BULLETIN OF THE BRAZILIAN MATHEMATICAL SOCIETY

Bull Braz Math Soc, New Series 37(4), 1-7

A simple proof of Sanov's theorem*

Imre Csiszár

Abstract. A simple self-contained proof of Sanov's theorem in τ-topology is given, well suited for a first course on large deviations.

Keywords: ??? .

Mathematical subject classification: ???.

1 Intoduction, notation

A simple proof of Sanov's theorem in τ-topology will be given. Its main ideas appear in Groeneboom et al. [4], a novelty is that the use of topological concepts is reduced to the most basic ones. In my experience, this proof is ideally suited for a first course on large deviations or a course on applications of information theory in probability and statistics. After my course at the 2005 Brasilian School of Probability, its organizer and host Antonio Galves, to whom I am pleased to thank for his hospitality, has encouraged me to publish this proof.

The set of all probability measures (PMs) on a given measurable space $(\mathcal{X}, \mathcal{F})$ will be denoted by $\mathcal{P}$, and the set of all partitions $\mathcal{A} = (A_1, \ldots, A_k)$ of $\mathcal{X}$ into a finite number of sets $A_i \in \mathcal{F}$ is denoted by $\prod$. For $P \in \mathcal{P}$, $\mathcal{A} \in \prod$, and $\epsilon > 0$, denote

$$U(P, \mathcal{A}, \epsilon) = \{P' \in \mathcal{P} : |P'(A_i) - P(A_i)| < \epsilon,\ i = 1, \ldots, k\}. \qquad (1)$$

The τ-topology on $\mathcal{P}$ is the coarsest topology in which the mappings $P \mapsto P(F)$ are continuous for all $F \in \mathcal{F}$; a base for this topology is the collection of the sets (1). The interior and closure in τ-topology of a set $\Gamma \subset \mathcal{P}$ are denoted by $int_\tau \Gamma$ and $cl_\tau \Gamma$.

Received 10 October 2005.

*This work was supported by the Hungarian National Foundation for Scientific Research under Grant T046376

For probability distributions on the finite set $\{1, \dots, k\}$, say $P = (p_1, \dots, p_k)$ and $Q = (q_1, \dots, q_k)$, information divergence (also called Kullback-Leibler distance or relative entropy) is defined as

$$D(P||Q) = \sum_{i=1}^{k} p_1 \log \frac{p_i}{q_i},$$

with the conventions $0 \log 0 = 0 \log \frac{0}{0} = 0$, $t \log \frac{t}{0} = +\infty$ if $t > 0$. The information divergence of PMs $P \in \mathcal{P}$ and $Q \in \mathcal{P}$ is defined as

$$D(P||Q) = \sup_{\mathcal{A} \in \prod} D(P^{\mathcal{A}}||Q^{\mathcal{A}}) \tag{2}$$

where $P^{\mathcal{A}} = (P(A_1), \dots, P(A_k))$, $Q^{\mathcal{A}} = (Q(A_1), \dots, Q(A_k))$. The well-known integral representation

$$D(P||Q) = \begin{cases} \int \log \frac{dP}{dQ}\, dP & \text{if } P \ll Q \\ +\infty & \text{otherwise} \end{cases}$$

will not be used in this paper.

The cardinality of a finite set T is denoted by $|T|$. The empirical distribution of an n-tuple $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$ is the PM $\hat{P}_{\mathbf{x}} \in \mathcal{P}$ defined by

$$\hat{P}_{\mathbf{x}}(F) = \frac{1}{n} |\{i : x_i \in F\}|, \ F \in \mathcal{F}.$$

The following version of Sanov's theorem (Sanov [5]) is addressed, see Dembo and Zeitouni [3], Theorem 6.2.10:

Sanov's Theorem. *For independent drawings from a distribution $Q \in \mathcal{P}$, the empirical distributions of the resulting samples satisfy the large deviations principle in τ-topology, with the good rate function $D(\,.\,||Q)$.*

This means that for $\Gamma \subset \mathcal{P}$

$$\liminf_{n\to\infty} \frac{1}{n} \log Q^n(\{\mathbf{x} : \hat{P}_{\mathbf{x}} \in \Gamma\}) \geqslant - \inf_{P \in int_\tau \Gamma} D(P||Q), \tag{3}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log Q^n(\{\mathbf{x} : \hat{P}_{\mathbf{x}} \in \Gamma\}) \leqslant - \inf_{P \in cl_\tau \Gamma} D(P||Q), \tag{4}$$

and the "divergence balls"

$$B(Q, \alpha) = \{P : D(P||Q) \leqslant \alpha\} \tag{5}$$

are compact in the τ-topology.

