Scaling of Model Approximation Errors and Expected Entropy Distances

Guido F. Montúfar
Department of Mathematics
Pennsylvania State University
University Park PA 16802 USA
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

July 17, 2012

Abstract

We compute the expected value of the Kullback-Leibler divergence to various fundamental statistical models with respect to canonical priors on the probability simplex. This yields information about the scaling of model approximation errors depending on the cardinality of the sample spaces, and it is a useful reference for more complicated statistical models such as restricted Boltzmann machines.

1 Introduction

Let p, q be probability distributions on a finite set $\mathcal{X}$. The *information divergence* or *relative entropy* or *Kullback Leibler divergence*

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

is a natural measure of dissimilarity between probability distributions that describes how easy it is to distinguish two distributions p and q by means of statistical experiments. In this paper we use the natural logarithm. The divergence is related to the log-likelihood: If p is an empirical distribution, summarizing the outcome of n statistical experiments, then the log-likelihood of a distribution q equals $-n(D(p\|q) + H(p))$. Hence, finding a *maximum likelihood estimator* q within some set of probability distributions $\mathcal{M}$ is the same as finding a minimizer of the divergence $D(p\|q)$ with q restricted to $\mathcal{M}$. The value of $D(p\|q)$ quantifies how well, or bad, the data can be described by q (and by $\mathcal{M}$).

Assume that $\mathcal{M}^{\mathrm{true}}$ is a set of probability distributions for which we do not have a simple mathematical description. We are interested in finding a model

$\mathcal{M}$ which does not necessarily include all distributions from $\mathcal{M}^{\text{true}}$, but which approximates them relatively well. What error magnitude should we accept from a good model?

To assess the expressive power of a model $\mathcal{M}$, we study properties of the function $p \mapsto D(p\|\mathcal{M}) = \inf_{q\in\mathcal{M}} D(p\|q)$. For example, the problem of finding the maximizers of this function corresponds to a worst case analysis. The problem of maximizing the divergence from a statistical model was first posed, with different motivation, in [1]. Since then, a lot of progress has been made, notably in the case where $\mathcal{M}$ is an exponential family [5, 4, 8], but also for discrete mixture models and restricted Boltzmann machines [6].

This worst case bound is not the only aspect that decides whether a given model is suited, but also the expected performance and *expected error* are of interest. This leads to the mathematical problem of computing the expectation value

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

where p is drawn from a probability density ψ on the probability simplex, called the *prior distribution*, or *prior* for short. The correct prior depends on the concrete problem at hand and is often difficult to determine. Given certain conditions on the prior, we also ask, how different is the worst case from the average case, and how much can this behavior be influenced by the choice of the model? We focus on the case that the prior ψ is the uniform distribution or a Dirichlet distribution. It turns out that in most cases the worst-case error is unbounded (as the number of elementary events grows), while the expected error is bounded. Our analysis leads to integrals that have been considered in a Bayesian framework for function estimation in [10], and we can take adventage of the tools developed there.

Our first observation is that, if ψ is the uniform prior, then the expected divergence from the uniform distribution is a monotone function of the system size N (the number of elementary events) and converges to the constant $1-\gamma \approx 0.4228 \approx 0.6099\log(2)$ as $N \to \infty$, where γ is the *Euler-Mascheroni* constant. Many natural statistical models contain the uniform distribution, and the expected divergence from such models is then bounded by the same constant. In comparison, for randomly chosen distributions p and q, the expected divergence $\langle D(p\|q)\rangle_{p,q}$ equals $1 - 1/N$. We show, for a class of models including the independence models, partition models, mixtures of product distributions with disjoint supports [6], and decomposable hierarchical models, that the expected divergence actually has the same limit $1-\gamma$, provided that the models remain *small* with respect to N (this is the case in most applications). In contrast, the maximum of the divergence from these models is at least $\log(N/(\dim\mathcal{M}+1))$, see [9]. For reasonable choices of the parameters, the results for Dirichlet priors are similar.

