Tatra Mt. Math. Publ. **39** (2008), 275–282

TATRA MOUNTAINS
Mathematical Publications

TRINITY OF CONDITIONAL LIMIT THEOREMS

Marian Grendár

ABSTRACT. Conditional Limit Theorem (CoLT) for Empirical Measures is a direct consequence of Sanov's Theorem. This note also discusses its counterpart, Conditional Limit Theorem for Sources (Data-sampling Distributions). The third CoLT concerns asymptotic conditional *joint* behavior of empirical measures and sources. Implications of the Theorems for associated ill-posed inverse problems are mentioned, as well.

1. Introduction

A threesome of Conditional Limit Theorems (CoLT's) is gathered here. CoLT for Empirical Measures is well-known in Shannon's Theory community, but not much outside. CoLT for Sources is a rather recent result. The third one, Joint CoLT, is new. Each of the Limit Theorems has a bearing for associated ill-posed inverse problem.

2. Conditional limit theorems

2.1. CoLT for types

In order to get into the subject, basic terminology and notation should be introduced.

Let there be a random variable X with probability mass function (pmf) r, and take values from a finite set $\mathcal{X} \triangleq \{x_1, x_2, \dots, x_m\}$, called alphabet, of m letters. Let $\mathcal{P}(\mathcal{X})$ be a set of all pmf's on $\mathcal{X}$. Let $\Pi \subseteq \mathcal{P}(\mathcal{X})$.

Let type, or n-type, be $\nu^n \triangleq [n_1, n_2, \dots, n_m]/n$, where n_i is the number of occurrences of the ith outcome in a random sample $X^n \triangleq X_1, X_2, \dots, X_n$ of size n.

2000 Mathematics Subject Classification: Primary 60F10; Secondary 60F15.
Keywords: information projection, Relative Entropy Maximization method, Bayesian non-parametric consistency, L-divergence, Maximum Non-parametric Likelihood method.
Supported by the VEGA grant No. 1/3016/06 and Australian Research Council grant no. DP0210999.

Thus, type is just another (and more apt) name for empirical measure induced by an iid sample of the length n. There are $\Gamma(\nu^n) \triangleq \frac{n!}{\prod_{i=1}^m n_i!}$ sequences that induce the same type. Let the sample be drawn from the source (data-sampling distribution) r. Probability $\pi(\nu^n; r)$ that the source r generates an n-type ν^n is just the standard multinomial probability: $\pi(\nu^n; r) \triangleq \Gamma(\nu^n) \exp(n \sum_{i=1}^{m} \nu_i^n \log q_i)$. Herafter, log stands for the natural logarithm. The key object of interest is the conditional probability $\pi(\nu^n \in A \,|\, \nu^n \in B; r)$ that there occurred a type in set A provided that a type from B has occurred. The probability in question is $\pi(\nu^n \in A \,|\, \nu^n \in B; r) = \frac{\pi(\nu^n \in A; r)}{\pi(\nu^n \in B; r)}$; provided that $\pi(\nu^n \in B; r) \neq 0$; for $A \subseteq B \subseteq \mathcal{P}(\mathcal{X})$. CoLT concerns asymptotic behavior of that probability. The information divergence (Kullback-Leibler "distance") of p with respect to q (both from $\mathcal{P}(\mathcal{X})$) is defined as $I(p||q) \triangleq \sum_{\mathcal{X}} p \log \frac{p}{q}$, with conventions that $0 \log 0 = 0$, $\log b/0 = +\infty$. The information projection $\hat{p}$ of q on Π is $\hat{p} \triangleq \arg\inf_{p\in\Pi} I(p||q)$. Finally, $I(\Pi||q)$ is the value of the I-divergence at an I-projection of q on Π. The support of $p \in \mathcal{P}(\mathcal{X})$ is a set $S(p) \triangleq \{x : p(x) > 0\}$. Topology induced by the standard topology on $\mathbb{R}^m$ is assumed on $\mathcal{P}(\mathcal{X})$.

Given that the source r produced an n-type from Π, it is of interest to know how the conditional probability/measure spreads among the n-types from Π; especially as n grows beyond any limit. For the set of a particular form, this question is answered by Conditional Limit Theorem for Types (ICoLT) which is also known as Conditional Weak Law of Large Numbers.

ICoLT can be established by means of Sanov's Theorem (ST).

