Online Bayesian Passive-Aggressive Learning*

Tianlin Shi
Institute for Interdisciplinary Information Sciences
Tsinghua University
Beijing, 100084 China

Jun Zhu
State Key Lab of Intelligent Technology and Systems
Tsinghua National Lab for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University
Beijing, 100084 China

Abstract

We present online Bayesian Passive-Aggressive (BayesPA) learning, a generic online learning framework for hierarchical Bayesian models with max-margin posterior regularization. We show that BayesPA subsumes the standard online Passive-Aggressive (PA) learning and extends naturally to incorporate latent variables for both parametric and nonparametric Bayesian inference, therefore providing great flexibility for explorative analysis. As an important example, we apply BayesPA to topic modeling and derive efficient online learning algorithms for max-margin topic models. We further develop nonparametric BayesPA topic models to infer the unknown number of topics in an online manner. Experimental results on 20newsgroups and a large Wikipedia multi-label dataset (with 1.1 millions of training documents and 0.9 million of unique terms in the vocabulary) show that our approaches significantly improve time efficiency while achieving comparable accuracy with the corresponding batch algorithms.

1. Introduction

With the fast growth of data volume and variety, it is becoming increasingly important to develop scalable machine learning algorithms, which can be categorized into two major categories — online/stochastic methods on a single core and parallel/distributed algorithms on multiple-cores or multiple machines. This paper focuses on online learning, a process of answering a sequence of questions (e.g., which category does a document belong to?) given knowledge of the correct answers (e.g., the true category labels) to previous questions. Such a process is especially suitable for the applications with streaming data. For the applications with large datasets, online learning algorithms can effectively explore data redundancy relative to the model to be learned, by repeatedly subsampling the data; and they often lead to faster convergence to satisfactory results than the corresponding batch learning algorithms. Among the many popular algorithms, online Passive-Aggressive (PA) learning (Crammer et al. 2006) provides a generic framework of performing online

* Correspondence should be addressed to J. Zhu. T. Shi is now with Stanford Artificial Intelligence Laboratory (SAIL), Stanford University, Stanford, CA, 94305

©2017 Tianlin Shi and Jun Zhu.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/ Attribution requirements are provided at http://jmlr.org/papers/v18/14-188.html
learning for large-margin methods (e.g., SVMs), with many applications in natural language processing and text mining (McDonald et al. 2005; Chiang et al., 2008). Though enjoying strong discriminative ability that is preferable for predictive tasks, existing online PA methods are formulated as a point estimate problem by optimizing some deterministic objective function. This may lead to some potential shortcomings. For example, a single large-margin model could fail to capture the rich underlying structure of complex data.

On the other hand, Bayesian methods enjoy greater flexibility in describing the possible underlying structures of complex data by incorporating a hierarchy of latent variables. Moreover, the recent progress on nonparametric Bayesian methods (Hjort 2010; Teh et al., 2006a) further provides an increasingly important framework that allows the Bayesian models to have an unbounded model complexity, e.g., an infinite number of components in a mixture model (Hjort, 2010) or an infinite number of units in a latent feature model (Ghahramani and Griffiths, 2005), and to adapt when the learning environment changes. In particular, adaptation to the changing environment is of great importance in online learning. For Bayesian models, one challenging problem is posterior inference, for which both variational and Monte Carlo methods can be too expensive to be applied to large-scale applications. To scale up Bayesian inference, much progress has been made on developing stochastic variational Bayes (Hoffman et al., 2013; Mimno et al., 2012) and stochastic gradient Monte Carlo (Welling and Teh, 2011; Ahn et al., 2012) methods, which repeatedly draw samples from a given finite dataset. To deal with the potentially unbounded streaming data, streaming variational Bayes methods (Broderick et al., 2013) have been developed as a general framework, with an application to topic models for learning latent topic representations. However, due to the generative nature, Bayesian models usually do not perform as well as discriminative models in predictive tasks such as classification.

Successful attempts have been made to bring large-margin learning and Bayesian methods together to solve supervised learning tasks using flexible Bayesian latent variable models. For example, maximum entropy discrimination (MED) made a significant advance in conjoining max-margin learning and Bayesian generative models, in the context of univariate prediction (Jaakkola et al., 1999) and structured output prediction (Zhu and Xing, 2009). Recently, much attention has been devoted to generalizing MED to incorporate latent variables and perform nonparametric Bayesian inference in various contexts, including topic modeling (Zhu et al., 2012), matrix factorization (Xu et al., 2013), and multi-task learning (Jebara, 2011; Zhu et al., 2014c). Regularized Bayesian inference (RegBayes) (Zhu et al. 2014c) provides a unified framework for Bayesian models to perform max-margin learning, where the max-margin principle is incorporated through adding posterior constraints to an information-theoretical optimization problem. By solving this problem involving latent variables, RegBayes can uncover structures that support the supervised learning tasks. RegBayes subsumes the standard Bayes’ rule and is more flexible in incorporating domain knowledge or max-margin constraints. Though flexible in discovering latent structures and powerful in discriminative predictions, posterior inference in such models remains a challenge. By exploring data augmentation techniques, recent progress has been made to develop efficient MCMC methods (Zhu et al., 2014b), which can also be implemented in distributed clusters (Zhu et al., 2013). However, these batch learning methods are not applicable to streaming data, and they do not explore the statistical redundancy in large-scale corpora either.

1. There are also many distributed methods for scalable Bayesian inference. Please see (Zhu et al., 2014a) for a review.
To address the above problems of both online PA on incorporating flexible latent structures and Bayesian max-margin models on scalable streaming inference, this paper presents online Bayesian Passive-Aggressive (BayesPA) learning, a general framework of performing online learning for Bayesian max-margin models. We show that online BayesPA subsumes the standard online PA when the underlying model is linear and the parameter prior is Gaussian (See Table 1 for its close relationships with streaming variational Bayes and RegBayes). We further show that one major significance of BayesPA is its natural generalization to incorporate a hierarchy of latent variables for both parametric and nonparametric Bayesian inference, therefore allowing online BayesPA to have the great flexibility of (nonparametric) Bayesian methods for explorative analysis as well as the strong discriminative ability of large-margin learning for predictive tasks. As concrete examples, we apply the theory of online BayesPA to topic modeling and derive efficient online learning algorithms for max-margin supervised topic models (Zhu et al., 2012). We further develop efficient online learning algorithms for the nonparametric max-margin topic models, an extension of the nonparametric topic models (Teh et al., 2006a) for predictive tasks. Extensive empirical results on real datasets demonstrate significant improvements on time efficiency and maintenance of comparable results with corresponding batch algorithms.

The paper is structured as follows. We discuss the related work in Section 2, and review the preliminary knowledge in Section 3. Then, we move on to the detailed description of BayesPA in Section 4 and present the theoretical analysis. Section 5 presents the concrete instantiations on topic modeling, and Section 6 presents the extensions to nonparametric topic models and multi-task learning. Section 7 presents experimental results. Finally, Section 8 concludes this paper with future directions discussed.

2. Related Work

As a well-established learning paradigm, online learning is of both theoretical and practical interest. The goal of online learning is to make a sequence of decisions, such as classification and regression, and use the knowledge extracted from previous correct answers to produce decisions on incoming ones. The root of online learning could be traced back to early studies of repeated games (Hannan, 1957), where an agent dynamically makes choices with the summary of past information. The idea became popular with the advent of Perceptron algorithms (Rosenblatt, 1958), which adopt an additive update rule for the classifier weights, and its multiplicative counterpart is the Winnow algorithm (Littlestone 1988). The class of online multiplicative algorithms was further generalized by Adaboost (Freund and Schapire, 1997) in a decision theoretic sense and is now widely applied to various fields of study (Arora et al., 2012).

First presented by Crammer et al. (2006) as a weight updating method, online Passive-Aggressive (PA) algorithms provide a generic online learning framework for max-margin models. In particular, they considered loss functions that enforce max-margin constraints, and showed that surprisingly simple update rules could be derived in closed forms. Motivated by online PA learning and to handle unbalanced training sets, Dredze et al. (2008) proposed confidence-weighted learning, which adopts an additive update rule for the classifier weights, and its multiplicative counterpart is the Winnow algorithm (Littlestone 1988). The class of online multiplicative algorithms was further generalized by Adaboost (Freund and Schapire, 1997) in a decision theoretic sense and is now widely applied to various fields of study (Arora et al., 2012).

First presented by Crammer et al. (2006) as a weight updating method, online Passive-Aggressive (PA) algorithms provide a generic online learning framework for max-margin models. In particular, they considered loss functions that enforce max-margin constraints, and showed that surprisingly simple update rules could be derived in closed forms. Motivated by online PA learning and to handle unbalanced training sets, Dredze et al. (2008) proposed confidence-weighted learning, which maintains a Gaussian distribution of the classifier weights at each round and replaces the max-margin constraint in PA with a probabilistic constraint enforcing confidence of classification. Within the same framework, Crammer et al. (2008) derived a new convex form of the constraint and demonstrated performance improvements through empirical evaluations.
The theoretical analysis of online learning typically relies on the notion of regret, which is the average loss incurred by an adaptive online learner on streaming data versus the best achievable through a single fixed model having the hindsight of all data (Murphy, 2012). It can be shown that the notion of regret is closely related to weak duality in convex optimization, which brings online learning to the algorithmic framework of convex repeated games (Shalev-Shwartz and Singer, 2006).

Although the classical regime of online learning is based on decision theory, recently much attention has been paid to the theory and practice of online probabilistic inference in the context of Big Data. Rooted either in variational inference or Monte Carlo sampling methods, there are broadly two lines of work on the topic of online Bayesian inference. Stochastic variational inference (SVI) (Hoffman et al., 2013) is a stochastic approximation algorithm for mean-field variational inference. By approximating the natural gradients in maximizing the evidence lower bound with stochastic gradients sampled from data points, Hoffman et al. (2013) demonstrated scalable inference of topic models on large corpora. Mimno et al. (2012) showed the performance of SVI could be improved through structured mean-field assumptions and locally collapsed variational inference. SVI is also applicable to the stochastic inference of nonparametric Bayesian models, such as hierarchical Dirichlet process (Wang et al., 2011; Wang and Blei, 2012).

There is also a large body of work on extending Monte Carlo methods to the online setting. A classic approach is sequential Monte Carlo methods (SMC) or particle filters (Doucet and Johansen, 2009), which arose from the numerical estimation of state-space models. Through Rao-Blackwellized particle filters (Doucet et al., 2000), one could obtain online inference algorithms for latent Dirichlet allocation (LDA) (Canini et al., 2009), and more recently Gao et al. (2016) presented a streaming (collapsed) Gibbs sampler with better accuracy. To tackle the sparsity issues and inadequate coverage of particles, Steinhardt and Liang (2014) leveraged “abstract particles” to represent entire regions of the sample space. Recently, Korattikara et al. (2014) introduced an approximate Metropolis-Hasting rule based on sequential hypothesis testing that allows accepting and rejecting samples using only a fraction of the data. As an alternative, Bardenet et al. (2014) proposed an adaptive sampling strategy of Metropolis-Hastings from a controlled perturbation of the target distribution. With elegant use of gradient information that Metropolis-Hastings algorithms neglected, a line of work (Welling and Teh 2011; Ahn et al., 2012; Patterson and Teh, 2013) also developed stochastic gradient methods based on Langevin dynamics.

While most online Bayesian inference methods have adopted a stochastic approximation of the posterior distribution by sub-sampling a given finite dataset, in many applications data arrives in stream so that the dataset is changing over time and its size is unknown. To relax the previous request on knowing the data size, Broderick et al. (2013) made streaming updates to the estimated posterior and demonstrated the advantage of streaming variational Bayes (SVB) over stochastic variational inference. As will be discussed in this paper, BayesPA also does not impose assumptions on the size of dataset and works on streaming data.

The idea to discriminatively train univariate or structured output classifiers was popularized by the seminal work on support vector machines (Vapnik, 1995) and max-margin Markov networks (aka structural SVMs) (Taskar et al., 2003). Since then, research on developing max-margin models with latent variables has received increasing attention, because of the promise to capture underlying complex structures of the problems. A promising line of work focused on Bayesian approaches, and one representative is maximum entropy discrimination (MED) (Jaakkola et al., 1999; Jebara 2001; Zhu and Xing, 2009), which learns distributions of model parameters discriminatively from labeled
Online Bayesian Passive-Aggressive Learning

3. Preliminaries

This section reviews the preliminary knowledge that is needed to develop online Bayesian Passive-Aggressive learning. The relationships with BayesPA will be summarized in Table 1 later.

