Journal of Machine Learning Research 1 (2014) 1-48 Submitted 4/00; Published 10/00

Online Bayesian Passive-Aggressive Learning

Tianlin Shi TIANLINSHI@GMAIL.COM
Institute for Interdisciplinary Information Sciences
Tsinghua University
Beijing, 100084 China

Jun Zhu DCSZJ@MAIL.TSINGHUA.EDU.CN
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

Editor:

Abstract

We present online Bayesian Passive-Aggressive (BayesPA) learning, a generic online learning framework for hierarchical Bayesian models with max-margin posterior regularization. We provide provable Bayesian regret bounds for both averaging classifiers and Gibbs classifiers. We show that BayesPA subsumes the standard online Passive-Aggressive (PA) learning and more importantly 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 resolve the unknown number of topics. Experimental results on 20newsgroups and a large Wikipedia multi-label data set (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 maintaining comparable results with the batch counterpart methods.

1. Introduction

In the Big Data era, it is becoming a norm that massive data corpora need to efficiently handled in many application areas, while standard batch learning algorithms may fail. This has lead to the fast growing interests in developing scalable online or distributed learning algorithms. This paper focuses on online learing, 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 a fixed large-scale data set, 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 batch counterpart algorithms. Among the many popular algorithms, online Passive-Aggressive (PA) learning (Crammer et al., 2006) provides a generic framework of performing online 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 pre-

dictive tasks, existing online PA methods are formulated as a point estimate problem by optimizing some deterministic objective function. This may lead to some inconvenience. For example, a single large-margin model is often less than sufficient in describing complex data, such as those with rich underlying structures.

On the other hand, Bayesian methods enjoy the great 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 Monte Carlo (Welling and Teh, 2011; Ahn et al., 2012) methods, which repeatedly draw samples from a given finite data set. 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 are lack of the discriminative ability of large-margin methods and are usually less than sufficient in performing discriminative tasks.

Successful attempts have been made to bring large-margin learning and Bayesian methods together. For example, maximum entropy discrimination (MED) (Jaakkola et al., 1999) made a significant advance in conjoining max-margin learning and Bayesian generative models, in the context of supervised learning 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., 2012, 2013), social link prediction (Zhu, 2012), and multi-task learning (Jebara, 2011; Zhu et al., 2011). Regularized Bayesian inference (RegBayes) (Zhu et al., 2014b) provides a unified framework for Bayesian models on performing max-margin learning, where the max-margin principle is incorporated through imposing posterior constraints to an information-theoretical optimization problem. 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., 2014a), 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 corpura either.

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 characterize the performance of BayesPA by providing regret bounds, for both the case when using an averaging classifier and

the case when using a Gibbs classifier. 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 data sets demonstrate significant improvements on time efficiency and maintenance of comparable results with the batch counterparts.

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. Section 5 presents the regret bounds. Section 6 presents the concrete instantiations on topic modeling, and Section 7 presents the extensions to nonparametric topic models and multi-task learning. Section 8 presents experimental results. Finally, Section 9 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 classifications 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 now widely applied to various fields of study (Arora et al., 2012).

As a member of the family of weight updating methods, online Passive-Aggressive (PA) algorithms provide a generic online learning framework for max-margin models, first presented by Crammer et al. (2006). 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 nature 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, 2012b).

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. For example, through Rao-Blackwellized particle filters (Doucet et al., 2000), one could obtain online inference algorithms for latent Dirichlet allocation (Canini et al., 2009). 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 data set, in many applications data arrives in stream so that the data set 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 data set 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). In the sequel, 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 data. MedLDA (Zhu et al., 2012) extended MED to infer latent topical structure from data with large margin constraints on the target posteriors. Similarly, nonparametric Bayesian max-margin models have also been developed, such as infinite SVMs (iSVM) (Zhu et al., 2011) for building SVM classifiers with latent mixture structure, and infinite latent SVMs (iLSVM) (Zhu et al., 2011) for automatic discovering predictive features for SVMs. Furthermore, the idea of nonparametric Bayesian learning has been widely applied to link prediction (Zhu, 2012), matrix factorization (Xu et al., 2012), etc. Regularized Bayesian inference (RegBayes) (Zhu et al., 2014b)

provides a unified framework for performing max-margin learning of (nonparametric) Bayesian models, where the max-margin principle was incorporated through imposing posterior constraints to a variational formulation of the standard Bayes' rule.

Max-margin Bayesian learning in the batch mode has already been one of the common challenges facing this class of models. Despite its general intractability, efficient algorithms have been proposed under different settings. One way is to solve the problem via variational inference under a mean-field (Zhu et al., 2012) or structured mean-field (Jiang et al., 2012) assumption. Recently, Zhu et al. (2014a) provided a key insight in deriving efficient Monte Carlo methods without making strict assumptions. Their technique is based on a data augmentation formulation of the expected margin loss. Based on similar techniques, fast inference algorithms have also been developed for generalized relational topic models (Chen et al., 2013), matrix factorization (Xu et al., 2013), etc. Data augmentation (DA) refers to the method of introducing augmented variables along with the observed data to make their joint distribution tractable. The technique was popularized in the statistics community by the well-known expectation-maximization algorithm (EM) (Dempster et al., 1977) for maximum likelihood estimation with missing data. For posterior inference, the technique is popularized by Tanner and Wong (1987) in statistics and by Swendsen and Wang (1987) for the Ising and Potts models in physics. For a broader introduction to DA methods, we refer the readers to Van Dyk and Meng (2001).

Finally, our conference version of the paper (Shi and Zhu, 2014) has introduced some preliminary work, which would be largely extended.

