Information Theory, Probability and Statistical Learning: A Festschrift in Honor of Andrew Barron

Editorial

Jason Klusowski[0000−0001−6484−8682]
Ioannis Kontoyiannis[0000−0001−7242−6375] and
Cynthia Rush[0000−0001−6857−2855]

Over the past 45 years, Andrew Barron has been a world-leading researcher in information theory, statistics, and statistical learning, and he is yet to show any signs of slowing down. His many important contributions, which cover an unusually broad range of questions in science and engineering, can roughly be categorized in three major groups: (*i*) information-theoretic methods in probability, (*ii*) core statistical theory and methodology, and (*iii*) statistical learning and neural networks.

Below we outline some of Barron's major contributions in each of these directions. The papers in this volume are similarly grouped into three corresponding parts.

1 Information theory in probability

The Shannon-McMillan-Breiman theorem. Barron's first major published research work [7] settled a fundamental open problem in information theory, which has significant implications for several of other fields as well. To set the stage, we recall that perhaps the deepest and most critical insight in Shannon's pioneering 1948 paper [33] was the so-called *asymptotic equipartition property* (AEP). The AEP provides a different way of viewing long realizations of symbols produced by a discrete random source with entropy rate H. It says that, for large n, all possible realizations of length n can be viewed as being either *typical* or *non-typical*. There are about 2^{nH} typical ones, all with approximately the same probability, 2^{-nH}, while the set of all non-typical realizations has vanishingly small probability.

Jason Klusowski
Princeton University, Princeton, NJ, USA, e-mail: jason.klusowski@princeton.edu

Ioannis Kontoyiannis
University of Cambridge, Cambridge, UK, e-mail: yiannis@maths.cam.ac.uk

Cynthia Rush
Columbia University, New York, NY, USA, e-mail: cynthia.rush@columbia.edu

The strongest version of the mathematical expression of the AEP is known as the Shannon-McMillan-Breiman theorem (SMBT), and it applies to all finite-valued, stationary and ergodic sources. It would be hard to overstate its importance to the development of information-theoretic results, particularly coding theorems, over the years. The SMBT also played a fundamental role in the evolution of isomorphism theory of dynamical systems, and it became the main link in the very fruitful connection between information theory and ergodic theory. Moreover, it was observed that the same mathematical computation – the identification of the exponential rate of decay of the Radon–Nikodym derivatives between the finite-dimensional marginals of two different stationary processes – corresponds to the optimal error exponent in a hypothesis test between the two underlying processes. After a handful of generalizations of Shannon's original result, the definitive version of the SMBT for processes with densities was established by Barron in 1985 [7].

The central limit theorem. Barron's second major research work, which appeared only a year later later, was on the information-theoretic approach to the central limit theorem (CLT). The idea of looking at the CLT through an information-theoretic lens has its roots in the early work of Linnik in the late 1950s [27, 28]. This was followed by a number of papers deriving increasing stronger results, and the first major breakthrough – the first general CLT in the sense of relative entropy, for sums of continuous random variables – was Barron's 1986 proof [8] of the following elegant statement: The relative entropy between the law of the standardized sum of independent and identically distributed (i.i.d.) random variables and the standard Gaussian converges to zero if and only if it is ever finite. Barron's derivation was first described in a technical report two years earlier [5], which has become a useful reference in its own right, as it contains the first rigorous development of an important collection of results (Stam's Fisher information inequality, de Bruijn's identity, and others) that have come to play a major role in information theory and its connections with other parts of analysis and probability in the past 20 years or so.

In 2004, Barron returned to the CLT and, in paper jointly written with Oliver Johnson [22], they established conditions under which explicit convergence rates could be derived for the entropic CLT. The same year, Artstein, Ball, Barthe and Naor [1] proved that the convergence in the entropic CLT is monotonic. Equivalently, they showed that the entropy of the standardized partial sums of i.i.d. continuous random variables *increases* to its maximum possible value under the obvious variance constraint, namely, that of the Gaussian. In 2007, Madiman and Barron [29] gave a new proof of this monotonicity. In fact, the provided a broad, unifying treatment of a number of important information-theoretic inequalities, including new generalizations of Stam's Fisher information inequality and Shannon's entropy power inequality.

Martingales and σ-algebras. In addition to usual information-theoretic functionals of entropy and mutual information, there are numerous other, different quantitative notions of "information" throughout science. In particular, in probability theory, the process by which more and more information is revealed via a sequence of

observations, is modeled via *filtrations*, namely, either increasing or decreasing sequences $\{\mathcal{F}_n\}$ of σ-algebras.

