Guido F. Montúfar; Johannes Rauh

Scaling of model approximation errors and expected entropy distances

Kybernetika, Vol. 50 (2014), No. 2, 234–245

Persistent URL: http://dml.cz/dmlcz/143791

KYBERNETIKA — VOLUME 50 (2014), NUMBER 2, PAGES 234–245

SCALING OF MODEL APPROXIMATION ERRORS AND EXPECTED ENTROPY DISTANCES

Guido F. Montúfar and Johannes Rauh

We compute the expected value of the Kullback–Leibler divergence of various fundamental statistical models with respect to Dirichlet priors. For the uniform prior, the expected divergence of any model containing the uniform distribution is bounded by a constant $1-\gamma$. For the models that we consider this bound is approached as the cardinality of the sample space tends to infinity, if the model dimension remains relatively small. For Dirichlet priors with reasonable concentration parameters the expected values of the divergence behave in a similar way. These results serve as a reference to rank the approximation capabilities of other statistical models.

Keywords: exponential families, KL divergence, MLE, Dirichlet prior

Classification: 62F25, 68T30

1. INTRODUCTION

Consider a finite set $\mathcal{X}$ of cardinality $|\mathcal{X}| = N$. The set $\Delta = \Delta_{N-1}$ of probability distributions on $\mathcal{X}$ can be identified with an $(N-1)$-simplex. A *(statistical) model* is any subset $\mathcal{M} \subseteq \Delta$. Given a distribution $p \in \Delta$, it is an important problem of statistics and machine learning to find the best approximation q within a model $\mathcal{M}$.

To quantify how good this approximation is, a natural choice is to use the *information divergence*, *relative entropy*, or *Kullback–Leibler divergence* from p to q, defined by

$$D(p\|q) := \sum_{i\in\mathcal{X}} p_i \ln \frac{p_i}{q_i}.$$

If p is an empirical distribution summarizing the outcome of n statistical experiments, then the log-likelihood of q equals $-n(D(p\|q) + H(p))$, where $H(p)$ is the Shannon entropy of p. Hence finding a *maximum likelihood estimate* q within a model $\mathcal{M}$ is the same as finding a minimizer of the divergence $D(p\|q)$ with q restricted to $\mathcal{M}$.

To assess the expressive power of a model $\mathcal{M}$, we study the function $p \mapsto D(p\|\mathcal{M}) = \inf_{q\in\mathcal{M}} D(p\|q)$. Finding the maximizers of this function corresponds to a worst-case analysis. This problem was first posed in [1] motivated by infomax principles in the context of neural networks. The case of exponential families has been studied in detail [4,

DOI: 10.14736/kyb-2014-2-0234

5, 8], and recently also discrete mixture models and restricted Boltzmann machines have been considered [6].

In addition to the worst-case error, the *expected error* is of interest. This leads to the mathematical problem of computing the expectation value

$$\langle D(p\|\mathcal{M})\rangle = \int_\Delta D(p\|\mathcal{M})\,\psi(p)\,\mathrm{d}p\,, \tag{1}$$

where p is drawn from a *prior* probability density ψ on the probability simplex Δ. We focus on Dirichlet priors. Our analysis leads to integrals that have been studied in the framework of Bayesian function estimation in [10], and we can take advantage of the tools developed there. It turns out that for many model classes the worst-case error diverges as the number of elementary events $N = |\mathcal{X}|$ tends to infinity, whereas the expected error remains bounded.

Our first observation is that, if ψ is the uniform prior, then the expected divergence from p to the uniform distribution is a monotone function of the system size N and converges to the constant $1-\gamma \approx 0.4228$ as $N \to \infty$, where γ is the *Euler–Mascheroni* constant. Many natural statistical models contain the uniform distribution and their expected divergence is bounded by the same constant. On the other hand, when p and q are chosen uniformly at random, the expected divergence $\langle D(p\|q)\rangle_{p,q}$ is equal to $1-1/N$.

We show that the expected divergence of a class of models including independence models, partition models, mixtures of product distributions with disjoint supports [6], and decomposable hierarchical models has the same limit, $1-\gamma$, provided the dimension of the models remains *small* with respect to N (the usual case in applications). For Dirichlet priors the results are similar (for reasonable choices of parameters). In contrast, as shown in [9], if $\mathcal{M}$ is an exponential family, then the maximum value of $D(\cdot\|\mathcal{M})$ is at least $\ln(N/(\dim(\mathcal{M})+1))$.

In Section 2 we define various models and collect basic properties of Dirichlet priors. Section 3 contains our main results: closed-form expressions for the expectation values of entropies and divergences. A discussion is given in Section 4.

2. PRELIMINARIES

2.1. Models from statistics and machine learning

As mentioned above, a *model* is any subset of the probability simplex Δ_{N-1}. The *support sets* of a model $\mathcal{M} \subseteq \Delta_{N-1}$ are the sets $\mathrm{supp}(p) = \{i \in \mathcal{X} \,|\, p_i > 0\}$ for all $p = (p_i)_{i\in\mathcal{X}}$ in $\mathcal{M}$.