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 $\boldsymbol{w} \in \mathbb{R}^K$ (e.g., the weights of a linear SVM) in an online manner with the instantaneous losses from arriving data $\{\boldsymbol{x}_t\}_{t\geq 0}$ ($\boldsymbol{x}_t \in \mathbb{R}^K$) and corresponding responses $\{y_t\}_{t\geq 0}$. The losses $\ell_\epsilon(\boldsymbol{w}; \boldsymbol{x}_t, y_t)$, as they consider, could be the hinge loss $(\epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t)_+$ for binary classification ($y_t \in \{0, 1\}$) or the ϵ-insensitive loss $(|y_t - \boldsymbol{w}^\top \boldsymbol{x}_t| - \epsilon)_+$ for regression ($y_t \in \mathbb{R}$), where ϵ is a hyper-parameter and $(x)_+ = \max(0, x)$. The online Passive-Aggressive update rule is then derived by defining the new weight $\boldsymbol{w}_{t+1}$ as the solution to the following optimization problem:

$$\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w} - \boldsymbol{w_t}||^2 \quad \text{s.t.: } \ell_\epsilon(\boldsymbol{w}; \boldsymbol{x}_t, y_t) = 0, \tag{1}$$

where $\|\cdot\|^2$ is the Euclidean 2-norm. Intuitively, if $\boldsymbol{w_t}$ suffers no loss from the new data, i.e., $\ell_\epsilon(\boldsymbol{w}_t; \boldsymbol{x}_t, y_t) = 0$, the algorithm *passively* assigns $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t$; otherwise, it aggressively projects $\boldsymbol{w_t}$ to the feasible zone 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 $\boldsymbol{w}^*$, 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

$$\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w} - \boldsymbol{w_t}||^2 + 2c \cdot \ell_\epsilon(\boldsymbol{w}; \boldsymbol{x}_t, y_t), \tag{2}$$

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 is $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \tau y_t \boldsymbol{x}_t$, where $\tau_t = \min(2c, \max(0, \epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t)/||\boldsymbol{x}_t||^2)$; and for the ϵ-insensitive loss, the update rule is $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \text{sign}(y_t - \boldsymbol{w}^\top \boldsymbol{x}_t)\tau_t \boldsymbol{x}_t$, where $\tau_t = \max(0, \epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t)/||\boldsymbol{x}_t||^2$.

3.2 Streaming Variational Bayes

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

$$p(\boldsymbol{w}|\{\boldsymbol{x}_t\}_{t=0}^T) \propto p(\boldsymbol{w}|\{\boldsymbol{x}_t\}_{t=0}^{T-1}) p(\boldsymbol{x}_T|\boldsymbol{w}).$$

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 $q(\boldsymbol{w}) = \mathcal{A}(\boldsymbol{X}, q_0(\boldsymbol{w}))$ based on data $\boldsymbol{X}$ and prior $q_0(\boldsymbol{w})$. Then, the recursive formula for approximate streaming update is:

$$q(\boldsymbol{w}|\{\boldsymbol{x}_t\}_{t=0}^T) = \mathcal{A}\left(\boldsymbol{x}_T, q(\boldsymbol{w}|\{\boldsymbol{x}_t\}_{t=0}^{T-1})\right).$$

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., 2014b) is to treat Bayesian inference as an optimization problem with an imposed loss function. Mathematically, RegBayes can be formulated as

$$\min_{q\in\mathcal{P}} \ \mathbf{KL}\Big[q(\boldsymbol{w})||p_0(\boldsymbol{w})\Big] - \mathbb{E}_{q(\boldsymbol{w})}\Big[\log p(\boldsymbol{X}|\boldsymbol{w})\Big] + 2c \cdot \ell\Big(q(\boldsymbol{w}); \boldsymbol{X}\Big), \tag{3}$$

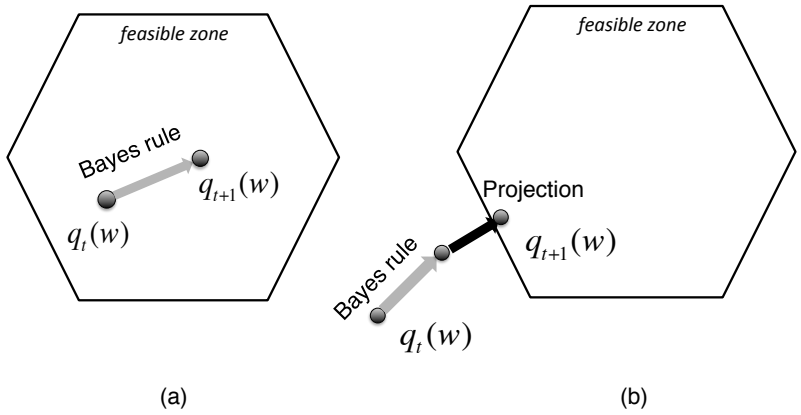

Figure 1: Graphical Illustration of BayesPA learning. **(a)**. Update *passively* by Bayes rule, if the resulting distribution suffer zero loss. **(b)** Otherwise, *aggressively* project the resulting distribution to the feasible zone of weights with zero loss.

where $\mathcal{P}$ is the probability simplex, $p(\boldsymbol{X}|\boldsymbol{w})$ is the likelihood function and $\mathbf{KL}$ is the Kullback-Leibler divergence.[1] Note that if $\ell(q(\boldsymbol{w}); \boldsymbol{X}) = 0$, then the optimal solution $q^*(\boldsymbol{w}) \propto p_0(\boldsymbol{w})p(\boldsymbol{X}|\boldsymbol{w})$, which is just the Bayes' rule. However, when $\ell(q(\boldsymbol{w}); \boldsymbol{X}) \neq 0$, RegBayes biases the inferred posterior towards discriminating the supervising side-information, with the parameter c controlling the extent of regularization. If the posterior regularization term $\ell(\cdot)$ is a convex function of $q(\boldsymbol{w})$ through the linear expectation operator, Zhu et al. (2014b) presented a general representation theorem to characterize the solution to problem (3). To distinguish from the posterior obtained via Bayes' rule, the solution to problem (3) is called *post-data posterior* (Zhu et al., 2014b). 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., 2014a), matrix factorization (Xu et al., 2012, 2013), link prediction (Zhu, 2012), etc.

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 for other learning tasks. We provide an extension in Section 7.2.