Remark 1. To make sure that $Q^n(\{\mathbf{x} : \hat{P}_\mathbf{x} \in \Gamma\})$ is well defined, usually a measurability condition is imposed on the permissible sets $\Gamma \subset \mathcal{P}$. Alternatively, Γ may be any subset of $\mathcal{P}$ if $Q^n(\{\mathbf{x} : \hat{P}_\mathbf{x} \in \Gamma\})$ is interpreted as inner measure in eq. (3) and as outer measure in eq. (4), see [4]. The proof in Section 2 covers also this stronger version of the theorem.

Remark 2. The lower bound (3) can be sharpened, replacing interior in τ-topology by interior in τ_0-topology, see Csiszár [1]; a base for that topology is the collection of sets $U_0(P, \mathcal{A}, \epsilon)$ defined for $\mathcal{A} = (A_1, \ldots, A_k) \in \prod$ and $\epsilon > 0$ by

$$U_0(P, \mathcal{A}, \epsilon) = \{P' \in U(P, \mathcal{A}, \epsilon),\ P'(A_i) = 0 \text{ if } P(A_i) = 0\}.$$

2 Proof of Sanov's theorem

The only prerequisites are two simple combinatorial lemmas, stated below. These are standard tools in information theory, and in a course on large deviations they are introduced early on, to prove a version of Sanov's theorem for the case when $\mathcal{X}$ is a finite set (see [2], Lemmas 1.2.2 and 1.2.6 or [3], Lemmas 2.1.2 and 2.1.9.)

Let $\mathcal{P}_n(k)$ denote the set of probability distributions on $\{1, \ldots, k\}$ of form $P = (\frac{n_1}{n}, \ldots, \frac{n_k}{n})$, with integers $n_1, \ldots, n_k$. For such P, let $\mathcal{T}_n(k)$ denote the set of those length-n sequences of elements of $\{1, \ldots, k\}$ in which each $i \in \{1, \ldots, k\}$ occurs n_i times.

Lemma 1. $|\mathcal{P}_n(k)| \leqslant (n+1)^k$.

Lemma 2. *For $P \in \mathcal{P}_n(k)$ and any distribution Q on $\{1, \ldots, k\}$,*

$$(n+1)^{-k} e^{-nD(P||Q)} \leqslant Q^n(\mathcal{T}_n(P)) \leqslant e^{-nD(P||Q)}.$$

We procced to prove

(i) the lower bound (3), in the stronger form mentioned in Remark 2;

(ii) the compactness in τ-topology of the divergence balls (5);

(iii) the upper bound (4).

The proof of (i), included to keep the paper self-contained, is the same as that of Lemma 4.1 in [1]. The proofs of (ii) and (iii) are simplified versions of corresponding proofs in [4], using only basic concepts from topology, in particular in the proof of the key inequality (8).

(i) The claim is that

$$\liminf_{n\to\infty} \frac{1}{n} \log Q^n(\{\mathbf{x} : \hat{P}_{\mathbf{x}} \in \Gamma\} \geqslant -D(P||Q) \tag{6}$$

for all PMs $P \in int_{\tau_0}(\Gamma)$, see Remark 2, or equivalently, for $P \in \mathcal{P}$ such that $U_0(P, \mathcal{A}, \epsilon) \subset \Gamma$ for some $\mathcal{A} \in \prod$ and $\epsilon > 0$. Now, pick $\bar{P}_n \in \mathcal{P}_n(k)$, $n = 1, 2, \ldots$ such that $\bar{P}_n \to P^{\mathcal{A}}$, with $\bar{P}_n(i) = 0$ when $P(A_i) = 0$. Then $|\bar{P}_n(i) - P(A_i)| < \epsilon_n$, $i = 1, \ldots, k$, for suitable $\epsilon_n \to 0$, thus for all n with $\epsilon_n \leqslant \epsilon$

$$\begin{aligned} Q^n(\{\mathbf{x} : \hat{P}_{\mathbf{x}} \in \Gamma\}) &\geqslant Q^n(\{\mathbf{x} : \hat{P}_{\mathbf{x}} \in U_0(P, \mathcal{A}, \epsilon_n)\}) \geqslant Q^n(\{\mathbf{x} : \hat{P}^{\mathcal{A}}_{\mathbf{x}} = \bar{P}_n\}) \\ &= (Q^{\mathcal{A}})^n(\mathcal{T}_n(\bar{P}_n)) \geqslant (n+1)^{-k} e^{-nD(\bar{P}_n||Q^{\mathcal{A}})}; \end{aligned}$$

the last inequality holds by Lemma 2. As the choice of $\bar{P}_n$ makes sure that $D(\bar{P}_n||Q^{\mathcal{A}}) \to D(P^{\mathcal{A}}||Q^{\mathcal{A}})$, and $D(P^{\mathcal{A}}||Q^{\mathcal{A}}) \leqslant D(P||Q)$ by (2), the claim (6) follows.