In 1991 [10], Barron used the chain rule for relative entropy to give a new proof of the following basic fact, connecting information-theoretic thinking with the probabilistic concept of filtrations: If $\{\mathcal{F}_n\}$ is an increasing family of σ-algebras, and μ_n, ν_n are the restrictions of μ, ν, respectively, to $\mathcal{F}_n$, $n \geq 1$, where μ, ν are probability measures on $\sigma(\cup_n \mathcal{F}_n)$, then $D(\mu_n \| \nu_n) \to D(\mu \| \nu)$ as $n \to \infty$, where D is the usual relative entropy. Moreover, he utilized this result to give a new, information-theoretic proof of both the almost sure and L^1 versions of the martingale convergence theorem for nonnegative martingales.

Some years later, in 2000 [13], Barron showed that, if $D(\mu_n \| \nu_n)$ is eventually finite, then the same "limit of information" result $D(\mu_n \| \nu_n) \to D(\mu \| \nu)$ holds in the case when the underlying family of σ-algebras $\{\mathcal{F}_n\}$ is decreasing. Then he used this to give a new characterization of *reverse information projections*. If E is a convex set of probability measures on a measurable space $(A, \mathcal{A})$, and $\mu \notin E$ is a probability measure on the same space then, roughly speaking, the reverse information projection ν^* of μ onto E is the probability measure that achieves $D(\mu \| \nu^*) = \inf_{\nu \in E} D(\mu \| \nu)$. Reverse information projections play a central role in the analysis of maximum likelihood estimation and in the construction of so-called e-variables [30] in statistics. Interestingly, he noted that similar arguments could be used to examine the convergence of a Markov chain to equilibrium, which brings us to our last topic.

Markov chains. One of the first probabilistic limit theorems after the CLT to be considered from an information-theoretic point of view, was the convergence of Markov chains to equilibrium. Early work in this area was done by Rényi [31], Kendall [25], and Fritz [21]. In 1997 [11], in part building on ideas in these earlier works, Barron outlined ways in which one could obtain information-theoretic bounds on the rate of convergence of the random-walk Metropolis sampler and the Metropolis-adjusted Langevin algorithm for Markov chain Monte Carlo (MCMC) simulation. First, he had a proof outline for establishing bounds on the convergence to zero of the relative Fisher information between the time-n distribution p_n of the chain and the target density p. Then he employed de Bruijn-style arguments similar to those in his earlier work [8, 5], to deduce corresponding results for the relative entropy $D(p_n \| p)$.

In 2000 [13], Barron returned to Markov chains and, using techniques similar to the "limits of information" theorems described above, he gave a new information-theoretic proof of the following elegant convergence theorem for reversible Markov chains: If a reversible Markov chain has unique unique invariant measure π then the relative entropy $D(\mu_n \| \pi)$ between its law μ_n and π converges to zero, if and only if it is eventually finite.

2 Statistical theory and methodology

Foundational results on the minimum description length principle. Starting with his 1985 Ph.D. thesis [6] and continuing throughout his career, Barron has made a number of profound and foundational contributions to the theory and applications of the Minimum Description Length (MDL) principle – a principle of statistical inference and information theory that formalizes Occam's Razor: Among all explanations of the observations at hand, select the *simplest* one. The MDL principle was introduced by Rissanen in 1978 [32] and much of the core MDL theory was developed in the 1980s and 90s by, among others, Jorma Rissanen, Andrew Barron, and Bin Yu, leading to their 1998 review paper [17].

In words, MDL is a collection of ideas and methods for model selection and statistical inference, based on the principle that the best model for any given dataset is the one that compresses it the most — that is, the one that leads to the shortest total code-length for both the model and the data. Formally, among all models M in a model class $\mathcal{M}$,

$$\text{best model} = \arg\min_{M \in \mathcal{M}} \Big\{ \underbrace{L(M)}_{\text{model complexity}} + \underbrace{L(x^n|M)}_{\text{data fit}} \Big\},$$

where x^n is the observed data, $L(M)$ is the number of bits needed to describe the model M, and $L(x^n|M)$ is the number of bits needed to describe the data based on M. Guided by information-theoretic principles and using tools from probability and statistics, Barron's work in the area – of which some important contributions are summarized below – clarified and formalized connections between MDL, Bayesian statistics, and universal data compression.