3.1 Online Passive-Aggressive Learning

Based on a decision-theoretic view, the goal of online supervised learning is to minimize the cumulative loss for a certain prediction task from the sequentially arriving training samples. Online Passive-Aggressive (PA) learning (Crammer et al. 2006) achieves this goal by updating some parametric model \( w \in \mathbb{R}^K \) (e.g., the weights of a linear SVM) in an online manner with the instantaneous losses from arriving data \( \{x_t\}_{t \geq 0} (x_t \in \mathbb{R}^K) \) and corresponding responses \( \{y_t\}_{t \geq 0} \), where \( K \) is the number of input features. The loss \( \ell(y_t; x_t, w, y_t) \), as they consider, could be the hinge loss \((\epsilon - y_t w^T x_t)_+\) for binary classification \((y_t \in \{0, 1\})\) or the \( \epsilon \)-insensitive loss \((|y_t - w^T x_t| - \epsilon)_+\) for regression \((y_t \in \mathbb{R})\), where \( \epsilon \) is a hyper-parameter and \((x)_+ = \max(0, x)\). The online Passive-Aggressive update rule is then derived by defining the new weight \( w_{t+1} \) as the solution to the
follows optimization problem:

\[
\min_w \frac{1}{2} ||w - w_t||^2 \quad \text{s.t.: } \ell_{\epsilon}(w; x_t, y_t) = 0,
\]

(1)

where \( || \cdot || \) is the Euclidean 2-norm. Intuitively, if \( w_t \) suffers no loss from the new data, i.e., \( \ell_{\epsilon}(w_t; x_t, y_t) = 0 \), the algorithm passively assigns \( w_{t+1} = w_t \); otherwise, it aggressively projects \( w_t \) to the feasible region of parameter vectors that attain zero loss on the new data. With provable regret bounds, Crammer et al. (2006) showed that online PA algorithms could achieve comparable results to the optimal classifier \( w^* \), which has the hindsight of all data. In practice, in order to account for inseparable training samples, soft margin constraints are often adopted and the resulting PA learning problem is

\[
\min_w \frac{1}{2} ||w - w_t||^2 + 2c \cdot \ell_{\epsilon}(w; x_t, y_t),
\]

(2)

where \( c \) is a positive regularization parameter and the constant factor 2 is included for simplicity as will be clear soon. For problems (1) and (2) with samples arriving one at a time, closed-form solutions can be derived (Crammer et al., 2006). For example, for the binary hinge loss the update rule for problem (2) is

\[
w_{t+1} = w_t + \tau_t y_t x_t,
\]

where \( \tau_t = \min(2c, \max(0, \epsilon - y_t w^{\top} x_t)/||x_t||^2) \); and for the \( \epsilon \)-insensitive loss, the update rule is

\[
w_{t+1} = w_t + \text{sign}(y_t - w^{\top} x_t) \tau_t x_t,
\]

where \( \tau_t = \max(0, \epsilon - y_t w^{\top} x_t)/||x_t||^2 \).

### 3.2 Streaming Variational Bayes

For Bayesian models, Bayes’ rule naturally leads to a streaming update procedure for online learning. Specifically, suppose the data \( \{x_t\}_{t \geq 0} \) are generated i.i.d. according to a distribution \( p(x|w) \) and the prior \( p(w) \) is given. Bayes’ theorem implies that the posterior distribution of \( w \) given the first \( T \) samples (\( T \geq 1 \)) satisfies

\[
p(w|\{x_t\}_{t=0}^T) \propto p(w|\{x_t\}_{t=0}^{T-1})p(x_T|w).
\]

In other words, the posterior after observing the first \( T - 1 \) samples is treated as the prior for the incoming data point.

This natural streaming Bayes’ rule, however, is often intractable to compute, especially for complex models (e.g., when latent variables are present). Streaming variational Bayes (SVB) (Broderick et al., 2013) suggests that a variational approximation should be adopted and it practically works well. Specifically, let \( \mathcal{A} \) be any algorithm that calculates an approximate posterior \( q(w) = \mathcal{A}(X, q_0(w)) \) based on data \( X \) and prior \( q_0(w) \). Then, the recursive formula for approximate streaming update is:

\[
q(w|\{x_t\}_{t=0}^T) = \mathcal{A}\left(x_T, q(w|\{x_t\}_{t=0}^{T-1})\right).
\]

The choice of \( \mathcal{A} \) can be problem-specific. For topic modeling, Broderick et al. (2013) showed that one may adopt mean-field variational Bayes (Wainwright and Jordan, 2008), expectation propagation (Minka, 2001), and one-pass posterior approximation algorithms using stochastic variational inference (Hoffman et al. 2013) or sufficient statistics update (Honkela and Valpola, 2003; Luts et al. 2013). By applying the streaming update in a distributed setting asynchronously, SVB could also be scaled up across multiple computer clusters (Broderick et al. 2013).
3.3 Regularized Bayesian Inference

The decision-theoretic and Bayesian view of learning can be reciprocal. For example, it would be beneficial to combine the flexibility of Bayesian models with the discriminative power of large-margin methods. The idea of regularized Bayesian inference (RegBayes) (Zhu et al., 2014c) is to treat Bayesian inference as an optimization problem with an imposed loss function for prediction tasks or a regularization term on the target posterior distribution in general. Mathematically, RegBayes for supervised learning can be formulated as

$$\min_{q \in \mathcal{P}} \text{KL}(q(w)||p_0(w)) - \mathbb{E}_{q(w)} \left[ \log p(X|w) \right] + 2c \cdot \ell(q(w); X, Y),$$

(4)

where $\mathcal{P}$ is the probability simplex, $p(X|w)$ is the likelihood function and KL is the Kullback-Leibler divergence.\(^2\) Note that if the loss is zero, i.e., $\ell(q(w); X, Y) = 0$, then the optimal solution is $q^∗(w) \propto p_0(w)p(X|w)$, which is just the Bayes’ rule. However, when $\ell(q(w); X, Y) \neq 0$, RegBayes biases the inferred posterior towards discriminating the supervising side-information $Y$, with the parameter $c$ controlling the extent of regularization. If the posterior regularization term $\ell(\cdot)$ is a convex function of $q(w)$ through the linear expectation operator, Zhu et al. (2014c) presented a general representation theorem to characterize the solution to problem (4). To distinguish from the posterior distribution obtained via Bayes’ rule, the solution to problem (4) is called post-data posterior (Zhu et al., 2014c). Many instantiations have been developed following the generic framework of RegBayes to demonstrate its superior performance than standard Bayesian models in various settings, such as topic modeling (Jiang et al. 2012; Zhu et al. 2014b), matrix factorization (Xu et al., 2013) and link prediction (Zhu. 2012), as well as the flexibility of performing robust Bayesian inference with noisy domain knowledge (Mei et al., 2014).

4. Bayesian Passive-Aggressive Learning

In this section, we present online Bayesian Passive-Aggressive (BayesPA) learning, a general perspective on online max-margin Bayesian inference. Without loss of generality, we consider binary classification. The techniques can be applied to other learning tasks. We provide an extension in Section 6.2

4.1 Online BayesPA Learning

Instead of updating a point estimate of $w$, online Bayesian PA (BayesPA) sequentially infers a new post-data posterior distribution $q_{t+1}(w)$, either parametric or nonparametric, on the arrival of a new data point $(x_t, y_t)$ by solving the following optimization problem:

$$\min_{q(w) \in \mathcal{F}_t} \text{KL}(q(w)||q_t(w)) - \mathbb{E}_{q(w)} \left[ \log p(x_t|w) \right],$$

s.t.: $\ell_t(q(w); x_t, y_t) = 0,$

(5)

---

2. We assume that the model space $\mathcal{W}$ is a complete separable metric space endowed with its Borel $\sigma$-algebra $\mathcal{B}(\mathcal{W})$. Let $P_0$ and $Q$ be probability measures on $\mathcal{W}$. The Kullback-Leibler (KL) divergence of the probability measure $Q$ with respect to the measure $P_0$ is defined as $\text{KL}(Q||P_0) = \int \frac{dQ}{dP_0}(w) \log \frac{dQ}{dP_0}(w) dP_0(w)$, where $\frac{dQ}{dP_0}(w)$ is the Radon-Nikodym derivative (Durrett 2010). It is required that $Q$ is absolutely continuous with respect to $P_0$ such that this derivative exists. In this paper, we further assume that $P_0$ is absolutely continuous with respect to some background measure $\mu$. Thus, there exists a density $p_0$ that satisfies $dP_0 = p_0 d\mu$ and there also exists a density $q$ that satisfies $dQ = q d\mu$. Then, the KL-divergence can be expressed as $\text{KL}(q||p_0) = \int q(w) \log \frac{q(w)}{p_0(w)} d\mu(w)$. 

where $\mathcal{F}_t$ can be some family of distributions or the whole probability simplex $\mathcal{P}$. For notational convenience, we denote the objective as $\mathcal{L}(q(\mathbf{w}))$. In words, we find a post-data posterior distribution $q_{t+1}(\mathbf{w})$ in the feasible region that is not only close to the current posterior $q_t(\mathbf{w})$ in terms of KL-divergence, but also has a high likelihood of explaining the new data. As a result, if Bayes’ rule already gives the posterior distribution $q_{t+1}(\mathbf{w}) \propto q_t(\mathbf{w})p(\mathbf{x}_t|\mathbf{w})$ that suffers no loss (i.e., $\ell_\epsilon = 0$), BayesPA passively updates the posterior following just Bayes’ rule; otherwise, BayesPA aggressively projects the new posterior to the feasible region of posteriors that attain zero loss. The passive and aggressive update cases are illustrated in Figure 1. We should note that when no likelihood is defined (e.g., $p(\mathbf{x}_t|\mathbf{w})$ is independent of $\mathbf{w}$), BayesPA will passively set $q_{t+1}(\mathbf{w}) = q_t(\mathbf{w})$ if $q_t(\mathbf{w})$ suffers no loss; otherwise, it will aggressively project $q_t(\mathbf{w})$ to the feasible region. We call it non-likelihood BayesPA.

In practical problems, the constraints in (5) could be unrealizable. To deal with such cases, we introduce the soft-margin version of BayesPA learning, which is equivalent to minimizing the objective function $\mathcal{L}(q(\mathbf{w}))$ in problem (5) with a regularization term, similar as in SVMs (Cortes and Vapnik, 1995):

$$q_{t+1}(\mathbf{w}) = \arg\min_{q(\mathbf{w}) \in \mathcal{F}_t} \mathcal{L}(q(\mathbf{w})) + 2c \cdot \ell_\epsilon(q(\mathbf{w}); \mathbf{x}_t, y_t).$$

(6)

For the max-margin classifiers that we consider, two types of loss functionals $\ell_\epsilon(q(\mathbf{w}); \mathbf{x}_t, y_t)$ are common:

3. Under the Bayesian setting, we regard Bayes’ rule as a natural update (passive), since no aggressive projection is applied.
### Online Bayesian Passive-Aggressive Learning

<table>
<thead>
<tr>
<th>Methods</th>
<th>Max-margin learning?</th>
<th>Bayesian inference?</th>
<th>Streaming update?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>SVB</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>RegBayes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>BayesPA</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

Table 1: The comparison between BayesPA and its various precursors, including online PA, streaming variational Bayes (SVB) and regularized Bayesian inference (RegBayes), in three different aspects.

1. **Averaging classifier**: assume that a post-data posterior distribution $q(w)$ is given, then an averaging classifier makes predictions using the sign rule $\hat{y}_t = \text{sign} \ E_q[w][w^\top x_t]$ when the discriminant function has the simple linear form, $f(x_t; w) = w^\top x_t$. For this classifier, its hinge loss is therefore defined as:

   $$\ell_{\text{ave}}^{\epsilon}(q(w); x_t, y_t) = (\epsilon - y_t E_q[w][w^\top x_t])_+.$$  

2. **Gibbs classifier**: assume that a post-data posterior distribution $q(w)$ is given, then a Gibbs classifier randomly draws a weight vector $w \sim q(w)$ to make predictions using the sign rule $\hat{y}_t = \text{sign} w^\top x_t$, when the discriminant function has the same linear form as above. For each single $w$, we can measure its hinge loss $(\epsilon - y_t w^\top x_t)_+$. To account for the randomness of $w$, the expected hinge loss of a Gibbs classifier is therefore defined as:

   $$\ell_{\text{gibbs}}^{\epsilon}(q(w); x_t, y_t) = E_q[w][(\epsilon - y_t w^\top x_t)_+]_+.$$  

They are closely connected via the following lemma due to the convexity of the function $(x)_+$.

