Asymptotics of Continuous Bayes for Non-i.i.d. Sources

Tor Lattimore[1] and Marcus Hutter[2]

[1] Department of Computing Science, University of Alberta
[2] Research School of Computer Science, Australian National University

Abstract

Clarke and Barron analysed the relative entropy between an i.i.d. source and a Bayesian mixture over a continuous class containing that source. In this paper a comparable result is obtained when the source is permitted to be both non-stationary and dependent. The main theorem shows that Bayesian methods perform well for both compression and sequence prediction even in this most general setting with only mild technical assumptions.

1 Introduction

We continue the work of Clarke and Barron [1] bounding the relative entropy between probability measures on infinite sequences and a Bayesian mixture over measures in some continuous class $\mathcal{M}$. Small values of relative entropy have a number of implications. Notably that prediction and compression using the Bayesian mixture is nearly as good as using the true unknown source. The contribution of this work is to show that Bayesian methods generalise to the case where data is not sampled identically and independently, a situation that is often encountered in practical problems. One application is online compression where a sequence of words is observed and should be compressed. If we assume the words are sampled from some probability distribution, which may be non-stationary and dependent, then near-optimal compression is obtained by arithmetic coding with respect to that distribution. Typically, however, the probability distribution from which the data is sampled is unknown. One approach in this case is to code with respect to the Bayesian mixture over some set of measures believed to contain the truth. Bounding the relative entropy between the truth and the Bayesian mixture is then equivalent to bounding the compression redundancy due to coding with respect to the wrong measure.

Another application is discriminative learning where a classifier should predict label data based on observations. Sometimes it is possible to model the joint distribution of the observations and labels together, but in many cases the observations are relatively unordered and modelling the conditional distribution of the label given the observation is easier. Our results show that Bayesian methods for discriminative learning perform well under only mild technical assumptions.

Suppose $\mathcal{M} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ is a set of measures on the space of infinite sequences over alphabet Ω. We let M be a Bayesian mixture over $\mathcal{M}$ with respect to some prior and analyse the relative entropy $D(P_\theta^n \| M^n)$ where P_θ^n and M^n are the distributions on the first n observations induced by P_θ and M^n respectively. Unlike in [1] we permit P_θ to be non-stationary and dependent, so P_θ^n is typically not a product measure.

Our main contribution is a proof under mild technical assumptions that the relative entropy can be bounded by

$$D(P_\theta^n \| M^n) \leq \ln \frac{1}{w(\theta)} + \frac{d}{2} \ln \frac{n}{2\pi} + \frac{1}{2} \ln \det \bar{J}_n(\theta) + o(1)$$

where $w(\theta)$ is the prior density of parameter θ and $\bar{J}_n(\theta)$ is the mean Fisher information matrix at θ. If P_θ^n is a product measure, then $\bar{J}_n(\theta)$ coincides with the usual Fisher information matrix and the result above is the same as that given in [1].

The main difficulty in generalising the proof in [1] to non-stationary dependent sources is the fact that the mean Fisher information matrix $\bar{J}_n(\theta)$ is now dependent on n. We start with some notation (Section 2). The main theorems

are presented (Section 3) followed by a variety of applications and discussion (Section 4). The proofs are found in Sections 5 and 6. We discuss some of the assumptions and special cases in Sections 7-9, including the case when the information matrix vanishes and a comparison to the known results in zero-dimensional (countable) families. For discussion and conclusions see Section 10.

2 Notation

We use $\ln$ for the natural logarithm and $A^\top$ for the transpose of matrix A. Suppose $A \in \mathbb{R}^{d \times d}$ and $x \in \mathbb{R}^d$. Then $|x|_2^2 := x^\top x$ is the standard 2-norm and $|x|_A^2 := x^\top A x$ is the norm with respect to A. Note that $|\cdot|_A$ is a norm only if A is positive definite, but occasionally we abuse notation by writing $|\cdot|_A$ even if A is not positive definite. We use $\|A\|_2$ for the spectral norm of A, which for positive definite matrices is the largest eigenvalue of A. It is easy to check that $|x|_2^2 \le |x|_A^2 \|A^{-1}\|_2$. The determinant of A is $\det A$. The indicator function $[\![expr]\!]$ is equal to 1 if $expr$ is true and 0 otherwise. For function $f : \mathbb{R}^n \to \mathbb{R}$ we write $\partial_i f$ for the partial derivative of f with respect to the ith coordinate. Higher order derivatives are denoted by $\partial_{i,j} f$ or $\partial_{i_1 \cdots i_k} f$. The Gamma function is $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$.

Probability spaces. Let $(\Omega, \mathcal{F})$ be a measurable space. Then the product measure space is $(\Omega^n, \mathcal{F}^n)$ where $\mathcal{F}^n$ is the σ-algebra generated by the n-fold tensor product of $\mathcal{F}$. Let $\mathcal{F}^\infty$ be the cylinder σ-algebra defined by

$$\mathcal{F}^\infty := \sigma \left\{ \bigcup_{n=1}^\infty \mathcal{F}^n \otimes \{\Omega^\infty\} \right\}.$$

Then $(\Omega^\infty, \mathcal{F}^\infty)$ is a measurable space. A probability measure P on this space may be thought of as a family of probability measures where $P^n : \mathcal{F}^n \to [0,1]$ is induced by restriction $P^n(A) = P(A \times \Omega^\infty)$. We think of P as a probability measure on infinite sequences in Ω^∞ and P^n to be the probability measure on the first n observations induced by P. If $\omega \in \Omega^\infty$, then $\omega_{1:n} = \omega_1 \omega_2 \cdots \omega_n \in \Omega^n$ is the projection of the first n components and $\omega_{<n} := \omega_{1:n-1}$.

Parameterised families and Bayes. Suppose $\Theta \subseteq \mathbb{R}^d$ and that P_θ is a family of probability measures parameterised by $\theta \in \Theta$. We assume there exists a measure ν such that ν^n is σ-finite for all n and the density of P_θ^n with respect to ν^n exists for all n and $\theta \in \Theta$, which is denoted by p_θ^n. Then by the definition of the density we have

$$(\forall A \in \mathcal{F}^n) \qquad P_\theta^n(A) = \int_A p_\theta^n(\omega) d\nu^n(\omega).$$

We denote expectations with respect to P_θ^n by $\mathbb{E}_\theta$ where n is always understood from the context. Variances are denoted by Var_θ. The n-step mean Fisher information matrix at $\theta \in \Theta$ is denoted by $\bar{J}_n(\theta) \in \mathbb{R}^{d \times d}$ and defined

$$\bar{J}_n(\theta)_{i,j} := -\frac{1}{n} \mathbb{E}_\theta \left[\partial_{i,j} \ln p_\theta^n \right],$$

which like all other quantities of interest is independent of the reference measure used to define the density. There are other definitions of the Fisher information matrix, all of which coincide under the weak assumption that derivatives and expectations can be exchanged [2, §18]. We are not aware of any interesting family of measures for which this assumption is not satisfied. Let $w : \Theta \to [0, \infty)$ be a prior probability density with respect to the Lebesgue measure, which satisfies $\int_\Theta w(\theta) d\theta = 1$. Then the density of the Bayesian mixture $m^n : \Omega^n \to [0, \infty)$ is defined by

$$m^n(\omega_{1:n}) := \int_\Theta w(\theta) p_\theta^n(\omega_{1:n}) d\theta.$$

The Bayes mixture measure $M^n : \mathcal{F}^n \to [0,1]$ may equivalently be defined by

$$M^n(A) = \int_A m^n(\omega) d\nu(\omega) \equiv \int_A \int_\Theta w(\theta) p_\theta^n(\omega) d\theta d\nu(\omega) = \int_\Theta w(\theta) P_\theta^n(A) d\theta.$$

2

If w and p_θ^n are continuous at θ, then P_θ^n is absolutely continuous with respect to M^n and the relative entropy between P_θ^n and M^n is defined by

$$D(P_\theta^n \| M^n) := \mathbb{E}_\theta \ln \frac{p_\theta^n}{m^n}.$$

We do not claim that $P_\theta \ll M$ and indeed this is not generally the case ($\mathcal{M} \equiv$ Bernoulli measures). The d-dimensional (non-degenerate) normal distribution has density

$$\mathcal{N}(\theta|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}|x - \mu|^2_{\Sigma^{-1}}\right)$$

where $\mu \in \mathbb{R}^d$ is the mean and Σ is the positive definite covariance matrix.

