DOOB'S PROOF OF MLE CONSISTENCY

OMAR HIJAB

Abstract. A short proof of the consistency of Fisher's maximum likelihood estimator, following Doob's proof, is presented.

Introduction

The consistency of Fisher's maximum likelihood estimator (MLE) [10] is a basic result discovered a century ago. This result is typically subsumed under the rubric of M-estimators, and derived via standard techniques dating back to Cramér [6] and Wald [23]. For a more recent treatment, van der Vaart [22] is standard.

In teaching this material, I found Doob's consistency proof [9] for the MLE more motivated and accessible to students, especially to students already exposed to information and entropy as part of a data science or machine learning course. The fact that the proof covers general parameter θ spaces and general state X spaces with no extra effort only adds to its appeal.

The aim of this note is to present this proof, together with some historical terminology remarks, in a self-contained manner. There is no novelty here except, perhaps, the conciseness of the presentation. By using the law of large numbers in its strong form, we derive MLE consistency in its strong form.

We start with the setup and assumptions, then state and prove consistency. After that, we attempt to give a coherent historical account of entropy terminology, which is an interesting story in itself, then we end by showing the assumptions hold for the multivariate normal case.

Surprisal and Information

Suppose a random variable X is sampled repeatedly, resulting in independent and identically distributed copies $X_1, X_2, \ldots$ of X. Suppose the density $p(X, \theta^*)$ of X is one of a parametric family of densities $p(X, \theta)$. Then θ^* is the *true parameter*.

More explicitly, let E denote the background expectation. Then the expectation corresponding to the parameter θ is

$$E_\theta(f(X)) = E(p(X, \theta) f(X)), \tag{1}$$

and the true expectation is

$$E_*(f(X)) = E(p(X, \theta^*) f(X)). \tag{2}$$

With this setup[1], the goal is to estimate or recover θ^* from $X_1, X_2, \ldots$.

Below, the phrase "with probability one" always refers to the true probability P_* corresponding to the true expectation E_*. The background expectation E plays no direct role, as it enters only through E_*.

[1]For more on this setup, see the multivariate normal case below.

Crucial to the analysis is the *surprisal* [17, 20]

$$S(X,\theta) = \log(1/p(X,\theta)). \tag{3}$$

This terminology is intuitive: Low probability equals high surprisal. Using the surprisal, the *relative information* is

$$I(\theta^*,\theta) = E_*(S(X,\theta) - S(X,\theta^*)) = E_*(\log(p(X,\theta^*)/p(X,\theta))). \tag{4}$$

This quantity appears in many fields, in different forms, and is often called "relative entropy" or "Kullback-Leibler divergence". Because of the widely differing terminology, below we give a coherent account of the terminology history of I.

Whatever it is called, the fundamental property of $I(\theta^*,\theta)$ is its ability to isolate the true density: $I(\theta^*,\theta) \geq 0$, and $I(\theta^*,\theta) = 0$ iff $p(X,\theta) \equiv p(X,\theta^*)$. In general, $I(\theta^*,\theta)$ is defined differently, and may be infinite, but is finite iff $S(X,\theta) - S(X,\theta^*)$ is P_* integrable, in which case (4) holds [7, 8].

The simplest example is a categorical random variable X distributed according to $\theta^* = (\theta^*_1, \theta^*_2, \ldots, \theta^*_d)$. If the values of X are 1, 2, ..., d, the mass function of X is $p(X,\theta^*) = \theta^*_X$. If $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ is any other distribution, the surprisal is $S(X,\theta) = \log(1/\theta_X)$, and

$$I(\theta^*,\theta) = \sum_{i=1}^{d} \theta^*_i \log(\theta^*_i/\theta_i). \tag{5}$$

The *likelihood* and *mean surprisal* are

$$L_n(\theta) = \prod_{k=1}^{n} p(X_k,\theta), \qquad \bar{S}_n(\theta) = \frac{1}{n}\log(1/L_n(\theta)) = \frac{1}{n}\sum_{k=1}^{n} S(X_k,\theta).$$

The mean surprisal is also called the negative-log-likelihood.

If $S(X,\theta)$ is P_* integrable for all θ, then θ^* minimizes $E_*(S(X,\theta))$. By the law of large numbers, $\bar{S}_n(\theta)$ converges to $E_*(S(X,\theta))$ as $n \to \infty$, with probability one. Hence it is natural to estimate θ^* by the minimizer of $\bar{S}_n(\theta)$ or, equivalently, the maximizer of $L_n(\theta)$: For each $n \geq 1$, a *maximum likelihood estimator (MLE)*, if it exists, is a parameter $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$ maximizing $L_n(\theta)$.

