Big Learning with Bayesian Methods

Jun Zhu, *Member, IEEE* , Jianfei Chen, Wenbo Hu

**Abstract**—Explosive growth in data and availability of cheap computing resources have sparked increasing interest in Big learning, an emerging subfield that studies scalable machine learning algorithms, systems, and applications with Big Data. Bayesian methods represent one important class of statistic methods for machine learning, with substantial recent developments on adaptive, flexible and scalable Bayesian learning. This article provides a survey of the recent advances in Big learning with Bayesian methods, termed Big Bayesian Learning, including nonparametric Bayesian methods for adaptively inferring model complexity, regularized Bayesian inference for improving the flexibility via posterior regularization, and scalable algorithms and systems based on stochastic subsampling and distributed computing for dealing with large-scale applications.

**Index Terms**—Big Bayesian Learning, Bayesian nonparametrics, Regularized Bayesian inference, Scalable algorithms

◆

1. **INTRODUCTION**

arXiv:1411.6370v1 [cs.LG] 24 Nov 2014

W

E live in an era of Big Data, where science, engi- neering and technology are producing massive

data streams, with petabyte and exabyte scales becom- ing increasingly common [45], [57], [154]. Besides the explosive growth in volume, Big Data also has high velocity, high variety, and high uncertainty. These complex data streams require ever-increasing process- ing speeds, economical storage, and timely response for decision making in highly uncertain environments. With the primary goal of building intelligent sys- tems that automatically improve from experiences, machine learning (ML) is becoming an increasingly important field to tackle the big data challenges [128], with an emerging field of *Big Learning*, which covers theories, algorithms and systems on addressing big

data problems.

# Big Learning Challenges

In big data era, machine learning needs to deal with the challenges of learning from complex situations with *large N, large P, large L*, and *large M* , where *N* is the data size, *P* is the feature dimension, *L* is the number of tasks, and *M* is the model size. Given that *N* is obvious, we explain the other factors below.

**Large** *P* : with the development of Internet, data sets with ultrahigh dimensionality have emerged, such as

the spam filtering data with trillion features [189] and the even higher-dimensional feature space via explicit kernel mapping [173]. Note that whether a learning problem is high-dimensional depends on the ratio between *P* and *N* . Many scientific problems with *P N* impose great challenges on learning, calling for effective regularization techniques to avoid overfitting and select salient features [64].

*J. Zhu, J. Chen and W. Hu are with TNList; State Key IST Lab; Dept. of Comp. Sci. & Tech.; Tsinghua University, Beijing, 100084 China. Email: {dcszj@, chenjian14@mails., hwb3@mails.}tsinghua.edu.cn*

*•*

**Large** *L*: many tasks involve classifying text or images into tens of thousands or millions of cate- gories. For example, the ImageNet [5] database con- sists of more than 14 millions of web images from

21 thousands of concepts, while with the goal of providing on average 1,000 images for each of 100+ thousands of concepts (or synsets) in WordNet; and the LSHTC text classification challenge 2014 aims to classify Wikipedia documents into one of 325,056 categories [2]. Often, these categories are organized in a graph, e.g., the tree structure in ImageNet and the DAG (directed acyclic graph) structure in LSHTC, which can be explored for better learning [28], [56].

**Large** *M* : with the availability of massive data, models with millions or billions of parameters are becoming common. Significant progress has been made on learning deep models, which have multi- ple layers of non-linearities allowing them to extract multi-grained representations of data, with successful applications in computer vision, speech recognition, and natural language processing. Such models include neural networks [84], auto-encoders [182], [108], and

probabilistic generative models [161], [155].

# Big Bayesian Learning

Though Bayesian methods have been widely used in ML, skepticism often arises when we talking about Bayesian methods for big data [94]. Practitioners also criticize that Bayesian methods are often too slow for even small-scaled problems, owning to many factors such as the non-conjugacy models with intractable in- tegrals. Nevertheless, Bayesian methods have several advantages on dealing with:

* + 1. **Uncertainty**: our world is an uncertain place be- cause of physical randomness, incomplete knowl- edge, ambiguities and contradictions. Bayesian methods provide a principled theory for combin- ing prior knowledge and uncertain evidence.
    2. **Flexibility**: Bayesian methods are intuitively sim- ple and flexible. Hierarchical Bayesian model-

and *p*(D) =

*p*0(**Θ**)*p*(D|**Θ**)d**Θ** is the marginal likeli-

ing offers a flexible tools for characterizing un-

where *p*0(·) is a prior distribution, chosen before see- ing any data;∫*p*(D|**Θ**) is the assumed likelihood model;

certainty, missing values, latent structures, and more.