The *K-mixture* of a model $\mathcal{M}$ is the union of all convex combinations of any K of its points: $\mathcal{M}^K := \{\sum_{k=1}^K \lambda_k p^{(k)} \,|\, \lambda_k \geq 0, \sum_k \lambda_k = 1, p^{(k)} \in \mathcal{M}\}$. Given a partition $\varrho = \{A_1, \ldots, A_K\}$ of $\mathcal{X}$ into K support sets of $\mathcal{M}$, the *K-mixture of $\mathcal{M}$ with disjoint supports ϱ* is the subset of $\mathcal{M}^K$ defined by

$$\mathcal{M}^\varrho = \left\{ \sum_{k=1}^K \lambda_k p^{(k)} \in \mathcal{M}^K \,\middle|\, p^{(k)} \in \mathcal{M},\ \mathrm{supp}(p^{(k)}) \subseteq A_k \text{ for all } k \right\}.$$

Let $\varrho = \{A_1, \dots, A_K\}$ be a partition of $\mathcal{X}$. The *partition model* $\mathcal{M}_\varrho$ consists of all $p \in \Delta_{N-1}$ that satisfy $p_i = p_j$ whenever i, j belong to the same block in the partition ϱ. Partition models are closures of convex exponential families with uniform reference measures. More generally, the closure of a convex exponential family is a set of the form (see [4])

$$\mathcal{M}_{\varrho,\nu} = \left\{ \sum_{k=1}^{K} \lambda_k \frac{\mathbb{1}_{A_k}\nu}{\nu(A_k)} \;\middle|\; \lambda_k \geq 0,\ \sum_{k=1}^{K} \lambda_k = 1 \right\},$$

where $\nu : \mathcal{X} \to (0, \infty)$ is a positive function on $\mathcal{X}$ called *reference measure*, and $\mathbb{1}_A$ is the indicator function of A. Note that all measures ν with fixed conditional distributions $\nu(\cdot|A_k) = \nu(\cdot)/\sum_{j\in A_k} \nu(j)$ on A_k, for all k, yield the same model. In fact, $\mathcal{M}_{\varrho,\nu}$ is the K-mixture of the set $\{\nu(\cdot|A_k) \,|\, k = 1, \dots, K\}$.

For a composite system with n variables $X_1, \dots, X_n$, the set of elementary events is $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, $|\mathcal{X}_i| = N_i$ for all i. A *product distribution* is a distribution of the form

$$p(x_1, \dots, x_n) = p_{\{1\}}(x_1) \cdots p_{\{n\}}(x_n) \qquad \text{for all } x \in \mathcal{X},$$

where $p_{\{i\}} \in \Delta_{N_i - 1}$. The *independence model* $\mathcal{M}_1$ is the set of all product distributions on $\mathcal{X}$. The support sets of the independence model are the sets of the form $A = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$ with $\mathcal{Y}_i \subseteq \mathcal{X}_i$ for each i.

Let $\mathcal{S}$ be a simplicial complex on $\{1, \dots, n\}$. The *hierarchical model* $\mathcal{M}_\mathcal{S}$ consists of all probability distributions that have a factorization of the form $p(x) = \prod_{S\in\mathcal{S}} \Phi_S(x)$, where Φ_S is a positive function that depends only on the S-coordinates of x. The model $\mathcal{M}_\mathcal{S}$ is called *reducible* if there exist simplicial subcomplexes $\mathcal{S}_1, \mathcal{S}_2 \subsetneq \mathcal{S}$ such that $\mathcal{S}_1 \cup \mathcal{S}_2 = \mathcal{S}$ and $\mathcal{S}_1 \cap \mathcal{S}_2$ is a simplex. In this case, the set $(\bigcup_{\mathcal{Y}\in\mathcal{S}_1} \mathcal{Y}) \cap (\bigcup_{\mathcal{Y}\in\mathcal{S}_2} \mathcal{Y})$ is called a *separator*. Furthermore, $\mathcal{M}_\mathcal{S}$ is *decomposable* if it can be iteratively reduced to simplices. Such an iterative reduction can be described by a *junction tree*, which is a tree (V, E) with vertex set the set of facets of $\mathcal{S}$ and with edge labels the separators. The independence model is an example of a decomposable model (where all separators are empty sets). We give another example in Fig. 1, and refer to [2] for more details. In general, the junction tree is not unique, but the multi-set of separators is unique.

For most models there is no closed-form expression for $D(\cdot\|\mathcal{M})$, since there is no closed formula for $\operatorname{arginf}_{q\in\mathcal{M}} D(p\|q)$. However, for some of the models mentioned above a closed formula does exist: The divergence of the independence model is called *multi-information* and satisfies

$$MI(X_1, \dots, X_n) = D(p\|\mathcal{M}_1) = -H(X_1, \dots, X_n) + \sum_{k=1}^{n} H(X_k). \tag{2}$$

If $n = 2$ it is also called the *mutual information* of X_1 and X_2. The divergence of the convex exponential family $\mathcal{M}_{\varrho,\nu}$ is given by (see [4, eq. (1)])

$$D(p\|\mathcal{M}_{\varrho,\nu}) = D\left(p \,\middle\|\, \sum_{k=1}^{K} p(A_k)\nu(x|A_k)\right). \tag{3}$$