3 Main Theorems

Our main result generalises Theorem 2.3 in [1]. The conditions (a), (b) and (d) below are standard regularity conditions also made in [1]. The conditions (c) and (e) are regularity conditions on the mean Fisher information matrix, which in this work depends on n as well as θ. The second result relaxes condition (e) at the cost of a slightly worse bound.

Theorem 1. *Let $\theta_\circ \in \Theta$ and $\bar{J}_n \equiv \bar{J}_n(\theta_\circ)$. Define a family of functions $\{f_n\} : \Theta \to \mathbb{R}$ by*

$$f_n(\theta) := \frac{1}{n} D(P_{\theta_\circ}^n \| P_\theta^n).$$

Assume that:

(a) w is continuous at $\theta_\circ$.

(b) $f_n(\theta)$ is twice differentiable at $\theta_\circ$.

(c) Entries in the Hessian of f_n are equicontinuous at $\theta_\circ$.

$$\lim_{\theta \to \theta_\circ} \sup_n |\partial_{i,j} f_n(\theta) - \partial_{i,j} f_n(\theta_\circ)| = 0.$$

(d) $\partial_i f_n(\theta_\circ) = 0$ for all i.

(e) $\limsup_{n \to \infty} \|\bar{J}_n^{-1}\|_2 < \infty$.

If $D_n \equiv D(P_{\theta_\circ}^n \| M^n)$, then

$$\limsup_{n \to \infty} \left(D_n - \frac{d}{2} \ln \frac{n}{2\pi} - \frac{1}{2} \ln \det \bar{J}_n \right) \le \ln \frac{1}{w(\theta_\circ)}.$$

The three components of the bound can be explained as follows. Assume P_θ is i.i.d., which implies that $\bar{J}_n = \bar{J}_1$ is independent of n. We want to approximate D_n by integrating the Bayes mixture over a region R_n containing $\theta_\circ$.

$$D_n = \mathbb{E}_{\theta_\circ} \ln \frac{p_{\theta_\circ}^n}{m^n} \le \mathbb{E}_{\theta_\circ} \ln \frac{p_{\theta_\circ}^n}{\int_{R_n} w(\theta) p_\theta^n d\theta} \approx \mathbb{E}_{\theta_\circ} \ln \frac{p_{\theta_\circ}^n}{\int_{R_n} w(\theta_\circ) p_{\theta_\circ}^n d\theta} = \ln \frac{1}{w(\theta_\circ)} + \ln \frac{1}{V(R_n)}$$

where $V(R_n)$ is the volume of R_n. The quality of the approximation depends on the choice of R_n with smaller regions leading to better approximations, but also smaller volumes. A single scalar parameter θ can usually be estimated with an accuracy of about $n^{-1/2}$ using n samples, which suggests that p_θ^n is approximately constant inside the cube of width $n^{-1/2}$ and dimension d. As n tends to infinity the volume of the cube shrinks to zero and the continuity of the prior justifies the approximation $w(\theta) \approx w(\theta_\circ)$. Choosing this as the region leads to $V(R_n) = n^{-d/2}$, which explicates the

$\frac{d}{2} \ln n$ component of the bound. The additional term depending on the information matrix is explained by making the above argument more precise. The insight is that $\bar{J}_1$ is a measure of the expected curvature of $\ln p^1_\theta$ at $\theta_\circ$. Then large $\bar{J}_1$ implies that R_n must be small for a good approximation. If the data is i.i.d., then the curvature is independent of n and $V(R_n) \propto n^{-d/2} \det \bar{J}_1^{1/2}$ is unsurprising. For non i.i.d. sources the curvature of $\ln p^n_\theta$ is possibly dependent on n and so appears in the bound in Theorem 1. An alternative explanation is found in [1].

By making an additional assumption we also obtain a stronger result in terms of the actual redundancy, rather than the expected redundancy.

Theorem 2. *Assume conditions (a-e) and additionally that*

$$(f) \qquad\qquad \frac{1}{n} \operatorname{Var}_{\theta_\circ} \left(\ln \frac{p^n_{\theta_\circ}}{p^n_\theta} \right)$$

is twice differentiable at $\theta_\circ$ and has equicontinuous second derivatives as in assumption (c) above. Then there exists a sequence of random variables $r_n : \Omega^\infty$ such that for all $\omega = \omega_1 \omega_2 \cdots \in \Omega^\infty$

$$\limsup_{n\to\infty} \left(\ln \frac{p^n_{\theta_\circ}(\omega_{1:n})}{m^n(\omega_{1:n})} - \frac{d}{2} \ln \frac{n}{2\pi} - \frac{1}{2} \ln \det \bar{J}_n - r_n(\omega) \right) \leq \ln \frac{1}{w(\theta_\circ)} + \ln 2 + \sqrt{2d}$$

where $\lim_{n\to\infty} \mathbb{E}_{\theta_\circ} |r_n| = 0$.

The condition (c) above is subtly different to the one suggested in [3, §3] where it is assumed that the average information matrix is equicontinuous at $\theta_\circ$. The version used here is require for technical reasons and may in fact be necessary. The condition (d) is very weak and follows from the standard regularity conditions that permit the exchange of the derivative and expectation. The assumption (e) that $\|\bar{J}_n^{-1}\|_2$ is uniformly bounded in n is unexpected because if the opposite is true, then $\mathcal{M}$ should in some direction be approximately flat near $\theta_\circ$, which is precisely when we might expect to do even better than Theorem 1 implies.

Theorem 3. *Let $\varepsilon > 0$. Under conditions (a-d) above it holds that*

$$\limsup_{n\to\infty} \left(D_n - \frac{d}{2} \ln \frac{n}{2\pi} - \frac{d}{2} \ln(\|\bar{J}_n\|_2 + \varepsilon) \right) \leq \ln \frac{1}{w(\theta_\circ)}.$$

Although the choice of ε is arbitrary, the $o(1)$ term hidden by the $\limsup$ depends on ε, which prevents $\varepsilon = 0$. Theorem 3 does not imply that D_n grows sub-logarithmically if $\|\bar{J}_n\|_2 = 0$. Later we show that D_n will typically grow logarithmically with n except when f_n is completely flat at $\theta_\circ$. The $\frac{d}{2} \ln(\|\bar{J}_n\|_2 + \varepsilon)$ term may be replaced by $\frac{1}{2} \ln \det A_n$ where $A_n = \bar{J}_n + \varepsilon I$ with I the identity matrix. See the proof for details. The proofs of Theorems 1, 2 and 3 are delayed until Sections 5 and 6.

4 Applications

The conditions of Theorems 1 and 3 are satisfied for many well-known sources such as categorical sources, Markov models and some types of discriminative learning. We present some simple examples, but first give some interpretations of the relative entropy as a measure of performance. Note that many of the applications considered below have been considered elsewhere in more specialized articles. The main objective of this section is to show that the new result provides comparable bounds on the performance of Bayes, even in these special cases.

