Proceedings of the

SEVENTH WORKSHOP ON INFORMATION THEORETIC METHODS IN SCIENCE AND ENGINEERING

Edited by
Jorma Rissanen, Petri Myllymäki, Teemu Roos,
& Narayana Prasad Santhanam

UNIVERSITY OF HELSINKI

University of Helsinki
Department of Computer Science
Series of Publications B
Report B-2014-4

PREFACE

The Seventh Workshop on Information Theoretic Methods in Science and Engineering (WITMSE 2014) took place on July 5–8, 2014, in Honolulu (HA), USA. The workshop was organized jointly by the University of Helsinki, the Helsinki Institute for Information Technology HIIT, and the University of Hawaii. This was the seventh workshop in the series which started in 2008. The first one, as well as the following two organized in 2009 and 2010, respectively, took place in Tampere, Finland. The following workshops, in 2011 through 2013, were held in Helsinki, Amsterdam, and Tokyo, respectively. In 2014, for the first time, the workshop was co-located with the annual IEEE Information Theory Symposium which took place during the week prior to WITMSE in Honolulu.

As the title of the workshop suggests, WITMSE seeks speakers from a variety of disciplines with emphasis on both theory and applications of information and coding theory with special interest in modeling. Since the beginning our plan has been, and still is, to keep the number of the participants small and to ensure the highest possible quality, which has been accomplished by inviting distinguished scholars as speakers.

The workshop was opened by Jorma Rissanen's keynote talk on "Entropy and Estimation of Random Maximum Likelihood Models". Plenary talks were given by Venkat Anantharam (UC Berkeley) "Entropy Power Inequalities: Results and Speculation" and Wojciech Szpankowski (Purdue) "Structural Information". The rest of the technical programme consisted of nineteen talks on various aspects of information theory and its applications.

Outside the technical sessions the program included a welcoming reception next to the wonderful beaches of Waikiki and a banquet dinner.

We would like to thank all the participants to our workshop. Some of the speakers kindly submitted written contributions or abstracts to these proceedings, for which we are particularly grateful. We also want to thank the NSF Science and Technology Center for Science of Information for sponsoring the workshop.

December 30, 2014
Helsinki, San Jose, and Honolulu
Workshop Co-Chairs
Jorma Rissanen,
Petri Myllymäki,
Teemu Roos,
& Narayanan P. Santhanam

Contents

ENTROPY POWER INEQUALITIES: RESULTS AND SPECULATION

Venkat Anantharam

University of California, Berkeley

ABSTRACT

Shannon's entropy power inequality characterizes the minimum differential entropy achievable by the sum of two independent random variables with fixed differential entropies. Since the pioneering work of Shannon, there has been a steady stream of results over the years, trying to understand the structure of Shannon's entropy power inequality, as well as trying to develop similar entropy power inequalities in other scenarios, such as for discrete random variables. We will discuss some aspects of this landscape in this talk. We will present old results, new results, and share some speculation about how to prove new kinds of entropy power inequalities.

STRUCTURAL INFORMATION

Wojciech Szpankowski

Purdue University

ABSTRACT

F. Brooks argued in his 2003 JACM paper on the challenges of computer sciences that there is "no theory that gives us a metric for the information embodied in structure". C. Shannon himself noticed this fifty years earlier in his 1953 paper. More generally, we lack an information theory of data structures (e.g., graphs, sets, social networks, chemical structures, biological networks). In this talk, we present some recent research results on structural information. We first propose some fundamental limits of information content for a wide range of data structures with correlated labels and then propose asymptotically optimal lossless compression algorithms achieving these limits for unlabeled graphs. Then we move to Markov fields and try to understand structural properties of large systems with local mutual dependencies and interaction. In particular, we focus on enumerating Markov field and universal types. Finally, we study capacity of a sequence to structure channel arising in protein folding applications. The channel itself is characterized by the Boltzmann distribution with a free parameter corresponding to temperature. Interestingly, capacity of such a channel exhibits an unusual phase transition with respect to temperature. We tackle most of these problems by complex analysis methods, thus within the realm of analytic information theory.

ON THE APPLICABILITY OF WAIC AND WBIC IN THE GAUSSIAN PROCESS FRAMEWORK

Tommi Mononen

Department of Information and Computer Science & O.V. Lounasmaa Laboratory,
Aalto University School of Science, P.O. Box 15400, FI-00076 Aalto, FINLAND,
tommi.j.mononen@aalto.fi

ABSTRACT

Many Gaussian process models are not analytically tractable but the computations have to be carried out using the slow sampling approach or using some approximative method (e.g. expectation propagation or Laplace approximation). In order to make reliable computation faster, we investigate the applicability of the singular information criteria WAIC and WBIC in the Gaussian process framework. Although the theoretical basis behind these criteria is parametric, the formulas themselves seem to be intuitively reasonable even in the non-parametric setting. The predictive model selection criterion WAIC approximates leave-one-out score and this kind of well performing method would be highly desirable from the Gaussian process viewpoint. The spin-offs of the log marginal likelihood approximation WBIC, could make sampling based methods faster. In this talk, we highlight and explain the cases where these criteria fail to perform well.

Part of this talk is based on [1].

1. REFERENCES

[1] T. Mononen, "A case study of the widely applicable bayesian information criterion and its optimality," *Statistics and Computing*, Apr. 2014, DOI: 10.1007/s11222-014-9463-3.

LOCAL SCORING RULES AND STATISTICAL INFERENCE IN UNNORMALIZED MODELS

Matthew Parry[1]

[1]Dept of Mathematics & Statistics, University of Otago,
P.O.Box 56, Dunedin 9054, NEW ZEALAND, mparry@maths.otago.ac.nz

ABSTRACT

A scoring rule is a principled way to assess probabilistic forecasts. Associated with every scoring rule is a divergence and a concave entropy. Conversely, every scoring rule can be generated by a concave entropy. Scoring rules can also be straightforwardly adapted to statistical inference. The resulting estimating equations are unbiased but typically entail some loss of efficiency. Local scoring rules are a class of scoring rules with the remarkable property that they do not depend on the normalization of the quoted probability distribution. Consequently, they allow inference in unnormalized statistical models, i.e. models in which the normalization is either difficult or impossible to compute. Local scoring rules provide a unifying framework in which to understand existing approaches to such intractable problems, for example pseudolikelihood, score matching and ratio matching.

1. INTRODUCTION

Scoring rules have long been used to evaluate probabilistic forecasts [1, 2]. If Q stands for the forecaster's distribution – Q is used to denote *quote* – then $S(x, Q)$ is the score given to the forecaster when outcome x is observed. Scoring rules fit into standard statistical decision theory: a scoring rule is a loss function where the action is to quote a probability distribution [3].

A key feature of scoring rules is that they can be crafted to elicit a forecaster's honestly held belief. Such a scoring rule is said to be *proper*. We can usefully overload the definition of a scoring rule by defining a forecaster's *expected score* when $X \sim P$ as $S(P, Q)$. A proper scoring rule has the property that $S(P, Q) \geq S(P, P)$ for $Q \neq P$. In other words, if a forecaster thinks $X \sim P$, they will minimize their expected score by quoting P.

Scoring rules connect naturally to information theory. Every proper scoring rules gives rise to a *divergence*

$$d(P, Q) := S(P, Q) - H(P), \tag{1}$$

where $H(P) := S(P, P)$ is the (concave) *entropy* associated with the scoring rule. As we will see, we can profitably reverse this connection and use a concave entropy function to generate a scoring rule.

2. STATISTICAL INFERENCE

Although scoring rules are formulated to evaluate predictions, they can be easily turned to the task of estimation. Given a parametric model Q_θ, we have the obvious estimator

$$\widehat{\theta}(x) = \arg\min_\theta S(x, Q_\theta). \tag{2}$$

Typically, this amounts to solving the estimating equation $\partial S(x, Q_\theta)/\partial\theta = 0$. When the scoring rule is proper, the estimating equation is unbiased [4] since at $\theta = \theta_0$, $0 = \partial S(Q_{\theta_0}, Q_\theta)/\partial\theta = \mathbb{E}_{\theta_0}\partial S(X, Q_\theta)/\partial\theta$. As a result, inference via scoring rules fits into the established theory of unbiased estimating equations. An important consequence is that typically there will be a loss of efficiency. If we define $D := \mathbb{E}_\theta\, \partial^2 S/\partial\theta^2$ and $J := \mathbb{E}_\theta\, (\partial S/\partial\theta)^2$, then the *Godambe* or *sandwich* information G cannot exceed the Fisher information F [5]:

$$G := DJ^{-1}D \leq F. \tag{3}$$

3. LOCAL SCORING RULES

The *logarithmic scoring rule* or *log score* is the simplest example of a scoring rule:

$$S(x, Q) = -\log q(x). \tag{4}$$

It is straightforward to see that using this for statistical inference amounts to maximum likelihood estimation. Furthermore, the divergence and entropy associated with the log score are simply the Kullback-Leibler divergence and Shannon entropy, respectively. It is also possible to show that the log score is the only scoring rule that depends on the value of the quoted probability distribution at the observed outcome x and no other (counterfactual) outcome.

For this reason, we call the log score *strongly local*. The main idea of this paper is the concept of a *local* scoring rule. A local scoring rule is a rule that depends on the quoted distribution at x and on a "neighbourhood" of x. It turns out that local scoring rules are of the form [6]

$$S(x, Q) = -\lambda \log q(x) + S_0(x, Q), \tag{5}$$

where $\lambda \geq 0$ and $S_0(x, Q)$ is a 0-homogeneous function of Q. Consequently, when $\lambda = 0$, we obtain a scoring rule that does not depend on the normalization of the quoted probability distribution.

4. ENTROPY AND SCORING RULES

Under mild regularity conditions, it can be shown [2, 7, 8] that $S(x, Q)$ is a scoring rule if and only if there exists a concave function $H(Q)$ such that

$$S(x, Q) = H(Q) + H^\star(x, Q) - H^\star(Q, Q), \quad (6)$$

where $H^\star(\cdot, Q)$ is a *subgradient* of H at x. Furthermore, $H(Q)$ is the entropy associated with the scoring rule. In practice, $H^\star(\cdot, Q)$ is often the gradient, in which case the scoring rule is uniquely defined by $H(Q)$.

The idea of locality amounts to requiring $H(Q) - H^\star(Q, Q) = \lambda$, where λ is a Q-independent constant. This essentially requires $H(Q)$ to be of the form $H(Q) = -\lambda \, q(x) \log q(x) + H_1(Q)$, where $H_1(Q)$ is a 1-homogeneous function of Q. Eq. (5) then follows, with $\lambda \geq 0$ required for propriety.

4.1. Continuous outcome spaces

Let $q(x)$ be a sufficiently differentiable and strictly positive probability density on a continuous outcome space and let ϕ be a 1-homogeneous concave function of $\{q(x), q'(x), \ldots, q^{(k)}(x)\}$ for all x. Then

$$H(Q) = \int \mathrm{d}x \, \phi\left(x, q(x), q'(x), \ldots, q^{(k)}(x)\right) \quad (7)$$

is 1-homogeneous and generates a local scoring rule of order $2k$ in the derivatives of $q(x)$. In this case, the neighbourhood of x is an infinitesimal neighbourhood about x.

We expect only rules of order 2 and 4 to be of practical use. Second order rules take the form [6, 9]

$$S(x, Q) = \left(-\frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial}{\partial q'} + \frac{\partial}{\partial q}\right)\phi[q], \quad (8)$$

where $\phi[q] = \phi(x, q, q')$. The simplest case, which occurs when $\phi[q] = -\frac{1}{2}\frac{q'^2}{q}$, was discovered by Almeida & Gidas [10] and by Hyvärinen [11], who dubbed it *score matching*:

$$S(x, Q) = \frac{q''(x)}{q(x)} - \frac{1}{2}\left(\frac{q'(x)}{q(x)}\right)^2. \quad (9)$$

4.2. Discrete outcome spaces

On a discrete outcome space, the neighbourhood is defined by an undirected graph G, i.e. y is in the neighbourhood of x if y is in the connection set of x. This relationship is also symmetric. Local scoring rules are then generated by the entropies of the form [12]

$$H(Q) = \sum_{K \in \mathcal{M}} \phi^K(Q_K), \quad (10)$$

where $\mathcal{M}$ is the set of maximal cliques, ϕ^K is 1-homogeneous and concave, and Q_K is the quoted distribution restricted to clique K. Specifically,

$$S(x, Q) = \sum_{K \in \mathcal{M}_x} \frac{\partial}{\partial q_x}\phi^K(Q_K), \quad (11)$$

where $\mathcal{M}_x$ are the maximal cliques containing x, and $q_x := q(x)$. Note that decomposition of the entropy into functions over cliques is directly analogous to the Hammersley-Clifford theorem for the factorization of a joint probabilty distribution on a graph [13].

