Machine Learning, 28, 7–39 (1997)

A Bayesian/Information Theoretic Model of Learning to Learn via Multiple Task Sampling*

JONATHAN BAXTER jon@syseng.anu.edu.au
Department of Mathematics, London School of Economics
and
Department of Computer Science, Royal Holloway College, University of London

Editors: Lorien Pratt and Sebastian Thrun

Abstract. A Bayesian model of learning to learn by sampling from multiple tasks is presented. The multiple tasks are themselves generated by sampling from a distribution over an environment of related tasks. Such an environment is shown to be naturally modelled within a Bayesian context by the concept of an *objective* prior distribution. It is argued that for many common machine learning problems, although in general we do not know the true (objective) prior for the problem, we do have some idea of a set of possible priors to which the true prior belongs. It is shown that under these circumstances a learner can use Bayesian inference to learn the true prior by learning sufficiently many tasks from the environment. In addition, bounds are given on the amount of information required to learn a task when it is simultaneously learnt with several other tasks. The bounds show that if the learner has little knowledge of the true prior, but the dimensionality of the true prior is small, then sampling multiple tasks is highly advantageous. The theory is applied to the problem of learning a common feature set or equivalently a low-dimensional-representation (LDR) for an environment of related tasks.

Keywords: Hierarchichal Bayesian Inference, Bias learning, Feature Learning, Neural Networks, Information Theory

1. Introduction

Hume's analysis shows that there is no *a priori* basis for induction. In a machine learning context, this means that a learner must be biased in some way for it to generalise well (Mitchell, 1990). Typically such bias is introduced by hand through the skill and insights of experts, but despite many notable successes, this process is clearly limited by the experts' abilities. Hence a desirable goal is to find ways of automatically *learning* the bias. As knowing the right bias makes the learning problem easier, learning the bias can be viewed as a form of *learning to learn*.

In this paper a Bayesian model of bias learning is introduced, based on the VC/PAC-type models of bias learning introduced in (Baxter, 1995b, Baxter, 1996b). The central assumption of all these models (including that of the present paper) is that the learner is embedded within an *environment* of related tasks. The learner is able to sample from the environment and hence generate multiple data sets corresponding to different tasks. The learner can then search for a hypothesis space that is appropriate for learning all the tasks. Learning problems which can naturally be viewed as belonging to a large class of related

* Author's present address: Department of Systems Engineering, Australian National University, Canberra 0200, Australia.
This work was supported by EPSRC grants #K70366 and #K70373.

tasks are things like face recognition (each individual face classifier can be thought of as a separate learning problem), speech recognition (the word classifiers are related) character recognition, fingerprint recognition and so on.

For the learner to be able to search for an hypothesis space, it must be provided with a family of hypothesis spaces from which to choose. In (Baxter, 1995b, Baxter, 1996b) it was shown that under certain restrictions on this family of hypothesis spaces (these restrictions are analogous to the "finite VC dimension" restrictions on ordinary learners), it is possible for the learner to sample sufficiently often from sufficiently many tasks to ensure that a hypothesis space containing hypotheses with small empirical loss on all the tasks will, with high probability, contain good solutions to novel tasks drawn from the same environment. Thus, in this formal sense, it is possible for a learner to learn its own bias.

Whether or not there actually exists a hypothesis space containing good solutions to all the tasks will depend upon the family of hypothesis spaces provided to the learner, or equivalently upon the *hyper-bias* of the learner. Such hyper-bias must be provided by hand, which appears to beg the question, "haven't you just replaced the problem of finding the right bias with the equally difficult problem of finding the right hyper-bias?" Part of the purpose of this paper is to show that for many classes of learning problems (in particular those that possess a common set of *features*, or equivalently, a common *Low Dimensional Representation* (LDR) or preprocessing), the task of finding the correct hyper-bias is considerably easier than that of finding the right bias, if multiple tasks can be sampled. Intuitively, the reason for this is that there is a lot more information in multiple tasks than there is in a single task, and so the hyper-bias can be more weakly specified than the bias.

Learning multiple related tasks not only enables bias learning—in the sense that it improves the learner's performance on novel tasks—but it also improves generalisation performance on the tasks in the training set. In particular, it was shown in (Baxter, 1995b) that if the learner is learning a common feature set (LDR) for an n task training set then the number of examples m required of each task to ensure good generalisation on average across all n tasks obeys

$$m = O\left(a + \frac{b}{n}\right). \tag{1}$$

Here a is a measure of the dimension of the smallest hypothesis space needed to learn all the tasks in the environment and b is a measure of the dimension of the space of possible representations available to the learner. "Good generalisation" means that the learner's performance in practice, on overage across all n tasks, will be close to its average performance on the training sets. Note that this is an agnostic definition of good generalisation because it does not assume that the learner actually performs well in training.

The $n = 1$ case of formula (1)—$m = O(a + b)$—is an upper bound on the number of examples that would be required for good generalisation in the ordinary, single task learning scenario, while the limiting case of $m = O(a)$ is an upper bound on the number of examples required if the correct preprocessing is already known. Thus, this formula shows that the upper bound on the number of examples required per task for good generalisation decays to the minimum possible as the number of tasks being learnt increases.

Although very suggestive, without a matching lower bound of the same form, it is not possible to conclude from (1) that learning multiple related tasks requires fewer examples per task for good generalisation than if those tasks are learnt independently. Unfortunately, lower bounds within a real-valued VC/PAC framework can only be obtained by making extra assumptions, such as that the function values are corrupted by noise (Bartlett, Long & Williamson, 1994), or that every algorithm within a certain class of algorithms performs well (Anthony & Bartlett, 1995). Without such assumptions it is possible to construct (albeit artificial) scenarios in which every function within some class encodes its identity at every point (Bartlett, Long & Williamson, 1994). The problem arises because the output of a real-valued function can potentially encode an infinite amount of information.

The Bayesian model introduced here is an alternative way to overcome these limitations. In particular the concept of information (in a Shannon sense) is more naturally modeled within a Bayesian framework than in a VC/PAC setting, and so one can precisely formulate questions such as "how much information is required to learn". By asking this kind of question rather than "how many examples are required to learn" we get away from the difficulties mentioned in the previous paragraph. Another advantage of the Bayesian framework is that it is much easier to formulate and analyse the effects of prior knowledge on the learning process. This is particularly important in bias learning where one is trying to understand how the process of acquiring prior knowledge can be automated.

The main novel feature of this model is that the traditional Bayes prior distribution is treated as *objective*, rather than subjective. The sample space of the objective prior represents the space of tasks in the environment, and sampling from the prior corresponds to selecting different learning tasks from the environment. The reason the prior is regarded as objective is because it is assumed that it can be sampled from, *i.e.* it represents some objective stochastic phenomenon, in contrast to subjective priors which reflect the prior *beliefs* of the learner.

The analogous question to "how many examples are required of each task in an n task training set" leading to the upper bound (1), is "how much information is required per task to learn n tasks?" By using the usual Shannon definition of information, it is shown in subsection 3.1 that if the learner already knows the true (objective) prior then there is no advantage to learning n tasks; that is, the expected amount of information needed to learn each task within an n task training set is the same as if the tasks are learnt separately. However, if the learner does not know the true prior (which is generally the case in bias learning, otherwise there is no need to do bias learning), but instead knows only that the prior is one of a set Π of possible priors, then we will see that the expected amount of information required per task to learn n tasks, $\overline{R}_{n,\pi^*}$, obeys

$$\overline{R}_{n,\pi^*} \doteq a' + b'(\pi^*)\frac{\log n}{n} + o\left(\frac{\log n}{n}\right) \tag{2}$$

where a' is the minimum amount of information possible (the amount the learner would require if it knew the true prior $\pi^* \in \Pi$) and $b'(\pi^*)$ is a local measure of the dimension of the space of possible priors Π at the point π^*. Here $f(n,\pi^*) \doteq g(n,\pi^*)$ means $f(n,\pi^*) = g(n,\pi^*)$ for all but a set of π^* of vanishingly small measure as $n \to \infty$, and $o(\log n/n)$ stands for a function $f(n)$ satisfying $f(n)/(\log n/n) \to 0$. The "vanishingly

small measure" referred to above is a measure on Π, the set of possible priors, and hence is itself a *hyper-prior* distribution. The hyper-prior has no physical meaning, it simply reflects the initial beliefs of the learner as to which *priors* are more likely. Thus, in the terminology of the present paper, the hyper-prior is a *subjective* distribution.

Comparing (2) and (1) and the meaning of a and b with their partners a' and b', we see that (2) partially realizes the aim of providing an exact bound justifying learning multiple related tasks. In particular, (2) shows that the information required to learn each task within an n task training set decays to the *minimum* possible as the number of tasks is made arbitrarily large. One way of interpreting this is that the effect of the learner's ignorance concerning the true (objective) prior can be made arbitrarily small by learning sufficiently many tasks, or equivalently that any uncertainty the learner may have about the appropriate bias to use for the environment can be made arbitrarily small by learning sufficiently many tasks.

The difference between the amount of information required by the learner to learn the nth task *after* already learning $n-1$ tasks, and the amount of information required to learn the nth task if the learner *knows* the true (objective) prior is analysed in subsection 3.2. In particular, defining the *cumulative loss* of the learner, $\overline{C}_{n,\pi^*}$ to be the sum of the extra information required when learning the first, second, ..., nth task, it is shown that

$$\overline{C}_{n,\pi^*} \doteq \frac{b'(\pi^*)}{2} \log n + o(\log n). \tag{3}$$

The form of this equation as $\log n$ multiplied by the dimension of the space of possible priors around the true prior π^* is similar to results from ordinary Bayesian inference (Clarke & Barron, 1990).

The results of section 3 are purely concerned with the amount of information required to learn each task within an n task training set, they do not address the problem of how the information is obtained. In section 4 it is assumed that each task takes the form of a probability distribution over an observation space, and the information about the task is obtained by sampling from this distribution. This model covers a multitude of learning scenarios, from pattern classification to density estimation (see section 2). The question of how much information is required to encode the m'th observation of each task in an n task training set, $\overline{L}_{n,m,\pi^*}$, after seeing the first $m-1$ observations of each task, is analysed. In particular, general results in terms of metric dimension are given in subsection 4.1 for the *cumulative loss*, $\overline{C}_{n,m,\pi^*} = \sum_{k=0}^{m} \overline{L}_{n,k+1,\pi^*}$.

These result are specialized in section 4.2 to hierarchical models with $a+b$ real parameters, b of which are hyper-parameters and the remaining a of which are model parameters. That is, each possible different prior in Π is obtained by fixing b of the total set of parameters to some value, and then each individual learning problem with respect to that prior is obtained by fixing the remaining a parameters to some value. These models are called (a,b)-*models*. Neural networks for learning LDRs are (almost) (a,b)-models; they are considered in section 4.3. A second example based on learning the parameters of a normal distribution is given in section 4.4.

In section 4.2 it is shown that for (a,b)-models, the cumulative loss in predicting novel examples of each task satisfies

$$\overline{C}_{n,m,\pi^*} \doteq \frac{\log m}{2}\left(a + \frac{b}{n}\right) + o\left(\log m\right). \tag{4}$$

Compare this with the situation in which each task is learnt independently:

$$\overline{C}_{n,m,\pi^*} \doteq \frac{\log m}{2}\left(a + b\right) + o\left(\log m\right), \tag{5}$$

and the optimal loss achievable if the true prior is known:

$$\overline{C}_{n,m,\pi^*} \doteq \frac{\log m}{2}\left(a\right) + o\left(\log m\right). \tag{6}$$

Again we find that the learner's loss decays to the minimum possible as the number of tasks grows. Note the reappearance of the factor $a + b/n$.

The rest of the paper is organized as follows. The Bayesian model of bias learning is introduced formally in section 2, along with a concrete example based on learning a feature map or low-dimensional representation (LDR) with a neural network. Equations (2) and (3) are derived in section 3 and the constants a' and $b'(\pi^*)$ are calculated for the neural network example, where contact is made between the Bayesian model results and the VC/PAC model results of (Baxter, 1995b). Equations (4), (5) and (6) are derived in section 4, along with more general versions based only upon metric dimension concepts. These results are again applied to the neural network example in section 4.3, and once again comparison is made with the VC/PAC model results. To demonstrate that this theory is more generally applicable than just the LDR example, a second example based on learning the parameters of a normal distribution is given in section 4.4.

1.1. Related Work

Several authors have made empirical studies of the idea that learning multiple related tasks should improve performance, see *e.g.* (Caruana, 1993, Abu-Mostafa, 1989, Mitchell & Thrun, 1994). Experimental verification of this for feedforward nets was also reported in (Baxter, 1995b). The additional assumption that the tasks are distributed according to an *objective* distribution is what allows us to perform a theoretical analysis of this idea. This assumption was also made in (Baxter, 1995b, Baxter, 1996b). However, note that the theoretical model presented here does not apply directly to the experimental results of (Caruana, 1993) because there the training sets are not generated independently for each task.