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

4.1 Online BayesPA Learning

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

$$\begin{aligned} \min_{q(\boldsymbol{w})\in\mathcal{F}_t} \ & \mathbf{KL}\Big[q(\boldsymbol{w})||q_t(\boldsymbol{w})\Big] - \mathbb{E}_{q(\boldsymbol{w})}\Big[\log p(\boldsymbol{x}_t|\boldsymbol{w})\Big] \\ \text{s.t.:} \ & \ell_\epsilon\Big(q(\boldsymbol{w}); \boldsymbol{x}_t, y_t\Big) = 0, \end{aligned} \tag{4}$$

where $\mathcal{F}_t$ can be some family of distributions or the probability simplex $\mathcal{P}$. In other words, we find a post-data posterior distribution $q_{t+1}(\boldsymbol{w})$ in the feasible zone that is not only close to $q_t(\boldsymbol{w})$ in terms of KL-divergence, but also has a high likelihood of explaining new data. As a result, if Bayes' rule already gives the posterior distribution $q_{t+1}(\boldsymbol{w}) \propto q_t(\boldsymbol{w})p(\boldsymbol{x}_t|\boldsymbol{w})$ 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 zone 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., $p(\boldsymbol{x}_t|\boldsymbol{w})$ is independent of $\boldsymbol{w}$), BayesPA will passively set $q_{t+1}(\boldsymbol{w}) = q_t(\boldsymbol{w})$ if $q_t(\boldsymbol{w})$ suffers no loss; otherwise, it will aggressively project $q_t(\boldsymbol{w})$ to the feasible zone. We call it *non-likelihood* BayesPA.

In practical problems, the constraints in (4) 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 $\mathcal{L}(q(\boldsymbol{w}))$ in problem (4) with a regularization term (Cortes and Vapnik, 1995):

$$q_{t+1}(\boldsymbol{w}) = \operatorname*{argmin}_{q(\boldsymbol{w})\in\mathcal{F}_t} \mathcal{L}\Big(q(\boldsymbol{w})\Big) + 2c \cdot \ell_\epsilon\Big(q(\boldsymbol{w}); \boldsymbol{x}_t, y_t\Big). \tag{5}$$

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

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

$$\ell_\epsilon^{\text{ave}}\Big(q(\boldsymbol{w}); \boldsymbol{x}_t, y_t\Big) = \Big(\epsilon - y_t \mathbb{E}_{q(\boldsymbol{w})}\Big[\boldsymbol{w}^\top \boldsymbol{x}_t\Big]\Big)_+.$$

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

$$\ell_\epsilon^{\text{gibbs}}\Big(q(\boldsymbol{w}); \boldsymbol{x}_t, y_t\Big) = \mathbb{E}_{q(\boldsymbol{w})}\Big[\Big(\epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t\Big)_+\Big].$$

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

Methods	Max-margin learning ?	Bayesian inference ?	Streaming update ?
PA	yes	no	yes
SVB	no	yes	yes
RegBayes	yes	yes	no
BayesPA	yes	yes	yes

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.

Lemma 1 *The expected hinge loss ℓ_ϵ^{gibbs} is an upper bound of the hinge loss ℓ_ϵ^{ave}, that is,*

$$\ell_\epsilon^{gibbs}\Big(q(\boldsymbol{w}); \boldsymbol{x}_t, y_t\Big) \geq \ell_\epsilon^{ave}\Big(q(\boldsymbol{w}); \boldsymbol{x}_t, y_t\Big).$$

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. 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 $q_0(\boldsymbol{w}) = \mathcal{N}(0, I)$, $\mathcal{F}_t = \mathcal{P}$ and we use the averaging classifier ℓ_ϵ^{ave}, the non-likelihood BayesPA subsumes the online PA.*

Proof The soft-margin version of BayesPA learning can be reformulated using a slack variable ξ_t:

$$\begin{aligned} q_{t+1}(\boldsymbol{w}) = \quad & \underset{q(\boldsymbol{w})\in\mathcal{P}}{\operatorname{argmin}} \ \mathrm{KL}\Big[q(\boldsymbol{w})||q_t(\boldsymbol{w})\Big] + 2c\cdot\xi_t \\ & \text{s.t.}: y_t\mathbb{E}_{q(\boldsymbol{w})}\left[\boldsymbol{w}^\top\boldsymbol{x}_t\right] \geq \epsilon - \xi_t, \ \ \xi_t \geq 0. \end{aligned} \tag{6}$$

Similar to Corollary 5 in Zhu et al. (2012), the optimal solution $q^*(\boldsymbol{w})$ of the above problem can be derived from its functional Lagrangian and has the following form:

$$q^*(\boldsymbol{w}) = \frac{1}{\Gamma(\tau_t^*, \boldsymbol{x}_t, y_t)} q_t(\boldsymbol{w}) \exp\Big(\tau_t^* y_t \boldsymbol{w}^\top \boldsymbol{x}_t\Big), \tag{7}$$

where the normalization term $\Gamma(\tau_t, \boldsymbol{x}_t, y_t) = \int_{\boldsymbol{w}} q_t(\boldsymbol{w}) \exp\Big(\tau_t y_t \boldsymbol{w}^\top \boldsymbol{x}_t\Big) d\boldsymbol{w}$, and τ_t^* is the optimal solution to the dual problem

$$\begin{aligned} & \max_{\tau_t} \ \epsilon\tau_t - \log\Gamma(\tau_t, \boldsymbol{x}_t, y_t) \\ & \text{s.t. } 0 \leq \tau_t \leq 2c. \end{aligned} \tag{8}$$

Using this primal-dual interpretation, we prove that for the normal prior $p_0(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}_0, I)$, the post-data posterior is also Gaussian: $q_t(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}_t, I)$ for some $\boldsymbol{\mu}_t$ in each round $t = 0, 1, 2, ...$ This can be shown by induction. By our assumption, $q_0(\boldsymbol{w}) = p_0(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}_0, I)$ is Gaussian. Assume for round $t \geq 0$, the distribution $q_t(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}_t, I)$. Then for round $t+1$, Eq. (7) suggests the distribution

$$q_{t+1}(\boldsymbol{w}) = \frac{\mathcal{C}}{\Gamma(\tau_t^*, \boldsymbol{x}_t, y_t)} \exp\left(-\frac{1}{2}||\boldsymbol{w} - (\boldsymbol{\mu}_t + \tau_t^* y_t \boldsymbol{x}_t)||^2\right),$$