In Section 2 we define the models that we are interested in and collect basic properties of the Dirichlet priors. Section 3 contains analytical results for expectation values of entropies and divergences from these models. The results are interpreted in Section 4. Proofs and calculations are deferred to Appendix A.

2 Preliminaries

2.1 Models from statistics and machine learning

We consider random variables on a finite set of elementary events $\mathcal{X}$, $|\mathcal{X}| = N$. The set of probability distributions on $\mathcal{X}$ is the $(N-1)$-simplex $\Delta_{N-1} \subset \mathbb{R}^N$. We call any subset $\mathcal{M} \subseteq \Delta_{N-1}$ that can be densely parametrized a model. The support sets of a model $\mathcal{M}$ are the support sets $\operatorname{supp}(p) = \{i \in \mathcal{X} \,|\, p_i > 0\}$ of points $p = (p_i)_{i \in \mathcal{X}}$ in $\mathcal{M}$.

The *k-mixture* of a model $\mathcal{M}$ is the union of all convex combinations of any k of its points, $\mathcal{M}^k := \{\sum_{i=1}^m \lambda_i p^{(i)} \,|\, \lambda_i \geq 0, \sum_i \lambda_i = 1, p^{(i)} \in \mathcal{M}\}$. The *k-mixture with disjoint supports* is the subset of $\mathcal{M}^k$ defined by

$$\mathcal{M}_0^k = \left\{ \sum_{i=1}^k \lambda_i p^{(i)} \in \mathcal{M}^k \,\middle|\, \operatorname{supp}(p^{(i)}) \cap \operatorname{supp}(p^{(j)}) = \emptyset \text{ for all } i \neq j \right\}.$$

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 of ϱ. Partition models are closures of convex exponential families with uniform reference measure. The closure of an arbitrary convex exponential family is of the form (see [4])

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

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 equal conditional distributions $\nu(\cdot|A_k)$ yield the same model. In fact, $\mathcal{M}_{\varrho,\nu}$ equals the K-mixture of the set $\{\nu(\cdot|A_k) : k = 1, \dots, K\}$.

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

$$p(x_1, \dots, x_n) = p_1(x_1) \cdots p_n(x_n),$$

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

Let $\mathcal{S}$ be a simplicial complex on $\{0, \dots, n\}$. 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-components of x. The model $\mathcal{M}_\mathcal{S}$ is called *reducible* if there exist simplicial subcomplexes $\mathcal{S}_1, \mathcal{S}_2 \subset \mathcal{S}$ 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*. $\mathcal{M}_\mathcal{S}$ is *decomposable* if it can be iteratively reduced into simplices. The reduction can be described by a *junction tree* (see [2]), which is a tree (V, E) with vertex set the set of facets of $\mathcal{S}$ and such that the following holds: If $(\mathcal{X}, \mathcal{Y})$ is an edge, then $\mathcal{X} \cap \mathcal{Y}$

is a separator, and if this edge is removed from the tree, then the two resulting trees are junction trees of two subcomplexes $\mathcal{S}_1$ and $\mathcal{S}_2$ separated by $\mathcal{X} \cap \mathcal{Y}$. In general the junction tree is not unique, but the multi-set of separators is unique. The independence model is an example of a decomposable model.

For most models it is not possible to find a closed formula 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 above mentioned models a closed formula does exist:

The divergence from the independence model is called *multi-information* and satisfies

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

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

$$D(p\|\mathcal{M}_{\varrho,\nu}) = D(p\| \sum_{k=1}^{K} p(A_k)\nu(x|A_k)) \;. \qquad (2)$$

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

$$D(p\|\mathcal{M}_\mathcal{S}) = \sum_{S\in V} H_p(X_S) - \sum_{S\in E} H_p(X_S) - H(p). \qquad (3)$$

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

2.2 Dirichlet prior

The Dirichlet distribution (or Dirichlet prior) with *concentration parameter* $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_N)$, $\alpha_i > 0$ for all i, is the probability distribution on Δ_{N-1} defined by $\mathrm{Dir}_{\boldsymbol{\alpha}}(p) := \frac{1}{\sqrt{N}} \frac{\Gamma(\sum_{i=1}^N \alpha_i)}{\prod_{i=1}^N \Gamma(\alpha_i)} \prod_{i=1}^N p_i^{\alpha_i - 1}$ for $p = (p_1,\dots,p_N) \in \Delta_{N-1}$, where Γ is the gamma function. We write $\alpha = \sum_{i=1}^N \alpha_i$.