ST. [7] *Let Π be an open set. Let r be such that $S(r) = \mathcal{X}$. Then,*

$$\lim_{n\to\infty} \frac{1}{n} \log \pi(\nu^n \in \Pi; r) = -I(\Pi||r).$$

Sanov's Theorem states that the probability $\pi(\nu^n \in \Pi; r)$ decays exponentially fast, with the decay rate given by the value of the information divergence at an I-projection of the source r on Π. For the proof see [7].

***I*CoLT.** [6] *Let Π be a convex, closed set. Let $B(\hat{p}, \epsilon)$ be a closed ϵ-ball defined by the total variation metric, centered at I-projection $\hat{p}$ of r on Π. Then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \pi\big(\nu^n \in B(\hat{p}, \epsilon) \,|\, \nu^n \in \Pi; r\big) = 1.$$

Informally, ICoLT states that if a dense rare set admits a unique I-projection, then asymptotical types conditionally concentrate just on it.

2.2. CoLT for sources

Let $\mathcal{Q} \subset \mathcal{P}(\mathcal{X})$ be a countably infinite set of sources. Let a Bayesian put his strictly positive prior probability mass function $\pi(\cdot)$ on $\mathcal{Q} \subseteq \mathcal{P}(\mathcal{X})$. Provided that $r \in \mathcal{Q}$, as the sample size n grows to infinity, the posterior distribution $\pi(\cdot|X^n = x^n; r)$ over $\mathcal{Q}$ is expected to concentrate in a neighborhood of the true source r. Whether and under what conditions this indeed happens is a subject of Bayesian nonparametric consistency investigations.

In what follows, r is not necessarily from $\mathcal{Q}$. The problem is to find the source(s) upon which the posterior concentrates.

The L-divergence $L(q\,||\,p)$ of $q \in \mathcal{P}(\mathcal{X})$ with respect to $p \in \mathcal{P}(\mathcal{X})$ is defined as $L(q\,||\,p) \triangleq -\sum_{i=1}^{m} p_i \log q_i$. The L-projection $\hat{q}$ of p on $\mathcal{Q}$ is $\hat{q} \triangleq \arg\inf_{q\in\mathcal{Q}} L(q||p)$. The value of L-divergence at an L-projection of p on $\mathcal{Q}$ is denoted by $L(\mathcal{Q}||p)$.

Sanov's Theorem for Sources (LST) provides rate of the exponential decay of the posterior probability.

***L*ST.** [15] *Let $\mathcal{N} \subset \mathcal{Q}$. As $n \to \infty$,*

$$\frac{1}{n} \log \pi(q \in \mathcal{N}|x^n; r) \to -\big\{L(\mathcal{N}\,||r) - L(\mathcal{Q}||r)\big\},$$

with probability one.

Proof of LST [15] is based on simple bounds; as it is short, we repeat it here.

Proof. Let $l_n(q) \triangleq \exp\Big(\sum_{l=1}^{n} \log q(X_l)\Big)$, $l_n(A) \triangleq \sum_{q\in A} l_n(q)$, and $\rho_n(q) \triangleq \pi(q) l_n(q)$, $\rho_n(A) \triangleq \sum_{q\in A} \rho_n(q)$. In this notation $\pi(q \in \mathcal{N}|x^n) = \frac{\rho_n(\mathcal{N})}{\rho_n(\mathcal{Q})}$. The posterior probability is bounded above and below as follows:

$$\frac{\hat{\rho}_n(\mathcal{N})}{\hat{l}_n(\mathcal{Q})} \le \pi\big(q \in \mathcal{N}|x^n; r\big) \le \frac{\hat{l}_n(\mathcal{N})}{\hat{\rho}_n(\mathcal{Q})},$$

where $\hat{l}_n(A) \triangleq \sup_{q\in A} l_n(q)$, $\hat{\rho}_n(A) \triangleq \sup_{q\in A} \rho_n(q)$.

$\frac{1}{n}\big(\log \hat{l}_n(\mathcal{N}) - \log \hat{\rho}_n(\mathcal{Q})\big)$ converges with probability one to $L\big(\mathcal{Q}||r\big) - L\big(\mathcal{N}||r\big)$. The same is the 'point' of a.s. convergence of $\frac{1}{n}$ log of the lower bound. □

LST says that almost surely the posterior probability $\pi\big(q \in \mathcal{N}|x^n; r\big)$ decays exponentially fast, with decay rate specified by the difference of the values of the two extremal L-divergences.

Let for $\epsilon > 0$, $\mathcal{N}^C_\epsilon(\mathcal{Q}) \triangleq \big\{q : L(q||r) - L(\mathcal{Q}||r) > \epsilon, q \in \mathcal{Q}\big\}$.