For example, the 1991 paper [16], co-written by Barron and his (by then, former) Ph.D. advisor, Tom Cover, contains some of Barron's seminal MDL results, including derivations of important convergence properties of MDL estimators based on two-part codes. Also, it is where Barron first introduces the key notion of the *index of resolvability*, which quantifies the best achievable trade-off between approximation error and model complexity when fitting data using a class of models.

In the same spirit, in the late 1990s, Xie and Barron [36, 35] obtained optimal asymptotic results for the minimax regret in universal prediction, data compression, and in sequential betting strategies. And around the same time, Yang and Barron [37], in what is by now widely regarded as a landmark paper for its generality, depth, and its unified approach, developed a general information-theoretic method for deriving minimax convergence rates in nonparametric estimation problems, showing how covering numbers, metric entropy, and relative entropy can be used to determine the optimal rates of convergence under the log-loss criterion.

Bayesian asymptotics. Bayesian statistics is a way of doing statistical inference that treats unknown quantities (like parameters or functions) as random variables with their own probability distributions, often referred to as prior beliefs. Bayesian

estimation and inference are based on the Bayesian posterior distribution, which describes the statistician's updated, 'posterior' beliefs after observing the data.

In two papers in the early 1990s [20, 19] with his first Ph.D. student, Bertrand Clarke, Barron helped to establish a Bayesian justification of MDL. These works provided an information-theoretic analysis of Bayesian statistical estimation, and they derived convergence rates for Bayesian posterior distributions. Specifically, [19] develops a precise asymptotic analysis of Bayesian inference using information-theoretic tools, and [20] shows that, in regular parametric models, Jeffreys' prior minimizes the maximum asymptotic entropy risk, making it least favorable in the information-theoretic sense.

Then, in 1999, Barron, Schervish, and Wasserman [18] provided general conditions under which Bayesian posterior distributions are consistent in nonparametric settings, meaning that the posterior asymptotically concentrates on the true data-generating distribution, as the number of observations increases. Foundational in Bayesian nonparametrics, this work offers a clean, general, and elegant framework for establishing posterior consistency without requiring restrictive assumptions. Moreover, it shows how to construct Bayesian priors that satisfy the assumptions needed for consistency. Tools and ideas build on Barron's early work in the area [9, 12].

In the same year, Barron, Birgé and Massart produced a very influential paper [14], in which they developed novel and powerful minimax bounds for model selection, for MLD-inspired, penalized-maximum-likelihood model selection criteria. In a very broad setting that includes nonparametric regression and density estimation, they introduced an "accuracy index" that quantifies the fundamental trade-off between the approximation error and the parameter dimension relative to sample size. In particular, this work provided minimax rate optimal – that is, *adaptive* – estimators in a variety of contexts.

Capacity-achieving sparse superposition codes. One of the key challenges in modern communication systems in general and in wireless communications in particular, is to devise coding schemes for transmitting information reliably from a sender to a receiver through a noisy channel [33]. Such coding schemes need to be computationally efficient, have low probability of decoding error, and allow for data rates close to the information-theoretically optimal limit, Shannon's channel capacity. A practical model of real world communication and one of the most widely studied and used, is the additive white Gaussian noise (AWGN) channel.

Based on ideas from high-dimensional, sparse linear regression, in 2012 [23], Barron and his Ph.D. student Anthony Joseph introduced and analyzed a class of practical codes for coding over the AWGN channel. These codes, referred to as sparse regression codes (SPARCs) [34], were initially shown to achieve capacity when maximum likelihood decoding is used. Then, in subsequent work by Barron, Cho and Joseph [24, 15], computationally efficient decoding schemes were developed for SPARCs. Therefore, SPARCs are the first provably efficient and capacity-achieving family of codes for the AWGN channel, and they are used for many modern practical communications tasks.

3 Statistical learning and neural networks

Barron's contributions to statistical learning and neural networks reflect a cohesive and far-reaching research program that integrates ideas from information theory, approximation theory, and statistical estimation. His work has helped establish a rigorous mathematical foundation for methods that are now central to modern machine learning, particularly in understanding how adaptive estimators and neural networks can achieve favorable performance in high-dimensional and nonparametric settings.

Information-theoretic foundations for statistical learning. In early work with his student Chyong-Hwa Sheu [4], Barron developed a framework for approximating the log-density of a distribution using sequences of regular exponential families constructed from polynomial, spline, and trigonometric basis functions. The central theoretical result demonstrated that maximum likelihood estimators within these families achieve the minimax-optimal rate of $O(n^{-2r/(2r+1)})$ in relative entropy, under the assumption that the true log-density has r square-integrable derivatives.

