Distributed Cooperative Bayesian Learning Strategies

Kenji Yamanishi
Theory NEC Laboratory, Real World Computing Partnership.
c/o C&C Research Laboratories, NEC Corporation,
1-1, 4-chome, Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 216, Japan.
yamanisi@sbl.cl.nec.co.jp

Abstract

This paper addresses the issue of designing an effective distributed learning system, in which a number of agents estimate the parameter of the target distribution in parallel, and the population learner (for short, the p-learner) combines their outputs to obtain a significantly better estimate. Such a system is important in speeding up learning. We propose as a type of a distributed learning system the *distributed cooperative Bayesian learning strategy* (DCB), in which each agent and the p-learner employ a probabilistic version of the Gibbs algorithm. We analyze DCB by giving upper bounds on its average logarithmic loss for predicting unseen data as functions of the number of examples that each agent observes and the size of agents. We thereby demonstrate the effectiveness of DCB in the sense that for some probability models, it performs approximately as well as the non-distributed optimal Bayesian strategy for polynomial agent size and sample size, achieving a significant speed-up of learning. We also consider the case where the hyperthesis class of probability distributions is hierarchically parameterized, and there is a feedback of information from the p-learner to agents. In this case we propose another type of DCB based on the Markov chain Monte Carlo method, and characterize its average prediction loss in terms of the number of feedback iterations as well as the agent size and sample size. Thereby we demonstrate that for the hierarchical Gaussian family DCB can also work approximately as well as the non-distributed optimal Bayesian strategy, while achieving a significant speed-up of learning.

COLT 97 Nashville, Tennesee, USA
Copyright 1997 ACM 0-89791-891-6/97/7..$3.50

1 INTRODUCTION

1.1 PROBLEM STATEMENT

We consider the situation where each example is generated according to an unknown parametric distribution, which we call the target distribution. We are concerned with the problem of learning the target distribution, equivalently, estimating the parameters specifying the target distribution, using a *distributed learning system*. A distributed learning system consists of a number of agent learners and the population learner (for short, the p-learner). Each agent learner independently observes a sequence of examples and outputs an estimate of the parameter specifying the target distribution. The p-learner doesn't have direct access to random examples, but only to a set of outputs of the agent learners. The p-learner combines the outputs of the agent learners in order to obtain a significantly better estimate of the parameter for the target distribution.

The purpose of designing distributed learning systems is twofold: 1) to speed up learning utilizing the parallelism (actually if a distributed learning system consists of s agent learners and can estimate the target distribution as well as a non-distributed learning system for the same sample size, then the distributed learning system runs s-times as fast as the non-distributed learning system), and 2) to meet the situation where there is sufficient communication bandwidth available for the agent learners to send their outputs to the p-learner, but not enough time or bandwidth for the p-learner to receive all the examples themselves.

We measure the performance of a distributed learning system in terms of the *average logarithmic loss* for predicting the probability of unseen data where the average is taken with respect to the target distribution belonging to a given parametric class. We wish to design a distributed learning system such that the average logarithmic loss is as small as possible, and eventually can predict future data approximately as well as the non-distributed optimal Bayesian learning strategy, which can observe all examples at once and attains the least average logarithmic loss.

This paper proposes two models of distributed learning; the *plain model* and the *hierarchical model*. The plain model deals with the case where each agent learner observes a data sequence generated according to the identical target distribution, and the value of the parameter specifying the target distribution is randomly generated according to a fixed prior distribution. The hierarchical model deals with the case where the target distribution according to which each data sequence is generated may not be identical over all agent learners, and the parameter values specifying the target distributions are randomly generated according to an identical prior distribution conditioned by a *hyper-parameter*, which is itself distributed according to a certain prior distribution. That is, in the hierarchical model the parameter for the target distribution and the hyper-parameter form a hierarchical structure in a probabilistic manner. In each of the two models we propose a specific type of distributed learning systems, and analyze its performance.

1.2 PREVIOUS WORK

The framework of distributed learning that we consider here is very much inspired by Kearns and Seung's seminal model of *population learning* ([7], see also the work by Nakamura et.al.[8]). Our model is similar with theirs in that each agent learner independently observes a data sequence and the p-learner has access only to the outputs of agent learners. The differences between our model and Kearns and Seung's one are as follows:

Kearns and Seung's model may be characterized by the following features:
1) The target to be learned is a deterministic rule.
2) The prediction loss is measured in terms of the 0-1 loss, which we may also call the discrete loss.
3) Each agent learner uses a *deterministic version* of the Gibbs algorithm, i.e., a deterministic hyperthesis is randomly chosen according to the uniform distribution from the set of hypertheses consistent with examples.
4) The p-learner uses the maximum likelihood strategy.
5) The performance of a distributed learning system is evaluated within the PAC (Probably Approximately Correct) learning model.
6) There is no feedback loop between the p-learner and agent learners.

In contrast, our model may be characterized by the following features:
1)' The target to be learned is a probability distribution. It is assumed that prior distributions of the target distribution exist, and form a hierarchical structure.
2)' The prediction loss is measured in terms of the average logarithmic loss.
3)' Each agent learner uses a *probabilistic version* of the Gibbs algorithm, i.e., a parameter value of a probabilistic model is randomly chosen according to the Bayes posterior distribution.
4)' The p-learner uses a simple averaging strategy or the Gibbs algorithm.
5)' The performance of a distributed learning system is evaluated in terms of the additional loss, defined as the difference between its average logarithmic loss and the average Bayes risk for the non-distributed Bayesian learning strategy.
6)' There is a feedback of information from the p-learner to agent learners (in the hierarchical model).

In summary, our model may be considered as a *probabilistic version of Kearns and Seung's model*, and also includes an extension of their model to the case where there are a hierarchical parameter structure and a feedback loop between each agent learner and the p-learner. In our model as above, we focus on the design and analysis of a specific distributed learning system, which we call, the *distributed cooperative Bayesian learning strategy* (DCB), in which agent learners and the p-learner hierarchically employ the *probabilistic version* of the Gibbs algorithm.

This work is also technically related to *hierarchical Bayesian inference* ([1]) and *Markov chain Monte Carlo (MCMC) method* ([5],[4],[12],[3],[10]). We apply MCMC to the iterative learning process induced by the feedback of information from the p-learner to agent learners in the hierarchical model. MCMC has mainly been applied to efficient approximations of analytically intractable Bayesian inference ([12],[3],[10],[15]). This work may suggest a possibility of a new application of MCMC to the design of a distributed learning system.

1.3 OVERVIEW OF RESULTS

In Section 2, we first give an exact mathematical form of the plain model. We also introduce the average logarithmic loss for a distributed learning system as its performance measure, and show that it is lower-bounded by the average Bayes risk for the non-distributed Bayesian learning strategy. Hence it turns out that the *additional loss*, which is defined as the difference between the average logarithmic loss for a distributed learning system and the average Bayes risk for the non-distributed Bayesian learning strategy, is the key quantity to be analyzed.

In the plain model we define DCB as the strategy in which each agent learner employs the Gibbs algorithm and the p-learner employs the simple averaging strategy. We investigate how well DCB works in the cases where the hypothesis class that each agent uses is a class of Gaussian distributions with a constant variance and a class of Poisson distributions. Theorem 6 and 8 give upper bounds on the additional loss for DCB as functions of 1) the number m of examples that each agent learner observes, and 2) the size s of agent learners. Thereby we demonstrate that DCB can work approximately as well as the non-distributed Bayesian learning strategy, while achieving a significant speed-up of learning.

In Section 3, we first define the hierarchical model of distributed learning systems and introduce DCB of this model as the strategy in which agent learners and the p-learner employ the Gibbs algorithm iteratively through the feedback of information from the p-learner to agent learners. Theorem 14 gives a general upper bound on the additional loss for DCB for the case where the parameter space is bounded. The bound is obtained as a function of 1) the number m of examples that each agent learner observes, 2) the size s of agent learners, and 3) the number N of feedback iterations between the p-learner to each agent learner.

Further Theorem 17 shows that for the case where the hyperthesis class is a class of Gaussian distribution with a hierarchical parameter structure, the expected variation distance between DCB and the non-distributed Bayesian learning strategy is upper-bounded by $O\left(\exp\left(-\frac{c_1 N \ln m}{\ln(sN)}\right)\right)+O\left(\frac{N \ln m}{\ln(sN)}\exp\left(-c_2 N\right)\right)$, where $0 < c_1, c_2 < \infty$ are constants. Thus for any $\varepsilon > 0$, if we set $N = O(\ln(1/\varepsilon))$, $m = O(1/\varepsilon)$, and $s = O(1/\varepsilon)$, then for arbitrarily small $\delta > 0$, the expected variation distance can be made at most $\varepsilon^{1-\delta}$ while the DCB runs $O((1/\varepsilon)/\ln(1/\varepsilon))$-times as fast as the non-distributed Bayesian learning strategy.

In Section 4, we discuss on extensions of DCB into the two cases; one is the multi-hierarchically parameterized case, and the other is the general decision-theoretic case where a loss function other than the logarithmic loss may be used as a distortion measure, and a general real-valued function may be used as a hyperthesis. Specifically, in the latter case, we relate a generalized version of DCB to Vovk's aggregating strategies ([13]) and the extended stochastic complexity ([14]).

2 PLAIN MODEL

2.1 MODEL

Let s be a positive integer. A *distributed learning system* $\mathcal{S}$ consists of s *agent learners* and a single *population learner* (for short, a *p-learner*). We call the number s the *size of agents.*

Let $\mathcal{D}$ be a measurable space. Let $\mathcal{C} = \{p(D|\theta) : \theta \in \Theta \subset \mathbf{R}^k\}$ be a given class of probability density functions over $\mathcal{D}$ specified by a k-dimensional real-valued parameter θ, belonging to a parameter space Θ. We call $\mathcal{C}$ the *hypothesis class.* Now we make the following assumption on the data generation:

Assumption 1
1) Each agent learner independently observes a sequence of examples, each of which is independently generated according to the identical target distribution *$p(D|\theta)$, belonging to $\mathcal{C}$.*
2) The value of the parameter θ specifying the target distribution is unknown and is generated according to a known prior distribution over Θ with a density function $\pi(\theta)$.