We will highlight especially the symmetric case $(\alpha_1,\dots,\alpha_N) = (a,\dots,a)$, which assigns no preferences to the elementary events. Observe that $\mathrm{Dir}_{(1,\dots,1)}$ is the uniform probability density on Δ_{N-1}. Furthermore, it is known that $\lim_{a\to 0} \mathrm{Dir}_{(a,\dots,a)}$ is uniformly concentrated in the point measures (it assigns mass $1/N$ to $p = \delta_x$, $x \in \mathcal{X}$), while $\lim_{a\to\infty} \mathrm{Dir}_{(a,\dots,a)}$ is concentrated in the uniform distribution $u := (1/N,\dots,1/N)$. In general, if $\boldsymbol{\alpha} \in \Delta_{N-1}$, then $\lim_{\kappa\to\infty} \mathrm{Dir}_{\kappa\boldsymbol{\alpha}}$ is the Dirac delta concentrated on $\boldsymbol{\alpha}$.

The Dirichlet distributions satisfy the following *aggregation property*: Consider a partition $\varrho = \{A_1,\dots,A_K\}$ of $\mathcal{X} = \{1,\dots,N\}$. If $p = (p_1,\dots,p_N) \sim \mathrm{Dir}_{(\alpha_1,\dots,\alpha_N)}$, then $(\sum_{i\in A_1} p_i,\dots,\sum_{i\in A_K} p_i) \sim \mathrm{Dir}_{(\sum_{i\in A_1}\alpha_i,\dots,\sum_{i\in A_K}\alpha_i)}$, see, e.g., [3]. We write $\boldsymbol{\alpha}^\varrho = (\alpha_1^\varrho,\dots,\alpha_K^\varrho)$, $\alpha_k^\varrho = \sum_{i\in A_k}\alpha_i$ 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 $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, $|\mathcal{X}| = N$, $\mathcal{X}_k = \{1, \dots, N_k\}$ we write $\boldsymbol{\alpha}^k = (\alpha_1^k, \dots, \alpha_{N_k}^k)$, $\alpha_j^k = \sum_{x \in \mathcal{X}\colon x_k = j} \alpha_x$ for the concentration parameter of the Dirichlet distribution induced on the $\mathcal{X}_k$-marginal $(\sum_{x \in \mathcal{X}\colon x_k=1} p(x), \dots, \sum_{x \in \mathcal{X}\colon x_k = N_k} p(x))$. Note that $\sum_{j=1}^{N_k} \alpha_j^k = \alpha$, and moreover, if $\alpha_x = 1$ for all $x \in \mathcal{X}$, then $\alpha_j^k = N/N_k$ for $j = 1, \dots, N_k$. For example, if p is drawn uniformly from the simplex of joint distributions Δ_{N-1}, then the sampled marginal probability distribution $p(y_k) = \sum_{x \in \mathcal{X}\colon x_k = y_k} p(x)$, $y_k \in \mathcal{X}_k$ is Dirichlet distributed in $\Delta_{N_k - 1}$ with concentration parameter $\boldsymbol{\alpha}^k = (N/N_k, \dots, N/N_k)$.

3 Expected entropies and divergences

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,

$$h(k) = \log(k) + \gamma + O(\frac{1}{k}),$$

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

The following theorems present formulas for expectation values of divergences from models as well as asymptotic results. The results are based on explicit solutions of the integrals, as done by [10]. The proofs are contained in Appendix A.