The Bayesian aspect of the model presented here is a special case of what is known as *hierarchical Bayesian inference* (see e.g (Berger, 1985, Berger, 1986, Good, 1980)). To the best of my knowledge the asymptotic analysis given in this paper for these models is new, as is the consideration of the effect of the difference in the number of hyper-parameters and model parameters, and the application of these results to representation or feature-map learning with neural networks. Hierarchical Bayesian inference has also been discussed in the context of neural networks by several authors (see *e.g.* (Mackay, 1991), although the techniques presented there are not explicitly identified as hierarchical Bayes). As far

as I know the idea of an objective prior has not been employed previously in Bayesian approaches to neural networks. For the most part hierarchical Bayes has been used to tune a small number of "nuisance" (hyper) parameters (such as the parameter λ controlling the trade-off between regularisation and data-misfit in regression networks (Mackay, 1991)), and this tuning has been based on learning a *single* task.

The asymptotic results for smooth Euclidean models given in section 4.2 could also be derived more directly from the results of (Clarke & Barron, 1990). The motivation behind the approach taken here (which is based on the ideas in (Haussler & Opper, 1995a)) is that it provides results for general metric spaces, not just Euclidean models, although this is at the expense of losing lower order terms in the asymptotic estimates. Theorem 1 can also be derived via quite different techniques as a special case of theorem 2 in (Haussler & Opper, 1995b) (which appeared as an earlier incarnation of the present paper (Baxter, 1996a) was being prepared).

1.2. Notation

The probability model treated throughout this paper is three-tiered. At the bottom level is Z which is assumed to be (at least) a complete separable metric space. All probability measures on Z are defined on the sigma-field of Borel subsets of Z. Z is the learner's interface with the environment—the learner receives all its data in the form of samples from Z. For example, in density estimation Z would just be the input space X, while in classification $Z = X \times Y$ where X is the input space and $Y = \{0, 1\}$.

The next level up in the hierarchy is Θ, which is the set of possible "states of nature" or "learning tasks" with which the learner might be confronted. For each $\theta \in \Theta$ there is a probability measure $P_{Z|\theta}$ on Z. It is assumed that there exists a fixed σ-finite measure ν that dominates $P_{Z|\theta}$ for each $\theta \in \Theta$. Θ is also assumed to be a complete separable metric space. At the highest level in the hierarchy is the set Π which represents the space of possible "priors" on Θ. For each $\pi \in \Pi$ there is a probability measure $P_{\Theta|\pi}$ on Θ. Again the $P_{\Theta|\pi}$'s are defined on the sigma field of Borel subsets of Θ and it is assumed that there exists a second measure μ dominating all $P_{\Theta|\pi}$. Finally, on Π there is a fixed probability measure P_Π: the "hyper-prior". As Θ is a complete separable metric space, the domain of P_Π can be taken to be the sigma field generated by the topology of weak convergence of the $P_{\Theta|\pi}$ measures. Let supp$P$ denote the support of measure P.

Where multiple instances of the same space need to be distinguished, the extra copies will be denoted by primes (Z') or tildes ($\tilde{Z}$).

Integration with respect to the measures ν and μ will be denoted by $\int_Z dz$ and $\int_\Theta d\theta$ respectively (ν and μ are not assumed to be Lebesgue measures—the notation is just for convenience). Integration with respect to the hyper-prior P_Π will be denoted $\int_\Pi p(\pi)\, d\pi$. The Radon-Nikodym derivative of any measure $P_{Z|\theta}$ at $z \in Z$, $\frac{dP_{Z|\theta}}{d\nu}(z)$ will be written interchangeably as $p(z|\theta)$ or $p_{Z|\theta}(z)$, and similarly $\frac{dP_{\Theta|\pi}}{d\mu}(\theta)$ will be written as $p(\theta|\pi)$ or $p_{\Theta|\pi}(\theta)$.

If f is a function on Z, then the expectation of f with respect to any random variable with distribution $P_{Z|\theta}$ will be denoted by $E_{Z|\theta} f(z) = \int_Z f(z) p(z|\theta)\, dz$. Similarly for functions defined on Θ and Π.

$n \times m$ matrices with elements from Z will be denoted by $z^{(n,m)}$:

$$z^{(n,m)} = \begin{matrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm}. \end{matrix} \tag{7}$$

The columns of $z^{(n,m)}$ will be denoted as z_i^n, so $z^{(n,m)} = [z_1^n \ldots z_m^n]$.

Let N denote the natural numbers.

2. The Basic Model

In Bayesian models of learning (see *e.g.* (Berger, 1985)) the learner receives data $z^n = z_1, \ldots, z_n$ which are observations on n random variables $Z^n = Z_1, \ldots, Z_n$. The Z_i are identically distributed and conditionally independent given the true state of nature θ. The learner does not know θ, but does know that θ belongs to a set of possible states of nature Θ. The learner begins with a prior distribution on Θ, $p(\theta)$, and upon receipt of the data z^n updates $p(\theta)$ to a posterior distribution $p(\theta|z^n)$ according to Bayes' rule:

$$p(\theta|z^n) = \frac{p(z^n|\theta)p(\theta)}{p(z^n)}, \tag{8}$$

where

$$p(z^n) = \int_\Theta p(z^n|\theta)p(\theta)\, d\theta. \tag{9}$$

2.1. Bayesian inference and Neural Networks

Pattern classification or regression with neural networks may be viewed as a special case of the above. To fix our ideas, consider the case of an MLP for recognising my face. The weights of the network correspond to the set of possible states of nature Θ, the true state of nature θ^* being an assignment of weights such that the output of the network is 1 when an example of my face is applied to its input, and 0 if anything else is applied to its input. The data $z^n = z_1, \ldots, z_n$ comes in the form of input-output pairs $z_i = (x_i, y_i)$ where each x_i is an example image and y_i is the correct class label (in this case either 0 or 1). As we are only interested in classification in this example, the input distribution $p(x)$ is not modeled, only the conditional distribution on class labels $p(y|x)$. Denoting the output of a network with weights θ by $f_\theta(x)$, and interpreting $f_\theta(x)$ as $p(y = 1|x)$, it can easily be shown (Bridle, 1989) that the probability of data set $z^n = (x_1, y_1), \ldots, (x_n, y_n)$ given weights θ is

$$p(z^n|\theta) = \prod_{i=1}^{n} p(x_i) e^{-E(z^n;\theta)} \tag{10}$$

where

$$E(z^n;\theta) = \sum_{i=1}^{n} y_i \log(f_\theta(x_i)) + (1-y_i)\log(f_\theta(x_i)). \tag{11}$$

Choosing a prior $p(\theta)$ (typically multivariate Gaussian or uniform over some compact set) for the weights and substituting (10) into (8) yields the posterior distribution on the weights $p(\theta|z^n)$. The posterior is the "output" of the learning process. It can be used to predict the class label of a novel input x^* by integrating:

$$p(y=1|x^*;z^n) = \int_\Theta f_\theta(x^*) p(\theta|z^n)\, d\theta. \tag{12}$$

Of course in general this integral cannot be calculated in closed form and so some kind of approximation procedure such as Markov-Chain Monte-Carlo must be used for its evaluation. In this paper we do not concern ourselves with such computational issues, except to note that the common practice of choosing the weights with minimal error is equivalent to approximating the posterior by a delta function at the maximum-likelihood weight setting.

2.2. Interpreting the Prior

In the example above the prior $p(\theta)$ is a purely *subjective* prior. A relatively weak prior was chosen reflecting our weak knowledge about appropriate weight settings for this problem. However, in the case of face recognition (and many other pattern recognition problems such as speech and character recognition) it is arguable that there exist *objective* priors. To see this, note that given our weak prior knowledge we are likely to have chosen a network large enough to solve *any* face recognition problem within some margin of error, not just the specific task: "recognise Jon ". Hence it is likely that there will exist weight settings $\theta_1, \theta_2, \theta_3, \ldots$ that will cause the network to behave as a classifier for 'Mary', 'Joe', 'males', 'smiling', 'big nose' and so on. In fact there should exist weight settings that correspond to nonexistent faces provided different examples of the face vary in a "face-like" way. Hence we can consider the space of all face classifiers, both real and fictitious, as represented by a particular subset Θ_{face} of all possible weight settings Θ. The *objective prior* $p(\theta)$ for face recognition is then characterised by the fact that its support is restricted to Θ_{face}. The restriction of the support is the most important aspect of the face prior. The actual numerical probabilities for each element $\theta \in \Theta_{\text{face}}$ could be chosen in a number of different ways, but for the sake of argument we can take them to be uniform or as corresponding to the general frequency of face-like classifier problems encountered in a particular person's environment. In general different people will have different environments and so there will actually be multiple different objective priors for the face recognition problem. However, this does not change the fact that the face prior is objective—it is objective precisely because it is defined by the *environment* of the learner and not by a set of subjective beliefs. Note also that

different people embedded within the same environment—say primarily Caucasian faces, or primarily Asian faces—will have essentially the same objective priors.

The usual subjective priors chosen in neural network applications (Gaussian or uniform on the weights) bear no resemblance to the objective prior discussed above: initializing the weights of a network according to a Gaussian prior typically does not cause the network to behave like some kind of face classifier, whereas initializing according to the objective prior by definition will induce such behaviour. Hence the use of subjective priors such as the Gaussian not only demonstrates our ignorance concerning the specific task at hand (*e.g.* learn to recognise Jon) but also demonstrates our ignorance concerning the true prior. That is, we typically have little idea which parameter settings θ correspond to face-like classifiers and which correspond to "random junk".

Should we care that we don't know the true prior? In short, yes. If we know the true prior then the task of learning any individual face is vastly simplified. A single positive example of my face is enough to set the posterior probability of any other individual face classifiers to zero (or very close to zero), and a few more examples with me smiling, frowning, bearded, clean-shaven, long-haired, short-haired and so on is enough to set the posterior probability of *every* other classifier (the smiling, frowning, *etc* classifiers) except the "Jon" classifier to zero. Contrast this with the usual subjective priors where typically thousands of examples and counter-examples of my face would have to be supplied to the network before a reasonably peaked posterior and hence reasonable generalisation could be achieved.

2.3. Learning the Prior

If knowing the true prior is such a great advantage then we should try to learn it. To do this an extra layer of inference must be added to the standard Bayesian model in the form of a *set* of candidate priors Π. Thus, each $\pi \in \Pi$ corresponds to some prior $p(\theta|\pi)$ on Θ. Realizability is assumed, so that the true objective prior $p(\theta|\pi^*)$ corresponds to some $\pi^* \in \Pi$. To complete the Bayesian picture a *subjective* hyper-prior $p(\pi)$ must be chosen for Π. The hyper-prior $p(\pi)$ is subjective, rather than objective, because it cannot be sampled, that is it does not correspond to some objective stochastic phenomenon in the way that the objective prior $p(\theta|\pi^*)$ does. Typically the learner will not have a strong preference for any particular prior and so we can follow the course taken in ordinary Bayesian inference under such circumstances and choose $p(\pi)$ to be non-informative or simply Gaussian with large variance or uniform over some compact set (assuming Π is Euclidean).

As the true prior $p(\theta|\pi^*)$ is objective it can *in principle* be sampled from to generate a sequence of training *tasks*[1] $\theta^n = \theta_1, \theta_2, \ldots, \theta_n$. A direct application of Bayes' rule (8) then gives the posterior probability of each prior:

$$p(\pi|\theta^n) = \frac{p(\theta^n|\pi)p(\pi)}{p(\theta^n)} \tag{13}$$

where $p(\theta^n|\pi) = \prod_{i=1}^n p(\theta_i|\pi)$ and $p(\theta^n) = \int_\Pi p(\theta^n|\pi)p(\pi)\, d\pi$.

Under appropriate conditions the posterior distribution will tend to a delta function over the true prior π^* as $n \to \infty$. Thus for large enough n the learner can be said to have *learnt the prior*.

2.4. Example: Learning a Low Dimensional Representation

For this model to work it has to be assumed that although the learner has no idea about the true prior, it can generate a class of priors Π containing the true prior π^*. This assumption is quite reasonable in the case of face recognition because it seems plausible that there exists a *low-dimensional representation* (LDR) or feature map for faces such that each face classifier can be implemented by a simple map (*e.g.* linear or nearest-neighbour) composed with the LDR. An LDR in its simplest form is just a fixed mapping from the (typically high-dimensional) input space to a much smaller dimensional space. One can think of the LDR as a preprocessing applied to the input data that extracts features that are important for classification. For example, in the case of face recognition it might be that to uniquely determine any face one only needs to know the distance between the eyes and the length of the nose. So an appropriate LDR would be a two-dimensional one that extracts these two features from an image. Although faces almost certainly cannot be represented solely by the inter-eye distance and nose length, it is highly plausible that some kind of LDR exists for the face recognition problem. It is similarly plausible that LDRs exist for other pattern recognition problems such as character and speech recognition.