5. EXAMPLES

5.1. Pseudolikelihood

Suppose $x = (x^1, \ldots, x^N)$ is an outcome from a product space. Let $x^{\setminus i} := \{x^k | k \neq i\}$. Then the *pseudolikelihood* [14] is

$$\mathrm{PL}(P; X = x) := \prod_i P(X^i = x^i | X^{\setminus i} = x^{\setminus i}). \quad (12)$$

Defining $y \in \mathrm{nhd}(x)$ if and only if $y^{\setminus i} = x^{\setminus i}$ for some i, then for each i and $y^{\setminus i}$, $K_{i,y^{\setminus i}} = \{x | x^{\setminus i} = y^{\setminus i}\}$ is a clique. Simplifying our notation so that $\phi^K(Q_K) = \phi_i(Q_K)$,

$$S(x, Q) = \sum_i S_i\left(x^i, Q(\cdot | X^{\setminus i} = x^{\setminus i})\right) \quad (13)$$

is a scoring rule, where the S_i are individual scoring rules for a single variable. Using the log score for each S_i justifies the use of the pseudolikelihood for inference:

$$S_{\mathrm{PL}}(x, Q) := -\ln \mathrm{PL}(Q; X = x) = \sum_i \ln \frac{q_{x^{\setminus i}}}{q_x}. \quad (14)$$

In the case of binary data outcomes, using the *Brier score*, $S(x, Q) = (x - Q(X = 1))^2$, for each S_i gives Hyvärinen's *ratio matching* method [15]. See [16] for a spatial modelling application.

5.2. Overdispersion

Overdispersion is the observation that statistical models do not always capture the amount of variation seen in the data. In many cases, overdispersion is due to the fact that there are unknown or unrecorded predictor variables. More subtly, overdispersion may indicate a breakdown of the assumption of independent observations. A phenomenological solution that is sometimes appropriate is to introduce a *dispersion parameter* ϕ to quantify the "anomalous" variation, namely $\mathrm{var}\, Y \to \phi \, \mathrm{var}\, Y$, where we expect $\phi > 1$.

The estimating equation for ϕ is often ad hoc. An obvious approach, however, is to suppose the effective number of observations is n/ϕ, leading to the updated probability model:

$$q(y) \to q(y|\phi) = \frac{q(y)^{1/\phi}}{Z(\phi)}. \quad (15)$$

Since the normalization $Z(\phi)$ will typically be impossible to compute, the usual methods of estimation will come up short. The local scoring rule in eq. (9), however, gives a delightfully simple expression:

$$\widehat{\phi} = -\frac{\ell(y)^2}{\ell'(y)}, \quad (16)$$

where $\ell(y) := q'(y)/q(y)$.

5.3. Sequential prediction

We end with a speculative application of local scoring rules. Suppose we observe $(x_1, \ldots, x_n)$ iid outcomes and wish to make a probabilistic prediction for x_{n+1}. Given a parametric model Q_θ, if $\widehat{\theta}_n$ is a consistent estimator for θ, then we would quote $Q_n := Q_{\widehat{\theta}_n}$ for x_{n+1}. When we are in the model, the desired result is $\mathbb{E}_\theta \, \mathrm{KL}(P, Q_n) \to \frac{1}{2} n^{-1}$, where KL is the Kullback-Leibler divergence.

Normalized maximum likelihood gives the optimal sequential prediction [17]:

$$q(x_{n+1}) = \frac{q(x_{n+1}|\widehat{\theta}_{n+1}(x_{1:n}, x_{n+1}))}{\sum_y q(y|\widehat{\theta}_{n+1}(x_{1:n}, y))}, \qquad (17)$$

where $\widehat{\theta}_{n+1}$ is the maximum likelihood estimate. Unfortunately, however, the denominator is often infinite. One might hope that there exists an appropriate local scoring rule that gives rise to a different estimator $\widehat{\theta}$ and a divergence which does not depend on normalization of the prediction. Admittedly, the loss of efficiency detailed in eq. (3) means we can expect only kn^{-1} convergence with $k \geq \frac{1}{2}$, but this may be an acceptable price to pay for tractability.

6. CONCLUSION

In addition to evaluating predictions, scoring rules provide a useful approach to statistical estimation. The requirement that a scoring rule be local gives rise to a surprising class of scoring rules that do not depend on the normalization of the quoted probability distribution. Consequently, inference can be carried out in models for which the normalization is either difficult or impossible to compute. Local scoring rules also appear to provide a unifying framework in which to understand existing approaches to such intractable problems. It would be interesting to see whether local scoring rules could be adapted to so called doubly intractable problems in Bayesian inference.

7. ACKNOWLEDGMENTS

It is a real pleasure to thank the organizers for their kind invitation to speak at WITMSE 2014. The work discussed here is joint work with Philip Dawid and Steffen Lauritzen. I acknowledge the financial support of a University of Otago travel grant.

8. REFERENCES

[1] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, pp. 1–3, 1950.

[2] J. McCarthy, "Measures of the value of information," *Proc. Nat. Acad. Sci.*, vol. 42, pp. 654–655, 1956.

[3] Peter D. Grünwald and Alexander Philip Dawid, "Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory," *Annals of Statistics*, vol. 32, pp. 1367–1433, 2004.

[4] A. Philip Dawid and Steffen L. Lauritzen, "The geometry of decision theory," in *Proceedings of the Second International Symposium on Information Geometry and its Applications*. University of Tokyo, 2005, pp. 22–28.

[5] V. P. Godambe, "An optimum property of regular maximum likelihood estimation," *Ann. Math. Statist.*, , no. 4, pp. 1208–1211, 1960.

[6] M. Parry, A. P. Dawid, and S. Lauritzen, "Proper local scoring rules," *Annals of Statistics*, vol. 40, pp. 561–592, 2012.

[7] A. D. Hendrickson and R. J. Buehler, "Proper scores for probability forecasters," *Ann. Math. Statist.*, vol. 42, pp. 1916–1921, 1971.

[8] Tilmann Gneiting and Adrian E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, pp. 359–378, 2007.

[9] Werner Ehm and Tilmann Gneiting, "Local proper scoring rules of order two," *Annals of Statistics*, vol. 40, pp. 609–637, 2012.

[10] M. P. Almeida and B. Gidas, "A variational method for estimating the parameters of mrf from complete or incomplete data," *The Annals of Applied Probability*, vol. 3, pp. 103–136, 1993.

[11] Aapo Hyvärinen, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning*, vol. 6, pp. 695–709, 2005.

[12] A. P. Dawid, S. Lauritzen, and M. Parry, "Proper local scoring rules on discrete sample spaces," *Annals of Statistics*, vol. 40, pp. 593–608, 2012.

[13] G. R. Grimmett, "A theorem about random fields," *Bull. Lond. Math. Soc.*, pp. 81–84, 1973.

[14] J. Besag, "Statistical analysis of non-lattice data," *J. Roy. Statist. Soc. Ser. D*, vol. 24, pp. 179–195, 1975.

[15] Aapo Hyvärinen, "Some extensions of score matching," *Computational Statistics and Data Analysis*, vol. 51, pp. 2499–2512, 2007.

[16] A. P. Dawid and M. Musio, "Estimation of spatial processes using local scoring rules," *Advances in Statistical Analysis*, pp. 173–179, 2013.

[17] Peter Grünwald, "A tutorial introduction to the minimum description length principle," in *Advances in Minimum Description Length: Theory and Applications*. 2005, MIT Press.

LOSSY IS LAZY

Susanne Still

Information and Computer Sciences, University of Hawaii at Mānoa,
1680 East-West Road, Honolulu, HI 96822, USA, sstill@hawaii.edu

ABSTRACT

Shannon's rate-distortion curve characterizes optimal lossy compression. I show here that the optimization principle that has to be solved to compute the rate-distortion function can be derived from a least effort principle: minimizing required thermodynamic effort necessitates the minimization of information (compatible with a given fidelity). Retaining less information costs less physical effort. In that sense, lossy compression is energy efficient, in other words, lossy is lazy.

1. INTRODUCTION

Rate distortion theory [1, 2, 3, 4] underlies much practical work in signal processing. It quantifies the rate at which data can be transmitted, given a tolerable level of fidelity. Shannon considered [1] "the set of messages of a long duration, say T seconds. The source is described by giving the probability density, in the associated space, that the source will select the message in question $[p(x)]$. A given communication system is described (from the external point of view) by giving the conditional probability $[p(y|x)]$ that if the message x is produced by the source the recovered message at the receiving point will be y."

Input messages, or input data, x, are then compressed into a representation, y, such that a certain desired level of fidelity is achieved, rather than perfect reconstruction. In other words, information is lost. In this process, some average distortion, $D[X,Y] := \langle d(x,y)\rangle_{p(x,y)}$ is encountered. The most efficient encoding compatible with a given quality of reproduction minimizes the mutual information[1] $I[X,Y] := \left\langle \ln\left[\frac{p(x,y)}{p(x)p(y)}\right]\right\rangle_{p(x,y)}$ under the constraint of fixed average distortion, D.

Shannon thus defined the rate, $R(D)$, of generating information compatible with a given distortion as the minimum of the mutual information under this constraint: [2]

$$R(D) := \min_{p(y|x)} I[X,Y] \quad \text{s.t. } D[X,Y] = D. \tag{1}$$

The minimum is taken over all possible communication systems, i.e. probabilistic assignments $p(y|x)$. The optimal rate is achievable, and algorithms exist for computing the rate-distortion function [3, 4].

This problem has a simple physical motivation, which I will now develop.

2. EFFORT OF CODING

Output messages are distributed according to $p(y)$.[3] However, when a specific input message, x, is given, then the corresponding code messages are distributed according to $p(y|x)$. Imagine a physical system which is changed from a state described by the distribution $p(y)$ to one described by $p(y|x)$. This change requires effort. How much effort?

The second law of thermodynamics tells us that we need to put in at least as much work as the resulting free energy difference, which is, on average over input x,

$$\Delta F[X,Y] := \langle F[p(y|x)]\rangle_{p(x)} - F[p(y)]\,, \tag{2}$$

where $F[p]$ denotes the *generalized*, or *nonequilibrium* free energy, $F[p] = \langle E\rangle_p + k_B T\langle \ln[p]\rangle_p$,[4] which has been used by a number of authors to describe nonequilibrium systems (see e.g. [5, 6, 7, 8, 9, 10, 11, 12, 13, 14], and references therein). It reduces to the thermodynamic equilibrium free energy when evaluated on the equilibrium distribution.

3. LEAST EFFORT PRINCIPLE

Typically, a representation of a quantity of interest is produced for some purpose, e.g. communication and reproduction of the original source data [1, 2, 4], or work extraction from a physical system [15, 16]. Let us define the function $u(x,y)$ to measure the general usefulness of a specific data representation. Its average value, $U[X,Y] := \langle u(x,y)\rangle_{p(x,y)}$ then quantifies the utility of the representation.

We are now in a position to state a *least effort principle* demanding that input data should be represented in such a way that the average free energy change (which is a lower bound on the effort) is minimized. Define the *least effort*,

[1] For simplicity, we measure information in units of the natural logarithm (nats). The shorthand $\langle\cdot\rangle_p$ denotes the average taken over the distribution p.

[2] Notation uses the convention in [4]: capital letters X and Y denote random variables. For visual clarity, all optimization problems appear without the constraints that ensure normalization and positivity of $p(y|x)$.

[3] Keep in mind that $p(y) = \langle p(y|x)\rangle_{p(x)}$.

[4] k_B is the Boltzmann constant, and T the temperature of a heat bath surrounding the system. The assumption is that the system exchanges only heat with the surroundings, and that the heat bath is large compared to the system which may be driven arbitrarily far from equilibrium by a change in external parameters. These parameter changes allow for doing work on and extracting work from the system.

$L(U)$, involved in representing x as y by the minimum free energy change compatible with utility U:

$$L(U) := \min_{p(y|x)} \Delta F[X,Y] \quad \text{s.t. } U[X,Y] = U. \tag{3}$$

The least effort function quantifies how conservative one can be with the expenditure of energy while achieving the intended utility. In other words, it measures how lazy one can afford to be.

