C. R. Acad. Sci. Paris, Ser. I 335 (2002) 763–766

Statistique/*Statistics*
(Probabilités/*Probability Theory*)

Sur l'estimation de l'entropie des lois à support dénombrable

Amor Keziou

LSTA, boîte courrier 158, 8A, Université Paris-6, 175, rue du Chevaleret, 75013 Paris, France

Reçu le 8 mars 2002 ; accepté après révision le 3 septembre 2002

Note présentée par Paul Deheuvels.

Résumé Soit P une loi de probabilité discrète sur un espace infini dénombrable $\mathcal{X}$. On étudie la vitesse de convergence presque sûre de l'estimateur « plug-in » de l'entropie $H := H(P)$ de la loi de probabilité inconnue P. On démontre aussi la convergence presque sûre de l'estimateur pour des variables aléatoires stationnaires ergodiques, et pour des variables aléatoires stationnaires α-mélangeantes sous une condition faible sur la queue de distribution de la loi P. ***Pour citer cet article : A. Keziou, C. R. Acad. Sci. Paris, Ser. I 335 (2002) 763–766.***

On entropy estimation for distributions with countable support

Abstract Suppose P is a discrete distribution on an infinite countable space $\mathcal{X}$. We study the almost surely convergence rate of the 'plug-in' estimate of the entropy $H := H(P)$ of the arbitrary distribution P. We prove also the consistency of the estimate for ergodic stationary random variables and for α-mixing stationary random variables under weak assumptions on the tail of the distribution P. ***To cite this article: A. Keziou, C. R. Acad. Sci. Paris, Ser. I 335 (2002) 763–766.***

1. Introduction et notations

Soit X une variable aléatoire discrète de loi de probabilité inconnue P sur un espace infini dénombrable $\mathcal{X}$. Pour tout x appartenant à $\mathcal{X}$, on note $p(x)$ la probabilité $P\{X = x\}$. L'entropie de Shannon de la loi P est définie par

$$H = H(P) := -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) = \mathbf{E}\big\{-\log_2 p(X)\big\},$$

où $\log_2$ est le logarithme de base 2, et $\mathbf{E}$ désigne l'espérance.

Adresse e-mail : keziou@ccr.jussieu.fr (A. Keziou).

S1631-073X(02)02559-1/FLA

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de loi P. L'estimateur « plug-in » $\widehat{H}_n$ de l'entropie H est défini par

$$\widehat{H}_n := H\big(\widehat{P}_n\big) = -\sum_{x\in\mathcal{X}} \widehat{p}_n(x)\log_2 \widehat{p}_n(x), \tag{1}$$

où $\widehat{P}_n$ est la mesure empirique définie par $\widehat{p}_n(x) := n^{-1}\sum_{i=1}^n \delta_{X_i}(x)$ pour tout $x \in \mathcal{X}$.

Bašarin a étudié les propriétés asymptotiques de l'estimateur plug-in $\widehat{H}_n$ de l'entropie de lois de probabilité discrètes à support fini (voir [3]). Antos et Kontoyiannis ont montré la convergence presque sûre (p.s.) universelle de $\widehat{H}_n$ dans le cas discret à support infini dénombrable. Ils ont montré aussi que pour toute vitesse de convergence $a_n \to 0$, il existe une loi P, telle que pour tout estimateur H_n de $H(P)$ on a $\limsup_{n\to\infty} \frac{\mathbf{E}|H_n - H(P)|}{a_n} = \infty$. Il n'existe donc pas de vitesse de convergence universelle pour les estimateurs de l'entropie (voir [1] et [2]). Le but de cette Note est d'expliciter des conditions sur de vastes classes $\mathcal{P}$ de lois P discrètes à support infini dénombrable, correspondant à des hypothèses naturelles dans le domaine du codage et de la compression universelle de sources avec alphabets grands ou infinis (voir [8,7] et les références ci-inclues) pour lesquelles on donnera des vitesses de convergence. Les références [1] et [2] fournissent des vitesses de convergence dans $\mathrm{L}_1(P)$ et $\mathrm{L}_2(P)$ sous des hypothèses assez contraignantes sur P, qui ne sont pas satisfaites par les modèles de Poisson et géométriques utilisés dans [8]. Nous démontrons aussi la convergence p.s. de l'estimateur dans le cas stationnaire ergodique et sationnaire α-mélangeant sous une condition faible sur la queue de distribution de P.