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

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

In the symmetric case $(\alpha_1, \dots, \alpha_N) = (a, \dots, a)$,

- $\langle H(p) \rangle = h(Na) - h(a)$

$$= \begin{cases} \log(Na) + \gamma - h(a) + O(1/Na) & \textit{for large } N \textit{ and const. } a \\ \log(N) + O(1/a) & \textit{for large } a \textit{ and arb. } N \\ O(aN) & \textit{as } a \to 0 \textit{ with bounded } N \\ h(c) + O(a) & \textit{as } a \to 0 \textit{ with } aN = c \end{cases}$$

- $\langle D(p \| u) \rangle = \log(N) - h(aN) + h(a)$

$$= \begin{cases} h(a) - \log(a) - \gamma + O(1/Na) & \textit{for large } N \textit{ and const. } a \\ O(1/a) & \textit{for large } a \textit{ and arb. } N \\ \log(N) + O(aN) & \textit{as } a \to 0 \textit{ with bounded } N \\ \log(N) - h(c) + O(a) & \textit{as } a \to 0 \textit{ with } aN = c. \end{cases}$$

The maximum of the (Shannon) entropy $H(p) = -\sum_i p_i \log p_i$ on the probability simplex Δ_{N-1} is attained at the uniform distribution u, which satisfies $H(u) = \log(N)$. For large N or a, the average entropy is close to the maximum value. It follows that in these cases the expected divergence from the uniform distribution u remains bounded. The fact that the expected entropy is close to the maximal entropy makes it difficult to estimate the entropy. See [7] for a discussion and possible solutions.

Theorem 2.

- *For any $q \in \Delta_{N-1}$, when $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$, then*

$$\langle D(p\|q)\rangle = \sum_{i=1}^{N} \frac{\alpha_i}{\alpha}(h(\alpha_i) - \log(q_i)) - h(\alpha)\ .$$

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

$$\langle D(p\|q)\rangle = \log(N) - h(aN) + h(a) + D(q\|u)\ .$$

  When $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$ and $q \sim \mathrm{Dir}_{\tilde{\boldsymbol{\alpha}}}$, then

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

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

- *For any $q \in \Delta_{N-1}$, when p is drawn uniformly from Δ_{N-1}, then*

$$\langle D(p\|q)\rangle = -\sum_{i=1}^{N} \frac{1}{N}\log(q_i) - h(N) + 1 = D(u\|q) + 1 - \gamma + O(1/N)\ .$$

The divergence is unbounded in $\Delta_{N-1} \times \Delta_{N-1}$, since $D(p\|q) = +\infty$ if p is not absolutely continuous with respect to q. Nevertheless, if $p, q \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$, then in the limit $N \to \infty$ the expected divergence $\langle D(p\|q)\rangle$ remains bounded, provided $\frac{1}{N}\sum_{i=1}^{N} \alpha_i = \alpha/N$ is bounded from below by a positive constant.

Consider a sequence of distributions $q_N \in \Delta_{N-1}$, $N \in \mathbb{N}$. As $N \to \infty$ the expected divergence $\langle D(\cdot\|q_N)\rangle$ with respect to the uniform prior is bounded from above by $1 - \gamma + \varepsilon$, $\varepsilon > 0$ if and only if $\limsup_{N\to\infty} D(u\|q_N) \le \varepsilon$. If $q_x \ge \frac{1}{N}e^{-\varepsilon}$ for all $x \in \mathcal{X}$, then $D(u\|q) \le \varepsilon$. Therefore, the expected divergence $\langle D(\cdot\|q_N)\rangle$ is unbounded only if the sequence q_N accumulates at the boundary of the probability simplex, and $\lim_{N\to\infty}\langle D(p\|q_N)\rangle \le 1 - \gamma + \varepsilon$ whenever q_N is in the subsimplex $\mathrm{conv}\{(1 - e^{-\varepsilon})\delta_x + e^{-\varepsilon}u\}_{x\in\mathcal{X}}$. The relative Lebesgue volume of this subsimplex in Δ_{N-1} is $(1 - e^{-\varepsilon})^{N-1}$.