Corollary. *Let there be a finite number of L-projections of r on $\mathcal{Q}$. As $n \to \infty$, $\pi\big(q \in \mathcal{N}^C_\epsilon(\mathcal{Q})|x^n; r\big) \to 0$, with probability one.*

The Corollary establishes posterior consistency in L-divergence. In words: the probability that r generates x^n such that the limit of the posterior probability $\lim_{n\to\infty} \pi\big(q \in \mathcal{N}^C_\epsilon(\mathcal{Q})|x^n; r\big) = 0$, is one.

Conditional Limit Theorem for Sources (LCoLT) is as well a direct consequence of LST.

LCoLT. *Let there be a unique L-projection $\hat{q}$ of r on $\mathcal{N}$. Let $B(\hat{q}, \epsilon)$ be an ϵ-ball defined by the total variation metric, centered at $\hat{q}$. Then, for $\epsilon > 0$,*

$$\lim_{n\to\infty} \pi\big(q \in B(\hat{q}, \epsilon) \,|\, q \in \mathcal{N}, \nu^n; r\big) = 1,$$

with probability one.

Thus, there is asymptotically conditionally (a.s) zero probability of sources other than those arbitrarily close to the L-projection $\hat{q}$ of r on $\mathcal{N}$. Conditioning is done by event of the form: r produced n-type ν^n and at the same time $q \in \mathcal{N}$ happened. Clearly, $\pi(q \in A, \nu^n; r) = \pi\big(\nu^n \,|\, q \in A\big)\pi(q \in A)$, where r (as always) is used as a reminder that the true source is r.

2.3. Joint CoLT

Consider the same, Bayesian, setting as in the previous Section 2.2.

Let $[\hat{p}, \hat{q}] \triangleq \arg\inf_{p\in\Pi, q\in\mathcal{Q}} I(p||q)$. Let $I(\Pi||\mathcal{Q})$ denote the value of the I-divergence at $[\hat{p}, \hat{q}]$.

Sanov's Theorem for pairs of types and sources

JST. *Let $\mathcal{N} \subset \mathcal{Q}$. Let $\Pi \subset \mathcal{P}(\mathcal{X})$. As $n \to \infty$,*

$$\frac{1}{n} \log \pi(q \in \mathcal{N}, \nu^n \in \Pi; r) \to -I(\Pi||\mathcal{N}),$$

with probability one.

Proof. $\pi(q \in \mathcal{N}, \nu^n \in \Pi; r) = \sum_{\nu^n\in\Pi} \sum_{q\in\mathcal{Q}} \pi(\nu^n \,|\, q)\pi(q)$. Employ the binding used at the proof of LST to bind the inner sum, accompanied by the binding used in the standard proof of the Sanov's Theorem [7] for the outer sum. □

JST directly implies the following Joint Conditional Limit Theorem (JCoLT):

JCoLT. *Let $\mathcal{N} \subset \mathcal{Q}$ admit unique $\hat{q}$ and let $\Pi \subseteq \mathcal{P}$ be a convex, closed set. Let $B(\hat{p}, \epsilon)$, $B(\hat{q}, \epsilon)$ be ϵ-balls defined by the total variation metric, centered at $\hat{p}$, $\hat{q}$, respectively. Then, for $\epsilon > 0$,*

$$\lim_{n\to\infty} \pi\big(\nu^n \in B(\hat{p}, \epsilon),\, q \in B(\hat{q}, \epsilon) \,|\, \nu^n \in \Pi,\, q \in \mathcal{N}; r\big) = 1.$$

with probability one.

3. Ill-posed inverse problems

Each of the CoLT's can be associated with a particular ill-posed inverse problem; the α, β and γ problem, respectively.

- The α-problem: given $\{\mathcal{X}, r, \Pi, n\}$ the objective is to select an n-type (one or more) from Π. If Π contains more than one n-type, the problem is under-determined, and in this sense ill-posed. ICoLT implies that (at least for sufficiently large n) the α-problem has to be solved by selecting the I-projection of r on Π, provided that Π is convex, closed. The method associated with this selection scheme is known as Relative Entropy Maximization (REM/MaxEnt).
- The β-problem: given $\left\{\mathcal{X}, \nu^n, \mathcal{N}, \pi(q)\right\}$ the objective is to select a source (one or more) from $\mathcal{N}$. LCoLT implies that for sufficiently large n the β-problem has to be solved by selecting the L-projection of ν^n on $\mathcal{Q}$. Note, that the L-projection is the Maximum a-posteriori probability (MAP) source, which is identical to the Maximum Non-parametric Likelihood (MNPL) source, since asymptotically prior does not matter. Elementary requirement of consistency implies that for finite n, MAPs have to be selected.
- The γ-problem: given $\left\{\mathcal{X}, \Pi, \mathcal{N}, n, \pi(q)\right\}$ the objective is to select a pair (one or more) $\nu^n \in \Pi$, $q \in \mathcal{N}$. JCoLT implies that for $n \to \infty$ the γ-problem has to be solved by selecting $[\hat{p}, \hat{q}]$.