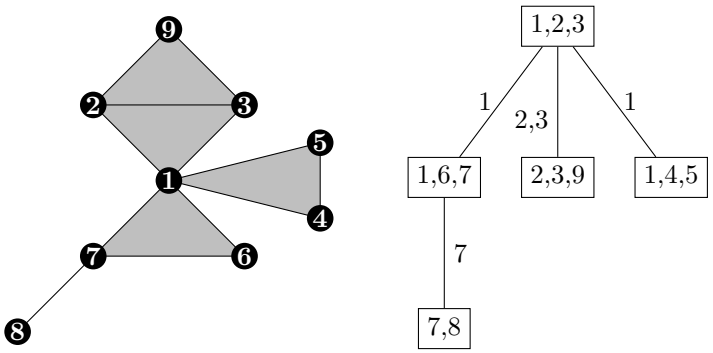

Fig. 1. An example of a decomposable model and its junction tree.

For a decomposable model $\mathcal{M}_{\mathcal{S}}$ with junction tree (V, E),

$$D(p\|\mathcal{M}_{\mathcal{S}}) = \sum_{S\in V} H_p(X_S) - \sum_{S\in E} H_p(X_S) - H_p(X), \tag{4}$$

where $H_p(X_S)$ denotes the joint entropy of the variables $\{X_i\}_{i\in S}$ under p.

2.2. Dirichlet priors

The Dirichlet prior with *concentration parameter* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N_{>0}$ is the probability distribution on Δ_{N-1} with density

$$\mathrm{Dir}_{\boldsymbol{\alpha}}(p) := \frac{1}{\sqrt{N}} \frac{\Gamma(\sum_{i=1}^N \alpha_i)}{\prod_{i=1}^N \Gamma(\alpha_i)} \prod_{i=1}^N p_i^{\alpha_i - 1} \quad \text{for all } p = (p_1, \ldots, p_N) \in \Delta_{N-1}, \tag{5}$$

where Γ is the gamma function. In particular, $\mathrm{Dir}_{(1,\ldots,1)}$ is the uniform probability density on Δ_{N-1}. Moreover, $\lim_{a\to 0} \mathrm{Dir}_{(a,\ldots,a)}$ assigns mass $1/N$ to δ_x for all $x \in \mathcal{X}$, and $\lim_{a\to\infty} \mathrm{Dir}_{(a,\ldots,a)}$ is concentrated at the uniform distribution $u := (1/N, \ldots, 1/N)$. For an arbitrary concentration parameter $\boldsymbol{\alpha}$ let $\alpha = \sum_{i=1}^N \alpha_i$. Then $\boldsymbol{\alpha}/\alpha \in \Delta_{N-1}$, and $\lim_{\kappa\to\infty} \mathrm{Dir}_{\kappa\boldsymbol{\alpha}}$ is the Dirac distribution concentrated at $\boldsymbol{\alpha}/\alpha$.

The Dirichlet distributions satisfy the following *aggregation property* (see, e.g., [3]): If $p = (p_1, \ldots, p_N) \sim \mathrm{Dir}_{(\alpha_1,\ldots,\alpha_N)}$ and $\varrho = \{A_1, \ldots, A_K\}$ is a partition of $\mathcal{X}$, then $(\sum_{i\in A_1} p_i, \ldots, \sum_{i\in A_K} p_i) \sim \mathrm{Dir}_{(\sum_{i\in A_1}\alpha_i, \ldots, \sum_{i\in A_K}\alpha_i)}$. We write $\boldsymbol{\alpha}^\varrho = (\alpha_1^\varrho, \ldots, \alpha_K^\varrho)$, $\alpha_k^\varrho = \sum_{i\in A_k} \alpha_i$ for the concentration parameter induced by the partition ϱ. The aggregation property is useful when treating marginals of composite systems. Given a composite system with $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, $|\mathcal{X}| = N$, $\mathcal{X}_k = \{1, \ldots, N_k\}$, we write $\boldsymbol{\alpha}^k = (\alpha_1^k, \ldots, \alpha_{N_k}^k)$, $\alpha_j^k = \sum_{x\in\mathcal{X}:\, x_k=j} \alpha_x$ for the concentration parameter of the Dirichlet distribution induced on the $\mathcal{X}_k$-marginal $(\sum_{x\in\mathcal{X}:\, x_k=1} p(x), \ldots, \sum_{x\in\mathcal{X}:\, x_k=N_k} p(x))$.

We will use the following Lemma 1 to evaluate the integral (1) in the case of Dirichlet priors, see [10, Theorem 5]. For any $k \in \mathbb{N}$ let $h(k) = 1 + \frac{1}{2} + \cdots + \frac{1}{k}$ be the kth *harmonic*

number. It is known that for large k,

$$h(k) = \ln(k) + \gamma + O(\tfrac{1}{k}), \tag{6}$$

where $\gamma \approx 0.57721$ is the *Euler–Mascheroni constant*. Moreover, $h(k) - \ln(k)$ is strictly positive and decreases monotonically. We also need the natural analytic extension of h to the non-negative reals, given by $h(z) = \frac{\partial}{\partial z} \ln(\Gamma(z+1)) + \gamma$, where Γ is the gamma function.

