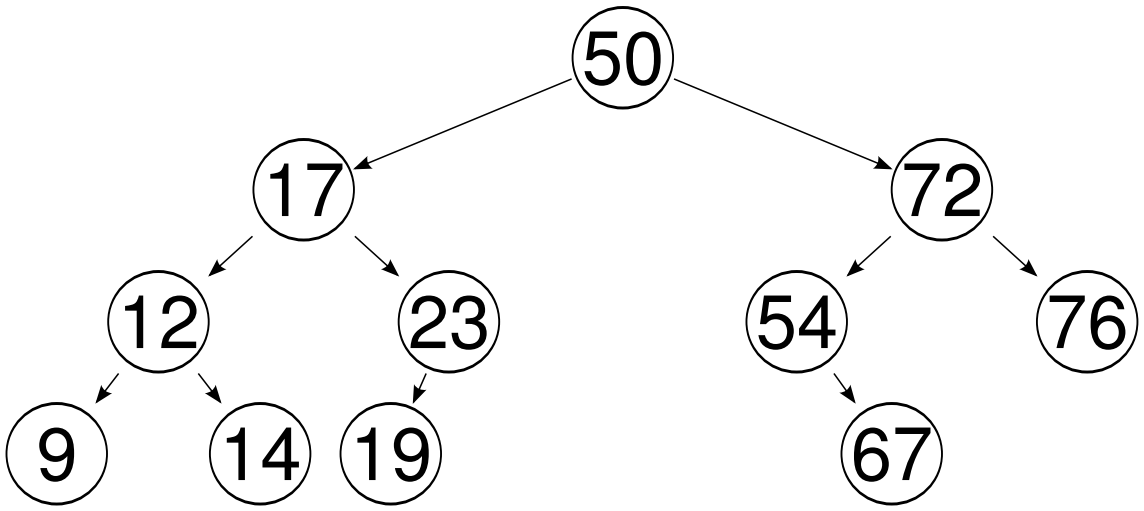
$\ell_1, \ell_2, \ldots, \ell_n.$

Then

$$\sum_{i=1}^{n} r^{-\ell_i} \leqslant 1.$$

Conversely, for a given set of natural numbers $\ell_1, \ell_2, \ldots, \ell_n$ satisfying the above inequality, there exists a uniquely decodable code over an alphabet of size r with those codeword lengths.

1.3 Example: binary trees



9, 14, 19, 67 and 76 are leaf nodes at depths of 3, 3, 3, 3 and 2, respectively.

Any binary tree can be viewed as defining a prefix code for the leaves of the tree. Kraft's inequality states that

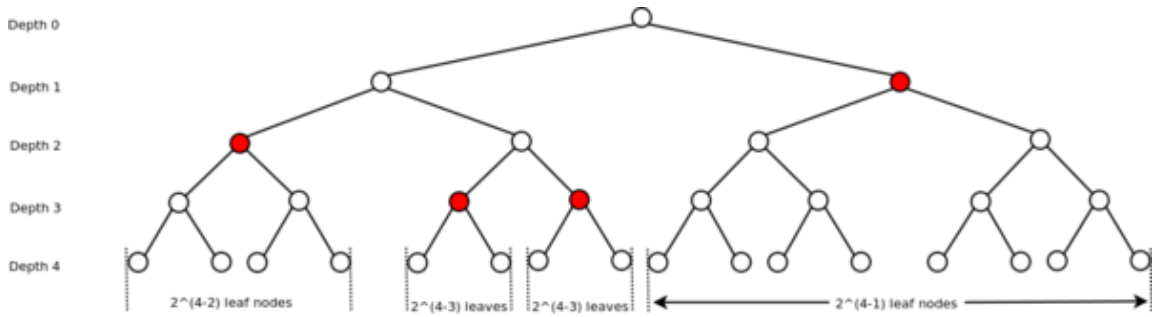
$$\sum_{\ell \in \text{leaves}} 2^{-\text{depth}(\ell)} \leqslant 1.$$

Here the sum is taken over the leaves of the tree, i.e. the nodes without any children. The depth is the distance to the root node. In the tree to the right, this sum is

$$\frac{1}{4} + 4\left(\frac{1}{8}\right) = \frac{3}{4} \leqslant 1.$$

1.4 Proof

1.4.1 Proof for prefix codes



Example for binary tree. Red nodes represent a prefix tree. The method for calculating the number of descendant leaf nodes in the full tree is shown.