L'estimateur « plug-in » défini par (1) s'écrit sous la forme :

$$\widehat{H}_n := -\frac{1}{n}\sum_{i=1}^n \log_2 \widehat{p}_n(X_i). \tag{2}$$

Cette écriture facilite l'étude de la vitesse de convergence p.s.

2. Résultats

On donne trois propositions essentielles aux démonstrations des théorèmes à suivre :

PROPOSITION 2.1. – *Pour tout* $0 \leqslant \alpha < 1/2$, $\delta \geqslant 0$, *on a* :

$$\lim_{n\to+\infty} n^{\alpha}(\log_2 n)^{\delta} \sup_{x\in\mathcal{X}} \big|\widehat{p}_n(x) - p(x)\big| = 0 \quad p.s.$$

PROPOSITION 2.2. – *Pour tout* $0 \leqslant \alpha < 1/2$, $\delta \geqslant 0$, *on a* :

$$\lim_{n\to+\infty} n^{\alpha}(\log_2 n)^{\delta} \sup_{t\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{p(X_i)<t\}} - P\big\{p(X) < t\big\}\right| = 0 \quad p.s.$$

Introduisons l'ensemble $A_n = A_n(C,\beta) := \{x \in \mathcal{X};\ p(x) \geqslant Cn^{-\beta}\}$.

PROPOSITION 2.3. – *Soient* α, β, γ, C, *tels que* $0 < \beta < \alpha < 1/2$, $0 \leqslant \gamma \leqslant \alpha - \beta < 1/2$ *et* $C > 0$. *Alors, pour tout* $\delta \geqslant 0$, *on a* : $\lim_{n\to+\infty} n^{\gamma}(\log_2 n)^{\delta} \sup_{x\in A_n} |\log_2 \widehat{p}_n(x) - \log_2 p(x)| = 0$ *p.s.*

Dans tout ce qui suit on suppose que l'entropie H de la loi P est finie. L'hypothèse suivante sur la distribution de la variable X permet de trouver une vitesse de convergence p.s. de l'estimateur $\widehat{H}_n$:

(H1) Il existe $\beta' > 0$ tel que $\sum_{n=1}^{+\infty} nP\{p(X) < n^{-\beta'}\} < \infty$.

THÉORÈME 2.4. – *Supposons* (H1). *Soit* $0 \leqslant \gamma < 1/2$ *et* $\delta \geqslant 0$. *S'il existe* α, β *et* $C > 0$ *vérifiant* : $0 < \beta < \alpha < 1/2$, $0 < \gamma < \alpha - \beta < 1/2$ *et*

$$\text{(H2)} \quad \lim_{n\to+\infty} n^{\gamma}(\log_2 n)^{1+\delta} P\big\{p(X) < Cn^{-\beta}\big\} = 0,$$

alors, on a presque sûrement : $\lim_{n\to+\infty} n^{\gamma}(\log_2 n)^{\delta}|\widehat{H}_n - H| = 0$.

Démonstration. – Définissons la suite $H_n := n^{-1}\sum_{i=1}^{n} - \log_2 p(X_i)$. Par application de l'inégalité de Hoeffding (voir [5], p. 191) sous l'hypothèse (H1), on démontre la convergence p.s. de $n^{\gamma}(\log_2 n)^{\delta}|H_n - H|$ vers 0. En effet, d'après le lemme de Borel–Cantelli, l'hypothèse (H1) implique : $-H \leqslant Y_i := -\log_2 p(X_i) - H \leqslant \beta' \log_2 n - H$. Par application de l'inégalité de Hoeffding, on obtient :

$$P\big\{n^{\gamma}(\log_2 n)^{\delta}|H_n - H| > \varepsilon\big\} := P\big\{|Y_1 + \cdots + Y_n| > n^{1-\gamma}(\log_2 n)^{-\delta}\varepsilon\big\} \leqslant 2\exp\left[-\frac{2\varepsilon^2 n^{1-2\gamma}}{\beta'^2 \log_2^{2\delta+2} n}\right].$$