Lemma 1. Let $\rho = \{A_1, \dots, A_K\}$ be a partition of $\mathcal{X} = \{1, \dots, N\}$, and let $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)$ be a vector of positive real numbers. Then

$$\int_{\Delta_{N-1}} \left(\sum_{i \in A_{k'}} p_i \right) \ln \left(\sum_{i \in A_{k'}} p_i \right) \prod_{i=1}^{N} p_i^{\alpha_i - 1} \, \mathrm{d}p = \int_{\Delta_{K-1}} p^*_{k'} \ln(p^*_{k'}) \prod_{k=1}^{K} (p^*_k)^{\alpha^\rho_k - 1} \, \mathrm{d}p^*$$
$$= \sqrt{N} \, \frac{\alpha^\rho_{k'} \prod_{k=1}^{K} \Gamma(\alpha^\rho_k)}{\Gamma(\alpha + 1)} \left(h(\alpha^\rho_{k'}) - h(\alpha) \right).$$

3. EXPECTED ENTROPIES AND DIVERGENCES

The following theorems contain formulas for the expectation value of the divergence from the models defined in the previous section and asymptotic expressions of these formulas. The results are based on explicit solutions of the integral (1) as derived by Wolpert and Wolf [10]. Recall that $h(z)$ denotes the analytic extension of the harmonic numbers, see eq. (6).

Theorem 1. If $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$, then:

- $\langle H(p) \rangle = h(\alpha) - \sum_{i=1}^{N} \frac{\alpha_i}{\alpha} h(\alpha_i)$,
- $\langle D(p \| u) \rangle = \ln(N) - h(\alpha) + \sum_{i=1}^{N} \frac{\alpha_i}{\alpha} h(\alpha_i)$,

where $\alpha = \sum_{i=1}^{N} \alpha_i$. In the symmetric case $(\alpha_1, \dots, \alpha_N) = (a, \dots, a)$,

- $\langle H(p) \rangle = h(Na) - h(a)$
$$= \begin{cases} \ln(Na) - h(a) + \gamma + O(1/Na) & \text{for large } N \text{ and const. } a \\ \ln(N) + O(1/a) & \text{for large } a \text{ and arb. } N \\ O(Na) & \text{as } a \to 0 \text{ with bounded } N \\ h(c) + O(a) & \text{as } a \to 0 \text{ with } Na = c, \end{cases}$$
- $\langle D(p \| u) \rangle = \ln(N) - h(Na) + h(a)$
$$= \begin{cases} h(a) - \ln(a) - \gamma + O(1/Na) & \text{for large } N \text{ and const. } a \\ O(1/a) & \text{for large } a \text{ and arb. } N \\ \ln(N) + O(Na) & \text{as } a \to 0 \text{ with bounded } N \\ \ln(N) - h(c) + O(a) & \text{as } a \to 0 \text{ with } Na = c. \end{cases}$$

P r o o f . The analytic formulas are [10, Theorem 7]. The asymptotic expansions are direct. □

The entropy $H(p) = -\sum_i p_i \ln p_i$ is maximized at the uniform distribution u, with $H(u) = \ln(N)$. For large N or large a, the average entropy is close to this maximum value and the expected divergence from the uniform distribution u is bounded. The fact that the expected entropy is close to the maximal value makes entropy estimation difficult. See [7] for a discussion and possible solutions.

Theorem 2a. If $q \in \Delta_{N-1}$ and $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$, then

$$\langle D(p\|q)\rangle_p = \sum_{i=1}^{N} \frac{\alpha_i}{\alpha}(h(\alpha_i) - \ln(q_i)) - h(\alpha) = D(\tfrac{\boldsymbol{\alpha}}{\alpha}\|q) + O(N/\alpha).$$

If $\boldsymbol{\alpha} = (a, \dots, a)$, then

$$\begin{aligned}\langle D(p\|q)\rangle_p &= D(u\|q) + h(a) + \ln(N) - h(Na)\\ &= D(u\|q) + (h(a) - \ln(a)) - \gamma + O(1/(Na)).\end{aligned}$$

Theorem 2b. If $p \in \Delta_{N-1}$ and $q \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$, then

$$\langle D(p\|q)\rangle_q = \sum_{i=1}^{N} p_i(\ln(p_i) - h(\alpha_i - 1)) + h(\alpha - 1).$$

If $\alpha_i > 1$ for all i, then

$$\langle D(p\|q)\rangle_q = D(p\|\tfrac{\boldsymbol{\alpha}}{\alpha}) + \sum_{i=1}^{N} O(1/(\alpha_i - 1)).$$

Theorem 2c. If $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$ and $q \sim \mathrm{Dir}_{\tilde{\boldsymbol{\alpha}}}$, then

- $\langle \sum_{i\in\mathcal{X}} p_i \ln(q_i)\rangle_{p,q} = \sum_{i=1}^{N} \frac{\alpha_i}{\alpha} h(\tilde{\alpha}_i - 1) - h(\tilde{\alpha} - 1)$,
- $\langle D(p\|q)\rangle_{p,q} = -\sum_{i=1}^{N} \frac{\alpha_i}{\alpha}(h(\tilde{\alpha}_i - 1) - h(\alpha_i)) + h(\tilde{\alpha} - 1) - h(\alpha)$.

If $\boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}$, then $\langle D(p\|q)\rangle_{p,q} = \frac{N-1}{\alpha}$.