Interpretations of relative entropy. For simplicity suppose Ω is countable, which allows us to replace the integrals by sums and think of the density p^n_θ as a distribution on Ω^n. The relative entropy between P^n and M^n is a versatile measure of the performance of M^n when predicting or compressing in place of P^n. If $\omega_{1:t} \in \Omega^t$, then the conditional density of measure p^t is written $p^t(\omega_t | \omega_{<t}) = p^t(\omega_{1:t}) / p^t(\omega_{<t})$. An important property is the chain rule, which says that

$$D_n \equiv D(P^n \| M^n) = \mathbb{E}_P \sum_{t=1}^n d_t$$

where

$$d_t(\omega_{<t}) := \mathbb{E}_P\left[\ln\frac{p^t(\omega_t|\omega_{<t})}{m^t(\omega_t|\omega_{<t})}\bigg|\omega_{<t}\right] \equiv \sum_{\omega_t\in\Omega} p^t(\omega_t|\omega_{<t})\ln\frac{p^t(\omega_t|\omega_{<t})}{m^t(\omega_t|\omega_{<t})}$$

is the $\mathcal{F}^{t-1}$-measurable random variable that is the relative entropy between the 1-step predictive distributions of P and M given $\omega_{<t}$. Therefore if $d_t(\omega_{<t})$ is small, then $m^t(\cdot|\omega_{<t})$ is close to $p^t(\cdot|\omega_{<t})$ and so predicting using the Bayesian mixture measure M is nearly as good as using the unknown P. If D_n grows sub-linearly, then d_t converges to zero in Cesàro average, which implies that M predicts almost as well as the optimal unknown distribution P.

More directly, suppose $\omega \in \Omega^\infty$ is sampled from an unknown P_θ and observed sequentially. An online compression algorithm that knows P_θ may use arithmetic coding to produce a code for $\omega_{1:n}$ that has an expected code-length of at most two bits more than the optimal value. If M is used rather than P, then the expected additional code-length is the relative entropy D_n. This means that substituting the Bayesian mixture into the arithmetic coding algorithm achieves an expected code-length of at most $D_n + 2$ bits more than the theoretical limit [4, 5, 6].

Categorical sources. Let $d \geq 1$ and $\Omega = \{0, 1, 2, \cdots, d\}$. An i.i.d. categorical source is a product measure on the infinite sequences Ω^∞ so that the probability of sampling symbol $0 \leq k \leq d$ depends only on k and not the preceding symbols. The space of categorical measures is parameterised by

$$\Theta = \left\{\theta \in [0,1]^d : |\theta|_1 \leq 1\right\}$$

where θ_k is the probability that the source parameterised by θ produces symbol k. Define $\theta_0 = 1 - |\theta|_1$ and

$$p_\theta(\omega_{1:n}) = \prod_{t=1}^n \theta_{\omega_t}.$$

Then since p_θ is a product measure we have for all n that $\bar{J}_n = \bar{J}_1$, which may easily be computed by hand to be

$$(\forall 1 \leq i,j \leq d) \qquad \bar{J}_n(\theta)_{i,j} = \begin{cases} \frac{1}{\theta_i} + \frac{1}{\theta_0} & \text{if } i = j \\ \frac{1}{\theta_0} & \text{otherwise.} \end{cases}$$

Then by Lemma 8 in the Appendix we compute the determinant to be $\det \bar{J}_n(\theta) = \prod_{k=0}^d \frac{1}{\theta_k}$. We are permitted to choose any prior density provided it is continuous, but a natural choice is Jeffrey's prior, which when it exists is defined to be proportional to the square root of the determinant of the information matrix. Jeffrey's prior is a non-informative prior that is parameterisation invariant. A nice property is that if w is chosen to be Jeffrey's prior, then the bound on D_n will be independent of $\theta_\circ$. For i.i.d. categorical sources Jeffrey's prior is the symmetric Dirichlet with parameter $\frac{1}{2}$.

$$w(\theta) = \Gamma\left(\frac{d+1}{2}\right)\pi^{-\frac{d+1}{2}}\sqrt{\prod_{k=0}^d \frac{1}{\theta_k}} \propto \sqrt{\det \bar{J}_n}.$$

After checking the conditions for Theorem 1 are satisfied we obtain

$$D(P_\theta^n \| M^n) \leq \frac{d+1}{2}\ln\pi + \ln\frac{1}{\Gamma(\frac{d+1}{2})} + \frac{d}{2}\ln\frac{n}{2\pi} + o(1).$$

Note that in this case the Bayesian mixture corresponds to the KT estimator for which the redundancy is already well-known [7, 8, **?**, 9]. It is also worth remarking on the choice of parameterisation. Perhaps the most natural approach would be to choose $\Theta \subseteq \mathbb{R}^{d+1}$ under the constraint that $|\theta_1| = 1$. Unfortunately the theorem cannot be applied for a variety of reasons. One is that there is no prior density on $\Theta \subset \mathbb{R}^{d+1}$ with respect to the $(d+1)$-dimensional Lebesgue measure. Another is that condition (d) is no longer satisfied. Even if the theorem held, the bound would still be worse by an additive $\frac{1}{2}\log n$. An overkill solution would be to allow Θ to be a d-dimensional manifold, but since

all conditions are local this is not too helpful in practise. One just ends up working on the chart level directly anyway, which essentially is what we do above.

So far we haven't made use of the generalisation to non-stationary dependent sources, but do so now by applying Theorem 1 to Markov models.

Markov models. One of the simplest commonly used examples of a non-i.i.d. process is the Markov chain. The redundancy of Bayes methods for Markov chains has been studied in detail in [10] (see Corollary 1) and [11]. Our general Theorem 1 implies a comparable result. Let $\Omega = \{1, 2, \cdots, N\}$ be a finite set of states. Then a Markov chain is characterised by a transition matrix $\theta : \mathbb{R}^{N \times N} \to [0, 1]$ with θ_j^k being the probability of transitioning from state j to state k. The vector $\theta_j = \{\theta_j^1 \cdots, \theta_j^N\} \in \mathbb{R}^N$ therefore satisfies $|\theta_j| = 1$ for all states j. The corresponding measure on the sequence of states is defined in terms of its density by

$$p_\theta(\omega_{1:n}) = \prod_{t=1}^n \theta_{\omega_{t-1}}^{\omega_t}$$

where we assume that the intial state $\omega_0 \in \Omega$ is known. Given the constraint that $|\theta_j| = 1$ allows us to view the parameter space as

$$\Theta = \left\{ \theta \in [0, 1]^{N(N-1)} : \sum_{k=1}^{N-1} \theta_j^k \leq 1, \forall j \right\}$$

where $\theta_j^N := 1 - \sum_{j=1}^{N-1} \theta_j^k$. The mean Fisher information matrix is no longer independent of n, but converges as $n \to \infty$. If $\pi_\theta(j) = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^n p_\theta(\omega_t = j|\omega_0)$ is the steady-state distribution of the Markov chain p_θ, then the determinant of the Fisher information matrix was shown in [10] to be

$$\det \bar{J}_n(\theta) = \prod_{j=1}^N \frac{\pi_\theta(j)^{N-1}}{\prod_{k=1}^N \theta_j^k}$$

Therefore if we choose $w(\theta) \propto \sqrt{\det \bar{J}_n}$, then

$$D(P_\theta^n \| M^n) \leq \frac{N(N-1)}{2} \ln \frac{n}{2\pi} + O(1),$$

which is known to be the minimax rate. For detailed discussion on the redundancy of Bayes methods for Markov chains see [10, 12] and [13].

Discriminative learning. We now consider a regression setting where a predictor should learn a (noisy) function $f : X \to Y$ from sequences of data x_t and targets y_t. At each time-step t the predictor observes data x_t in X and should predict a distribution over labels y_t in Y. The sampled label is then observed and the cycle repeats. It can occur that the sequence of observations is unpredictable and the joint distribution $P(x, y)$ may be hard to model. In this case we may prefer discriminative learning where $P(y|x)$ is modelled for each x. Since x may be arbitrary we cannot reasonably assume that the distribution of the labels y are independent and identically distributed and so the work of Clarke and Barron does not apply. In order to apply our main theorem we need to be more specific and parameterise $\mathcal{M} = \{P_\theta(\cdot|x)\}$ where $\theta \in \Theta$. The most natural example is linear regression described below.