Observations related to least effort coding have previously come up in the context of language [17, 18]. The effort of the speaker was modeled as the entropy of the code signals, while the effort for the listener was modeled as conditional entropy of the objects of reference, given the signal [18]. The combined effort was minimized, and the relative importance of the two terms was controlled by a parameter. At a critical value, Zipf's law [17] was retrieved at a phase transition [18]. While related in general spirit, the measure used in [18] is not the same as the physical effort discussed here.[5]

4. RATE-DISTORTION CURVE IS A LEAST EFFORT FUNCTION

Let a physical system that is in a state described by $p(y)$ have internal energy $E(y)$, and let the energy associated with the state described by $p(y|x)$ be denoted by $E_x(y)$. Write the difference as $\mathcal{E}(x,y) := E_x(y) - E(y)$, and denote its average by $E[X,Y] := \langle \mathcal{E}(x,y) \rangle_{p(x,y)}$. Then the least effort involved in the change $p(y) \to p(y|x)$, averaged over all x, is given by

$$\begin{aligned} \Delta F[X,Y] &= \langle E_x(y) \rangle_{p(y|x)p(x)} - k_B T H[Y|X] \\ &\quad - \langle E(y) \rangle_{p(y)} + k_B T H[Y] \qquad (4) \\ &= E[X,Y] + k_B T I[X,Y] \,. \qquad (5) \end{aligned}$$

Now consider the case that the average energy does not change, i.e. $\langle E_x(y) \rangle_{p(y|x)p(x)} = \langle E(y) \rangle_{p(y)}$, in other words, $E[X,Y] = 0$. A simple example is given by a particle in a double well potential. For simplicity of the exposition, make the potential rectangular, having an energy barrier of infinite energy between two wells of identical width and identical energy, E_0. Coarse grain the position of the particle so that $y = 0$ ($y = 1$) denotes the particle in the left (right) well. Then $E(y=0) = E(y=1) = E_0$, and hence $\langle E(y) \rangle_{p(y)} = E_0$. The particle can be forced into either well by deformation of the potential. Let $x \in \mathbb{R}$, and let the protocol that achieves this preparation of y depend on x, so that, at the end of the protocol, $y = \theta(x)$. Let, e.g.,

$$E_x(y) = \begin{cases} E_0 & \text{if } y = \theta(x) \\ \infty & \text{else} \end{cases} \,, \tag{6}$$

and

$$p(y|x) = \begin{cases} 1 & \text{if } y = \theta(x) \\ 0 & \text{else} \end{cases} \,. \tag{7}$$

Then $\langle E_x(y) \rangle_{p(y|x)} = E_0$, which is independent of x, and therefore $\langle E_x(y) \rangle_{p(y|x)p(x)} = E_0$, for any $p(x)$.

For classical systems and measurements, things can often be set up in such a way that the assumption $E[X,Y] = 0$ is valid. It could, however be violated by quantum entanglement, and also possibly in living, metabolizing agents. Both of these areas are outside the scope of this paper.

Under the assumption that the average energy does not change, the free energy change is proportional to mutual information:

$$\Delta F[X,Y] = k_B T I[X,Y] \,. \tag{8}$$

The least effort principle thus dictates minimization of mutual information.

The optimization problem in Eq. (3) can be solved using the method of Lagrange multipliers. The constraint is added to the objective function, with a Lagrange multiplier that effectively controls the trade-off between minimal effort and achieved utility. For data compression, utility is related to fidelity and can be identified with negative distortion.

A least effort data compression then has to solve

$$\min_{p(y|x)} \Big(\Delta F[X,Y] + \lambda D[X,Y] \Big) \,. \tag{9}$$

Comparison with Eq. (8) reveals that this is equivalent to $\min_{p(y|x)} \big(I[X,Y] + \lambda \bar{D}[X,Y] \big)$, where $\bar{D} = D/k_B T$ is the distortion measured in units of $k_B T$. The solution to this problem lies on the rate-distortion curve, $R(\bar{D})$, as we can see from comparison with the optimization problem in Eq. (1). This shows that the rate-distortion curve is a least effort function.

This finding is similar, but not identical to the formal mapping of the rate-distortion function onto free energy minimization in multiphase chemical equilibrium [3], and to the statements in [19, 20], where large deviations theory was used to show a formal analogy between the rate-distortion function and the free energy of a chain of particles, i.e. the minimum amount of work needed to compress the chain. These formal analogies are based on identifying the distortion function with physical aspects of a corresponding system, e.g. its energy. It was pointed out in [20] that these formal analogies have some interpretational freedom. Specifically, the interpretation of the Lagrange multiplier that effectively controls the trade-off between distortion and compression depends on the details of the analogy. In the mechanical analogy, it can be interpreted either as inverse temperature [19], or as a conjugate force [20]. In contrast, the derivation given above retains explicitly the distortion constraint and shows that physical temperature adjusts the units by rescaling the distortion measure, or, alternatively, by rescaling the trade-off parameter.

[5] Written in the notation used here, the effort in [18] was quantified by $\lambda H[X|Y] + (1-\lambda) H[Y]$, where the parameter λ weights how much listener and speaker contribute to the total effort. $H[Y] = -\langle \log[p(y)] \rangle_{p(y)}$ denotes Shannon entropy, and $H[X|Y] = -\langle \log[p(x|y)] \rangle_{p(x,y)}$ conditional entropy.

5. CHANNEL CAPACITY

The output messages y can also be interpreted as measurement outcomes. If the measurement is useful, then the observer learns something about x when given y. In the absence of y, the observer's best guess about x is expressed by the prior probability $p(x)$, but when the measurement is received, this changes to the posterior distribution $p(x|y) = p(y|x)p(x)/p(y)$ (Bayes' rule) [21]. Changing of the observer's knowledge state from prior to posterior comes at a cost; it takes a certain amount of effort to implement this change. By the same arguments as above, the minimum amount of work that has to be done (on average) is given by the free energy difference $\langle F[p(x|y)]\rangle_{p(y)} - F[p(x)]$.

This quantity also determines the maximum amount of work that can be *extracted* from a physical system which is (partially) described by x, by exploiting knowledge of y. By convention, energy flowing *into* the system is positive, while energy flowing *out* of the system has a negative sign. Hence, at most $F[p(x)] - \langle F[p(x|y)]\rangle_{p(y)}$ can be extracted as work. Assuming once again no change in average energy, i.e. $\langle E(x)\rangle_{p(x)} = \langle E_y(x)\rangle_{p(x|y)p(y)}$, we have

$$F[p(x)] - \langle F[p(x|y)]\rangle_{p(y)} = -k_B T I[X,Y] \,. \qquad (10)$$

Therefore, maximization of extractable work motivates maximization of mutual information.[6]

A simple example in which the condition $\langle E(x)\rangle_{p(x)} = \langle E_y(x)\rangle_{p(x|y)p(y)}$ holds is that of measuring the x-position of a particle in a box connected to a heat bath at temperature T. Let the length of the box be L. Then $p(x) = 1/L$ inside the box and zero outside. The energy of the particle does not depend on its x-position within the box, where it is given by the particle's kinetic energy, E_K, but the walls of the box pose an infinite energy barrier. Thus we may write:

$$E(x) = \begin{cases} E_K & \forall x \in [0,L] \\ \infty & \forall x \notin [0,L] \end{cases} \,. \qquad (11)$$

The average energy is $\langle E(x)\rangle_{p(x)} = E_K$.

Knowing that the particle is confined e.g. to the left side of the box results in a posterior of $p(x|y = \text{“LEFT”}) = 2/L$ for x between 0 and $L/2$, and zero outside that range (similarly for $y =$ “RIGHT”). This distribution describes a particle in a box of half of the length, but otherwise the same as the original box. The particle's energy is then E_K inside the range of the smaller box, and infinite outside that range:

$$E_{y=\text{“LEFT”}}(x) = \begin{cases} E_K & \forall x \in [0,L/2] \\ \infty & \forall x \notin [0,L/2] \end{cases} \,, \qquad (12)$$

and

$$E_{y=\text{“RIGHT”}}(x) = \begin{cases} E_K & \forall x \in (L/2,L] \\ \infty & \forall x \notin (L/2,L] \end{cases} \,, \qquad (13)$$

with an expected value of $\langle E_y(x)\rangle_{p(x|y)p(y)} = E_K$.[7]

Now, assume that the distribution $p(y|x)$, which describes the data representation method, or, alternatively, the measurement apparatus, be fixed. Then ask for the physical system that best matches the measurement apparatus in the sense that it allows for maximum work extraction, given the measurement (on average). Eq. (10) tells us that the answer is given by Shannon's channel capacity:

$$C = \max_{p(x)} I[X,Y] \,. \qquad (14)$$

For a fixed channel, one chooses the source which would allow for exploiting the knowledge obtained from the messages y towards maximum work extraction.

These are two sides of a coin: communicating more information costs more effort, but the more informative a measurement is about the state of a physical system, the more work that can be extracted from the system given the measurement outcome.

6. LEAST EFFORT MAXIMUM WORK EXTRACTION

Imagine two correlated systems, $\mathcal{X}$ and $\mathcal{Z}$ with mutual information $I[X,Z]$. An observer measures x, and represents it by y, which is obtained with probability $p(y|x)$. This representation, or measurement, can then be used to extract work from system $\mathcal{Z}$.

Knowledge of system $\mathcal{Z}$, given y, is expressed by the probability distribution $p(z|y)$. By the same arguments as above, the maximum amount of extractable work (averaged over all measurements) is given by the free energy difference $F[p(z)] - \langle F[p(z|y)]\rangle_{p(y)}$. Under the assumption that the average energy does not change, this is given by $-k_B T I[Y,Z]$.

The least effort data representation method which maximizes extractable work thus solves

$$\min_{p(y|x)} \left(I[X,Y] - \alpha I[Y,Z] \right) \,, \qquad (15)$$

The Lagrange multiplier α controls the trade-off between work extractable from system $\mathcal{Z}$ (which one wants to maximize) and necessary effort to represent system $\mathcal{X}$ (which one wants to minimize). We recognize Eq. (15) as the *Information Bottleneck* (IB) method [22], hereby lending IB a new thermodynamic motivation: it finds the least effort representation of system $\mathcal{X}$ that allows for maximum work extraction from a correlated system $\mathcal{Z}$.

If $\mathcal{X}$ and $\mathcal{Z}$ are kept at two different temperatures, T_X and T_Z, then α can be interpreted as the ratio T_Z/T_X: the larger the temperature difference, the more beneficial it is to keep relevant information, as it can be traded off for more extractable work.

7. SUMMARY

Least effort coding leads to data representations that lie on the rate distortion curve. Least effort is measured by

[6] Be reminded of the sign. $I[X,Y]$ is a non-negative quantity. Extracted work comes with a negative sign. Thus, more work can be extracted when $I[X,Y]$ is larger.

[7] This holds for all $p(y)$, because $\langle E_y(x)\rangle_{p(x|y)} = E_K$, which is independent of y.

the average free energy difference, quantifying the least amount of physical work necessary to change a system described by the average output distribution to one described by the specific output distribution necessary to produce output messages when the input is known. Information loss has to do with thermodynamic efficiency: least effort is proportional to mutual information (under the assumption that there is no average energy change). Codes that are efficient in a rate-distortion sense are also energetically efficient. In that sense, lossy compression is lazy compression, because it minimizes physical effort.

Channel capacity, on the other hand, represents the maximum amount of work that could be extracted from a source (on average) given the channel's output messages. In contrast to the above, where the source is fixed and the optimization is over encoding schemes, here the channel is given. The maximum is then taken over all possible sources, thus optimizing over physical systems for the best match in terms of possible work extraction.

The Information Bottleneck method provides the means of finding a minimum effort compression (or data representation) that allows for maximum work extraction from another system by exploiting correlations.

8. ACKNOWLEDGEMENTS

I am grateful for support from the Foundational Questions Institute (FQXi). I thank Arne Grimsmo and Rob Shaw for inspiring discussions that initiated this research, and Toby Berger, Antonio Celani, Gavin Crooks, David Sivak and Elan Stopnitzky for extremely valuable comments on the manuscript.

9. REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[2] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec*, vol. 4, no. 142-163, pp. 1, 1959.

[3] T. Berger, "Rate distortion theory: A mathematical basis for data compression," 1971.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 2nd edition, 2006.

[5] F. Schlögl, "On stability of steady states," *Zeitschrift für Physik*, vol. 243, no. 4, pp. 303–310, 1971.

[6] J. Schnakenberg, "Network theory of microscopic and macroscopic behavior of master equation systems," *Reviews of Modern physics*, vol. 48, no. 4, pp. 571, 1976.

[7] R. Shaw, *The dripping faucet as a model chaotic system*, Aerial Press, 1984.

[8] B. Gaveau and L. S. Schulman, "A general framework for non-equilibrium phenomena: The master equation and its formal consequences," *Phys. Lett. A*, vol. 229, no. 6, pp. 347–353, 1997.

[9] H. Qian, "Relative Entropy: Free Energy Associated with Equilibrium Fluctuations and Nonequilibrium Deviations," *Phys. Rev. E*, vol. 63, pp. 042103, 2001.

[10] G. E. Crooks, "Beyond Boltzmann-Gibbs statistics: Maximum entropy hyperensembles out-of-equilibrium," *Phys. Rev. E*, vol. 75, pp. 041119, 2007.

[11] M. Esposito and C. Van den Broeck, "Second law and landauer principle far from equilibrium," *EPL (Europhysics Letters)*, vol. 95, no. 4, pp. 40004, 2011.

[12] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks, "Thermodynamics of prediction," *Physical Review Letters*, vol. 109, pp. 120604, 2012.

[13] S. Deffner and C. Jarzynski, "Information processing and the second law of thermodynamics: An inclusive, hamiltonian approach," *Phys. Rev. X*, vol. 3, pp. 041003, Oct 2013.

[14] B. Gaveau, L. Granger, M. Moreau, and L. S. Schulman, "Relative entropy, interaction energy and the nature of dissipation," *Entropy*, vol. 16, no. 6, pp. 3173–3206, 2014.