P r o o f . Theorem 2a follows from $D(p\|q) = -H(p) - \sum_{i=1}^{N} p_i \ln(q_i)$, Theorem 1, and Lemma 1. Similarly, Theorems 2b and 2c follow from Lemma 1 by which

$$\int_{\Delta_{N-1}} \ln(p_i) \prod_{j=1}^{N} p_j^{\alpha_j - 1}\,\mathrm{d}p \Big/ \int_{\Delta_{N-1}} \prod_{j=1}^{N} p_j^{\alpha_j - 1}\,\mathrm{d}p = h(\alpha_i - 1) - h(\alpha - 1).$$ □

Consider a sequence of distributions $q^{(N)} \in \Delta_{N-1}$, $N \in \mathbb{N}$. For uniform priors, as $N \to \infty$ the expected divergence $\langle D(p\|q^{(N)})\rangle_p$ is bounded from above by $1 - \gamma + c$, $c > 0$ if and only if $\limsup_{N\to\infty} D(u\|q^{(N)}) \le c$. It is easy to see that $D(u\|q) \le c$ when

$q_x \geq \frac{1}{N} e^{-c}$ for all $x \in \mathcal{X}$. Therefore, the expected divergence is unbounded as N tends to infinity only if the sequence $q^{(N)}$ accumulates at the boundary of the probability simplex. In fact, $\lim_{N\to\infty} \langle D(p\|q^{(N)})\rangle_p \leq 1 - \gamma + c$ whenever $q^{(N)}$ is in the subsimplex $\Delta^c_{N-1} = \operatorname{conv}\{(1-e^{-c})\delta_x + e^{-c}u\}_{x\in\mathcal{X}}$ for all N. For any given N the Lebesgue volume of this subsimplex satisfies $\operatorname{vol}\Delta^c_{N-1}/\operatorname{vol}\Delta_{N-1} = (1-e^{-c})^{N-1}$.

For arbitrary Dirichlet priors with concentration parameters $\boldsymbol{\alpha}^{(N)}$ depending on N, the expectation value $\langle D(p\|q^{(N)})\rangle_p$ remains bounded in the limit $N \to \infty$, provided $D(\frac{\boldsymbol{\alpha}^{(N)}}{\alpha^{(N)}}\|q^{(N)})$ does and $\alpha_i^{(N)}$ remains bounded from below by a positive constant.

If $p, q \sim \operatorname{Dir}_{\boldsymbol{\alpha}^{(N)}}$, then $\langle D(p\|q)\rangle_{p,q}$ remains bounded in the limit $N \to \infty$, provided $\frac{\alpha^{(N)}}{N}$ remains bounded from below by a positive constant.

Theorem 3. Consider a system of n random variables $X_1, \ldots, X_n$ with joint probability distribution p. If $p \sim \operatorname{Dir}_{\boldsymbol{\alpha}}$, then

- $\langle H(X_k)\rangle = h(\alpha) - \sum_{j=1}^{N_k} \frac{\alpha_j^k}{\alpha} h(\alpha_j^k)$,
- $\langle MI(X_1, \ldots, X_n)\rangle = (n-1)h(\alpha) + \sum_{i=1}^{N} \frac{\alpha_i}{\alpha} h(\alpha_i) - \sum_{k=1}^{n}\sum_{j=1}^{N_k} \frac{\alpha_j^k}{\alpha} h(\alpha_j^k)$.

If $\boldsymbol{\alpha} = (a, \ldots, a)$,

- $\langle H(X_k)\rangle = h(Na) - h(\frac{N}{N_k}a)$,
- $\langle MI(X_1, \ldots, X_n)\rangle = (n-1)h(Na) + h(a) - \sum_{k=1}^{n} h(\frac{N}{N_k}a)$.

If, moreover, Na/N_k is large for all k (e.g., a remains bounded from below by some $\varepsilon > 0$ and *(i)* all N_k become large, or *(ii)* all N_k remain bounded and n becomes large), then

- $\langle H(X_k)\rangle = \ln(N_k) + O(N_k/Na)$,
- $\langle MI(X_1, \ldots, X_n)\rangle = h(a) - \ln(a) - \gamma + O(n \max_k N_k/Na)$.

Proof. This is a corollary to Theorem 1, the aggregation property of the Dirichlet priors, and the formula (2) for the multi-information. □

If Na/N_k is large for all k, then the expected entropy of a subsystem is also close to its maximum, and hence the expected multi-information is bounded. This follows also from the fact that the independence model contains the uniform distribution, and hence $D(p\|\mathcal{M}_1) \leq D(p\|u)$.