First, let us show that the Kraft inequality holds whenever S is a prefix code.

Suppose that $\ell_1 \leqslant \ell_2 \leqslant \cdots \leqslant \ell_n$. Let A be the full r -ary tree of depth ℓ_n (thus, every node of A at level $< \ell_n$ has r children, while the nodes at level ℓ_n are leaves). Every word of length $\ell \leqslant \ell_n$ over an r -ary alphabet corresponds to a node in this tree at depth ℓ . The i th word in the prefix code corresponds to a node v_i ; let A_i be the set of all leaf nodes (i.e. of nodes at depth ℓ_n) in the subtree of A rooted at v_i . That subtree being of height $\ell_n - \ell_i$, we have

$$|A_i| = r^{\ell_n - \ell_i}.$$

Since the code is a prefix code, those subtrees cannot share any leaves, which means that

$$A_i \cap A_j = \varnothing, \quad i \neq j.$$

Thus, given that the total number of nodes at depth ℓ_n is r^{ℓ_n} , we have

$$\left| \bigcup_{i=1}^{n} A_i \right| = \sum_{i=1}^{n} |A_i| = \sum_{i=1}^{n} r^{\ell_n - \ell_i} \leqslant r^{\ell_n}$$

from which the result follows.

Conversely, given any ordered sequence of n natural numbers,

$$\ell_1 \leqslant \ell_2 \leqslant \cdots \leqslant \ell_n$$

satisfying the Kraft inequality, one can construct a prefix code with codeword lengths equal to each ℓ_i by choosing a word of length ℓ_i arbitrarily, then ruling out all words of greater length that have it as a prefix. There again, we shall interpret this in terms of leaf nodes of an r -ary tree of depth ℓ_n . First choose any node from the full tree at depth ℓ_1 ; it corresponds to the first word of our new code. Since we are building a prefix code, all the descendants of this node (i.e., all words that have this first word as a prefix) become unsuitable for inclusion in the code. We consider the descendants at depth ℓ_n (i.e., the leaf nodes among the descendants); there are $r^{\ell_n - \ell_1}$ such descendant nodes that are removed from consideration. The next iteration picks a (surviving) node at depth ℓ_2 and removes $r^{\ell_n - \ell_2}$ further leaf nodes, and so on. After n iterations, we have removed a total of

$$\sum_{i=1}^{n} r^{\ell_n - \ell_i}$$

nodes. The question is whether we need to remove more leaf nodes than we actually have available — r^{ℓ_n} in all — in the process of building the code. Since the Kraft inequality holds, we have indeed

$$\sum_{i=1}^{n} r^{\ell_n - \ell_i} \leqslant r^{\ell_n}$$

and thus a prefix code can be built. Note that as the choice of nodes at each step is largely arbitrary, many different suitable prefix codes can be built, in general.

1.4.2 Proof of the general case

Now, we will prove that the Kraft inequality holds whenever S is a uniquely decodable code. (The converse needs not be proven, since we have already proven it for prefix codes, which is a stronger claim.)

Consider the generating function in inverse of x for the code S

$$F(x) = \sum_{i=1}^{n} x^{-|s_i|} = \sum_{\ell=\min}^{\max} p_\ell \, x^{-\ell}$$

in which p_ℓ —the coefficient in front of $x^{-\ell}$ —is the number of distinct codewords of length ℓ . Here min is the length of the shortest codeword in *S*, and max is the length of the longest codeword in *S*.

Consider all *m*-powers S^m, in the form of words $s_{i_1} s_{i_2} \dots s_{i_m}$, where $i_1, i_2, \dots, i_m$ are indices between 1 and *n*. Note that, since *S* was assumed to uniquely decodable, $s_{i_1} s_{i_2} \dots s_{i_m} = s_{j_1} s_{j_2} \dots s_{j_m}$ implies $i_1 = j_1, i_2 = j_2, \dots, i_m = j_m$. Because of this property, one can compute the generating function $G(x)$ for S^m from the generating function $F(x)$ as

$$\begin{aligned} G(x) = (F(x))^m &= \left(\sum_{i=1}^{n} x^{-|s_i|} \right)^m \\ &= \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_m=1}^{n} x^{-\left(|s_{i_1}|+|s_{i_2}|+\cdots+|s_{i_m}|\right)} \\ &= \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_m=1}^{n} x^{-|s_{i_1} s_{i_2} \cdots s_{i_m}|} = \sum_{\ell=m\cdot\min}^{m\cdot\max} q_\ell \, x^{-\ell} \ . \end{aligned}$$