* + 1. **Adaptivity**: The dynamics and uncertainty of Big Data require that our models should be adaptive

when the learning scenarios change. Nonpara- metric Bayesian methods provide elegant tools to deal with situations in which phenomena con- tinue to emerge as data are collected [85]. More-

hood (or evidence), often involving an intractable in- tegration problem that requires approximate inference as detailed below. The year 2013 marks the 250th an- niversary of Thomas Bayes’ essay on how humans can sequentially learn from experience, steadily updating their beliefs as more data become available [63].

A useful variational formulation of Bayes’ rule is

over, the Bayesian updating rule and its variants are sequential in nature and suitable for dealing

min

*q*(**Θ**)∈P

KL(*q*(**Θ**)ǁ*p*0(**Θ**)) − E*q*[log *p*(D|**Θ**)]*,* (2)

with big data streams.

* + 1. **Overfitting**: Although the data volume grows exponentially, the predictive information grows slower than the amount of Shannon informa- tion [30], while our models are becoming increas- ingly large by leveraging powerful computers, such as the deep networks with billions of param- eters. It implies that our models are increasing their capacity faster than the amount of infor- mation that we need to fill them with, therefore causing serious overfitting problems that call for effective regularization [170].

Therefore, Bayesian methods are becoming increas- ingly relevant in the big data era [190] to protect high capacity models against overfitting, and to allow models adaptively updating their capacity. However, the application of Bayesian methods to big data prob- lems runs into a computational bottleneck that needs to be addressed with new (approximate) inference methods. This article aims to provide a literature survey of the recent advances in big learning with Bayesian methods, including nonparametric Bayesian methods, regularized Bayesian inference, scalable in- ference algorithms and systems based on stochastic subsampling and distributed computing.

1. **BASICS OF BAYESIAN METHODS**

The general blueprint of Bayesian data analysis [67] is that a Bayesian model expresses a generative process of the data that includes hidden variables, under some statistical assumptions. The process specifies a joint probability distribution of the hidden and observed random variables. Given a set of observed data, data analysis is performed by *posterior inference*, which computes the conditional distribution of the hidden variables given the observed data.

# Bayes’ Theorem

At the core of Bayesian methods is Bayes’ theorem (a.k.a Bayes’ rule). Let **Θ** be the model parameters and be the given data set. The Bayesian posterior distribution is

D

*p*(**Θ** ) = *p*0(**Θ**)*p*(D|**Θ**) *,* (1)

|D

*p*(D)

where is the space of all distributions that make

the objective well-defined. It can be shown that the optimum solution to [(2)](#_bookmark0) is identical to the Bayesian posterior. In fact, if we add the constant term log *p*( ), the problem is equivalent to minimizing the KL- divergence between *q*(**Θ**) and the Bayesian posterior *p*(**Θ** ), which is non-negative and takes 0 if and only if *q* equals to *p*(**Θ** ). The variational interpretation is significant in two aspects: (1) it provides a basis for variational Bayes methods; and (2) it provides a starting point to make Bayesian methods more flexible by incorporating a rich set of posterior constraints. We will make these clear soon later.

P

D

|D

|D

It is noteworthy that *q*(**Θ**) represents the density of a general post-data posterior in the sense of [75, pp.15], not necessarily corresponding to a Bayesian posterior induced by Bayes’ rule. As we shall see in Section 3.2, when we introduce additional constraints, the post-data posterior *q*(**Θ**) is different from the Bayesian posterior *p*(**Θ** ), and moreover, it could even not be obtainable by the conventional Bayesian inference via Bayes’ rule. In the sequel, in order to distinguish *q*( ) from the Bayesian posterior, we will call it post-data posterior. The optimization formula- tion in [(2)](#_bookmark0) implies that Bayes’ rule is an information projection procedure that projects a prior density to a post-data posterior by taking account of the observed data. In general, Bayes’s rule is a special case of the principle of minimum information [193].