Figure 1 illustrates how in the case of neural-networks the assumption that there exists an LDR for the tasks in the environment can be translated into a specification for the set of possible priors Π. Referring to the figure, the input space has dimension d, the LDR is implemented by a two-layer sigmoidal net with l hidden units followed by k output units (any bounded k-dimensional LDR can be approximated to arbitrarily high accuracy by such a two-layer structure (see *e.g.* (Hornik, 1991))), while each individual classifier task is assumed to be a sigmoidal map composed with the output of the LDR. Thus each $\theta \in \Theta$ divides into two parts: $\theta = (\theta_{\rm LDR}, \theta_{\rm OUT})$, where $\theta_{\rm LDR}$ are the hidden layer weights and $\theta_{\rm OUT}$ are the weights of the output map. Assuming that the true preprocessing for the environment corresponds to some assignment of weights to the hidden layers, $\theta_{\rm LDR} = \theta^*_{\rm LDR}$, the true prior can be written as

$$p(\theta_{\rm LDR}, \theta_{\rm OUT}) = \delta(\theta_{\rm LDR} - \theta^*_{\rm LDR}) f(\theta_{\rm OUT}) \tag{14}$$

where δ is the Dirac delta-function and $f(\theta_{\rm OUT})$ is some distribution over the output weights that generates the different tasks in the environment. Thus it is reasonable to take Π to be the set of all priors that are a delta function over some $\theta_{\rm LDR}$, with the distribution $f(\theta_{\rm OUT})$ over the output weights:

$$\Pi = \left\{ \delta(\theta_{\rm LDR} - \hat{\theta}_{\rm LDR}) f(\theta_{\rm OUT}) : \hat{\theta}_{\rm LDR} \in \Theta_{\rm LDR} \right\}. \tag{15}$$

Thus Π is equivalent to the set of possible weights in the hidden layers, $\Theta_{\rm LDR}$. Note that assuming Π is of this form means that the learner must know the true distribution

$f(\theta_{\mathrm{OUT}})$ on the output weights. If the learner does not know $f(\theta_{\mathrm{OUT}})$ but does know that $f(\theta_{\mathrm{OUT}})$ belongs to a parameterised set of distributions Π_{OUT} (*e.g.* multi-variate Gaussians with unknown means and covariance), then the parameters parameterizing Π_{OUT} can be adjoined to Π, which will then be of the form

$$\Pi = \left\{ \delta(\theta_{\mathrm{LDR}} - \hat{\theta}_{\mathrm{LDR}}) f_{\pi_{\mathrm{OUT}}}(\theta_{\mathrm{OUT}}) : \hat{\theta} \in \Theta_{\mathrm{LDR}}, \pi_{\mathrm{OUT}} \in \Pi_{\mathrm{OUT}} \right\}. \tag{16}$$

2.5. Hyper-parameters outnumber parameters

In this model knowing the true prior is equivalent to knowing the correct hidden layer weights (and the true parameters for the output weight distribution). So if the true prior is known, learning any individual task is simply a matter of estimating the output weights for a single node (which is a just a linear regression or linear classification problem). Thus the output layer weights are *model parameters* while the hidden layer weights (and the parameters of the output weight distribution) are the *model hyper-parameters*. In contrast to other techniques for Bayes learning with neural networks in which there are at most a handful of hyper-parameters (see *e.g.* (Mackay, 1991)), here the hyper-parameters vastly outnumber the model parameters. This happens because we have assumed that the learner's uncertainty concerning the true model (or equivalently, the true prior) is large, while the dimensionality of the true model is in fact quite low. For many real-world learning environments this seems to be a plausible assumption. For example, for the environment of face-recognition problems, we have a fairly large uncertainty concerning the true model, but human performance on these kinds of problems (*e.g.* our ability to recognise faces from single examples) shows that the true model must be very small. Another example is speech recognition. Considering all individual spoken words as constituting a "speech recognition environment", it is true that we have little idea of what the true model is for this environment, but again human performance suggests that the true model must be small. Many other pattern recognition problems are arguably best modeled by a two-tier inference structure in which the hyper-parameters vastly outweigh the model parameters. In the remainder of the paper we will see how such a two-tiered structure can lead to great improvements in learning performance if multiple tasks, rather than just a single task, are learnt.

3. Learning Multiple Tasks

Having set up the model of Bayesian bias learning in the previous section, we can now tackle the questions posed in the introduction: "How much information is required per task to learn n tasks simultaneously?" and "How much extra information is required to learn a sequence of tasks when the true prior is unknown?".

3.1. Learning n tasks simultaneously

Note that if the learner already knows the true prior $p(\theta|\pi^*)$, then the expected amount of information required per task to learn n tasks is

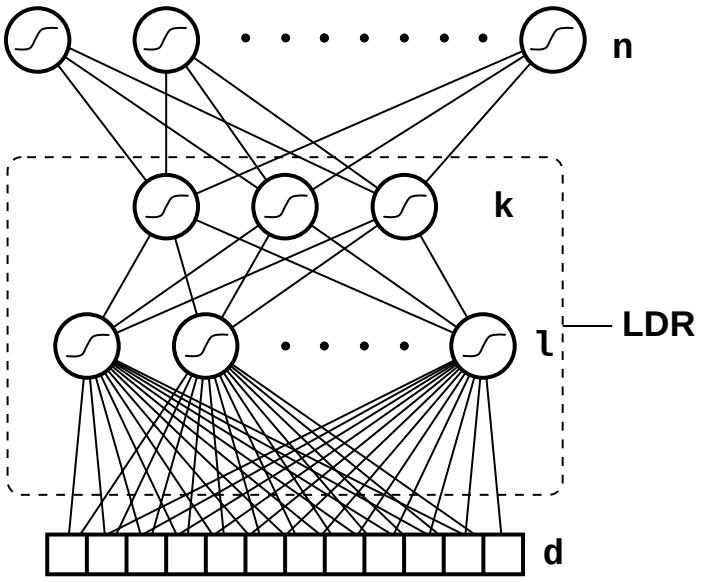

Figure 1. A neural network for learning low dimensional representations (LDRs) via multi-task sampling. Each output node corresponds to a different task. Each task is assumed to be implementable by composing a squashed linear map with a fixed LDR. The LDR is assumed to be implementable by a single-hidden-layer sigmoidal net. The LDR weights are *hyper-parameters*, while the the output weights for a single node are ordinary model parameters.

$$\frac{H(P_{\Theta^n|\pi^*})}{n} = H(P_{\Theta|\pi^*}) \tag{17}$$

because $P_{\Theta^n|\pi^*} = P^n_{\Theta|\pi^*}$ and entropy is additive over products of independent distributions (here $H(P_{\Theta|\pi^*}) = -E_{\Theta|\pi^*} \log p(\theta|\pi^*)$ is the *entropy* of the true prior). As $H(P_{\Theta|\pi^*})$ is the expected amount of information required to learn a single task, (17) shows that there is no advantage to learning multiple tasks if the true prior is known.

If the true prior is unknown, but the learner is in possession of a family of priors Π containing the true prior $P_{\Theta|\pi^*}$, then the expected amount of information required per task to learn n tasks is

$$\overline{R}_{n,\pi^*} := \frac{H_{\pi^*}(P_{\Theta^n})}{n}, \tag{18}$$

where $H_{\pi^*}(P_{\Theta^n}) := -E_{\Theta^n|\pi^*} \log p(\theta^n)$ where

$$p(\theta^n) = \int_\Pi p(\theta^n|\pi)p(\pi)\, d\pi \tag{19}$$

is the density of the *induced* or *mixture* prior on θ^n, P_{Θ^n}. Note that $-\log p(\theta^n)$ is (within one bit) the optimal amount of information required to encode the n tasks θ^n under the

distribution $p(\theta^n)$, and that $p(\theta^n)$ is the *consistent* distribution for the learner to use, given its prior beliefs as encapsulated in the hyper-prior $p(\pi)$. As the tasks are selected according to the true prior $P_{\Theta|\pi^*}$, we see that $H_{\pi^*}(P_{\Theta^n})/n$ is indeed the expected amount of information required (per task) to learn n tasks.

Rather than tackling $\overline{R}_{n,\pi^*}$ directly it is more convenient to analyse the expected difference between the information required to learn n tasks using the true prior $p(\theta^n|\pi^*)$ and the information required to learn n tasks using the induced prior $p(\theta^n)$. This quantity is

$$\int_{\Theta^n} p(\theta^n|\pi^*) \log \frac{p(\theta^n|\pi^*)}{p(\theta^n)}\, d\theta^n =: D_K(P_{\Theta^n|\pi^*} \| P_{\Theta^n}), \tag{20}$$

which is the *Kullback-Liebler divergence* between the true and induced distributions on Θ^n. Note that if we know $D_K(P_{\Theta^n|\pi^*} \| P_{\Theta^n})$, we can recover $\overline{R}_{n,\pi^*}$ from the relation

$$\overline{R}_{n,\pi^*} = \frac{1}{n} D_K(P_{\Theta^n|\pi^*} \| P_{\Theta^n}) + H(P_{\Theta|\pi^*}) \tag{21}$$

To bound $D_K(P_{\Theta^n|\pi^*} \| P_{\Theta^n})$ the following definitions are needed.

Definition 1 *For any $\pi, \pi' \in \Pi$, let $\Delta_H(\pi, \pi')$ denote the squared Hellinger distance between the two priors $P_{\Theta|\pi}$ and $P_{\Theta|\pi'}$:*

$$\Delta_H(\pi, \pi') := \int_\Theta \left[\sqrt{p(\theta|\pi)} - \sqrt{p(\theta|\pi')} \right]^2 d\theta \tag{22}$$

and let $\Delta_K(\pi, \pi')$ denote the Kullback-Liebler divergence between the two priors $P_{\Theta|\pi}$, $P_{\Theta|\pi'}$:

$$\Delta_K(\pi, \pi') := D_K\left(P_{\Theta|\pi} \| P_{\Theta|\pi'}\right) = \int_\Theta p(\theta|\pi) \log \frac{p(\theta|\pi)}{p(\theta|\pi')}\, d\theta. \tag{23}$$

Let $B_\varepsilon(\pi) := \{\pi' \colon \Delta_H^{1/2}(\pi, \pi') \le \varepsilon\}$, i.e. *the closed Hellinger ball of radius ε around π. For all $\pi \in \Pi$, define the* local metric dimension of π *by*

$$\dim_{P_\Pi}(\pi) := \lim_{\varepsilon \to 0} \frac{-\log P_\Pi(B_\varepsilon(\pi))}{\log \frac{1}{\varepsilon}} \tag{24}$$

whenever the limit exists (P_Π is the subjective (hyper) prior probability distribution on Π).

Note that $(\Pi, \Delta_H^{1/2})$ is a metric space while (Π, Δ_K) is not (Δ_K is asymmetric and does not satisfy the triangle inequality). Also, $\Delta_K(\pi, \pi') \ge \frac{1}{2}\Delta_H(\pi, \pi')$ always (see *e.g.* (Haussler & Opper, 1995a)). To get a feel for the meaning of $\dim_{P_\Pi}(\pi)$, observe that if $\Pi = R^d$ and P_Π has a continuous density $p(\pi)$, then for any $\pi \in R^d$ with $p(\pi) > 0$, $\dim_{P_\Pi}(\pi) = d$.

Definition 2 *Let (X, Σ, P) be a measure space and $f, g \colon N \times X \to R$ be two real-valued functions on $N \times X$ such that for all $n \in N$, $f(n, \cdot)$ and $g(n, \cdot)$ are measurable functions on X. Set $X_n := \{x \colon f(n, x) = g(n, x)\}$ for each $n \in N$. We say*

$$f(n,x) \doteq_{(X,P)} g(n,x) \tag{25}$$

if $\lim_{n\to\infty} P(X_n) = 1$. *This will be abbreviated to* $f(n,x) \doteq g(n,x)$ *when* X *and* P *are clear from the context.*

Theorem 1 *If there exists* $\alpha < \infty$ *such that for all* $\pi, \pi' \in \Pi$,

$$\Delta_K(\pi,\pi') \leq \alpha \Delta_H(\pi,\pi'), \tag{26}$$

and $\dim_{P_\Pi}(\pi)$ *exists for almost all* (P_Π) $\pi \in \Pi$, *then*

$$\frac{D_K(P_{\Theta^n|\pi^*} \| P_{\Theta^n})}{\log n} \doteq_{(\Pi,P_\Pi)} \frac{\dim_{P_\Pi}(\pi^*)}{2} + o(1), \tag{27}$$

where $o(g(n))$ *for any function* $g(n)$ *stands for a function* $f(n)$ *for which* $\lim_{n\to\infty} f(n)/g(n) = 0$.

Proof: See appendix A. ■

Note that if

$$\sup_{\pi,\pi^* \in \Pi \text{ and } \theta \in \Theta} \frac{p(\theta|\pi)}{p(\theta|\pi^*)} < \infty \tag{28}$$

then there exists $\alpha < \infty$ such that $\Delta_K(\pi,\pi') \leq \alpha \Delta_H(\pi,\pi')$ (Haussler & Opper, 1995a).

Theorem 2 *Under the same conditions as theorem 1,*

$$\overline{R}_{n,\pi^*} \doteq \frac{\dim_{P_\Pi}(\pi^*)}{2} \frac{\log n}{n} + H(P_{\Theta|\pi^*}) + o\left(\frac{\log n}{n}\right). \tag{29}$$

Proof: The theorem follows directly from (21) and theorem 1. ■

Note that this result is not quite as strong as it looks on face value because the set of priors for which

$$\overline{R}_{n,\pi^*} = \frac{\dim_{P_\Pi}(\pi^*)}{2} \frac{\log n}{n} + H(P_{\Theta|\pi^*}) + o\left(\frac{\log n}{n}\right) \tag{30}$$

fails can *vary with* n, even though its measure becomes vanishingly small as $n \to \infty$. This implies that for any individual $\pi^* \in \Pi$, (30) may fail for infinitely many n. However, if the sum over all n of the P_Π measure of the sets of π^* for which (30) fails is finite, then by Borel-Cantelli, for all but a set of π of P_Π measure zero, (30) will fail only *finitely* often.