Here, similarly as before, q_ℓ — the coefficient in front of $x^{-\ell}$ in $G(x)$ — is the number of words of length ℓ in S^m . Clearly, q_ℓ cannot exceed r^ℓ . Hence for any positive *x*,

$$(F(x))^m \le \sum_{\ell=m\cdot\min}^{m\cdot\max} r^\ell \, x^{-\ell} \ .$$

Substituting the value *x* = *r* we have

$$(F(r))^m \le m \cdot (\max - \min) + 1$$

for any positive integer m . The left side of the inequality grows exponentially in m and the right side only linearly. The only possibility for the inequality to be valid for all m is that $F(r) \le 1$. Looking back on the definition of $F(x)$ we finally get the inequality.

$$\sum_{i=1}^{n} r^{-\ell_i} = \sum_{i=1}^{n} r^{-|s_i|} = F(r) \le 1 \ .$$

1.4.3 Alternative construction for the converse

Given a sequence of n natural numbers,

$$\ell_1 \leqslant \ell_2 \leqslant \cdots \leqslant \ell_n$$

satisfying the Kraft inequality, we can construct a prefix code as follows. Define the i^{th} codeword, C_i, to be the first ℓ_i digits after the radix point (e.g. decimal point) in the base *r* representation of

$$\sum_{j=1}^{i-1} r^{-\ell_j} .$$

Note that by Kraft's inequality, this sum is never more than 1. Hence the codewords capture the entire value of the sum. Therefore, for *j* > *i*, the first ℓ_i digits of *Cj* form a larger number than *Ci*, so the code is prefix free.

1.5 Notes

[1] Cover, Thomas M.; Thomas, Joy A. (2006), "Data Compression", *Elements of Information Theory* (PDF) (2nd ed.), John Wiley & Sons, Inc, pp. 108–109, ISBN 0-471-24195-4, doi:10.1002/047174882X.ch5

1.6 References

- Kraft, Leon G. (1949), *A device for quantizing, grouping, and coding amplitude modulated pulses*, Cambridge, MA: MS Thesis, Electrical Engineering Department, Massachusetts Institute of Technology.

- McMillan, Brockway (1956), "Two inequalities implied by unique decipherability", *IEEE Trans. Information Theory*, **2** (4): 115–116, doi:10.1109/TIT.1956.1056818.

1.7 See also

Chaitin's constant, Canonical Huffman code.

Chapter 2

Gibbs' inequality

In information theory, **Gibbs' inequality** is a statement about the mathematical entropy of a discrete probability distribution. Several other bounds on the entropy of probability distributions are derived from Gibbs' inequality, including Fano's inequality. It was first presented by J. Willard Gibbs in the 19th century.

2.1 Gibbs' inequality

Suppose that

$$P = \{p_1, \ldots, p_n\}$$

is a probability distribution. Then for any other probability distribution

$$Q = \{q_1, \ldots, q_n\}$$

the following inequality between positive quantities (since the p_i and q_i are positive numbers less than one) holds[1]:68