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

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

If $(\alpha_1,\dots,\alpha_N) = (a,\dots,a)$ *(symmetric Dirichlet),*

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

If, moreover, Na/N_k *is large for all* k *(this happens, for example, when* a *remains bounded from below by some* $\varepsilon > 0$ *and* (i) *all* N_k *become large, or* (ii) *all* N_k *are bounded and* n *becomes large), then:*

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

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* $\varrho = \{A_1,\dots,A_K\}$ *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* $p \sim \mathrm{Dir}_{\boldsymbol{\alpha}}$*, then*

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

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

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

If furthermore $N \gg K$*, then*

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

Partition models (with $\nu = u$) also contain the uniform distribution, and therefore the expected divergence is again bounded. In contrast, the maximal divergence is $\max_{p\in\Delta_{N-1}} D(p\|\mathcal{M}_\varrho) = \max_k \log(N_k)$. The result for mixtures of product distributions of disjoint supports is similar:

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

- *If $p \sim \mathrm{Dir}_{(\alpha_1,\dots,\alpha_N)}$, then the expected divergence to $\mathcal{M}^K_{1,\varrho}$ is*

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

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

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

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

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

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

The k-mixture of binary product distributions with disjoint supports is contained in the restricted Boltzmann machine model with $k-1$ hidden nodes, see [6]. Hence Theorem 5 gives bounds for the expected divergence to these models.

Theorem 6. *For a decomposable model $\mathcal{M}_{\mathcal{S}}$ with junction tree (V,E), if $p \sim \mathrm{Dir}_{(\alpha_1,\dots,\alpha_N)}$, then*

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

where $\alpha^S_j = \sum_{x\colon x_S = j}\alpha_x$ for $j\in\mathcal{X}_S$. If p is drawn uniformly at random, then

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

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

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

4 Discussion

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

- For a low-dimensional model $\mathcal{M} \subset \Delta_{N-1}$ and large N, 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 very easy to construct counter-examples: Using space-filling curves, it is possible to construct one-dimensional models $\mathcal{M}$ with an arbitrary 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 $\alpha = (a, \dots, a)$, then the expected divergence from $\mathcal{M}_{\varrho,\nu}$ is minimal, if and only if $\nu = u$. 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:

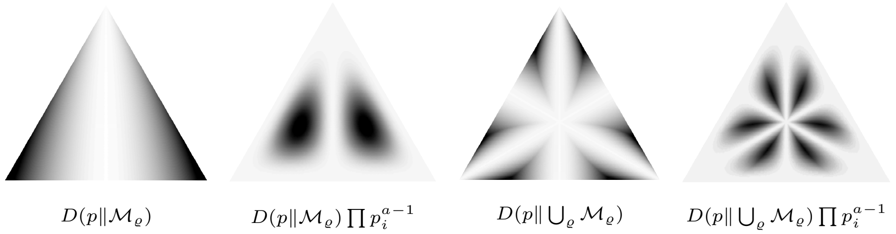
- For any exponential family $\mathcal{E}$ there is a partition model $\mathcal{M}$ of the same dimension such that $\langle D(p\|\mathcal{E})\rangle \geq \langle D(p\|\mathcal{M})\rangle$.

The statement is, of course, true for zero-dimensional exponential families, i.e., models that consist 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 $\max_{p\in\Delta_{N-1}} D(p\|\mathcal{E}) \geq \max_{p\in\Delta_{N-1}} D(p\|\mathcal{M})$.

Our findings may be biased by the fact that all the models treated in Section 3 are examples of 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, and let $\mathcal{M}_\Upsilon = \bigcup_{\varrho\in\Upsilon} \mathcal{M}_\varrho$ be the union of the corresponding partition models. Our interest in these models comes from the fact that such models are contained in more difficult models with hidden variables, like restricted Boltzmann machines and deep belief networks. Figure 1 compares a single partition model on three states with the union of all partition models for bipartitions.



$D(p\|\mathcal{M}_\varrho)$ $D(p\|\mathcal{M}_\varrho)\prod p_i^{a-1}$ $D(p\|\bigcup_\varrho \mathcal{M}_\varrho)$ $D(p\|\bigcup_\varrho \mathcal{M}_\varrho)\prod p_i^{a-1}$

Figure 1: From left to right: Divergence to a partition model with two blocks on $\mathcal{X} = \{1, 2, 3\}$. Same, multiplied by a symmetric Dirichlet density with parameter $a = 5$. Divergence to the union of the three partition models with two blocks on $\mathcal{X} = \{1, 2, 3\}$. Same, multiplied by the symmetric Dirichlet density with $a = 5$. The shading is scaled on each image individually.