Setting $a = H(P_{\Theta|\pi^*})$ and $b = \dim_{P_\Pi}(\pi^*)$, theorem 2 shows that the expected amount of information required per task to learn an n task training set approaches

$$a + \frac{b \log n}{2n}, \tag{31}$$

except for a set of priors of vanishingly small measure as $n \to \infty$, which in turn approaches a—the minimum amount of information required to learn a task on average (a is the amount of information required if the true prior is known, *c.f.* (17)). Observe that the advantage in learning n tasks is controlled by the relative size of a and b, and is greatest when $b \gg a$. As b is a measure of our uncertainty concerning the true prior, the greatest advantage in learning multiple tasks occurs when the true model is small, but we have little idea about what the true model should be. It is a plausible hypothesis that many pattern recognition problems (such as speech, face and character recognition) fit this bill.

3.2. Learning n tasks sequentially

Consider the same set-up as above, but now instead of learning the n tasks simultaneously, the learner receives each task one at a time. So for each $n = 1, 2, \ldots$ the learner has already seen $n-1$ tasks, $\theta^{n-1} = (\theta_1, \ldots, \theta_{n-1})$, drawn according to the true prior $p(\theta|\pi^*)$. The learner then:

- generates the posterior distribution on Π, $p(\pi|\theta^{n-1})$ according to Bayes' rule (13),
- uses the posterior distribution to generate a predictive distribution on Θ,

$$p(\theta|\theta^{n-1}) = \int_{\Pi} p(\theta|\pi) p(\pi|\theta^{n-1})\, d\pi, \tag{32}$$

- and suffers a loss, $\overline{L}_{n,\pi^*}$, equal to the expected amount of extra information needed to encode each task using the predictive distribution $p(\theta|\theta^{n-1})$, over and above the amount of information that would be required if it was using the true prior:

$$\overline{L}_{n,\pi^*} := E_{\Theta^{n-1}|\pi^*} E_{\Theta'|\pi^*} \log \frac{p(\theta'|\pi^*)}{p(\theta'|\theta^{n-1})}. \tag{33}$$

Note that $\overline{L}_{n,\pi^*}$ is the expected loss of the learner over all initial sequences θ^{n-1} and over all new tasks θ'. The quantity analysed in this section is the *cumulative loss*

$$\overline{C}_{n,\pi^*} := \sum_{k=0}^{n-1} \overline{L}_{k+1,\pi^*}, \tag{34}$$

i.e. the total loss incurred by the learner after n steps of the above process.

Theorem 3 *Under the same conditions as theorem 1,*

$$\overline{C}_{n,\pi^*} \doteq \frac{\dim_{P_\Pi}(\pi^*)}{2} \log n + o(\log n). \tag{35}$$

Proof: Direct calculation shows that

$$\overline{C}_{n,\pi^*} = D_K(P_{\Theta^n|\pi^*} \| P_{\Theta^n}) \tag{36}$$

where P_{Θ^n} is the mixture prior on Θ^n induced by the hyper-prior P_Π (recall equation (19)). The result now follows from theorem 1. ■

The nice thing about (35) is that the cumulative loss only diverges logarithmically, so the expected loss per trial, $\overline{C}_{n,\pi^*}/n$, tends to zero at a rate $\log n/n$.

3.3. Example: learning an LDR

Recall from section 2.4 that for the problem of learning a Low Dimensional Representation (LDR), Θ was split into $(\Theta_{\rm LDR}, \Theta_{\rm OUT})$. Each prior $\pi \in \Pi$ was chosen to be a delta function over some $\theta_{\rm LDR}$, multiplied by a fixed distribution $f(\theta_{\rm OUT})$ over $\Theta_{\rm OUT}$. In order to apply the results of the previous subsection the delta function needs to be smoothed out, otherwise the correct prior is identifiable from the observation of a single task[2] θ. So instead assume the prior corresponding to each π is of the form

$$p(\theta_{\rm OUT}, \theta_{\rm LDR}|\pi) := p(\theta_{\rm LDR}|\pi) f(\theta_{\rm OUT}) \tag{37}$$

where $p(\theta_{\rm LDR}|\pi)$ is a Gaussian with small variance σ_Π and mean $\theta_{\rm LDR}(\pi)$. In addition, for $H(P_{\Theta|\pi})$ to be well defined (*i.e.* finite) the output weights $\theta_{\rm OUT}$ need to be quantized, so let each weight w be coded with k bits and (somewhat arbitrarily) choose the distribution f over the discretized $\Theta_{\rm OUT}$ to be uniform for each prior π. Denote the number of weights in $\Theta_{\rm LDR}$ by $W_{\rm LDR}$ and the number of weights in $\Theta_{\rm OUT}$ by $W_{\rm OUT}$. Finally, choose the hyper-prior distribution P_Π on Π to be uniform over some compact subset of $\Theta_{\rm LDR}$.

A simple calculation shows the Hellinger and Kullback-Liebler distances to be given by

$$\Delta_H(\pi, \pi') = 2\left(1 - \exp\left(-\frac{1}{8\sigma_\Pi^2}\|\theta_{\rm LDR}(\pi) - \theta_{\rm LDR}(\pi')\|^2\right)\right), \tag{38}$$

$$\Delta_K(\pi, \pi') = \frac{1}{2\sigma_\Pi^2}\|\theta_{\rm LDR}(\pi) - \theta_{\rm LDR}(\pi')\|^2 \tag{39}$$

Note that as $\Delta_H(\pi, \pi') \to 0$, $\Delta_H(\pi, \pi') \to \frac{1}{4\sigma_\Pi^2}\|\theta_{\rm LDR}(\pi) - \theta_{\rm LDR}(\pi')\|^2$. Substituting this expression into the definition of $\dim_{P_\Pi}(\pi)$ we find

$$\dim_{P_\Pi}(\pi) = W_{\rm LDR} \tag{40}$$

for all $\pi \in \Pi$. Trivially, $H(P_{\Theta|\pi}) = kW_{\rm OUT}$ for all $\pi \in \Pi$. The fact that the prior on Π is compactly supported coupled with the use of Gaussian priors on Θ ensures that $\Delta_K(\pi, \pi')$ is bounded above by $\alpha\Delta_H(\pi, \pi')$ for all π, π' and some $\alpha < \infty$. Hence the conditions of theorem 2 are satisfied and we have

$$\overline{R}_{n,\pi^*} \doteq \frac{W_{\rm LDR}}{2}\frac{\log n}{n} + kW_{\rm OUT} + o\left(\frac{\log n}{n}\right). \tag{41}$$

The similarity of this expression to the upper bound on the number of examples required per task for good generalisation in a PAC sense of $O(W_{\rm OUT} + W_{\rm LDR}/n)$ is noteworthy

(see (Baxter, 1995b) for a derivation of the latter expression). Note how the amount of information required to learn each task decays to $kW_{\rm OUT}$ as the number of tasks being learnt increases. $kW_{\rm OUT}$ is the *minimum* amount of information necessary to learn an individual task, *i.e.* the amount of information needed if the true prior is known. Note also that the advantage in learning multiple tasks is greatest if $W_{\rm LDR} \gg W_{\rm OUT}$, *i.e.* if the number of hyper-parameters greatly outweighs the number of model parameters.

4. Sampling multiple tasks

Theorems 2 and 3 were derived under the assumption that the learner receives information about the tasks θ directly. In fact $\overline{R}_{n,\pi^*}$ is (within one query) the average number of *queries* the learner will require per task to identify n tasks if the queries are restricted to be of the form "is $\theta^n \in A$" where A is any subset of Θ^n and the learner uses the best possible querying strategy.

In general the learner will not be able to query in this way, but will instead receive information about the parameters θ indirectly via a training set $z^m = (z_1, \ldots, z_m)$, sampled i.i.d. according to $p(z|\theta)$. If the learner is learning n tasks simultaneously then it will receive n such samples (called an *(n,m)-sample* in (Baxter, 1995b, Baxter, 1995a)):

$$z^{(n,m)} = \begin{matrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{matrix} \tag{42}$$

Each row of $z^{(n,m)}$ is sampled according to $p(z|\theta_i)$ where $\theta_1, \ldots, \theta_n$ are the n tasks being learnt. Let $Z^{(n,m)}$ denote the set of all such $z^{(n,m)}$. The correct hierarchical Bayes approach to learning the n tasks $\theta_1, \ldots, \theta_n$ is to use the hyper prior P_Π to generate a prior distribution on Θ^n via

$$\begin{aligned} p(\theta^n) &= \int_\Pi p(\theta^n|\pi)p(\pi)\, d\pi \\ &= \int_\Pi p(\pi) \prod_{i=1}^n p(\theta_i|\pi)\, d\pi \end{aligned}$$

and then the posterior $p(\theta^n|z^{(n,m)})$ can be computed according to Bayes' rule

$$\begin{aligned} p(\theta^n|z^{(n,m)}) &= \frac{p(z^{(n,m)}|\theta^n)p(\theta^n)}{p(z^{(n,m)})} \\ &= \frac{p(\theta^n)\prod_{i=1}^n \prod_{j=1}^m p(z_{ij}|\theta_i)}{p(z^{(n,m)})} \end{aligned} \tag{43}$$

where $p(z^{(n,m)}) = \int_{\Theta^n} p(\theta^n) \prod_{i=1}^n \prod_{j=1}^m p(z_{ij}|\theta_i)\, d\theta^n$.

4.1. Loss as the extra information required to predict the next observation

One way to measure the advantage in learning n tasks together is by the rate at which the learner's loss in predicting novel examples decays for each task. This is the same as the approach taken in section 3.2, but now we are considering the more realistic situation in which the learner receives information about each task θ_i indirectly via a sample z^m from $P_{Z|\theta_i}$. So fix the number of tasks n, sample n tasks $\theta^n = \theta_1, \ldots, \theta_n$ according to the true prior $P_{\Theta|\pi^*}$, and then for each $m = 1, 2, \ldots$ the learner has already seen $m-1$ examples of each task

$$z^{(n,m-1)} = \begin{matrix} z_{11} & \cdots & z_{1m-1} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm-1} \end{matrix} \tag{44}$$

where each row is drawn according to $P^{m-1}_{Z|\theta_i}$ (or equivalently, each column is drawn according to $P_{Z^n|\theta^n}$). The learner then:

- generates the posterior distribution on Θ^n, $p(\theta^n|z^{(n,m-1)})$ according to Bayes' rule (43),
- uses the posterior distribution to generate a predictive distribution on Z^n,

$$p(z^n|z^{(n,m-1)}) = \int_{\Theta^n} p(z^n|\theta^n)p(\theta^n|z^{(n,m-1)})\, d\theta^n, \tag{45}$$

- and suffers a loss, $\overline{L}_{n,m}$, equal to the expected amount of extra information needed *per task* to encode a novel example of each task using the predictive distribution $p(z^n|z^{(n,m-1)})$, over and above the amount of information that would be required if it was using the true distribution, $p(z^n|\theta^n)$:

$$\overline{L}_{n,m} := \frac{1}{n} E_{Z^n|\theta^n} \log \frac{p(z^n|\theta^n)}{p(z^n|z^{(n,m)})}. \tag{46}$$

Note that

$$\overline{L}_{n,1} := \frac{1}{n} E_{Z^n|\theta^n} \log \frac{p(z^n|\theta^n)}{p(z^n)}, \tag{47}$$

where $p(z^n)$ is the learner's initial distribution on Z^n before any data has arrived,

$$p(z^n) = \int_{\Theta^n} p(z^n|\theta^n)p(\theta^n)\, d\theta^n = \int_{\Pi} \int_{\Theta^n|\pi} p(z^n|\theta^n)p(\theta^n|\pi)\, d\theta^n\, p(\pi) d\pi \tag{48}$$

To understand better the meaning of $\overline{L}_{n,m}$, consider the loss associated with learning a single classification task. In this case $Z = X \times \{0, 1\}$. If we assume that only the conditional

distribution on class labels is affected by the model, then $p(z|\theta) = p(x)p(y|x,\theta)$, and for the predictive distribution, $p(z|z^m) = p(x)p(y|x,z^m)$. Let $\alpha(x) := p(y=1|x,\theta)$ and $\beta(x) := p(y=1|x,z^m)$. Substituting these expressions into (46) and simplifying yields

$$\overline{L}_{1,m} = E_X \left[\alpha(x) \log \frac{\alpha(x)}{\beta(x)} + (1-\alpha(x)) \log \frac{1-\alpha(x)}{1-\beta(x)} \right]. \tag{49}$$

The expression in square brackets is zero if $\alpha(x) = \beta(x)$, *i.e.* if the conditional distributions on class labels are the same for the true and predictive distributions. It increases slowly as $\alpha(x)$ and $\beta(x)$ diverge.