[15] L. Szilard, "On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings," *Z. Phys.*, vol. 53, pp. 840–856, 1929.

[16] J. V. Koski, V. F. Maisi, T. Sagawa, and J. P. Pekola, "Experimental study of mutual information in a Maxwell Demon," *arXiv preprint arXiv:1405.1272*, 2014.

[17] G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA, 1949.

[18] R. F. i Cancho and R. V. Solé, "Least effort and the origins of scaling in human language," *Proceedings of the National Academy of Sciences*, vol. 100, no. 3, pp. 788–791, 2003.

[19] N. Merhav, "An identity of chernoff bounds with an interpretation in statistical physics and applications in information theory," *Information Theory, IEEE Transactions on*, vol. 54, no. 8, pp. 3710–3721, 2008.

[20] N. Merhav, "Another look at the physics of large deviations with application to rate-distortion theory," *arXiv:0908.3562*, 2009.

[21] H. Jeffreys, *Theory of Probability*, Oxford University Press, third edition, 1998.

[22] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annual Allerton Conference*, B. Hajek and R. S. Sreenivas, Eds. 1999, pp. 368–377, University of Illinois, Available at http://xxx.arXiv.cornell.edu/abs/physics/0004057.

RATE-DISTORTION ANALYSIS FOR AN EPSILON-INSENSITIVE LOSS FUNCTION

Kazuho Watanabe

Department of Computer Science and Engineering, Toyohashi University of Technology,
1-1 Hibarigaoka Tempaku-cho Toyohashi, Aichi 441-8580, JAPAN, wkazuho@cs.tut.ac.jp

ABSTRACT

Explicit evaluation of the rate-distortion function has rarely been achieved when it is strictly greater than its Shannon lower bound. In this paper, we consider the rate-distortion function for the distortion measure defined by an ε-insensitive loss function. We first present the Shannon lower bound for this distortion measure and provide a necessary condition for its tightness. Then, focusing on the Laplacian and Gaussian sources, we prove that the rate-distortion functions of these sources are strictly greater than their Shannon lower bounds and obtain analytic upper bounds for the rate-distortion functions. Small distortion limits of the bounds and approximate computation of the rate-distortion function suggest that the Shannon lower bound provides a good approximation to the rate-distortion function for the ε-insensitive distortion measure.

1. INTRODUCTION

In source coding, the rate-distortion function $R(D)$ of a source shows the minimum information rate required to reconstruct the source outputs with average distortion not exceeding D. Rate-distortion functions have been explicitly evaluated for various sources and distortion measures. The Shannon lower bound (SLB) $R_L(D)$ plays an important role in the explicit evaluation of rate-distortion functions of difference distortion measures. A common approach is to derive $R_L(D)$ and examine the condition for $R(D)$ to coincide with $R_L(D)$ [1, 2]. There have been, however, only several results when $R(D) > R_L(D)$ for all D. In this case, direct explicit evaluation of $R(D)$ has been achieved only in limited cases such as discrete memoryless finite-alphabet sources [2] and a class of sources under an absolute-magnitude distortion measure [3, 4, 5]. There have also been indirect approaches. Rose proposed a deterministic annealing algorithm to generate $R(D)$ based on the fact that under the squared distortion measure, the optimal reconstruction is purely discrete when $R(D) > R_L(D)$ [6]. Buzo *et al.* obtained upper and lower bounds for $R(D)$ under the Itakura-Saito distortion measure [7].

In this paper, we focus on the ε-insensitive loss function as a distortion measure, which was introduced to support vector machines for regression [8]. We obtain the SLB for this difference distortion measure, which is analytically evaluable for arbitrary sources with finite differential entropy. Then, we examine the condition for the rate-distortion function to coincide with the SLB. Taking the Laplacian and Gaussian sources as specific examples, we prove that the rate-distortion functions of these sources lie strictly above their SLBs for all D when $\varepsilon > 0$ and derive analytic upper bounds for the rate-distortion functions. Investigation of small distortion limit of these upper bounds shows that the SLB has the accuracy of $O(\varepsilon^2)$ as $D \to 0$ in both sources. We then apply the learning algorithm of the finite mixture of ε-insensitive component distributions developed in [9] and approximately compute the rate-distortion function. The approximate computation also suggests that the SLB provides a good approximation to the rate-distortion function.

2. RATE-DISTORTION FUNCTION FOR THE ε-INSENSITIVE DISTORTION MEASURE

2.1. Rate-Distortion Function

Let X and Y be random variables on $\mathbf{R}$ and $d(x, y)$ be the non-negative distortion measure between x and y. The rate-distortion function $R(D)$ of the source $X \sim p(x)$ with respect to the distortion d is defined by

$$R(D) = \inf_{q(y|x):E[d(X,Y)]\leq D} I(q), \tag{1}$$

where

$$I(q) = \int\int q(y|x)p(x)\log\frac{q(y|x)}{\int q(y|x)p(x)dx}dxdy$$

is the mutual information and E denotes the expectation with respect to $q(y|x)p(x)$. $R(D)$ shows the minimum achievable rate for the i.i.d. source with the density $p(x)$ under the given distortion measure d [2, 1].

The above minimization problem can be reformulated as a minimization problem over the reproduction density $q(y)$,

$$\inf_{q(y)}\left[-\int p(x)\log\int\exp(sd(x,y))q(y)dydx\right], \tag{2}$$

where $s \leq 0$ is a parameter [2, 10]. Then, if there exists $q_s(y)$ that achieves the infimum in Eq. (2), $R(D)$ is parametrically given by

$$\begin{aligned} R(D_s) &= -\int p(x)\log\int\exp(sd(x,y))q_s(y)dydx + sD_s, \\ D_s &= \int\int p(x)q_s(y|x)d(x,y)dxdy, \end{aligned} \tag{3}$$

where the optimal conditional density of reconstruction, $q_s(y|x)$ is defined by

$$q_s(y|x) = \frac{q_s(y)\exp(sd(x,y))}{\int q_s(y)\exp(sd(x,y))dy}. \quad (4)$$

In Eq. (3), $R(D)$ is parameterized by $s \leq 0$, which corresponds to the slope of the tangent of $R(D)$ at $(D_s, R(D_s))$ and hence is referred to as the slope parameter [2].

From the properties of the rate-distortion function $R(D)$, we know that $R(D) > 0$ for $0 < D < D_{\max}$, where

$$D_{\max} = \inf_y \int p(x)d(x,y)dx, \quad (5)$$

and $R(D) = 0$ for $D \geq D_{\max}$ [2, p. 90].

2.2. ε-Insensitive Loss Function

In this paper, we focus on the following difference distortion measure defined by the ε-insensitive loss function ρ_ε,

$$d(x,y) = \rho_\varepsilon(x-y), \quad (6)$$

where

$$\rho_\varepsilon(z) = \begin{cases} |z| - \varepsilon, & (|z| \geq \varepsilon), \\ 0, & (|z| < \varepsilon). \end{cases}$$

This loss function with $\varepsilon > 0$ was introduced to support vector regression in order to provide a sparsity inducing mechanism [8, 11]. We denote the rate-distortion function for this distortion measure by $R^{(\varepsilon)}(D)$ and the maximum distortion $D_{\max}$ in Eq. (5) by $D^{(\varepsilon)}_{\max}$.

2.3. Shannon Lower Bound

Generally for difference distortion measures, Shannon obtained a lower bound to $R(D)$, which is referred to as the Shannon lower bound (SLB) [2, p. 92]. For the ε-insensitive distortion measure, it is parametrically expressed as

$$R^{(\varepsilon)}(D_s) \geq R^{(\varepsilon)}_L(D_s) = h(p) - h(g_s), \quad (7)$$

$$D_s = \int \rho_\varepsilon(x) g_s(x) dx, \quad (8)$$

where $h(p) = -\int p(x)\log p(x)dx$ is the differential entropy of the probability density p and g_s is the probability density function defined by[1]

$$g_s(x) = \frac{e^{s\rho_\varepsilon(x)}}{\int e^{s\rho_\varepsilon(z)}dz}. \quad (9)$$

We explicitly evaluate $h(g_s)$ to obtain the SLB. The density g_s is explicitly given by

$$g_s(x) = \begin{cases} \frac{1}{C_s}, & (|x| < \varepsilon), \\ \frac{1}{C_s} e^{s(|x|-\varepsilon)}, & (|x| \geq \varepsilon), \end{cases} \quad (10)$$

where

$$C_s = 2\frac{1+|s|\varepsilon}{|s|}. \quad (11)$$

[1] We omit the dependency on ε in notations unless we put $\varepsilon = 0$.

Its differential entropy is evaluated as,

$$h(g_s) = \log\left(2\frac{1+|s|\varepsilon}{|s|}\right) + \frac{1}{1+|s|\varepsilon}. \quad (12)$$

The slope parameter s is related to the average distortion D_s by Eq. (8), which is rewritten as,

$$D_s = \frac{1}{(1+\varepsilon|s|)|s|}. \quad (13)$$

Solving this for $|s|$ yields $|s| = \frac{-D_s + \sqrt{D_s^2 + 4D_s\varepsilon}}{2D_s\varepsilon}$. Putting this back into Eq. (12), from Eq. (7), we obtain the following theorem.

Theorem 1 *The rate-distortion function for the ε-insensitive distortion measure in Eq. (6) satisfies $R^{(\varepsilon)}(D) \geq R^{(\varepsilon)}_L(D)$ for all D, where*

$$R^{(\varepsilon)}_L(D) = h(p) - \log\left(1 + \tilde{D} + \sqrt{\tilde{D}^2 + 2\tilde{D}}\right) - \log(2\varepsilon) + \tilde{D} - \sqrt{\tilde{D}^2 + 2\tilde{D}},$$

$\tilde{D} = \frac{D}{2\varepsilon}$ and $h(p)$ is the differential entropy of the source density.

2.4. Condition for $R^{(\varepsilon)}(D) = R^{(\varepsilon)}_L(D)$

For any negative value of the slope parameter s, the lower bound $R^{(\varepsilon)}_L(D_s)$ coincides with $R^{(\varepsilon)}(D_s)$ if and only if the condition

$$p(x) = \int q(y) g_s(x-y)dy, \quad (14)$$

holds for all x and a valid density function $q(y)$ [2, p. 94]. The condition in Eq. (14) is equivalent to

$$P(\omega) = Q(\omega)G_s(\omega), \quad (15)$$

where P, Q and G_s are the Fourier transforms (characteristic functions) of p, q and g_s respectively.

The Fourier transform of g_s is specifically given by

$$G_s(\omega) = \frac{s^2}{s^2+\omega^2} \cdot \frac{\varepsilon|s|\frac{\sin(\omega\varepsilon)}{\omega\varepsilon} + \cos(\omega\varepsilon)}{1+|s|\varepsilon} \equiv L_{|s|}(\omega) \cdot M^{(\varepsilon)}_{|s|}(\omega). \quad (16)$$

Here, the first factor, defined as $L_{|s|}(\omega) = \frac{s^2}{s^2+\omega^2}$, is the characteristic function of the Laplace distribution with parameter $|s|$ whose density function is $l_{|s|}(x) = \frac{|s|}{2}e^{s|x|}$. The second factor, $M^{(\varepsilon)}_{|s|}(\omega) = \frac{\varepsilon|s|\frac{\sin(\omega\varepsilon)}{\omega\varepsilon}+\cos(\omega\varepsilon)}{1+|s|\varepsilon}$, is the characteristic function of the mixture of the delta distributions (on $-\varepsilon$ and ε with equal weight) and the uniform distribution on $[-\varepsilon, \varepsilon]$ mixed with the proportion $1 : \varepsilon|s|$. More specifically, the density function $m^{(\varepsilon)}_{|s|}(x)$ of this mixture is expressed as

$$\frac{1}{1+\varepsilon|s|}\frac{\delta(x-\varepsilon)+\delta(x+\varepsilon)}{2} + \frac{\varepsilon|s|}{1+\varepsilon|s|}u_{[-\varepsilon,\varepsilon]}(x),$$

where δ is the Dirac delta function and $u_{[-\varepsilon,\varepsilon]}$ is the density function of the uniform distribution on $[-\varepsilon, \varepsilon]$. Hence, Eq. (16) means that the density g_s is given by the convolution $l_{|s|} * m^{(\varepsilon)}_{|s|}$ of $l_{|s|}$ and $m^{(\varepsilon)}_{|s|}$. Summarizing Eqs. (15) and (16), we see that for the ε-insensitive distortion measure, the condition for $R^{(\varepsilon)}_L(D)$ to coincide with $R^{(\varepsilon)}(D)$ is the existence of a valid characteristic function $Q(\omega)$ satisfying

$$P(\omega) = Q(\omega) L_{|s|}(\omega) M^{(\varepsilon)}_{|s|}(\omega), \tag{17}$$

for the characteristic function $P(\omega)$ of the source distribution.