**Lemma 1** The expected hinge loss $\ell_{\epsilon}^{\text{gibbs}}$ is an upper bound of the hinge loss $\ell_{\epsilon}^{\text{ave}}$, that is,

$$\ell_{\epsilon}^{\text{gibbs}}(q(w); x_t, y_t) \geq \ell_{\epsilon}^{\text{ave}}(q(w); x_t, y_t).$$

BayesPA is deeply connected to its various precursors reviewed in Section 3, as summarized in Table 1. First, BayesPA is a natural Bayesian extension of online PA, which is explicated via the following theorem. Note that this theorem only serves as a connection to PA, and will not be relied on by our algorithms, where we are more interested in the cases with nontrivial likelihood models to describe the underlying structures (e.g., topic structures), which are typically out-of-reach for the standard PA. Nevertheless, the idea of the proof details would later be applied to develop practical BayesPA algorithms for topic models. Therefore, we include the complete proof here.

**Theorem 2** If the prior is normal $q_0(w) = N(w_0, I)$, $\mathcal{F}_t = \mathcal{P}$, and we use the averaging classifier $\ell_{\epsilon}^{\text{ave}}$, the non-likelihood BayesPA subsumes the online PA.
The optimal solution is

Using this primal-dual interpretation, we prove that for the normal prior solution to the dual problem where the normalization term \( \Gamma(\tau_t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}||w - (\mu_t + \tau_t y_t x_t)||^2\right) \), the distribution \( q^*(w) \) of the above problem can be derived from its functional Lagrangian and has the following form:

\[
q^*(w) = \frac{1}{\Gamma(\tau^*_t, x_t, y_t)} q_t(w) \exp\left(\tau^*_t y_t w^\top x_t\right),
\]

where the normalization term \( \Gamma(\tau_t, x_t, y_t) = \int_w q_t(w) \exp\left(\tau_t y_t w^\top x_t\right) dw \), and \( \tau^*_t \) is the optimal solution to the dual problem

\[
\begin{aligned}
&\max_{\tau_t} & & \epsilon \tau_t - \log \Gamma(\tau_t, x_t, y_t) \\
&\text{s.t.} & & 0 \leq \tau_t \leq 2c.
\end{aligned}
\]

Using this primal-dual interpretation, we prove that for the normal prior \( p_0(w) = \mathcal{N}(w_0, I) \), the post-data posterior is also Gaussian: \( q_t(w) = \mathcal{N}(\mu_t, I) \) for some \( \mu_t \) in each round \( t = 0, 1, 2, \ldots \). This can be shown by induction. By our assumption, \( q_0(w) = p_0(w) = \mathcal{N}(w_0, I) \) is Gaussian. Assume for round \( t \geq 0 \), the distribution \( q_t(w) = \mathcal{N}(\mu_t, I) \). Then for round \( t + 1 \), Eq. (8) suggests the distribution

\[
q_{t+1}(w) = \frac{C}{\Gamma(\tau^*_t, x_t, y_t)} \exp\left(-\frac{1}{2}||w - (\mu_t + \tau^*_t y_t x_t)||^2\right),
\]

where the constant \( C = \exp(y_t \tau^*_t \mu_t^\top x_t + \frac{1}{2} \tau^*_t y_t x_t^\top x_t) \). Therefore, the distribution \( q_{t+1}(w) = \mathcal{N}(\mu_{t+1}, y_t x_t, I) \), and the normalization term is \( \Gamma(\tau_{t+1}, x_t, y_t) = (\sqrt{2\pi})^K \exp(\tau_{t+1} y_t x_t^\top \mu_{t+1} + \frac{1}{2} \tau_{t+1} y_t x_t^\top x_t) \) for any \( \tau_t \in [0, 2c] \).

Next, we show that \( \mu_{t+1} = \mu_t + \tau^*_t y_t x_t \) is the optimal solution of the online PA update rule (Crammer et al., 2006). To see this, we replace \( \Gamma(\tau_t, x_t, y_t) \) in problem (9) with our derived form. Ignoring constant terms, we obtain the dual problem

\[
\begin{aligned}
&\max_{\tau_t} & & \epsilon \tau_t - \frac{1}{2} \tau_t^2 x_t^\top x_t - y_t \tau_t \mu_t^\top x_t \\
&\text{s.t.} & & 0 \leq \tau_t \leq 2c,
\end{aligned}
\]

which is exactly the dual form of the online PA update equation:

\[
\mu_{t+1}^{PA} = \arg \min_{\mu} \frac{1}{2}||\mu - \mu_t||^2 + 2c \cdot \xi_t \\
\text{s.t.} \ y_t \mu^\top x_t \geq \epsilon - \xi_t, \ \xi_t \geq 0.
\]

The optimal solution is \( \mu_{t+1}^{PA} = \mu_t + \tau^*_t y_t x_t \). Note that \( \tau^*_t \) is the optimal solution of dual problem (10) shared by both PA and BayesPA. Therefore, we conclude that \( \mu_{t+1} = \mu_{t+1}^{PA} \).

Second, suppose some algorithm \( \mathcal{A} \) is capable of solving problem (6), then it would produce streaming updates to the posterior distribution. For averaging classifiers, it is easy to modify the proof of theorem 2 to derive the update rule of BayesPA, which is presented in the following lemma.
Lemma 3 If $\mathcal{F}_t = \mathcal{P}$ and we use the averaging classifier with loss functional $\ell^\text{ave}_c$, the update rule of online BayesPA is

$$q_{t+1}(w) = \frac{1}{\Gamma(\tau^*_t, x_t, y_t)} q_t(w)p(x_t|w) \exp \left( \tau^*_t y_t w^\top x_t \right),$$  \hspace{1cm} (11)

where $\Gamma(\tau, x_t, y_t)$ is the normalization term $\Gamma(\tau, x_t, y_t) = \int_w q_t(w)p(x_t|w) \exp \left( \tau y_t w^\top x_t \right) dw$, and $\tau^*_t$ is the optimal solution to the dual problem

$$\max_{\tau_t} \epsilon \tau_t - \log \Gamma(\tau_t, x_t, y_t) \quad \text{s.t.} \quad 0 \leq \tau_t \leq 2c.$$  

For Gibbs classifiers, we have the following lemma to characterize its streaming update rule.

Lemma 4 If $\mathcal{F}_t = \mathcal{P}$ and we use the Gibbs classifier with loss functional $\ell^\text{gibbs}_e$, the update rule of online BayesPA is

$$q_{t+1}(w) = \frac{q_t(w)p(x_t|w) \exp \left( -2c \left( \epsilon - y_t w^\top x_t \right) \right)}{\Gamma(x_t, y_t)},$$  \hspace{1cm} (12)

where $\Gamma(x_t, y_t)$ is the normalization constant.

In both update rules (11) and (12), the post-data posterior $q_t(w)$ in the previous round $t$ can be treated as a prior, while the newly observed data and the loss it incurs provide a likelihood and an un-normalized pseudo-likelihood respectively. Due to the analytical form, the BayesPA update rule (12) is sequential in nature, similar to the conventional Bayes rule (3). Therefore, the sequential posterior at time $T$ is the same as the batch posterior that observes all the data instances up to $T$.

For the case with averaging classifiers, Eq. (11) reduces to the streaming Bayesian update problem if there is no loss functional (i.e., $\ell_e = 0$). Therefore, BayesPA is an extension to streaming variational Bayes (SVB) with imposed max-margin posterior constraints.

Finally, the update formulation (6) is essentially a RegBayes problem with a single data point $(x_t, y_t)$. Although RegBayes inference is normally intractable, we would show later in the paper how to use variational approximation to bypass the difficulty for specific settings. This would lead to variational approximation algorithm $\mathcal{A}$ for the streaming update of post-data posterior.

Besides treating a single data point at a time, a useful technique in practice to reduce the noise in data is to use mini-batches. Suppose that we have a mini-batch of data points at time $t$ with an index set $B_t$. Let $X_t = \{x_d\}_{d \in B_t}$, $Y_t = \{y_d\}_{d \in B_t}$. The online BayesPA update equation for this mini-batch can be defined in a natural way:

$$\min_{q \in \mathcal{F}_t} \text{KL} \left[ q(w) || q_t(w) \right] - E_q(w) \left[ \log p(X_t|w) \right] + 2c \cdot \ell_e \left( q(w); X_t, Y_t \right),$$

where $\ell_e(q(w); X_t, Y_t) = \sum_{d \in B_t} \ell_e(q(w); x_d, y_d)$ is the cumulative loss functional on the mini-batch $B_t$. Like the vanilla PA methods (Crammer et al., 2006), BayesPA on mini-batches may not have closed-form update rules, and numerical optimization methods are needed to solve this new formulation.
4.2 BayesPA Learning with Latent Structures

To expressively explain complex real-world data, Bayesian models with latent structures have been extensively developed. The latent structures could typically be characterized by a hierarchy of variables, which are generally grouped into two sets—local latent variables $h_d$ ($d \geq 0$) that characterize the hidden structures of each observed data $x_d$ and global variables $\mathcal{M}$ that capture the common properties shared by all data.

As illustrated in Figure 2, BayesPA learning with latent structures aims to update the distribution of $\mathcal{M}$ as well as the classifier weights $w$, based on each incoming mini-batch $(X_t, Y_t)$ and their corresponding latent variables $H_t = \{h_d\}_{d \in B_t}$. Because of the uncertainty in $H_t$, our posterior approximation algorithm $A$ would first infer the joint posterior distribution $q_{t+1}(w, \mathcal{M}, H_t)$ from

$$
\min_{q \in \mathcal{F}_t} L(q(w, \mathcal{M}, H_t)) + 2c \cdot \ell_e(q(w, \mathcal{M}, H_t); X_t, Y_t),
$$

where $L(q) = \text{KL}[q(w, \mathcal{M}, H_t) || q(t(w, \mathcal{M})p_0(H_t))] - \mathbb{E}_{q(w, \mathcal{M}, H_t)}[\log p(X_t|w, \mathcal{M}, H_t)]$ and $\ell_e(q(w, \mathcal{M}, H_t); X_t, Y_t)$ is some cumulative loss functional on the mini-batch data incurred by some classifiers on the latent variables $H_t$ and/or global variables $\mathcal{M}$. As in the case without latent variables, both averaging classifier and Gibbs classifier can be used.

Next, algorithm $A$ produces the approximate posterior $q_{t+1}(w, \mathcal{M})$. In general we would not obtain a closed-form posterior distribution by marginalizing out $H_t$, especially in dealing with some involved models like MedLDA (Zhu et al., 2012). The intractability is bypassed through the mean-field assumption $q(w, \mathcal{M}, H_t) = q(w)q(\mathcal{M})q(H_t)$. Specifically, algorithm $A$ solves problem (13) using an iterative procedure and obtain the optimal distribution $q^*(w)q^*(\mathcal{M})q^*(H_t)$. Then it sets $q_{t+1}(w, \mathcal{M}) = q^*(w)q^*(\mathcal{M})$ and proceeds to next round. Concrete examples of this method will be discussed in Section 5 and Section 6.

5. Online Max-Margin Topic Models

In this section, we apply the theory of online BayesPA to topic modeling. We first review the basic ideas of max-margin topic models, and develop online learning algorithms based on BayesPA with averaging and Gibbs classifiers respectively.
5.1 Basics of MedLDA

A max-margin topic model consists of a latent Dirichlet allocation (LDA) model (Blei et al., 2003) for describing the underlying topic representations of document content and a max-margin classifier for making predictions. Specifically, LDA is a hierarchical Bayesian model that treats each document as an admixture of $K$ topics, $\Phi = \{\phi_k\}_{k=1}^K$, where each topic $\phi_k$ is a multinomial distribution over a given $W$-word vocabulary. The generative process of the $d$-th document ($1 \leq d \leq D$) is described as follows:

- Draw a topic mixture proportion vector $\theta_d | \alpha \sim \text{Dir}(\alpha)$.
- For the $i$-th word in document $d$, where $i = 1, 2, ..., n_d$,
  - draw a latent topic assignment $z_{di} \sim \text{Mult}(\theta_d)$.
  - draw the word instance $x_{di} \sim \text{Mult}(\phi_{z_{di}})$.

where $\text{Dir}()$ is the Dirichlet distribution and $\text{Mult}()$ is the multinomial distribution. For Bayesian LDA, the topics are also drawn from a Dirichlet distribution, i.e., $\phi_k \sim \text{Dir}(\gamma)$.