·

|D

# Bayesian Methods in Machine Learning

Bayesian statistics has been applied to almost ev- ery ML task, ranging from the single-variate regres- sion/classification to the structured output predic- tions and to the unsupervised/semi-supervised learn- ing scenarios [31]. In essence, however, there are several basic tasks that we briefly review below.

**Prediction**: After training, Bayesian models make predictions using the distribution:

*p*(**x**|D) = ∫ *p*(**x***,* **Θ**|D)d**Θ** = ∫ *p*(**x**|**Θ***,* D)*p*(**Θ**|D)d**Θ***,* (3)

where *p*(**x Θ***,* ) is often simplified as *p*(**x Θ**) due to the i.i.d assumption of the data when the model is

| D |

given. Since the integral is taken over the posterior distribution, the training data is considered.

**Model Selection**: Model selection is a fundamental problem in statistics and machine learning [96]. Let **M** be a family of models, where each model is in- dexed by a set of parameters **Θ**. Then, the marginal likelihood of the model family (or model evidence) is

∫

*p*(D|**M**) =

*p*(D|**Θ**)*p*(**Θ**|**M**)d**Θ***,* (4)

show that the negative of the objective in [(2)](#_bookmark0) is a lower bound of the evidence (i.e., log-likelihood):

log *p*(D) ≥ E*q*[log *p*(**Θ***,* D)] − E*q*[log *q*(**Θ**)]*.* (5) Then, variational Bayesian methods maximize the Evidence Lower BOund (ELBO):

max E*q*[log *p*(**Θ***,* )] E*q*[log *q*(**Θ**)]*,* (6)

D −

*q*∈P

whose solution is the target posterior if no assump- tions are made.

where *p*(**Θ M**) is often assumed to be uniform if no

|

strong prior exists.

For two different model families **M**1 and **M**2, the ratio of model evidences *κ* = *p*(D|**M**1 ) is called Bayes factor [98]. The advantage of using Bayes factors for model selection is that it automatically and naturally

D| 2

*p*( **M** )

includes a penalty for including too much model structure [31, Chap 3]. Thus, it guards against overfit- ting. For models where an explicit version of the likeli- hood is not available or too costly to evaluate, approx- imate Bayesian computation (ABC) can be used for model selection in a Bayesian framework [79], [180], while with the caveat that approximate-Bayesian es- timates of Bayes factors are often biased [157].

# Approximate Bayesian Inference

The computational difficulties in Bayesian inference arise from the intractability of high-dimensional inte- grals as involved in the posterior and in Eq.s [(3,](#_bookmark1) [4).](#_bookmark2) These are typically not only analytically intractable but also difficult to obtain numerically. Common prac- tice resorts to approximate methods, which can be grouped into two categories[1](#_bookmark3) — variational methods and Monte Carlo methods.

* + 1. *Variational Bayesian Methods*

Variational methods have a long history in physics, statistics, control theory and economics. In machine learning, variational formulations appear naturally in regularization theory, maximum entropy estimates, and approximate inference in graphical models. We refer the readers to the seminal book [183] and the nice short overview [95] for more details. A variational method basically consists of two parts:

* + - 1. cast the problems as some optimization problems;
      2. find an approximate solution when the exact solution is not feasible.

For Bayes’ rule, we have provided a variational for- mulation in [(2),](#_bookmark0) which is equivalent to minimizing the KL-divergence between the variational distribution

*q*(**Θ**) and the target posterior *p*(**Θ**|D). We can also

However, in many cases it is intractable to calculate

the target posterior. Therefore, to simplify the opti- mization, the variational distribution is often assumed to be in some parametric family, e.g., *q****φ***(**Θ**), and has some mean-field representation

Y

*q****φ***(**Θ**) = *qφi* (**Θ***i*)*,* (7)

*i*

where **Θ***i* represent a partition of **Θ**. Then, the problem transforms to find the best parameters ***φ***ˆ that

{ }

maximize the ELBO, which can be solved with nu- merical optimization methods. For example, with the factorization assumption, coordinate descent is often used to iteratively solve for *φi* until reaching some local optimum. Once a variational approximation *q*∗ is

found, the Bayesian integrals can be approximated by

replacing *p*(**Θ** ) by *q*∗. In many cases, the model **Θ**

|D

consists of parameters ***θ*** and hidden variables **h**. Then,

if we make the (structured) mean-field assumption that *q*(***θ****,* **h**) = *q*(***θ***)*q*(**h**), the variational problem can be solved by a variational Bayesian EM algorithm [24], which alternately updates *q*(**h**) at the variational Bayesian E-step and updates *q*(***θ***) at the variational Bayesian M-step.