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

Next, we show that $\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \tau_t^* y_t\boldsymbol{x}_t$ is the optimal solution of the online PA update rule (Crammer et al., 2006). To see this, we replace $\Gamma(\tau_t, \boldsymbol{x}_t, y_t)$ in problem (8) with our derived form. Ignoring constant terms, we obtain the dual problem

$$\begin{array}{ll}\max\limits_{\tau_t} & \epsilon\tau_t - \frac{1}{2}\tau_t^2\boldsymbol{x}_t^\top\boldsymbol{x}_t - y_t\tau_t\boldsymbol{\mu}_t^\top\boldsymbol{x}_t \\ \text{s.t.:} & 0 \le \tau_t \le 2c,\end{array} \tag{9}$$

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

$$\begin{array}{c}\boldsymbol{\mu}_{t+1}^{\text{PA}} = \arg\min\limits_{\boldsymbol{\mu}} \ \frac{1}{2}||\boldsymbol{\mu} - \boldsymbol{\mu}_t||^2 + 2c\cdot\xi_t \\ \text{s.t.} \ \ y_t\boldsymbol{\mu}^\top\boldsymbol{x}_t \ge \epsilon - \xi_t, \ \ \xi_t \ge 0.\end{array}$$

The optimal solution is $\boldsymbol{\mu}_{t+1}^{\text{PA}} = \boldsymbol{\mu}_t + \tau_t^* y_t\boldsymbol{x}_t$. Note that τ_t^* is the optimal solution of dual problem (9) shared by both PA and BayesPA. Therefore, we conclude that $\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_{t+1}^{\text{PA}}$. ■

Second, suppose some algorithm $\mathcal{A}$ is capable of solving problem (5), 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 be presented in the following lemma.

Lemma 3 *If $\mathcal{F}_t = \mathcal{P}$ and we use the averaging classifier with loss functional ℓ_ϵ^{ave}, the update rule of online BayesPA is*

$$q_{t+1}(\boldsymbol{w}) = \frac{1}{\Gamma(\tau_t^*, \boldsymbol{x}_t, y_t)} q_t(\boldsymbol{w})p(\boldsymbol{x}_t|\boldsymbol{w}) \exp\Big(\tau_t^* y_t\boldsymbol{w}^\top\boldsymbol{x}_t\Big), \tag{10}$$

where $\Gamma(\tau_t, \boldsymbol{x}_t, y_t)$ is the normalization term

$$\Gamma(\tau_t, \boldsymbol{x}_t, y_t) = \int_{\boldsymbol{w}} q_t(\boldsymbol{w})p(\boldsymbol{x}_t|\boldsymbol{w}) \exp\Big(\tau_t y_t\boldsymbol{w}_t^\top\boldsymbol{x}_t\Big) d\boldsymbol{w}$$

and τ_t^* is the optimal solution to the dual problem

$$\begin{array}{ll}\max\limits_{\tau_t} & \epsilon\tau_t - \log\Gamma(\tau_t, \boldsymbol{x}_t, y_t) \\ \text{s.t.} & 0 \le \tau_t \le 2c.\end{array}$$

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 ℓ_ϵ^{gibbs}, the update rule of online BayesPA is*

$$q_{t+1}(\boldsymbol{w}) = \frac{q_t(\boldsymbol{w})p(\boldsymbol{x}_t|\boldsymbol{w}) \exp\Big(-2c\left(\epsilon - y_t\boldsymbol{w}^\top\boldsymbol{x}_t\right)_+\Big)}{\Gamma(\boldsymbol{x}_t, y_t)}, \tag{11}$$

where $\Gamma(\boldsymbol{x}_t, y_t)$ is the normalization constant.

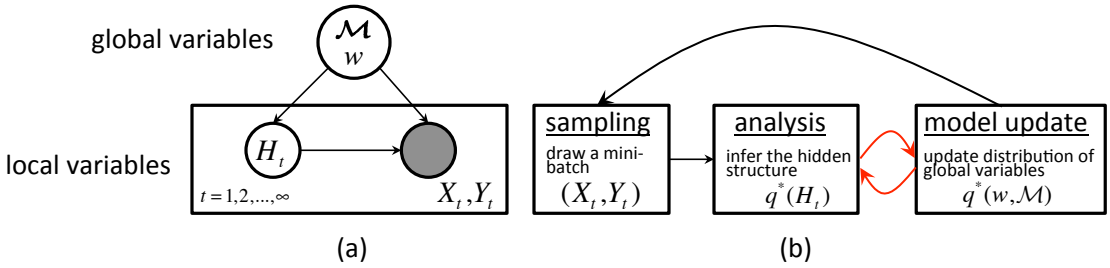

Figure 2: Graphical illustrations of: (a) the abstraction of models with latent structures; and (b) the procedure of BayesPA learning with latent structures.

In both update rules in Eq. (10) and Eq. (11), the post-data posterior $q_t(\boldsymbol{w})$ 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. Note that if there is no loss functional (i.e., $\ell_\epsilon = 0$), both Eq. (10) and Eq. (11) reduce to the streaming Bayesian update problem. Therefore, BayesPA is an extension to streaming variational Bayes (SVB) with imposed max-margin posterior constraints.