Given a document set $X = \{x_d\}_{d=1}^D$. Let $Z = \{z_d\}_{d=1}^D$ and $\Theta = \{\theta_d\}_{d=1}^D$ denote all the topic assignments and topic mixing vectors. LDA infers the posterior distribution $p(\Phi, \Theta, Z | X) \propto p(\Phi, Z)p(\Theta)p(X | \Phi)$ via Bayes’ rule. From a variational point of view, the posterior is identical to the solution of the optimization problem:

$$\min_{q \in P} \text{KL}[q(\Phi, \Theta, Z) || p(\Phi, \Theta, Z | X)].$$

The advantage of the variational formulation of Bayesian inference lies in the convenience of posing restrictions on the post-data distribution with a regularization term. For supervised topic models (Blei and McAuliffe, 2010; Zhu et al., 2012), such a regularization term could be a loss function of a prediction model $w$ on the data $X = \{x_d\}_{d=1}^D$ and response signals $Y = \{y_d\}_{d=1}^D$. As a regularized Bayesian (RegBayes) model, MedLDA infers a distribution of the latent variables $Z$ as well as classification weights $w$ by solving the problem:

$$\min_{q \in P} \mathcal{L}(q(w, \Phi, \Theta, Z)) + 2c \sum_{d=1}^D \ell_q(q(w, z_d); x_d, y_d),$$

where $\mathcal{L}(q(w, \Phi, \Theta, Z)) = \text{KL}[q(w, \Phi, \Theta, Z) || p(w, \Phi, \Theta, Z | X)]$. To specify the loss function, a linear discriminant function needs to be defined with respect to $w$ and $z_d$

$$f(w, z_d) = w^\top \tilde{z}_d,$$

where $\tilde{z}_{dk} = \frac{1}{n_d} \sum z_{di} = k$ is the average frequency of assigning the words in document $d$ to topic $k$. Based on the discriminant function, both averaging classifiers with the hinge loss

$$\ell^\text{ave}_q(q(w, z_d); x_d, y_d) = (\epsilon - y_d \mathbb{E}_{q(w, z_d)}[f(w, z_d)])_+,$$

4. Without causing confusion, we slightly abused the notation $K$ to denote the topic number (i.e., the latent dimension) in topic models.
and Gibbs classifiers with the expected hinge loss

\[ E_g(q(w, z_d); x_d, y_d) = \mathbb{E}_{q(w, z_d)} \left[ (\epsilon - y_d f(w, z_d))^+ \right], \quad (16) \]

have been proposed, with extensive comparisons reported in Zhu et al. (2014b) using batch learning algorithms. As commented in (Chang and Blei, 2010), defining the classifier on \( \bar{\theta}_d \), instead of \( \theta_d \), enforces words and labels to share the same topics. Besides, it retains the conjugacy structure between the Dirichlet prior of \( \theta \) and the multinomial likelihood of generating \( z \); and it will allow us to integrate out \( \theta \) for efficient collapsed inference.

### 5.2 Online MedLDA

We first apply online BayesPA to MedLDA with averaging classifiers, which will be referred to as paMedLDA\textsuperscript{ave} for convenience. During inference, we integrate out the local variables \( \Theta_t \) using the conjugacy between a Dirichlet prior and a multinomial likelihood (Griffiths and Steyvers, 2004; Teh et al., 2006b), which potentially improves the inference accuracy. Then we have the global variables \( \mathcal{M} = \Phi \) and local variables \( H_t = Z_t \). The latent BayesPA rule (13) becomes:

\[
\begin{align*}
\min_{q,\xi_d} & \quad \text{KL} \left[ q(w, \Phi, Z_t) \mid \mid q_t(w, \Phi) p_0(Z_t) p(X_t | \Phi, Z_t) \right] + 2c \sum_{d \in B_t} \xi_d, \\
\text{s.t.} & \quad y_d \mathbb{E}_{q(w, z_d)} [w^\top \bar{z}_d] \geq \epsilon - \xi_d, \quad \xi_d \geq 0, \quad \forall d \in B_t, \\
& \quad q(w, \Phi, Z_t) \in \mathcal{P}.
\end{align*}
\]

Since directly solving the above problem is intractable, we would impose a mild mean-field assumption

\[ q(w, \Phi, Z_t) = q(w) q(\Phi) q(Z_t). \]

Now, problem (17) can be solved using an iterative procedure that alternately updates each factor distribution (Jordan et al., 1998), as detailed below:

1. **Update global \( q(\Phi) \):** By fixing the distributions \( q(Z_t) \) and \( q(w) \), we can ignore irrelevant terms and solve

\[
\begin{align*}
\min_{q(\Phi)} & \quad \text{KL} \left[ q(\Phi) q(Z_t) \mid \mid q_t(\Phi) p_0(Z_t) p(X_t | \Phi, Z_t) \right].
\end{align*}
\]

The optimal solution has the following closed form:

\[
q^*(\Phi_k) \propto q_t(\Phi_k) \exp \left( \mathbb{E}_{q(Z_t)} \left[ \log p_0(Z_t) p(X_t | Z_t, \Phi) \right] \right), \quad k = 1, 2, ..., K. \quad (18)
\]

If initially the prior is \( q_0(\Phi_k) = \text{Dir}(\Delta_{k1}^0, ..., \Delta_{kW}^0) \), then by induction the inferred distributions in each round are also in the family of Dirichlet distributions, namely, \( q_t(\Phi_k) = \text{Dir}(\Delta_{k1}^t, ..., \Delta_{kW}^t) \). Using equation (18), we can derive

\[
q^*(\Phi_k) = \text{Dir}(\Delta_{k1}^*, ..., \Delta_{kW}^*), \quad (19)
\]

where \( \Delta_{kw}^* = \Delta_{kw}^t + \sum_{d \in B_t} \sum_{i=1}^{n_d} \gamma_{di}^k \cdot I[x_{di} = w] \) for all words \( w \) (1 \( \leq w \leq W \) in the vocabulary and \( \gamma_{di}^k = \mathbb{E}_{q(z_d)} [z_{di} = k] \) is the probability of assigning each word \( x_{di} \) to topic \( k \).
2. **Update global weight** $q(w)$: Keeping all the other distributions fixed, $q(w)$ can be solved as

$$
\min_{q(w)} \text{KL}[q(w) || q_t(w)] + 2c \sum_{d \in B_t} \xi_d,
$$

subject to:

$$
y_d \mathbb{E}_{q(w)}[w]^T \tilde{z}_d \geq \epsilon - \xi_d, \quad \xi_d \geq 0, \quad \forall d \in B_t,
$$

where $\tilde{z}_d = \mathbb{E}_{q(z_d)}[\tilde{z}_d]$ is the expectation of topic assignments under the fixed distribution $q(Z)$. Similar to Proposition 2 in MedLDA (Zhu et al., 2012), the optimal solution is attained by solving the Lagrangian form with respect to $q(w)$, which gives

$$
q^*(w) = \frac{1}{Z(\tau^*_d)} q_t(w) \exp \left( \sum_{d \in B_t} \tau^*_d y_d w^T \tilde{z}_d \right),
$$

where the Lagrange multipliers $\tau^*_d (d \in B_t)$ are obtained by solving the dual problem

$$
\max_{0 \leq \tau_d \leq 2c} \epsilon \sum_{d \in B_t} \tau_d - \log Z(\tau_d).
$$

For the common spherical Gaussian prior $q_0(w) = \mathcal{N}(0, \sigma^2 I)$, by induction the distribution $q_t(w) = \mathcal{N}(\mu_t, \sigma^2 I)$ at each round. So equation (20) gives $q^*(w) = \mathcal{N}(\mu^*, \sigma^2 I)$, where

$$
\mu^* = \mu_t + \sigma^2 \sum_{d \in B_t} \tau_d^* y_d \tilde{z}_d.
$$

Furthermore, the dual problem becomes,

$$
\max_{0 \leq \tau_d \leq 2c} \epsilon \sum_{d \in B_t} \tau_d - \sum_{d_1, d_2 \in B_t} \frac{1}{2} \sigma^2 \tau_{d_1} \tau_{d_2} \tilde{z}_{d_1}^T \tilde{z}_{d_2} - \mu^T \sum_{d \in B_t} y_d \tau_d \tilde{z}_d,
$$

which is identical to the Lagrangian dual of the classical PA problem with mini-batch $B_t$ (expressed in the equivalent constrained form by introducing slack variables)

$$
\min_{\mu} \frac{||\mu - \mu_t||^2}{2\sigma^2} + 2c \sum_{d \in B_t} \left( \epsilon - y_d \mu^T \tilde{z}_d \right)_+.
$$

This equivalence suggests that we could rely on contemporary PA techniques to solve for $\mu^*$. In particular, for instances coming one at a time (i.e., $B_t = \{t\}$, $\forall t$), we have the closed-form solution

$$
\tau^*_t = \min \left\{ 2c, \frac{(\epsilon - y_t \mu_t^T \tilde{z}_t)_+}{||\tilde{z}_t||^2} \right\},
$$

whose computation requires $O(K)$ time; for mini-batches, we could adapt methods solving linear SVM to either the dual (22) or primal (23) problem, which by state-of-the-arts require complexity $O(\text{poly}(\epsilon^{-1}) K)$ per training instance in order to obtain $\epsilon$-accurate solutions. Here we choose a gradient-based method similar to Shalev-Shwartz et al. (2011). Specifically, we first set $\mu^1 = \mu_t$, and then take the gradient steps $i = 2, 3, \ldots$ until $\mu^i$ converges to $\mu^*$. Let
where annealing rate $\rho_i = \sigma^2 i^{-1}$ and

$$\nabla_i = \frac{\mu^i - \mu^i}{\sigma^2} - 2c \sum_{d \in A_i^i} y_d \hat{Z}_d.$$

Correspondingly, we can derive the gradient-based update rule for the dual parameters. Imagine that we implicitly maintain the relationship $\mu = \mu_t + \sigma^2 \sum_{d \in B_t} \tau_d y_d \hat{Z}_d$. Then the following update rule for $\tau_d$ ($d \in B_t$) naturally implies the update rule (24) for $\mu$:

$$\tau_d^i \leftarrow \begin{cases} (1 - \frac{1}{i}) \tau_d^i + \frac{2c}{i} & \text{for } d \in A_i^i \\ (1 - \frac{1}{i}) \tau_d^i & \text{for } d \notin A_i^i. \end{cases}$$

Therefore, the gradient steps adaptively adjust the contribution of each latent $\hat{Z}_d$ to $\mu$ based on the loss it incurs. Furthermore, the annealing makes sure that $0 \leq \tau_d^i \leq 2c$ for all $i$. Since the problem (22) is concave, it can be guaranteed that $\tau_d^i$ converges to $\tau_d^*$. This correspondence would be used in updating $q(Z_t)$.

3. Update local $q(Z_t)$: Fixing all the other distributions, we aim to solve

$$\min_{q(Z_t)} \text{KL}[q(Z_t)||p_0(Z_t)p(X_t|Z_t, \Phi)] + 2c \sum_{d \in B_t} \xi_d,$$

s.t.: $y_d \mu^* \tau_d^* \mathbb{E}_{q(z_d)}[\hat{Z}_d] \geq \epsilon - \xi_d, \quad \xi_d \geq 0, \quad \forall d \in B_t,$

where $\mu^* = \mathbb{E}_{q(w)}[w]$ is the expectation of $w$ under the fixed distribution $q(w)$. Unlike the weight $w$, the expectation over $Z_t$ during optimization is intractable due to the combinatorial space of values. Instead, we adopt the same approximation strategy as MedLDA (Zhu et al., 2012): fix $\xi, \tau_d$ at the previous global step, and use the approximate solution

$$q^*(Z_t) = p_0(Z_t)p(X_t|Z_t, \Phi) \exp \left( \sum_{d \in B_t} \tau_d^* y_d \mu^* \tau_d^* \hat{Z}_d \right).$$

Then the expectation of $\hat{Z}_d$, as needed in the global updates, could be approximated by samples from the distribution $q^*(Z_t)$. Specifically, we use Gibbs sampling with the conditional distribution

$$q(z_{di} = k | Z_{di}^{-di}) \propto (\alpha + C_{dk}^{-di}) \exp \left( \Lambda_{k, x_{di}} + \sum_{d \in B_t} n_d^{-1} y_d \tau_d^* m_k^* \right).$$

where $\Lambda_{z_{di}, x_{di}} = \mathbb{E}_{q(\Phi)}[\log(\Phi_{z_{di}, x_{di}})] = \Psi(\Delta_{z_{di}, x_{di}})^* - \Psi(\sum_w \Delta_{z_{di}, w}^*)$ (note that $\Psi(\cdot)$ is the digamma function) and $C_{dk}^{-di}$ is a vector with the $k$-th entry being the number of words in document $d$ (except the $i$-th word) that are assigned to topic $k$.

