Entropy Level Testing from a Large Deviations Principle.

Valérie Girardin* **and Philippe Regnault**⋆

* *Laboratoire de Mathématiques N. Oresme, UMR6139, Université de Caen Basse Normandie, Campus II, BP 5186, 14032 Caen, France.*
girardin@math.unicaen.fr

⋆ *Laboratoire de Mathématiques de Reims, EA4535, Université de Reims Champagne-Ardenne, UFR Sciences Exactes et Naturelles, BP 1039, 51687 Reims, France.*
philippe.regnault@univ-reims.fr

RÉSUMÉ. *Dans cet exposé, nous présenterons un principe de grandes déviations (PGD) pour la suite des estimateurs empiriques de l'entropie de Shannon d'une probabilité sur un ensemble fini. Nous utiliserons quelques arguments de géométrie de l'information pour établir une expression explicite de la fonction de taux gouvernant le PGD. Nous montrerons que les tests de niveau d'entropie obtenus par seuillage de la fonction de taux sont consistants et que leur puissance croÃ®t exponentiellement vite avec la taille de l'échantillon.*

ABSTRACT. *In this talk, we will present a large deviations principle (LDP) for the sequence of empirical estimators of Shannon entropy of a probability distribution on a finite set. We will use some arguments of information geometry to establish a closed form expression for the rate function governing the LDP. Then, we will show that the entropy level tests built by thresholding the rate function are consistent with power growing exponentially fast with the size of the sample.*

MOTS-CLÉS : *Entropie de Shannon, estimation par plug-in, principe de grandes déviations, géométrie de l'information, tests de niveau d'entropie.*

KEYWORDS: *Shannon entropy, plug-in estimation, large deviations principle, information geometry, entropy level tests.*

1. Empirical estimation of Shannon entropy

The concept of entropy has been introduced in the field of probability in Shannon [14] by defining

$$\mathbb{S}(P) := -\sum_{i\in E} P(i)\log P(i),$$

for any P belonging to the set $\mathcal{D}$ of all probability distributions supported by the finite set E, with the convention $0\log 0 = 0$.

Kullback and Leibler [9] introduced what is now called the Kullback-Leibler divergence of a distribution Q relative to another P as

$$\mathbb{K}(Q|P) := \sum_{i\in E} Q(i)\log\frac{Q(i)}{P(i)}$$

with the conventions $0\log(0/a) = 0$, and $a\log(a/0) = +\infty$, for $0 \le a \le 1$.

Entropy appears through $\mathbb{K}(P|U) = \mathbb{S}(U) - \mathbb{S}(P)$ as a measure of the variation of information from U to P, where U is the uniform distribution on E. The entropy of a probability distribution is widely used in numerous fields involving random variables, such as large deviations theory or computer science ; see Cover and Thomas [3] for properties of entropy and an overview of possible topics.

When only observations are available, the need to estimate entropy arises. Given $(X_1,\dots,X_n)$ an i.i.d. n-sample of $P \in \mathcal{D}$, the empirical estimator of $\mathbb{S}(P)$ is

$$\begin{aligned}\widehat{S}_n &:= \frac{1}{n}\sum_{k=1}^{n}\log\widehat{P}_n(X_i),\\ &= \mathbb{S}(\widehat{P}_n),\end{aligned}$$

where $\widehat{P}_n$ is the empirical distribution associated to the sample.

Basharin [2] proves that $\widehat{S}_n$ is biased but strongly consistent and asymptotically normal. As a particular case of a complicated series scheme of observations, Zubkov [16] shows that asymptotic normality holds only if P is not uniform on E, that is if the entropy is not maximum ; see also Harris [8] and the references therein. We give a simple proof of the following result including all the asymptotic properties of $\widehat{S}_n$ in Girardin and Regnault [7].

Theorem 1. *The empirical estimator $\widehat{S}_n = \mathbb{S}(\widehat{P}_n)$ is a strongly consistent estimator of the entropy $\mathbb{S}(P)$. Moreover :*

if P is not uniform, then $\sqrt{n}[\widehat{S}_n - \mathbb{S}(P)]$ converges in distribution to a centered normal distribution with explicit variance ;

if $P = U$ is uniform, then $2n[\widehat{S}_n - \mathbb{S}(U)]$ converges to $\sum_{i=1}^{N}\beta_i Y_i$, where all the Y_i are independent and $\chi^2(1)$-distributed random variables and $\beta_i \in \mathbb{R}$ for $i \in [\![1, N]\!]$.

2. LDP for empirical estimators of Shannon entropy

From classical Sanov's large deviations principle and contraction principle (see e.g., Dembo and Zeitouni [5]), we show that the sequence $(\widehat{S}_n)_{n\in\mathbb{N}}$ of empirical estimators of $\mathbb{S}(P)$ satisfies an LDP. We give a closed form expression for the good rate function governing the LDP through arguments of information geometry ; see Amari and Nagaoka [1], Sgarro [13], Csiszàr [4], and Regnault [12] for details about information geometry and its use for LDP. The families of escort distributions defined by

$$\mathcal{E}_P^k(i) = \frac{P(i)^k}{\sum_{j\in E} P(j)^k}, \qquad i \in E,$$

where $k \in \mathbb{R}^*$, play a prominent role in the LDP ; see Girardin and Regnault [7] for both a detailed study of their entropic properties and the proof of theorem below.