The above condition leads to a necessary condition for $R^{(\varepsilon)}(D_s) = R^{(\varepsilon)}_L(D_s)$.

Lemma 1 *Given any $s \leq 0$, if $R^{(\varepsilon)}(D_s) = R^{(\varepsilon)}_L(D_s)$ then $R^{(0)}(D_s) = R^{(0)}_L(D_s)$, that is, the SLB coincides with the rate-distortion function under the absolute distortion measure.*

2.5. General Upper Bound

Let us turn to upper bounds for $R^{(\varepsilon)}(D)$. Since $\rho_\varepsilon(x) \leq \rho_0(x) = |x|$, we have a trivial upper bound,

$$R^{(\varepsilon)}(D) \leq R^{(0)}(D),$$

where $R^{(0)}(D)$ is the rate-distortion function for the absolute-magnitude distortion measure, $d(x, y) = |x - y|$.

Another more informative upper bound is obtained by taking $q(y|x) = g_s(y-x)$, where g_s is defined by Eq. (9), in the original rate-distortion problem in Eq. (1) [2, p. 103]. This yields the following upper bound,

$$R^{(\varepsilon)}(D_s) \leq R^{(\varepsilon)}_U(D_s) = h(r_s) - h(g_s), \tag{18}$$

where

$$r_s(y) = (g_s * p)(y) = \int g_s(y - x) p(x) dx \tag{19}$$

and D_s is given by Eq. (8) and further by Eq. (13). Note in Eq. (18) that the term $h(g_s)$ is common to the SLB and is specifically given by Eq. (12).

Since g_s is defined by ρ_ε as in Eq. (9), $h(r_s)$ is analytically intractable for many sources. Hence, we create a further upper bound which is analytically obtained for any sources with finite variance.

Let $v_p \equiv \int x^2 p(x) dx - \left(\int x p(x) dx\right)^2$ and $v^{(\varepsilon)}_s \equiv \int x^2 g_s(x) dx$, which is specifically evaluated as,

$$v^{(\varepsilon)}_s = \frac{2}{C_s} \left\{ \frac{\varepsilon^3}{3} + \frac{1}{|s|} \left(\varepsilon^2 + \frac{2}{|s|} \varepsilon + \frac{2}{|s|^2} \right) \right\}.$$

Then, the maximum entropy principle of the Gaussian distribution yields the following upper bound to $R^{(\varepsilon)}_U(D)$, which is referred to as the Gaussian entropy bound.

Lemma 2 *For $s \leq 0$, $R^{(\varepsilon)}(D_s) \leq R^{(\varepsilon)}_U(D_s) \leq R^{(\varepsilon)}_{GE}(D_s)$, where*

$$R^{(\varepsilon)}_{GE}(D_s) = \frac{1}{2} \log \left(2\pi e (v_p + v^{(\varepsilon)}_s) \right) - h(g_s). \tag{20}$$

In the next sections, we will evaluate these upper bounds for the Laplacian and Gaussian sources to examine the tightness of the general lower bound obtained in Theorem 1.

3. LAPLACIAN AND GAUSSIAN SOURCES

3.1. Laplacian Source

In this subsection, we consider the Laplacian source with parameter α,

$$p(x) = l_\alpha(x) = \frac{\alpha}{2} e^{-\alpha|x|}. \tag{21}$$

The SLB for this source is given by Theorem 1 with the differential entropy, $h(p) = 1 - \log \frac{\alpha}{2}$. The maximum distortion in Eq. (5) is

$$D^{(\varepsilon)}_{\max} = \int \rho_\varepsilon(x) p(x) dx = \frac{1}{\alpha} e^{-\alpha\varepsilon}. \tag{22}$$

For the absolute-magnitude distortion measure ($\varepsilon = 0$),

$$R^{(0)}(D) = R^{(0)}_L(D) = -\log(\alpha D), \quad (0 \leq D \leq 1/\alpha), \tag{23}$$

holds [2, p. 95, Example 4.3.2.1] because the condition in Eq. (17) is satisfied by $M^{(0)}_{|s|}(\omega) = 1$ and $Q(\omega) = \frac{\alpha^2}{|s|^2} + \left(1 - \frac{\alpha^2}{|s|^2}\right) \frac{\alpha^2}{\alpha^2 + \omega^2}$, which is the Fourier transform of the valid probability density, $q(y) = \frac{\alpha^2}{|s|^2} \delta(y) + \left(1 - \frac{\alpha^2}{|s|^2}\right) l_\alpha(y)$. For $\varepsilon > 0$, however, $R^{(\varepsilon)}(D)$ is strictly greater than $R^{(\varepsilon)}_L(D)$ for all D [12].

To obtain an analytic upper bound for $R^{(\varepsilon)}(D)$, we evaluate $h(r_s)$ in Eq. (18). Since the differential entropy of $r_s(y)$ is not analytically simplified any more, we evaluate it from above to obtain an upper bound for $R^{(\varepsilon)}(D_s)$.

Let us define $B_s \equiv \frac{s}{\alpha - s} \frac{1}{\alpha} e^{-2\alpha\varepsilon} + \frac{s^2(\alpha - s) - 2\alpha^3}{(\alpha^2 - s^2)s\alpha}$ and $E_s \equiv \frac{s}{\alpha - s} \frac{1 + \alpha\varepsilon}{\alpha^2} e^{-2\alpha\varepsilon} + \frac{s}{s + \alpha} \frac{1 + \alpha\varepsilon}{\alpha^2} + \frac{2\alpha^2}{\alpha^2 - s^2} \frac{1 - s\varepsilon}{s^2}$. Then, we have the further upper bound, which we will refer to as the analytic upper bound in Section 4 [12].

Theorem 2 *For the Laplacian source density in Eq. (21), $R^{(\varepsilon)}_L(D_s) < R^{(\varepsilon)}(D_s) \leq R^{(\varepsilon)}_U(D_s) \leq R^{(\varepsilon)}_{AU}(D_s)$, where*

$$R^{(\varepsilon)}_{AU}(D_s) \equiv -\log \frac{c_s}{2C_s} - \frac{\alpha\varepsilon}{C_s} B_s + \frac{\alpha}{C_s} E_s - h(g_s). \tag{24}$$

In the low distortion limit, $D \to 0$ and $|s| \to \infty$, we have ([12]),

$$R^{(\varepsilon)}_L(0) < R^{(\varepsilon)}(0) \leq R^{(\varepsilon)}_L(0) + \frac{(\alpha\varepsilon)^2}{2} + O(\varepsilon^3).$$

3.2. Gaussian Source

In this subsection, we consider the Gaussian source with mean zero and variance σ^2,

$$p(x) = \sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}. \tag{25}$$

The differential entropy of this source is $h(p) = \frac{1}{2}\{1 + \log(2\pi\sigma^2)\}$. Since $R^{(0)}(D) > R^{(0)}_L(D)$ holds for all D [3], Lemma 1 and Lemma 2 lead to the following theorem.

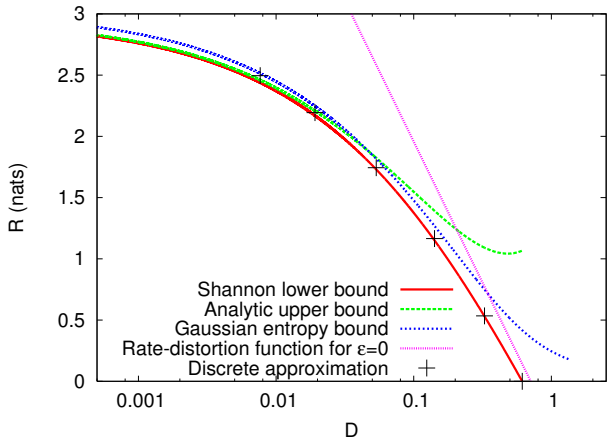

Figure 1. Rate-distortion bounds for the Laplacian source.

Theorem 3 *For the Gaussian source density in Eq. (25),* $R_L^{(\varepsilon)}(D_s) < R^{(\varepsilon)}(D_s) \le R_U^{(\varepsilon)}(D_s) \le R_{GE}^{(\varepsilon)}(D_s)$*, where*

$$R_{GE}^{(\varepsilon)}(D_s) = \frac{1}{2}\log\left(2\pi e(\sigma^2 + v_s^{(\varepsilon)})\right) - h(g_s). \quad (26)$$

In the limit, $|s| \to \infty$, $v_s^{(\varepsilon)} \to \varepsilon^2/3$. This means that the SLB given by Theorem 1 provides an approximation to $R^{(\varepsilon)}(D)$ with accuracy $\frac{\varepsilon^2}{6\sigma^2}$ as $D \to 0$.

4. NUMERICAL EVALUATION

Figure 1 depicts the functions $R_L^{(\varepsilon)}(D)$ and $R_{AU}^{(\varepsilon)}(D)$ for the Laplacian source in Eq. (21) with $\alpha = \sqrt{2}$ when $\varepsilon = 0.1$. It also shows $R_{GE}^{(\varepsilon)}(D)$ in Eq. (20) with $v_p = 2/\alpha^2 = 1$ and the trivial upper bound $R^{(0)}(D)$ given by Eq. (23). It is observed that the analytic upper bound and the SLB are very close to each other for small distortion ($D < 0.01$). The analytic upper bound becomes looser than the Gaussian entropy bound for large distortion ($0.05 < D$) while the trivial upper bound is relatively more informative about $R^{(\varepsilon)}(D)$ in the vicinity of $D_{\max}^{(\varepsilon)}$ when combined with the SLB.

Then, restricting the reconstruction distribution to be a discrete distribution with finite mass points, we estimated the parameters $\{a_k; a_k \ge 0, k = 1, \cdots, K, \sum_{k=1}^K a_k = 1\}$ and $\{y_k \in R; k = 1, \cdots, K\}$ of the finite mixture model, $\sum_{k=1}^K a_k g_s(x - y_k)$, by the learning algorithm developed in [9], and approximately computed the 6 points on the rate-distortion function corresponding to $|s| = 1.25$, 2.5, 5, 10, 20, and 40. The obtained approximations are very close to the SLB (Fig.1).

5. CONCLUSION

In this extended abstract, we have shown upper and lower bounds for the rate-distortion function of the ε-insensitive distortion measure. We derived the SLB, which is applicable to any source densities. Focusing on the Laplacian and Gaussian sources, we have proved that the rate-distortion functions for these sources are strictly greater than the corresponding SLBs for all D and provided upper bounds for the rate-distortion functions, which are proved to have accuracy of $O(\varepsilon^2)$ in the small distortion limit. We have demonstrated through numerical evaluation that the SLB is very accurate in the small distortion region while it also provides a good approximation to $R^{(\varepsilon)}(D)$ for the high distortion region as the approximate computation suggests.

6. ACKNOWLEDGMENTS

This work was supported in part by JSPS KAKENHI Grant Numbers 23700175 and 25120014.

7. REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, 1991.

[2] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Englewood Cliffs, NJ: Prentice-Hall, 1971.

[3] H. H. Tan and K. Yao, "Evaluation of rate-distortion functions for a class of independent identically distributed sources under an absolute magnitude criterion," *IEEE Transactions on Information Theory*, vol. IT-21, no. 1, pp. 59–64, 1975.

[4] K. Yao and H. H. Tan, "Absolute error rate-distortion functions for sources with constrained magnitudes," *IEEE Transactions on Information Theory*, vol. IT-24, no. 4, pp. 499–503, 1978.

[5] K. Watanabe and S. Ikeda, "Rate-distortion function for gamma sources under absolute-log distortion measure," in *Proc. of ISIT*, 2013, pp. 2557–2561.

[6] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1939–1952, 1994.

[7] A. Buzo, F. Kuhlmann, and C. Rivera, "Rate-distortion bounds for quotient-based distortions with application to Itakura-Saito distortion measures," *IEEE Transactions on Information Theory*, vol. IT-32, no. 2, pp. 141–147, 1986.

[8] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

[9] K. Watanabe, "Vector quantization using mixture of epsilon-insensitive components," in *Proc. of ICONIP*. 2013, pp. 85–92, Springer.

[10] R. M. Gray, *Entropy and Information Theory (2nd Edition)*, Springer, 2011.

[11] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, 2008.

[12] K. Watanabe, "Rate-distortion bounds for an ε-insensitive distortion measure," in *Proc. of ITW*. 2013, pp. 679–683, IEEE.

DISCOVERY PHENOMENON AND INFORMATION CRITERIA

SUMIO WATANABE[1]

[1] Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology
Mailbox G5-19, Nagatsuta Midoriku Yokohama 226-8502, JAPAN, swatanab@dis.titech.ac.jp

ABSTRACT

Many learning machines which have hidden variables or hierarchical structures contain singularities in parameter spaces. The likelihood function near singularities can not be approximated by any Gaussian function. At a regular point which is in a neighborhood of a singularity, the non-Gaussian likelihood function becomes to be Gaussian as the number of training samples increases. In many statistical models, a singularity corresponds to a smaller model, hence the transition of the likelihood function from non-Gaussian to Gaussian represents the discovery process of the structure of a true distribution. The conventional information criteria AIC and DIC are made on the assumption that the likelihood function is Gaussian, resulting that they can not be applied to observation of discovery process. Recently, a new information criterion WAIC was devised based on singular learning theory which holds for both Gaussian and non-Gaussian likelihood functions. In this paper, we experimentally compare AIC, DIC with WAIC in discovery process, and show that the generalization loss can be estimated by WAIC but not by either AIC or DIC.