Consistency

To achieve consistency, there are three natural assumptions, the first of which is

A. *(Identifiability):* $\theta \neq \theta'$ implies $p(X,\theta) \neq p(X,\theta')$, hence $I(\theta^*,\theta) > 0$. Without identifiability, one cannot hope to extract θ^* from $p(X,\theta^*)$.

A function $F(\theta)$ is *proper* if its sublevel sets $F(\theta) \leq c$ are compact subsets of the parameter space, for every level c. A well-known elementary fact is the existence of minimizers of continuous proper functions. Since $\hat{\theta}_n$ should minimize $\bar{S}_n(\theta)$, this leads us to the second assumption

B. *(Properness):* $S(X,\theta)$ is P_* integrable for all θ and there is a continuous proper $F(\theta)$ such that, with probability one, the inequality

$$F(\theta) \leq \bar{S}_n(\theta), \qquad \text{for all } \theta, \tag{6}$$

holds for all but finitely many n.

As a consequence of **B**, with probability one, an MLE $\hat{\theta}_n$ exists for all but finitely many n, $E_*(S(X,\theta))$ is proper, and hence $I(\theta^*,\theta)$ is proper. Because of this, before establishing **B**, it is best, as a reality check, to first establish properness of $I(\theta^*,\theta)$.

We will need the convergence of $\bar{S}_n(\theta)$ to $E_*(S(X,\theta))$ across θ, in the sense

$$\bar{S}_n(\theta) \to E_*(S(X,\theta)), \qquad \text{as } n \to \infty, \qquad \text{for all } \theta, \tag{7}$$

with probability one. For this we need the third assumption

C. *(Continuity):* The parameter space is separable and there is a continuous $f(\theta,\theta')$ and a P_* integrable $g(X)$ satisfying

$$|S(X,\theta) - S(X,\theta')| \le f(\theta,\theta')g(X) \qquad \text{and} \qquad f(\theta,\theta) = 0.$$

Then **B**, **C** and a separability argument establishes the validity of (7).

B, **C** typically hold for *exponential families* [5, 14, 18, 22], the densities $p(X,\theta)$ where $S(X,\theta)$ is a sum of products of the form $f(\theta)g(X)$.

Theorem. *Assume a population has a distribution depending on a parameter θ, and let θ^* be the population's true parameter. Assume also the relative information $I(\theta^*,\theta)$ is finite for all θ. Then, under **A**, **B**, **C**, with probability one, an MLE sequence $\hat{\theta}_n$ may be defined for all but finitely many n, and any such sequence converges to θ^* as $n \to \infty$.*

By the law of large numbers,

$$\frac{1}{n}\sum_{k=1}^{n} g(X_k) \to E_*(g(X)), \qquad \text{as } n \to \infty, \tag{8}$$

with probability one. Fix an outcome for which (6), (7), (8) hold, except, for (6), for finitely many n. We show $\hat{\theta}_n \to \theta^*$ as $n \to \infty$ for this outcome.

By (6),

$$F(\hat{\theta}_n) \le \bar{S}_n(\hat{\theta}_n) \le \bar{S}_n(\theta^*), \qquad \text{for all but finitely many } n.$$

Since the right side is bounded as $n \to \infty$ and $F(\theta)$ is proper, the sequence $\hat{\theta}_n$ lies in a compact set in the parameter space.

Now suppose $\hat{\theta}_n \to \theta$ along a subsequence. By **C**,

$$\bar{S}_n(\theta) - \bar{S}_n(\theta^*) \le \bar{S}_n(\theta) - \bar{S}_n(\hat{\theta}_n) \le f(\theta,\hat{\theta}_n) \cdot \frac{1}{n}\sum_{k=1}^{n} g(X_k). \tag{9}$$

By (8), the left side of (8) is bounded as $n \to \infty$, and $f(\theta,\hat{\theta}_n) \to 0$ along a subsequence, hence the right side of (9) goes to zero along a subsequence. By (7), the left side of (9) converges to

$$E_*(S(X,\theta) - S(X,\theta^*)) = I(\theta^*,\theta),$$

hence $I(\theta^*,\theta) \le 0$. But $I(\theta^*,\theta) \ge 0$, hence $I(\theta^*,\theta) = 0$. From **A**, this implies $\theta = \theta^*$, establishing the result.