For a given N and $0 \leq k \leq N/2$ let Υ_k be the set of all partitions of $\{1, \dots, N\}$ into two blocks of cardinalities k and $N-k$. For different values of a and N we computed $D(p\|\mathcal{M}_{\Upsilon_1})$ for 10 000 distributions sampled from $\mathrm{Dir}_{(a,\dots,a)}$, $D(p\|\mathcal{M}_{\Upsilon_2})$ for 20 000 distributions sampled from $\mathrm{Dir}_{(a,\dots,a)}$, and $D(p\|\mathcal{M}_{\Upsilon_{N/2}})$ for 20 000 distributions sampled from the uniform prior. The results are shown in Figure 2.

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 2 right.

For $N = 22$ we computed $D(p\|\mathcal{M}_{\Upsilon_{N/2}})$ for 500 uniformly sampled distributions (in this case $|\Upsilon_{N/2}| = 352\,716$), and found $\langle D(p\|\mathcal{M}_{\Upsilon_{N/2}})\rangle \approx 0.1442$ (with variance 0.0032), which is well below the corresponding expectation values for $\mathcal{M}_{\Upsilon_1}$ and $\mathcal{M}_{\Upsilon_2}$. 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 this case, the model $\mathcal{M}_{\Upsilon_k}$ has (Hausdorff) dimension only one, but it is a union of exponentially many one-dimensional exponential families.

A Computations and proofs

The analytic formulas in Theorem 1 are [10, Theorem 7]. The asymptotic expansions are direct.

The proof of Theorem 2 makes use of the following Lemma, see [10, Theorem 3]:

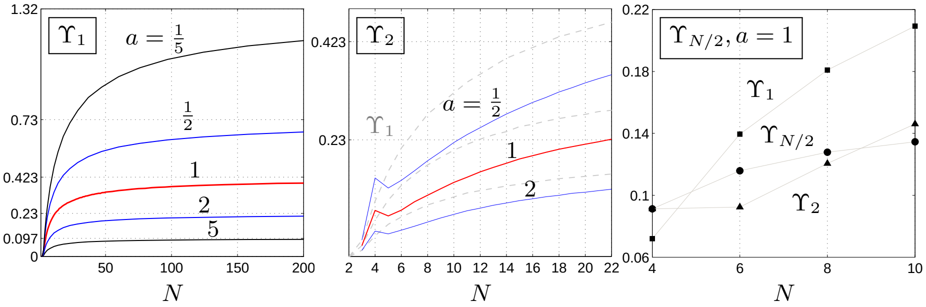


Figure 2: Expected divergence (numerically) from various unions of bipartition models with respect to $\mathrm{Dir}_{(a,\dots,a)}$, for different system sizes N and values of the concentration parameter a. Left: Union of all bipartition models with blocks of cardinalities 1 and $(N-1)$. The y-ticks are located at $h(a) - \log(a) - \gamma$, which are the limits of the expected divergence from single bipartition models, see Theorem 4. Middle: Union of all bipartition models with blocks of cardinalities 2 and $(N-2)$. The peak at $N = 4$ is caused by the fact that there are only 3 different partitions when $N = 4$, instead of $\binom{N}{2}$. The dashed plot indicates corresponding results from the left figure. Right: Comparison of the expected divergence from the two previous models and the union of all $\binom{N}{N/2}/2$ bipartition models with two blocks of cardinalities $N/2$, for $a = 1$ and even N.