Linear Basis Function Regression. Let $x_1, x_2, \cdots, x_n$ be an observed sequence of data where $x_k \in X$ and $\Phi : X \to [0, 1]^d$ be a set of bounded basis function. Then for $\theta \in \Theta \equiv \mathbb{R}^d$ define a model for the targets $y_{1:n} \in \mathbb{R}^n \equiv \Omega^n$ by

$$p_\theta(y_{1:n}|x_{1:n}) = \prod_{t=1}^n \mathcal{N}(\theta^\top \Phi(x_t), \beta^{-1})$$

where the noise parameter $\beta > 0$ is known. The Hessian of f_n in the statement of Theorem 3 is easily computed to be

$$\partial_{i,j} f_n(\theta) = \frac{\beta}{n} \sum_{t=1}^n \Phi(x_t)_i \Phi(x_t)_j,$$

6

which is independent of θ. Therefore conditions (b-c) are trivially satisfied. For condition (a) we simply choose the prior by

$$w(\theta) := \mathcal{N}(\theta | \mu, \Sigma),$$

which is continuous everywhere for all $\mu \in \mathbb{R}^d$ and non-degenerate Σ. Other priors are possible, but the Gaussian prior permits efficient computation of the posterior and predictive distributions [14]. Now

$$\bar{J}_n(\theta_\circ)_{i,j} = \partial_{i,j} f_n(\theta_\circ) = \frac{\beta}{n} \sum_{t=1}^n \Phi_i(x_t) \Phi_j(x_t).$$

Since $\Phi(x) \in [0, 1]^d$ the spectral norm of the information matrix is bounded by $\|\bar{J}_n(\theta)\|_2 \leq d\|\bar{J}_n(\theta)\|_{\max} \leq d \cdot \beta$. Therefore by Theorem 3 we obtain for all $\varepsilon > 0$ that

$$\limsup_{n \to \infty} \left(D_n - \frac{d}{2} \ln \frac{n}{2\pi} - \frac{d}{2} \ln (d \cdot \beta + \varepsilon) \right) \leq \ln \frac{1}{\mathcal{N}(\theta_\circ | \mu, \Sigma)} = \frac{1}{2} |\theta_\circ - \mu|_{\Sigma^{-1}}^2 + \frac{d}{2} \ln 2\pi + \frac{1}{2} \ln \det \Sigma.$$

Under the assumption that the model is not mis-specified this shows that Bayesian linear regression converges to the truth with low cumulative expected error.

5 Proof of Theorems 1 and 2

As suggested in [3] we roughly follow the proof in [1], carefully adapting each step to the non i.i.d. case where necessary.

Lemma 4. *Let $R_n \subseteq \Theta$ be a Lebesgue measurable region with non-zero measure and define a probability measure on R_n by*

$$c_n := \int_{R_n} \exp\left(-n f_n(\theta)\right) d\theta$$

$$\phi_n(\theta) := \frac{1}{c_n} \exp\left(-n f_n(\theta)\right),$$

Then

$$\ln \frac{p_{\theta_\circ}^n}{m^n} \leq \ln \frac{1}{w(\theta_\circ)} + \ln \frac{1}{c_n} + \int_{R_n} \phi_n(\theta) \left(\ln \frac{w(\theta_\circ)}{w(\theta)} \right) d\theta + \int_{R_n} \phi_n(\theta) \left(\ln \frac{p_{\theta_\circ}^n}{p_\theta^n} - n f_n(\theta) \right) d\theta \qquad (\star)$$

$$D_n \leq \ln \frac{1}{w(\theta_\circ)} + \ln \frac{1}{c_n} + \int_{R_n} \phi_n(\theta) \left(\ln \frac{w(\theta_\circ)}{w(\theta)} \right) d\theta \qquad (\star\star)$$

Before presenting the proof we want to remark on one aspect of the difference between our proof and [1]. They replaced $n f_n(\theta) \equiv D(P_{\theta_\circ}^n \| P_\theta^n)$ with its Taylor series in the first step above when defining $\phi_n(\theta)$. With our definition we obtain a straight-forward bound on the expected redundancy with almost no assumptions. The properties of $f_n(\theta)$ can then be used to control the dominant $\ln \frac{1}{c_n}$ term, with strong assumptions leading to strong results. Note that the integral term in $(\star\star)$ vanishes asymptotically under the assumptions that w is continuous at $\theta_\circ$ and R_n contracts to the point $\theta_\circ$ as n tends to infinity.

Proof. We follow the proof of Theorem 2.3 in [1], but replace the ϕ_n in their proof with the version defined above.

$$\ln\left(\frac{p_{\theta_\circ}^n}{m^n}\right) - \ln\frac{1}{w(\theta_\circ)} \stackrel{(a)}{=} -\ln\left(\frac{m^n}{w(\theta_\circ)p_{\theta_\circ}^n}\right)$$

$$\stackrel{(b)}{=} -\ln\int_\Theta \frac{w(\theta)p_\theta^n}{w(\theta_\circ)p_{\theta_\circ}^n}d\theta$$

$$\stackrel{(c)}{\leq} -\ln\int_{R_n} \frac{w(\theta)p_\theta^n}{w(\theta_\circ)p_{\theta_\circ}^n}d\theta$$

$$\stackrel{(d)}{=} \ln\frac{1}{w(\theta_\circ)} - \ln\int_{R_n}\phi_n(\theta)\frac{w(\theta)p_\theta^n}{w(\theta_\circ)p_{\theta_\circ}^n}c_n\exp\left(nf_n(\theta)\right)d\theta$$

$$\stackrel{(e)}{\leq} \int_{R_n}\phi_n(\theta)\ln\left(\frac{w(\theta_\circ)p_{\theta_\circ}^n}{w(\theta)p_\theta^n}\frac{1}{c_n}\exp\left(-nf_n(\theta)\right)\right)d\theta$$

$$\stackrel{(f)}{=} \ln\frac{1}{c_n} + \int_{R_n}\phi_n(\theta)\left(\ln\frac{p_{\theta_\circ}^n}{p_\theta^n} - nf_n(\theta) + \ln\frac{w(\theta_\circ)}{w(\theta)}\right)d\theta$$

where (a), (b) and (f) are immediate from definitions and rearrangement. (c) by positivity of the quantity inside the integral and by restricting the integral to the region R_n. (d) by the introduction of $\phi_n(\theta)$. (e) is true by Jensen's inequality. The proof of ($\star\star$) follows by taking the expectation with respect to $P_{\theta_\circ}$ leading to

$$D_n \equiv \mathbb{E}_{\theta_\circ}\ln\left(\frac{p_{\theta_\circ}^n}{m^n}\right) \leq \ln\frac{1}{w(\theta_\circ)} + \ln\frac{1}{c_n} + \int_{R_n}\phi_n(\theta)\left(\ln\frac{w(\theta_\circ)}{w(\theta)}\right)d\theta$$

where we used the fact that $nf_n(\theta) \equiv \mathbb{E}_{\theta_\circ}\ln p_{\theta_\circ}^n/p_\theta^n$. $\square$

We now prove Theorem 1. First we choose R_n in such a way that $R_n \to \{\theta_\circ\}$ as n tends to infinity. We then bound c_n by approximating $f_n(\theta)$ using a Taylor series expansion about $\theta_\circ$. Since R_n contracts to a point and w is continuous, the term inside the integral in Lemma 4 vanishes.

