ON THE IMPORTANCE OF ENTROPY

Martin Janžura

Keywords: Entropy, I-divergence, maximum likelihood, I-projection, large deviations, conditional limit theorem.

Abstract: The aim of the paper consists in demonstrating the relevance of the fundamental information-theoretic concepts, namely the entropy and the I-divergence, for both the statistical inference and the limit theorems of probability theory.

Abstrakt: Cílem je ukázat význam základních pojmů teorie informace, tj. entropie a I-divergence, jak pro statistické úlohy tak i pro limitní věty teorie pravděpodobnosti.

1 Introduction

Entropy, as a fundamental concept arising from statistical physics, was originally understood as a thermodynamical property of heat engines (see, e.g., [8] or [12]). Later, in the pioneering Shannon's paper [13] it was introduced as a crucial quantity of information theory. Its relevance as a measure of uncertainty (ignorance, information) was early recognized and widely exploited (see [3], [4], [10]). As the number of different names for the same concept indicates, the more general quantity of I-divergence (relative entropy, Kullback-Leibler number, information gain) became also widely used (see again [3], [4], or [11], [14], and the references therein).

The purpose of the present paper consists in showing even more fundamental importance of the concepts for the area of mathematical statistics and probability theory. We intend to illustrate how the quantities are inherent, especially, for the multidimensional joint distributions, and how they arise naturally in various situations.

Similar contents as here can be found, e.g., in [8]. For many topics of the present paper, in particular for the limit theorems, [3] or [4] are the basic references. For exponential distributions, closely related to the maximum entropy principle, see [2], or, more generally, [5]. For statistical problems see [11] or [14]. The presented results can be also generalized from the I.I.D. case to the random processes or fields, see [6], [7], [9],and [15].

2 Basic definitions and properties

Let us consider a finite state space $\mathcal{X} = \{x_1, \dots, x_M\}$.

By $\mathcal{P}(\mathcal{X})$ we denote the class of all probability measures on $\mathcal{X}$, and by $\mathcal{F}(\mathcal{X})$ the class of all real-valued functions on $\mathcal{X}$. In particular, by $R \in \mathcal{P}(\mathcal{X})$ we denote the *uniform* distribution, i.e., $R(x) = \frac{1}{M}$ for every $x \in \mathcal{X}$.

Let us recall the formulas for the *entropy* and the *I-divergence*, respectively, namely

$$H(P) = \int -\log P \,\mathrm{d}P = \sum_{x\in\mathcal{X}} -\log P(x)\, P(x),$$

for $P \in \mathcal{P}(\mathcal{X})$, and

$$I(P|Q) = \int \log \frac{P}{Q}\,\mathrm{d}P = \sum_{x\in\mathcal{X}} \log \frac{P(x)}{Q(x)}\, P(x)$$

for $P, Q \in \mathcal{P}(\mathcal{X})$, providing the terms are well defined. Otherwise we set $I(P|Q) = \infty$.

Proposition 1. For $P, Q \in \mathcal{P}(\mathcal{X})$ it holds

i) $H(P) \in [0, \log M]$;

ii) $H(P) = 0$ iff $P(x_i) = 1$ for some $i \in \{1, \dots, M\}$ and $P(x_j) = 0$ for every $j \neq i$;

iii) $H(P) = \log M$ iff $P = R$.

iv) $I(P|Q) \in [0, \infty]$;

v) $I(P|Q) = 0$ iff $P = Q$;

vi) $I(P|Q) = \infty$ iff $P \not\ll Q$.

Proof.

$H(P) \geq 0$ follows from $-\log P(x) \geq 0$ for every $x \in \mathcal{X}$. Thus $H(P) = 0$ can occur only if $-\log P(x) \equiv 0$ which proves ii). $H(P) \leq \log M$ and iii) follow from iv) and v) with $Q = M$.

iv) and v) are due to Jensen's inequality, vi) is obvious. □

Remark.

Due to the above properties of the entropy, which is minimal for non-random case and maximal for the uniform distribution, it can be understood as a measure of *uncertainty* contained in the probability distribution. Namely, if the entropy is zero then the output is sure, if it is maximal then all possible outputs are equally likely.

The I-divergence, on the other hand, can serve as a *distance*, being equal to zero for identical distributions, and maximal, equal to ∞, if the first distribution is not supported on the same set as the second one.