Theorem 4. Let $\varrho = \{A_1, \ldots, A_K\}$ be a partition of $\mathcal{X}$ into sets of cardinalities $|A_k| = L_k$, and let ν be a reference measure on $\mathcal{X}$. If $p \sim \operatorname{Dir}_{\boldsymbol{\alpha}}$, then

$$\langle D(p\|\mathcal{M}_{\varrho,\nu})\rangle = \sum_{i=1}^{N} \frac{\alpha_i}{\alpha}\left((h(\alpha_i) - \ln(\nu_i)) - \sum_{k=1}^{K} \frac{\alpha_k^{\varrho}}{\alpha}\left(h(\alpha_k^{\varrho}) - \ln(\nu(A_k))\right)\right),$$

where $\alpha_k^\varrho = \sum_{i\in A_k} \alpha_i$. If $\boldsymbol{\alpha} = (a, \ldots, a)$, and (wlog) $\nu(A_k) = L_k/N$,

$$\langle D(p\|\mathcal{M}_{\varrho,\nu})\rangle = h(a) - \sum_{k=1}^{K} \frac{L_k}{N}(h(L_k a) - \ln(L_k)) + D(u\|\nu).$$

If, moreover, $N \gg K$, then

$$\langle D(p\|\mathcal{M}_{\varrho,\nu})\rangle = h(a) - \ln(a) - \gamma + D(u\|\nu) + O(1/N).$$

P r o o f . This follows from eq. (3). □

If the reference measure ν is uniform, then $\mathcal{M}_{\varrho,\nu}$ is a partition model and contains the uniform distribution. In these cases the expected divergence is bounded. In contrast, the maximal divergence is $\max_{p\in\Delta_{N-1}} D(p\|\mathcal{M}_\varrho) = \max_k \ln(N_k)$.

For mixtures of products with disjoint supports the result is similar:

Theorem 5. Consider a composite system of n random variables with state space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, $|\mathcal{X}| = N$, $|\mathcal{X}_k| = N_k$. Let $\varrho = \{A_1, \ldots, A_K\}$ be a partition of $\mathcal{X}$ into K support sets $A_k = \mathcal{X}_{1,k} \times \cdots \times \mathcal{X}_{n,k}$, $k = 1, \ldots, K$ of the independence model, and let $\mathcal{M}_1^\varrho$ be the model containing all mixtures of K product distributions $p^{(1)}, \ldots, p^{(K)}$ with $\mathrm{supp}(p^{(k)}) \subseteq A_k$.

- If $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$, then

$$\langle D(p\|\mathcal{M}_1^\varrho)\rangle = \sum_{i=1}^{N} \frac{\alpha_i}{\alpha}(h(\alpha_i) - h(\alpha)) + \sum_{k=1}^{K}(|G_k| - 1)\frac{\alpha_k^\varrho}{\alpha}(h(\alpha_k^\varrho) - h(\alpha)) - \sum_{k=1}^{K}\sum_{j\in G_k}\sum_{x_j\in\mathcal{X}_{j,k}} \frac{\alpha^{k,x_j}}{\alpha}(h(\alpha^{k,x_j}) - h(\alpha)),$$

  where $\alpha_k^\varrho = \sum_{x\in A_k} \alpha_x$, $\alpha^{k,x_j} = \sum_{y\in A_k\colon y_j = x_j} \alpha_y$, and $G_k \subseteq \{1, \ldots, n\}$ is the set of variables that take more than one value in the block A_k.

- If the system is homogeneous, $|\mathcal{X}_i| = N_1$ for all i, A_k is a cylinder set of cardinality $|A_k| = N_1^{m_k}$, $m_k = |G_k|$ for all k, and $(\alpha_1, \ldots, \alpha_N) = (a, \ldots, a)$, then

$$\langle D(p\|\mathcal{M}_1^\varrho)\rangle = h(a) + \sum_{k=1}^{K} N_1^{m_k - n}((m_k - 1)h(N_1^{m_k} a) - m_k h(N_1^{m_k - 1} a)).$$

- If $\frac{N_1^{m_k-1}a}{m_k}$ is large for all k, then

$$\langle D(p\|\mathcal{M}_1^\varrho)\rangle = h(a) - \ln(a) - \gamma + O\big(\max_k \frac{m_k}{N_1^{m_k-1}a}\big).$$

Proof. The unique solution $q \in \operatorname{arginf}_{q' \in M_1^e} D(p\|q')$ satisfies $p(A_k) = q(A_k)$ and $q(\cdot|A_k) \in \operatorname{arginf}_{q' \in \mathcal{M}_1} D(p(\cdot|A_k)\|q')$ (see [6]). This implies

$$D(p\|\mathcal{M}_1^e) = \sum_{i=1}^{K} \sum_{x \in A_i} p(x) \ln \frac{p(x)p(A_i)^{n-1}}{\prod_{j=1}^{n} (\sum_{y \in A_i : y_j = x_j} p(y))}.$$

□

The k-mixture of binary product distributions with disjoint supports is a submodel of the restricted Boltzmann machine model with $k-1$ hidden units, as shown in [6]. Hence Theorem 5 gives bounds for the expected divergence of restricted Boltzmann machines.

Theorem 6. Consider a decomposable model $\mathcal{M}_{\mathcal{S}}$ with junction tree (V, E).

- If $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$, then

$$\langle D(p\|\mathcal{M}_{\mathcal{S}})\rangle = -\sum_{S \in V} \sum_{j \in \mathcal{X}_S} \frac{\alpha_j^S}{\alpha} h(\alpha_j^S) + \sum_{S \in E} \sum_{j \in \mathcal{X}_S} \frac{\alpha_j^S}{\alpha} h(\alpha_j^S) + \sum_{i=1}^{N} \frac{\alpha_i}{\alpha} h(\alpha_i),$$

  where $\alpha_j^S = \sum_{x\colon x_S = j} \alpha_x$ for $j \in \mathcal{X}_S$.