Then we draw $J$ samples $\{Z_{ij}^{(j)}\}_{j=1}^J$ using Eq. (25), discard the first $\beta$ ($0 \leq \beta < J$) burn-in samples, and approximate $\hat{Z}_{dk}$ with the empirical sum $(J - \beta)^{-1} \sum_{j=\beta+1}^J \sum_{i} [z_{di}^{(j)} = k]$. 


Algorithm 1 Online MedLDA

1: Let \( q_0(w) = N(0; \sigma^2 I) \), \( q_0(\phi_k) = \text{Dir}(\gamma), \forall k \).
2: for \( t = 0 \rightarrow \infty \) do
3:   Set \( q_t(\Phi, w) = q_t(\Phi)q_t(w) \). Initialize \( Z_t \).
4:   for \( i = 1 \rightarrow I \) do
5:     Draw samples \( \{Z_t^{(j)}\}_{j=1}^J \) from Eq. (25).
6:     Discard the first \( \beta \) burn-in samples (\( \beta < J \)).
7:     Use the rest \( J - \beta \) samples to update \( q(\Phi, w) \) following Eq.s (19, 20).
8:   end for
9:   Set \( q_{t+1}(\Phi, w) = q(\Phi, w) \).
10: end for

At each round \( t \) of BayesPA optimization during training, we run the global and local updates alternately until convergence, and assign \( q_t(\Phi, w) = q^*(\Phi)q^*(w) \), as outlined in Algorithm 1. To make predictions on testing data, we use the mean \( \mu \) as the classification weight and apply the prediction rule. The inference of \( \bar{\varepsilon} \) for testing documents is similar to Zhu et al. (2014b). First, we draw a single sample of \( \Phi \), and for each test document \( d \), we infer the MAP of \( \theta_d \). Then, we directly run the sampling of \( z_d \) until the burn-in stage is completed, and use the average of several samples to compute \( \bar{z}_d \). Then the prediction rule is applied on \( \mu \) and \( \bar{z}_d \).

5.3 Online Gibbs MedLDA

In this section, we apply the theory of BayesPA to Gibbs MedLDA. As shown in Zhu et al. (2014b), using Gibbs classifiers admits efficient inference algorithms by exploring data augmentation (DA) techniques (Tanner and Wong, 1987; Polson and Scott, 2011). Based on this insight, we will develop our efficient online inference algorithms for Gibbs MedLDA. We denote the model by \( \text{paMedLDA}^{\text{gibbs}} \). Specifically, let \( \zeta_d = \epsilon - y_df(w, z_d) \) and \( \psi(y_d|z_d, w) = e^{-2\zeta_d} \). By Lemma 4, the optimal solution to problem (13) is

\[
q_{t+1}(w, \mathcal{M}, H_t) = \frac{q_t(w, \mathcal{M})p_0(H_t)p(X_t|H_t, \mathcal{M})\psi(Y_t|H_t, w)}{\Gamma(X_t, Y_t)},
\]

where \( \psi(Y_t|H_t, w) = \prod_{d \in B_t} \psi(y_d|h_d, w) \) and \( \Gamma(X_t, Y_t) \) is a normalization constant. The basic idea of DA is to construct conjugacy between prior and data during inference by introducing augmented variables. Specifically, we would use the following equality (Zhu et al., 2014b):

\[
\psi(y_d|z_d, w) = \int_0^\infty \psi(y_d, \lambda_d|z_d, w)d\lambda_d,
\]

where \( \psi(y_d, \lambda_d|z_d, w) = (2\pi\lambda_d)^{-1/2}\exp\left(-\frac{(y_d - \lambda_d z_d)^2}{2\lambda_d}\right) \). Equality (26) essentially implies that the collapsed posterior \( q_{t+1}(w, \Phi, Z_t) \) is a marginal distribution of

\[
q_{t+1}(w, \Phi, Z_t, \lambda_t) = \frac{p_0(Z_t)q_t(w, \Phi)p(X_t|Z_t, \Phi)\psi(Y_t, \lambda_t|Z_t, w)}{\Gamma(X_t, Y_t)},
\]

where \( \psi(Y_t, \lambda_t|Z_t, w) = \prod_{d \in B_t} \psi(y_d, \lambda_d|z_d, w) \) and \( \lambda_t = \{\lambda_d\}_{d \in B_t} \) are augmented variables locally associated with individual documents. In fact, the augmented distribution \( q_{t+1}(w, \Phi, Z_t, \lambda_t) \) is the
solution to the problem:

\[
\min_{q \in \mathcal{P}} \mathcal{L}(q(w, \Phi, Z_t, \lambda_t)) - \mathbb{E}_q \left[ \log \psi(Y_t, \lambda_t | Z_t, w) \right], \tag{27}
\]

where \( \mathcal{L}(q(w, \Phi, Z_t, \lambda_t)) = \text{KL}[q(w, \Phi, Z_t, \lambda_t) \| q_\theta(w, \Phi) p_0(Z_t)] - \mathbb{E}_q[\log p(X_i | Z_t, \Phi)] \). In fact, this objective is an upper bound of that in the original problem (13) (See Appendix A for details).

Again, with the mild mean-field assumption that \( q(w, \Phi, Z_t, \lambda_t) = q(w, \Phi) q(Z_t, \lambda_t) \), we can solve problem (27) via an iterative procedure that alternately updates each factor distribution (Jordan et al., 1998), as detailed below.

1. **Global Update**: By fixing the distribution of local variables, \( q(Z_t, \lambda_t) \), and ignoring irrelevant terms, the optimal distribution of \( w \) and \( \Phi \) can be shown to have the induced factorization form, \( q(w, \Phi) = q(w) q(\Phi) \). For \( q(\Phi) \), the update rule is exactly (19). For \( q(w) \), we have the update rule

\[
q_{t+1}(w) \propto q_t(w) \exp \left( \mathbb{E}_{q(Z_t, \lambda_t)} \left[ \log p_0(Z_t) \psi(Y_t, \lambda_t | Z_t, w) \right] \right).
\]

If the initial prior is normal \( q_0(w) = \mathcal{N}(w; \mu^0, \Sigma^0) \), by induction we can show that the inferred distribution in each round is also a normal distribution, namely, \( q_t(w) = \mathcal{N}(w; \mu^t, \Sigma^t) \).

Indeed, the optimal solution of \( q(w) \) to problem (27) is

\[
q^*(w) = \mathcal{N}(w; \mu^*, \Sigma^*),
\]

where the posterior parameters are computed as

\[
\Sigma^* = (\Sigma^t)^{-1} + c^2 \sum_{d \in B_t} \mathbb{E}_{q(z_d, \lambda_d)} \left[ \lambda_d^{-1} z_d z_d^\top \right],
\]

\[
\mu^* = \Sigma^* \left( (\Sigma^t)^{-1} \mu^t + c \sum_{d \in B_t} \mathbb{E}_{q(z_d, \lambda_d)} \left[ y_d \left( 1 + c \lambda_d^{-1} \right) z_d \right] \right).
\]

For the sequential update rule, we simply set \( \mu^{t+1} = \mu^* \) and \( \Sigma^{t+1} = \Sigma^* \).

2. **Local Update**: Given the distribution of global variables, \( q(\Phi, w) \), the mean-field update equation for \( (Z_t, \lambda_t) \) is

\[
q(Z_t, \lambda_t) \propto p_0(Z_t) \prod_{d \in B_t} \frac{1}{\sqrt{2 \pi \lambda_d}} \exp \left( \sum_{i \in [n_d]} \Lambda_{x_{di}, x_{di}} - \mathbb{E}_{q(\Phi, w)} \left[ \frac{(\lambda_d + c \zeta_d)^2}{2 \lambda_d} \right] \right),
\]

where \( \Lambda \) admits the same definition as in (25). But it is impossible to evaluate the expectation in the global update using \( q(Z_t, \lambda_t) \) because of the huge number of configurations for \( (Z_t, \lambda_t) \). As a result, we turn to Gibbs sampling and estimate the required expectations using multiple empirical samples. This hybrid strategy has shown promising performance for LDA (Mimno et al., 2012). Specifically, the conditional distributions used in the Gibbs sampling are as follows:

18
Algorithm 2 Online Gibbs MedLDA

1: Let \( q_0(w) = \mathcal{N}(0, \sigma^2 I) \), \( q_0(\phi_k) = \text{Dir}(\gamma), \forall k \).
2: for \( t = 0 \rightarrow \infty \) do
3: \hspace{1em} Set \( q(\Phi, w) = q_t(\Phi)q_t(w) \). Initialize \( Z_t \).
4: \hspace{1em} for \( i = 1 \rightarrow I \) do
5: \hspace{2em} Draw samples \( \{Z_t^{(j)}, \lambda_t^{(j)}\}_{j=1}^J \) from Eq.s (29, 30).
6: \hspace{2em} Discard the first \( \beta \) burn-in samples (\( \beta < J \)).
7: \hspace{2em} Use the rest \( J - \beta \) samples to update \( q(\Phi, w) \) following Eq.s (19, 28).
8: \hspace{1em} end for
9: \hspace{1em} Set \( q_{t+1}(\Phi, w) = q(\Phi, w) \).
10: end for

- For \( Z_t \): By canceling out common factors, the conditional distribution of one variable \( z_{di} \) given \( Z_t^{\neg di} \) and \( \lambda_t \) is
  \[
  q(z_{di} = k | Z_t^{\neg di}, \lambda_t) \propto (\alpha + C_{dk}) \exp \left( \frac{c_{pd}(\alpha + \lambda_d) \mu_k^*}{n_d \lambda_d} + \Lambda_{k,xd} - \frac{c^2(\mu_k^* + \Sigma^*_{k,k} + 2(\mu_k^* + \Sigma^*_{k,k})^\top C_{dk})}{2n_d^2 \lambda_d} \right),
  \]
  equation (29)

where \( \Sigma^*_{k,k} \) is the \( k \)-th column of \( \Sigma^* \).

- For \( \lambda_t \): Let \( \zeta_d = \epsilon - y_d \bar{z}_d^\top \mu^* \). The conditional distribution of each variable \( \lambda_d \) given \( Z_t \) is
  \[
  q(\lambda_d | Z_t) \propto \frac{1}{\sqrt{2\pi \lambda_d}} \exp \left( -\frac{c^2 \bar{z}_d^\top \Sigma^* \bar{z}_d + (\lambda_d + c\zeta_d)^2}{2\lambda_d} \right)
  = \mathcal{IG} \left( \frac{1}{2}, 1, c^2 \left( \zeta_d + \bar{z}_d^\top \Sigma^* \bar{z}_d \right) \right),
  \]
  equation (30)

a generalized inverse gaussian distribution (Devroye, 1986). Therefore, \( \lambda_d^{-1} \) follows an inverse gaussian distribution, that is,

  \[
  q(\lambda_d^{-1} | Z_t) = \mathcal{IG} \left( \frac{1}{c\sqrt{\zeta_d^2 + \bar{z}_d^\top \Sigma^* \bar{z}_d}}, 1 \right)
  \]

from which we can draw a sample in constant time (Michael et al., 1976).

For training, we run the global and local updates alternately until convergence at each round of PA optimization, as outlined in Alg. 2. To make predictions on testing data, we then draw one sample of \( \hat{w} \) as the classification weight and apply the prediction rule. The inference of \( \bar{z} \) for testing documents is the same as online MedLDA.

6. Extensions

In the above topic models, we assume that the number of topics (i.e., \( K \)) is pre-specified, which is unfortunately unknown in practice. An extra procedure is often applied to select a good \( K \) value.
We now present extensions of online MedLDA to automatically determine the unknown $K$ values by leveraging nonparametric Bayesian techniques. We also present an extension of these models for multi-task learning.