* + 1. *Monte Carlo Methods*

Monte Carlo (MC) methods represent a diverse class of algorithms that rely on repeated random sampling to compute the solution to problems whose solution space is too large to explore systematically or whose systemic behavior is too complex to model. The ba- sic idea of MC methods is to draw a set of i.i.d samples {**Θ***i*}*N* from a target distribution *p*(**Θ**) and

*i*=1

use the empirical distribution · 1 Σ*N* ·

*p*ˆ( ) = *δ***Θ** ( )*,*

*N*

*i*=1

*i*

to approximate the target distribution, where *δ***Θ***i* ( ) is the delta-Dirac mass located at **Θ***i*. Consider the common operation on calculating the expectation of some function *φ* with respect to a given distribution. Let *p*(**Θ**) = *p*¯(**Θ**)*/Z* be the density of a probability distribution that we compute *p*¯(**Θ**) pointwise up to a normalizing constant *Z*. The expectation of interest is

·

∫

*I* =

*φ*(**Θ**)*p*(**Θ**)d**Θ***.* (8)

1. Both maximum likelihood estimation (MLE),

**Θ**ˆ MLE =

Replacing *p*(·) by

*p*ˆ(·), we get the unbiased Monte

argmax**Θ** *p*(*D|***Θ**), and maximum a posterior estimation (MAP),

**Θ**ˆ MAP = argmax**Θ** *p*0(**Θ**)*p*(*D|***Θ**), can be seen as the third type of

Carlo estimate of this quantity:

*N*

Σ

approximation methods to do Bayesian inference. We omit them since they examine only a single point, and so can neglect the potentially large distributions in the integrals.

*I*ˆMC =

1

*N*

*i*=1

*φ*(**Θ***i*)*.* (9)

Asymptotically, when *N* → ∞ the estimate

*I*ˆMC

with multiple modes. Another useful technique to

will almost surely converge to *I* by the strong law of large numbers. In practice, however, we often cannot sample from *p* directly. Many methods have been developed, such as rejection sampling and importance sampling, which however often suffer from severe limitations in high dimensional spaces. We refer the readers to the book [156] and the review article [15] for details. Below, we introduce Markov chain Monte Carlo (MCMC), a very general and powerful frame- work that allows sampling from a broad family of dis- tributions and scales well with the dimensionality of the sample space. More importantly, many advances have been made on scalable MCMC methods for Big Data, which will be discussed later.

An MCMC method constructs an ergodic *p*- stationary Markov chain sequentially. Once the chain has converged (i.e., finishing the burn-in phase), we can use the samples to estimate *I*. The Metropolis- Hastings algorithm [125], [83] constructs such a chain by using the following rule to transit from the current state **Θ***t* to the next state **Θ***t*+1:

1. draw a candidate state **Θ**j from a proposal distri-

develop simpler or more efficient MCMC methods is data augmentation [174], [62], [135], which intro- duces auxiliary variables to transform marginal de- pendency into a set of conditional independencies. For Gibbs samplers, the useful techniques to improve the convergence include blockwise Gibbs sampling and partially collapsed Gibbs (PCG) sampling [181]. A PCG sampler is as simple as an ordinary Gibbs sam- pler, but often improves the convergence by replacing some of the conditional distributions of an ordinary Gibbs sampler with conditional distributions of some marginal distributions.

# FAQ

Common questions regarding Bayesian methods are:

# Q: Why should I use Bayesian methods?

**A:** There are many reasons for choosing Bayesian methods, as discussed in the Introduction. A formal theoretical argument is provided by the classic de Finitti theorem, which states that: If (**x**1*,* **x**2*, . . .* ) are infinitely exchangeable, then for any *N*

bution *q*(**Θ Θ***t*);

|

1. compute the acceptance probability:

*p*(**x**1*, . . . ,* **x***N* ) =

∫ .Y*N*

*p*(**x***i*|***θ***)Σ

d*P* (***θ***) (11)

*A*(**Θ**j*,* **Θ***t*) ¾ min .