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i=1}^{n} p_i \log_2 q_i$$

with equality if and only if

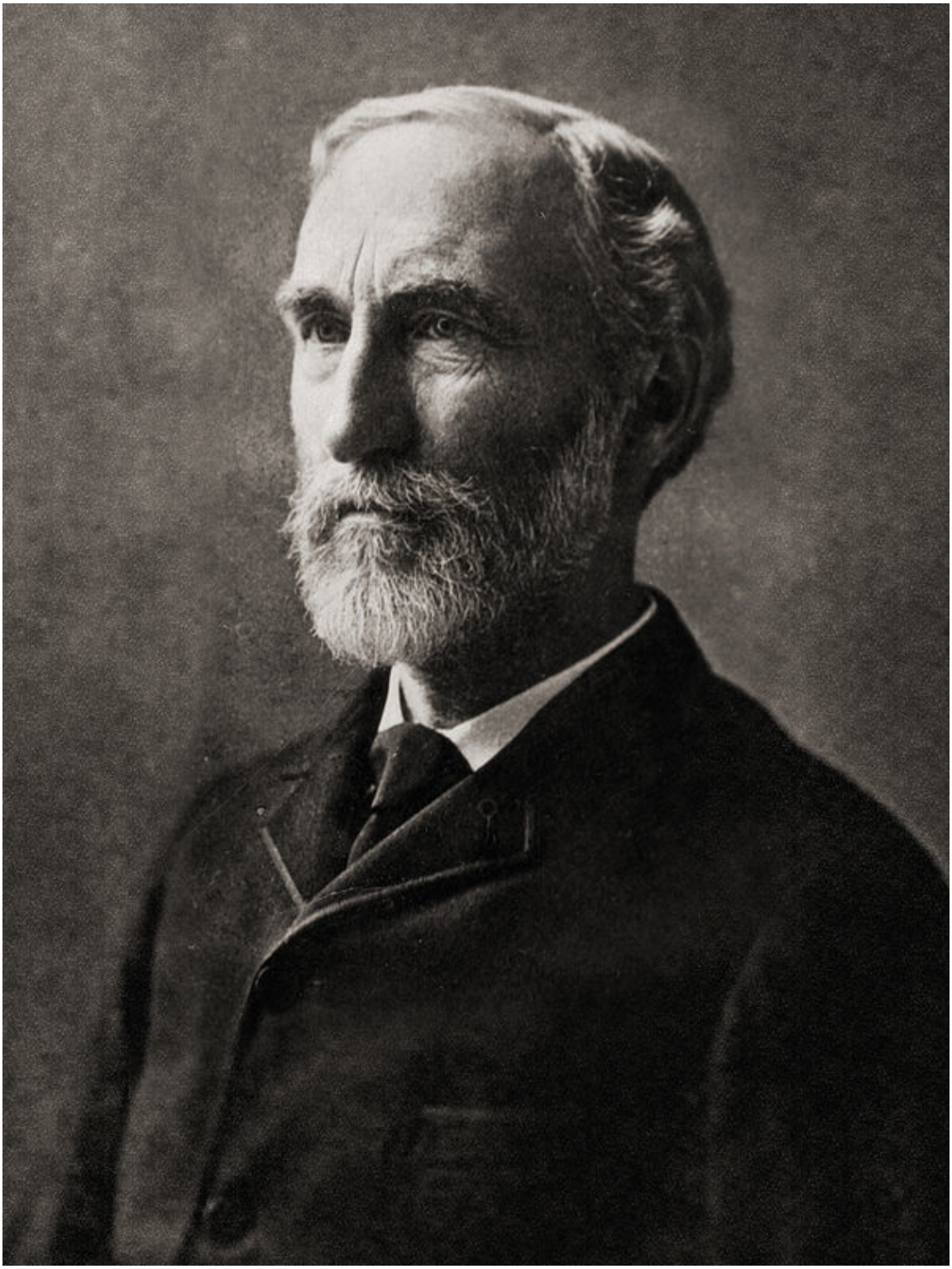
$$p_i = q_i$$

for all *i*. Put in words, the information entropy of a distribution P is less than or equal to its cross entropy with any other distribution Q.

The difference between the two quantities is the Kullback–Leibler divergence or relative entropy, so the inequality can also be written:[2]:34

$$D_{\mathrm{KL}}(P\|Q) \equiv \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i} \geq 0.$$

Note that the use of base-2 logarithms is optional, and allows one to refer to the quantity on each side of the inequality as an "average surprisal" measured in bits.