1. INTRODUCTION

Many learning machines which have hidden variables or hierarchical structures are not regular statistical models. In fact, they have a lot of parameters whose Fisher information matrices are not invertible. Such a parameter is called a singularity and a statistical model which contains a singularity is called a singular model. One might think a singular model is special or exceptional, however, in fact, almost all learning machines are singular. Artificial neural networks, normal mixtures, hidden Markov models, Boltzmann machines, and Bayesian networks are main examples. Deep learning of such machines utilizes singularities.

In a regular statistical model, a parameter that represents a smaller model is a regular point of the larger model. On the other hand, in a singular model, it is not. In this paper, we study a singular model and analyze a case when a true parameter is a regular point in the neighborhood of a singularity. Then, if the number of training samples is small, the estimated parameter seems to be singular, whereas, if it becomes large, the true and regular parameter can be distinguished from singularities. We call such statistical event a *discovery phenomenon*.

It is well known that the generalization loss of a regular statistical model can be estimated by AIC[1] or DIC[2, 3]. However, in the discovery phenomenon, neither AIC nor DIC is applicable. Recently, a new information criterion, WAIC[4, 5], was devised based on singular learning theory, and it was proved that WAIC can be used to estimate the generalization loss even if a statistical model is singular or even if a true distribution is not realizable by a statistical model. Researches of applying WAIC to hierarchical Bayesian modelling are reported [6, 7, 8].

In this paper, we experimentally compare AIC, DIC, and WAIC in a discovery phenomenon and quantitatively show that WAIC can estimate changing of the generalization loss, where as either AIC or DIC not.

2. INFORMATION CRITERIA

Let $q(x)$ be a probability density function of $x \in \mathbb{R}^N$, and $X_1, X_2, ..., X_n$ be random variables which are independently taken from $q(x)$. A statistical model and a prior are respectively denoted by $p(x|w)$ and $\varphi(w)$, where w is a parameter contained in $\mathbb{R}^d$. Note that, in this research, we study the case when the parameter space has finite dimension and the training samples are independent and identical. The set of training samples is denoted by

$$X^n = (X_1, X_2, ..., X_n).$$

The posterior distribution is defined by

$$p(w|X^n) = \frac{1}{Z}\varphi(w)\prod_{i=1}^{n} p(X_i|w),$$

where Z is the normalizing constant. The average and variance with respect to the posterior distribution are respectively denoted by $\mathbb{E}_w[\ \ ]$ and $\mathbb{V}_w[\ \ ]$. The predictive distribution is defined by

$$p(x|X^n) = \mathbb{E}_w[p(x|w)].$$

The generalization and training losses are respectively defined by

$$\begin{aligned} G &= -\mathbb{E}_X[\log p(X|X^n)], \\ T &= -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i|X^n). \end{aligned}$$

Both G and T are random variables because they are functions of X^n. The minus log likelihood function is defined

by

$$L_n(w) = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i|w).$$

The entropy and the empirical entropy are respectively defined by

$$\begin{aligned} S &= -\mathbb{E}_X[\log q(X)], \\ S_n &= -\frac{1}{n}\sum_{i=1}^{n} \log q(X_i). \end{aligned}$$

In this paper, we mainly study information criteria which estimate the generalization loss. AIC, DIC, and WAIC are respectively defined by

$$\begin{aligned} \text{AIC} &= T + \frac{d}{n}, \\ \text{DIC} &= L_n(\overline{w}) + 2\{\mathbb{E}_w[L_n(w)] - L_n(\overline{w})\}, \\ \text{WAIC} &= T + \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}_w[\log p(X_i|w)], \end{aligned}$$

where $\overline{w} = \mathbb{E}_w[w]$. If a true distribution is realizable by a statistical model, that is to say, if there exists a parameter w_0 such that $q(x) = p(x|w_0)$, and if Fisher information matrix at w_0 is positive definite, then

$$\begin{aligned} \mathbb{E}[G] &= \frac{d}{2n} + o(\frac{1}{n}), \\ \mathbb{E}[\text{AIC}] &= \mathbb{E}[G] + o(\frac{1}{n}), \\ \mathbb{E}[\text{DIC}] &= \mathbb{E}[G] + o(\frac{1}{n}). \end{aligned}$$

Even if a true distribution is not realizable by or even if the regularity condition is not satsified,

$$\begin{aligned} \mathbb{E}[G] &= \frac{\lambda}{n} + o(\frac{1}{n}), \\ \mathbb{E}[\text{WAIC}] &= \mathbb{E}[G] + O(\frac{1}{n^2}), \end{aligned}$$

where λ is the real log canonical threshold [4]. From the mathematical point of view, information criteria AIC, DIC, and WAIC are all based on asymptotic theory, hence they require the sufficiently large number of training samples. However, both AIC and DIC are based on the assumption that the posterior distribution can be approximated by some normal distribution, whereas WAIC is not. In this paper, we experimentally study whether such different assumptions affect their performance as unbiased estimators of the generalization loss or not.

3. AN EXPERIMENT

In this paper, we study a discovery phenomenon in a normal mixture model. Let $g(x_1, x_2)$ be a normal distribution of $(x_1, x_2) \in \mathbb{R}^2$ whose mean is equal to zero and whose covariance matrix is the identity matrix.

A normal mixture model is defined by

$$\begin{aligned} p(x_1, x_2|w) &= (1-a)g(x_1, x_2) \\ &\quad + ag(x_1 - b_1, x_2 - b_2), \end{aligned}$$

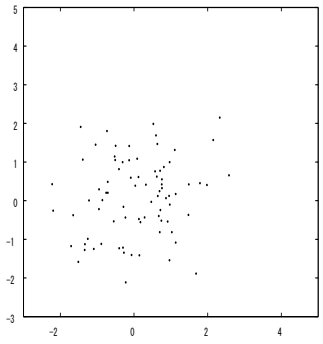

Figure 1. n=80, Training samples

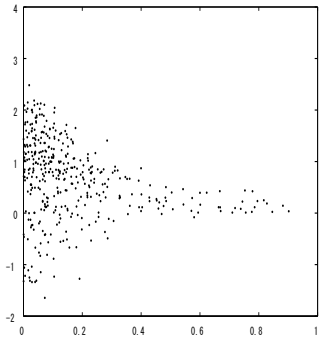

Figure 2. n=80, Posterior distribution

where $w = (a, b_1, b_2)$ is a parameter which satisfies

$$0 \leq a \leq 1, \quad (b_1, b_2) \in \mathbb{R}^2.$$

For a prior distribution, we employ

$$\varphi(a, b_1, b_2) = \frac{1}{2\pi\sigma^2} \exp\Big(-\frac{1}{2\sigma^2}(b_1^2 + b_2^2)\Big),$$

where $\sigma = 10$. If $0 < a \leq 1$ and $b = (b_1, b_2) \neq 0$, then Fisher information matrix at w is positive definite, hence the posterior distribution can be approximated by a normal distribution if the number of training samples is sufficiently large. On the other hand,

$$W_0 = \{(a, b_1, b_2) \ ; \ a = 0 \text{ or } b_1 = b_2 = 0\}.$$

is the set of singularities. In this paper, we study a case when a true parameter is a regular point which is in a neighborhood of W_0,

$$q(x) = p(x|0.5, 0.3, 0.3).$$

In this case, $q(x)$ consists of two normal distributions whose centers are $(0, 0)$ and $(0.3, 0.3)$. They are different distributions, however, almost overlap each other. If the number of training samples is not sufficiently large, then the estimated distribution seems to be made of one normal distribution.

For a given set of training samples, parameters were taken from the posterior distribution using the conventional Metropolis method. The initial parameter was set

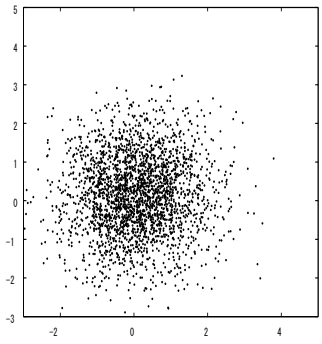

Figure 3. n=2560, Training samples

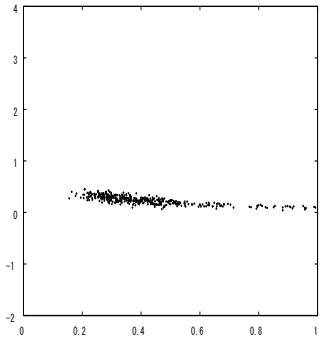

Figure 4. n=2560, Posterior distribution

as the true parameter $(0.5, 0.3, 0.3)$. After the 5000 burn-in sampling steps, 2000 parameters were taken every 25 Metropolis steps. One trial step was given by a normal distribution such that the acceptance probabilities were between 0.1 and 0.8.

Then the set of training samples and the posterior distribution for $n = 80$ are respectively shown in Fig.1 and Fig.2. In Fig.2, parameters are shown on two dimensional space (a, b_1) which are extracted from the three dimensional space (a, b_1, b_2). In the case $n = 80$, the training samples seem to be taken from one normal distribution. The posterior distribution is made of neighborhoods of a set of singularities, W_0.

The set of training samples and posterior distribution for $n = 2560$ are respectively shown in Fig.3 and Fig.4. In this case, the posterior distribution is in a neighborhood of the true parameter $(05, 0.3, 0.3)$. The change of the posterior distribution from singular Fig.2 to regular Fig.4 is called *phase transition* in statistical physics.

In Fig.5, experimental results for AIC, DIC, and WAIC are shown. The horizontal line shows the numbers of training samples,

$$n = 20, 40, 80, 160, 320, 640, 1280, 2560.$$

For each number of training samples, the vertical line shows average values of

—— : $n(\text{AIC} - S_n)$
– o – :$n(\text{WAIC} - S_n)$

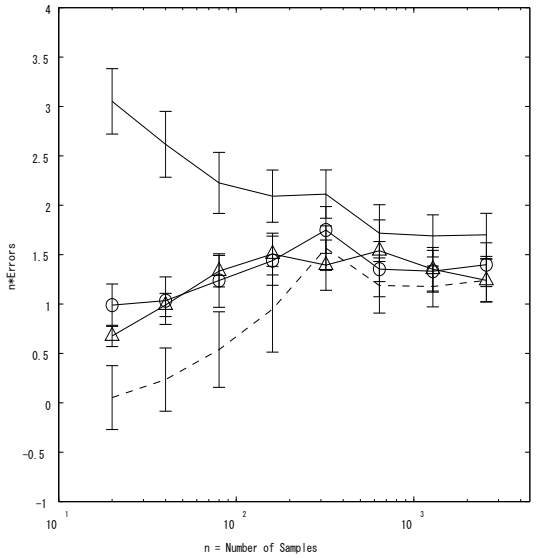

Figure 5. From top to bottom, AIC, WAIC, G, and DIC

–Δ– : $n(G - S)$
- - - : $n(\text{DIC} - S_n)$

are shown in the figure, where average values were calculated using 200 independent sets of training samples. Since the true parameter is a regular point, when n tends to infinity,

$$n\mathbb{E}[\text{AIC} - S_n] \to d/2,$$
$$n\mathbb{E}[\text{WAIC} - S_n] \to d/2,$$
$$n\mathbb{E}[G - S] \to d/2,$$
$$n\mathbb{E}[\text{DIC} - S_n] \to d/2,$$

where $d = 3$ is the dimension of a parameter. If the true parameter would be contained in a set of singularities W_0, then $\lambda = 1/2$, resulting that

$$n\mathbb{E}[\text{AIC} - S_n] \to 1/2,$$
$$n\mathbb{E}[\text{WAIC} - S_n] \to 1/2,$$
$$n\mathbb{E}[G - S] \to 1/2,$$
$$n\mathbb{E}[\text{DIC} - S_n] \to 1/2.$$

For cases $n = 20, 40, 80, 160$, AICs were larger than G, DICs were smaller than G, and WAICs could estimate G. For cases $n = 320, 640, 1280, 2560$, AICs, DICs, and WAICs could estimate G. It seems that a critical point of discovery was between $n = 160$ and 320.

4. DISCUSSION

Let us discuss information criteria in discovery phenomenon from three viewpoints.

Firstly, we compared the ranges of applicability of information criteria. In this experiment, the true distribution is regular for and realizable by a statistical model, although the true parameter is in the neighborhood of singularities. It was experimentally shown that WAIC can be

used for $n \geq 40$, whereas AIC and DIC $n \geq 320$. This result suggests that WAIC is applicable more widely than AIC and DIC. In the more complex and deeper learning machines used in practical applications, parameter regions near singularities have more important roles, where WAIC will be more useful.