*p*¯(**Θ**j)*q*(**Θ***t*|**Θ**j) *p*¯(**Θ** )*q*(**Θ**j|**Θ** )

*.* (10)

*i*=1

for some random variable ***θ*** and probability measure

*t t*

Σ1*,*

*P* . The infinite exchangeability is an often satisfied

1. draw *γ* ∼ Uniform[0*,* 1]. If *γ < A*(**Θ**j*,* **Θ***t*) set

j

**Θ***t*+1 ← **Θ** , otherwise set **Θ***t*+1 ← **Θ***t*.

Note that for Bayesian models, each MCMC step involves an evaluation of the full likelihood to get the (unnormalized) posterior *p*¯(**Θ**), which can be pro- hibitive for big learning with massive data sets. We will revisit this problem later.

One special type of MCMC methods is the Gibbs sampling [69], which iteratively draws samples from local conditionals. Let **Θ** be a *M* -dimensional vector. The standard Gibbs sampler performs the following steps to get a new sample **Θ**(*t*+1):

1. draw a sample *θ*(*t*+1) ∼ *p*(*θ*1|*θ*(*t*)*,* · · · *, θ*(*t*));

property. For example, any i.i.d data are infinitely exchangeable. Moreover, the data whose ordering information is not informative is also infinitely ex- changeable, e.g., the commonly used bag-of-words representation of documents [36] and images [112].

# Q: How should I choose the prior?

**A:** There are two schools of thought, namely, objec- tive Bayes and subjective Bayes. For objective Bayes, an improper noninformative prior (e.g., the Jeffreys prior [91] and the maximum-entropy prior [90]) is used to capture ignorance, which admits good fre- quentist properties. In contrast, subjective Bayesian methods embrace the influence of priors. A prior may

1. for

1 2 *M*

have some parameters *λ*. Since it is often difficult to

*j* = 2 : *M* − 1, draw a sample

*θ*(*t*+1) ∼ *p*(*θj* |*θ*(*t*+1)*,* · · · *, θ*(*t*+1)*, θt*

· · · *, θ* );

elicit an honest prior, e.g., setting the true value of *λ*, two practical methods are often used. One is hierar-

*j* 1 *j*−1 *j*+1 *M*

*t*

3) draw a sample *θ*(*t*+1) ∼ *p*(*θM* |*θ*(*t*+1)*,* · · · *, θ*(*t*+1)); One issue with MCMC methods is that the con-

*M*

1

*M* −1

vergence rate can be prohibitively slow even for conventional applications. Extensive efforts have been spent to improve the convergence rates. For exam- ple, hybrid Monte Carlo methods explore gradient information to improve the mixing rates when the model parameters are continuous, with representa- tive examples of Langevin dynamics and Hamilto- nian dynamics [136]. Other improvements include population-based MCMC methods [89] and annealing methods [73] that can sometimes handle distributions

chical Bayesian methods, which assume a hyper-prior on *λ* and define the prior as a marginal distribution: *p*0(**Θ**) = ∫ *p*0(**Θ**|*λ*)*p*(*λ*)d*λ.* (12)

Though *p*(*λ*) may have hyper-parameters as well, it is commonly believed that these parameters will have a weak influence as long as they are far from the like- lihood model, thus can be fixed at some convenient values or put another layer of hyper-prior.

Another method is *empirical Bayes*, which adopts a data-driven estimate *λ*ˆ and uses *p*0(**Θ** *λ*ˆ) as the prior. Empirical Bayes can be seen as an approximation

|

**Example 2: Logistic-Normal Prior** A logistic- normal distribution [11] provides one way to impose correlation structure among the multiple dimensions of ***θ***. It is defined as follows:



*α*

*θi Zij*

*Wij*

*N Li*

*φk*

*K*

*β*



*μ,*Σ *ηi*

*Zij*

*Wij*

*N Li*

*φk*

*K*

*β*



*α*

*θi Zij*

*Wij Yi*

*N Li*

*φk η*

*K*

*β γ*

*eηk*

(13)

***η*** ∼ N (***µ****,* Σ)*, θk* = Σ

*j*

*eηj .*

(a)

(b)

(c)

This prior has been used to develop correlated topic models (or logistic-normal topic models) [34], which can infer the correlation structure among topics. How- ever, the flexibility pays cost on computation, needing

