Book Review_____

Information Theory, Inference, and Learning Algorithms—David J. C. MacKay (Cambridge, U.K.: Cambridge Univ. Press, 2003, 628 + xxii pp.) *Reviewed by Sergio Verdú.*

Best known in our circles for his key role in the renaissance of low-density parity-check (LDPC) codes, David MacKay has written an ambitious and original textbook. Almost every area within the purview of these TRANSACTIONS can be found in this book: data compression algorithms, error-correcting codes, Shannon theory, statistical inference, constrained codes, classification, and neural networks. The required mathematical level is rather minimal beyond a modicum of familiarity with probability. The author favors exposition by example, there are few formal proofs, and chapters come in mostly self-contained morsels richly illustrated with all sorts of carefully executed graphics. With its breadth, accessibility, and handsome design, this book should prove to be quite popular.

Highly recommended as a primer for students with no background in coding theory, the set of chapters on error-correcting codes are an excellent brief introduction to the elements of modern sparse-graph codes: LDPC, turbo, repeat–accumulate, and fountain codes are described clearly and succinctly.

As a result of the author's research on the field, the nine chapters on neural networks receive the deepest and most cohesive treatment in the book. Under the umbrella title of Probability and Inference we find a medley of chapters encompassing topics as varied as the Viterbi algorithm and the forward–backward algorithm, Monte Carlo simulation, independent component analysis, clustering, Ising models, the saddle-point approximation, and a sampling of decision theory topics.

The chapters on data compression offer a good coverage of Huffman and arithmetic codes, and we are rewarded with material not usually encountered in information theory textbooks such as hash codes and efficient representation of integers.

The expositions of the memoryless source coding theorem and of the achievability part of the memoryless channel coding theorem stick closely to the standard treatment in [1], with a certain tendency to oversimplify. For example, the source coding theorem is verbalized as:

> "N i.i.d. random variables each with entropy $H(X)$ can be compressed into more than $NH(X)$ bits with negligible risk of information loss, as $N \rightarrow \infty$; conversely if they are compressed into fewer than $NH(X)$ bits it is virtually certain that information will be lost."

Although no treatment of rate-distortion theory is offered, the author gives a brief sketch of the achievability of rate $C/(1 - h(p_b))$ with bit-error rate p_b, and the details of the converse proof of that limit are left as an exercise. Neither Fano's inequality nor an operational definition of capacity put in an appearance.

Perhaps his quest for originality is what accounts for MacKay's proclivity to fail to call a spade a spade. Almost-lossless data compression is called "lossy compression;" a vanilla-flavored binary hypoth-

esis testing problem with known hypotheses becomes a "pattern recognition" problem; frequentist (non-Bayesian) statistics is called "sampling theory;" dependent random variables are "correlated;" entropy of a "random variable," of an "outcome" and of an "ensemble" are used interchangeably.

A key feature of this text is its collection of over 400 problems, some of them including solutions. The author's tremendous effort in this very important department is commendable. An icon indicates those exercises that are particularly recommended, and a difficulty rating between 1 and 5 is given for each problem: Showing that entropy is not more than log of the cardinality is a 3; proving the Slepian–Wolf theorem (no reference to the literature given) is also a 3. In fact, a weakness of the text is that pointers to the literature are very uneven. The list of references concentrates on those topics in which the author has published research, namely, neural networks, inference, and error-correcting codes. Other than textbooks and Shannon's 1948 paper [2], I could only find four references to the information theory literature.[1]

MacKay chooses an informal lecture-like writing style: "you, gentle reader," "let me tell you why," "last term collapses in a puff of smoke," "I am frequently asked," "exact marginalization […] is a macho activity," which many students will find engaging. However, I must say that some attempts at humor, such as entitling a subsection "KABOOM!," a quip about the intelligence of the 43rd President of the United States, including O. J. Simpson in the index, or the companion website [4] comparing this book to *Harry Potter and the Philosopher's Stone*, come across as rather sophomoric.

Several asides sprinkled throughout the book provide fun reading: why the base-10 unit of information is called the *ban* (in England); how the brain deconvolves optical signals to improve the resolution of the retina; how race influences the imposition of the death penalty in the United States; how the ISBN (International Standard Book Number) code works; or how to crack the Enigma cipher used by the Nazis. Yet another topic not usually covered in information theory books is the subject of a chapter entitled "Why have sex?" I suspect that connoisseurs of British slang may find that the punch line of that chapter contains more than one meaning.

With over 250 occurrences of *Bayes[ian]* and frequent jeremiads against the purported evils and dangers of non-Bayesian statistics, MacKay's faith in the credo of Bayesian statistics will be obvious to even the most casual reader. Those unfamiliar with the acrimony of the religious battles between "orthodox" and Bayesian statisticians may be surprised at the tone of statements such as "who cares about moments?—only sampling theory statisticians who are barking up the wrong tree," or "the Bayesian answer is always better than the sampling theory answer; and often much, much better." Framed in a box is the maxim:

> "Always write down the probability of everything."

[1]Aside from [2], the only referenced Shannon work is the unpublished Bell Labs report [3]. In this forgotten gem from June 1944, Shannon derives the maximum *a posteriori* probability (MAP) solution for two nonequiprobable signals received in nonwhite Gaussian noise. This is the first derivation of the general matched filter (difference signal applied to the inverse of the noise covariance operator). Contemporaneously, North in June 1943 and Middleton–VanVleck in May 1944 came up with the matched filter as the max-SNR solution for the far simpler white noise case. Traditionally, the dawn of the application of hypothesis testing to communication theory is ascribed to the work of Kotelnikov, Woodward, and others in the mid–1950s. It is time to revise history and credit Claude Shannon with the inception of this keystone of communication theory!

preceded, for the delight of mischievous non-Bayesians, by "There is a general rule which helps immensely in confusing probability problems:" Abiding by the golden rule, MacKay admits (without any apparent sign of irony) that the use of a uniform prior on the bias of one-euro coins made in Belgium "seems reasonable to me."

There is something to be said for someone who is passionate about a scientific approach and is willing to stick his neck out to spread the gospel. However, one-way thinking has its dangers. Information theory is a prime example of a discipline that has benefitted enormously from Bayesian thinking, and yet it has also produced resounding, distinctly non-Bayesian, successes such as the minimum-description-length (MDL) principle and the individual-sequence approach to universal optimality. These are topics that would seem to fit nicely in a text that covers both information theory and statistics. Yet, MacKay dismisses universality as only practical when the class of sources is severely restricted, and concludes his one-page treatment of MDL stating that it is a concept that is useful for motivating priors.

David MacKay and Cambridge University Press should be congratulated for the attractive look of this volume and for its unusually affordable pricing[2] ($35 at a major online bookseller). Idiosyncratic, occasionally sloppy, but never dull, this thoroughly original book is just a lot of fun.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, Jul.–Oct. 1948.
[3] ——, "The best detection of pulses; Bell Laboratories memorandum June 22, 1944," in *Claude Elwood Shannon: Collected Papers*, N. J. A. Sloane and A. D. Wyner, Eds. Piscataway, NJ: IEEE Press, 1993, pp. 148–150.
[4] D. J. C. Mackay. Comparison of Information Theory, Inference, and Learning Algorithms with Harry Potter. [Online]. Available: http://www.inference.phy.cam.ac.uk/mackay/itila/Potter.html
[5] ——, (2003) Information Theory, Inference, and Learning Algorithms. [Online]. Available: http://www.inference.phy.cam.ac.uk/mackay/itila/book.html

[2] For examination purposes the text, along with software and demos, is available online [5].