La série de terme général $2\exp[-2\varepsilon^2 n^{1-2\gamma}/\beta'^2 \log_2^{2\delta+2} n]$ est convergente pour tout $\varepsilon > 0$ et le lemme de Borel–Cantelli implique la convergence p.s. de $n^{\gamma}(\log_2 n)^{\delta}|H_n - H|$ vers 0. La suite de la démonstration consiste à montrer la convergence de $n^{\gamma}(\log_2 n)^{\delta}|\widehat{H}_n - H_n|$ vers 0 p.s. En utilisant des majorations similaires à celles utilisées dans [4], p. 86, on obtient :

$$\begin{aligned} n^{\gamma}(\log_2 n)^{\delta}\big|\widehat{H}_n - H_n\big| &\leqslant n^{\gamma}(\log_2 n)^{\delta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\in A_n\}}\big|\log_2 \widehat{p}_n(X_i) - \log_2 p(X_i)\big| \\ &\quad + n^{\gamma}(\log_2 n)^{\delta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\in A_n^c\}}\big|\log_2 \widehat{p}_n(X_i) - \log_2 p(X_i)\big| := A + B. \end{aligned}$$

A est majoré par $n^{\gamma}(\log_2 n)^{\delta}\sup_{x\in A_n}|\log_2 \widehat{p}_n(x) - \log_2 p(x)|$. D'après la Proposition 2.3 ce dernier est négligeable, donc le terme A tend vers 0 p.s.

On a :

$$\begin{aligned} B &\leqslant n^{\gamma}(\log_2 n)^{\delta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{p(X_i)<Cn^{-\beta}\}}\big|\log_2 \widehat{p}_n(X_i)\big|r \\ &\quad + n^{\gamma}(\log_2 n)^{\delta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{p(X_i)<Cn^{-\beta}\}}\big|\log_2 p(X_i)\big| := C + D. \end{aligned}$$

Or $C \leqslant n^{\gamma}(\log_2 n)^{1+\delta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{p(X_i)<Cn^{-\beta}\}}$. Donc d'après la Proposition 2.2, sous la condition (H2), le terme C tend vers 0 p.s. Le terme D lui aussi tend vers 0 p.s. En effet

$$\begin{aligned} D &= n^{\gamma}(\log_2 n)^{\delta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{n^{-\beta'}\leqslant p(X_i)<Cn^{-\beta}\}}\big|\log_2 p(X_i)\big| \\ &\leqslant n^{\gamma}\beta'(\log_2 n)^{1+\delta}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{p(X_i)<Cn^{-\beta}\}}. \end{aligned}$$

D'après la Proposition 2.2, sous la condition (H2), le dernier terme tend vers 0 p.s. □

2.1. Remarques

1. L'hypothèse (H1) est utilisée pour appliquer l'inégalité de Hoeffding et montrer la convergence presque sûre de $n^{\gamma}(\log_2 n)^{\delta}|H_n - H|$ vers 0. L'hypothèse (H2) est utiliséée pour montrer que $n^{\gamma}(\log_2 n)^{\delta}|\widehat{H}_n - H_n|$ est négligeable.

2. Les références [1] et [2] montrent que l'erreur $\mathrm{L}^1(P)$ de l'estimateur $\widehat{H}_n$ est de l'ordre de $n^{-(q-1)/q}$ si la loi P vérifie $c_1/i^q \leqslant p(i) \leqslant c_2/i^q$ pour tout $i \in \mathbb{N}^*$ avec $1 < q < 2$ et $c_1, c_2 > 0$ (voir [1], Théorème 7). Cette condition n'est pas vérifiée pour les lois géométriques et les lois de Poisson.
3. Les hypothèses (H1) et (H2) contrôlent la probabilité d'observer une modalité correspondant à une petite probabilité, ce qui est plus faible et plus naturelle qu'une hypothèse sur les queues de distribution.

La condition suivante permet de démontrer la convergence p.s. de l'estimateur pour des variables aléatoires stationnaires ergodiques et pour des variables aléatoires stationnaires α-mélangeantes :

(H3) Il existe $0 < \beta < 1/2$ et $C > 0$ tels que $\lim_{n\to+\infty}(\log_2 n)P\{p(X) < Cn^{-\beta}\} = 0$.