The α- and β-problem are, in a sense, opposite to each other. In the α-problem, the source (data-sampling distribution) is known, and the objective is to select type(s) from given set Π, which (supposedly) characterizes studied phenomenon. On the contrary, the β-problem assumes a given type ν^n, and the objective is to select a source from given set $\mathcal{N}$, which might or might not contain the true source. Note that LCoLT makes it necessary to formulate the β-problem in the Bayesian context; i.e., a prior has to be put on the set of sources.

ICoLT provides probabilistic justification of application of REM for the α-problem, as it also does for the Maximum Probability (MaxProb) method [12] in the same context. LCoLT justifies application of MAP for the β-problem. Note that asymptotically, MAP turns into MNPL, the same way as MaxProb turns into REM.

The γ-problem merges the two problems together. It captures the situation where a non-parametric Bayesian has a set of empirical measures (instead of just one such a measure) and a set of sources, over which he puts the prior. JCoLT implies that the objective of selecting a pair of type and source, should be for $n \to \infty$ attained by selecting the joint projection $[\hat{p}, \hat{q}]$.

4. Notes on literature

For historical developments on Sanov's Theorem and ICoLT see [1]–[3], [8], [9], [17]–[19], [21], [24]–[27], [31], among many others. For an extension of ST to the continuous case cf. [17], [6], [18]. Extension of ST and ICoLT to the case of feasible set admitting non-unique I-projection was studied in [16].

For surveys on Bayesian non-parametric consistency check [11], [29] among others. See also [28], [20], [30].

An inverse of Sanov's Theorem has been established by Ganesh and O'Connell [10] for the case of sources with finite alphabet, by means of formal large-deviations approach. Unaware of their work, the present author developed in [13] Sanov's Theorem for n-sources, for both discrete and continuous alphabet and applied it to conditioning by rare sources problem and criterion choice problem. The present form of LST was established in [15], in a more general setting of continuous sources. There, also an extension of LST to the case of the set of sources admitting non-unique L-projection was presented.

For a discussion of a justification of REM via ICoLT see [4], [5].

Implications of CoLTs for empirical estimation (cf. [22], [23]) are discussed in [14].

Acknowledgements. Hospitality of the School of Computer Science and Engineering of the University of New South Wales, Sydney, where this work was completed is gratefully acknowledged. Special thanks to Arthur Ramer. It is a pleasure to thank George Judge for stimulating discussions.

REFERENCES

[1] BÁRTFAI, P.: *On a conditional limit theorem*, Progress in Statistics **1** (1974), 85–91.

[2] COVER, T.—THOMAS, J.: *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[3] CSISZÁR, I.: *Sanov property, generalized I-projection and a conditional limit theorem*, Ann. Probab. **12** (1984), 768–793.

[4] CSISZÁR, I.: *An extended Maximum Entropy principle and a Bayesian justification*. In: Bayesian Statistics 2, Proceedings of the Second Valencia International Meeting, September 6–10, 1983, Valencia, Spain, (Bernardo, J. M., DeGroot, M. H., Lindley, D. V., Smith, A. F. M., eds.), Elsevier, New York, 1985, pp. 83–98.

[5] CSISZÁR, I.: *Maxent, mathematics and information theory*. In: Maximum Entropy and Bayesian Methods, Proceedings of the 15th International Workshop, Santa Fe, NM, USA, July 31–August 4, 1995 (Hanson, K. M., Silver, R. N., eds.), Kluwer, Dordrecht, 1996, pp. 35–50.

[6] CSISZÁR, I.: *The method of types*, IEEE Trans. Inf. Theory **44** (1998), 2505–2523.