The quantity analysed in this section is again the *cumulative risk*:

$$\overline{C}_{n,m,\pi^*} := \sum_{k=0}^{m-1} E_{\Theta^n|\pi^*} E_{Z^{(n,k)}|\theta^n} \overline{L}_{n,k+1}, \tag{50}$$

i.e. the *expected* total loss incurred by the learner after m steps of the above process. Note that the expectation is over all sequences of n tasks θ^n and all (n,k)-samples drawn according to $p(z^n|\theta^n)$.

Definition 3 *For any $n = 1, 2, \ldots$, and for all $\theta^n, \tilde{\theta}^n \in \Theta^n$, define $\dim_{P_{\Theta^n}}(\theta^n)$, $\Delta_H(\theta^n, \tilde{\theta}^n)$ and $\Delta_K(\theta^n, \tilde{\theta}^n)$ by replacing all occurrences of Π by Θ^n and all occurrences of Θ by Z^n in definition 1.*

Theorem 4 *For this theorem fix $n \in N$ and take all limiting behaviour to be with respect to m. Suppose there exists $\alpha < \infty$ such that for all $\theta, \tilde{\theta} \in \Theta$,*

$$\Delta_K(\theta, \tilde{\theta}) \leq \alpha \Delta_H(\theta, \tilde{\theta}), \tag{51}$$

and that $\dim_{P_{\Theta^n}}(\theta^n)$ exists for almost all (P_{Θ^n}) $\theta^n \in \Theta^n$. Then,

$$\overline{C}_{n,m,\pi^*} \doteq_{(\Pi, P_\Pi)} \frac{\log m}{2n} E_{\Theta^n|\pi^*} \dim_{P_{\Theta^n}}(\theta^n) + o(\log m). \tag{52}$$

Proof: Direct calculation shows that

$$\overline{C}_{n,m,\pi^*} = \frac{1}{n} E_{\Theta^n|\pi^*} D_K(P_{Z^{(n,m)}|\theta^n} \| P_{Z^{(n,m)}}). \tag{53}$$

As $\Delta_K(\theta^n, \tilde{\theta}^n) = \sum_{i=1}^n \Delta_K(\theta_i, \tilde{\theta}_i)$, the condition $\Delta_K(\theta, \tilde{\theta}) \leq \alpha \Delta_H(\theta, \tilde{\theta})$ ensures that $\Delta_K(\theta^n, \tilde{\theta}^n) \leq n\alpha \Delta_H(\theta^n, \tilde{\theta}^n)$. So the conditions of theorem 1 are satisfied (with Θ^n replaced by $Z^{(n,m)}$, Π replaced by Θ^n, and n replaced by m). Hence,

$$\frac{D_K(P_{Z^{(n,m)}|\theta^n} \| P_{Z^{(n,m)}})}{\log m} \doteq_{(\Theta^n, P_{\Theta^n})} \frac{\dim_{P_{\Theta^n}}(\theta^n)}{2} + o(1). \tag{54}$$

More specifically, equation (54) means that for all $n = 1, 2, \ldots$, there exists $f(m)$ such that $f(m) \to 0$ and the sets

$$\Theta_m^n := \left\{ \theta^n \in \Theta^n : \frac{D_K(P_{Z^{(n,m)}|\theta^n} \| P_{Z^{(n,m)}})}{\log m} = \frac{\dim_{P_{\Theta^n}}(\theta^n)}{2} + f(m) \right\} \tag{55}$$

satisfy $P_{\Theta^n}(\Theta_m^n) \to 1$ as $m \to \infty$. As $P_{\Theta^n}(\Theta_m^n) = E_\Pi P_{\Theta^n|\pi}(\Theta_m^n)$, we must have

$$P_{\Theta^n|\pi}(\Theta_m^n) \doteq_{(\Pi, P_\Pi)} 1 + o(1) \tag{56}$$

for each n. Hence

$$E_{\Theta^n|\pi^*} \frac{D_K(P_{Z^{(n,m)}|\theta^n} \| P_{Z^{(n,m)}})}{\log m} \doteq_{(\Pi, P_\Pi)} E_{\Theta^n|\pi^*} \frac{\dim_{P_{\Theta^n}}(\theta^n)}{2} + o(1), \tag{57}$$

which completes the proof. ■

Theorem 4 gives an expression for the expected *cumulative* risk for a learner that is simultaneously learning n tasks using a hierarchical model. In contrast, if the learner does not take account of the fact that the n tasks are related, then each time it comes to learn a new task it will start with the same prior $p(\theta) = \int_\Pi p(\theta|\pi)p(\pi)\, d\pi$. In this case the learner's expected cumulative risk when learning n tasks is given by

$$\overline{C}_{n,m,\pi^*} \doteq \frac{\log m}{2} E_{\Theta|\pi^*} \dim_{P_\Theta}(\theta) + o(\log m) \tag{58}$$

(the proof of this is similar to the proof of theorem 4). Thus the difference between the learner's risk when taking task relatedness into account (52) vs. ignoring task relatedness (58) is to first order controlled by the difference between

$$\frac{1}{n} E_{\Theta^n|\pi^*} \dim_{P_{\Theta^n}}(\theta^n) \tag{59}$$

and

$$E_{\Theta|\pi^*} \dim_{P_\Theta}(\theta). \tag{60}$$

In the next section expressions (59) and (60) are calculated for a general class of hierarchical models that includes the LDR model.

4.2. Dimension of $(\mathbf{a}, \mathbf{b})$-models

Definition 4 *Let (X, ρ) be a metric space. We say a second metric ρ'* locally dominates *ρ at x if there exists $\varepsilon, c, c' > 0$ such that for all $y \in B_\varepsilon(x, \rho)$ (the ε-ball around x under ρ),*

$$c\rho'(x, y) \leq \rho(x, y) \leq c'\rho'(x, y). \tag{61}$$

Definition 5 *An (a, b)-model is a hierarchical model in which $\Pi = R^b$, $\Theta = R^a \times R^b$ and the following conditions hold:*

1. *The priors $p(\theta|\pi)$ are of the form*

$$p(\theta = (x^a, x^b)|\pi) = \delta(x^b - \pi)g_\pi(x_a) \tag{62}$$

   where $\delta(\cdot)$ is the b-dimensional Dirac delta function and g_π is a continuous function on R^a.

2. *The hyper-prior P_Π has a continuous density $p(\pi)$ and the true prior π^* has positive density $p(\pi^*)$.*

3. *The conditional distributions $p(z|\theta)$ are twice continuously differentiable functions of θ.*

4. *$\Delta_H^{1/2}$ is locally dominated by the Euclidean distance $\|\cdot\|$ on Θ, except possibly for a set of θ of $P_{\Theta|\pi^*}$-measure zero.*

5. *There exists an $\alpha < \infty$ such that for all $\theta, \tilde{\theta} \in \Theta$, $\Delta_K(\theta, \tilde{\theta}) \leq \alpha\Delta_H(\theta, \tilde{\theta})$.*

Conditions 1–3 of an (a, b)-model formalize the idea of a smooth hierarchical model in which there are $a + b$ parameters, b of which are effectively hyper-parameters and are fixed by the prior and the remaining a of which are model parameters. Conditions 4 and 5 are technical restrictions needed to make the proofs go through. In many cases the following results would still hold without these restrictions, but different proof techniques would be required. Recall that supp$P$ is the smallest closed set of P-probability 1.

Theorem 5 *In an (a, b)-model, for all θ^n in the interior of* suppP_{Θ^n} *(except for a set of $P_{\Theta^n|\pi^*}$-measure zero),*

$$\dim_{P_{\Theta^n}}(\theta^n) = na + b, \tag{63}$$

In addition, for any π, if θ^n is in the interior of supp$(P_{\Theta^n|\pi})$*, then*

$$\dim_{P_{\Theta^n|\pi}}(\theta^n) = na, \tag{64}$$

again except for a set of $P_{\Theta^n|\pi^*}$-measure zero.

Proof: See appendix B. ■

Note that the set of θ^n not covered by the first part of theorem 5 has P_{Θ^n} measure zero because P_{Θ^n} is absolutely continuous with respect to $P_{\Theta^n|\pi^*}$ and the P_{Θ^n} measure of the boundary of supp(P_{Θ^n}) is zero (because g_π is continuous). A similar conclusion applies to the $P_{\Theta^n|\pi^*}$ measure of the set of θ^n not covered by the second part of the theorem.

The requirement that θ^n be in the *interior* of supp(P_{Θ^n}) in theorem 5 is sometimes necessary. To see this, consider a distribution P on $[0, 1]$ that has an analytic density with one zero at $x = 1/2$. In this case supp$(P) = [0, 1]$ and the interior of supp$(P) = [0, 1] - \{1/2\}$. For any x in the interior of supp(P), $\dim_P(x) = 1$, but for $x = 1/2$, $\dim_P(x) = 3$.

Theorem 6 *In an (a, b)-model, the learner's cumulative risk (50) satisfies*

$$\overline{C}_{n,m,\pi^*} \doteq \frac{\log m}{2}\left(a + \frac{b}{n}\right) + o(\log m) \tag{65}$$

if the tasks are learnt hierarchically, and

$$\overline{C}_{n,m,\pi^*} \doteq \frac{\log m}{2}(a+b) + o(\log m) \tag{66}$$

if they are learnt independently. Furthermore, if the true prior is known then

$$\overline{C}_{n,m,\pi^*} \doteq \frac{\log m}{2}(a) + o(\log m)\,. \tag{67}$$

Proof: Equation (65) follows immediately from theorem 4 and the first part of theorem 5 (noting the comment after theorem 5), while equation (66) follows from equation (58) and the first part of theorem 5 with $n = 1$. Equation (67) follows by replacing $\dim_{P_\Theta^n}$ by $\dim_{P_\Theta^n|\pi^*}$ in theorem 4, and then applying the second part of theorem 5. ■

Theorem 6 shows that the hierarchical approach always does better asymptotically in an (a, b)-model (even for $n = 2$), and is most advantageous when the hyper-parameters dominate the parameters ($b \gg a$). Comparing (65) with (67), we see that the effect of lack of knowledge of the true prior can be made arbitrarily small by learning enough tasks simultaneously, the same conclusion that was reached in section 3.

The following theorem gives sufficient conditions for $\|\cdot\|$ to locally dominate $\Delta_H^{1/2}$ in an (a, b)-model.

Theorem 7 *If the map $P_{Z|\theta} \mapsto \theta$ is continuous (*i.e. $P_{Z|\theta} \to P_{Z|\theta_0} \Rightarrow \theta \to \theta_0$ *where convergence on the left is weak convergence) on some open set containing θ_0, and the Fisher information matrix*

$$J(\theta) = E_{Z|\theta}\left[\frac{\partial}{\partial\theta_i}\log p(z|\theta)\frac{\partial}{\partial\theta_j}\log p(z|\theta)\right]_{i,j=1,\ldots,a+b} \tag{68}$$

exists and is positive definite at θ_0, then $\|\cdot\|$ locally dominates $\Delta_H^{1/2}$ at θ_0.

Proof: See appendix C ■

4.3. Learning an LDR revisited

Consider the LDR model of section 2.4. Set $a = W_{\rm OUT}$ and $b = W_{\rm LDR}$, where $W_{\rm OUT}$ is the number of weights in an output node and $W_{\rm LDR}$ is the number of weights in the LDR (recall Figure 1). Assume the priors $p(\theta|\pi)$ are given by equation (62). Suppose that the weights are restricted to lie in some compact subset of R^{a+b} (so that the hyper-prior $p(\pi)$ has compact support and so do the functions g_π). To complete the model,

suppose that for each $\theta \in R^{a+b}$, $p(z|\theta)$ is of the form $p(y=1,x|\theta) = p(x)f_\theta(x)$ and $p(y=0,x|\theta) = p(x)(1-f_\theta(x))$, where $f_\theta(x)$ is the output of the network with weights θ and input x, and $p(x)$ is a continuous density on some compact subset of R^d. Assume the sigmoid is $\sigma(x) = \tanh(x)$, except at the output node where $\sigma(x) = (1+\tanh(x))/2$.

Theorem 8 *For the neural-network LDR model as above, the cumulative risk (50) satisfies*

$$\overline{C}_{n,m,\pi^*} \doteq \frac{\log m}{2}\left(W_{\rm OUT} + \frac{W_{\rm LDR}}{n}\right) + o\left(\log m\right), \tag{69}$$

Proof: The theorem would follow immediately from theorem 6 if the neural-network LDR model was an (a,b)-model. Indeed, conditions 1,2,3 and 5 of definition 5 all hold (condition 5 is the only nontrivial one—it holds because of the compactness assumptions and the boundedness of $\tanh(x)$). Unfortunately, condition 4 does not hold because there are various weight-vector transformations that leave the network invariant—such as hidden-node permutations and sign-flips of all incoming and outgoing weights at a node. This also causes the continuity assumption to fail in theorem 7. Let $[\theta]$ denote the set of all weight vectors that produce the same behaviour as θ. Fefferman (Feferman, 1994) showed that for all but a set of weights of Lebesgue measure zero, node permutations and sign-flips are the *only* transformations that leave a multi-layer $\tanh$ network invariant. Hence, for almost all (P_Π) priors π^*, and for almost all ($P_{\Theta|\pi^*}$) parameters θ, $[\theta]$ is finite.