We call the model in which Assumption 1 holds the *plain model.*

In a *distributed learning system $\mathcal{S}$* in the plain model agent learners and the p-learner perform as follows: Let a positive integer m be given. For each i ($i = 1, \cdots, s$), the ith agent learner takes as input a sequence $D_i^m = D_{i1} \cdots D_{im}$ of m examples, then outputs an estimate of θ. Letting $\hat{\theta}_i$ be an output of the ith agent learner, the p-learner takes $\hat{\theta}_1, \cdots, \hat{\theta}_s$ as input and outputs an estimate $\hat{\theta}$ as a function of $\hat{\theta}_1, \cdots, \hat{\theta}_s$. The output of the p-learner is an output of the distributed learning system in the plain model. Note that the p-learner doesn't have direct access to random examples, but only to a set of outputs of agent learners. (See Figure 1.)

Although we assume, for the sake of analytical simplicity, that the sample size m is uniform over all agent learners, the model can be immediately extended into a general case where the sample size is not uniform.

Let $D^{sm} = D_1^m \cdots D_s^m$. For a distributed learning system $\mathcal{S}$, let $q(\theta|D^{sm})$ be the probability density function according to which the output $\hat{\theta}$ of the p-learner is generated. Then we can think of a mixture distribution $\int p(D|\theta)q(\theta|D^{sm})d\theta$ as an average probability of D (for given D^{sm}) determined by $\mathcal{S}$. Hence we measure the performance of $\mathcal{S}$ in terms of its *average logarithmic loss* $L_{s,m}(\mathcal{S})$ for predicting the probability of unseen data, defined by

$$L_{s,m}(\mathcal{S}) \overset{\text{def}}{=} E_\theta E_{D^{sm}|\theta} E_{D|\theta}\left[-\ln \int p(D|\theta)q(\theta|D^{sm})d\theta\right],$$

where $E_\theta, E_{D^{sm}|\theta}$, and $E_{D|\theta}$ denote the expectations taken with respect to $\pi(\theta), p(D^{sm}|\theta)$, and $p(D|\theta)$, respectively.

Let the *Bayes posterior density function* over Θ from D^{sm} be $p^*(\theta|D^{sm})$, i.e.,

$$p^*(\theta|D^{sm}) = \frac{\pi(\theta)\prod_{i=1}^s \prod_{j=1}^m p(D_{ij}|\theta)}{\int \pi(\theta)\prod_{i=1}^s \prod_{j=1}^m p(D_{ij}|\theta)d\theta}.$$

We define the *average Bayes risk* $L^*_{s,m}$ for sample size sm by

$$L^*_{s,m} \overset{\text{def}}{=} E_\theta E_{D^{sm}|\theta} E_{D|\theta}\left[-\ln \int p(D|\theta)p^*(\theta|D^{sm})d\theta\right],$$

which is obtained by plugging $q(\theta|D^{sm}) = p^*(\theta|D^{ms})$ to the formula of $L_{s,m}(\mathcal{S})$. The *non-distributed Bayesian learning strategy* (in the plain model) is referred here to as the strategy which takes as input D^{sm} at once (not in parallel) and outputs θ according to $p^*(\theta|D^{sm})$. We can think of $L^*_{s,m}$ as the average logarithmic loss for the non-distributed Bayesian learning strategy for sample size sm. Note that the non-distributed Bayesian learning strategy is optimal in the sense that it attains the least average logarithmic loss for prediction from the training examples of size sm (see e.g. [2]).

Below we show a general relationship between $L_{s,m}(\mathcal{S})$ and $L^*_{s,m}$.

Lemma 2 *For a distributed learning system $\mathcal{S}$, let $q(\theta|D^{sm})$ be the probability density function according to which the output $\hat{\theta}$ of the p-learner in $\mathcal{S}$ is generated, and let*

$$\begin{aligned} m^*(D|D^{sm}) &\overset{\text{def}}{=} \int p(D|\theta)p^*(\theta|D^{sm})d\theta, \\ m_q(D|D^{sm}) &\overset{\text{def}}{=} \int p(D|\theta)q(\theta|D^{sm})d\theta, \\ D(m^* \parallel m_q) &\overset{\text{def}}{=} \int m^*(D|D^{sm})\ln\frac{m^*(D|D^{sm})}{m_q(D|D^{sm})}dD. \end{aligned}$$

Then for any distributed learning system $\mathcal{S}$, the following equation holds.

$$L_{s,m}(\mathcal{S}) = L^*_{s,m} + E_{D^{sm}}\left[D(m^* \parallel m_q)\right], \qquad (1)$$

where $E_{D^{sm}}$ denotes the expectation taken with respect to $p(D^{sm}) = \int \pi(\theta)p(D^{sm}|\theta)d\theta$.

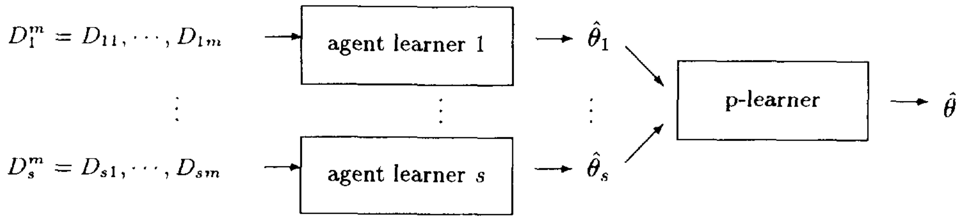


Figure 1: Distributed Learning System: Plain Model

Proof of Lemma 2. Observe first that for any function f,

$$\begin{aligned}&E_\theta E_{D^{sm}|\theta}E_{D|\theta}[f]\\ &=\int d\theta\pi(\theta)\int dD^{sm}p(D^{sm}|\theta)\int dDp(D|\theta)f\\ &=\int dD^{sm}\int dD\int d\theta p(D^{sm})p(D|\theta)\left(\frac{\pi(\theta)p(D^{sm}|\theta)}{p(D^{sm})}\right)f\\ &=\int dD^{sm}p(D^{sm})\int dD\int d\theta p(D|\theta)p^*(\theta|D^{sm})f\\ &=E_{D^{sm}}E_{D|D^{sm}}[f],\end{aligned}$$

where $E_{D|D^{sm}}$ and $E_{D^{sm}}$ denote the expectations taken with respect to $m^*(D|D^{sm}) = \int p(D|\theta)p^*(\theta|D^{sm})d\theta$, and $p(D^{sm}) = \int \pi(\theta)p(D^{sm}|\theta)d\theta$, respectively. The above relation immediately yields

$$\begin{aligned}L_{s,m} &= L^*_{s,m} + E_\theta E_{D^{sm}|\theta}E_{D|\theta}\left[\ln\frac{m^*(D|D^{sm})}{m_q(D|D^{sm})}\right]\\ &= L^*_{s,m} + E_{D^{sm}}E_{D|D^{sm}}\left[\ln\frac{m^*(D|D^{sm})}{m_q(D|D^{sm})}\right]\\ &= L^*_{s,m} + E_{D^{sm}}[D(m^*||m_q)].\end{aligned}$$

This completes the proof of Lemma 2. □

Since $D(m^* \parallel m_q) \geq 0$, we immediately see from (1) that

$$L_{s,m}(\mathcal{S}) \geq L^*_{s,m},$$

where the equality holds if and only if

$$q(\theta|D^{sm}) = p^*(\theta|D^{sm}), \qquad (2)$$

for every θ and D^{sm}. A distributed learning system which realizes (2) is *ideal* in the sense that it can realize the same learning performance as the non-distributed Bayesian learning strategy, while it utilizes the parallelism of the system to make the speed of learning s-times as fast as non-distributed Bayesian learning strategy. However, (2) doesn't necessarily hold for most distributed learning systems with $s(\geq 2)$ agent learners. Then the effectiveness of a distributed learning system is measured in terms of how the average logarithmic loss for the system is deviated from $L^*_{s,m}$. Hence we may measure the performance of a distributed learning system $\mathcal{S}$ in terms of $L_{s,m}(\mathcal{S}) - L^*_{s,m}$, which we call the *additional average logarithmic loss* (for short, *additional loss*) for $\mathcal{S}$. We wish to design a distributed learning system such that the additional loss is as small as possible, while keeping s large enough to attain a significant speed-up of learning over the non-distributed Bayesian learning strategy.

2.2 ALGORITHM

The performance of a distributed learning system depends on what types of algorithms are used for agent learners and the p-learner. Now we introduce the distributed cooperative Bayesian strategy in the plain model as a specific type of distributed learning systems.

Definition 3 *We define the* distributed cooperative Bayesian learning strategy *(in the plain model), which we abbreviate as* DCB, *as the distributed learning system satisfying the following:*
1) Each agent learner employs the Gibbs algorithm, *i.e., the ith agent learner takes as input a sequence $D_i^m = D_{i1}\cdots D_{im}$ and outputs the parameter value $\hat\theta_i$ chosen randomly according to the Bayes posterior density function $p(\theta_i|D_i^m)$, which is calculated as*

$$p(\theta|D_i^m) = \frac{\pi(\theta)\prod_{j=1}^m p(D_{ij}|\theta)}{\int \pi(\theta)\prod_{j=1}^m p(D_{ij}|\theta)d\theta} \quad (i = 1,\cdots,s).$$

2) The p-learner outputs the parameter value of the following form;

$$\hat\theta = \frac{1}{s}\sum_{i=1}^s \hat\theta_i. \qquad (3)$$

Remarks.

1) Each agent learner in DCB employs the Gibbs algorithm for m examples while the non-distributed Bayesian learning strategy employs the Gibbs algorithm for sm examples. Hence we see that the computational complexity for DCB is $(1/s)$-times as large as that for the non-distributed Bayesian learning strategy, assuming that the computation time is linear in the sample size and the size of agents. Then we say that *DCB runs s-times as fast as the non-distributed Bayesian learning strategy.*

2) The output of the p-learner in the plain model may be generalized to the form;

$$f(\hat\theta_1,\cdots,\hat\theta_s), \qquad (4)$$

where f is a function from Θ^s to Θ, which must be chosen appropriately depending on the hypothesis class $\mathcal{C}$. For example, one can expect that there exists some case where the DCB with $f(\hat\theta_1,\cdots\hat\theta_s) = (\prod_{i=1}^s \hat\theta_i)^{1/s}$ works better than (3).