Proof of Theorem 1. Fix some $K \in \mathbb{N}$ and define

$$R_n := \left\{\theta : n\left|\theta - \theta_\circ\right|_{\bar{J}_n}^2 < K\right\}.$$

By condition (e) in the theorem statement we have that $\|\bar{J}_n^{-1}\|_2$ is uniformly bounded by some constant c. Recalling the definition of R_n and Lemma 7 we obtain

$$\lim_{n\to\infty}\sup_{\theta\in R_n} n|\theta-\theta_\circ|_2^2 \leq \lim_{n\to\infty}\sup_{\theta\in R_n} n|\theta-\theta_\circ|_{\bar{J}_n}^2\|\bar{J}_n^{-1}\|_2 \leq \lim_{n\to\infty} K\|\bar{J}_n^{-1}\|_2 \leq K\cdot c.$$

Therefore R_n contracts to the point-set $\{\theta_o\}$ as n tends to infinity. To bound $\ln\frac{1}{c_n}$ we approximate $f_n(\theta)$ by a second order Taylor series about $\theta_\circ$. It is immediate from the definition that $f_n(\theta_\circ) = 0$. The first derivative is the expected value of the score function, which vanishes by assumption (d).

$$\partial_i f_n(\theta_\circ) = 0.$$

The second derivative at $\theta_\circ$ is the information matrix $\bar{J}_n$ so

$$f_n(\theta) = \frac{1}{2}|\theta-\theta_\circ|_{\bar{J}_n}^2 + |\theta-\theta_\circ|_{h_n(\theta)}^2$$

where $h_n : \Theta \to \mathbb{R}$ satisfies $\lim_{\theta \to \theta_\circ} \sup_n \|h_n(\theta)\|_2 = 0$ by Lemma 6 in the Appendix. We now bound c_n by comparing to the multivariate normal. First we deal with the remainder term.

$$
\begin{aligned}
E_n &:= \inf_{\theta \in R_n} \exp\left(-n\,|\theta - \theta_\circ|^2_{h_n(\theta)}\right) \\
&\geq \exp\left(-\sup_{\theta \in R_n} n\|h_n(\theta)\|_2^2\|\bar{J}_n^{-1}\|_2\,|\theta - \theta_\circ|^2_{\bar{J}_n}\right) \\
&\geq \exp\left(-\sup_{\theta \in R_n} \frac{\|h_n(\theta)\|_2^2\|\bar{J}_n^{-1}\|_2}{K}\right) \xrightarrow[R_n \to \{\theta_\circ\}]{n \to \infty} 1
\end{aligned}
$$

where the first inequality follows from Lemma 7, the second from the definition of R_n. The convergence follows from the equicontinuity of the remainder term and the fact that R_n contracts to $\{\theta_\circ\}$. Therefore

$$
\begin{aligned}
c_n &\equiv \int_{R_n} \exp\left(-nf_n(\theta)\right) d\theta \\
&= \int_{R_n} \exp\left(-\frac{1}{2}n|\theta - \theta_\circ|^2_{\bar{J}_n} - n\,|\theta - \theta_\circ|^2_{h_n(\theta)}\right) d\theta \\
&\geq E_n\sqrt{\frac{(2\pi)^d}{\det(n\bar{J}_n)}} \int_{R_n} \mathcal{N}(\theta|\theta_\circ, (n\bar{J}_n)^{-1})d\theta,
\end{aligned}
$$

which is proportional to a normal integral over R_n with mean $\theta_\circ$ and variance $(n\bar{J}_n)^{-1}$. Therefore by a multivariate form of Chebyshev's inequality (Lemma 5 in the Appendix or see [15]) we can bound c_n by

$$
\begin{aligned}
c_n &\geq E_n\sqrt{\frac{(2\pi)^d}{\det(n\bar{J}_n)}}\,(1 - d/K) \\
&= E_n\sqrt{\left(\frac{2\pi}{n}\right)^d \frac{1}{\det \bar{J}_n}}\,(1 - d/K).
\end{aligned}
\tag{1}
$$

Therefore by Lemma 4

$$
D_n \leq \ln\frac{1}{w(\theta_\circ)} + \ln\frac{1}{1 - d/K} + \frac{d}{2}\ln\frac{n}{2\pi} + \frac{1}{2}\ln\det\bar{J}_n + \ln\frac{1}{E_n} + \int_{R_n}\phi_n(\theta)\left(\ln\frac{w(\theta_\circ)}{w(\theta)}\right)d\theta.
$$

By the continuity of w and the fact that R_n converges to a point it follows that

$$
\lim_{K\to\infty}\lim_{n\to\infty}\int_{R_n}\phi_n(\theta)\ln\frac{w(\theta_\circ)}{w(\theta)}d\theta = 0.
$$

Therefore sending K to infinity leads to.

$$
\limsup_{n\to\infty}\left(D_n - \frac{d}{2}\ln\frac{n}{2\pi} - \frac{1}{2}\ln\det\bar{J}_n\right) \leq \frac{1}{w(\theta_\circ)}
$$

as required. $\qquad\square$

Proof of Theorem 2. We use the same R_n as in the proof of Theorem 1 and apply the first part of Lemma 4 to obtain.

$$
\ln\frac{p_{\theta_\circ}^n}{m^n} \leq \ln\frac{1}{w(\theta_\circ)} + \ln\frac{1}{c_n} + \int_{R_n}\phi_n(\theta)\left(\ln\frac{w(\theta_\circ)}{w(\theta)}\right)d\theta + \int_{R_n}\phi_n(\theta)\left(\ln\frac{p_{\theta_\circ}^n}{p_\theta^n} - nf_n(\theta)\right)d\theta.
\tag{2}
$$

The $\ln\frac{1}{c_n}$ term and the first integral are constant random variables and are bounded as they were in Theorem 1. We only need to show that

$$
r_n := \int_{R_n}\phi_n(\theta)\left(\ln\frac{p_{\theta_\circ}^n}{p_\theta^n} - nf_n(\theta)\right)d\theta
$$

9

converges to zero in L^1. By an application of Fubini's theorem the expectations can be exchanged to yield

$$\mathbb{E}_{\theta_\circ}|r_n| = \int_{R_n} \phi_n(\theta) \mathbb{E}_{\theta_\circ} \left| \ln \frac{p_{\theta_\circ}^n}{p_\theta^n} - n f_n(\theta) \right| d\theta$$

$$\leq \int_{R_n} \phi_n(\theta) \left(\mathrm{Var}_{\theta_\circ} \ln \frac{p_{\theta_\circ}^n}{p_\theta^n} \right)^{\frac{1}{2}} d\theta.$$

We now use assumption (f) to control the variance term by taking a second degree Taylor series expansion about $\theta_\circ$. If $\theta \in R_n$, then

$$\mathrm{Var}_{\theta_\circ} \left(\ln \frac{p_{\theta_\circ}^n}{p_\theta^n} \right) = n\, |\theta - \theta_\circ|_{\bar{J}_n}^2 + n\, |\theta - \theta_\circ|_{h_n(\theta)}^2$$

$$\leq n\, |\theta - \theta_\circ|_{\bar{J}_n} \left(1 + \|\bar{J}_n^{-1}\|_2 \|h_n(\theta)\|_2^2 \right)$$

$$\leq K(1 + \|\bar{J}_n^{-1}\|_2 \|h_n(\theta)\|_2^2)$$

where the second derivative of the variance at $\theta_\circ$ is twice the information matrix with remainder $h_n : \Theta \to \mathbb{R}$ satisfying. $\lim_{\theta \to \theta_\circ} \sup_n \|h_n(\theta)\|_2 = 0$ by Lemma 6. In the last two steps we used Lemma 7 and the definition of R_n. Therefore

$$\lim_{n \to \infty} \mathbb{E}_{\theta_\circ}|r_n| \leq \lim_{n \to \infty} \int_{R_n} \phi_n(\theta) \left(\mathrm{Var}_{\theta_\circ} \ln \frac{p_{\theta_\circ}^n}{p_\theta^n} \right)^{\frac{1}{2}} d\theta$$