- If p is drawn uniformly at random, then

$$\langle D(p\|\mathcal{M}_{\mathcal{S}})\rangle = 1 - \sum_{S \in V} h(N/N_S) + \sum_{S \in E} h(N/N_S).$$

- If N/N_S is large for all $S \in V \cup E$, then

$$\langle D(p\|\mathcal{M}_{\mathcal{S}})\rangle = 1 - \gamma + O\big(\max_S N/N_S\big).$$

Proof. This follows from eq. (4) and $|V| - |E| - 1 = 0$. □

4. DISCUSSION

We have shown that the values of $\langle D(p\|\mathcal{M})\rangle$ are very similar for different models $\mathcal{M}$ in the limit of large N, provided the Dirichlet concentration parameters α_i remain bounded and the model dimension remains relatively small. In particular, if $\alpha_i = 1$ for all i, then $\langle D(p\|\mathcal{M})\rangle \approx 1 - \gamma$ for large N holds for $\mathcal{M} = \{u\}$, independence models, decomposable models, partition models, and mixtures of product distributions on disjoint supports (for reasonable values of the hyperparameters N_k and L_k). Some of these models are contained in each other, but the expected divergences do not differ much. The general phenomenon seems to be the following:

- If N is large and if $\mathcal{M} \subset \Delta_{N-1}$ is low-dimensional, then the expected divergence is $\langle D(p\|\mathcal{M})\rangle \approx 1 - \gamma$, when p is uniformly distributed on Δ_{N-1}.

Of course, this is not a mathematical statement, because it is easy to construct counter-examples: Space-filling curves can be used to construct one-dimensional models with an arbitrarily low value of $\langle D(p\|\mathcal{M})\rangle$ (for arbitrary N). However, we expect that the statement is true for most models that appear in practice. In particular, we conjecture that the statement is true for restricted Boltzmann machines.

In Theorem 4, if $\alpha = (a, \ldots, a)$, then the expected divergence from a convex exponential family $\mathcal{M}_{\varrho,\nu}$ is minimal if and only if $\nu \propto u$. In this case $\mathcal{M}_{\varrho,\nu}$ is a partition model. We conjecture that partition models are optimal among all (closures of) exponential families in the following sense:

- For any exponential family $\mathcal{E}$ there is a partition model $\mathcal{M}$ of the same dimension such that $\langle D(p\|\mathcal{E})\rangle \geq \langle D(p\|\mathcal{M})\rangle$, when $p \sim \mathrm{Dir}_{(a,\ldots,a)}$.

The statement is of course true for any zero-dimensional exponential family, which consists of a single distribution. The conjecture is related to the following conjecture from [9]:

- For any exponential family $\mathcal{E}$ there is a partition model $\mathcal{M}$ of the same dimension such that $\max_{p\in\Delta_{N-1}} D(p\|\mathcal{E}) \geq \max_{p\in\Delta_{N-1}} D(p\|\mathcal{M})$.

Computations

Our findings may be biased by the fact that all models treated in Section 3 are exponential families. As a slight generalization we did computer experiments with a family of models which are not exponential families, but unions of exponential families.

Let Υ be a family of partitions of $\{1, \ldots, N\}$ and let $\mathcal{M}_\Upsilon = \bigcup_{\varrho\in\Upsilon} \mathcal{M}_\varrho$ be the union of the corresponding partition models. We are interested in these models, because they can be used to study more difficult models, like restricted Boltzmann machines and deep belief networks. Figure 2 compares a single partition model with the union of all partition models on three states.

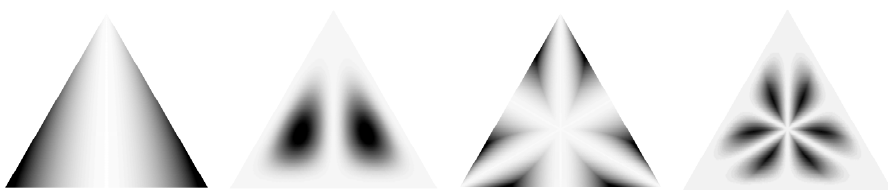

Fig. 2. From left to right: Divergence from the distributions on $\mathcal{X} = \{0, 1, 2\}$ to a partition model with two blocks. Same, scaled by the symmetric Dirichlet density with $a = 5$. Divergence from the distributions on $\mathcal{X} = \{0, 1, 2\}$ to the union of three partition models. Same, scaled by the symmetric Dirichlet density with $a = 5$. The shading is scaled on each image individually. Integrals over this kind of densities are plotted in Figure 3.