Terminology History

The relative information $I(\theta^*,\theta)$ is often called "relative entropy" [5, 14]. The term "entropy", originating in thermodynamics, was coined by Clausius in his 1865 paper [4]. Boltzmann, in his 1872 paper [2], derived his H theorem, using a quantity H proportional to the negative of entropy. Gibbs, in his 1902 book [11], building on Boltzmann's 1877 combinatorial interpretation [3] and Planck's 1901 paper [16], flipped Boltzmann's sign convention, writing the entropy H as done today.

Shannon, in his 1948 paper [19], independently defined

$$H(p) = -\sum_{i=1}^{d} p_i \log p_i.$$

Von Neumann had earlier defined quantum entropy in his 1932 book [15], and when Shannon consulted him around 1939–1940 about terminology for H, von Neumann suggested "entropy", which Shannon adopted [21]. However, in his remark "H is then, for example, the H in Boltzmann's H theorem", Shannon conflates his H with Boltzmann's H.

The relative information $I(\theta^*, \theta)$ is jointly convex in (θ^*, θ), a fact that is apparent in the special case (5), whereas $H(p)$ above is strictly concave in p. Because of this, calling both quantities "entropy" creates a mismatch. Indeed, based on the curvature, it is natural to call the *negative* of (5) the *relative entropy*, as information and entropy are opposites. The mismatch is also in Python: `entropy(p)` returns the concave H, but `entropy(p,q)` returns the convex I.

This error is a common psychological conflation: When measuring the temperature, are we measuring how cold it is, or how hot it is? Of course we are doing both, but this doesn't mean cold equals hot. In fact, cold equals minus hot.

While many authors' terminology conflicts with ours, some authors align partially with our usage. Schervish, in his 1995 text [18], uses "Kullback-Leibler information" for I, which is more descriptive than the widely used but opaque "Kullback-Leibler divergence" [1]. Kullback himself, in his 1959 monograph [13], refers to I as an "information function". Finally, Donsker-Varadhan, in their 1975 large deviations paper [8], denote the rate function as I rather than H, derive its properties in full generality, yet give it no name.

Regarding Doob's 1934 proof, this was at the dawn of abstract probability theory, within the milieu developed by Birkhoff, Daniell, Fréchet, Hopf, Khintchine, Kolmogorov, Lévy, Radon, and Wiener, and within a year of Kolmogorov's 1933 foundational extension theorem [12].

Multivariate Normal

Let μ be a point in $\mathbf{R}^d$, Q a positive definite symmetric $d \times d$ matrix, and $\theta = (\mu, Q)$. Let $v \cdot w$ be the dot product in $\mathbf{R}^d$. With x in $\mathbf{R}^d$, the *normal density with mean μ and variance Q* is $p(x, \theta) = e^{-S(x,\theta)}$ where

$$S(x, \theta) = \frac{1}{2} \log \det(2\pi Q) + \frac{1}{2}(x - \mu) \cdot Q^{-1}(x - \mu). \tag{10}$$

Given vectors v, w in $\mathbf{R}^d$, the *symmetric tensor product* $v \otimes w$ is the symmetric matrix whose ij-th entry is $(v_i w_j + v_j w_i)/2$. A random variable X in $\mathbf{R}^d$ is *normally distributed with parameter θ* if its density is $p(x, \theta)$. For such X,

$$E(X) = \mu, \qquad E(X \otimes X) = Q + \mu \otimes \mu. \tag{11}$$

Suppose X is normally distributed with parameter $\theta^* = (\mu^*, Q^*)$, and let X_1, X_2, ... be independent and identically distributed copies of X, as in (1), (2), all defined on some outcome space with background expectation E. This setup is possible[2] if $E = E_*$ and the parametric family is $p(X, \theta)/p(X, \theta^*)$. With this parametric family, the surprisal is $S(X, \theta) - S(X, \theta^*)$. With this recalibration, the

[2]This is a special case of the Kolmogorov extension theorem.

main result's statement and proof are unchanged. Then **A** follows from (11), and **C** from (10) with $g(X) = 1 + |X|^2$.