$$\leq \lim_{n \to \infty} \int_{R_n} \phi_n(\theta) \left(K(1 + \|\bar{J}_n^{-1}\|_2 \|h_n(\theta)\|_2^2) \right)^{\frac{1}{2}} d\theta \qquad (3)$$

$$= \sqrt{K} \qquad (4)$$

where the final inequality follows since R_n contracts to a point and because $\|\bar{J}_n^{-1}\|_2$ is uniformly bounded. Therefore $\lim_{n \to \infty} \mathbb{E}_{\theta_\circ}|r_n| = 0$. We cannot send K to infinity like in the proof of Theorem 1. Instead we simply fix $K = 2d$ and insert Equations (1) and (3) into (2) to obtain for all $\omega = \omega_1 \omega_2 \cdots \in \Omega^\infty$

$$\limsup_{n \to \infty} \left(\ln \frac{p_{\theta_\circ}^n(\omega_{1:n})}{m^n(\omega_{1:n})} - \frac{d}{2} \ln \frac{n}{2\pi} - \frac{1}{2} \ln \det \bar{J}_n - r_n(\omega) \right) \leq \ln \frac{1}{w(\theta_\circ)} + \ln 2 + \sqrt{2d}$$

as required. $\qquad \square$

6 Proof of Theorem 3

In the previous section we used the assumption (e) that $\|\bar{J}_n^{-1}\|_2$ was uniformly bounded by a constant to show that the critical region R_n contracts to $\theta_\circ$ as n tends to infinity. This result can be guaranteed by defining the region R_n with respect to a different norm. Define a positive definite matrix $A_n := \bar{J}_n + \varepsilon I$ with $I \in \mathbb{R}^{d \times d}$ the identity matrix and $\varepsilon > 0$ to be chosen later. The matrix A_n is positive definite since it is the sum of two positive definite matrices. Furthermore $A_n > \bar{J}_n$. The assumption may now be eliminated by using the norm $|\theta - \theta_\circ|_{A_n}$ to define the critical region rather than $|\theta - \theta_\circ|_{\bar{J}_n}$. The only component that requires checking is the bound on c_n. The critical region becomes

$$R_n := \left\{ n\, |\theta - \theta_\circ|_{A_n}^2 \leq K \right\}.$$

We now show that R_n contracts to a point as n tends to infinity. It is an easy consequence of the spectral decomposition theorem that the smallest eigenvalue of A_n satisfies $\lambda_{\min} \geq \varepsilon$ and $\|A_n^{-1}\|_2 = \frac{1}{\lambda_{\min}} \leq \frac{1}{\varepsilon}$, which is precisely the condition required for R_n to contract to the point $\theta_\circ$. Therefore

$$\limsup_{n \to \infty} \left(D_n - \frac{d}{2} \ln \frac{n}{2\pi} - \frac{1}{2} \ln \det A_n \right) \leq \ln \frac{1}{w(\theta_\circ)}$$

The determinant of A_n may asymptotically be arbitrarily larger than $\det \bar{J}_n$, which appeared in Theorem 1. If λ_i is the ith eigenvalue of $\bar{J}_n$, then $\lambda_i + \varepsilon$ is the ith eigenvalue of A_n and an easy bound on $\det A_n$ is

$$\det A_n = \prod_{i=1}^{d}(\lambda_i + \varepsilon) \overset{(a)}{\leq} \prod_{i=1}^{d}(\|\bar{J}_n\|_2 + \varepsilon) = (\|\bar{J}_n\|_2 + \varepsilon)^d$$

where (a) follows by the inequality $\lambda_i \leq \|\bar{J}_n\|_2$. Therefore

$$\limsup_{n\to\infty}\left(D_n - \frac{d}{2}\ln\frac{n}{2\pi} - \frac{d}{2}\ln(\|\bar{J}_n\|_2 + \varepsilon)\right) \leq \ln\frac{1}{w(\theta_\circ)}$$

as required. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

7   Weakening the assumptions

The continuity of the prior w at $\theta_\circ$ is required to ensure that w assigns non-zero probability to environments in a region of $\theta_\circ$. That f_n is twice differentiable is required to define the curvature of $\mathbb{E}_{\theta_\circ}\ln p_\theta$. Condition (d) was used to show that the 1st order terms in the Taylor series of f_n vanish. The assumption is standard and very weak. For example, it holds for all finite Ω, as well as exponential families in their canonical form. A nice discussion with examples may be found in [2, §18]. The equicontinuity condition is necessary to ensure that the Taylor approximation of f_n is uniformly accurate in n. Actually a counter-example when $\partial_{i,j}f_n$ is not equicontinuous is not hard to construct. Let $\Omega = \{0, 1\}$, $\Theta = [0, 1]$ and $w(\theta) = 1$ and $a_n \geq 0$ be a sequence of constants to be chosen later. Define probability mass in terms of its conditionals by

$$p_\theta(1|\omega_{<n}) := \min\left\{1, \theta + a_n\left(\theta - \frac{1}{2}\right)^2\right\},$$

which depends on n, but not $\omega_{<n}$. For $\theta_\circ = \frac{1}{2}$ we have $p_{\theta_\circ}(1|\omega_{<n}) = \frac{1}{2}$ for all n, but if a_n is chosen to converge to infinity, then $\lim_{n\to\infty}p_\theta(1|\omega_{<n}) = 1$ for all $\theta \neq \theta_\circ$. Based on this, if a_n is chosen to converge to infinity sufficiently fast, then width of the interval in which p_θ^n is close to $p_{\theta_\circ}^n$ can be made arbitrarily small (Fig. 1) and so

$$\lim_{n\to\infty}\left(D_n - \frac{1}{2}\ln\frac{n}{2\pi}\right) = \infty.$$

The information matrix at $\theta_\circ = \frac{1}{2}$ can be found via a straight-forward application of the definition and is $\bar{J}_n(\theta_\circ) = 1/4$, which is positive (definite) and independent of n. Therefore all conditions of the theorem are met except for equicontinuity and yet the result does not hold. This should not be surprising. The equicontinuity condition ensures that the approximation of f_n via its second order Taylor series has uniform error, which is crucial to the proof.

8   Higher Order Derivatives

In the main theorems we used the second order approximation of the log likelihood function to control the redundancy, but if $\bar{J}_n = 0$, then Theorem 1 cannot be applied and Theorem 3 seems suboptimal since, regardless of how ε is chosen, the bound on D_n still increases like $\frac{d}{2}\log n$. Here we consider the case where the second (and maybe higher) derivatives vanish. For the remainder of this section we fix an even k and assume that $\partial_{i_1\cdots i_j}f_n(\theta_\circ) = 0$ for all $j < k$. Then by the kth order Taylor expansion we let

$$f_n(\theta) \approx \frac{1}{k!}\sum_{i_1\cdots i_k}\partial_{i_1\cdots i_k}f_n(\theta_\circ)(\theta - \theta_\circ)^i$$
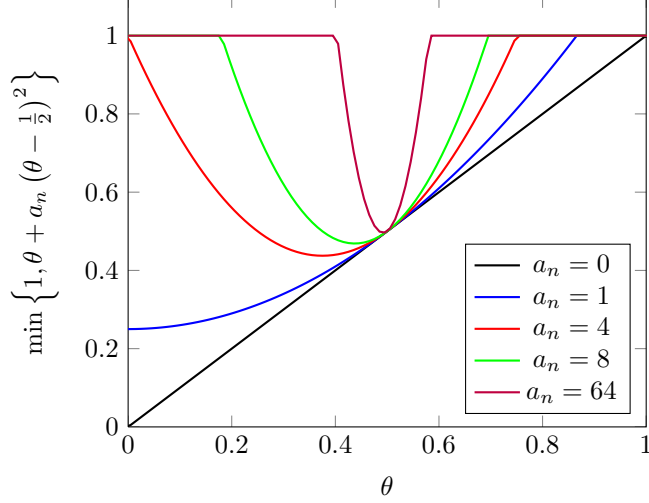
Figure 1: Conditional probability mass $p_\theta(1|\omega_{<n})$ for different a_n

where for $x \in \mathbb{R}^d$ and $i = i_1 \cdots i_k$ we use the multi-index notation $x^i \equiv \prod_{i=1}^{k} x_{i_i}$. Then define

$$\Lambda_n := \max_{\substack{x \in \mathbb{R}^d \\ |x|_k = 1}} \sum_{i_1 \cdots i_k} \partial_{i_1 \cdots i_k} f_n(\theta_\circ) x^i$$

where $\Lambda_n \in (0, \infty)$ because the derivative is positive definite and the set $\{x \in \mathbb{R}^d : |x|_k = 1\}$ is compact. Then

$$\sum_{i_1 \cdots i_k} \partial_{i_1 \cdots i_k} f_n(\theta_o)(\theta - \theta_\circ)^i \leq \Lambda_n \sum_{i=1}^{d} (\theta_i - \theta_{\circ i})^k.$$