3 Maximum entropy principle

As a rule, whenever we have no information about the distribution of some random phenomenon, we turn to the uniform distribution. It is justified by the fact that we have *no reason* for preferring any particular output. Thus,

in the light of the above Remark, we opt for the distribution with maximum entropy.

Such approach can be extended (see, e.g., [10] as the standard reference) to the general *maximum entropy principle* (MAXENT). Suppose we have only a partial information about the distribution, namely $P \in \mathcal{E}$ where $\mathcal{E} \subset \mathcal{P}(\mathcal{X})$.

Then, applying the MAXENT, we seek for

$$\overline{P}_{\mathcal{E}} \in \operatorname{argmax}_{P\in\mathcal{E}} H(P)$$

or, more generally,

$$\overline{P}_{\mathcal{E}} \in \operatorname{argmin}_{P\in\mathcal{E}} I(P|Q)$$

where $Q \in \mathcal{P}(\mathcal{X})$ is some fixed reference probability measure.

Usually, the first definition, which, after all, agrees with the latter one for the uniform $Q = R$, is meant by the maximum entropy principle.

Example (maximum entropy with linear constraints). Let us consider a collection of statistics $\boldsymbol{f} = \{f_j\}_{j\in\mathcal{K}}$ with $|\mathcal{K}| < \infty$, where $f_j \in \mathcal{F}(\mathcal{X})$ for every $j \in \mathcal{K}$. Moreover, in order to guarantee the basic *regularity (identifiability) condition*, we assume the system $(1,\ \{f_j\}_{j\in\mathcal{K}})$ to be linearly independent.

Now, for a collection of constants $\boldsymbol{m} = \{m_j\}_{j\in\mathcal{K}}$ we denote

$$\mathcal{E} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{f}) = \{P \in \mathcal{P}; \int f_j \, \mathrm{d}P = m_j \quad \text{for every } j \in \mathcal{K}\}.$$

Further, let us introduce the *exponential distribution* P^α given by

$$P^\alpha(x) = \exp\left\{\sum_{j\in\mathcal{K}} \alpha_j \, f_j(x) - c(\alpha)\right\}$$

where $\alpha = (\alpha_j)_{j\in\mathcal{K}} \in R^{\mathcal{K}}$ is a parameter, and the appropriate normalizing constant is given by $c(\alpha) = \log \sum_{x\in\mathcal{X}} \exp\left\{\sum_{j\in\mathcal{K}} \alpha_j \, f_j(x)\right\}$.

Then, we may deduce the following properties:

i) There is a one-to-one relation between the parameter α and the exponential distribution P^α. Namely, for $P^\alpha = P^\beta$ we have $\langle \alpha - \beta, \boldsymbol{f}\rangle = \text{const.}$, and the statement holds thanks to the identifiability condition above.

ii) Let $P^\alpha,\ P^\beta \in \mathcal{E}$. Then $\alpha = \beta$. We observe

$$0 \le I(P^\alpha|P^\beta) + I(P^\beta|P^\alpha) = \langle \beta - \alpha,\ \int \boldsymbol{f} \, \mathrm{d}P^\beta - \int \boldsymbol{f} \, \mathrm{d}P^\alpha\rangle = 0.$$

Hence $P^\alpha = P^\beta$, and, due to i), we have $\alpha = \beta$.

iii) Let $P^\alpha \in \mathcal{E}$. Then $\overline{P}_{\mathcal{E}}$ is given uniquely, and $\overline{P}_{\mathcal{E}} = P^\alpha$.

As it is well-known, we have

$$0 \le I(P|P^\alpha) = c(\alpha) - \langle \alpha,\ \boldsymbol{m}\rangle - H(P) = H(P^\alpha) - H(P)$$

where, by Proposition 1.v), the inequality turns into equality iff $P = P^{\alpha}$.

We may conclude that whenever there exists the *exponential representative* $P^{\overline{\alpha}} \in \mathcal{E} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{f})$ then it is given uniquely and satisfies the MAXENT. Thus, the exponential distributions maximize the entropy under the linear constraints, and, since the linear constraints are rather standard form of partial information (see, e.g., the moment conditions in mathematical statistics), the exponential families are well justified as probability models.