Example 4 Let $\mathcal{D} = \mathbf{R}$ and $\mathcal{C} = \{N(\theta, \sigma^2 : D) : \theta \in (-\infty, \infty),\ \sigma \text{ is known }\}$. Here $N(\theta, \sigma^2 : D)$ denotes the Gaussian distribution with density with mean θ and variance σ^2 : $(1/\sqrt{2\pi}\sigma)\exp\{-\frac{(D-\theta)^2}{2\sigma^2}\}$. (*Note*: We may write $N(\theta, \sigma)$ instead of $N(\theta, \sigma^2 : D)$ when D is trivial from the context.) Let the prior density of θ be $N(\theta_0, \sigma_0^2 : \theta)$ where θ_0 and $\sigma_0 (< \infty)$ are known. Then the ith agent learner outputs $\hat{\theta}_i$ chosen randomly according to the following distribution:

$$\begin{aligned}\hat{\theta}_i &\sim p(\theta_i | D_i^m) \\ &= N\left(\frac{\sigma_0^2 \sum_{j=1}^m D_{ij} + \sigma^2\theta_0}{\sigma_0^2 m + \sigma^2}, \frac{\sigma^2\sigma_0^2}{\sigma_0^2 m + \sigma^2}\right) \\ & \qquad (i = 1, \cdots, s).\end{aligned}$$

If the p-learner outputs $\hat{\theta}$ calculated as in (3), then it can be easily verified that

$$\begin{aligned}\hat{\theta} &\sim q(\theta | D^{sm}) \\ &= N\left(\frac{\sigma_0^2(\sum_{i=1}^s \sum_{j=1}^m D_{ij}) + s\sigma^2\theta_0}{s(\sigma_0^2 m + \sigma^2)}, \frac{\sigma^2\sigma_0^2}{s(\sigma_0^2 m + \sigma^2)}\right).\end{aligned}$$

On the other hand, the Bayes posterior density $p^*(\theta|D^{sm})$ is given by

$$p^*(\theta|D^{sm}) = N\left(\frac{\sigma_0^2(\sum_{i=1}^s \sum_{j=1}^m D_{ij}) + \sigma^2\theta_0}{sm\sigma_0^2 + \sigma^2}, \frac{\sigma^2\sigma_0^2}{sm\sigma_0^2 + \sigma^2}\right).$$

In this case $q(\theta|D^{sm})$ and $p^*(\theta|D^{sm})$ differ each other when $s \geq 2$ and $\sigma_0 < \infty$. Note that it is straightforward to extend the above analysis to the case where θ is multi-dimensional.

Example 5 Let $\mathcal{D} = \mathbf{Z}^+ \cup \{0\}$ and let $\mathcal{C} = \{\frac{e^{-\theta}\theta^D}{D!} : \theta > 0\}$ be a class of Poisson distributions. Let the prior distribution of θ be the Gamma distribution with a density function $G(\alpha, \beta) = \theta^{\alpha-1}e^{-\frac{\theta}{\beta}}/\beta^\alpha\Gamma(\alpha)$, where $\alpha > 0$ and $\beta > 0$ are given, and Γ denotes the gamma function. Then each agent learner outputs $\hat{\theta}_i$ chosen randomly according to the following distribution:

$$\begin{aligned}\hat{\theta}_i &\sim p(\theta_i | D_i^m) \\ &= G\left(\alpha + \sum_{j=1}^m D_{ij}, \left(m + \frac{1}{\beta}\right)^{-1}\right) \\ & \qquad (i = 1, \cdots, s).\end{aligned}$$

If the p-learner outputs $\hat{\theta}$ calculated as in (3), then the distribution of $\hat{\theta}$ is given by

$$\begin{aligned}\hat{\theta} &\sim q(\theta | D^{sm}) \\ &= G\left(s(\alpha - 1) + 1 + \sum_{i=1}^s \sum_{j=1}^m D_{ij}, \left(sm + \frac{s}{\beta}\right)^{-1}\right).\end{aligned}$$

On the other hand, the Bayes posterior density $p^*(\theta|D^{sm})$ is given by

$$p^*(\theta|D^{sm}) = G\left(\alpha + \sum_{i=1}^s \sum_{j=1}^m D_{ij}, \left(sm + \frac{1}{\beta}\right)^{-1}\right).$$

In this case $q(\theta|D^{sm})$ and $p^*(\theta|D^{sm})$ differ each other when $s \neq 1$.

2.3 ANALYSIS

This section analyzes the performance of DCB in term of its additional loss. Theorem 6 gives an upper bound on the additional loss for the DCB in Example 4.

Theorem 6 *For the DCB for the Gaussian family with a constant variance as in Example 4, for sufficiently large s and m, the additional loss for the DCB is evaluated as follows:*

$$\begin{aligned}& L_{s,m}(\text{DCB}) - L^*_{s,m} \\ &= \frac{\sigma^2\sigma_0^2}{2(m\sigma_0^2 + \sigma^2)^2}\left(\frac{s-1}{s}\right)^2 (1 + o(1)), \qquad (5)\end{aligned}$$

where $o(1)$ goes to zero as sm goes to infinity.

Remarks.

1) Recall that DCB runs s-times as fast as the non-distributed Bayesian learning strategy. On the other hand, as seen from (5), the main term of the additional loss for the DCB is an increasing function of s. Thus Eq.(5) precisely quantifies the trade-off between the average prediction accuracy for the DCB and the degree of speed-up of learning over the non-distributed Bayesian learning strategy. Further we see that as s increases for fixed m, the main term of the additional loss for DCB would approach $\sigma^2\sigma_0^2/2(m\sigma_0^2 + \sigma^2)^2$, which can be thought of a loss inevitable due to the parallelism.

2) Specifically, for any $\varepsilon > 0$, if we set $m = O((1 - \varepsilon)/\sqrt{\varepsilon})$ and $s = O(1/\varepsilon)$, then the additional loss for the DCB can be made at most ε, while the DCB runs $O(1/\varepsilon)$-times as fast as the non-distributed Bayesian learning strategy.

3) Consider the case where we are allowed to use a *uniform prior* for θ, which can be thought of the Gaussian prior $N(\theta_0, \sigma_0^2)$ with $\sigma_0 = \infty$. Then it is immediately seen that this prior makes $L_{s,m}(\text{DCB})$ completely coincide with $L^*_{s,m}$. Thus the DCB in this case has the exactly same prediction performance as the non-distributed Bayesian learning strategy.

Proof of Theorem 6. We start with a lemma on a general result on the Kullback-Leibler divergence.

Lemma 7 *For $p = N(\mu_1, \sigma_1^2)$ and $q = N(\mu_2, \sigma_2^2)$, the divergence $D(p||q)$ is given by*

$$D(p||q) = \ln\frac{\sigma_2}{\sigma_1} + \frac{1}{2\sigma_2^2}(\sigma_1^2 + (\mu_1 - \mu_2)^2) - \frac{1}{2}.$$

Next observe that $m_q(D|D^{sm})$ and $m^*(D|D^{sm})$ are calculated as follows.

$$\begin{aligned}& m_q(D|D^{sm}) \\ &= N\left(\frac{\sigma_0^2\sum_{i,j} D_{ij} + s\sigma^2\theta_0}{s(\sigma_0^2 m + \sigma^2)}, \sigma^2 + \frac{\sigma^2\sigma_0^2}{s(m\sigma_0^2 + \sigma^2)}\right), \\ & m^*(D|D^{sm}) \\ &= N\left(\frac{\sigma_0^2\sum_{i,j} D_{ij} + \sigma^2\theta_0}{sm\sigma_0^2 + \sigma^2}, \sigma^2 + \frac{\sigma^2\sigma_0^2}{sm\sigma_0^2 + \sigma^2}\right).\end{aligned}$$

Using Lemma 7 we see that $D(m^*||m_q)$ is expanded as follows:

$$
\begin{aligned}
&D(m^* \parallel m_q)\\
&= \frac{1}{2}\ln\frac{\sigma^2+\sigma^2\sigma_0^2/(sm\sigma_0^2+\sigma^2)}{\sigma^2+\sigma^2\sigma_0^2/s(m\sigma_0^2+\sigma^2)}\\
&\quad+\frac{\sigma^2+\sigma^2\sigma_0^2/s(m\sigma_0^2+\sigma^2)}{2(\sigma^2+\sigma^2\sigma_0^2/(sm\sigma_0^2+\sigma^2))}\\
&\quad+\frac{\sigma^4\sigma_0^4\left(\sum_{i,j}D_{ij}-sm\theta_0\right)^2\left(\frac{s-1}{s}\right)^2}{2\left(\sigma^2+\frac{\sigma^2\sigma_0^2}{s(m\sigma_0^2+\sigma^2)}\right)(m\sigma_0^2+\sigma^2)^2(sm\sigma_0^2+\sigma^2)^2}-\frac{1}{2}\\
&= \frac{\sigma^2}{2(m\sigma_0^2+\sigma^2)^2}\left(\frac{s-1}{s}\right)^2\left(\frac{1}{sm}\sum_{i,j}D_{ij}-\theta_0\right)^2\\
&\quad\times(1+o(1)), \qquad (6)
\end{aligned}
$$

where $o(1)$ goes to zero uniformly with respect to D^{sm} as sm goes to infinity. Notice here that it is immediately proven that

$$E_{D^{sm}}\left[\left(\frac{1}{sm}\sum_{i,j}D_{ij}-\theta_0\right)^2\right]=\sigma_0^2. \qquad (7)$$

Taking an expectation of (6) with respect to D^{sm} and then using (7) yield

$$
\begin{aligned}
&E_{D^{sm}}\left[D(m^* \parallel m_q)\right]\\
&= \frac{\sigma^2\sigma_0^2}{2(m\sigma_0^2+\sigma^2)^2}\left(\frac{s-1}{s}\right)^2(1+o(1)).
\end{aligned}
$$

Combining this fact with Lemma 2 yields (5). This completes the proof of Theorem 6. □

Next Theorem 8 gives an asymptotic bound on the additional loss for the DCB in Example 5.

Theorem 8 *For the DCB for the Poisson family with $\alpha = 1$ in the prior density as in Example 5, for sufficiently large s and m, the additional loss for the DCB is evaluated as follows.*

$$L_{s,m}(\mathrm{DCB})-L^*_{s,m}\le\frac{1}{m}\left(\frac{s-1}{s}\right)(1+o(1)), \qquad (8)$$

where $o(1)$ goes to zero as sm goes to infinity.