We replace the critical region used in the proof of Theorem 1 with

$$R_n := \left\{ \theta : n\Lambda_n \sum_{i=1}^{d} (\theta_i - \theta_{\circ i})^k < K \right\}.$$

The proof goes through unchanged except for the computation of c_n. We define a measure on $\mathbb{R}^d$ by

$$G_n(\theta) := \frac{1}{\eta_n} \exp\left(-\frac{n\Lambda_n}{k!} \sum_{i=1}^{d} (\theta_i - \theta_{\circ i})^k \right)$$

where the η_n is the normalisation constant that can be determined by the usual methods for computing Gaussian integrals.

$$\eta_n := \int_{\mathbb{R}^d} \exp\left(-\frac{n\Lambda_n}{k!} \sum_{i=1}^{d} (\theta_i - \theta_{\circ i})^k \right) d\theta = \left(\int_{-\infty}^{\infty} \exp\left(-\frac{n\Lambda_n}{k!} x^k \right) dx \right)^d = \left(\frac{2}{k} \left(\frac{k!}{n\Lambda_n} \right)^{1/k} \Gamma\left(\frac{1}{k} \right) \right)^d.$$

where the last step follows by substituting $y = n\Lambda_n x^k / k!$ and the definition of the Gamma function. Then

$$c_n = \int_{R_n} \exp\left(-n f_n(\theta) \right) d\theta \approx \eta_n \int_{R_n} G_n(\theta) d\theta$$

12

where the approximation is due to the lower order terms when substituting the Taylor series expansion. The integral may be bounded naively by computing the kth moments of G in each direction, Markov's inequality and the union bound. We omit the details. This approach leads to the bound

$$\int_{R_n} G(\theta) d\theta \geq 1 - \frac{d^2(k-1)!}{K} \stackrel{K \to \infty}{\longrightarrow} 1,$$

which importantly is independent of n. Like in the proof of Theorem 1 we can send K to infinity in the final stage of the proof.

$$c_n \geq \eta_n \left(1 - \frac{d^2(k-1)!}{K}\right)$$

Finally we apply Lemma 4 to obtain $\limsup_{n \to \infty} \left(D_n - \ln \frac{1}{\eta_n}\right) \leq \ln \frac{1}{w(\theta_\circ)}$ and so

$$\limsup_{n \to \infty} \left(D_n - \frac{d}{k} \ln n - \frac{d}{k} \ln \Lambda_n\right) \leq \ln \frac{1}{w(\theta_\circ)} + \frac{d}{k} \ln \frac{1}{k!} + d \ln \frac{k}{2\Gamma(\frac{1}{k})}.$$

Most interesting is the dependence on n, which if Λ_n is assumed to be constant in n is logarithmic, regardless of k. Larger k only decrease the multiplicative constant. Note that for $k = 2$ we have $\frac{d}{2} \ln \Lambda_n = \frac{d}{2} \ln \|\bar{J}_n\|_2 \geq \frac{1}{2} \ln \det \bar{J}_n$, which featured in Theorem 1. An obscure case is when the kth derivatives of $f_n(\theta)$ vanish for all k, but where $f_n(\theta)$ is itself non-zero. This can occur, for example if $\Omega = \{0, 1\}$, $\Theta = [0, 1]$, $w(\theta) = 1$ and $p_\theta^n(\omega_{1:n})$ is i.i.d. with $p_\theta^1(0) = \exp(-\exp(-1/\theta))$ where $p_0^1(0) := 1$. In this case D_n can be shown to grow with order $\log \log n$.

9 Zero Dimensional Families

All results have been proven for $d \geq 1$. Here we briefly compare to the case when $\mathcal{M} = \{P_1, P_2, \cdots\}$ is a countable family of measures. It is no longer possible, desirable or necessary to define the Bayes mixture in terms of densities or a continuous prior. Instead $w : \mathbb{N} \to [0, 1]$ is a probability mass function and the Bayes mixture is defined by

$$M(A) := \sum_{i=1}^{\infty} w(i) P_i(A).$$

Then it is trivial to bound $D_n = D(P_k^n \| M^n) \leq \ln \frac{1}{w(k)}$, which is independent of n. $\mathcal{M}$ may be parameterised by $\Theta = \mathbb{R}^+$ where $P_\theta = P_k$ for $\theta \in [k-1, k)$. With this parameterisation the log likelihood is constant in a region about $\theta_\circ = k - \frac{1}{2} \in (k-1, k)$, but in this case even the bound shown in the previous section guarantees only sub-logarithmic redundancy, when in fact it should be constant.

10 Conclusion

Our main contribution is a generalisation of Theorem 2.3 in [1] to the case where sources are permitted to be dependent and non-stationary. Under only mild assumptions we obtain the same bound on the relative entropy between the Bayesian mixture and $P_\theta \in \mathcal{M}$ of order $\frac{d}{2} \log n$ where d is the dimension of the hypothesis class $\mathcal{M}$. The results were applied to discriminative learning and Markov chains for which the results in [1] do not apply. If the mean Fisher information has unbounded spectral norm as n tends to infinity, then Theorem 3 can be applied to obtain nearly the same bound as Theorem 1. We showed that if the information matrix vanishes, then higher-order approximations of the log-likelihood lead to similar bounds on the redundancy. Our results can usefully be applied beyond discriminative learning and Markov sources presented here and suggest Bayesian reinforcement learning as a natural and important example of non-i.i.d.sources [3]. Interesting future work is to generalise the other results in [1].

Acknowledgment

References

[1] B. Clarke, A. Barron, Information-theoretic asymptotics of Bayes methods, IEEE Transactions on Information Theory 36 (1990) 453–471.

[2] P. Grünwald, The Minimum Description Length Principle, MIT Press, 2007.

[3] M. Hutter, Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability, Springer, Berlin, 2005.

[4] J. Rissanen, G. Langdon, Arithmetic coding, IBM Journal of research and development 23 (2) (1979) 149–162.

[5] I. Witten, R. Neal, J. Cleary, Arithmetic coding for data compression, Communications of the ACM 30 (6) (1987) 520–540.

[6] D. MacKay, Information theory, inference and learning algorithms, Cambridge university press, 2003.

[7] R. Krichevsky, V. Trofimov, The performance of universal encoding, Information Theory, IEEE Transactions on 27 (2) (1981) 199–207.

[8] J. Rissanen, Universal coding, information, prediction, and estimation, Information Theory, IEEE Transactions on 30 (4) (1984) 629–636.

[9] R. Begleiter, R. El-Yaniv, Superior guarantees for sequential prediction and lossless compression via alphabet decomposition, Journal of Machine Learning Research 7 (2006) 379–411.