Finally, the update formulation (5) is essentially a RegBayes problem with a single data point $(\boldsymbol{x}_t, y_t)$. 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 $\boldsymbol{X}_t = \{\boldsymbol{x}_d\}_{d \in B_t}, \boldsymbol{Y}_t = \{y_d\}_{d \in B_t}$. The online BayesPA update equation for this mini-batch can be defined in a natural way:

$$\min_{q \in \mathcal{F}_t} \mathbf{KL}\Big[q(\boldsymbol{w}) || q_t(\boldsymbol{w})\Big] - \mathbb{E}_{q(\boldsymbol{w})}\Big[\log p(\boldsymbol{X}_t | \boldsymbol{w})\Big] + 2c \cdot \ell_\epsilon\Big(q(\boldsymbol{w}); \boldsymbol{X}_t, \boldsymbol{Y}_t\Big),$$

where $\ell_\epsilon(q(\boldsymbol{w}); \boldsymbol{X}_t, \boldsymbol{Y}_t) = \sum_{d \in B_t} \ell_\epsilon(q(\boldsymbol{w}); \boldsymbol{x}_d, y_d)$. Like 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-word 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* $\boldsymbol{h}_d$ $(d \geq 0)$ that characterize the hidden structures of each observed data $\boldsymbol{x}_d$ and *global variables* $\boldsymbol{\mathcal{M}}$ 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 $\boldsymbol{\mathcal{M}}$ as well as the classifier weights $\boldsymbol{w}$, based on each incoming mini-batch $(\boldsymbol{X}_t, \boldsymbol{Y}_t)$ and their corresponding latent variables $\boldsymbol{H}_t = \{\boldsymbol{h}_d\}_{d \in B_t}$. Because of the uncertainty in $\boldsymbol{H}_t$, our posterior

approximation algorithm $\mathcal{A}$ would first infer the joint posterior distribution $q_{t+1}(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t)$ from

$$\min_{q \in \mathcal{F}_t} \mathcal{L}\Big(q(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t)\Big) + 2c \cdot \ell_\epsilon\Big(q(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t); \boldsymbol{X}_t, \boldsymbol{Y}_t\Big), \tag{12}$$

where $\mathcal{L}(q) = \text{KL}[q(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t)||q_t(\boldsymbol{w}, \boldsymbol{\mathcal{M}})p_0(\boldsymbol{H}_t)] - \mathbb{E}_{q(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t)}[\log p(\boldsymbol{X}_t|\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t)]$ and $\ell_\epsilon(q(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t); \boldsymbol{X}_t, \boldsymbol{Y}_t)$ is some cumulative loss functional on the min-batch data incurred by some classifiers on the latent variables $\boldsymbol{H}_t$ and/or global variables $\boldsymbol{\mathcal{M}}$. As in the case without latent variables, both averaging classifier and Gibbs classifier can be used.

In the sequel, algorithm $\mathcal{A}$ produces the approximate posterior $q_{t+1}(\boldsymbol{w}, \boldsymbol{\mathcal{M}})$. In general we would not obtain a closed-form posterior distribution by marginalizing out $\boldsymbol{H}_t$, especially in dealing with some involved models like MedLDA (Zhu et al., 2012). The intractability is bypassed through the mean-field assumption $q(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t) = q(\boldsymbol{w})q(\boldsymbol{\mathcal{M}})q(\boldsymbol{H}_t)$. Specifically, algorithm $\mathcal{A}$ solves problem (12) using an iterative procedure and obtain the optimal distribution $q^*(\boldsymbol{w})q^*(\boldsymbol{\mathcal{M}})q^*(\boldsymbol{H}_t)$. Then it sets $q_{t+1}(\boldsymbol{w}, \boldsymbol{\mathcal{M}}) = q^*(\boldsymbol{w})q^*(\boldsymbol{\mathcal{M}})$ and proceeds to next round. Concrete examples of this method will be discussed in Section 6 and Section 7.

5. Theoretical Analysis

In this section, we provide theoretical analysis for BayesPA learning. We consider the fully observed case where no latent structures are assumed, and leave the more complex case with hidden structures as future work. Specifically, we prove *regret bounds* (Murphy, 2012; Shalev-Shwartz and Singer, 2006), which relate the cumulative loss attained by our algorithms on *any* sequence of incoming samples to that by a fixed distribution of models $p(\boldsymbol{w})$. Such bounds guarantee that the loss ℓ_ϵ of the online learning algorithms cannot be too much larger compared to the loss ℓ_ϵ^* of any fixed predictor chosen with hindsight of all data.

In BayesPA learning, however, not only do we desire a model $\boldsymbol{w}$ with low cumulative loss, we also want $\boldsymbol{w}$ to have high cumulative likelihood. To capture this fact, we generalize the notion of regret as follows.

Definition 5 ***(Bayesian Regret)*** *The Bayesian Regret at observing $(\boldsymbol{x}_t, y_t)$ with the current model $q(\boldsymbol{w})$ is defined as*

$$\mathcal{R}_c(q(\boldsymbol{w}); \boldsymbol{x}_t, y_t) = -\mathbb{E}_{q(\boldsymbol{w})}[\log p(\boldsymbol{x}_t|\boldsymbol{w})] + 2c \cdot \ell_\epsilon(q(\boldsymbol{w}_t); \boldsymbol{x}_t, y_t) \tag{13}$$

where c is a parameter determining the trade-off between likelihood and loss, characterized by some loss function $\ell_\epsilon(q(\boldsymbol{w}_t); \boldsymbol{x}_t, y_t)$.

We will use the notation $\mathcal{R}_c^{\text{ave}}$ for the Bayesian regret if choosing the averaging loss $\ell_\epsilon^{\text{ave}}$ and use the notation $\mathcal{R}_c^{\text{gibbs}}$ if choosing the Gibbs loss $\ell_\epsilon^{\text{gibbs}}$. For non-likelihood BayesPA, the regret is naturally reduced to be $\mathcal{R}_c(q(\boldsymbol{w}); \boldsymbol{x}_t, y_t) = \ell_\epsilon(q(\boldsymbol{w}_t); \boldsymbol{x}_t, y_t)$. Our below analysis considers the full BayesPA. The main results are applicable for non-likelihood BayesPA, as remarked later.