### 6.1 Online Nonparametric MedLDA

We first present online nonparametric MedLDA for resolving the unknown number of topics, based on the theory of hierarchical Dirichlet process (HDP) (Teh et al., 2006a).

#### 6.1.1 Batch MedHDP

A two-level HDP provides an extension to LDA that allows for a nonparametric inference of the unknown topic numbers. The generative process of HDP can be summarized using a stick-breaking construction (Wang and Blei, 2012b), where the stick lengths $\bar{\pi} = \{\bar{\pi}_k\}_{k=1}^\infty$ are generated as:

$$\pi_k = \pi_k \prod_{i< k} (1 - \pi_i), \; \bar{\pi}_k \sim \text{Beta}(1, \eta), \; \text{for} \; k = 1, ..., \infty,$$

and the topic mixing proportions are generated as $\theta_d \sim \text{Dir}(\alpha \pi)$, for $d = 1, ..., D$. Each topic $\phi_k$ is a sample from a Dirichlet base distribution, i.e., $\phi_k \sim \text{Dir}(\gamma)$. After we get the topic mixing proportions $\theta_d$, the generation of words is the same as in the standard LDA.

To extend the HDP topic model for predictive tasks, we introduce a classifier $w$ that is drawn from a Gaussian process, $\mathcal{GP}(0, \Sigma)$, where the covariance function is $\Sigma(w, w') = \sigma^2 \mathbb{I}[w = w']$. We still define the linear discriminant function in the same form as Eq. (14). Since the number of words in a document is finite, the average topic assignment vector $\bar{z}_d$ has only a finite number of non-zero elements, and the dot product in Eq. (14) is in fact finite. Therefore, given the latent topic assignments, the conditional posterior of $w$ is in fact a multivariate Gaussian distribution.

Let $\bar{\pi} = \{\bar{\pi}_k\}_{k=1}^\infty$. We define maximum entropy discrimination HDP (MedHDP) topic model as solving the following RegBayes problem to infer the joint post-data posterior $q(w, \bar{\pi}, \Phi, \Theta, Z)$:

$$\min_{q \in \mathcal{P}} \mathcal{L}(q(w, \bar{\pi}, \Phi, \Theta, Z)) + 2c \sum_{d=1}^D \ell_d \left( q(w, z_d); x_d, y_d \right),$$

where $\mathcal{L}(q(w, \bar{\pi}, \Phi, \Theta, Z)) = \text{KL}[q(w, \bar{\pi}, \Phi, \Theta, Z) || p(w, \bar{\pi}, \Phi, \Theta, Z | X)]$ is the objective corresponding to the standard Bayesian inference under the variational formulation of Bayes’ rule. The loss function could be either (15) or (16). We call the resulting model with the averaging classifier MedHDP$^{\text{ave}}$ and that with the Gibbs classifier MedHDP$^{\text{gibbs}}$.

Since MedHDP is a new model, we would briefly discuss the corresponding inference problem. For the inference of MedHDP$^{\text{gibbs}}$, we can use Gibbs sampling based on Chinese Restaurant Franchise (Teh et al., 2006a; Wang and Blei, 2012a) with modifications similar to the techniques introduced in Zhu et al. (2014b). For MedHDP$^{\text{ave}}$, the current state-of-the-art for inferring max-margin Bayesian models with averaging loss resorts to mean-field assumptions and variational inference. Notice that classical mean-field derivation would fail due to the potentially unbounded space of variables. However, it is possible to incorporate Gibbs sampling into mean-field update equations to explore the unbounded space (Welling et al. 2008; Wang and Blei 2012b) and therefore bypass the difficulty. In this paper, we would not focus on developing inference algorithms for MedHDP.$^5$

---

$^5$ Given $\bar{\pi}$, $\pi$ can be computed via the stick breaking process.
but instead attain batch MedHDP algorithms from the corresponding BayesPA methods, as will be clear at the end of each subsection below.

6.1.2 Online MedHDP

To apply the ideas of BayesPA to develop online MedHDP algorithms, we have the global variables \( \mathcal{M} = (\bar{\pi}, \Phi) \), and the local variables \( H_t = (\Theta_t, Z_t) \). As in online MedLDA, we marginalize out \( \Theta_t \) by conjugacy. Furthermore, to simplify the sampling scheme, we introduce another set of auxiliary latent variables \( S_t = \{s_d\}_{d \in B_t} \), where \( s_d = \{s_{dk}\}_{k=1}^{\infty} \) and each element \( s_{dk} \) represents the number of occupied tables serving dish \( k \) in a Chinese restaurant process (CRP) (Teh et al., 2006a; Wang and Blei, 2012b). By definition, we have \( p(Z_t, S_t|\bar{\pi}) = \prod_{d \in B_t} p(s_d, z_d|\bar{\pi}) \) and

\[
p(s_d, z_d|\bar{\pi}) \propto \prod_{k=1}^{\infty} S(n_d\bar{z}_{dk}, s_{dk})(\alpha \pi_k)^{s_{dk}},
\]

where \( S(a, b) \) are unsigned Stirling numbers of the first kind (Antoniak, 1974). It is not hard to verify that \( p(z_d|\bar{\pi}) = \sum_{s_d} p(s_d, z_d|\bar{\pi}) \). After this “collapse-and-augment” procedure, we now have the local variables \( H_t = (Z_t, S_t) \). The global variables remain intact. The new BayesPA problem is now:

\[
\min_{q \in \mathcal{F}_t} \mathcal{L}(q(w, \bar{\pi}, \Phi, H_t)) + 2c \sum_{d \in B_t} \ell_q(w; x_t, y_t),
\]

where \( \mathcal{L}(q(w, \bar{\pi}, \Phi, H_t)) = \text{KL}[q(w, \bar{\pi}, \Phi, H_t) || q_t(w, \bar{\pi}, \Phi) p(Z_t, S_t|\bar{\pi}) p(X_t|Z_t, \Phi)] \). As in online MedLDA, we adopt the mean-field approximation \( q(w, \bar{\pi}, \Phi, H_t) = q(w)q(\bar{\pi})q(\Phi)q(H_t) \) and solve problem (33) via an iterative procedure detailed below.

1. **Global Update:** By fixing the distribution of local variables, \( q(H_t) \), and ignoring the irrelevant terms, we have same mean-field update equations (19) for \( \Phi \) and (21) for \( w \) with the averaging loss. For global variable \( \bar{\pi} \), we have

\[
q^*(\bar{\pi}_k) \propto q_t(\bar{\pi}_k) \prod_{d \in B_t} \exp \left( \mathbb{E}_{q(h_d)} \left[ \log p(s_d, z_d|\bar{\pi}) \right] \right).
\]

By induction, we can show that \( q_t(\bar{\pi}_k) = \text{Beta}(u^t_k, v^t_k) \) is a Beta distribution at each step, and the update equation is

\[
q^*(\bar{\pi}_k) = \text{Beta}(u^*_k, v^*_k),
\]

where \( u^*_k = u^t_k + \sum_{d \in B_t} \mathbb{E}_{q(s_d)}[s_{dk}] \) and \( v^*_k = v^t_k + \sum_{d \in B_t} \mathbb{E}_{q(s_d)}[\sum_{j > k} s_{dj}] \) for \( k = \{1, 2, \ldots\} \) and \( u^0_k = 1, v^0_k = \eta \).

Since \( Z_t \) contains only a finite number of discrete variables, we only need to maintain and update the above global distributions for a finite number of topics.

2. **Local Update:** Fixing the global distribution \( q(w, \bar{\pi}, \Phi) \), we get the mean-field update equation for \((Z_t, S_t)\):

\[
q^*(Z_t, S_t) \propto \tilde{q}(Z_t, S_t) \tilde{q}(Z_t)
\]

21
have the following sampling scheme. Notice in the local updates, \( \bar{\pi} \) that the expectation operator is essentially collapsing \( \bar{\pi} \) and \( \tau \). Point to tackle is the potentially unbounded sample space of \( Z \) from (Wang and Blei, 2012b) and adopt an approximation for \( \tilde{\pi} \), where

\[
\tilde{q}(Z_t, S_t) = \exp \left( \sum_{d \in B_t} \tau_d y_d \mathbb{E}[w]^T \bar{z}_d \right),
\]

and \( \tau_d (d \in B_t) \) are the dual variables computed in the global update. The most cumbersome point to tackle is the potentially unbounded sample space of \( Z_t \) and \( S_t \). We take the ideas from (Wang and Blei, 2012b) and adopt an approximation for \( \tilde{q}(Z_t, S_t) \):

\[
\tilde{q}(Z_t, S_t) \approx \mathbb{E}_{\theta^*}(\Phi^* \theta^*) \left[ p(X|\Phi, Z_t)p(Z_t, S_t|\bar{\pi}) \right]. \tag{37}
\]

Computing the expectation regarding \( \bar{\pi} \) in (37) turns out to be difficult. However, imagine that the expectation operator is essentially collapsing \( \bar{\pi} \) out from the joint distribution

\[
\tilde{q}(\bar{\pi}, Z_t, S_t) \approx \mathbb{E}_{\theta^*}(\Phi) \left[ q^*(\bar{\pi})p(X|\Phi, Z_t)p(Z_t, S_t|\bar{\pi}) \right]. \tag{38}
\]

Now we propose to uncollapse \( \bar{\pi} \) and sample the local variables from

\[
q^*(\bar{\pi}, Z_t, S_t) \propto \tilde{q}(\bar{\pi}, Z_t, S_t) \tilde{q}(Z_t). \tag{39}
\]

Notice in the local updates, \( \bar{\pi} \) is only an auxiliary variable. Putting all the pieces together, we have the following sampling scheme.

- **For** \( Z_t \): Let \( K \) be the current inferred number of topics. The conditional distribution of one variable \( z_{di} \) given all other local variables can be derived from (36) with \( s_d \) marginalized out for convenience.

\[
q(z_{di} = k|Z_t^{-di}, \bar{\pi}) \propto \frac{(\alpha \pi_k + C_{dk}^{-di})(C_{kx,di}^{-di} + \Delta_{kx,di}^t)}{\sum_w (C_{kw}^{-di} + \Delta_{kw}^t)} \exp \left( \sum_{d \in B_t} n_d^{-1} y_d \tau_d k^* \right). \tag{40}
\]

Besides, for \( k > K \) and symmetric Dirichlet prior \( \gamma \), (40) converge to a single rule

\[
q(z_{di} = k|Z_t^{-di}, \bar{\pi}) \propto \alpha \pi_k / W,
\]

and therefore the total probability of assigning a new topic is

\[
q(z_{di} > K|Z_t^{-di}, \bar{\pi}) \propto \alpha \left( 1 - \sum_{k=1}^{K} \pi_k \right) / W.
\]

- **For** \( S_t \): The conditional distribution of \( s_{dk} \) given \( (Z_t, \bar{\pi}, \lambda_t) \) can be derived from the joint distribution (32):

\[
q(s_{dk}|Z_t, \bar{\pi}) \propto S(n_d z_{dk}, s_{dk})(\alpha \pi_k)^{s_{dk}}. \tag{41}
\]

- **For** \( \bar{\pi} \): It can be derived from (36) that given \( (Z_t, S_t) \), each \( \bar{\pi}_k \) follows the beta distribution

\[
\bar{\pi}_k \sim \text{Beta}(a_k, b_k), \tag{42}
\]

where \( a_k = u_k^* + \sum_{d \in B_t} s_{dk} \) and \( b_k = v_k^* + \sum_{d \in B_t} \sum_{j > k} s_{dj} \).
Similar to online MedLDA, we iterate the above steps till convergence for training. For testing, the learned model is essentially a finite MedLDA, and we use the same scheme as that of online MedLDA.

Notice that if we run online MedHDP for only one round \((T = 1)\) and use the entire dataset as mini-batch \(|B| = D\), iterating the above steps till converge in fact solves the batch MedHDP problem Eq. (31). We call this batch version MedHDP\textsuperscript{ave}, and will use it as a baseline algorithm.

6.1.3 Online Gibbs MedHDP

For Gibbs MedHDP, the only difference is the loss functional \(\ell_{e}\), which is reflected in the sampling of local variables. As in online Gibbs MedLDA, we can facilitate more efficient inference by adopting the same data augmentation technique with the augmented variables \(\lambda_t\). Then the local variables are \((Z_t, S_t, \lambda_t)\) and the global variables are unchanged. We then use the mean field assumption \(q(w, \bar{\pi}, \Phi, H_t) = q(w)q(\bar{\pi})q(\Phi)q(H_t)\) and compute the iterative steps as follows.