Similar arguments to those used in the proof of lemma 12 and theorem 7 can be used to show that finiteness of $[\theta]$ and positive definiteness of $J(\theta)$ ensures that there exist δ, c, c' such that for all $0 < \varepsilon < \delta$,

$$\bigcup_{\theta' \in [\theta]} B_{c\varepsilon}\left(\theta', \|\cdot\|\right) \subseteq B_\varepsilon\left(\theta, \Delta_H^{1/2}\right) \subseteq \bigcup_{\theta' \in [\theta]} B_{c'\varepsilon}\left(\theta', \|\cdot\|\right). \tag{70}$$

A slightly modified version of the proof of theorem 5 can then be used to show that theorem 5 holds in this case as well, which coupled with theorem 4 proves (69). Hence, the only thing left to show is that for almost all (P_Π) priors π^*, and for almost all ($P_{\Theta|\pi^*}$) parameters θ, $J(\theta)$ is positive definite. Note that $J(\theta)$ is always nonnegative-definite (to see this observe that

$$[J(\theta)]_{ij} = \left.\frac{\partial^2}{\partial\theta_i\partial\theta_j} D_K(P_{\theta'}\|P_\theta)\right|_{\theta'=\theta} \tag{71}$$

and use the fact that $D_K(P\|Q) \geq 0$ with equality if and only if $P = Q$ a.s.) So suppose that $\det[J(\theta)] = 0$ on a set of θ of positive probability. But $\det[J(\theta)]$ is analytic, hence if it is zero on a set of positive probability it must be zero everywhere. But in that case there must exist a smooth re-parameterization $\phi = \phi(\theta)$ of smaller dimension than θ such that $P_\theta = P_{\phi(\theta)}$, which violates the finiteness of $[\theta]$ a.e. Hence $J(\theta)$ is positive definite almost everywhere. ■

If the true model has a small set of features then $W_{\rm OUT}$ is small ($W_{\rm OUT}$ is always just the number of features plus 1 for the threshold, *c.f.* Figure 1). If our uncertainty concerning the

correct set of features is large then the LDR net will have to be large and so $W_{\rm LDR}$ will be large. Equation (69) shows that under these circumstances multiple task learning is most advantageous.

4.4. Learning the prior on the mean of a Gaussian

To demonstrate the wider applicability of this Bayesian multi-task sampling model, in this section we consider an altogether simpler model: that of learning the prior on the mean of a Gaussian.

So let $Z = R^b$, $\Theta = R^b \times R$, $\Pi = R^b$ and

$$\begin{aligned} p(z|\theta = (\mu, \sigma)) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\|z-\mu\|^2}{2\sigma^2} \right\}, \\ p(\mu, \sigma|\pi) &= \delta(\mu - \pi) U(\sigma), \\ p(\pi) &= B(\pi), \end{aligned}$$

where $U(\sigma)$ is the uniform distribution on $[1, 2]$, and $B(\pi)$ is the uniform distribution on the unit ball in R^b. In this model the prior fixes the mean of the distribution on Z, and then each learning problem corresponds to a different value of the variance σ, which is uniformly distributed in $[1, 2]$. In this case the true prior π^* is the mean of the distributions in the environment.

Theorem 7 holds for all θ in this model (see (Clarke & Barron, 1990)), and so condition 5 of the definition of an (a, b)-model ($(1, b)$ in this case) holds. Conditions 1,2 and 3 hold trivially, and the use of compact support for the mean and variance ensures condition 4 holds. Hence, this is a $(1, b)$ model and so a direct application of theorem 6 yields

$$\overline{C}_{n,m,\pi^*} \doteq \frac{\log m}{2} \left(1 + \frac{b}{n} \right) + o\left(\log m\right), \tag{72}$$

if n tasks are learnt hierarchically.

5. Conclusion

The problem of learning appropriate domain-specific bias via multi-task sampling has been modeled from a Bayesian/Information-Theoretic viewpoint. The approach shows that in certain high-dimensional, essentially non-parametric modeling scenarios, most of the model parameters are more appropriately regarded as hyper-parameters. Performing hierarchical Bayesian inference within such a model, using multiple task sampling, is asymptotically much more efficient than a non-hierarchical approach.

There are many interesting avenues for further research. Much more experimental work needs to be done to verify that bias learning actually works in practice. An ideal place to start would be learning domains in which there are a large number of related tasks and for which traditional approaches based on hand-coded feature sets have already produced good results. Face recognition, speech recognition and fingerprint recognition all fit this

description. One way to test the theory would be to try to learn feature sets for these domains using the neural net architecture described in figure 1.

Caruana (Caruana, 1993) has observed that adding extra output nodes to a single-hidden layer net and training them to perform correctly on related tasks can improve performance on a reference problem. This scenario is not covered by the Bayesian model presented here, nor by the VC/PAC type models of (Baxter, 1995b, Baxter, 1996b), because these models assume that independent training sets are available for each output node. It would be interesting to derive theoretically the behaviour observed by Caruana.

Another open problem is to determine the conditions under which Jeffrey's prior is the optimal *hyper-prior* to use for the hierarchical models discussed here. This question has only recently been settled for ordinary Bayes models (Barron & Clarke, 1994). Another important question is to what extent the assumption of *realizability* (*i.e.* $\pi^* \in \Pi$) can be relaxed. Also, the results of (Haussler & Opper, 1995b) can be used to derive asymptotic bounds on the KL divergence even when the model is infinite dimensional. It would be interesting to apply those results to the hierarchical case.

Appendix A

Proof of theorem 1

Let $I(\Pi;\Theta^n)$ denote the *mutual information* between Π and Θ^n (*i.e.* $I(\Pi;\Theta^n) := E_{\Pi^*} D_K(P_{\Theta^n|\pi^*} \| P_{\Theta^n})$).

Theorem 9 ((Haussler & Opper, 1995a), theorem 1) *For all* $n \geq 1$,

$$-E_{\Pi^*} \log E_{\Pi} e^{-\frac{n}{4}\Delta_H(\pi^*,\pi)} \leq I(\Pi;\Theta^n) \leq -E_{\Pi^*} \log E_{\Pi} e^{-n\Delta_K(\pi,\pi^*)}. \tag{A.1}$$

Using the assumption of theorem 1 that $\Delta_K(\pi,\pi') \leq \alpha\Delta_H(\pi,\pi')$ we have:

$$-E_{\Pi^*} \log E_{\Pi} e^{-\frac{n}{4}\Delta_H(\pi^*,\pi)} \leq E_{\Pi^*} D_K(P_{\Theta^n|\pi^*} \| P_{\Theta^n}) \leq -E_{\Pi^*} \log E_{\Pi} e^{-n\alpha\Delta_H(\pi,\pi^*)} \tag{A.2}$$

For any pair of random variables W and V and any real-valued function $u(w,v)$, we have the following inequality due to Feynman:

$$-E_V \log E_W e^{u(w,v)} \leq -\log E_W e^{E_V u(w,v)}. \tag{A.3}$$

Using (A.3) we can effectively "lop off" the expectation over Π^* in the upper bound of (A.2) to give an upper bound on $D_K(P_{\Theta^n|\pi^*} \| P_{\Theta^n})$.

Lemma 10 *For all* $n \geq 1$ *and* $\pi^* \in \Pi$,

$$D_K(P_{\Theta^n|\pi^*} \| P_{\Theta^n}) \leq -\log E_{\Pi} e^{-n\alpha\Delta_H(\pi,\pi^*)} \tag{A.4}$$

Proof: The proof is via the same chain of inequalities used to prove the upper-bound in theorem 9.

$$\begin{aligned}
D_K(P_{\Theta^n|\pi^*}\|P_{\Theta^n}) &= E_{\Theta^n|\pi^*}\log\frac{p(\theta^n|\pi^*)}{E_\Pi p(\theta^n|\pi)} \\
&= -E_{\Theta^n|\pi^*}\log E_\Pi e^{\log\frac{p(\theta^n|\pi)}{p(\theta^n|\pi^*)}} \\
&\leq -\log E_\Pi e^{E_{\Theta^n|\pi^*}\log\frac{p(\theta^n|\pi)}{p(\theta^n|\pi^*)}} \\
&= -\log E_\Pi e^{-D_K(P_{\Theta^n|\pi^*}\|P_{\Theta^n|\pi})} \\
&= -\log E_\Pi e^{-n\Delta_K(\pi,\pi^*)} \\
&\leq -\log E_\Pi e^{-n\alpha\Delta_H(\pi,\pi^*)}.
\end{aligned}$$

The penultimate line follows because the KL divergence is additive over the product of independent distributions (see *e.g.* (Cover & Thomas, 1991)). ■

Lemma 11 *If* $\dim_{P_\Pi}(\pi^*)$ *exists then for any* $0<\alpha<\infty$*,*

$$\lim_{n\to\infty}\frac{-\log E_\Pi e^{-n\alpha\Delta_H(\pi,\pi^*)}}{\log n} = \frac{\dim_{P_\Pi}(\pi^*)}{2}. \tag{A.5}$$

Proof: The arguments used in the proof of lemma 11 are similar to those used in (Haussler & Opper, 1995a) for proving corresponding global metric entropy bounds. Setting $\varepsilon = \frac{1}{\sqrt{\alpha n}}$, we have

$$\frac{-\log E_\Pi e^{-n\alpha\Delta_H(\pi,\pi^*)}}{\log n} = \frac{-\log E_\Pi e^{-\left(\frac{1}{\varepsilon}\Delta_H^{1/2}(\pi,\pi^*)\right)^2}}{-2\log\varepsilon-\log\alpha}. \tag{A.6}$$

Set ε sufficiently small to ensure that $-2\log\varepsilon-\log\alpha>0$. Now,

$$\begin{aligned}
-\log E_\Pi e^{-\left(\frac{1}{\varepsilon}\Delta_H^{1/2}(\pi,\pi^*)\right)^2} &= -\log\left(\int_{B_\varepsilon(\pi^*)} p(\pi)e^{-\left(\frac{1}{\varepsilon}\Delta_H^{1/2}(\pi,\pi^*)\right)^2}d\pi \right. \\
&\qquad \left. + \int_{B_\varepsilon^c(\pi^*)} p(\pi)e^{-\left(\frac{1}{\varepsilon}\Delta_H^{1/2}(\pi,\pi^*)\right)^2}d\pi\right) \\
&\leq -\log\left(\int_{B_\varepsilon(\pi^*)} p(\pi)e^{-1}\,d\pi \right. \\
&\qquad \left. + \int_{B_\varepsilon^c(\pi^*)} p(\pi)e^{-\left(\frac{1}{\varepsilon}\Delta_H^{1/2}(\pi,\pi^*)\right)^2}d\pi\right) \\
&\leq -\log\left[\frac{1}{e}P_\Pi(B_\varepsilon(\pi^*))\right] \\
&= -\log P_\Pi\left(B_\varepsilon(\pi^*)\right)+1,
\end{aligned}$$

and so

$$\begin{aligned}\limsup_{\varepsilon\to 0}\frac{-\log E_\Pi e^{-\left(\frac{1}{\varepsilon}\Delta_H^{1/2}(\pi,\pi^*)\right)^2}}{-2\log\varepsilon-\log\alpha} &\le \limsup_{\varepsilon\to 0}\frac{-\log P_\Pi(B_\varepsilon(\pi^*))+1}{-2\log\varepsilon-\log\alpha}\\ &= \frac{\dim_{P_\Pi}(\pi^*)}{2}.\end{aligned}$$

To get a matching lower bound note that for all $r>0$,

$$\begin{aligned}-\log E_\Pi e^{-\left(\frac{1}{\varepsilon}\Delta_H^{1/2}(\pi,\pi^*)\right)^2} &= -\log\left(\int_{B_r(\pi^*)} p(\pi)e^{-\left(\frac{1}{\varepsilon}\Delta_H^{1/2}(\pi,\pi^*)\right)^2}d\pi\right.\\ &\quad\left.+\int_{B_r^c(\pi^*)} p(\pi)e^{-\left(\frac{1}{\varepsilon}\Delta_H^{1/2}(\pi,\pi^*)\right)^2}d\pi\right)\\ &\ge -\log\left[P_\Pi\left(B_r(\pi^*)\right)+e^{-\left(\frac{r}{\varepsilon}\right)^2}\right].\end{aligned}$$

Setting $r=\varepsilon^{1-\delta}$ for any $0<\delta<1$ gives

$$-\log E_\Pi e^{-\left(\frac{1}{\varepsilon}\Delta_H^{1/2}(\pi,\pi^*)\right)^2} \ge -\log\left(P_\Pi\left(B_{\varepsilon^{1-\delta}}(\pi^*)\right)+e^{-\frac{1}{\varepsilon^{2\delta}}}\right) \tag{A.7}$$

Now, if $\dim_{P_\Pi}(\pi^*)$ exists then we know that $P_\Pi\left(B_{\varepsilon^{1-\delta}}(\pi^*)\right)$ decreases no faster than some power of $\varepsilon^{1-\delta}$, which for small enough ε will dominate $e^{-\frac{1}{\varepsilon^{2\delta}}}$, because the latter expression decreases faster than any fixed polynomial in ε as $\varepsilon\to 0$. Thus

$$\lim_{\varepsilon\to 0}\frac{-\log\left(P_\Pi\left(B_{\varepsilon^{1-\delta}}(\pi^*)\right)+e^{-\frac{1}{\varepsilon^{2\delta}}}\right)}{-\log\varepsilon}=(1-\delta)\dim_{P_\Pi}(\pi^*), \tag{A.8}$$

and so

$$\liminf_{\varepsilon\to 0}\frac{-\log E_\Pi e^{-\left(\frac{1}{\varepsilon}\Delta_H^{1/2}(\pi,\pi^*)\right)^2}}{-2\log\varepsilon-\log\alpha}\ge\frac{1-\delta}{2}\dim_{P_\Pi}(\pi^*) \tag{A.9}$$

for all $0<\delta<1$. Letting $\delta\to 0$ finishes the proof of lemma 11. ■

Without loss of generality, we may assume from now on that $\dim_{P_\Pi}(\pi)$ exists for all $\pi\in\Pi$ (by assumption $\dim_{P_\Pi}(\pi)$ exists except for a set of P_Π measure zero, so we can just remove all those π whose dimension is undefined).