A distinctive feature of this analysis was its decomposition of the relative entropy into approximation and estimation error terms –essentially a bias-variance trade-off in the language of information theory. Increasing the number of basis functions reduces approximation error but increases estimation error, and balancing the two yields the minimax rate. More broadly, the work revealed that exponential families, interpreted through information projections (as relative entropy minimizers), offer a unified framework for nonparametric density estimation.

Over a decade later, Barron explored information-theoretic methods for combining statistical models. In joint work with his student Gilbert Leung [26], he investigated mixtures of least-squares projections and introduced unbiased risk estimators to analyze their performance. The paper established sharp oracle inequalities of the form:

$$\text{Risk(mixture)} \leq \min_{M \in \mathcal{M}} \left\{ \text{Risk(model } M) + O\left(\frac{\log |\mathcal{M}|}{n} \right) \right\}, \tag{1}$$

where $\mathcal{M}$ denotes the model class considered.

This work extended Barron's broader agenda of using information-theoretic principles to structure and analyze adaptive statistical methods. Like many of his earlier results, the model averaging framework shared the same underlying emphasis on balancing approximation and estimation. The estimated risk of the mixture decomposes into three interpretable components: a weighted average of the risks of individual models, a variance reduction term capturing the stabilizing effect of combining diverse estimators, and a complexity penalty reflecting the cost of adaptively assigning weights. This work helped catalyze a large body of subsequent work on statistical aggregation and exponential weighting.

Universal approximation and complexity bounds for neural networks. Among Barron's most foundational contributions are his theoretical results on the approximation capabilities and statistical complexity of neural networks. His seminal 1993

paper [2] demonstrated that single-hidden-layer neural networks with sigmoidal activation functions can achieve an integrated squared error of order $O(1/n)$, where n is the number of hidden units. This rate holds for a class of functions satisfying a novel smoothness condition—marking one of the first rigorous demonstrations that neural networks could mitigate the curse of dimensionality under appropriate assumptions.

The key innovation was Barron's introduction of a smoothness class defined via the decay of the Fourier transform, rather than traditional Sobolev norms, together with a clever application of the probabilistic method. Specifically, functions f satisfying

$$\int |\omega| \, |\hat{f}(\omega)| \, d\omega < C_f,$$

can be approximated with dimension-independent accuracy by neural networks. This condition ensures a form of "spectral smoothness," controlling high-frequency content in a way that permits effective approximation. The class of such functions—now known as the *Barron class*—has become a cornerstone in neural network approximation theory and continues to shape modern research.

Barron showed that, while classical n-term basis expansions typically yield approximation rates of order $O(n^{-2/d})$ for input dimension d, neural networks can attain rates of order $O(C_f^2/n)$, independent of d. This striking contrast helps explain the empirical success of neural networks in high-dimensional settings and provided an early theoretical framework for understanding it.

In a 1994 follow-up [3], Barron extended these approximation results to the statistical learning setting. He analyzed both the approximation error due to finite network size and the estimation error arising from finite data. He showed that the integrated mean squared error for estimating a function f using a neural network trained on N samples in d-dimensional input space could be bounded as

$$O\left(\frac{C_f^2}{n}\right) + O\left(\frac{nd}{N} \log N\right),$$

reflecting goodness-of-fit and model complexity relative to sample size. Balancing these terms by choosing

$$n \sim C_f \left(\frac{N}{d \log N}\right)^{1/2},$$

yields an overall risk of order

$$O\left(C_f \left(\frac{d}{N} \log N\right)^{1/2}\right).$$

This result provided one of the earliest and clearest demonstrations that neural networks can achieve statistically efficient learning in moderately high-dimensional regimes.