Secondly, the leave-one-out cross validation (LOOCV) is an alternative method how to estimate the generalization loss. LOOCV requires n different posterior distributions, which needs huge computational costs. The importance sampling LOOCV (ISLOOCV) was introduced whose computational cost is same as WAIC. It was proved that WAIC, LOOCV, ISLOOCV are asymptotically equivalent to each other [5]. In using ISLOOCV, a divergence problem of the posterior importance was pointed out [9, 6]. It is the future study to compare WAIC, LOOCV, and ISLOOCV.

And thirdly, for the mathematical proofs of WAIC and WBIC, we need the assumption that the dimension of the parameter space is finite and that the training samples are independent and identical. Numerical study for the real log canonical threshold of RBF network was reported [10]. The inifnite dimensional case was studied by using Gaussian process case [11]. It is important to clarify the necesarry and sufficient condition for the use of WAIC and WBIC is the important problem for the future study.

5. CONCLUSION

In a discovery phenomenon, information criteria AIC, DIC, and WAIC were compared as the unbiased estimator of the generalization loss. Neither AIC nor DIC estimated the generalization loss if the true parameter was in the neighborhood of singularities, whereas WAIC did.

Acknowledgment. This research was partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 23500172.

6. REFERENCES

[1] Hirotugu Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[2] D.J.Spiegelhalter, N.G. Best, B.P.Carlin, and A.Linde, "Bayesian measures of model complexity and fit," *Journal of Royal Statistical Society, Series B*, vol. 64, no. 4, pp. 583–639, 2002.

[3] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde, "The deviance information criterion: 12 years on," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2014, DOI: 10.1111/rssb.12062.

[4] Sumio Watanabe, *Algebraic geometry and statistical learning theory*, Cambridge University Press, Cambridge, UK, 2009.

[5] Sumio Watanabe, "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory," *Journal of Machine Learning Research*, vol. 11, pp. 3571–3591, 2010.

[6] Aki Vehtari and Janne Ojanen, "A survey of bayesian predictive methods for model assessment, selection and comparison," *Statistics Surveys*, vol. 6, pp. 142–228, 2012.

[7] A.Gelman, J.B.Carlin, H.S.Stern, D.B. Dunson, A.Vehtari, and D.B.Rubin, *Bayesian Data Analysis, 3rd Edition*, Chapman and Hall/CRC, Boca Raton, 2013.

[8] Andrew Gelman, Jessica Hwang, and Aki Vehtari, "Understanding predictive information criteria for bayesian models," *Statistics and Computing*, 2013, DOI 10.1007/s11222-013-9416-2.

[9] Ilenia Epifani, Steven N. MacEachern, and Mario Peruggia, "Case-deletion importance sampling estimators: Central limit theorems and related results," *Electronic Journal of Statistics*, vol. 2, pp. 774–806, 2008, DOI: 10.1214/08-EJS259.

[10] Satoru Tokuda, Kenji Nagata, and Masato Okada, "A numerical analysis of learning coefficient in radial basis function network," *IPSJ Transactions on Mathematical Modeling and Its Applications*, vol. 6, no. 3, pp. 117–123, 2013.

[11] Tommi Mononen, "A case study of the widely applicable bayesian information criterion and its optimality," *Statistics and Computing*, vol. DOI:10.1007/s11222-014-9463, 2014.

COMPRESSION AND PREDICTION FOR LARGE ALPHABET I.I.D AND MARKOV MODELS

Xiao Yang and Andrew R. Barron

Department of Statistics, Yale University
24 Hillhouse Ave, New Haven, CT, USA
Correspondence: xiao.yang@yale.edu, andrew.barron@yale.edu

ABSTRACT

This paper considers coding and predicting sequences of random variables generated from a large alphabet. We start by proposing a simple coding distribution formulated by a product of tilted Poisson maximized likelihood distributions which achieves close to optimal performance. Then we extend to Markov models, and in particular, tree sources. A context tree based algorithm is designed which seeks for the greatest savings in codelength when constructing the tree.

1. INTRODUCTION

Non-vanishing per symbol redundancy renders large alphabet compression mission impossible. However, distributions living on large alphabets usually display a decaying trend. For example, in Chinese, a subset of 964 characters covers 90% inputs in Chinese[1] though the vocabulary size is more than 100,000 in total.

Coding and prediction of strings of random variables generated from an i.i.d model have been considered for the large alphabet setting with the restriction that the ordered count list rapidly decreasing [3], or satisfies an envelope class property [4][5]. Although this i.i.d model is not the best for compression or prediction when there is dependence between successive characters, it serves as an analytical tool that more complicated models can be based on, and helps understand the behavior of coding and predictive distributions.

Suppose a string of random variables $\underline{X}=(X_1,\ldots,X_N)$ is generated independently from a discrete alphabet $\mathcal{A}$ of size m. Here the string length N can be random. Then $\underline{X}$ is a member of the set $\mathcal{X}^*$ of all finite length strings

$$\mathcal{X}^* = \bigcup_{n=0}^{\infty} \{x^n=(x_1,\ldots,x_n) : x_i \in \mathcal{A}, i=1,\ldots,n\}.$$

Our goal is to code/predict the string $\underline{X}$.

Now suppose given N, each random variable X_i is generated independently according to a probability mass function in a parametric family $\mathcal{P}_\Theta = \{P_{\underline{\theta}}(x) : \underline{\theta} \in \Theta \subset R^m\}$ on $\mathcal{A}$. That is

$$P_{\underline{\theta}}(X_1,\ldots,X_N|N=n) = \prod_{i=1}^{n} P_{\underline{\theta}}(X_i),$$

for $n = 1, 2, \ldots$ We are interested in the class of all distributions with $P_{\underline{\theta}}(j) = \theta_j$ parameterized by the simplex $\Theta=\{\underline{\theta}=(\theta_1,\ldots,\theta_m): \theta_j \geq 0, \sum_{j=1}^m \theta_j = 1, j=1,\ldots,m\}$.

Let $\underline{N} = (N_1,\ldots,N_m)$ denote the vector of counts for symbol $1,\ldots,m$. The observed sample size N is the sum of the counts $N = \sum_{j=1}^m N_j$. Then $P_{\underline{\theta}}(\underline{X})$ have factorizations based on the distribution of the counts

$$P_{\underline{\theta}}(\underline{X}) = P(\underline{X}|\underline{N})\, P_{\underline{\theta}}(\underline{N}).$$

The first factor is the uniform distribution on the set of strings with given counts. The vector of counts $\underline{N}$ forms a sufficient statistic for $\underline{\theta}$. In the particular case of all i.i.d. distributions parameterized by the simplex, the distribution $P_{\underline{\theta}}(\underline{N}|N=n)$ is the $multinomial(n,\underline{\theta})$ distribution.

In the above, there is a need for a distribution of the total count N. Of particular interest is the case that the total count is taken to be $Poisson$, because then the resulting distribution of individual counts are independent.

Poisson sampling is a standard technique to simplify analysis [6][7]. Here we consider the target family $\mathcal{P}_\Lambda^m = \{P_{\underline{\lambda}}(\underline{N}) : \lambda_j \geq 0,\ j=1,\ldots,m\}$, in which $P_{\underline{\lambda}}(\underline{N})$ is the product of $Poisson(\lambda_j)$ distribution for $N_j, j=1,\ldots,m$. It makes the total count $N \sim Poisson(\lambda_{sum})$ with $\lambda_{sum} = \sum_{j=1}^m \lambda_j$ and yields the $multinomial(n,\underline{\theta})$ distribution by conditioning on $N=n$, where $\theta_j = \lambda_j/\lambda_{sum}$. And the induced distribution on $\underline{X}$ is

$$P_{\underline{\lambda}}(\underline{X}) = P(\underline{X}|\underline{N}) P_{\underline{\lambda}}(\underline{N}).$$

Adopting the conventional definition for *regret*, we have

$$R(Q, P_{\hat{\underline{\lambda}}}, \underline{X}) = \log \frac{1}{Q(\underline{X})} - \log \frac{1}{P_{\hat{\underline{\lambda}}}(\underline{X})},$$

where $P_{\hat{\underline{\lambda}}}(\underline{X}) = \max_{\underline{\lambda}\in\Lambda}(P_{\underline{\lambda}}(\underline{X}))$, and $\log$ is logarithm base 2.

Here we can construct Q by choosing a probability distribution for the counts and then use the uniform distribution for the distribution of strings given the counts, written as $P_{unif}(\underline{X}|\underline{N})$. Then the regret becomes

$$R(Q, P_{\hat{\underline{\lambda}}}, \underline{X}) = \log \frac{P_{\hat{\underline{\lambda}}}(\underline{N})}{Q(\underline{N})}.$$

And the problem becomes: given the family $\mathcal{P}_\Lambda^m$, how to choose Q to minimize the maximized regret

$$\min_Q \max_{\underline{X}} R(Q, P_{\hat{\underline{\lambda}}}, \underline{X}) = \min_Q \max_{\underline{N}} \log \frac{P_{\hat{\underline{\lambda}}}(\underline{N})}{Q(\underline{N})}.$$

For the regret, the maximum can be restricted to a set of counts instead of the whole space. A traditional choice being $S_{m,n} = \{(N_1, \ldots, N_m) : \sum_{j=1}^m N_j = n, N_j \geq 0,\ j = 1, \ldots, m\}$ associated with a given sample size n, in which case the minimax regret is

$$\min_Q \max_{\underline{N} \in S_{m,n}} \log \frac{P_{\hat{\underline{\lambda}}}(\underline{N})}{Q(\underline{N})}.$$

As is familiar in universal coding [8][9], the normalized maximum likelihood (NML) distribution

$$Q_{nml}(\underline{N}) = \frac{P_{\hat{\underline{\lambda}}}(\underline{N})}{C(S_{m,n})}$$

is the unique pointwise minimax strategy when $C(S_{m,n}) = \sum_{\underline{N} \in S_{m,n}} P_{\hat{\underline{\lambda}}}(\underline{N})$ is finite, and $\log C(S_{m,n})$ is the minimax value. When m is large, the NML distribution can be unwieldy to compute for compression or prediction. Instead we will introduce a slightly suboptimal coding distribution that makes the counts independent and show that it is nearly optimal for every $S_{m,n'}$ with n' not too different from a target n. Indeed, we advocate that our simple coding distribution is preferable to use computationally when m is large even if the sample size n were known in advance.

To produce our desired coding distribution we make use of two basic principles. One is that the multinomial family of distributions on counts matches the conditional distribution of $N_1, \ldots, N_m$ given the sum N when unconditionally the counts are independent Poisson. Another is the information theory principle [10][11][12] that the conditional distribution given a sum (or average) of a large number of independent random variables is approximately a product distribution, each of which is the one closest in relative entropy to the unconditional distribution subject to an expectation constraint. This minimum relative entropy distribution is an exponential tilting of the unconditional distribution.

In the Poisson family with distribution $\lambda_j^{N_j} e^{-\lambda_j}/N_j!$, exponential tilting (multiplying by the factor e^{-aN_j}) preserves the Poisson family (with the parameter scaled to $\lambda_j e^{-a}$). Those distributions continue to correspond to the multinomial distribution (with parameters $\theta_j = \lambda_j/\lambda_{sum}$) when conditioning on the sum of counts N. A particular choice of $a = \ln(\lambda_{sum}/N)$ provides the product of Poisson distributions closest to the multinomial in regret. Here for universal coding, we find the tilting of individual maximized likelihood that makes the product of such closest to the Shtarkov's NML distribution. This greatly simplifies the task of approximate optimal universal compression and the analysis of its regret.

Indeed, applying the maximum likelihood step to a Poisson count k produces a maximized likelihood value of $M(k) = k^k e^{-k}/k!$. We call this maximized likelihood the *Stirling ratio*, as it is the quantity that Stirling's approximation shows near $(2\pi k)^{-1/2}$ for k not too small. We find that this $M(k)$ plays a distinguished role in universal large alphabet compression, even for sequences with small counts k. Although M has an infinite sum by itself, it is normalizable when tilted for every positive a. The *tilted Stirling ratio distribution* is

$$P_a(N_j) = \frac{N_j^{N_j} e^{-N_j}}{N_j!} \frac{e^{-aN_j}}{C_a}, \tag{1}$$

with the normalizer $C_a = \sum_{k=0}^\infty k^k e^{-(1+a)k}/k!$.

The coding distribution we propose and analyze is simply the product of those tilted one-dimensional maximized Poisson likelihood distributions for a properly chosen a

$$Q_a(\underline{N}) = P_a^m(\underline{N}) = P_a(N_1) \cdots P_a(N_m).$$

If it is known that the total count is n, then the regret is a simple function of n and the normalizer C_a. The choice of the tilting parameter a^* given by the moment condition $\mathbf{E}_{Q_a} \sum_{j=1}^m N_j = n$ minimizes the regret over all positive a. Moreover, value of a^* depends only on the ratio between the size of the alphabet and the total count m/n. Details about finding a^* can be found in [3].