[7] CSISZÁR, I.—SHIELDS, P.: *Notes on information theory and statistics: A tutorial*, Foundations and Trends in Communications and Information Theory **1** (2004), 1–111.
[8] DEMBO, A.—ZEITOUNI, O.: *Large Deviations Techniques and Applications* (2nd ed.), Appl. Math., Vol. 38, Springer-Verlag, New York, 1998.
[9] ELLIS, R. S.: *The theory of large deviations: from Boltzmann's 1877 calculation to equilibrium macrostates in 2D turbulence*, Phys. D **1** (1999), 106–136.
[10] GANESH, A.—O'CONNELL, N.: *An inverse of Sanov's Theorem*, Statist. Probab. Lett. **42** (1999), 201–206.
[11] GHOSAL, A.—GHOSH, J. K.—RAMAMOORTHI, R. V.: *Consistency issues in Bayesian nonparametrics.* Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri, Marcel Dekker, Stat., Textb. Monogr. **158** (1999), 639–667.
[12] GRENDÁR, M. JR.—GRENDÁR, M.: *What is the question that MaxEnt answers: a probabilistic interpretation.* In: Bayesian Inference and Maximum Entropy Methods in Science and Engineering (Mohammad-Djafari, A., ed.), AIP, Melville, 2001, pp. 83–94. (also, arXiv:math-ph/0009020).
[13] GRENDÁR, M.: *Conditioning by rare sources*, Acta Univ. M. Belii Ser. Math. **12** (2005), 19–29.
[14] GRENDÁR, M.—JUDGE, G.: *Large deviations theory and empirical estimator choice*, Econometric Rev. **27** (2008), 513–525.
[15] GRENDÁR, M.: *L-divergence consistency for a discrete prior*, J. Statist. Res. **40** (2006), 73–76.
[16] GRENDÁR, M.: *Conditional equiconcentration of types.* In: Focus on Probability Theory (L. R. Velle, ed.), NSP, New York, 2006, pp. 73–89.
[17] GROENEBOOM, P.—OOSTERHOFF, J.—RUYMGAART, F. H.: *Large deviation theorems for empirical probability measures*, Ann. Probab. **7** (1979), 553–586.
[18] HARREMOËS, P.: *Information topologies with applications*, Bolyai Soc. Math. Stud., Vol. 16, Springer, New York, 2007, pp. 113–150.
[19] LEONARD, CH.—NAJIM, J.: *An extension of Sanov's theorem: Application to the Gibbs conditioning principle*, Bernoulli **8** (2002), 721–743.
[20] KLEIJN, B. J. K.—VAN DER VAART, A. W.: *Misspecification in infinite-dimensional Bayesian statistics*, Ann. Statist. **34** (2006), 837–877.
[21] LEWIS, J. T.—PFISTER, C.-E.—SULLIVAN, W. G.: *Entropy, concentration of probability and conditional theorems*, Markov Process. Related Fields **1** (1985), 319–386.
[22] MITTELHAMMER, R. C.—JUDGE, G. G.—MILLER D. J.: *Econometric Foundations*, CUP, Cambridge, 2000.
[23] OWEN, A. B.: *Empirical Likelihood*, Chapman-Hall/CRC, New York, 2001.
[24] SANOV, I. N.: *On the probability of large deviations of random variables*, Mat. Sbornik **42** (1957), 11–44.
[25] VAN CAMPENHOUT, J. M.—COVER, T. M.: *Maximum entropy and conditional probability*, IEEE Trans. Inf. Theory **27** (1981), 483–489.
[26] VASICEK, O. A.: *A conditional law of large numbers*, Ann. Probab. **8** (1980), 142–147.
[27] VINCZE, I.: *On the maximum probability principle in statistical physics.* In: Progress in statistics, Vol. I, II Coll. Math. Soc. J. Bolyai **9** (1972), pp. 869–893.
[28] WALKER, S.: *New appraoches to Bayesian consistency*, Ann. Statist. **32** (2004), 2028–2043.
[29] WALKER, S.—LIJOI, A.— PRÜNSTER, I.: *Contributions to the understanding of Bayesian consistency*, ICER Working papers–Applied Mathematics Series **13** (2004).

[30] WATANABE, S.: *Information-theoretic aspects of inductive and deductive inference*, IBM J. Res. Dev. **4** (1960), 208–231.

[31] ZABEL, S.: *Rates of convergence for conditional expectations* Ann. Probab. **8** (1980), 928–941.

Received September 27, 2006

Department of Mathematics
FPV UMB
Tajovského 40
974 01 Banská Bystrica
SLOVAKIA

Institute of Measurement Science
Slovak Academy of Sciences (SAS)
Dubravská cesta 9
841 04 Bratislava
SLOVAKIA

Institute of Mathematics and
Computer Science
Severná 5
974 01 Banská Bystrica
SLOVAKIA
E-mail: marian.grendar@savba.sk