Remarks.

1) As seen from (8), the main term of the additional loss for the DCB is an increasing function of s. As with Example 4, Eq.(5) quantifies the trade-off between the average prediction accuracy for the DCB and the degree of speed-up of learning over the non-distributed Bayesian learning strategy. Further we see that as s increases for fixed m, the main term of the additional loss for DCB would approach $1/m$, which can be thought of a loss inevitable due to the parallelism.

2) Specifically, for any $\varepsilon > 0$, if we set $m = O((1-\varepsilon)/\varepsilon)$ and $s = O(1/\varepsilon)$, then the additional loss for the DCB can be made at most ε, while the DCB runs $O(1/\varepsilon)$-times as fast as the non-distributed Bayesian learning strategy.

Proof of Theorem 8. Writing $\sum_{i,j}D_{i,j}$ as τ for the sake of notational simplicity, observe first that $m_q(D|D^{sm})$ and $m^*(D|D^{sm})$ are calculated as follows.

$$
\begin{aligned}
m^*(D|D^{sm}) &= \frac{(sm+1/\beta)^{1+\tau}(D+\tau)!}{(1+sm+1/\beta)^{D+\tau}D!\tau!}\\
m_q(D|D^{sm}) &= \frac{(sm+s/\beta)^{1+\tau}(D+\tau)!}{(1+sm+s/\beta)^{D+\tau}D!\tau!}.
\end{aligned}
$$

Using these formulas a simple calculation shows that

$$
\begin{aligned}
&\ln\frac{m^*(D|D^{sm})}{m_q(D|D^{sm})}\\
&= D\ln\frac{1+sm+s/\beta}{1+sm+1/\beta} \qquad (9)\\
&\quad+\left(1+\sum_{i,j}D_{i,j}\right)\ln\frac{(1+sm+s/\beta)(sm+1/\beta)}{(1+sm+1/\beta)(sm+s/\beta)}. \qquad (10)
\end{aligned}
$$

Notice here that as for the expectation of D we have

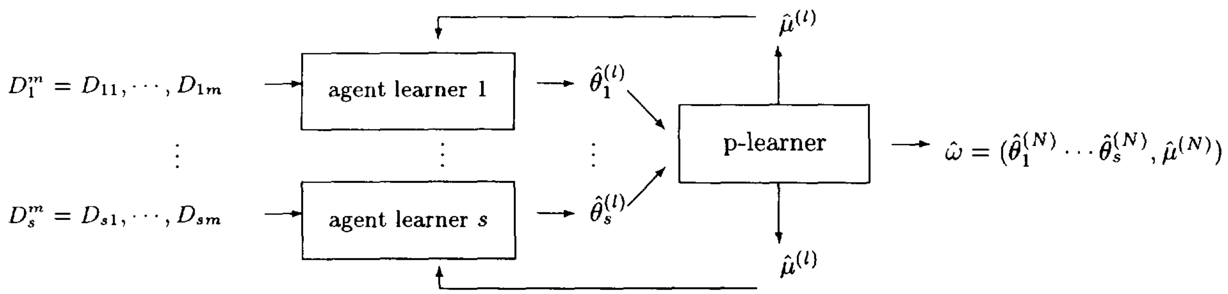
$$
\begin{aligned}
&E_{D^{sm}}E_{D|D^{sm}}[D]\\
&= E_{D^{sm}}\left[\sum_D D\int p(D|\theta)p^*(\theta|D^{sm})d\theta\right]\\
&= E_{D^{sm}}\left[\int\theta p^*(\theta|D^{sm})d\theta\right] \qquad (11)\\
&= E_{D^{sm}}\left[\left(1+\sum_{i,j}D_{i,j}\right)(sm+1/\beta)^{-1}\right] \qquad (12)\\
&= (1+\beta sm)(sm+1/\beta)^{-1} \qquad (13)\\
&= \beta(1+o(1)), \qquad (14)
\end{aligned}
$$

where $o(1)$ goes to zero as sm goes to infinity. In deriving (11) we have used the fact that the mean of the Poisson distribution $e^{-\theta}\theta^D/D!$ is θ. In deriving (12) and (13) we have used the fact that the mean of the Gamma distribution $G(\alpha,\beta)$ is $\alpha\beta$. Further notice that the term (10) is always less than zero. Thus taking an expectation of (9) with respect to D and D^{sm} and then plugging (14) into the expected form of (9) yields

$$
\begin{aligned}
&E_{D^{sm}}E_{D|D^{sm}}\left[\ln\frac{m^*(D|D^{sm})}{m_q(D|D^{sm})}\right]\\
&\le E_{D^{sm}}E_{D|D^{sm}}\left[D\ln\frac{1+sm+s/\beta}{1+sm+1/\beta}\right]\\
&= \frac{1}{m}\left(\frac{s-1}{s}\right)(1+o(1)),
\end{aligned}
$$

where we have used the fact that $\ln\frac{1+sm+s/\beta}{1+sm+1/\beta}=(s-1)(1+o(1))/\beta sm$ and $o(1)$ goes to zero as sm goes to infinity.

Combining this with Lemma 2 yields (8). This completes the proof of Theorem 8. □



$D_1^m = D_{11}, \cdots, D_{1m}$
agent learner 1
$\hat{\theta}_1^{(l)}$
$D_s^m = D_{s1}, \cdots, D_{sm}$
agent learner s
$\hat{\theta}_s^{(l)}$
p-learner
$\hat{\mu}^{(l)}$
$\hat{\mu}^{(l)}$
$\hat{\omega} = (\hat{\theta}_1^{(N)} \cdots \hat{\theta}_s^{(N)}, \hat{\mu}^{(N)})$

Figure 2: Distributed Learning System: Hierarchical Model

3 HIERARCHICAL MODEL

3.1 MODEL

In this section we consider the case where a probability distribution from which examples are generated is not identical over all agent learners, but there exists a hyper-parameter that relates all the distributions one another.

A distributed learning system $\mathcal{S}$ in this case consists of s agent learners and a single p-leaner as in the plain model. The only difference between this case and the plain model is an assumption on the probability distribution of examples which agent learners observe. Below we give a mathematical form of this assumption. Let $C = \{p(D|\theta) : \theta \in \Theta \subset \mathbf{R}^k\}$ be a given hypothesis class of probability density functions specified by a k-dimensional real-valued parameter θ, belonging to a parameter space Θ.

Assumption 9
1) Each agent independently observes a sequence of examples each of which is independently generated according to the target distribution $p(D|\theta_i)$ $(i = 1, \cdots, s)$*, belonging to C.*
2) All the parameters θ_i $(i = 1, \cdots, s)$ are unknown, and each of them is independently generated according to a prior distribution over Θ with a density function $p(\theta|\mu)$, where μ is an unknown hyper-parameter.
3) The hyper-parameter μ is generated according to a known prior distribution with a density function $\pi(\mu)$.

We call the model in which Assumption 9 holds the *hierarchical model.* We set $\omega = (\theta_1, \cdots, \theta_s, \mu)$, which is a $(ks+1)$-dimensional real-valued vector. Let Θ and $\mathcal{M}$ be the ranges of θ and μ, respectively. We denote the range of ω as $\Omega = \Theta^s \times \mathcal{M}$.

Let positive integers m and N be given. A *distributed learning system $\mathcal{S}$* in the hierarchical model makes the following iteration process.

Let an initial value $\mu^{(0)}$ be given. At the lth iteration step $(l = 1, 2, \cdots, N)$, for each i $(i = 1, \cdots, s)$, the ith agent learner takes a sequence $D_i^m = D_{i1} \cdots D_{im}$ of m examples and $\mu^{(l-1)}$ as input, and outputs an estimate $\hat{\theta}_i^{(l)}$ of θ_i. The p-learner takes $\hat{\theta}_1^{(l)} \cdots, \hat{\theta}_s^{(l)}$ as input and outputs an estimate $\hat{\mu}^{(l)}$ of μ. The output $\hat{\mu}^{(l)}$ of the p-learner is sent back to all of the agent learners. This process goes on iteratively with respect to l. The parameter value $\hat{\omega} = (\hat{\theta}_1^{(N)}, \cdots, \hat{\theta}_s^{(N)}, \hat{\mu}^{(N)})$ obtained after N iterations is an output of the distributed learning system $\mathcal{S}$ in the hierarchical model. (See Figure 2.)

Let $D^{sm} = D_1^m \cdots D_s^m$. Let $q(\omega|D^{sm})$ be the probability density according to which the output $\hat{\omega}$ of $\mathcal{S}$ with N iterations is generated. Let $\mathbf{D} = (D_1 \cdots D_s)$ where D_i denotes a random variable representing an example that the ith agent learner observes. We write the distribution of $\mathbf{D}$ as

$$p(\mathbf{D}|\omega) = p(D_1 \cdots D_s|\omega) = \prod_{i=1}^{s} p(D_i|\theta_i). \qquad (15)$$

We measure the performance of a distributed learning system $\mathcal{S}$ by the *average logarithmic loss* $L_{s,m,N}$ for predicting the probability of unseen data, which is defined by

$$L_{s,m,N}(\mathcal{S}) \overset{\text{def}}{=} E_\omega E_{D^{sm}|\omega} E_{\mathbf{D}|\omega}\left[-\ln m_q(\mathbf{D}|D^{sm})\right],$$

where

$$m_q(\mathbf{D}|D^{sm}) \overset{\text{def}}{=} \int p(\mathbf{D}|\omega) q(\omega|D^{sm}) d\omega. \qquad (16)$$

We also define the *average Bayes risk* similarly with the plain model, by

$$L^*_{s,m} \overset{\text{def}}{=} E_\omega E_{D^{sm}|\omega} E_{\mathbf{D}|\omega}\left[-\ln m^*(\mathbf{D}|D^{sm})\right],$$

where

$$m^*(\mathbf{D}|D^{sm}) \overset{\text{def}}{=} \int p(\mathbf{D}|\omega) p^*(\omega|D^{sm}) d\omega, \qquad (17)$$

and $p^*(\omega|D^{sm})$ is the *Bayes posterior density function* of ω from D^{sm} defined as

$$p^*(\omega|D^{sm}) \qquad (18)$$
$$= \frac{\pi(\mu)\prod_{i=1}^{s}\left(p(\theta_i|\mu)\prod_{j=1}^{m} p(D_{ij}|\theta_i)\right)}{\int \pi(\mu)\prod_{i=1}^{s}\left(p(\theta_i|\mu)\prod_{j=1}^{m} p(D_{ij}|\theta_i)\right) d\mu d\theta_1 \cdots d\theta_s}.$$

The *non-distributed Bayesian learning strategy* (in the hierarchical model) is referred here to as the strategy which takes as input D^{sm} at once (not in parallel) and outputs ω according to $p^*(\omega|D^{sm})$. We can think of $L^*_{s,m}$ as the average logarithmic loss for the non-distributed Bayesian learning strategy. We define the *additional* *average logarithmic loss* (for short, *additional loss*) for $\mathcal{S}$ by $L_{s,m,N}(\mathcal{S}) - L^*_{s,m}$. It can be proven as with Lemma 2 that the additional loss is not less than zero, and is zero if and only if $q(\omega|D^{sm}) = p^*(\omega|D^{sm})$ for every ω and D^{sm}. We wish to design a distributed learning system such that its average logarithmic loss is as small as possible.