1. **Global Update**: The same as online MedHDP, except that the update rule for \(w\) is now (28) for the Gibbs classifier.

2. **Local Update**: This step involves drawing samples of the local variables. We develop a Gibbs sampler, which iteratively draws \(S_t\) from the local conditional in (41), draws \(\bar{\pi}\) from the conditional in (42), and draws the augmented variables \(\lambda_t\) from the conditional in (30). For \(Z_t\), we explain the sampling procedure in detail. Specifically, we infer \(Z_t\) through

\[
q^*(Z_t, S_t, \lambda_t, \bar{\pi}) \propto \hat{q}(\bar{\pi}, Z_t, S_t)q(Z_t, \lambda_t),
\]

where

\[
\hat{q}(Z_t, \lambda_t) = \prod_{d \in B_t} \frac{1}{\sqrt{2\pi\lambda_d}} \exp \left( -\frac{\sum_{i \in [n_d]} \Lambda_{z_{di}, x_{di}} - \mathbb{E}_{\Phi, w} \left[ \frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d} \right]}{\lambda_d} \right),
\]

The Gibbs sampling for each variable \(z_{di}\) is

\[
q(z_{di} = k | Z_t^{-di}, \lambda_t, \bar{\pi}) \propto \frac{\alpha\pi_k + \sum_{d \in B_t} \Delta_{kz_{di}}^t}{\sum_{w} (C_{kw}^{-di} + \Delta_{kw}^t)} \exp \left( \frac{-c\mu_k^* + \sum_{i \in [n_d]} \Lambda_{z_{di}, x_{di}} - \mathbb{E}_{\Phi, w} \left[ \frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d} \right]}{n_d\lambda_d} \right),
\]

while the probability of sampling a new topic is

\[
q(z_{di} > K | Z_t^{-di}, \bar{\pi}) \propto \alpha \left( 1 - \sum_{k=1}^{K} \pi_k \right) / W.
\]

Again, we iterate the above steps till convergence for training and the testing is the same as online MedHDP. A batch version algorithm can be attained by setting \(T = 1\) and \(|B| = D\), which we denote as MedHDP\textsuperscript{gibbs}.
6.2 Multi-task Learning

The above models have been presented for classification. The basic ideas can be applied to solve other learning tasks, such as regression and multi-task learning (MTL). We use multi-task learning as an example. The primary assumption of multi-task learning is that by sharing statistical strength in a joint learning procedure, multiple related tasks can be mutually enhanced or some main tasks can be improved. MTL has many applications. We consider one scenario for multi-label classification. In this task, a set of binary classifiers are trained, each of which identifies whether a document \( x_d \) belongs to a specific category \( y_d^q \in \{+1, -1\} \). These binary classifiers are allowed to share common latent representations and therefore could be attained via a modified BayesPA update equation:

\[
\min_{q \in \mathcal{F}_t} L \left( q(w, \mathcal{M}, H_t) \right) + 2c \sum_{q=1}^{Q} \ell_q \left( q(w, \mathcal{M}, H_t); X_t, Y_t^\tau \right),
\]

where \( Q \) is the total number of tasks. We can then derive the multi-task version of Passive-Aggressive topic models, denoted by \( \text{paMedLDA}_{\text{ave}} \) and \( \text{paMedLDA}_{\text{gibbs}} \), in a way similar as in Section 5. We can further develop the nonparametric multi-task MedLDA topic models in a way similar as in Section 6.1 and the online PA learning algorithms. We denote the nonparametric online models by \( \text{paMedHDP}_{\text{ave}} \) and \( \text{paMedHDP}_{\text{gibbs}} \), according to whether the task-specific classifier is averaging or Gibbs.

7. Experiments

We demonstrate the efficiency and prediction accuracy of online MedLDA, online Gibbs MedLDA and their extensions on the 20Newsgroup (20NG) and a large Wikipedia dataset. A sensitivity analysis of the key parameters is also provided. Following the same setting in Zhu et al. (2012), we remove a standard list of stop words. All of the experiments are done on a normal computer with single-core clock rate up to 2.4 GHz.

7.1 Classification on 20Newsgroup

We perform multi-class classification on the entire 20NG dataset with all the 20 categories. The training set contains 11,269 documents, with the smallest category having 376 documents and the biggest category having 599 documents. The testing set contains 7,505 documents, with the smallest and biggest categories having 259 and 399 documents respectively. We adopt the “one-vs-all” strategy (Rifkin and Klautau, 2004) to combine binary classifiers for multi-class prediction tasks.

We use shorthand notations \( \text{paMedLDA}_{\text{ave}} \) and \( \text{paMedLDA}_{\text{gibbs}} \) for online MedLDA and online Gibbs MedLDA respectively. The corresponding batch learning algorithms are MedLDA (Zhu et al., 2012) and Gibbs MedLDA (\( \text{MedLDA}_{\text{gibbs}} \)) (Zhu et al., 2014b), which is a MedLDA model with Gibbs classifiers. We use collapsed Gibbs sampling to solve MedLDA, which is exactly the gMedLDA model proposed by Jiang et al. (2012). We also choose a state-of-the-art online unsupervised topic model as the baseline, the sparse inference for LDA (spLDA) (Mimno et al., 2012), which has been demonstrated to be superior than the stochastic variational LDA (Hoffman et al., 2013) in prediction performance. To perform the supervised tasks, we learn a linear SVM with the topic representations using LIBSVM (Chang and Lin, 2011). We refer the readers to (Zhu et al., 2012) for the performance of other batch-learning-based supervised topic models, such as sLDA.
(Blei and McAuliffe 2010) and DiscLDA (Lacoste-Julien et al., 2008), as well as SVM classifiers on the raw inputs, which are generally either inferior in accuracy or slower in efficiency than MedLDA. For all the LDA-based topic models, we use symmetric Dirichlet priors \( \alpha = \frac{1}{K} \cdot 1 \) and \( \gamma = 0.45 \cdot 1 \). For BayesPA with Gibbs classifiers, the parameters were set at \( \epsilon = 164 \), \( c = 1 \), and \( \sigma^2 = 1 \). The models’ performance is not sensitive to the choice of these parameters in wide ranges as shown in Zhu et al. (2014b). For BayesPA with averaging classifiers, the parameters determined by cross validation are \( \epsilon = 16 \), \( c = 500 \), and \( \sigma^2 = 10^{-3} \). For reasons explained in section 7.3, we set the mini-batch size \( |B| = 1 \) for the averaging classifier and \( |B| = 512 \) for the Gibbs classifier.

We first analyze how many processed documents are sufficient for each model to converge. Figure 3 shows the test errors with respect to the number of passes through the 20NG training set, where the number of topics is set at \( K = 80 \) and the other parameters of BayesPA are set at
$(I, J, \beta) = (1, 2, 0)$. As we can observe, by solving a series of latent BayesPA problems, both paMedLDA^{ave} and paMedLDA^{gibbs} fully explore the redundancy of documents and converge in less than one pass, while their corresponding batch algorithms (i.e., MedLDA and MedLDA^{gibbs}) need many passes as burn-in steps. Besides, compared with the online learning algorithms for unsupervised topic models (i.e., spLDA+SVM), BayesPA topic models use supervising side information from each mini-batch, and therefore exhibit a faster convergence rate in discrimination ability. The convergence performance of BayesPA models is significantly better than that of the unsupervised spLDA.

Next, we study each model’s best performance possible and the corresponding training time. To allow for a fair comparison, we train each model until the relative change of its objective is less than $10^{-4}$. Figure 4 shows the prediction accuracy and training time of LDA-based models on the whole dataset with varying numbers of topics. We can see that BayesPA topic models are more than an order of magnitude faster than their corresponding batch algorithms in training time due to the power of online learning. paMedLDA^{ave} is faster than paMedLDA^{gibbs}, because it does not need to update the covariance matrix of classifier weights $w$. But the tradeoff is that for averaging models, they are more sensitive to the initial choice of $\sigma^2$, and therefore we need to use cross-validation to determine the best choice of variance beforehand. Furthermore, thanks to the merits of structured mean-field inference, which does not impose strict assumptions on the independence of latent variables, BayesPA topic models parallel their batch alternatives in accuracy. Moreover, all the supervised models are significantly better than the unsupervised spLDA in classification, reaching the state-of-the-art performance on the tested datasets.

Table 2 visualizes the learnt topic representation by paMedLDA^{ave} and paMedLDA^{gibbs}. For the displayed categories, we plot the corresponding classifier’s topic distribution averaged over the positive examples and top words from the topic matrix. As we can see, the average topic distributions become increasingly sparse as more and more data are observed. Eventually, the averaged topic distribution for each category contains only 1~2 non-zero entries and meanwhile different categories have quite diverse average topic distributions, therefore showing strong discriminative ability of the topic representations in distinguishing different categories. Such sparse and discriminative patterns are similar to what have been shown in batch settings (Zhu et al. 2012, 2014b).

7.2 Extensions

We now present the experimental results on the extensions of BayesPA topic models. We first present the results of nonparametric topic modeling on the same 20NG dataset. Then we demonstrate multi-task learning on a large Wikipedia dataset with more than 1 million documents and about 1 million unique terms.

7.2.1 Nonparametric Topic Modeling

Recall the nonparametric extensions of BayesPA topic models paMedHDP^{ave} and paMedHDP^{gibbs}. To validate the advantage of online learning, we test them against their corresponding batch algorithms (i.e., the models MedHDP^{ave} and MedHDP^{gibbs}) on the 20NG corpus. We also include an unsupervised baseline as the comparison model, the truncation-free HDP topic model (tfHDP) (Wang and Blei, 2012b), which is first used to discover the latent topic representations, and then combined with a linear SVM classifier for document categorization. For all HDP-based models, the following parameter setting is used: $\alpha = 5 \cdot 1$, $\gamma = 0.5 \cdot 1$ and $\eta = 1$. As an initial number of topic numbers...
Figure 4: Classification accuracy and running time of various models with respect to the number of topics on the 20NG dataset.

for HDP to start with, we choose $K = 20$. We observed that the training time and the prediction accuracy do not depend heavily on the initial number of topics. The other parameters of BayesPA are the same.

Figure 3 shows the convergence of $\text{paMedHDP}^{\text{ave}}$ and $\text{paMedHDP}^{\text{gibbs}}$, and Figure 4 plots the accuracy and time together with the inferred topic numbers, where the length of the horizontal bars represents the variance of the inferred topic numbers. The results are summarized from five different runs. As we can see, the nonparametric extensions of BayesPA topic models also dramatically improve time efficiency and converge to corresponding batch algorithms in prediction performance. Furthermore, the averaging models are again faster to train because they do not need to update the covariance matrix of classifier weights.
<table>
<thead>
<tr>
<th>Category</th>
<th>#Observation</th>
<th>Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>alt.altheism</td>
<td>1 64 512 4096 11269</td>
<td></td>
</tr>
<tr>
<td>Topic distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top words</td>
<td>T8 writes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T11 god</td>
<td></td>
</tr>
<tr>
<td></td>
<td>article</td>
<td></td>
</tr>
<tr>
<td></td>
<td>don people</td>
<td></td>
</tr>
<tr>
<td></td>
<td>time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>good</td>
<td></td>
</tr>
<tr>
<td>comp.graphics</td>
<td>1 64 512 4096 11269</td>
<td></td>
</tr>
<tr>
<td>Topic distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top words</td>
<td>T19 image</td>
<td></td>
</tr>
<tr>
<td></td>
<td>graphics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>file</td>
<td></td>
</tr>
<tr>
<td></td>
<td>jpeg files</td>
<td></td>
</tr>
<tr>
<td></td>
<td>software</td>
<td></td>
</tr>
<tr>
<td>misc.forsale</td>
<td>1 64 512 4096 11269</td>
<td></td>
</tr>
<tr>
<td>Topic distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top words</td>
<td>T4 sale</td>
<td></td>
</tr>
<tr>
<td></td>
<td>offer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>shipping</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mail dos</td>
<td></td>
</tr>
<tr>
<td></td>
<td>price</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>#Observation</th>
<th>Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>alt.altheism</td>
<td>1 64 512 4096 11269</td>
<td></td>
</tr>
<tr>
<td>Topic distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top words</td>
<td>T11 god</td>
<td></td>
</tr>
<tr>
<td></td>
<td>people</td>
<td></td>
</tr>
<tr>
<td></td>
<td>writes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>article</td>
<td></td>
</tr>
<tr>
<td></td>
<td>don</td>
<td></td>
</tr>
<tr>
<td></td>
<td>time</td>
<td></td>
</tr>
<tr>
<td>comp.graphics</td>
<td>1 64 512 4096 11269</td>
<td></td>
</tr>
<tr>
<td>Topic distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top words</td>
<td>T14 graphics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>image</td>
<td></td>
</tr>
<tr>
<td></td>
<td>file</td>
<td></td>
</tr>
<tr>
<td></td>
<td>program</td>
<td></td>
</tr>
<tr>
<td></td>
<td>images</td>
<td></td>
</tr>
<tr>
<td></td>
<td>writes</td>
<td></td>
</tr>
<tr>
<td>misc.forsale</td>
<td>1 64 512 4096 11269</td>
<td></td>
</tr>
<tr>
<td>Topic distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top words</td>
<td>T17 sale</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mail dos</td>
<td></td>
</tr>
<tr>
<td></td>
<td>good offer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>shipping</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Visualization of the learnt topics by paMedLDA$^\text{ave}$ and paMedLDA$^\text{gibbs}$. See text for details.
7.2.2 Multi-Task Classification