Theorem 6 (A regret bound for BayesPA with Gibbs classifiers) *Let the initial prior be $q_0(\boldsymbol{w})$. For all $t \in \{0, 1, ..., T-1\}$, define the exponential family of tilted distributions*

$$q_{t,\tau}(\boldsymbol{w}) \propto q_t(\boldsymbol{w}) \exp(\tau \mathcal{T}(\boldsymbol{w}, \boldsymbol{x}_t, y_t))$$

with parameter τ and sufficient statistics

$$\mathcal{T}(\boldsymbol{w}, \boldsymbol{x}_t, y_t) = -\log p(\boldsymbol{x}_t|\boldsymbol{w}) + \frac{1}{2c}(\epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t)_+.$$

If the Fisher information of $\mathcal{T}(\boldsymbol{w}, \boldsymbol{x}_t, y_t)$ about τ satisfies

$$J_{\mathcal{T}(\boldsymbol{w},\boldsymbol{x}_t,y_t)}(\tau) = \mathbb{V}_{q_{t,\tau}(\boldsymbol{w})}\left[\mathcal{T}(\boldsymbol{w}, \boldsymbol{x}_t, y_t)\right] \leq R \tag{14}$$

for all parameters $0 < \tau < 2c$ and some constant $R > 0$. Then for any fixed distribution $p(\boldsymbol{w})$, the regret of BayesPA is bounded as

$$\sum_{t=0}^{T-1} \mathcal{R}_c^{\text{gibbs}}(q_t(\boldsymbol{w}); \boldsymbol{x}_t, y_t) \leq \sum_{t=0}^{T-1} \mathcal{R}_c^{\text{gibbs}}(p(\boldsymbol{w}); \boldsymbol{x}_t, y_t) + \mathbf{KL}[p(\boldsymbol{w})||p_0(\boldsymbol{w})] + 2c^2 RT. \tag{15}$$

Proof According to Lemma 4, the update rule for the distribution $q(\boldsymbol{w})$ is

$$q_{t+1}(\boldsymbol{w}) = \frac{1}{\Gamma(\boldsymbol{x}_t, y_t)} q_t(\boldsymbol{w}) p(\boldsymbol{x}_t|\boldsymbol{w}) e^{-2c\left(\epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t\right)_+},$$

where $\Gamma(\boldsymbol{x}_t, y_t)$ is the partition function

$$\Gamma(\boldsymbol{x}_t, y_t) = \int_{\boldsymbol{w}} q_t(\boldsymbol{w}) p(\boldsymbol{x}_t|\boldsymbol{w}) e^{-2c\left(\epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t\right)_+} d\boldsymbol{w}.$$

The proof idea is to relate the loss $\ell_\epsilon^{\text{gibbs}}(\boldsymbol{x}_t)$ in each round with the difference of prior $q_t(\boldsymbol{w})$ and posterior $q_{t+1}(\boldsymbol{w})$, which is

$$\mathbf{KL}[p(\boldsymbol{w})||q_t(\boldsymbol{w})] - \mathbf{KL}[p(\boldsymbol{w})||q_{t+1}(\boldsymbol{w})] = -\mathcal{R}_c^{\text{gibbs}}(p(\boldsymbol{w}); \boldsymbol{x}_t, y_t) - \log \Gamma(\boldsymbol{x}_t, y_t). \tag{16}$$

Construct a canonical exponential family $q_{t,\tau}(\boldsymbol{w})$ parameterized by τ through tilting the base distribution $q_t(\boldsymbol{w})$ as follows:

$$q_{t,\tau}(\boldsymbol{w}) = q_t(\boldsymbol{w}) \exp\Big(\frac{\tau}{2c}(\log p(\boldsymbol{x}_t|\boldsymbol{w}) - 2c(\epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t)_+) - \log f(\tau)\Big),$$

where the log partition function

$$\log f(\tau) = \log\left(\int_{\boldsymbol{w}} q_t(\boldsymbol{w}) \Big(p(\boldsymbol{x}_t|\boldsymbol{w})^{\frac{1}{2c}} e^{-(\epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t)_+}\Big)^\tau d\boldsymbol{w}\right).$$

According to properties of exponential family, the first-order and second-order derivatives can be related to its cumulants:

$$\frac{\partial}{\partial \tau} \log f(\tau) = \mathbb{E}_{q_{t,\tau}(\boldsymbol{w})}\left[\frac{1}{2c}\log p(\boldsymbol{x}_t|\boldsymbol{w}) - \Big(\epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t\Big)_+\right] = -\frac{1}{2c}\mathcal{R}_c^{gibbs}(q_{t,\tau}(\boldsymbol{w}); \boldsymbol{x}_t, y_t),$$

$$\frac{\partial^2}{\partial \tau^2} \log f(\tau) = \mathbb{V}_{q_{t,\tau}(\boldsymbol{w})}\left[\frac{1}{2c}\log p(\boldsymbol{x}_t|\boldsymbol{w}) - \Big(\epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t\Big)_+\right] = J_{\mathcal{T}(\boldsymbol{w},\boldsymbol{x}_t,y_t)}(\tau).$$

Using the fact that $\log f(0) = 0$ and applying Taylor's theorem with the Lagrange's remainder [2] at $\tau = 0$, we have a second-order expression

$$\log f(\tau) = -\frac{1}{2c}\mathcal{R}_c^{\text{gibbs}}(q_t(\boldsymbol{w}); \boldsymbol{x}_t, y_t)\tau + \frac{1}{2}J_{\mathcal{T}}(\hat{\tau})\tau^2,$$

for some $0 \leq \hat{\tau} \leq \tau$. By assumption, $J_{\mathcal{T}}(\hat{\tau}) \leq R$,

$$\log f(2c) \leq \mathcal{R}_c^{\text{gibbs}}(q_t(\boldsymbol{w}); \boldsymbol{x}_t, y_t) + 2c^2 R,$$

Since $\Gamma(\boldsymbol{x}_t, y_t) = f(2c)$, Eq. (16) can be lower bounded as

$$\mathbf{KL}[p(\boldsymbol{w})||q_t(\boldsymbol{w})] - \mathbf{KL}[p(\boldsymbol{w})||q_{t+1}(\boldsymbol{w})] \geq -\mathcal{R}_c^{\text{gibbs}}(p(\boldsymbol{w}); \boldsymbol{x}_t, y_t) + \mathcal{R}_c^{\text{gibbs}}(q_t(\boldsymbol{w}); \boldsymbol{x}_t, y_t) - 2c^2 R.$$