Fig. 1. Graphical illustration of (a) LDA [36]; (b) logistic- normal topic model [34]; and (c) supervised LDA.

to the hierarchical approach, where *p*(*λ*) is approx- imated by a delta-Dirac mass *δλ*ˆ(*λ*). One common choice is maximum marginal likelihood estimate, that

is, *λ*ˆ = argmax*λ p*( *λ*). Empirical Bayes has been

D|

applied in many problems, including variable sec-

tion [70] and nonparametric Bayesian methods [123]. Recent progress has been made on characterizing the conditions when empirical Bayes merges with the Bayesian inference [148] as well as the convergence rates of empirical Bayes methods [58].

In practice, another important consideration is the tradeoff between model capacity and computational cost. If a prior is conjugate to the likelihood, the pos- terior inference will be relatively simpler in terms of computation and memory demands, as the posterior belongs to the same family as the prior.

# Example 1: Dirichlet-Multinomial Conjugate Pair

Let **x** 0*,* 1 *V* be a one-hot representation of a discrete variable with *V* possible values. It is easy to verify that for the multinomial likelihood,

∈ { }

*p*(**x**|***θ***) =

Q*V θxk* , the conjugate prior is a Dirichlet distribu-

tion, *p*0(***θ***|***α***) = Dir(***α***) =

*Z*

*k*=1 *θk k*

, where ***α*** is

scalable algorithms to learn large topic graphs [48].

1. **BIG BAYESIAN LEARNING**

Though much more emphasis in big Bayesian learning has been put on scalable algorithms and systems, substantial advances have been made on *adaptive* and *flexible* Bayesian methods. This section reviews nonparametric Bayesian methods for adaptively in- ferring model complexity and regularized Bayesian inference for improving the flexibility via posterior regularization, while leaving the large part of scalable algorithms and systems to next sections.

# Nonparametric Bayesian Methods

For parametric Bayesian models, the parameter space is pre-specified. No matter how the data changes, the number of parameters is fixed. This restriction may cause limitations on model capacity, especially for big data applications, where it may be difficult or even counter-productive to fix the number of parameters a priori. For example, a Gaussian mixture model with

well; however, it may be sub-optimal to use the same

*k*=1 *k*

1 Q*V*

*α* −1

a fixed number of clusters may fit the given data set

the hyper-parameter and *Z* is the normalization factor.

In fact, the posterior distribution is Dir(***α*** + **x**).

A popular Bayesian model that explores such con-

jugacy is latent Dirichlet allocation (LDA) [36], as illustrated in Fig. [1(a).](#_bookmark4) LDA posits that each document **w***i* is an admixture of a set of *K* topics, of which each topic ***ψ****k* is a unigram distribution over a given vocabulary. The generative process is as follows:

1. draw *K* topics ***ψ****k* ∼ Dir(***β***)
2. for each document *i* ∈ [*N* ]:
   1. draw a topic mixing vector ***θ****i* ∼ Dir(***α***)
   2. for each word *j* ∈ [*Li*] in document *i*:
      1. draw a topic assignment *zij* ∼ Multi(***θ****i*)
      2. draw a word *wij* ∼ Multi(***ψ****zij* ).

LDA has been popular in many applications. How- ever, a conjugate prior can be restrictive. For example, the Dirichlet distribution does not impose correlation between different parameters, except the normaliza- tion constraint. In order to obtain more flexible mod- els, a non-conjugate prior can be chosen.

number of clusters if more data comes under a slightly

changed distribution. It would be ideal if the cluster-

ing models can figure out the unknown number of clusters automatically. Similar requirements on auto- matical model selection exist in feature representation learning [29] or factor analysis, where we would like the models to automatically figure out the dimension of latent features (or factors) and maybe also the topological structure among features (or factors) at different abstraction levels [7].

Nonparametric Bayesian (NPB) methods provide an elegant solution to such needs on automatic adapta- tion of model capacity when learning a single model. Such adaptivity is obtained by defining stochastic processes on rich measure spaces. Classical examples include Dirichlet process (DP), Indian buffet process (IBP), and Gaussian process (GP). Below, we briefly review DP and IBP. We refer the readers to the ar- ticles [74], [72], [132] for a nice overview and the textbook [85] for a comprehensive treatment.