Soit $(\alpha_n)_{n\in\mathbb{N}}$ la suite de coefficients de mélange fort définie par : $\alpha_0 := 1/2$ et pour tout $n \in \mathbb{N}$; $\alpha_n := \sup_{k\in\mathbb{Z}} \alpha(\mathcal{F}_k, \sigma(X_{k+n}))$ où $\mathcal{F}_k := \sigma(X_i, i \leqslant k)$ est la sigma-algèbre engendrée par $(X_i, i \leqslant k)$ et α est le coefficient de mélange fort défini pour toutes tribus $\mathcal{A}$, $\mathcal{B}$ par :

$$\alpha(\mathcal{A}, \mathcal{B}) := 2\sup\big\{\big|P(A \cap B) - P(A)P(B)\big|;\ (A, B) \in \mathcal{A} \times \mathcal{B}\big\}.$$

THÉORÈME 2.5. –

(a) *Soit $X_1, \ldots, X_n$ une suite de variables aléatoires stationnaires et ergodiques de loi P. Sous l'hypothèse* (H3)*, l'estimateur $\widehat{H}_n$ est p.s. convergent.*

(b) *Soit $X_1, \ldots, X_n$ une suite de variables aléatoires stationnaires. Sous l'hypothèse* (H3)*, si la suite de mélange fort $(\alpha_n)_{n\in\mathbb{N}}$ vérifie : $\sum_{n\geqslant 0} \frac{\alpha_n}{n+1} < \infty$, alors, l'estimateur $\widehat{H}_n$ est p.s. convergent.*

Démonstration. – Pour démontrer le Théorème 2.5, il suffit de démontrer la convergence p.s. uniforme de $|\widehat{p}_n(x) - p(x)|$ vers 0 pour les deux cas (a) et (b) (voir la démonstartion du Théorème 2.4). Sans perte de généralité on suppose que l'espace $\mathcal{X}$ est l'ensemble $\mathbb{N}$. On a :

$$\forall \varepsilon > 0,\ \exists x_0(\varepsilon) \in \mathcal{X},\ \exists n_0(\varepsilon, x_0) > 0,\ \forall n \geqslant n_0 : \quad \sup_{x\in\mathcal{X}} \big|\widehat{p}_n(x) - p(x)\big| \leqslant \sup_{x<x_0} \big|\widehat{p}_n(x) - p(x)\big| + \varepsilon.$$

Le cas (a) : $\sup_{x<x_0} |\widehat{p}_n(x) - p(x)|$ converge vers 0 p.s. car les variables $\delta_{X_1}(x) - p(x), \ldots, \delta_{X_n}(x) - p(x)$ sont stationnaires et ergodiques si $X_1, \ldots, X_n$ le sont.

Le cas (b) : on a convergence p.s. de $\sup_{x<x_0} |\widehat{p}_n(x) - p(x)|$ vers 0 (voir [6], p. 55). □

Remerciements. Je tiens à remercier vivement le professeur M. Broniatowski pour ses discussions et suggestions qui ont mené à améliorer cette Note.

Références bibliographiques

[1] A. Antos, I. Kontoyiannis, Convergence properties of functional estimates for discrete distributions, Random Structures Algorithms 1 (2001) 163–193.
[2] A. Antos, I. Kontoyiannis, Estimating the entropy of discrete distributions, IEEE Internat. Sympos. Inform. Theory 1 (2001) 45–51.
[3] G.P. Bašarin, On a statistical estimate for the entropy of a sequence of independent random variables, Theory Probab. Appl. 4 (1959) 333–336.
[4] E. Guerre, Méthodes non paramétriques d'analyse des séries temporelles multivariées : estimation de mesures de dépendance, Doc. d'univ., Math., Paris 6, 1993.
[5] D. Pollard, Convergence of Stochastic Processes, Springer-Verlag, 1984.
[6] E. Rio, Théorie asymptotique des processus aléatoires faiblement dépendants, Springer-Verlag, 2000.
[7] S. Verdú, Fifty years of Shannon theory, IEEE Trans. Inform. Theory 44 (6) (1998) 2057–2078. Information theory: 1948–1998.
[8] E.H. Yang, Y. Jia, Universal lossless coding of sources with large or unbounded alphabets, in: I. Althfor, et al. (Eds.), Numbers, Information and Complexity, Kluwer Academic, 2000, pp. 421–442.