Summing over all $t = 0, 1, 2, ..., T-1$ and neglecting $\mathbf{KL}[p(\boldsymbol{w})||q_T(\boldsymbol{w})]$, we can obtain (15). ■

Theorem 7 (A regret bound for BayesPA with averaging classifiers) *Let the initial prior be $q_0(\boldsymbol{w})$. For all $t \in 0, 1, ..., T-1$, define the expoential family of tilted distributions,*

$$q_{t,\tau,u}(\boldsymbol{w}) \propto q_t(\boldsymbol{w}) \exp(u\mathcal{U}(\boldsymbol{w}, \boldsymbol{x}_t, y_t) + \tau\mathcal{T}(\boldsymbol{w}, \boldsymbol{x}_t, y_t))$$

with two parameters τ, u and the sufficient statistics,

$$\mathcal{U}(\boldsymbol{w}, \boldsymbol{x}_t, y_t) = \log p(\boldsymbol{x}_t|\boldsymbol{w}), \text{ and } \mathcal{T}(\boldsymbol{w}, \boldsymbol{x}_t, y_t) = y_t \boldsymbol{w}^\top \boldsymbol{x}_t.$$

If the Fisher information satisfies,

$$J_{\mathcal{U}(\boldsymbol{w}, \boldsymbol{x}_t, y_t)} = \mathbb{V}_{q_{t,\tau,u}}[\mathcal{U}(\boldsymbol{w}, \boldsymbol{x}_t, y_t)] \leq S$$

and

$$J_{\mathcal{T}}(\boldsymbol{w}, \boldsymbol{x}_t, y_t) = \mathbb{V}_{q_{t,\tau,u}}[\mathcal{T}(\boldsymbol{w}, \boldsymbol{x}_t, y_t)] \leq R$$

for all $(\tau, u) \in (0, 2c) \times (0, 1)$. Then for any fixed distribution $p(\boldsymbol{w})$, the regret of BayesPA satisfies,

$$\sum_{t=0}^{T-1} \mathcal{R}_c^{\text{ave}}(q_t(\boldsymbol{w}); \boldsymbol{x}_t, y_t) \leq \sum_{t=0}^{T-1} \mathcal{R}_c^{\text{ave}}(p(\boldsymbol{w}); \boldsymbol{x}_t, y_t) + \mathbf{KL}[p(\boldsymbol{w})||p_0(\boldsymbol{w})] + \left(\frac{S}{2} + 2c^2 R\right) T. \quad (17)$$

Proof The proof is similar to that of Gibbs classifiers. According to Lemma 3, for BayesPA with averaging classifiers, we have the streaming update rule

$$q_{t+1}(\boldsymbol{w}) = \frac{1}{\Gamma(\tau_t; \boldsymbol{x}_t, y_t)} q_t(\boldsymbol{w}) p(\boldsymbol{x}_t|\boldsymbol{w}) e^{\tau_t y_t \boldsymbol{w}^\top \boldsymbol{x}_t},$$

where $\Gamma(\tau_t; \boldsymbol{x}_t, y_t)$ is the partition function and τ_t is the solution of the dual problem:

$$\max_{0 \leq \tau \leq 2c} \epsilon\tau - \log \Gamma(\tau; \boldsymbol{x}_t, y_t). \quad (18)$$

2. The theorem states that if a function $g : \mathbb{R}^k \to \mathbb{R}$ is $k+1$ times differentiable in a closed ball B, then for $\boldsymbol{x}_0, \boldsymbol{x} \in B$, $\exists c \in (0, 1)$ such that $f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \partial f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^\top \partial^2 f[c\boldsymbol{x}_0 + (1-c)\boldsymbol{x}](\boldsymbol{x} - \boldsymbol{x}_0)$.

By definition, the partition function is $\Gamma(\tau_t; \boldsymbol{x}_t, y_t) = \int_{\boldsymbol{w}} q_t(\boldsymbol{w}) p(\boldsymbol{x}_t|\boldsymbol{w}) e^{\tau_t y_t \boldsymbol{w}_t \boldsymbol{x}_t} d\boldsymbol{w}$. Construct the two parameter exponential family

$$q_{t,\tau,u}(\boldsymbol{w}) = q_t(\boldsymbol{w}) \exp\Big(u \log p(\boldsymbol{x}_t|\boldsymbol{w}) + \tau y_t \boldsymbol{w}^\top \boldsymbol{x}_t - \log f(\tau, u)\Big),$$

where the log partition function $f(\tau, u)$ is

$$\log f(\tau, u) = \log \int_{\boldsymbol{w}} q_t(\boldsymbol{w}) p(\boldsymbol{x}_t|\boldsymbol{w})^u e^{\tau y_t \boldsymbol{w} \boldsymbol{x}_t} d\boldsymbol{w}.$$

Using Taylor's theorem again at the origin $(0, 0)$, we have

$$\begin{aligned}\log f(\tau, u) = \mathbb{E}_{q_t(\boldsymbol{w})} \left[\log p(\boldsymbol{x}_t|\boldsymbol{w})\right] u + \mathbb{E}_{q_t(\boldsymbol{w})} \left[y_t \boldsymbol{w}^\top \boldsymbol{x}_t\right] \tau \\ + \frac{1}{2}\Big(\mathbb{V}_{q_{\hat{\tau},\hat{u}}} \left[y_t \boldsymbol{w} \boldsymbol{x}_t\right] \tau^2 + \mathbb{V}_{q_{\hat{\tau},\hat{u}}} \left[\log p(\boldsymbol{x}_t|\boldsymbol{w})\right] u^2\Big),\end{aligned}$$

for some $0 < \hat{\tau} < \tau$ and $0 < \hat{u} < u$. Using our assumption on the fisher information and the fact that $\Gamma(\tau; \boldsymbol{x}_t, y_t) = f(\tau, 1)$, we have the bound

$$\epsilon\tau - \log \Gamma(\tau; \boldsymbol{x}_t, y_t) \geq -\mathbb{E}_{q_t(\boldsymbol{w})} \left[\log p(\boldsymbol{x}_t|\boldsymbol{w})\right] + \Big(\epsilon - \mathbb{E}_{q_t(\boldsymbol{w})} \left[y_t \boldsymbol{w}^\top \boldsymbol{x}_t\right]\Big) \tau - \frac{1}{2} R \tau^2 - \frac{1}{2} S. \quad (19)$$