(ii) Let $\mathcal{M}$ denote the set of all finitely additive set functions on $(\mathcal{X}, \mathcal{F})$ with values in the interval $[0, 1]$. Clearly, $\mathcal{M}$ is a closed subset of the set $[0, 1]^{\mathcal{F}}$ of all functions $f : \mathcal{F} \to [0, 1]$ endowed with the product topology, the coarsest one in which all mappings $f \mapsto f(F)$ $(F \in \mathcal{F})$ are continuous. As $[0, 1]^{\mathcal{F}}$ is compact (Tychonoff's theorem, see [3], p.345), so is also $\mathcal{M}$.

The definition (2) of information divergence extends unchanged to P and Q in $\mathcal{M}$. Clearly, for any partition $\mathcal{A} \in \prod$, the subset $\{P \in \mathcal{M} : D(P^{\mathcal{A}}||Q^{\mathcal{A}}) \leqslant \alpha\}$ of $\mathcal{M}$ is closed, hence compact, and therefore

$$K = \{P \in \mathcal{M} : D(P||Q) \leqslant \alpha\},$$

the intersection of the above sets for all $\mathcal{A} \in \prod$, is compact, too. When $Q \in \mathcal{P}$, this compact set K is a subset of $\mathcal{P}$ and hence equals the divergence ball (5). Indeed, if $P \in \mathcal{M}$ is not σ-additive, there exists a decreasing sequence of sets $F_n \in \mathcal{F}$ with empty intersection and $\lim_{n\to\infty} P(F_n) > 0$, while the σ-additivity of Q implies $Q(F_n) \to 0$. It follows that $D(P^{\mathcal{A}_n}||Q^{\mathcal{A}_n}) \to \infty$ for the partitions $\mathcal{A}_n = (F_n, F_n^c)$, hence $D(P||Q) = \infty$ thus $P \notin K$.

This completes the proof of the claim (ii), since the subspace topology of $\mathcal{P} \subset \mathcal{M} \subset [0, 1]^{\mathcal{F}}$ equals the τ-topology, by definition.

(iii) With the notation

$$\Gamma^{\mathcal{A}} = \{P^{\mathcal{A}} : P \in \Gamma\}, \quad \Gamma(\mathcal{A}) = \{P \in \mathcal{P} : P^{\mathcal{A}} \in \Gamma^{\mathcal{A}}\}, \tag{7}$$

it holds for each partition $\mathcal{A} \in \prod$ that

$$Q^n(\{\mathbf{x} : \hat{P}_{\mathbf{x}} \in \Gamma\}) \leqslant Q^n(\{\mathbf{x} : \hat{P}_{\mathbf{x}} \in \Gamma(\mathcal{A})\}) = Q^n(\{\mathbf{x} : \hat{P}_{\mathbf{x}}^{\mathcal{A}} \in \Gamma(\mathcal{A}) \cap \mathcal{P}_n(k)\})$$

$$\leqslant (n+1)^k \max_{\bar{P} \in \Gamma(\mathcal{A}) \cap \mathcal{P}_n(k)} (Q^{\mathcal{A}})^n(\mathcal{T}_n(\bar{P})) \leqslant (n+1)^k e^{-n \inf_{P \in \Gamma} D(P^{\mathcal{A}} || Q^{\mathcal{A}})},$$

the last two inequalities by Lemmas 1 and 2. Hence

$$\begin{aligned} \limsup_{n \to \infty} \frac{1}{n} \log Q^n(\{\mathbf{x} : \hat{P}_{\mathbf{x}} \in \Gamma\}) &\leqslant \inf_{\mathcal{A} \in \prod} [- \inf_{P \in \Gamma} D(P^{\mathcal{A}} || Q^{\mathcal{A}})] \\ &= - \sup_{\mathcal{A} \in \prod} \inf_{P \in \Gamma} D(P^{\mathcal{A}} || Q^{\mathcal{A}}). \end{aligned}$$

The key part of the proof of (4) is to show that

$$\sup_{\mathcal{A} \in \prod} \inf_{P \in \Gamma} D(P^{\mathcal{A}} || Q^{\mathcal{A}}) \geqslant \inf_{P \in cl_\tau \Gamma} D(P || Q). \tag{8}$$

Assuming w.l.o.g. that the left hand side is finite, it suffices to show that

$$cl_\tau \Gamma \cap B(Q, \alpha) \neq \emptyset \tag{9}$$

whenever

$$\alpha > \sup_{\mathcal{A} \in \prod} \inf_{P \in \Gamma} D(P^{\mathcal{A}} || Q^{\mathcal{A}}); \tag{10}$$

recall that $B(Q, \alpha)$ denotes the divergence ball (5).