Lemma 7. *Let $\{A_1,\dots,A_K\}$ be a partition of $\mathcal{X} = \{1,\dots,N\}$, let $\alpha_1,\dots,\alpha_N$ be positive reals, and let $\alpha^k = \sum_{i\in A_k}\alpha_i$ for $k = 1,\dots,K$. Then*

$$\int_{\Delta_{N-1}} \big(\sum_{i\in A_k} p_i\big)\log\big(\sum_{i\in A_k} p_i\big)\prod_{i=1}^N p_i^{\alpha_i - 1}\,\mathrm{d}p = \int_{\Delta_{K-1}} p_k^*\log(p_k^*)\prod_{k'=1}^K (p_{k'}^*)^{\alpha^{k'}-1}\,\mathrm{d}p^*$$
$$= \frac{\alpha^k\prod_{k'=1}^K\Gamma(\alpha^{k'})}{\Gamma(\alpha+1)}\big(h(\alpha^k) - h(\alpha)\big)\,.$$

Proof of Theorem 2. The first statement follows from

$$\int_{\Delta_{N-1}} \log(q_i)p_i\prod_i p_i^{n_i}\,\mathrm{d}p \Big/ \int_{\Delta_{N-1}} \prod_i p_i^{n_i}\,\mathrm{d}p = \log(q_i)\frac{(n_i+1)}{(N+n)}$$

and $D(p\|q) = -H(p) - \sum_i p_i\log(q_i)$. By Lemma 7,

$$\int_{\Delta_{N-1}} \log(q_i)\prod_i q_i^{n_i}\,\mathrm{d}q \Big/ \int_{\Delta_{N-1}} \prod_i q_i^{n_i}\,\mathrm{d}q = h(n_i) - h(N+n-1)\,,$$

and the remaining statements follow. □

Theorem 3 is a corollary to Theorem 1, the aggregation property of the Dirichlet priors and the formula (1) for the multi-information. Theorem 4 follows from (2), and Theorem 6 follows from (3). Similarly, Theorem 5 follows from the equality

$$D(p\|\mathcal{M}_0) = \sum_{i=1}^{K} \sum_{x\in A_i} p(x) \log \frac{p(x)p(A_i)^{n-1}}{\prod_{j=1}^{n}(\sum_{y\in A_i : y_j = x_j} p(y))} ,$$

which can be derived as follows: The unique solution $q \in \operatorname{arginf}_{q'\in M_{1,\varrho}^{K}} D(p\|q')$ satisfies $p(A_i) = q(A_i)$, and $q(\cdot|A_i) \in \operatorname{arginf}_{q'\in\mathcal{M}_1} D(p(\cdot\|A_i)\|q')$.

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

References

[1] N. Ay, “An information-geometric approach to a theory of pragmatic structuring,” *Annals of Probability*, vol. 30, pp. 416–436, 2002.

[2] M. Drton, B. Sturmfels, and S. Sullivant, *Lectures on Algebraic Statistics*, 1st ed., ser. Oberwolfach Seminars. Birkhuser, Basel, 2009, vol. 39.

[3] B. A. Frigyik, A. Kapila, and M. R. Gupta, “Introduction to the Dirichlet distribution and related processes,” Department of Electrical Engineering University of Washington, Tech. Rep., 2010.

[4] F. Matúš and N. Ay, “On maximization of the information divergence from an exponential family,” in *Proceedings of the 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 *2011 IEEE International Symposium on Information Theory Proceedings (ISIT2011)*, 2011.

[6] G. F. Montúfar, J. Rauh, and N. Ay, “Expressive power and approximation errors of restricted Boltzmann machines,” in *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, 2011, pp. 415–423, available at http://books.nips.cc/papers/files/nips24/NIPS2011_0307.pdf.

[7] I. Nemenman, F. Shafee, and W. Bialek, “Entropy and inference, revisited,” in *NIPS*, 2001, pp. 471–478.

[8] J. Rauh, “Finding the maximizers of the information divergence from an exponential family,” Ph.D. dissertation, Universitt Leipzig, 2011.

[9] ——, “Optimally approximating exponential families,” *submitted*, 2012, available at http://arxiv.org/abs/1111.0483.

[10] D. Wolpert and D. Wolf, “Estimating functions of probability distributions from a finite set of samples,” *Physical Review E*, vol. 52, no. 6, pp. 6841–6854, 1995.