We test paMedLDAmt\textsuperscript{ave}, paMedLDAmt\textsuperscript{gibbs} and their nonparametric extensions on a large Wiki dataset. The Wiki dataset is built from the Wikipedia set used in PASCAL LSHC challenge 2012\textsuperscript{6}. The Wiki dataset is a collection of documents with labels up to 20 different kinds, while the data distribution among the labels is balanced. The training set consists of 1.1 millions of wikipedia documents and the testing test consists of 5,000 documents. The vocabulary contains 917,683 unique terms. To measure performance, we use F1 score, the harmonic mean of precision and recall.

As baseline batch algorithms, we include MedLDAmt, a recent multi-task extension of Gibbs MedLDA (Zhu et al., 2013). Since MedHDP is a new model, there is no existing implementation of multi-task batch versions. So we instead extended MedHDP to support multi-task inference. We call this model MedHDPmt.

We use the same validation scheme as previous to select batchsize $|B| = 1, c = 5000, \sigma^2 = 10^{-6}$ for paMedLDAmt\textsuperscript{ave}; We choose $|B| = 512, c = 1, \sigma^2 = 1$ for paMedLDAmt\textsuperscript{gibbs}. For both models, the Dirichlet parameters are $\alpha = 0.8 \cdot 1, \gamma = 0.5 \cdot 1$, and $\epsilon = 1$. We use $K = 40$ topics for MedLDA models. The nonparametric extensions use exactly the same parameter settings except that $\alpha = 5 \cdot 1, \eta = 1$ and we do not need to specify the topic number $K$.

Figure 5 shows the F1 scores of various models as a function of training time. We find that BayesPA topic models produce comparable results with their batch counterparts, but the training time is significantly less. With either Gibbs or averaging classifiers, BayesPA is about two orders of magnitude faster than their batch counterparts. Therefore, BayesPA topic models could potentially be applied to large-scale multi-class settings.

\textsuperscript{6} See http://lshtc.iit.demokritos.gr/.
Figure 6: Test errors of paMedLDA<sub>ave</sub> (left) and paMedLDA<sub>gibbs</sub> (right) with different batch sizes on the 20NG dataset.

7.3 Sensitivity Analysis

We provide further discussions on BayesPA learning for topic models. We analyze the models’ sensitivity to some key parameters.

Batch Size $|B|$: Figure 6 presents the test errors of BayesPA topic models (paMedLDA<sub>ave</sub>, paMedLDA<sub>gibbs</sub>) as a function of training time on the entire 20NG dataset with various batch sizes. The number of topics is fixed at $K = 40$. We can see that the convergence speeds of different algorithms vary. First of all, the batch algorithms suffer from multiple passes through the dataset and therefore are much slower than the online alternatives. Second, we could observe that algorithms with medium batch sizes ($|B| = 64$ or $256$) converge faster. If we choose a batch size too small, for example, $|B| = 1$, each iteration would not provide sufficient evidence for the update of global variables; if the batch size is too large, each mini-batch becomes redundant and the convergence rate also decreases. By comparing the two figures, we find that paMedLDA<sub>ave</sub> runs faster than paMedLDA<sub>gibbs</sub>. This is because for averaging classifiers, we do not update the covariance of the classifier weights, which requires frequent matrix inverse operations. Furthermore, paMedLDA<sub>ave</sub> appears to be more robust against change in batchsize. Similarly, Figure 7 shows the sensitivity experiment of the batchsize parameter in paMedHDP models. The results are similar to paMedLDA models, that is, a moderate batchsize leads to faster convergence.

Number of iterations $I$ and samples $J$: Since the time complexity of Algorithm 2 is linear in both $I$ and $J$, we would like to know how these parameters influence the quality of the trained models. First, we analyze which setting of $(I, J)$ guarantees good performance. Fix $\beta = 0, K = 40$. Figure 8 presents the results. First, the number of samples $J$ does not have a large effect on the accuracy. Second, the performance of paMedLDA<sub>gibbs</sub> and paMedHDP<sub>gibbs</sub> is not sensitive the number of optimization round, but paMedLDA<sub>ave</sub> and paMedHDP<sub>ave</sub> suffers largely if $I = 1$. This is because with averaging classifiers, the sampling of latent variable $Z$ relies not only on global
parameters, but also on a local variable $\tau$, so more optimization rounds are needed. The training time of all models scale linearly in terms of $I$ and $J$.

\[
\begin{array}{ccc}
J & \beta & 0 & 2 & 4 \\
1 & 0.802 & & & \\
3 & 0.803 & 0.803 & & \\
5 & 0.805 & 0.805 & 0.803 & \\
\end{array}
\]

(a)

\[
\begin{array}{ccc}
J & \beta & 0 & 2 & 4 \\
1 & 0.783 & & & \\
3 & 0.803 & 0.799 & & \\
5 & 0.808 & 0.803 & 0.792 & \\
\end{array}
\]

(b)

Table 3: Effect of the number of local samples and burn-in steps for (a). paMedLDA$^\text{ave}$; and (b). paMedLDA$^\text{gibbs}$.

Notice that the first $\beta$ samples are discarded as burn-in steps. To understand how large $\beta$ is sufficient, we consider the settings of the pairs $(J, \beta)$ and check the prediction accuracy of Algorithm 2 for $K = 40$. Based on the sensitivity analysis of $I$ and $J$, we fix $I = 1$ for paMedLDA$^\text{gibbs}$, paMedHDP$^\text{gibbs}$ and $I = 2$ for paMedLDA$^\text{ave}$, paMedHDP$^\text{ave}$. The results are shown in Table 3. We can see that accuracy scores closer to the diagonal of the table are relatively lower, while settings with the same number of kept samples, e.g. $(J, \beta) = (3, 0), (5, 2),(9, 6)$, yield similar results. Therefore, the number of kept samples exhibits a more significant role in the performance of BayesPA topic models than the number of burn-in steps.

**Number of Topics.** Finally, we observed an accuracy degradation phenomenon when running one-pass BayesPA with a large number of topics on small datasets. To see this effect, we test on a small binary classification dataset, which consists of the documents from the two groups alt.atheism and talk.religion.misc in the 20 newsgroup dataset. The training set contains 1,597
documents with a split of 795/802 over the two categories and the test set contains 401 documents with a split of 204/197. For clarity, we report the results of the online BayesPA with a Gibbs MedLDA classifier, whose hyper-parameters are set as in the multi-class setting.

As shown in Figure 9 (the red curve), when the number of topics increases to relatively large, the classification accuracy drops significantly for Gibbs MedLDA with a single pass of the dataset. We believe this phenomenon is due to the fact that with a limited number of data points and a large number of parameters, the online BayesPA fails to converge in just one pass. Two possible strategies can be used to make this algorithm practically useful:

- **Multiple Scans**: In this strategy, we simply scan the dataset for multiple passes. As shown in Figure 9 this method works well — with five passes, BayesPA can learn the topic model with more than 100 topics well and lead to improved accuracy when $K$ is in the range of $[50, 100]$. However, one disadvantage of this strategy is that it takes more computational resources, linear to the number of scans.

- **Ghost Samples**: In this strategy, we create ghost copies of the data, that is, for every data point $(x_d, y_d)$, we make $L - 1$ exact clones. Then in our global update equation, each aggregation
8. Conclusions and Future Work

We present online Bayesian Passive-Aggressive (BayesPA) learning as a new framework for max-margin Bayesian inference on streaming data. For fixed but large-scale datasets, online BayesPA effectively explores the statistical redundancy by repeatedly drawing samples and leads to faster convergence. We show that BayesPA subsumes the online PA, and more significantly generalizes naturally to incorporate latent variables and to perform nonparametric Bayesian inference, therefore providing great flexibility for explorative analysis. Based on the ideas of BayesPA, we develop efficient online learning algorithms for max-margin topic models as well as their nonparametric extensions which can automatically infer the unknown topic numbers. Empirical experiments on 20Newsgroups and a large-scale Wikipedia multi-label dataset demonstrate significant improvements on time efficiency, while maintaining comparable results.

As for future work, we are interested in developing highly scalable, distributed (Broderick et al., 2013) BayesPA learning paradigms, which will better meet the demand of processing massive real data available today. We are also interested in applying BayesPA to develop efficient algorithms for more sophisticated max-margin Bayesian models, such as the latent feature relational models (Zhu, 2012). Finally, a theoretical analysis of the regret bounds for BayesPA needs to be systematically carried out for both averaging and Gibbs classifiers.
Acknowledgements

This work is supported by National Key Foundation R&D Projects (No. 2013CB329403), the National Natural Science Foundation of China (NSFC) (Nos. 61620106010, 61322308, 61332007), the NSFC and German Research Foundation (DFG) in Project Crossmodal Learning (No. NSFC 61621136008 / DFG TRR-169), and the National Youth Top-notch Talent Support Program. J.Z is also partly supported by the collaborative projects from Intel and Tencent.

Appendix A.

We show the objective in (27) is an upper bound of that in (13), that is,

\[
L(q(w, \Phi, Z_t, \lambda_t)) - E_q[\log(\psi(Y_t, \lambda_t|Z_t, w))] \geq L(q(w, \Phi, Z_t)) + 2c \sum_{d \in B_t} E_q[(\xi_d)_+],
\]

where \(L(q) = KL[q||q_t(w, \Phi)q_0(Z_t)]\).

Proof We first have

\[
L(q(w, \Phi, Z_t, \lambda_t)) = E_q\left[\log \frac{q(\lambda_t|w, \Phi, Z_t)q(w, \Phi, Z_t)}{q_t(w, \Phi, Z_t)}\right],
\]

and

\[
L(q(w, \Phi, Z_t)) = E_q\left[\log \frac{q(w, \Phi, Z_t)}{q_t(w, \Phi, Z_t)}\right].
\]

Comparing these two equations and canceling out common factors, we know that in order for (45) to make sense, it suffices to prove

\[
\mathbb{H}[q] - E_q'[\log(\psi(Y_t, \lambda_t|Z_t, w))] \geq 2c \sum_{d \in B_t} E_q'[\xi_d]_+ \tag{46}
\]

is uniformly true for any given \((w, \Phi, Z_t)\), where \(\mathbb{H}(\cdot)\) is the entropy operator and \(q' = q(\lambda_t|w, \Phi, Z_t)\). The inequality (46) can be reformulated as

\[
E_{q'}\left[\log \frac{q'(Y_t, \lambda_t|Z_t, w)}{\psi(Y_t, \lambda_t|Z_t, w)}\right] \geq 2c \sum_{d \in B_t} E_q'[(\xi_d)_+] \tag{47}
\]

Exploiting the convexity of the function \(\log(\cdot)\), i.e.

\[-E_{q'}\left[\log \frac{\psi(Y_t, \lambda_t|Z_t, w)}{q'(Y_t, \lambda_t|Z_t, w)}\right] \geq -\log \int_{\lambda_t} \psi(Y_t, \lambda_t|Z_t, w) d\lambda_t,
\]

and utilizing the equality (26), we then have (47) and therefore prove (45).

References