To this end, we first show that

$$cl_\tau \Gamma = \bigcap_{\mathcal{A} \in \prod} cl_\tau \Gamma(\mathcal{A}), \tag{11}$$

see (7). The inclusion $\subset$ is obvious since $\Gamma \subset \Gamma(\mathcal{A})$. The reverse inclusion means that if $P \in cl_\tau \Gamma(\mathcal{A})$ for each $\mathcal{A} \in \prod$ then all τ-neighborhoods $U(P, \mathcal{A}, \epsilon)$ of P intersect Γ. To verify this, fix any $\mathcal{A} = (A_1, \ldots, A_k) \in \prod$, and pick a PM P' in $U(P, \mathcal{A}, \epsilon) \cap \Gamma(\mathcal{A})$ which is nonempty due to $P \in cl_\tau \Gamma(\mathcal{A})$. Then $P' \in \Gamma(\mathcal{A})$ implies by (7) that $P'(A_i) = \tilde{P}(A_i),\ i = 1, \ldots, k$, for some $\tilde{P} \in \Gamma$, moreover, this and $P' \in U(P, \mathcal{A}, \epsilon)$ imply by (1) that $\tilde{P} \in U(P, \mathcal{A}, \epsilon)$, thus $U(P, \mathcal{A}, \epsilon)$ intersects Γ as claimed.

Next we show that, for each $\mathcal{A} = (A_1, \ldots, A_k) \in \prod$,

$$\Gamma(\mathcal{A}) \cap B(Q, \alpha) \neq \emptyset. \tag{12}$$

On account of (10), there exists $\tilde{P} \in \Gamma$ such that $D(\tilde{P}^{\mathcal{A}}||Q^{\mathcal{A}}) < \alpha$. From such a $\tilde{P}$ construct a PM $P \in \Gamma(\mathcal{A})$ via

$$P(F) = \sum_{i=1}^{k} \frac{\tilde{P}(A_i)}{Q(A_i)} Q(F \cap A_i), \quad F \in \mathcal{F}; \tag{13}$$

if $Q(A_i) = 0$ for some i (when also $\tilde{P}(A_i) = 0$ as $D(\tilde{P}^{\mathcal{A}}||Q^{\mathcal{A}})$ is finite) the corresponding term in (13) is set equal to 0. This P belongs to $\Gamma(\mathcal{A})$ due to $P^{\mathcal{A}} = \tilde{P}^{\mathcal{A}}$, and the claim (13) follows, as P belongs also to $B(Q, \alpha)$, due to $D(P||Q) = D(\tilde{P}^{\mathcal{A}}||Q^{\mathcal{A}})$. The last equality, obvious from the integral representation of $D(P||Q)$, easily follows also directly from the definition (2) of $D(P||Q)$, because (13) implies $D(P^{\mathcal{B}}||Q^{\mathcal{B}}) = D(\tilde{P}^{\mathcal{A}}||Q^{\mathcal{A}})$ for each partition $\mathcal{B} \in \prod$ that refines $\mathcal{A}$.

For any finite collection of partitions $\mathcal{A}_i \in \prod$, $i = 1, \ldots, m$, and $\mathcal{A} \in \prod$ refining each $\mathcal{A}_i$, clearly each $\Gamma(\mathcal{A}_i)$ contains $\Gamma(\mathcal{A})$. Hence (12) implies

$$\bigcap_{i=1}^{m} (\Gamma(\mathcal{A}_i) \cap B(Q, \alpha)) \neq \emptyset. \tag{14}$$

Finally, the sets $cl_\tau \Gamma(\mathcal{A}) \cap B(Q, \alpha)$, $\mathcal{A} \in \prod$ are compact in τ-topology due to the compactness of $B(Q, \alpha)$, and any finite collection of them has nonempty intersection by (14). It follows that the intersection of all these sets is also nonempty. This and (11) complete the proof of (9), and thereby of (4).

References

[1] I. Csiszár, Sanov property, generalized I-projctions and a conditional limit theorem. *Ann. Probab.* **12** (1984), 768–793.

[2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Akadémiai Kiadó, Budapest and Academic Press, Orlando, (1981).

[3] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications, Second Edition.* Springer, (1998).

[4] P. Groeneboom, J. Oosterhoff and F.H. Ruymgaart, Large deviation theorems for empirical probability measures. *Ann. Probab.* **7** (1979), 553–586.

[5] I.N. Sanov, On the probability of large deviations of random variables. *Selected Translations in Mathematical Statistics and Probability* **1** (1961), 213–224. (Russian original: *Mat. Sb.* 1957.)

Imre Csiszár
Institute of Mathematics
Hungarian Academy of Sciences
H-1364 Budapest, P.O. Box 127
HUNGARY

E-mail: csiszar@renyi.hu