[10] K. Atteson, The asymptotic redundancy of Bayes rules for Markov chains, Information Theory, IEEE Transactions on 45 (6) (1999) 2104–2109.

[11] J.-i. Takeuchi, T. Kawabata, A. R. Barron, Properties of jeffreys mixture for markov sources, Information Theory, IEEE Transactions on 59 (1) (2013) 438–457.

[12] J.-i. Takeuchi, Fisher information determinant and stochastic complexity for Markov models, in: Information Theory, 2009. ISIT 2009. IEEE International Symposium on, IEEE, 2009, pp. 1894–1898.

[13] L. Davisson, R. McEliece, M. Pursley, M. Wallace, Efficient universal noiseless source codes, Information Theory, IEEE Transactions on 27 (3) (1981) 269–279.

[14] C. Bishop, Pattern recognition and machine learning.

[15] K. Ferentinos, On Tchebycheff's type inequalities., Trabajos de Estadística e Investigación Operativa 33 (1) (1982) 125–132.

A   Technical Results

Lemma 5 ([15])**.** *Suppose $\theta \sim \mathcal{N}(\theta_\circ, \Sigma)$ where $\theta_\circ \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite. Then for all $\delta > 0$*

$$P(|\theta - \theta_\circ|^2_{\Sigma^{-1}} \leq \delta) \geq 1 - \frac{d}{\delta}.$$

The following lemma is required to uniformly bound the approximation error of the Taylor series of $\ln P^n_\theta / P^n_{\theta_\circ}$. We presume similar results appear elsewhere, but include the statement and proof for completeness and because references seem hard to find.

Lemma 6. *Let $\theta_\circ \in \Theta \subset \mathbb{R}^d$ and $f_n : \Theta \to \mathbb{R}$ and $H_n(\theta) \in \mathbb{R}^{d \times d}$ be the Hessian of f_n at θ. Suppose also that*

(a) f_n is twice differentiable at $\theta_\circ \in \Theta$.

(b) $\partial_i f_n(\theta_\circ) = 0$ and $f_n(\theta_\circ) = 0$ for all $1 \le i \le d$.

(c) $H_n(\theta)_{i,j}$ is equicontinuous at $\theta_\circ$ for all $1 \le i, j \le d$.

Then there exists a family of functions $h_n : \Theta \to \mathbb{R}^{d \times d}$ such that

$$f_n(\theta) = \frac{1}{2} |\theta - \theta_\circ|^2_{H_n(\theta_\circ)} + |\theta - \theta_\circ|^2_{h_n(\theta)}$$

where $\lim_{\delta \to 0} \sup_n \sup_{\theta \in N^2_\delta} \|h_n(\theta)\|_2 = 0$ and $N^2_\delta := \{\theta : |\theta - \theta_\circ|_2 < \delta\}$.

Proof. For $\theta \in N^2_\delta$ a first order Taylor series of f_n about $\theta_\circ$ leads to

$$f_n(\theta) = (\theta - \theta_\circ)^\top S_n(\theta)(\theta - \theta_\circ)$$

where

$$S_n(\theta)_{i,j} \in \left\{ \frac{1}{2} H_n(\bar\theta)_{i,j} : \bar\theta \in N^2_\delta \right\}$$

Now define $h_n(\theta) := S_n(\theta) - \frac{1}{2} H_n(\theta_\circ)$, which implies that $f_n(\theta) = \frac{1}{2}|\theta - \theta_\circ|_{H_n} + |\theta - \theta_\circ|_{h_n(\theta)}$ and

$$\lim_{\delta \to 0} \sup_n \sup_{\theta \in N^2_\delta} \|h_n(\theta)\|_2 \overset{(a)}{\le} d \lim_{\delta \to 0} \sup_n \sup_{\theta \in N^2_\delta} |h_n(\theta)|_{\max}$$

$$\overset{(b)}{=} d \lim_{\delta \to 0} \sup_n \sup_{\theta \in N^2_\delta} \max_{i,j} \left| S_n(\theta)_{i,j} - \frac{1}{2} H_n(\theta_\circ)_{i,j} \right|$$

$$\overset{(c)}{\le} d \max_{i,j} \lim_{\delta \to 0} \sup_n \sup_{\theta \in N^2_\delta} \left| \frac{1}{2} H_n(\theta)_{i,j} - \frac{1}{2} H_n(\theta_\circ)_{i,j} \right|$$

$$\overset{(d)}{=} 0$$

where (a) is follows from the bound $\| \cdot \|_2 \le d |\cdot|_{\max}$ for matrices of dimension d. (b) by the definition of $h_n(\theta)$. (c) by the definition of $S_n(\theta)$ and exchanging the max with the limit, which is valid because the indices i and j range over finite set. (d) by equicontinuity of the Hessian at θ_o. $\qquad\square$

Lemma 7. *If $A \in \mathbb{R}^{d \times d}$ is positive definite and $x \in \mathbb{R}^d$, then $|x|^2_2 \le |x|^2_A \|A^{-1}\|_2$. If $B \in \mathbb{R}^{d \times d}$ and $x \in \mathbb{R}^d$, then $|x|^2_B \le |x|^2_2 \|B\|^2_2$.*

Proof. Let $\lambda_{\min}$ be the smallest eigenvalue of A and $A = U^\top D U$ be the spectral decomposition of A where the diagonal of D consists of the eigenvalues of A and $U^{-1} = U^\top$. Then $|x|^2_2 = x^\top U^\top U x \le x^\top U^\top D U x / \lambda_{\min} = |x|^2_A / \lambda_{\min}$. Complete the first part by checking that $\|A^{-1}\|_2 = 1/\lambda_{\min}$. The second part is proven by noting that

$$\frac{|x|^2_B}{|x|^2_2} \le \max_{y:|y|^2_2 = |x|^2_2} \frac{|y|^2_B}{|y|^2_2} \equiv \|B\|^2_2.$$

Rearrange to complete the result. $\qquad\square$

Lemma 8. *Let $\sum_{j=0}^d \theta_j = 1$. Then the matrix $A \in \mathbb{R}^{d \times d}$ defined by*

$$A_{j,k} = \frac{[\![j = k]\!]}{\theta_j} + \frac{1}{\theta_0}$$

has determinant $\det A = \prod_{j=0}^d \frac{1}{\theta_j}$.

Proof. Define $\alpha_j = 1/\theta_j$ and subtract the last row of A from all previous rows to obtain a block matrix B with the same determinant as A of the form

$$
B = \begin{pmatrix} W & X \\ Y & Z \end{pmatrix} = \left(\begin{array}{ccccc|c}
\alpha_1 & 0 & \cdots & & 0 & -\alpha_d \\
0 & \alpha_2 & & & & -\alpha_d \\
\vdots & & \ddots & & & -\alpha_d \\
0 & \cdots & 0 & \alpha_{d-1} & & -\alpha_d \\
\hline
\alpha_0 & \alpha_0 & \alpha_0 & & \alpha_0 & \alpha_d + \alpha_0
\end{array} \right)
$$

Then

$$
\det A = \det W \det(Z - Y W^{-1} X)
$$

$$
= \left(\prod_{j=1}^{d-1} \alpha_j \right) \left(\alpha_d + \alpha_0 + \sum_{j=1}^{d-1} \frac{\alpha_d \alpha_0}{\alpha_j} \right)
$$

$$
= \left(\prod_{j=0}^{d} \alpha_j \right) \left(\frac{1}{\alpha_0} + \frac{1}{\alpha_d} + \sum_{j=1}^{d-1} \frac{1}{\alpha_j} \right)
$$

$$
= \left(\prod_{j=0}^{d} \alpha_j \right) \left(\sum_{i=0}^{d} \theta_i \right) = \prod_{j=0}^{d} \alpha_j \equiv \prod_{j=0}^{d} \frac{1}{\theta_j}
$$

as required. $\qquad\square$