In this section, we also introduce another performance measure $d(\mathcal{S})$ for the sake of analytical simplicity. For a distributed learning system $\mathcal{S}$, we define $d(\mathcal{S})$ by

$$d(\mathcal{S}) \overset{\text{def}}{=} E_{D^{sm}} \left[\int |m_q(\mathbf{D}|D^{sm}) - m^*(\mathbf{D}|D^{sm})| d\mathbf{D} \right],$$

where the notations of $m_q(\mathbf{D}|D^{sm})$ and $m^*(\mathbf{D}|D^{sm})$ follow (16) and (17), respectively. The expectation $E_{D^{sm}}$ is taken with respect to $p(D^{sm}) = \int \pi(\omega) p(D^{sm}|\omega) d\omega$.

We can think of $d(\mathcal{S})$ as the expected variation distance between the average distribution for $\mathcal{S}$ and that for the non-distributed Bayesian learning strategy. Thus $d(\mathcal{S})$ can be thought of as an analogue of the additional loss $L_{s,m,N}(\mathcal{S}) - L^*_{s,m}$.

3.2 ALGORITHM

For the hierarchical model, we define a version of DCB in the plain model as follows.

Definition 10 *We define the* distributed cooperative Bayesian strategy *(in the hierarchical model), which we abbreviate as* DCB, *as the distributed learning system that makes the following iteration N times(N is given): At the lth iteration ($l = 1, 2, \cdots, N$),*
1) Each agent learner employs the Gibbs algorithm, *i.e., the ith agent learner takes as input a sequence $D_i^m = D_{i1} \cdots D_{im}$ and the latest estimate $\hat{\mu}^{(l-1)}$ of μ, then outputs the parameter value $\hat{\theta}_i^{(l)}$ chosen randomly according to the Bayes posterior distribution with a density function $p(\theta_i|D_i^m, \hat{\mu}^{(l-1)})$, which is calculated as*

$$p(\theta_i|D_i^m, \hat{\mu}^{(l-1)}) = \frac{p(\theta_i|\hat{\mu}^{(l-1)}) \prod_{j=1}^m p(D_{ij}|\theta_i)}{\int p(\theta_i|\hat{\mu}^{(l-1)}) \prod_{j=1}^m p(D_{ij}|\theta_i) d\theta_i}.$$

2) The p-learner also employs the Gibbs algorithm, i.e., it takes the latest estimates $\hat{\theta}_1^{(l)}, \cdots, \hat{\theta}_s^{(l)}$ as input, and then outputs the parameter value $\hat{\mu}^{(l)}$ chosen randomly according to the Bayes posterior distribution with a density function $p(\mu|\hat{\theta}_1^{(l)}, \cdots, \hat{\theta}_s^{(l)})$, which is calculated as

$$p(\mu|\hat{\theta}_1^{(l)}, \cdots, \hat{\theta}_s^{(l)}) = \frac{\pi(\mu) \prod_{i=1}^s p(\hat{\theta}_i^{(l)}|\mu)}{\int \pi(\mu) \prod_{i=1}^s p(\hat{\theta}_i^{(l)}|\mu) d\mu}.$$

The output of the system is $\hat{\omega} = (\hat{\theta}_1^{(N)}, \cdots, \hat{\theta}_s^{(N)}, \hat{\mu}^{(N)})$.

Remarks.

1) Let the outputs of agent learners and the p-learner at the lth iteration be $\hat{\theta}_i^{(l)}$ ($i = 1, \cdots, s$) and $\hat{\mu}^{(l)}$, respectively. Let us set $\hat{\omega}^{(l)} = (\hat{\theta}_1^{(l)}, \cdots, \hat{\theta}_s^{(l)}, \hat{\mu}^{(l)})$ ($l = 1, \cdots, N$). Then the process

$$\hat{\omega}^{(1)} \to \hat{\omega}^{(2)} \to \cdots \to \hat{\omega}^{(N)} \tag{19}$$

forms a Markov chain. It is immediately proven that the stationary distribution for this Markov chain is the Bayes posterior density $p^*(\omega|D^{sm})$ as in (18). The iteration process (19) can be thought of as a *Markov chain Monte Carlo* process.

2) Each agent learner in DCB employs the Gibbs algorithm for m examples N times, while the non-distributed Bayesian learning strategy employs the Gibbs algorithm for sm examples just at once (not in parallel). Hence we see that the computational complexity for DCB is (N/s)-times as large as that for the non-distributed Bayesian learning strategy, assuming that the computation time is linear in the sample size and the size of agents. Then we say that *DCB runs (s/N)-times as fast as the non-distributed Bayesian learning strategy.*

3) DCB has another computational merit in comparison with the non-distributed Bayesian learning strategy. As will be seen in Example 11,12, there exist some cases where the joint posterior distribution $p(\omega|D^{sm}) = p(\theta_1, \cdots, \theta_s, \mu|D^{sm})$ for the non-distributed Bayesian learning strategy is analytically hard to calculate, while the conditional distributions $p(\theta_1, \cdots \theta_s|D^m, \mu)$ and $p(\mu|\theta_1, \cdots, \theta_s)$ can be straightforwardly calculated. For such cases the DCBs are computationally tractable while the non-distributed Bayesian learning strategy is not.

Example 11 Let $\mathcal{D} = \{0, 1\}$ and let $\mathcal{C} = \{p(1|\theta) = \theta,\ p(0|\theta) = 1 - \theta\ :\ \theta \in [c, 1 - c]\}$ be a class of Bernoulli distributions for which the parameter value is restricted to $[c, 1 - c]$ for a given $c \in (0, 1/2)$. Let the prior distribution of θ_i with a hyper-parameter μ be $p(\theta_i|\mu) = D_c(\mu, n - \mu : \theta_i)$ ($i = 1, \cdots, s$) where n is a given positive real number, and $D_c(\alpha, \beta : \theta) \overset{\text{def}}{=} \theta^\alpha (1 - \theta)^\beta / \int_c^{1-c} \theta'^\alpha (1 - \theta')^\beta d\theta'$, where $\alpha, \beta > 0$, and $0 < c < 1$ are given. Let the prior distribution of μ be the uniform distribution over $[0, n]$. For each i, let m_i be the number of examples such that $D_{ij} = 1$ in the sequence $D_i^m = D_{i1} \cdots D_{im}$. Then we have

$$p(\theta_i|D_i^m, \mu) = D_c(m_i + \mu,\ m + n - m_i - \mu : \theta_i) \quad (i = 1, \cdots, s),$$

$$p(\mu|\theta_1, \cdots, \theta_s) = \left(\prod_{i=1}^s \theta_i\right)^\mu \left(\prod_{i=1}^s (1 - \theta_i)\right)^{n-\mu} \ln\left(\prod_{i=1}^s \frac{\theta_i}{1 - \theta_i}\right) \times \left(\left(\prod_{i=1}^s \theta_i\right)^n - \left(\prod_{i=1}^s (1 - \theta_i)\right)^n\right)^{-1}.$$

We can implement the iteration process in DCB using these conditional distributions.

Example 12 Let $\mathcal{D} = \mathbf{R}$ and let $\mathcal{C} = \{N(\theta, \sigma^2 : D) : \theta \in \mathbf{R},\ \sigma^2 \text{ is constant}\}$ be a class of Gaussian distributions with a constant variance. Let the prior distribution of θ_i with a hyper-parameter μ be $p(\theta_i|\mu) = N(\mu, \sigma_\theta^2 : \theta_i)$ ($i = 1, \cdots, s$) where σ_θ^2 is given. Let the prior distribution of μ be $N(\mu_0, \sigma_0^2 : \mu)$ where μ_0 and σ_0^2 are given. Then we have

$$p(\theta_i|D_i^m, \mu) = N\left(\frac{\sigma_\theta^2 \sum_{j=1}^m D_{ij} + \sigma^2 \mu}{m\sigma_\theta^2 + \sigma^2},\ \frac{\sigma_\theta^2 \sigma^2}{m\sigma_\theta^2 + \sigma^2}\right)$$

$$(i = 1, \cdots, s),$$

$$p(\mu|\theta_1, \cdots, \theta_s) = N\left(\frac{\sigma_\theta^2\mu_0 + \sigma_0^2\sum_{i=1}^s \theta_i}{\sigma_\theta^2 + s\sigma_0^2}, \frac{\sigma_\theta^2\sigma_0^2}{\sigma_\theta^2 + s\sigma_0^2}\right).$$

We can implement the iteration process in DCB using these conditional distributions.

3.3 ANALYSIS

First we consider the case where the parameter space is bounded. We start with the following assumption and notation.

Assumption 13 *The hyperthesis class $\mathcal{C} = \{p(D|\theta) : \theta \in \Theta\}$ satisfies the condition: There exist some positive numbers $\underline{c}$ and $\bar{c}$ such that for all D, for all θ, $0 < \underline{c} \leq p(D|\theta) \leq \bar{c} < \infty$.*

Below we denote $K(\cdot, \cdot)$ as a Markov chain transition kernel on a state space Ω, i.e., for $\omega, \omega' \in \Omega$, $K(\omega, \omega')$ denotes the transition probability density from ω to ω'.