Josiah Willard Gibbs

2.2 Proof

Since

$$\log_2 a = \frac{\ln a}{\ln 2}$$

it is sufficient to prove the statement using the natural logarithm (ln). Note that the natural logarithm satisfies

$$\ln x \leq x - 1$$

for all $x > 0$ with equality if and only if $x=1$.
Let I denote the set of all i for which p_i is non-zero. Then

$$-\sum_{i\in I} p_i \ln \frac{q_i}{p_i} \geq -\sum_{i\in I} p_i \left(\frac{q_i}{p_i} - 1\right)$$

$$= -\sum_{i\in I} q_i + \sum_{i\in I} p_i$$

$$\geq 0.$$

So

$$-\sum_{i\in I} p_i \ln q_i \geq -\sum_{i\in I} p_i \ln p_i$$

and then trivially

$$-\sum_{i=1}^{n} p_i \ln q_i \geq -\sum_{i=1}^{n} p_i \ln p_i$$

since the right hand side does not grow, but the left hand side may grow or may stay the same.
For equality to hold, we require:

1. $\frac{q_i}{p_i} = 1$ for all $i \in I$ so that the approximation $\ln \frac{q_i}{p_i} = \frac{q_i}{p_i} - 1$ is exact.
2. $\sum_{i\in I} q_i = 1$ so that equality continues to hold between the third and fourth lines of the proof.

This can happen if and only if

$$p_i = q_i$$

for $i = 1, ..., n$.

2.3 Alternative proofs

The result can alternatively be proved using Jensen's inequality or log sum inequality.

2.4 Corollary

The entropy of P is bounded by:[1]:68

$$H(p_1, \ldots, p_n) \leq \log n.$$

The proof is trivial - simply set $q_i = 1/n$ for all i.

2.5 See also

- Information entropy

2.6 References

[1] Pierre Bremaud (6 December 2012). *An Introduction to Probabilistic Modeling*. Springer Science & Business Media. ISBN 978-1-4612-1046-7.

[2] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. ISBN 978-0-521-64298-9.

2.7 Text and image sources, contributors, and licenses

2.7.1 Text

- **Kraft–McMillan inequality** *Source:* https://en.wikipedia.org/wiki/Kraft%E2%80%93McMillan_inequality?oldid=795574646 *Contributors:* Michael Hardy, Ixfd64, Charles Matthews, Dina, Giftlite, Roo72, Spoon!, Pearle, Jheald, Xiaoyanggu, Oleg Alexandrov, Milez, Arunkumar, Mikm, Reetep, YurikBot, Dantheox, David Pal, Javalenok, FilipeS, ElPoojmar, Thijs!bot, Heysan, Magioladitis, David Eppstein, Twotonkatrucks, Camrn86, Cooperh, Akshatsinghal, Vitz-RS, Leena34, Gamall Wednesday Ida, Mko3okm, Mjoachimiak, Addbot, Yobot, Li3939108, Citation bot, Xqbot, Mr.gondolier, D'ohBot, Darij, MaxDel, MondalorBot, Trappist the monk, Patmorin, RjwilmsiBot, ChuispastonBot, Ahmahran, Deacon Vorbis, Quinton Feldberg and Anonymous: 21
- **Gibbs' inequality** *Source:* https://en.wikipedia.org/wiki/Gibbs'{}_inequality?oldid=790709123 *Contributors:* Michael Hardy, Giftlite, Burn, Jheald, Nigosh, Srleffler, Reetep, SmackBot, Maksim-e~enwiki, Mcld, Bluebot, Chungc, Thermochap, D.H, ReviewDude, Mild Bill Hiccup, Addbot, MARijlaarsdam, SpBot, Lightbot, RVS, Erik9bot, ⍰⍰, JimbobHolton, Vieque, Deacon Vorbis and Anonymous: 18

2.7.2 Images

- **File:AVLtreef.svg** *Source:* https://upload.wikimedia.org/wikipedia/commons/0/06/AVLtreef.svg *License:* Public domain *Contributors:* Own work *Original artist:* User:Mikm
- **File:Josiah_Willard_Gibbs_-from_MMS-.jpg** *Source:* https://upload.wikimedia.org/wikipedia/commons/c/c7/Josiah_Willard_Gibbs_-from_MMS-.jpg *License:* Public domain *Contributors:* Frontispiece of *The Scientific Papers of J. Willard Gibbs*, in two volumes, eds. H. A. Bumstead and R. G. Van Name, (London and New York: Longmans, Green, and Co., 1906) *Original artist:* Unknown. Uploaded by Serge Lachinov (обработка для wiki)
- **File:Kraft_inequality_example.png** *Source:* https://upload.wikimedia.org/wikipedia/commons/0/04/Kraft_inequality_example.png *License:* Public domain *Contributors:* Own work *Original artist:* Mjoachimiak

2.7.3 Content license

- Creative Commons Attribution-Share Alike 3.0