From lemmas 11 and 10,

$$\limsup_{n\to\infty}\frac{D_K(P_{\Theta^n|\pi}\|P_{\Theta^n})}{\log n}\le\frac{\dim_{P_\Pi}(\pi)}{2}. \tag{A.10}$$

Applying lemma 11 to equation (A.2) and invoking Fatou's lemma twice gives

$$\lim_{n\to\infty}\frac{E_\Pi D_K(P_{\Theta^n|\pi}\|P_{\Theta^n})}{\log n}=E_\Pi\frac{\dim_{P_\Pi}(\pi)}{2}. \tag{A.11}$$

Now let

$$\Pi_{\text{supbad}} := \left\{ \pi \in \Pi : \limsup_{n\to\infty} D_K(P_{\Theta^n|\pi} \| P_{\Theta^n}) < \frac{\dim_{P_\Pi}(\pi)}{2} \right\} \tag{A.12}$$

Suppose that $P_\Pi(\Pi_{\text{supbad}}) > 0$. Then,

$$\begin{aligned}
E_\Pi \frac{\dim_{P_\Pi}(\pi)}{2} &= \limsup_{n\to\infty} E_\Pi \frac{D_K(P_{\Theta^n|\pi} \| P_{\Theta^n})}{\log n} \qquad \text{(by (A.11))} \\
&\leq \limsup_{n\to\infty} E_{\Pi_{\text{supbad}}} \frac{D_K(P_{\Theta^n|\pi} \| P_{\Theta^n})}{\log n} + \limsup_{n\to\infty} E_{\Pi^c_{\text{supbad}}} \frac{D_K(P_{\Theta^n|\pi} \| P_{\Theta^n})}{\log n} \\
&\leq E_{\Pi_{\text{supbad}}} \limsup_{n\to\infty} \frac{D_K(P_{\Theta^n|\pi} \| P_{\Theta^n})}{\log n} + E_{\Pi^c_{\text{supbad}}} \limsup_{n\to\infty} \frac{D_K(P_{\Theta^n|\pi} \| P_{\Theta^n})}{\log n} \\
&< E_{\Pi_{\text{supbad}}} \frac{\dim_{P_\Pi}(\pi)}{2} + E_{\Pi^c_{\text{supbad}}} \frac{\dim_{P_\Pi}(\pi)}{2} \qquad \text{(by assumption and (A.10))} \\
&= E_\Pi \frac{\dim_{P_\Pi}(\pi)}{2},
\end{aligned}$$

a contradiction. Thus $P_\Pi(\Pi_{\text{supbad}}) = 0$. Hence, for almost all π,

$$\limsup_{n\to\infty} \frac{D_K(P_{\Theta^n|\pi} \| P_{\Theta^n})}{\log n} = \frac{\dim_{P_\Pi}(\pi)}{2}. \tag{A.13}$$

Now, for each $n = 1, 2, \ldots$ and $\varepsilon > 0$ let

$$\Pi_{n,\varepsilon} = \{\pi : \frac{D_K(P_{\Theta^n|\pi} \| P_{\Theta^n})}{\log n} < \frac{\dim_{P_\Pi}(\pi)}{2} - \varepsilon\}. \tag{A.14}$$

Suppose that $\limsup_{n\to\infty} P_\Pi(\Pi_{n,\varepsilon}) = \kappa > 0$. So there exists an infinite sequence of integers $n_1 < n_2 < \ldots$ such that $P_\Pi(\Pi_{n_i,\varepsilon}) \geq \kappa$. From (A.10) we know that for any δ, $0 < \delta < \varepsilon\kappa$, there exists $k > 0$ such that for all $i > k$,

$$\frac{D_K(P_{\Theta^{n_i}|\pi} \| P_{\Theta^{n_i}})}{\log n_i} < \frac{\dim_{P_\Pi}(\pi)}{2} + \varepsilon\kappa - \delta. \tag{A.15}$$

Hence, for all $i > k$,

$$\begin{aligned}
E_\Pi \frac{D_K(P_{\Theta^{n_i}|\pi} \| P_{\Theta^{n_i}})}{\log n_i} &= E_{\Pi_{n_i,\varepsilon}} \frac{D_K(P_{\Theta^{n_i}|\pi} \| P_{\Theta^{n_i}})}{\log n_i} + E_{\Pi^c_{n_i,\varepsilon}} \frac{D_K(P_{\Theta^{n_i}|\pi} \| P_{\Theta^{n_i}})}{\log n_i} \\
&< E_{\Pi_{n_i,\varepsilon}} \left(\frac{\dim_{P_\Pi}(\pi)}{2} - \varepsilon \right) + E_{\Pi^c_{n_i,\varepsilon}} \left(\frac{\dim_{P_\Pi}(\pi)}{2} + \varepsilon\kappa - \delta \right) \\
&< E_{\Pi_{n_i,\varepsilon}} \frac{\dim_{P_\Pi}(\pi)}{2} - \varepsilon\kappa + E_{\Pi^c_{n_i,\varepsilon}} \frac{\dim_{P_\Pi}(\pi)}{2} + \varepsilon\kappa - \delta \\
&= E_\Pi \frac{\dim_{P_\Pi}(\pi)}{2} - \delta.
\end{aligned}$$

and so

$$\begin{aligned} E_{\Pi}\frac{\dim_{P_{\Pi}}(\pi)}{2} &= \lim_{i\to\infty} E_{\Pi}\frac{D_K(P_{\Theta^{n_i}|\pi}\|P_{\Theta^{n_i}})}{\log n_i} \\ &\leq E_{\Pi}\frac{\dim_{P_{\Pi}}(\pi)}{2} - \delta, \end{aligned}$$

which is a contradiction and so the assumption $\limsup_{n\to\infty} P_{\Pi}(\Pi_{n,\varepsilon}) > 0$ must be false. Hence for all $\varepsilon > 0$, $\lim_{n\to\infty} P_{\Pi}(\Pi_{n,\varepsilon}) = 0$. Setting

$$\Pi'_{n,\varepsilon} := \left\{\pi\colon \frac{D_K(P_{\Theta^n|\pi}\|P_{\Theta^n})}{\log n} < \frac{\dim_{P_{\Pi}}(\pi)}{2} - \varepsilon \text{ or } \frac{D_K(P_{\Theta^n|\pi}\|P_{\Theta^n})}{\log n} > \frac{\dim_{P_{\Pi}}(\pi)}{2}\right\}, \tag{A.16}$$

we have proved so far that $\lim_{n\to\infty} P_{\Pi}(\Pi'_{n,\varepsilon}) = 0$ for all $\varepsilon > 0$. Now define $n_0(1) = 1$ and for all $m > 1$,

$$n_0(m) = \min_{n_0}\colon P_{\Pi}\left(\Pi'_{n,\frac{1}{m}}\right) \leq \frac{1}{m} \quad \forall n \geq n_0. \tag{A.17}$$

Note that $\Pi'_{n,\frac{1}{m+1}} \supseteq \Pi'_{n,\frac{1}{m}}$ so $n_0(m)$ is an increasing function of m. For all $n \geq 1$ define $m_0(n) = \max_m\colon n_0(m) \leq n$ (with $m_0(n) = \infty$ if there is no maximium). Note that $m_0(n) \to \infty$ and so $\frac{1}{m_0(n)} \in o(1)$. Let

$$\Pi'_n = \left\{\pi\colon \frac{D_K(P_{\Theta^n|\pi}\|P_{\Theta^n})}{\log n} < \frac{\dim_{P_{\Pi}}(\pi)}{2} - \frac{1}{m_0(n)} \text{ or } \frac{D_K(P_{\Theta^n|\pi}\|P_{\Theta^n})}{\log n} > \frac{\dim_{P_{\Pi}}(\pi)}{2}\right\}. \tag{A.18}$$

By definition $P_{\Pi}(\Pi'_n) \leq \frac{1}{m_0(n)}$, hence $P_{\Pi}(\Pi'_n) \to 0$. Thus

$$\frac{D_K(P_{\Theta^n|\pi}\|P_{\Theta^n})}{\log n} \doteq \frac{\dim_{P_{\Pi}}(\pi)}{2} + o(1). \tag{A.19}$$

■

Appendix B

Proof of theorem 5

Lemma 12 *Set $\theta^n = (\theta_1, \ldots, \theta_n)$. If $\Delta_H^{1/2}$ is locally dominated by $\|\cdot\|$ at each θ_i then there exists $c, c', \delta > 0$, such that for all $0 < \varepsilon < \delta$,*

$$B_{c\varepsilon}(\theta^n, \|\cdot\|) \subseteq B_{\varepsilon}\left(\theta^n, \Delta_H^{1/2}\right) \subseteq B_{c'\varepsilon}(\theta^n, \|\cdot\|) \tag{B.1}$$

Proof: Let

$$\begin{aligned} D(\theta,\tilde{\theta}) &:= \int_Z \left[p(z|\theta)p(z|\tilde{\theta})\right]^{1/2} dz \\ D(\theta^n,\tilde{\theta}^n) &:= \int_Z \left[p(z^n|\theta^n)p(z^n|\tilde{\theta}^n)\right]^{1/2} dz^n = \prod_{i=1}^n D(\theta_i,\tilde{\theta}_i). \end{aligned}$$

Note that $\Delta_H(\theta,\tilde{\theta}) = 2\left(1 - D(\theta,\tilde{\theta})\right)$ and $\Delta_H(\theta^n,\tilde{\theta}^n) = 2\left(1 - \prod_{i=1}^n D(\theta_i,\tilde{\theta}_i)\right)$. Now suppose that for all i, $\Delta_H(\theta_i,\tilde{\theta}_i) \leq \varepsilon/n$. Hence $D(\theta_i,\tilde{\theta}_i) \geq 1-\varepsilon/2n \Rightarrow \prod_{i=1}^n D(\theta_i,\tilde{\theta}_i) \geq (1-\varepsilon/2n)^n \geq 1-\varepsilon/2 \Rightarrow \Delta_H(\theta^n,\tilde{\theta}^n) \leq \varepsilon$. Next suppose that $\Delta_H(\theta^n,\tilde{\theta}^n) \leq \varepsilon$. Hence $\prod_{i=1}^n D(\theta_i,\tilde{\theta}_i) \geq 1-\varepsilon/2 \Rightarrow D(\theta_i,\tilde{\theta}_i) \geq 1-\varepsilon/2$ for each i, because $D(\theta,\tilde{\theta}) \leq 1$ always. Thus $\Delta_H(\theta_i,\tilde{\theta}_i) \leq \varepsilon$ for all i. These two results show that

$$\begin{aligned} B_{\varepsilon/\sqrt{n}}\left(\theta_1,\Delta_H^{1/2}\right) \times \ldots \times B_{\varepsilon/\sqrt{n}}\left(\theta_n,\Delta_H^{1/2}\right) &\subseteq B_\varepsilon\left(\theta^n,\Delta_H^{1/2}\right) \\ &\subseteq B_\varepsilon\left(\theta_1,\Delta_H^{1/2}\right) \times \ldots \times B_\varepsilon\left(\theta_n,\Delta_H^{1/2}\right). \end{aligned} \tag{B.2}$$

Hence, by the local domination of $\Delta_H^{1/2}$ by $\|\cdot\|$ at each θ_i, there exists c, c' such that for sufficiently small ε,

$$\begin{aligned} B_{c\varepsilon/\sqrt{n}}(\theta_1,\|\cdot\|) \times \ldots \times B_{c\varepsilon/\sqrt{n}}(\theta_n,\|\cdot\|) &\subseteq B_\varepsilon(\theta^n,\|\cdot\|) \\ &\subseteq B_{c'\varepsilon}(\theta_1,\|\cdot\|) \times \ldots \times B_{c'\varepsilon}(\theta_n,\|\cdot\|), \end{aligned} \tag{B.3}$$

which implies that there exists c, c' such that

$$B_{c\varepsilon}(\theta^n,\|\cdot\|) \subseteq B_\varepsilon\left(\theta^n,\Delta_H^{1/2}\right) \subseteq B_{c'\varepsilon}(\theta^n,\|\cdot\|). \tag{B.4}$$