For $\omega = (\theta_1, \cdots, \theta_s, \mu), \omega' = (\theta'_1, \cdots, \theta'_s, \mu') \in \Omega$, for a Markov chain transition kernel K, for $S \subset \Omega$, for any positive integer ℓ, we define $K^{(\ell)}(\omega, S)$ inductively by

$$\begin{aligned} K^{(\ell)}(\omega, S) &= \int K^{(\ell-1)}(\omega, \omega')K(\omega', S)d\omega', \\ &\quad (\ell = 2, 3, \cdots), \\ K^{(1)}(\omega, S) &= K(\omega, S). \end{aligned}$$

Theorem 14 gives a general upper bound on the average logarithmic loss for DCB for the case where the parameter space Ω is bounded.

Theorem 14 *Suppose that the parameter space $\Omega = \{\omega = (\theta_1, \cdots, \theta_s, \mu)\}$ is bounded. Also suppose that the prior density π over Ω is everywhere positive and continuous. Then for any DCB in the hierarchical model, for some $0 < C < \infty$, for any positive integer ℓ, we have the following upper bound on $d(\mathrm{DCB})$:*

$$d(\mathrm{DCB}) \leq C' E_{D^{sm}}\left[\rho_{s,m}^N\right]. \tag{20}$$

Here $0 < \rho_{s,m} < 1$ is a function of D^{sm}, calculated as

$$\rho_{s,m} = 1 - \int \inf_{\theta'_1, \cdots, \theta'_s} K^{(\ell)}(\omega', \omega)d\mu d\theta_1 \cdots d\theta_s,$$

and where

$$K(\omega', \omega) = \left(\prod_{i=1}^s p(\theta_i|D_i^m, \mu)\right) \cdot p(\mu|\theta'_1, \cdots, \theta'_s).$$

Under Assumption 13 for $\mathcal{C}$ in addition to the above conditions, for any DCB in the hierarchical model, for some $0 < C', \kappa < \infty$, for any positive integer ℓ, we have the following upper bound on the additional loss for DCB:

$$L_{s,m,N}(\mathrm{DCB}) - L^*_{s,m} \leq C'\kappa^s E_{D^{sm}}\left[\rho_{s,m}^N\right], \tag{21}$$

where the notation of $\rho_{s,m}$ is as above.

Specifically, if there exists some $0 < \rho < 1$ such that $\sup_{D^{sm}} \rho_{s,m} \leq \rho$, then for any $\varepsilon > 0$, by setting $s = O(\ln(1/\varepsilon))$ and $N = O((\ln(1/\varepsilon))/(\ln(1/\rho)))$, the additional loss for DCB and the distance $d(\mathrm{DCB})$ can be made at most ε while DCB runs $O(\ln(1/\rho))$-times as fast as the non-distributed Bayesian learning strategy.

Remark.

Theorem 14 shows that the additional loss for DCB converges to zero exponentially in the iteration number N. This implies that unlike the plain model, DCB in the hierarchical model can make its additional loss arbitrarily small by increasing the number N of feedback iterations. Theorem 14 characterizes the degree of speed-up of learning by DCB and its additional loss in terms of the quantity $\rho_{s,m}$, which depends on s, m, and the hyperthesis class.

In order to prove Theorem 14, we give Lemma 15 and Proposition 16. Lemma 15 relates the additional loss and $d(\mathcal{S})$ to the rate of convergence of MCMC.

Lemma 15 *For any distributed learning system $\mathcal{S}$ in the hierarchical model, let q be the probability density on Ω according to which the output of $\mathcal{S}$ with N feedback iterations is generated, and let p^* be the Bayes posterior density on Ω for the non-distributed Bayesian learning strategy. Then for some $0 < C < \infty$, the following inequality holds.*

$$d(\mathrm{DCB}) \leq CE_{D^{sm}}[d(q, p^*)], \tag{22}$$

where $d(q, p^*) \overset{\mathrm{def}}{=} \int |q(\omega|D^{sm}) - p^*(\omega|D^{sm})|d\omega$.

Further under Assumption 13 for $\mathcal{C}$, for any distributed learning system $\mathcal{S}$ in the hierarchical model, for some $0 < C', \kappa < \infty$, the following inequality holds.

$$L_{s,m,N}(\mathcal{S}) - L^*_{s,m} \leq C'\kappa^s E_{D^{sm}}[d(p^*, q)]. \tag{23}$$

(The proof of Lemma 15 is in Appendix.)

Recall that the iteration process in DCB induces a Markov chain with a stationary distribution $p^*(\omega|D^{sm})$. Proposition 16 gives an upper bound on the rate of convergence for the Markov chain Monte Carlo process.

Proposition 16 ([9]). *Let K be a Markov chain transition kernel over a bounded state space Ω and π be the stationary distribution of the Markov chain. Suppose that there are some probability measure Q over Ω, some positive integer ℓ, and some $\varepsilon > 0$ such that for all $\omega \in \Omega$, for all $S \in \Omega$,*

$$K^{(\ell)}(\omega, S) \geq \varepsilon Q(S). \tag{24}$$

Then for any initial distribution π_0, the distribution π_N of the Markov chain after N steps satisfies the following inequality:

$$d(\pi_N, \pi) \leq 2(1 - \varepsilon)^{\lfloor N/\ell \rfloor}, \tag{25}$$

where $\lfloor x \rfloor$ denotes the largest integer that doesn't exceed x.

Proof of Theorem 14. The Markov chain kernel induced by the iteration process in DCB is written as

$$K(\omega', \omega) = p(\mu|\theta'_1, \cdots, \theta'_s)\prod_{i=1}^s p(\theta_i|D_i^m, \mu),$$

where $\omega = (\theta_1, \cdots, \theta_s, \mu)$ and $\omega' = (\theta'_1, \cdots, \theta'_s, \mu')$. Observe that for any positive integer ℓ, for all $\omega, \omega' \in \Omega$, we have

$$K^{(\ell)}(\omega', \omega) \geq \varepsilon \cdot \frac{\inf_{\theta'_1 \cdots \theta'_s} K^{(\ell)}(\omega', \omega)}{\varepsilon},$$

where $\varepsilon \overset{\text{def}}{=} \int \inf_{\theta'_1, \cdots, \theta'_s} K^{(\ell)}(\omega', \omega) d\omega$. Notice here that $K^{(\ell)}(\omega', \omega)$ depends on $\theta'_1, \cdots, \theta'_s$ but not on μ'. Hence setting $Q(S) \overset{\text{def}}{=} \int_S (\inf_{\theta'_1 \cdots \theta'_s} K^{(\ell)}(\omega', \omega)/\varepsilon) d\omega$ for any $S \subset \Omega$, we see that Q forms a distribution over Ω, and thus (24) is satisfied. By Proposition 16, setting $\rho_{s,m} = 1 - \varepsilon$ gives

$$d(q, p^*) \leq 2\rho_{s,m}^N. \tag{26}$$

Combining (26) with (22) and (23) in Lemma 15 yields (20) and (21), respectively. This completes the proof of Theorem 14. □

Theorem 14 can be applied to Example 11, but cannot be applied to the case where the parameter space is unbounded or the probability value of each example may not be uniformly lower-bounded away from zero, as in Example 12. Below we consider the latter case to give another type of analysis for DCB with respect to the measure $d(\mathcal{S})$.

Theorem 17 gives a concrete upper bound on $d(\text{DCB})$ for Example 12.

Theorem 17 *For the DCB for the hierarchical Gaussian family as in Example 12, for some $0 < c_1, c_2 < \infty$ independent of s, m, and N, we have*

$$d(\text{DCB}) \leq O\left(e^{-\frac{c_1 N \ln m}{\ln(sN)}}\right) + O\left(\frac{N \ln m}{\ln(sN)} e^{-c_2 N}\right). \tag{27}$$

Let $c = \min\{c_1, c_2\}$. Specifically, for any $\varepsilon > 0$, if we set $N = O((1/c)\ln(1/\varepsilon))$, $m = O(1/\varepsilon)$, and $s = (1/\varepsilon)$, then for arbitrary small $\delta > 0$, $d(\text{DCB})$ can be made at most $\varepsilon^{1-\delta}$ while the DCB runs $O((1/\varepsilon)/\ln(1/\varepsilon))$-times as fast as the non-distributed Bayesian learning strategy.

Remarks.

1) By Theorem 17, $d(\text{DCB})$ can be made arbitrarily small by increasing N even when m is fixed, while the additional loss for DCB in the plain model cannot be made arbitrarily small for fixed m (see Theorem 6,8). This implies that the feedback of information from the p-learner to agent learners in the hierarchical model is so effective that DCB can work approximately as well as the the non-distributed Bayesian learning strategy as N becomes sufficiently large even for fixed m.

2) Theorem 17 guarantees that for the hierarchical Gaussian family as in Example 12, the DCB can attain significant speed-up of learning over the non-distributed Bayesian learning strategy, with the number of iterations, sample size, and agent size polynomial in $1/\varepsilon$ for prediction accuracy ε.

The outline of the proof of Theorem 17 basically follows Rosenthal's proof in ([10]) on the rate of convergence of the *variance component model*. Notice, however, that the model we deal with can be thought of as a simplified variant of the variance composed model, and thus our analysis is somewhat specific, and hence it makes our bound tighter than Rosenthal's.

In order to prove Theorem 17, we give Proposition 18 and Lemma 19, 20. Note that for the case where the parameter space is bounded, we can apply Proposition 16 to obtain bounds (20) and (21). For the case where the parameter space is not bounded, however, it is usually impossible to require the condition in Proposition 16 that Eq.(24) hold for an arbitrary initial value $\omega \in \Omega$. Proposition 18, which was first proven by Rosenthal, enables us to do analysis similarly with Proposition 16 by requiring that the same type equation as Eq.(24) hold for an arbitrary initial value in some *subset* of Ω rather than in the whole state space Ω.