4 I.I.D. sequences

Let us consider a sequence $\overline{x}_n = (x_1, \dots, x_n)$ where $x_i \in \mathcal{X}$ for every $i = 1, \dots, n$. Let us denote by $N^{\overline{x}_n}(y) = \sum_{i=1}^{n} \delta(y, x_i)$ the *number of occurrences* of the state $y \in \mathcal{X}$ in the sequence $\overline{x}_n$. Consequently, we shall denote by $P^{\overline{x}_n} = \frac{1}{n} N^{\overline{x}_n}$ the *empirical distribution* induced by the sequence $\overline{x}_n$.

Now, we understand the sequence $\overline{x}_n$ as a collection of data obtained from a sequence of I.I.D. random variables with a one-body marginal distribution $P \in \mathcal{P}(\mathcal{X})$. Then, for the joint distribution, we may easily observe

$$P_n(\overline{x}_n) = \prod_{i=1}^{n} P(x_i) = \prod_{y \in \mathcal{X}} P(y)^{N^{\overline{x}_n}(y)} = \exp\{-n[D(P^{\overline{x}_n}|P) + H(P^{\overline{x}_n})]\}.$$

As a result we may express the *log-likelihood* as

$$-\frac{1}{n} \log P_n(\overline{x}_n) = D(P^{\overline{x}_n}|P) + H(P^{\overline{x}_n}).$$

Thus, we may conclude that, first, the empirical distribution is a sufficient statistics for the joint distribution, and, moreover,the joint distribution depends on the empirical distribution just through the above quantities, namely the *entropy* and the *I-divergence*. Let us emphasize that such relation is not imposed or artificial, it is apparently natural and inherent for the joint probability distributions.

5 Statistical problems

5.1 Parameter estimation

In this section let us consider a *parametric family* of probability distributions $\{P^{\theta}\}_{\theta \in \Theta}$ where $P^{\theta} \in \mathcal{P}(\mathcal{X})$ for every $\theta \in \Theta$. For the sake of simplicity we shall assume $P^{\theta} > 0$ for every $\theta \in \Theta$ (we may, e.g., imagine the exponential family as introduced in Section 3).

Based on a data sequence $\overline{x}_n = (x_1, \dots, x_n)$, we may define the *maximum likelihood estimate* (MLE) standardly as : $\hat{\theta}^n = \arg\max_{\theta \in \Theta} \log P_n^{\theta}(\overline{x}_n)$.

But, due to the above formula for the log-likelihood, we have also

$$\hat{\theta}^n = \arg\min_{\theta\in\Theta} D(P^{\overline{x}_n}|P^\theta).$$

Thus, the MLE can be alternatively understood as the *Minimum I-divergence estimate* which provides us with an additional justification for the MLE: we seek for such value of the parameter that makes the theoretical distribution as close (in the sense of I-divergence) to the empirical one as possible. Conversely, if we stay within the framework of "minimum distance estimation" then the distance measured by the I-divergence is privileged since it yields the maximum likelihood estimation with all its favourable properties.

Remark(Consistency). The consistency of the MLE is in general rather well-known. Nevertheless, it can be simultaneously derived from the minimum I-divergence approach. Let us give a sketch of the proof:

Suppose $\theta_0 \in \Theta$ is the true parameter. Then, obviously,

$$\theta_0 = \arg\min_{\theta\in\Theta} D(P^{\theta_0}|P^\theta).$$

Denote $L_n(\theta) = D(P^{\overline{x}_n}|P^\theta)$ for every $n = 1, ...,$ and $L_0(\theta) = D(P^{\theta_0}|P^\theta)$. Then, due to the law of large numbers, we obtain $L_n(\theta) \longrightarrow_{n\to\infty} L_0(\theta)$ a.s.$[P]$ point-wise for every $\theta \in \Theta$. But, in order to prove

$$\hat{\theta}^n = \arg\min_{\theta\in\Theta} L_n(\theta) \longrightarrow_{n\to\infty} \arg\min_{\theta\in\Theta} L_0 = \theta_0$$

we need the above convergence uniform at least on every compact set. The latter is satisfied if all the functions L_n are convex, which is the case, e.g., of the exponential families.