For **B** we review background matrix results. Given symmetric matrices Q, R, S, we say $Q \geq 0$ if $v \cdot Qv \geq 0$ for all v, and $Q \geq R$ if $Q - R \geq 0$. If $Q \geq 0$ and $R \geq 0$, then $\text{trace}(QR) \geq 0$. If $Q \geq 0$ and $R \geq S$, then $\text{trace}(QR) \geq \text{trace}(QS)$. Then

$$v \otimes w = w \otimes v, \quad v \cdot Qw = \text{trace}(Q(v \otimes w)), \quad 2v \otimes w \leq \epsilon v \otimes v + \frac{1}{\epsilon} w \otimes w, \quad \epsilon > 0.$$

By the eigenvalue decomposition, every symmetric matrix is a linear combination of matrices of the form $v \otimes v$.

The determinant $\det(Q)$ is the product of the eigenvalues of Q. With Q invertible symmetric and t a scalar,

$$\det(Q + tv \otimes v) = \det(Q) \left(1 + tv \cdot Q^{-1} v\right).$$

If further $Q + tv \otimes v$ is invertible, then

$$(Q + tv \otimes v)^{-1} = Q^{-1} - \frac{t}{1 + tv \cdot Q^{-1} v} \cdot (Q^{-1} v) \otimes (Q^{-1} v).$$

These are the matrix facts used below.

Define $f(\mu) = Q^* + (\mu - \mu^*) \otimes (\mu - \mu^*)$ and

$$F(\mu, Q) = \frac{1}{2} \log \det(2\pi Q) + \frac{1}{2} \text{trace}(Q^{-1} f(\mu)). \tag{12}$$

By (10) and (11), $S(X, \theta)$ is P_* integrable and $E_*(S(X, \theta)) = F(\mu, Q)$.

Lemma. *Let $f(\mu)$ be a symmetric matrix-valued function such that trace$(f(\mu))$ is proper and $f(\mu) \geq \delta I$ for some $\delta > 0$. Then $F(\mu, Q)$ is proper.*

Proof. Since $F(\mu, Q)$ is proper iff $F(\mu, \delta Q)$ is proper, and, up to a constant, $F(\mu, \delta Q)$ corresponds to $f(\mu)/\delta$, we may assume $\delta = 1$. Now suppose $2F(\mu, Q) \leq c$. It is enough to establish $\epsilon I \leq Q \leq I/\epsilon$ for some ϵ, since this confines Q to a compact subset of the space of positive definite matrices, and implies

$$c \geq 2F(\mu, Q) \geq \log \det(2\pi\epsilon) + \epsilon \, \text{trace}(f(\mu)),$$

which confines μ to a compact subset of $\mathbf{R}^d$.

For $\lambda > 0$, set $h(\lambda) = \log \lambda + 1/\lambda$. Then $h(\lambda) \geq 1$. If $h(\lambda) \leq c$ with $c \geq 1$, then $\lambda \leq e^c$. Since $\lambda \log \lambda \geq -1/e$, $h(\lambda) \leq c$ also implies $\lambda \geq (1 - 1/e)/c \geq e^{-c}$. Summarizing, with $\epsilon = e^{-c}$. $h(\lambda) \leq c$ implies $\epsilon \leq \lambda \leq 1/\epsilon$.

Let $\lambda_1, \lambda_2, \ldots, \lambda_d$ be the eigenvalues of Q. Since $f(\mu) \geq I$,

$$\sum_{i=1}^{d} h(\lambda_i) = \log \det(Q) + \text{trace}(Q^{-1}) \leq 2F(\mu, Q) - d \log(2\pi) \leq c.$$

Since $h(\lambda) \geq 1$, $h(\lambda_i) \leq c$, hence $\epsilon \leq \lambda_i \leq 1/\epsilon$, establishing $\epsilon I \leq Q \leq I/\epsilon$. □

From the Lemma, $E_*(S(X, \theta))$ is proper, hence $I(\theta^*, \theta)$ is proper. Let $0 < \epsilon < 1$. By the law of large numbers, with probability one, we have

$$\frac{1}{n} \sum_{k=1}^{n} (X_k - \mu^*) \otimes (X_k - \mu^*) \geq (1 - \epsilon) Q^*$$

for all but finitely many n, and

$$\left(\mu^* - \frac{1}{n}\sum_{k=1}^{n} X_k\right) \otimes \left(\mu^* - \frac{1}{n}\sum_{k=1}^{n} X_k\right) \leq \epsilon^2 Q^*,$$

for all but finitely many n.