Proposition 18 ([10]). *Let K be a Markov chain transition kernel over a state space Ω and π be the stationary distribution of the Markov chain. Suppose that there are measurable subsets $R_1, R_2 \subset \Theta$, some probability measure Q over Ω, some positive integer ℓ_0, and some $\varepsilon_1, \varepsilon_2 > 0$ such that for all $\omega \in R_1$,*

$$K^{(\ell_0)}(\omega, R_2) \geq \varepsilon_1, \tag{28}$$

and for all $\omega_2 \in R_2$, for all $S \in \Omega$,

$$K(\omega, S) \geq \varepsilon_2 Q(S). \tag{29}$$

Then for any initial distribution π_0 supported entirely in R_1, the distribution π_N of the Markov chain after N steps satisfies

$$\begin{aligned} & d(\pi_N, \pi) \\ & \leq (1 - \varepsilon_1\varepsilon_2)^{\lfloor N/(\ell_0+1)\rfloor} + A + 2\lfloor N/(\ell_0 + 1)\rfloor B, \end{aligned} \tag{30}$$

where

$$\begin{aligned} A &= 1 - \pi(R_1), \\ B &= 1 - \inf_{\omega \in R_1} K^{(\ell_0+1)}(\omega, R_1). \end{aligned}$$

Lemma 19 gives a concrete method for constructing R_1, R_2, Q, ℓ_0 to satisfy the conditions in Proposition 18 for the model as in Example 12 .

Lemma 19 *Let $\overline{D} = (1/sm)\sum_{i=1}^{s}\sum_{j=1}^{m} D_{ij}$. For the model in Example 12, define $R_1, R_2 \subset \Omega$ by*

$$R_1 \overset{\text{def}}{=} \left\{\omega \in \Omega : \left|\overline{D} - \frac{1}{s}\sum_{i=1}^{s}\theta_i\right| \leq N^{\frac{1}{2}}\right\}, \tag{31}$$

$$R_2 \overset{\text{def}}{=} \left\{\omega \in \Omega : \left|\overline{D} - \frac{1}{s}\sum_{i=1}^{s}\theta_i\right| \leq \frac{2}{\sqrt{s}}\right\}. \tag{32}$$

Also let Q be the probability measure which first chooses μ randomly according to the uniform distribution over the set

$$I = \left[\overline{D} - 2/\sqrt{s},\ \overline{D} + 2/\sqrt{s}\right],$$

then chooses θ_i independently randomly according to

$$N\left(\frac{\sigma_\theta^2 \sum_{j=1}^{m} D_{ij} + \sigma^2\mu}{m\sigma_\theta^2 + \sigma^2}, \frac{\sigma_\theta^2\sigma^2}{m\sigma_\theta^2 + \sigma^2}\right) \quad (i = 1, \cdots, s).$$

Then for sufficiently large s, for the Markov chain that the DCB in Example 12 induces, all the conditions (28),(29) in Proposition 18 are satisfied by R_1, R_2, Q as above, some positive integers $\varepsilon_1, \varepsilon_2$ independent of s, m, N, and $\ell_0 = O((\ln(sN))/(\ln m))$.

(The proof of Lemma 19 is in Appendix.)

Lemma 20 shows concrete evaluation of A and B as in Proposition 18, which can be easily proven by considering the tail of the normal distribution.

Lemma 20 *For $R_1, R_2, Q, \varepsilon_1, \varepsilon_2, \ell_0$ as in Lemma 19, A and B as in (30) in Proposition 18 are both upper-bounded by expressions of the form $d_1 e^{-d_2 N}$, with $d_1, d_2 > 0$ independent of s, m, and N.*

Proof of Theorem 17. From Proposition 18 and Lemma 19, 20, we see that for the DCB in Example 12 with N feedback iterations, for some $0 < c_1, c_2 < \infty$ independent of s, m, and N, the following inequality holds.

$$d(q, p^*) \leq O\left(e^{-\frac{c_1 N \ln m}{\ln(sN)}}\right) + O\left(\frac{N \ln m}{\ln(sN)} e^{-c_2 N}\right). \quad (33)$$

Plugging (33) into (22) in Lemma 15 yields (27). This completes the proof of Theorem 17. □

4 Extensions of DCB

4.1 MULTI-HIERARCHICAL MODEL

First we consider an extension of DCB into the hierarchical model with multi hyper-levels. Recall that the parameters $\theta = (\theta_1 \cdots \theta_s)$ and μ have a hierarchical structure in DCB in the hierarchical model. We denote this structure as

$$\theta|\mu.$$

We can consider a general hierarchical version of this structure, having k stages as follows:

$$\xi_1|\xi_2|\cdots|\xi_k.$$

Then the hierarchical structure implies the following probability relations among parameters:

$$\begin{aligned} & p(\xi_i|D^{sm}, \xi_j, \ (j \neq i)) \\ & = \begin{cases} p(\xi_1|D^{sm}, \xi_2), \\ p(\xi_i|\xi_{j-1}, \xi_{j+1}) \quad (i = 2, \cdots, k-1), \\ p(\xi_k|\xi_{k-1}). \end{cases} \end{aligned}$$

We may obtain a variant of DCB which uses these conditional distributions hierarchically to form an iteration process called the *Gibbs sampler* ([4]). In it each iteration step the parameter values are sampled in the order: $\xi_1 \to \xi_2 \to \cdots \to \xi_k$. After a given number of iterations the variant of DCB outputs an estimate of $(\xi_1, \cdots, \xi_k)$.

4.2 GENERAL DECISION-THEORETIC MODEL

Next we consider an extension of DCB to the general decision-theoretic scenario. In Section 2-3, we have developed a theory in the case where the hypothesis class is a class of probability distributions and the prediction loss is measured in terms of the logarithmic loss. We can extend our theory into the general case where the hypothesis class is a class of parametric real-valued functions of the form of $C = \{f_\theta(x) : \theta \in \Theta \subset \mathbf{R}^k\}$ and the prediction loss is measure in terms of a general loss function, which we write as $L(D : f_\theta)$ where $D = (x, y)$. For example, the square loss is written as $L(D : f_\theta) = (y - f_\theta(x))^2$.

Below let us focus on the plain model. Let the range of the loss function L be $[0, 1]$. We assume that the parameter θ is generated according to the prior distribution with density $\pi(\theta)$. For given $D^{sm} = D_1^m \cdots D_s^m$, $D_i^m = D_{i1} \cdots D_{im}$ $(i = 1, \cdots, s)$, let us calculate $p(\theta_i|D_i^m)$ by

$$\begin{aligned} p(\theta_i|D_i^m) &= \frac{\pi(\theta)e^{-\lambda \sum_{j=1}^m L(D_{ij}:f_\theta)}}{\int \pi(\theta)e^{-\lambda \sum_{j=1}^m L(D_{ij}:f_\theta)} d\theta} \\ & \quad (i = 1, \cdots, s), \end{aligned}$$

where λ is a positive real number depending on the loss function L. Here we obtain a variant of DCB in which each agent learner outputs $\hat{\theta}_i$ chosen randomly according to $p(\theta_i|D_i^m)$ as above, and the p-learner calculates $\hat{\theta} = f(\hat{\theta}_1, \cdots, \hat{\theta}_s)$ using some appropriate deterministic function f.

In a decision-theoretic scenario, we might be concerned with *predicting* unseen data y for given input x rather than estimating the parameter θ. Below we describe how to predict future data using DCB. Let $q(\theta|D^{sm})$ denote the density function according to which the output $\hat{\theta}$ of DCB is generated. Rewrite $L(D : f_\theta)$ for $D = (x, y)$ as $L(y : f_\theta(x))$. On receiving x, DCB calculates the quantity:

$$\Delta(y) \stackrel{\text{def}}{=} \frac{1}{\lambda} \ln \int q(\theta|D^{sm}) e^{-\lambda L(y:f_\theta(x))} d\theta \quad (y = 0, 1), \quad (34)$$

and then predicts y with $\hat{y}$ such that

$$L(0, \hat{y}) \leq \Delta(0), \quad \text{and} \quad L(1, \hat{y}) \leq \Delta(1). \quad (35)$$

It is known (see e.g., [13], [6]) that under some smoothness conditions for a loss function, there exists a range of λ such that the prediction of y with $\hat{y}$ as in (35) is well-defined so that the prediction loss is upper-bounded by $\Delta(y)$ uniformly with respect to a correct value $y \in [0, 1]$. For example, $\lambda = 1/2$ for the square loss (see [13],[6]).

When $q(\theta|D^{sm})$ equals

$$p^*(\theta|D^{sm}) = \frac{\pi(\theta) \prod_{i=1}^s \prod_{j=1}^m e^{-\lambda L(D_{ij}:f_\theta)}}{\int \pi(\theta) \prod_{i=1}^s \prod_{j=1}^m e^{-\lambda L(D_{ij}:f_\theta)} d\theta},$$

the prediction strategy based on (34) becomes equivalent to Vovk's *aggregating strategy* ([13],[6]), and the quantity (34) is related to the notion of *Extended Stochastic Complexity (ESC)* ([14]). We can expect that the prediction strategy based on (34) works approximately as well as the aggregating strategy. It remains for future study to derive any bounds on the additional loss for the DCB for concrete examples.

The above general decision-theoretic variant of DCB in the plain model can also be readily extended to that in the hierarchical model.

5 CONCLUDING REMARKS

We have developed the plain model and the hierarchical model for distributed learning, as probabilistic versions of Kearns and Seung's model of population learning. Within these models we have proposed DCBs and analyzed their performance in terms of their additional losses (and the expected variation distance) as functions of sample size, agent size and iteration number. For some concrete hyperthesis classes we have demonstrated that DCB works approximately as well as the non-distributed Bayesian learning strategy, while attaining a significant speed-up of learning.

The following issues remain for future study.

1) Any other effective strategy of the p-learner in the plain model? In the plain model we have analyzed only for the case where the p-learner takes a simple averaging strategy (3). It is an interesting question when any other form of (4) works effectively.

2) Relating the analysis of DCB to rapidly mixing Markov chains. As seen in Section 3.3, the additional loss for DCB in the hierarchical model is related to the rate of convergence on MCMC that the DCB induces. On the other hand, in theoretical computer science, the theory of *rapidly mixing Markov chains* ([11]) has been explored to investigate the rate of convergence of general Markov chains. It would be interesting to develop another type of analysis of DCB on the basis of the theory of rapidly mixing Markov chains.

References

[1] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, 1985.

[2] B.S. Clarke and A.R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, IT-36, pp.453-471, 1990.