References

1. S. Artstein, K. Ball, F. Barthe, and A. Naor. Solution of Shannon's problem on the monotonicity of entropy. *J. Amer. Math. Soc.*, 17(4):975–982, April 2004.
2. A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945, 1993.
3. A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.
4. A. R. Barron and C.-H. Sheu. Approximation of density functions by sequences of exponential families. *Ann. Statist.*, 19(3):1347–1369, 1991.
5. A.R. Barron. Monotonic central limit theorem for densities. NSF Technical Report no. 50, Department of Statistics, Stanford University, March 1984. Available at `statistics.stanford.edu/resources/technical-reports`.
6. A.R. Barron. *Logically smooth density estimation*. PhD thesis, Department of Electrical Engineering, Stanford University, Stanford, CA, September 1985.
7. A.R. Barron. The strong ergodic theorem for densities: Generalized Shannon-Mcmillan-Breiman theorem. *Ann. Probab.*, 13(4):1292–1303, November 1985.
8. A.R. Barron. Entropy and the central limit theorem. *Ann. Probab.*, 14(1):336–342, January 1986.
9. A.R. Barron. The exponential convergence of posterior probabilities with implications for bayes estimators of density functions. Technical report, Department of Statistics, University of Illinois Champaign, IL, 1988.
10. A.R. Barron. Information theory and martingales. In *1991 IEEE International Symposium on Information Theory (ISIT)*, Budapest, Hungary, June 1991.
11. A.R. Barron. Information theory in probability, statistics, learning, and neural nets. In Y. Freundand and R.E. Schapire, editors, *Proceedings of the Tenth Annual Conference on Computational Learning Theory (COLT)*, Nashville, Tennessee, July 1997. Available at `www.stat.yale.edu/~arb4/publications_files/COLT97.pdf`.
12. A.R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. *Bayesian Statistics*, 6, 1998.
13. A.R. Barron. Limits of information, Markov chains, and projection. In *2000 IEEE International Symposium on Information Theory (ISIT)*, Sorrento, Italy, June 2000.
14. A.R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, February 1999.
15. A.R. Barron and S. Cho. High-rate sparse superposition codes with iteratively optimal estimates. In *2012 IEEE International Symposium on Information Theory (ISIT)*, pages 120–124, Cambridge, MA, July 2012. IEEE.
16. A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054, 1991.
17. A.R. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6):2743–2760, October 1998. Information theory: 1948–1998.
18. A.R. Barron, M.J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2):536–561, 1999.
19. B.S. Clarke and A.R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory*, 36(3):453–471, May 1990.
20. B.S. Clarke and A.R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Statist. Plann. Inference*, 41(1):37–60, 1994.
21. J. Fritz. An information-theoretical proof of limit theorems for reversible Markov processes. In *Trans. Sixth Prague Conf. Information Theory, Statist. Decision Functions, Random Processes (Tech. Univ., Prague, 1971; dedicated to the memory of Antonín Špaček)*, pages 183–197. Academia, Prague, 1973.
22. O. Johnson and A.R. Barron. Fisher information inequalities and the central limit theorem. *Probab. Theory Related Fields*, 129(3):391–409, July 2004.

23. A. Joseph and A.R. Barron. Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity. *IEEE Trans. Inform. Theory*, 58(5):2541–2557, May 2012.
24. A. Joseph and A.R. Barron. Fast sparse superposition codes have near exponential error probability for $R < C$. *IEEE Trans. Inform. Theory*, 60(2):919–942, 2013.
25. D.G. Kendall. Information theory and the limit-theorem for Markov chains and processes with a countable infinity of states. *Ann. Inst. Statist. Math.*, 15(1):137–143, May 1963.
26. G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.
27. Ju.V. Linnik. An information-theoretic proof of the central limit theorem with Lindeberg conditions. *Theory Probab. Appl.*, 4:288–299, 1959.
28. Ju.V. Linnik. On certain connections of the information theory of C. Shannon and R. Fisher with the theory of symmetrization of random vectors. In *Trans. Second Prague Conf. Information Theory, Statist. Decision Functions, Random Processes*, pages 313–327, New York, NY, 1960. Academic Press.
29. M. Madiman and A.R. Barron. Generalized entropy power inequalities and monotonicity properties of information. *IEEE Trans. Inform. Theory*, 53(7):2317–2329, July 2007.
30. A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statist. Sci.*, 38(4):576–601, 2023.
31. A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 547–561. Univ. California Press, Berkeley, CA, 1961.
32. J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.
33. C.E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27(3):379–423, 623–656, 1948.
34. R. Venkataramanan, S. Tatikonda, and A.R. Barron. Sparse regression codes. *Foundations and Trends in Communications and Information Theory*, 15(1-2):1–195, June 2019.
35. Q. Xie and A.R. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Trans. Inform. Theory*, 43(2):646–657, March 1997.
36. Q. Xie and A.R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Inform. Theory*, 46(2):431–445, March 2000.
37. Y. Yang and A.R. Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.