5.2 Testing hypotheses

Similarly as for the estimates, the likelihood ratio test may be understood as tests based on the I-divergence. In particular, for $\Theta = R^K$, let us consider the test of a simple hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \Theta \setminus \{\theta_0\}$. Then

$$2nD(P^{\hat{\theta}^n}|P^{\theta_0}) \Longrightarrow_{n\to\infty} \chi^2_K \quad \text{in distribution} \quad [P^{\theta_0}].$$

More generally, for a subspace or a hyperplane $\Theta_0 \subset\subset \Theta$, we may consider the affine hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta \setminus \Theta_0$. Then

$$2nD(P^{\hat{\theta}^n}|P^{\hat{\theta}^n_0}) \Longrightarrow_{n\to\infty} \chi^2_{K-\dim(\Theta_0)} \quad \text{in distribution} \quad [P^\theta]$$

for every $\theta \in \Theta_0$, where

$$\hat{\theta}^n_0 = \arg\min_{\theta\in\Theta_0} D(P^{\overline{x}_n}|P^\theta).$$

For more detailed treatment see, e.g., [9] or [14].

6 I.I.D. sequences and types

From Section 3 we know $P_n(\overline{x}_n) = \exp\{-n[D(P^{\overline{x}_n}|P) + H(P^{\overline{x}_n})]\}$ providing $x_1, \dots, x_n$ are drawn I.I.D. according to $P \in \mathcal{P}(\mathcal{X})$.

Now, let $\overline{x}_n \in T_n(Q) = \{\overline{x}_n : P^{\overline{x}_n} = Q\}$ for some $Q \in \mathcal{P}$, we say $\overline{x}_n$ is in the *type class* of Q, then

$$P_n(\overline{x}_n) = \exp\{-n[D(Q|P) + H(Q)]\}.$$

In particular, for $\overline{x}_n \in T_n(P)$ we have $P_n(\overline{x}_n) = \exp\{-nH(P)\}$.

At the same time, by combinatorial arguments we can obtain $|T_n(P)| \doteq \exp\{nH(P)\}$ or, more precisely

$$\exp\{nH(P) + o(n)\} \le |T_n(P)| \le \exp\{nH(P)\}$$

(see, e.g. [3]), and, therefore

$$P_n[T_n(P)] \doteq 1$$

while, in general,

$$P_n[T_n(Q)] \doteq \exp\{-nD(Q|P)\}.$$

Thus, we may conclude that the joint distribution P_n is approximately supported on the set $T_n(P) \subset \mathcal{X}^n$, that contains approximately $e^{nH(P)}$ configurations. And each of the configurations has the equal probability $e^{-nH(P)}$. In such a way the joint distribution $P_n$is essentially determined by the entropy $H(P)$.

As a corollary we obtain the *law of large numbers*. Namely, for $g : \mathcal{X} \to R$ we obtain

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) = \int g dP^{\overline{x}_n} \doteq \int g dP.$$

7 Limit theorems

From the preceding section we know that every joint distribution P_n is concentrated on the configurations from its own type class $T_n(P)$. Nevertheless, we are still interested in the behaviour of P_n outside its type class. Such behaviour is again characterized by the entropy and related notions.

Definition(I-projection). For $\mathcal{E} \subset \mathcal{P}(\mathcal{X})$ we denote $D(\mathcal{E}|P) = \inf_{Q\in\mathcal{E}} D(Q|P)$. If there exists $P_* = \arg\min_{Q\in\mathcal{E}} D(Q|P)$ we call it *I-projection*.

Due to compactness of $\mathcal{P}(X)$ and continuity of $D(\bullet|P)$ we observe that P_* is attained if $\mathcal{E} \subset \mathcal{P}(\mathcal{X})$ is closed.

The proofs of the following results can be found. e.g., in [3] and [4], or, in the most general form, also in [7].

Theorem 1(Large deviations). If $\mathcal{E} = \mathrm{cl}(\mathrm{int}\mathcal{E})$ then

$$-\frac{1}{n}\log P_n(P^{\overline{x}_n} \in \mathcal{E}) \longrightarrow_{n\to\infty} D(P_*|P).$$

□

The above theorem has many interesting consequences, as an example let us introduce the following result.

Corollary(Test power). When testing a simple hypothesis $\mathcal{H}_0 : P = P^0$ against a simple alternative $\mathcal{H}_1 : P = P^1$ with a level (first kind error) equal to $\alpha \in (0, 1)$ we obtain for the second kind error

$$-\frac{1}{n}\log \beta_n \longrightarrow_{n\to\infty} D(P^0|P^1).$$

□

The theorem on large deviations proclaims that, whenever the distance $D(\mathcal{E}|P)$ is positive, the probability $P_n(P^{\overline{x}_n} \in \mathcal{E})$ is very small, it tends to zero exponentially fast. The following important and nice theorem shows what will happen if we still "enforce" the distribution to concentrate on such a "small" set.