[3] A.E. Gelfand and A.F.M. Smith, "Sampling-based approach to calculating marginal densities," *J.Am. Statist. Assoc.*, vol.85, pp.398-409, 1990.

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayes restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-6, pp.721-741, 1984.

[5] W.K. Hastings, "Monte Carlo sampling method using Markov chains and their applications," *Biometrika*, vol.57, pp.97-109, 1970.

[6] D. Haussler, J. Kivinen, and M. Warmuth, "Tight worst-case loss bounds for predicting with expert advice," *Computational Learning Theory: Second European Conference, EuroCOLT'95*, Springer, pp.69-83, 1995.

[7] M. Kearns and H.S. Seung, "Learning from a population of hypotheses," *Proceedings of the Sixth Annual ACM Conference of Computational Learning Theory*, ACM Press, pp.101-110, 1993.

[8] A. Nakamura, N. Abe, and J. Takeuchi, "Efficient distribution-free population learning of simple concepts," *ALT'96 Algorithmic Learning Theory*, Springer, pp.500-515, 1994.

[9] E. Nummelin, *General irreducible Markov chains and non-negative operators*, Cambridge University Press, 1984.

[10] J. Rosenthal, "Rates of convergence for Gibbs sampling for variance component models," Technical report No.9322, Univ. of Toronto, Dept. of Statistics, 1993.

[11] A. Sinclair, *Algorithms for Random Generation & Counting: A Markov Chain Approach*, Birkhauser, 1993.

[12] M.A. Tanner and H.W. Wong, "The calculation of posterior distributions by data augmentation," *Jr. American Statist. Assoc.*, vol.82, pp.528-550, 1987.

[13] V.G. Vovk. "Aggregating strategies," *Proceedings of the Third Annual Workshop on Computational Learning Theory,* Morgan Kaufmann, pp.371-386, 1990.

[14] K. Yamanishi, "A decision-theoretic extension of stochastic complexity and its approximation to learning," submitted for publication, 1995.

[15] K. Yamanishi, "A randomized approximation of the MDL for stochastic models with hidden variables," *Proceedings of the Ninth Annual Conference on Computational Learning Theory,* ACM Press, pp.99-109, 1996.

APPENDIX

Proof of Lemma 15. Eq.(22) can be straightforwardly proven as follows.

$$\begin{aligned}
&d(\mathcal{S})\\
&= E_{D^{sm}}\left[\int |m_q(\mathbf{D}|D^{sm}) - m^*(\mathbf{D}|D^{sm})|d\mathbf{D}\right]\\
&= E_{D^{sm}}\left[\int \left|\int p(\mathbf{D}|\omega)(q(\omega|D^{sm}) - p^*(\omega|D^{sm}))d\omega\right| d\mathbf{D}\right]\\
&\le E_{D^{sm}}\left[\int \left(\int p(\mathbf{D}|\omega)\,|q(\omega|D^{sm}) - p^*(\omega|D^{sm})|\,d\omega\right) d\mathbf{D}\right]\\
&= E_{D^{sm}}\left[\int |q(\omega|D^{sm}) - p^*(\omega|D^{sm})|d\omega \int p(\mathbf{D}|\omega)d\mathbf{D}\right]\\
&= E_{D^{sm}}\left[d(q,p^*)\right].
\end{aligned}$$

Next, in order to prove (23), we start by observing that the average logarithmic loss for any distributed learning system in the hierarchical model is written as follows.

$$\begin{aligned}
&-\ln\int p(\mathbf{D}|\omega)q(\omega|D^{sm})d\omega\\
&= -\ln\int p(\mathbf{D}|\omega)p^*(\omega|D^{sm})d\omega
\end{aligned}$$

$$+\ln \frac{\int p(\mathbf{D}|\omega)p^*(\omega|D^{sm})d\omega}{\int p(\mathbf{D}|\omega)q(\omega|D^{sm})d\omega}. \tag{36}$$

We can further upper bound $\ln \frac{\int p(\mathbf{D}|\omega)p^*(\omega|D^{sm})d\omega}{\int p(\mathbf{D}|\omega)q(\omega|D^{sm})d\omega}$ as follows.

$$\begin{aligned}
&\ln \frac{\int p(\mathbf{D}|\omega)p^*(\omega|D^{sm})d\omega}{\int p(\mathbf{D}|\omega)q(\omega|D^{sm})d\omega} \\
&\leq \frac{\int p(\mathbf{D}|\omega)p^*(\omega|D^{sm})d\omega}{\int p(\mathbf{D}|\omega)q(\omega|D^{sm})d\omega} - 1 && (37)\\
&= \frac{\int p(\mathbf{D}|\omega)(p^*(\omega|D^{sm}) - q(\omega|D^{sm}))d\omega}{\int p(\mathbf{D}|\omega)q(\omega|D^{sm})d\omega} \\
&\leq \frac{(\sup_{\mathbf{D},\omega} p(\mathbf{D}|\omega)) \int |q(\omega|D^{sm}) - p^*(\omega|D^{sm})|d\omega}{\inf_{\mathbf{D},\omega} p(\mathbf{D}|\omega)} \\
&= (\overline{c}/\underline{c})^s d(q,p^*) && (38)\\
&= \kappa^s d(q,p^*), && (39)
\end{aligned}$$

where we set $\kappa = (\overline{c}/\underline{c})$. We have used the fact that $\ln x \leq x - 1$ $(\forall x > 0)$ to derive (37) and have used the facts that $\sup_{\mathbf{D},\omega} p(\mathbf{D}|\omega) = (\sup_{D,\theta} p(D|\theta))^s = \overline{c}^s$ and $\inf_{\mathbf{D},\omega} p(\mathbf{D}|\omega) = (\inf_{D,\theta} p(D|\theta))^s = \underline{c}^s$ (by Assumption 13) to derive (38). Combining (39) with (36) we see that for some $0 < C' < \infty$,

$$\begin{aligned}
&-\ln \int p(\mathbf{D}|\omega)q(\omega|D^{sm})d\omega \\
&\leq -\ln \int p(\mathbf{D}|\omega)p^*(\omega|D^{sm})d\omega + C'\kappa^s d(q,p^*).
\end{aligned}$$

Taking an expectation of the both sides of the above inequality with respect to the joint distribution of ω, D^{sm}, and $\mathbf{D}$ gives (23). This completes the proof of Lemma 15. □

Proof of Lemma 19. We start with the following claim.

Claim. *Let the initial value $\omega^{(0)}$ be in R_1 as in (31). Then at each iteration of the Markov chain (19) for the DCB, $|\overline{D} - (1/s)\sum_{i=1}^s \theta_i|$ gets multiplied by $1/(1 + (\sigma_\theta^2/\sigma^2)m)$ up to $1/\sqrt{s} + O(1/s)$ with at least uniform probability independent of s, m, and N.*

Proof of Claim. Let $\omega^{(l)} = (\theta_1^{(l)}, \cdots, \theta_s^{(l)}, \mu^{(l)})$ be the sampled value at the l-th iteration of the Markov chain for the DCB. Then we can immediately see that the mean of $(1/s)\sum_{i=1}^s \theta_i^{(l+1)}$ is $\frac{\sigma_\theta \overline{D} + \sigma^2 \mu^{(l)}}{\sigma_\theta^2 m + \sigma^2}$. Using Chebyshev's inequality we can prove that $|\overline{D} - (1/s)\sum_{i=1}^s \theta_i^{(l+1)}|$ is within $1/2\sqrt{s}$ of $\frac{1}{1+(\sigma_\theta^2/\sigma^2)m}|\mu^{(l)} - \overline{D}|$, with at least uniform probability independent of s, m, N and D^{sm}. Since the mean of $\mu^{(l)}$ is

$$\frac{\sigma_\theta^2 \mu_0 + \sigma_0^2 \sum_{i=1}^s \theta_i^{(l)}}{\sigma_\theta^2 + s\sigma_0^2} = \frac{1}{s}\sum_{i=1}^s \theta_i^{(l)} + O\left(\frac{1}{s}\right),$$

we can prove using Chebyshev's inequality again that $|\mu^{(l)} - \overline{D}|$ is within $1/2\sqrt{s} + O(1/s)$ of $|(1/s)\sum_{i=1}^s \theta_i^{(l)} - \overline{D}|$, with uniform probability. Thus we see that $|\overline{D} - (1/s)\sum_{i=1}^s \theta_i^{(l+1)}|$ is within $1/\sqrt{s} + O(1/sm)$ of $|\overline{D} - (1/s)\sum_{i=1}^s \theta_i^{(l)}|$ times $1/(1 + (\sigma_\theta^2/\sigma^2)m)$ with uniform probability, say ε_1. This implies that Claim holds.

From the above Claim it can be readily proven that for any $\omega \in R_1$, for $\ell_0 = O((\ln(sN))/(\ln m))$, with probability at least ε_1 independent of s, m, and N, we have $|\overline{D} - (1/s)\sum_{i=1}^s \theta_i^{(\ell_0)}| \leq (1/m)^{\ell_0} N^{\frac{1}{2}} + 1/\sqrt{s} + O(1/sm) \leq 2/\sqrt{s}$ for sufficiently large s. This implies that Eq.(28) holds for ε_1, R_1, R_2, and ℓ_0 as above.

Next let us define ε' by

$$\begin{aligned}
\varepsilon' \overset{\text{def}}{=} & \min_{\mu,x} \left\{ \left(\frac{4}{\sqrt{s}}\right) N\left(\frac{\sigma_\theta^2\mu_0 + \sigma_0^2 sx}{\sigma_\theta^2 + s\sigma_0^2}, \frac{\sigma_\theta^2\sigma_0^2}{\sigma_\theta^2 + s\sigma_0^2} : \mu\right)\right. \\
& \left. : |\mu - \overline{D}| \leq \frac{2}{\sqrt{s}},\ |x - \overline{D}| \leq \frac{2}{\sqrt{s}} \right\}.
\end{aligned}$$

Then we see that for the probability measure Q as in Lemma 19, for all $\omega \in R_2$, for all $\omega' \in \Omega$, we have $K(\omega, \omega') \geq \varepsilon' Q(\omega')$. Since ε' is bounded below by $\varepsilon_2 > 0$ independent of s, m, and N, we see that Eq.(29) holds for ε_2, R_2, and Q as above. This completes the proof of Lemma 19. □