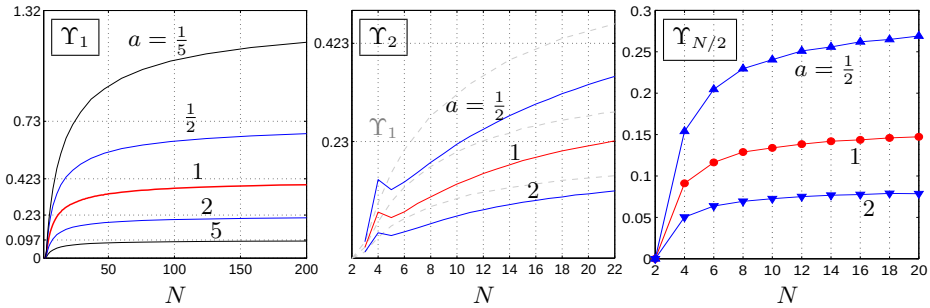

Fig. 3. Numerical approximation of the expected divergence of $\mathcal{M}_{\Upsilon_k}$ from $p \sim \mathrm{Dir}_{(a,\ldots,a)}$, for different system sizes N and values of a. Left: The case $k = 1$. The y-ticks are located at $h(a) - \ln(a) - \gamma$, which are the limits of the expected divergence from single bipartition models, see Theorem 4. Middle: The case $k = 2$. The peak at $N = 4$ emerges, because there are only 3 different partitions, instead of $\binom{4}{2}$. The dashed plot indicates corresponding results from the left figure. Right: The expected divergence of the union of all $\binom{N}{N/2}/2$ bipartition models with two blocks of cardinalities $N/2$, for even N.

For a given N and $0 \leq k \leq N/2$ let Υ_k be the set of all partitions of $\{1, \ldots, N\}$ into two blocks of cardinalities k and $N-k$. For different values of a and N we sampled 10 000 distributions from $\mathrm{Dir}_{(a,\ldots,a)}$ for which we then computed the sample-average values of $D(p\|\mathcal{M}_{\Upsilon_1})$, $D(p\|\mathcal{M}_{\Upsilon_2})$, and $D(p\|\mathcal{M}_{\Upsilon_{N/2}})$ as approximations of the expectation values. The results are shown in Figure 3.

In the first two cases the expected divergence seems to tend to the asymptotic value of $\langle D(p\|u)\rangle$. Observe that $\langle D(p\|\mathcal{M}_{\Upsilon_1})\rangle \geq \langle D(p\|\mathcal{M}_{\Upsilon_2})\rangle$, unless $N = 4$. Intuitively this makes sense for two reasons: First, for $\varrho_1 \in \Upsilon_1$ and $\varrho_2 \in \Upsilon_2$, using Theorem 4 one can show that $\langle D(p\|\mathcal{M}_{\varrho_1})\rangle \geq \langle D(p\|\mathcal{M}_{\varrho_2})\rangle$; and second, the cardinality of Υ_2 is much larger than the cardinality of Υ_1 if $N \geq 4$. For small values of N this intuition may not always be correct. For example, for $N = 8$, the expected divergence from $\mathcal{M}_{\Upsilon_{N/2}}$ is larger than the one from $\mathcal{M}_{\Upsilon_2}$, although in this case $|\Upsilon_{N/2}| = 35$ and $|\Upsilon_2| = 28$, see Figure 3 right.

We expect that, for large N, it is possible to make $\langle D(p\|\mathcal{M}_{\Upsilon_k})\rangle$ much smaller than $\langle D(p\|u)\rangle$ by choosing $k \approx N/2$. In these cases the models $\mathcal{M}_{\Upsilon_k}$ have (Hausdorff) dimension only one, but they are unions of exponentially many one-dimensional exponential families.

ACKNOWLEDGMENT.

J. Rauh is supported by the VW Foundation. G. Montúfar is supported in part by DARPA grant FA8650-11-1-7145.

(Received February 25, 2013)

REFERENCES

[1] N. Ay: An information-geometric approach to a theory of pragmatic structuring. Ann. Probab. *30* (2002), 416–436.

[2] M. Drton, B. Sturmfels, and S. Sullivant: Lectures on Algebraic Statistics. Birkhäuser, Basel 2009.

[3] B. A. Frigyik, A. Kapila, and M. R. Gupta: Introduction to the Dirichlet Distribution and Related Processes. Technical Report, Department of Electrical Engineering University of Washington, 2010.

[4] F. Matúš and N. Ay: On maximization of the information divergence from an exponential family. In: Proc. WUPES'03, University of Economics, Prague 2003, pp. 199–204.

[5] F. Matúš and J. Rauh: Maximization of the information divergence from an exponential family and criticality. In: Proc. ISIT, St. Petersburg 2011, pp. 903–907.

[6] G. Montúfar, J. Rauh, and N. Ay: Expressive power and approximation errors of restricted Boltzmann machines. In: Advances in NIPS 24, MIT Press, Cambridge 2011, pp. 415–423.

[7] I. Nemenman, F. Shafee, and W. Bialek: Entropy and inference, revisited. In: Advances in NIPS 14, MIT Press, Cambridge 2001, pp. 471–478.

[8] J. Rauh: Finding the Maximizers of the Information Divergence from an Exponential Family. Ph.D. Thesis, Universität Leipzig 2011.

[9] J. Rauh: Optimally approximating exponential families. Kybernetika *49* (2013), 199–215.

[10] D. Wolpert and D. Wolf: Estimating functions of probability distributions from a finite set of samples. Phys, Rev. E *52* (1995), 6841–6854.

Guido F. Montúfar, Department of Mathematics, Pennsylvania State University, University Park PA 16802. U. S. A.
e-mail: gfm10@psu.edu

Johannes Rauh, Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, 04103 Leipzig. Germany.
e-mail: jrauh@mis.mpg.de