Theorem 2. *The sequence of estimators $(\widehat{S}_n)_{n\in\mathbb{N}^*}$ satisfies the large deviation principle*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\widehat{S}_n \in A) = -\inf_{s\in A} I_{\mathbb{S}}(s, P),$$

for all Borel sets $A \subseteq [0, \log(N+1)]$ such that $\mathring{A} \neq \emptyset$, with good rate function $I_{\mathbb{S}}$ defined by

$$I_{\mathbb{S}}(s,P) = \begin{cases} -s - \log p & \text{if } 0 \leq s \leq \log m, \\ \mathbb{K}(\mathcal{E}_P^k|P) & \text{if } \log(m) < s \leq \log(|E|+1), \text{ with } k > 0 \text{ such that } \mathbb{S}(\mathcal{E}_P^k) = s, \\ +\infty & \text{otherwise,} \end{cases}$$

where m is the number of modes of P, with weight p and $|E|$ is the cardinal of the support of P.

3. Entropy level testing

In goodness-of-fit testing, statistics based on the difference between the entropy of distributions are usual for discriminating between distributions. Now applied to all classical distributions, they have been introduced by Vasicek [15] for testing normality, in relation with the maximum entropy principle ; see Girardin and Lequesne [6] for more details and Girardin and Regnault [7] for fields of application.

For testing

$$H_0 : \text{"}\mathbb{S}(P) = s_0\text{"} \qquad \text{against} \qquad H_1 : \text{"}\mathbb{S}(P) = s_1\text{"},$$

we choose the statistic

$$\mathbb{K}(\mathcal{S}_{\widehat{S}_n}|\mathcal{S}_{s_0}) := \inf_{P:\mathbb{S}(P)=s_0} I_{\mathbb{S}}(\widehat{S}_n, P)$$

which appears as the Kullback-Leibler divergence between entropic spheres $\mathcal{S}_{\widehat{S}_n} := \mathbb{S}^{-1}(\{\widehat{S}_n\})$ and $\mathcal{S}_{s_0} := \mathbb{S}^{-1}(\{s_0\})$ in information geometry. We reject the null hypothesis when $\mathbb{K}(\mathcal{S}_{\widehat{S}_n}|\mathcal{S}_{s_0})$ is greater than a threshold depending on the number n of observations. The error of the first kind is shown to decrease with $1/n$. The error of the second kind is shown to decrease exponentially fast with the divergence $\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1})$; see Girardin and Regnault [7] for the proof of the following result.

Theorem 3. *Let the entropy level test with rejection region given by*

$$R_n = \left\{ \mathbb{K}(\mathcal{S}_{\widehat{S}_n}|\mathcal{S}_{s_0}) \geq \delta_n \right\},$$

with $\delta_n = (|E| + 2)\log[(n+1)/n]$.

The errors of the first kind $\alpha_n := \sup_{P \in \mathcal{S}_{s_0}} P^{\otimes n}(R_n)$ *and of the second kind* $\beta_n := \sup_{P \in \mathcal{S}_{s_1}} P^{\otimes n}(E^n \backslash R_n)$ *of this test satisfy*

$$\alpha_n \leq 1/(n+1) \quad \textit{and} \quad \limsup_{n\to\infty} \frac{1}{n} \log \beta_n \leq -\mathbb{K}(\mathcal{S}_{s_0}|\mathcal{S}_{s_1}).$$

4. Bibliographie

Amari, S. and Nagaoka, H. (2000) *Methods of information geometry*. Oxford University Press.

Basharin, G. P. (1959) On a Statistical Estimation for the Entropy of a Sequence of Independent Random Variables. *Theory of Probability and its Applications*, **4**, 333–336.

Cover, L. and Thomas, J. (1991) *Elements of Information Theory.* Wiley series in telecommunications, New-York.

Csiszár, I. (1975) I-Divergence Geometry of Probability Distributions and Minimization Problems. *Annals of Probability*, **3**, 141–158.

Dembo, A. and Zeitouni, O. (1998) *Large Deviations Techniques and Applications*, 2nd edition. Springer, New York.

Girardin, V. and Lequesne, J. (2013) Relative Entropy Versus Entropy Difference in Goodness-of-Fit Tests. Application to Pareto Fitting. *Preprint Université de Caen Basse Normandie*, France.

Girardin, V. and Regnault P. (2011) *Large Deviation Principle for the Entropy of a Finite Distributions, with Applications, Preprint Université de Caen Basse Normandie*, France.

Harris, B. (1977) The statistical estimation of entropy in the non-parametric case. *Colloquia Mathematica Societatis JaÂ¡nos Bolyai*, **16**, 323–355, North-Holland, Amsterdam.

Kullback, S. and Leibler, R. A. (1951) On Information and Sufficiency. *Annals of Mathematical Statistics*, **29**, 79–86.

Pardo, L. (2006) *Statistical Inference Based on Divergence Measures* Chapman & Hall/CRC, Taylor and Francis Group, LLC.

Pielou (1967) The Use of Information Theory in the Study of the Diversity of Biological Populations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **4**, 163-177.

Regnault, P. (2011) *Différents problèmes liés à l'estimation de l'entropie de Shannon d'une loi, d'un processus de Markov*, PhD thesis, Université de Caen Basse Normandie, France.

Sgarro, A. (1978) An Informational Divergence Geometry for Stochastic Matrices. *Calcolo*, **15**, 41–49.

Shannon, C. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**, 379–423, 623–656.

Vasicek, O. (1976) A test for normality based on sample entropy *Journal of the Royal Statistical Society* **38**, 54–59.

Zubkov, A. M. (1973) Limit Distribution for a Statistical Estimator of the Entropy. *Theory of Probability and its Applications*, **18**, 611–618.