As compared to i.i.d class, Markov sources are richer and more realistic. Suppose given N, each random variable X_i is generated according to a probability mass function depending on its *context* (string of symbols preceding it). Following Willems et al' notations in [13], a tree source can be determined by a context set $\mathcal{S}$. Elements of $\mathcal{S}$ are strings of symbols from $\mathcal{A}$ or concatenation of "others" and suffixes of the existing contexts. "others" represents complements of the contexts in $\mathcal{S}$ with a common *parent*. The collection of distributions is $\mathcal{P}_{\Theta_\mathcal{S}} = \{P_{\underline{\theta}_s}(x) : \underline{\theta}_s \in \Theta_\mathcal{S}, s \in \mathcal{S}\}$, where $\Theta_\mathcal{S}$ is the parameter set defined later. For simplicity, we require the order of the model no larger than $T \in \{0, 1, 2, \ldots\}$, so $\mathcal{S} \in \mathcal{C}_T$, where $\mathcal{C}_T$ is the class of tree sources with order T or less.

For each context $s \in \mathcal{S}$ with a given $\mathcal{S}$, let θ_{sx} denote the probability of symbol $x \in \mathcal{A}$ showing up after s, for all $x \in \mathcal{A}$. Then $\underline{\theta}_{s\cdot} = (\theta_{s1}, \ldots, \theta_{sm})$ lies in the set

$$\Theta_\mathcal{S} = \{\underline{\theta}_{s\cdot} = (\theta_{s1}, \ldots, \theta_{sm}) : x \in \mathcal{A}, \theta_{sx} \geq 0, \sum_{x \in \mathcal{A}} \theta_{sx} = 1\}.$$

Again, we could take advantage of factorizations based on the distribution of the counts $\mathbf{N}_\mathcal{S} = (\underline{N}_{s\cdot})_{s \in \mathcal{S}}$, where $\underline{N}_{s\cdot} = (N_{s1}, \ldots, N_{sm})$ is the count for all symbols given context $s \in \mathcal{S}$, and Pick the distribution for the total count to be $Poisson$. It leads to the target family $\mathcal{P}_\Lambda^{|\mathcal{S}|m} = \{P_{\underline{\lambda}}(\mathbf{N}_\mathcal{S}) : \lambda_{sj} \geq 0,\ j = 1, \ldots, m, s \in \mathcal{S}\}$, in which $P_{\underline{\lambda}}(\mathbf{N}_\mathcal{S})$ is the product of $Poisson(\lambda_{sj})$ distribution for $N_{sj}, j = 1, \ldots, m$ and $s \in \mathcal{S}$.

There are two sources of costs involved in using a tree model. One is the *coding cost* for the string given the tree. The other is the *description cost* $D(\mathcal{S})$ for describing the tree. Overall, we want to find Q which uses shorter

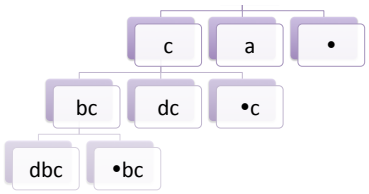

Figure 1. An example context tree with $\mathcal{A} = \{a, b, c, d\}$ where • represents "others".

codelength for sequences generated from an unknown tree source $\mathcal{S} \in \mathcal{C}_T$. That is, to minimize

$$\min_{\mathcal{S}\in\mathcal{C}_T} \left(\log 1/Q(\underline{X}|\mathcal{S}) + D(\mathcal{S})\right).$$

We use the same coding distribution as given in equation (1) for count variables conditional on each given context s. The coding distribution for the counts given s is simply the product

$$Q_{a_s}(\underline{N}_{s\cdot}) = P^m_{a_s}(\underline{N}_{s\cdot}) = P_{a_s}(N_{s1}) \cdots P_{a_s}(N_{sm}), \quad (2)$$

with a properly chosen a_s for each context $s \in \mathcal{S}$. Using the product of tilted distribution P_{a_s} as a coding distribution, the regret is simply a sum of the individual regrets.

To construct the tree, we adopt a method similar to Rissanen's approach in [14]. Using the total codelength to evaluate the performance of different models and coding distributions, we adopt a greedy algorithm to build the context tree with details discussed in Section 3.3. An illustrative example tree is given in Fig.1.

2. I.I.D CLASS

Theorem 1. *The regret of using a product of tilted Sterling Ratio distributions Q_a for a given vector of counts $\underline{N} = (N_1, \ldots, N_m)$ is*

$$R\left(Q_a, \mathcal{P}^m_\Lambda, \underline{N}\right) = aN \log e + m \log C_a.$$

Let $S_{m,n}$ be the set of count vectors with total count n be defined as before, then

$$\max_{\underline{N}\in S_{m,n}} R\left(Q_a, \mathcal{P}^m_\Lambda, \underline{N}\right) = an \log e + m \log C_a. \quad (3)$$

Let a^* be the choice of a satisfying the following moment condition

$$\mathbf{E}_{P_a} \sum_{j=1}^{m} N_j = m\, \mathbf{E}_{P_a} N_1 = n. \quad (4)$$

Then a^* is the minimizer of the regret in expression (3). Write $R_{m,n} = \min_a R(Q_a, \mathcal{P}^m_\Lambda, S_{m,n})$.

When $m = o(n)$, the $R_{m,n}$ is near $\frac{m}{2} \log \frac{ne}{m}$ with

$$\begin{aligned} -d_1 \frac{m}{2} \log e &\le R_{m,n} - \frac{m}{2} \log \frac{ne}{m} \\ &\le m \log(1 + \sqrt{\frac{m}{n}}), \end{aligned}$$

where $d_1 = O\left((\frac{m}{n})^{1/3}\right)$.

When $n = o(m)$, the $R_{m,n}$ is near $n \log \frac{m}{ne}$ as follows.

$$\begin{aligned} m \log\left(1 + (1 - d_2)\frac{n}{m}\right) &\le R_{m,n} - n \log \frac{m}{ne} \\ &\le m \log\left(1 + \frac{n}{m} + d_3\right) \end{aligned}$$

where $d_2 = O(\frac{n}{m})$, and $d_3 = \frac{1}{2\sqrt{\pi}} \frac{n^2 e^2}{m(m-ne)}$.

When $n = bm$, the $R_{m,n} = cm$, where the constant $c = a^* b \log e + \log C_{a^*}$, and a^* is such that $\mathbf{E}_{P_a} N_1 = b$.

Proof. Details of proof can be found in [3]. □

Remark 1: The regret depends only on the number of parameters m, the total counts n and the tilting parameter a. The optimal tilting parameter is given by a simple moment condition in equation (4).

Remark 2: The regret $R_{m,n}$ is close to the minimax level in all three cases listed in Theorem 1. The main terms in the $m = o(n)$ and $n = o(m)$ cases are the same as the minimax regret given in [15] except the multiplier for $\log(ne/m)$ here is $m/2$ instead of $(m-1)/2$ for the small m scenario. For the $n = bm$ case, the $R_{m,n}$ is close to the minimax regret in [15] numerically.

3. TREE SOURCE

3.1. Coding cost

The coding distribution for a given tree is the product of all the $Q_{a_s}(\underline{N}_{s\cdot})$, i.e.

$$Q^{\mathcal{S}}_a(\mathbf{N}_{\mathcal{S}}) = \prod_{s\in\mathcal{S}} Q_{a_s}(N_{s\cdot}).$$

Let $S_{m,n,\mathcal{S}} = \{\mathbf{N}_{\mathcal{S}} : \sum_{s\in\mathcal{S}} \sum_{j=1}^{m} N_{sj} = n, N_{sj} \ge 0,\ j = 1, \ldots, m, s \in \mathcal{S}\}$.

Corollary 1. *Using independent tilted Stirling ratio distribution $Q^{\mathcal{S}}_a$ to code the counts in $S_{m,n,\mathcal{S}}$, the regret equals*

$$\max_{\mathbf{N}_{\mathcal{S}}\in S_{m,n,\mathcal{S}}} R(\mathcal{P}^{|\mathcal{S}|m}_\Lambda, Q^{\mathcal{S}}_a, \mathbf{N}_{\mathcal{S}}) = \sum_{s\in\mathcal{S}} (a_s N_s \log e + m \log C_{a_s}).$$

This can be easily seen by applying the definition.

3.2. Description cost

To describe a given context set $\mathcal{S}$, we use the following rule

$$D(S) = 1 + N_{branches}\left(1 + \log|\mathcal{A}|\right),$$

where $N_{branches}$ is the number of "labeled" branches in the tree. Here "labeled" means having a specified symbol in the alphabet. For instance, $N_{branches} = 5$ in the example tree.

The first bit is used to describe if the model is nondegenerate (i.i.d or Markov). For each branch other than "others", we first use 1 bit to say if it is nondegenerate, and then $\log m$ bits to convey which symbol it is. Our example tree uses $1 + 5(1 + \log 4) = 16$ bits.

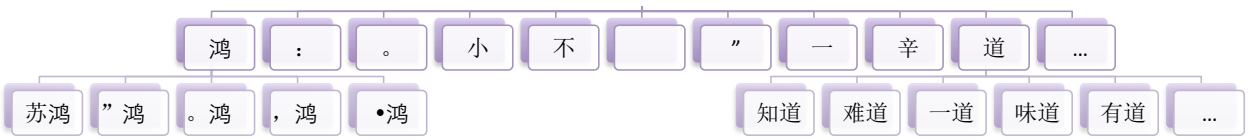

Figure 2. Context tree for *Fortress Beseiged*.

3.3. Using codelength to construct the tree

Here we use the example in Fig.1 to illustrate how we construct the tree. Starting from a null tree (the i.i.d model), we first choose the single symbol (c) that produces the most savings (if any) in codelength. Next, we consider two possible leaves: one is another symbol in the first level (a) that achieves the most savings; the other is to extend to the second level based on the symbols just found. After calculating possible savings produced by these two candidates, we pick again the one with larger savings. Continue in this fashion until no more savings is available or the maximum number of symbols to condition on (T) is reached, the context tree is built. "others" represents contexts with the same parent that are not picked up. It includes b and d in the first level in the example tree.

4. A REAL EXAMPLE

We apply the proposed method to a contemporary Chinese novel translated as Fortress Besieged. The book contains 216,601 characters encoded in GB18030, the largest official Chinese character set which contains 70,244 characters.

The i.i.d model uses 1,954,777 bits. For the tree model, the first single character to condition on saves 12,631. We restrict the order of the Markov model to be no larger than 5, but it turns out no context exceeding two characters shows up. There are 342 branches in the tree, among which 95 are in the first level, and 5 of them extends to the second level. In fact, second level branches are picked up only after most first level ones are chosen. A small part of the tree is displayed in Fig.2. The dots on the right stand for the rest of the model that cannot be shown. And the blank cell in the middle of the first level is the space symbol. The total savings amount to 401,922 bits (about 20.56%) as compared to the i.i.d model.

5. CONCLUSION

We consider a compression and prediction problem under both large alphabet i.i.d model and bounded tree models, and design a method to construct the context tree. Combining this method with tilted Stirling ratio distribution, we have a convenient and efficient way for compression and prediction.

6. REFERENCES

[1] Institue of Applied Linguistice Ministry of Education, "2007 report on language use in china," Nov 2008.

[2] X. Yang and A.R. Barron, "Large alphabet compression and predictive distributions through poissonization and tilting," *arXiv:1401.3760v1 [stat.ME]*, 2013.

[3] A. Garivier S. Boucheron and E. Gassiat, "Coding on countably infinite alphabets," *IEEE Transactions on Information Theory*, vol. 55, no. 1, Jan 2009.

[4] A. Orlitsky J. Acharya, A. Jafarpour and A. T. Suresh, "Poissonization and universal compression of envelope classes," *IEEE International Symposium on Information Theory*, 2014.

[5] William Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, Wiley, 1950.

[6] H. Das J. Archarya, "Tight bounds for universal compression of large alphabets," *IEEE International Symposium on Information Theory*, 2013.

[7] Yu. M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, no. 3, pp. 175–186, 1987.

[8] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Transactions on Information Theory*, vol. 43, pp. 646–657, May 1997.

[9] I. Csiszar, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, Feb 1975.

[10] I. Csiszar, "Sanov property, generalized I-projection and a conditional limit theorem," *The Annals of Probability*, vol. 12, no. 3, pp. 768–793, Jan 1984.

[11] J. V. Campenhout and T.M. Cover, "Maximum entropy and conditional probability," *IEEE Transactions on Information Theory*, vol. 27, no. 4, July 1981.

[12] F. M J Willems, Y.M. Shtarkov, and T.J. Tjalkens, "The context-tree weighting method: basic properties," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[13] J Rissanen, "A Universal Data Compression Syetem," *IEEE Transactions on Information Theory*, vol. 29, no. 5, pp. 656–664, 1983.

[14] W. Szpankowski and M. J. Weinberger, "Minimax redundancy for large alphabets," *Information Theory Proceedings*, June 2010.