■

Now fix $\hat{\theta}^n = (\hat{\theta}_{a1},\hat{\theta}_{b1},\ldots,\hat{\theta}_{an},\hat{\theta}_{bn})$. By property 4 of an (a,b)-model (definition 5), with $P_{\Theta^n|\pi^*}$ probabilty 1, $\|\cdot\|$ locally dominates $\Delta_H^{1/2}$ at each $\theta_i = (\theta_{ai},\theta_{bi})$. Again by the definition of an (a,b)-model,

$$\begin{aligned} P_{\Theta^n}\left(B_\varepsilon\left(\hat{\theta}^n,\Delta_H^{1/2}\right)\right) &= \int\limits_{B_\varepsilon\left(\hat{\theta}^n,\Delta_H^{1/2}\right)} \int_\Pi p(\theta^n|\pi)p(\pi)\,d\pi\,d\theta^n \\ &= \int_{\pi\in R^b} p(\pi) \int\limits_{B_\varepsilon\left(\hat{\theta}^n,\Delta_H^{1/2}\right)} \delta(\theta_{b1}-\pi)\ldots\delta(\theta_{bn}-\pi) \\ &\qquad g_\pi(\theta_{a1})\ldots g_\pi(\theta_{an})\,d\theta_{b1}\ldots d\theta_{bn}d\theta_{a1}\ldots d\theta_{an}d\pi \\ &\leq \int_\Pi p(\pi) \int\limits_{B^\pi_{c'\varepsilon\sqrt{n}}\left(\hat{\theta}^n,\|\cdot\|\right)} g_\pi(\theta_{a1})\ldots g_\pi(\theta_{an})\,d\theta_{a1}\ldots d\theta_{an}d\pi \end{aligned} \tag{B.5}$$

where $B^{\pi}_{c'\varepsilon\sqrt{n}}\left(\hat{\theta}^n, \|\cdot\|\right) := \{\theta^n_\pi = (\theta_{a1}, \pi, \ldots, \theta_{an}, \pi) \colon \|\theta^n_\pi - \hat{\theta}^n\| \leq c'\varepsilon\sqrt{n}\}$, and we have invoked lemma 12. The condition that $\hat{\theta}^n$ be in the interior of $\text{supp}(P_{\Theta^n})$ in the statement of theorem 5 means $\int_\Pi p(\hat{\theta}^n|\pi)p(\pi)\,d\pi > 0$, or

$$\int_\Pi p(\pi) \prod_{i=1}^{n} \delta(\hat{\theta}_{bi} - \pi) g_\pi(\hat{\theta}_{ai})\,d\pi > 0. \tag{B.6}$$

This can only hold if there is some $\hat{\pi}$ such that $p(\hat{\pi}) > 0$ and $\hat{\theta}_{bi} = \hat{\pi}$ and $g_{\hat{\pi}}(\hat{\theta}_{ai}) > 0$ for all $i = 1 \ldots n$. Hence (B.5) is an integral over an $na + b$ dimensional ball of a function $p(\pi)g_\pi(\theta_{a1}) \ldots g_\pi(\theta_{an})$ that is positive at the center $(\hat{\theta}_{a1}, \hat{\pi}, \ldots, \hat{\theta}_{an}, \hat{\pi})$. By assumption, $p(\cdot)$ and $g_\pi(\cdot)$ are continuous and so for small enough ε, (B.5) will be bounded above by $K\varepsilon^{na+b}$ for some $K > 0$. A similar argument, using the left-hand inequality in lemma 12, shows that $P_{\Theta^n}\left(B_\varepsilon\left(\hat{\theta}^n, \Delta_H^{1/2}\right)\right) \geq K'\varepsilon^{na+b}$ which shows that

$$\dim_{P_{\Theta^n}}(\theta^n) = na + b, \tag{B.7}$$

as required for the first part of theorem 5. The second part of theorem 5 follows from a similar argument. ■

Appendix C

Proof of theorem 7

$$\begin{aligned}
\Delta_H(\theta, \tilde{\theta}) &= \int_Z \left[p(z|\theta)^{\frac{1}{2}} - p(z|\tilde{\theta})^{\frac{1}{2}}\right]^2 dz \\
&= 2\left(1 - \int_Z \left[p(z|\theta)p(z|\tilde{\theta})\right]^{\frac{1}{2}} dz\right)
\end{aligned} \tag{C.1}$$

By assumption $p(z|\theta)$ is twice differentiable and so

$$\begin{aligned}
\int_Z \left[p(z|\theta)p(z|\tilde{\theta})\right]^{\frac{1}{2}} dz &= \int_Z p(z|\theta)^{\frac{1}{2}} \Bigg[p(z|\theta)^{\frac{1}{2}} + \frac{1}{2}p(z|\theta)^{-\frac{1}{2}}(\theta_i - \tilde{\theta}_i)\frac{\partial}{\partial\theta_i}p(z|\theta) \\
&\quad -\frac{1}{4}p(z|\theta)^{-\frac{3}{2}}(\theta_i - \tilde{\theta}_i)\frac{\partial}{\partial\theta_i}p(z|\theta)\frac{\partial}{\partial\theta_j}p(z|\theta)(\theta_j - \tilde{\theta}_j) \\
&\quad +\frac{1}{2}p(z|\theta)^{-\frac{1}{2}}(\theta_i - \tilde{\theta}_i)\frac{\partial^2}{\partial\theta_i\partial\theta_j}p(z|\theta)(\theta_j - \tilde{\theta}_j)\Bigg] dz \\
&\quad +O(\|\theta - \tilde{\theta}\|^3) \\
&= 1 - \frac{1}{4}\langle\theta - \tilde{\theta}|J(\theta)|\theta - \tilde{\theta}\rangle + O(\|\theta - \tilde{\theta}\|^3),
\end{aligned} \tag{C.2}$$

where

$$\begin{aligned}
[J(\theta)]_{ij} :&= \int_Z p(z|\theta)^{-1}\frac{\partial}{\partial\theta_i}p(z|\theta)\frac{\partial}{\partial\theta_j}p(z|\theta)\,dz \\
&= E_{Z|\theta}\left[\frac{\partial}{\partial\theta_i}\log p(z|\theta)\frac{\partial}{\partial\theta_j}\log p(z|\theta)\right]
\end{aligned}$$

which is the *Fisher information matrix* at θ. In the above derivation the Einstein summation convention of summing over repeated indices has been used. Substituting (C.2) in (C.1) gives

$$\Delta_H(\theta,\tilde{\theta}) = \frac{1}{2}\langle\theta-\tilde{\theta}|J(\theta)|\theta-\tilde{\theta}\rangle + O(\|\theta-\tilde{\theta}\|^3). \tag{C.3}$$

Let $\lambda_{\min}(\theta)$ and $\lambda_{\max}(\theta)$ denote the minimum and maximum eigenvalues of $J(\theta)$. By assumption $J(\theta)$ is positive definite, so $\lambda_{\min}(\theta) > 0$. Working in the basis in which $J(\theta)$ is diagonal gives

$$\lambda_{\min}(\theta)\|\tilde{\theta}-\tilde{\theta}\|^2 \leq \langle\theta-\tilde{\theta}|J(\theta)|\theta-\tilde{\theta}\rangle \leq \lambda_{\max}(\theta)\|\tilde{\theta}-\tilde{\theta}\|^2, \tag{C.4}$$

which coupled with (C.3) yields

$$\begin{aligned}\left[\tfrac{\lambda_{\min}(\theta)}{2}\right]^{\frac{1}{2}}\|\theta-\tilde{\theta}\| + O(\|\theta-\tilde{\theta}\|^{\frac{3}{2}} &\leq \Delta_H^{\frac{1}{2}}(\theta,\tilde{\theta}) \\ &\leq \left[\tfrac{\lambda_{\max}(\theta)}{2}\right]^{\frac{1}{2}}\|\theta-\tilde{\theta}\| + O(\|\theta-\tilde{\theta}\|^{\frac{3}{2}}).\end{aligned} \tag{C.5}$$

By assumption, the map $P_{Z|\theta} \mapsto \theta$ is continuous in the topology of weak convergence, which implies it is continuous in the topology generated by the Hellinger distance, and hence for any $\varepsilon > 0$ there will exist a $\delta > 0$ such that if $\Delta_H^{1/2}(\theta,\tilde{\theta}) < \delta$, then $\|\theta-\tilde{\theta}\| < \varepsilon$. Combined with (C.5), this proves that $\Delta_H^{1/2}$ is locally dominated by $\|\cdot\|$. ■

Acknowledgments

This work was supported by EPSRC grants #K70366 and #K70373. Thanks to Martin Anthony and John Shawe-Taylor for useful discussions and to John Howard for pointing out the connection with hierarchical Bayes techniques. Thanks also to two anonymous referees for their helpful remarks.

Notes

1. In reality the prior cannot be directly sampled to get $\theta_1, \theta_2, \ldots$, only the conditional distributions $p(z|\theta_1), p(z|\theta_2), \ldots$ can be sampled. This is discussed further in section 4, however for the moment the fiction that we have direct access to the parameters will be maintained.
2. We will put the delta function back in the next section where we consider the more realistic scenario in which the learner receives information about θ in the form of examples z chosen according to $p(z|\theta)$, rather than receiving θ directly.

References

Abu-Mostafa, Y.S. (1989). Learning from Hints in Neural Networks. *Journal of Complexity*, 6:192–198.

Anthony, Martin & Bartlett, Peter. (1995). Function learning from interpolation. In *Proceedings of the Second European Conference on Computational Learning Theory*, Barcelona. Springer-Verlag.

Barron, Andrew & Clarke, Bertrand. (1994). Jeffreys' Prior is Asymptotically Least Favourable under Entropy Risk. *Journal of Statistical Planning and Inference*, 41:37–60.

Bartlett, Peter, Long, Philip & Williamson, Bob. (1994). Fat-Shattering and the Learnability of Real-Valued Functions. In *Proccedings of the Seventh ACM Conference on Computational Learning Theory*, New York. ACM Press.

Baxter, Jonathan. (1995a). A Model of Bias Learning. Technical Report LSE-MPS-97, London School of Economics, Centre for Discrete and Applicable Mathematics. Submitted for publication.

Baxter, Jonathan. (1995b). Learning Internal Representations. In *Proceedings of the Eighth International Conference on Computational Learning Theory*, pages 311–320, Santa Cruz, California. ACM Press.

Baxter, Jonathan. (1996a). A Bayesian/Information Theoretic Model of Bias Learning. In *Proccedings of the Ninth ACM Conference on Computational Learning Theory*, New York. ACM Press.

Baxter, Jonathan. (1996b). Learning Model Bias. In *Advances in Neural Information Processing Systems 8*, pages 169–175.

Berger, James O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Berger, James O. (1986) Multivariate Estimation: Bayes, Empirical Bayes, and Stein Approaches. *SIAM*.

Bridle, J.S. (1989). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F Fogelman-Soulie and J Herault, editors, *Neurocomputing: Algorithms, Architectures*. Springer Verlag, New York.

Caruana, Richard. (1993). Learning Many Related Tasks at the Same Time with Backpropagation. In *Advances in Neural Information Processing 5*.

Clarke, Bertrand & Barron, Andrew. (1990). Information-Theoretic Asymptotics of Bayes Methods. *IEEE Transactions on Information Theory*, 36:453–471.

Cover, T.M. & Thomas, J.A. (1991). *Elements of Information Theory*. John Wiley & Sons, Inc., New York.

Fefferman, Charles. (1994). Reconstructing a neural network from its output. *Rev. Mat. Iberoamericana*, 10:507–555.

Good, I.J. (1980). Some History of the Hierarchical Bayesian Methodology. In J M Bernado, M H De Groot, D V Lindley, and A F M Smith, editors, *Bayesian Statistics II*. University Press, Valencia.

Haussler, David & Opper, Manfred. (1995a). General Bounds on the Mutual Information Between a Parameter and n Conditionally Independent Observations. In *Proccedings of the Eighth ACM Conference on Computational Learning Theory*, New York. ACM Press.

Haussler, David & Opper, Manfred. (1995b). Mutual Information, Metric Entropy and Risk in Estimation of Probability Distributions. Submitted to Annals of Statistics.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257.

Mackay, David. (1991). Bayesian Interpolation. *Neural Computation*, 4:415–447.

Mackay, David. (1991). The Evidence Framework Applied to Classification Networks. *Neural Computation*, 4:698–714.

Mitchell, Tom M. (1990). The need for biases in learning generalisations. In Tom G Dietterich and Jude Shavlik, editors, *Readings in Machine Learning*. Morgan Kaufmann.

Mitchell, Tom M. & Thrun, Sebastian. (1994). Learning One More Thing. Technical Report CMU-CS-94-184, CMU.

Pratt, Lori Y. (1992). Discriminability-based transfer between neural networks. In Stephen J Hanson, Jack D Cowan, and C Lee Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 204–211, San Mateo. Morgan Kaufmann.