For $0 < \epsilon < 1/2$, let

$$f_\epsilon(\mu) = (1-2\epsilon)Q^* + (1-\epsilon)(\mu-\mu^*) \otimes (\mu-\mu^*).$$

Then, as above, $\text{trace}(f_\epsilon(\mu))$ is proper in μ, and $f_\epsilon(\mu) \geq \delta I$ for some $\delta > 0$. Hence

$$F_\epsilon(\mu, Q) = \frac{1}{2}\log\det(2\pi Q) + \frac{1}{2}\text{trace}(Q^{-1} f_\epsilon(\mu))$$

is proper. By expanding the square, with probability one

$$\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu) \otimes (X_k - \mu) \geq f_\epsilon(\mu), \qquad \text{for all } \mu,$$

for all but finitely many n. Since this implies $\bar{S}_n(\theta) \geq F_\epsilon(\theta)$ for all but finitely many n, with probability one, this establishes **B**.

By differentiating $\bar{S}_n(\mu + t\nu, Q)$ and $\bar{S}_n(\mu, Q + tv \otimes v)$ at $t = 0$, the MLE is

$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} X_k, \qquad \hat{Q}_n = \frac{1}{n}\sum_{k=1}^{n}(X_k - \hat{\mu}_n) \otimes (X_k - \hat{\mu}_n).$$

References

[1] Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, Cham (2006)

[2] Boltzmann, L.: Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissenschaftliche Classe. Abteilung II **66**, 275–370 (1872)

[3] Boltzmann, L.: Über die beziehung zwischen dem zweiten hauptsatze der mechanischen wärmetheorie und der wahrscheinlichkeitsrechnung respektive den sätzen über das wärmegleichgewicht. Wiener Berichte **76**, 373–435 (1877)

[4] Clausius, R.: Über verschiedene für die anwendung bequeme formen der hauptgleichungen der mechanischen wärmetheorie. Annalen der Physik und Chemie **125**, 353–400 (1865)

[5] Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley-Interscience (2006)

[6] Cramér, H.: Mathematical Methods of Statistics. Princeton University Press (1946)

[7] Dembo, A., Zeitouni, O.: Large Deviations Techniques and Applications, *Stochastic Modelling and Applied Probability*, vol. 38, 2nd edn. Springer, New York (1998)

[8] Donsker, M.D., Varadhan, S.R.S.: Asymptotic evaluation of certain markov process expectations for large time, I. Communications on Pure and Applied Mathematics **28**(1), 1–47 (1975)

[9] Doob, J.L.: Probability and statistics. Transactions of the American Mathematical Society **36**, 759–775 (1934)

[10] Fisher, R.A.: On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London, Series A **222**, 309–368 (1922). DOI 10.1098/rsta.1922.0009

[11] Gibbs, J.W.: Elementary Principles in Statistical Mechanics. Yale University Press, New Haven (1902)

[12] Kolmogorov, A.N.: Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin (1933)

[13] Kullback, S.: Information Theory and Statistics. Wiley, New York (1959)

[14] Murphy, K.P.: Probabilistic Machine Learning: An Introduction. MIT Press (2022)

[15] von Neumann, J.: Mathematische Grundlagen der Quantenmechanik. Springer, Berlin (1932)

[16] Planck, M.: Ueber das gesetz der energieverteilung im normalspectrum. Annalen der Physik **309**(3), 553–563 (1901). DOI 10.1002/andp.19013090310
[17] Samson, E.W.: Fundamental natural concepts of information theory. Report E5079, Communications Laboratory, Electronics Research Division, Air Force Cambridge Research Center, Cambridge, MA (1951)
[18] Schervish, M.J.: Theory of Statistics. Springer, New York (1995)
[19] Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**(3), 379–423 (1948)
[20] Tribus, M.: Thermostatics and Thermodynamics: An Introduction to Energy, Information and States of Matter, With Engineering Applications. D. Van Nostrand Company, Princeton, NJ (1961)
[21] Tribus, M., McIrvine, E.C.: Energy and information. Scientific American **225**(3), 179–190 (1971). DOI 10.1038/scientificamerican0971-179
[22] van der Vaart, A.W.: Asymptotic Statistics, *Cambridge Series in Statistical and Probabilistic Mathematics*, vol. 3. Cambridge University Press, Cambridge (1998). DOI 10.1017/CBO9780511802256
[23] Wald, A.: Note on the consistency of the maximum likelihood estimate. Annals of Mathematical Statistics **20**(4), 595–601 (1949). DOI 10.1214/aoms/1177729952