Theorem 2 (Conditional limit theorem). Let $\mathcal{E}$ be a closed convex subset of $\mathcal{P}(X)$ and $P \notin \mathcal{E}$. Then

$$P_n(\bullet|P^{\overline{x}_n} \in \mathcal{E}) \Longrightarrow_{n\to\infty} P_*^{\infty} \quad \text{(weakly)}.$$

□

The conditional limit theorem proofs, e.g., one of the fundamental results of statistical physics.

Remark (the second law of thermodynamics). For $P = R$ we have $P_* = \arg\max_{Q\in\mathcal{E}} H(Q)$. That means the conditional distribution attains (at infinity) the maximum entropy.

Usually we have the set $\mathcal{E}$ given by a linear constraint (see Section 3), i.e. $\mathcal{E} = \{Q \in \mathcal{P}(\mathcal{X}); \int E dQ \le c\}$. Then we obtain the exponential distribution $P_* \propto \exp\{\alpha E\} \cdot P$ as the I-projection.

Remark (importance sampling). The above conditional limit theorem can be also used for approximating the conditional distribution, e.g., for the simulation method known as "importance sampling".

Namely, for estimating some "small" probability $P(A)$ we can use $\widehat{P(A)} = \frac{1}{k}\sum_{i=1}^{k} I_A(y_i)\frac{P(y_i)}{Q(y_i)}$ where $y_1, \ldots, y_k$ are drawn I.I.D. with some $Q \in \mathcal{P}(\mathcal{X})$.

From the minimum variance point of view, the optimal choice is $Q = P(\bullet|A)$, and, whenever we have $A = A_n = \{\overline{x}_n; P^{\overline{x}_n} \in \mathcal{E}\}$, we may use the approximation $P_n(\bullet|A_n) \doteq P_*^n$ (see [1]).

References

[1] Antoch, J. (2006) *O simulaci řídkých jevů.* In: Robust 2006 (Eds.: J. Antoch, G. Dohnal), JČMF, Praha.

[2] Barndorff–Nielsen, O. E. (1978) *Information and Exponential Families in Statistical Theory.* Wiley , New York.

[3] Cover T.M., Thomas J.A. (1991) *Elements of Information Theory.* Wiley, New York.

[4] Csiszár I.,Körner J. (1981) *Information Theory.* Akademiai Kiado, Budapest.

[5] Csiszár I. and Matúš F. (2006) *Generalize maximum likelihood estimates for exponential families.* Probability Theory and Related Fields **141**, p. 213–246.

[6] Föllmer, H.(1973) *On entropy and information gain in random fields*, Z. Wahrs. verw. Geb. **26**, 207–217.

[7] Georgii, H.-O. (1993) *Large deviations and maximum entropy principle for interacting random fields on Zd.* Ann. Prob. **21**, 1845–1875.

[8] Georgii, H.-O. (2003) *Probabilistic aspects of entropy.* In: Entropy(Eds.: A. Greven, G. Keller, G. Warnecke), Princeton University Press , 37–54.

[9] Janžura M. (1997) *Asymptotic results in parameter estimation for Gibbs random fields.* Kybernetika **33** , 2, 133–159.

[10] Jaynes E. T. (1982) *On the rationale of the maximum entropy methods.* Proc. IEEE **70**, 939–952.

[11] Liese F., Vajda I. (2006) *On divergences and informations in statistics and information theory* , IEEE Transactions on Information Theory **52**, 10 (2006), 4394-4412.

[12] Sethna J.P. (2006) *Statistical Mechanics: Entropy, Order Parameters, and Complexity.* Oxford University Press.

[13] Shannon, C.E. (1948) *A mathematical theory of communication.* Bell System Techn. J. **27**, 379–423, 623–657.

[14] Vajda I. (1989) *Theory of Statistical Inference and Information.* Kluwer Academic Publishers, Dordrecht.

[15] Winkler G. (1995)*Image Analysis, Random Fields and Dynamic Monte Carlo Methods.* Springer–Verlag, Berlin .

Acknowledgement: Supported by grant GA ČR No.201/06/1323 and Research Center DAR (MŠMT ČR Project No. 1M0572).

Address: Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic

E-mail: janzura@utia.cas.cz