The optimal solution for the lower bound is $\tau^* = \min\{2c, (\epsilon - \mathbb{E}_{q_t(\boldsymbol{w})}[y_t \boldsymbol{w}^\top \boldsymbol{x}_t])/R\}$. Now, assume that the current round $q_t(\boldsymbol{w})$ suffers non-zero loss and consider the difference

$$\begin{aligned}&\mathbf{KL}\left[p(\boldsymbol{w})||q_t(\boldsymbol{w})\right] - \mathbf{KL}\left[p(\boldsymbol{w})||q_{t+1}(\boldsymbol{w})\right] \\ &\quad = \int_{\boldsymbol{w}} p(\boldsymbol{w}) \tau_t \left(y_t \boldsymbol{w}_t^\top \boldsymbol{x}_t - \epsilon\right) d\boldsymbol{w} + \mathbb{E}_{p(\boldsymbol{w})}[\log p(\boldsymbol{x}_t|\boldsymbol{w})] + \Big(\epsilon\tau_t - \log \Gamma(\tau_t; \boldsymbol{x}_t, y_t)\Big) \\ &\quad \geq -\mathcal{R}_c^{\text{ave}}(p(\boldsymbol{w}); \boldsymbol{x}_t, y_t) + \Big(\epsilon\tau_t - \log \Gamma(\tau_t; \boldsymbol{x}_t, y_t)\Big).\end{aligned} \quad (20)$$

Notice that the second term in Eq. (20) is exactly the optimization objective in the dual problem (18). Therefore, if $(\epsilon - \mathbb{E}_{q_t(\boldsymbol{w})}[y_t \boldsymbol{w}^\top \boldsymbol{x}_t]) \geq 2cR$, we have $\tau^* = 2c$ and use (19) to show

$$\mathcal{R}_c^{\text{ave}}(q_t(\boldsymbol{w}); \boldsymbol{x}_t, y_t) \leq \mathcal{R}_c^{\text{ave}}(p(\boldsymbol{w}); \boldsymbol{x}_t, y_t) + \mathbf{KL}[p||q_t] - \mathbf{KL}[p||q_{t+1}] + 2c^2 R + \frac{S}{2}.$$

If $(\epsilon - \mathbb{E}_{q_t(\boldsymbol{w})}[y_t \boldsymbol{w}^\top \boldsymbol{x}_t]) < 2cR$, we obtain

$$\begin{aligned}\ell_\epsilon^{\text{ave}}(q_t(\boldsymbol{w}); \boldsymbol{x}_t, y_t) &\leq 2\sqrt{cR \cdot \tfrac{1}{2c}\Big(\mathcal{R}_c^{\text{ave}}(p(\boldsymbol{w}); \boldsymbol{x}_t, y_t) + \mathbf{KL}[p(\boldsymbol{w})||q_t(\boldsymbol{w})] - \mathbf{KL}[p(\boldsymbol{w})||q_{t+1}(\boldsymbol{w})] + S/2\Big)} \\ &\leq cR + \tfrac{1}{2c}\Big(\mathcal{R}_c^{\text{ave}}(p(\boldsymbol{w}); \boldsymbol{x}_t, y_t) + \mathbf{KL}[p(\boldsymbol{w})||q_t(\boldsymbol{w})] - \mathbf{KL}[p(\boldsymbol{w})||q_{t+1}(\boldsymbol{w})] + S/2\Big).\end{aligned}$$

where we have used the geometric inequality. Summing over all $t = 0, 1, 2, ..., T-1$ gives (17) and further relax it by neglecting $\mathbf{KL}[p(\boldsymbol{w})||q_T(\boldsymbol{w})]$, we then derive Eq. (17).

■

Remark 1. The bounds (15) and (17) both imply that the regrets satisfy

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{R}_c(q_t(\boldsymbol{w});\boldsymbol{x}_t,y_t) \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathcal{R}_c(p(\boldsymbol{w});\boldsymbol{x}_t,y_t) + \frac{1}{T}\mathbf{KL}[p(\boldsymbol{w})||q_0(\boldsymbol{w})] + \text{const.}$$

When $T \to \infty$, the asymptotic average regret of BayesPA is at most larger than that of the optimal batch learner by a constant factor.

Remark 2. Interestingly, $\mathbf{KL}[p(\boldsymbol{w})||q_0(\boldsymbol{w})] + \sum_{t=0}^{T-1}\mathcal{R}_c(p(\boldsymbol{w});\boldsymbol{x}_t,y_t)$ is the RegBayes objective function of the batch learner with T data samples. In other words, if there exists a batch learner $p(\boldsymbol{w})$ who achieves a small objective, so can BayesPA learning.

Remark 3. For non-likelihood BayesPA, the regret is $\mathcal{R}_c(q_t(\boldsymbol{w});\boldsymbol{w}_t,y_t) = \ell_\epsilon(q_t(\boldsymbol{w});\boldsymbol{w}_t,y_t)$, which recovers the notion of regret in the classical sense. As a special case, theorems 6 and 7 also hold true.

Remark 4. Both theorem 6 and 7 assume the Fisher information is bounded by a constant factor. In other words, each data point does not cause abrupt change in paramter estimate. This is a practical assumption for online learning because to allow for reasonable inference, the concept space should be sufficiently restricted. Detecting abrupt change points (Adams and MacKay, 2007) in streaming data is beyond the scope of this paper.

6. 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.

6.1 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, $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_k\}_{k=1}^K$, where each topic $\boldsymbol{\phi}_k$ is a multinomial distribution over a given W-word vocabulary.[3] The generative process of the d-th document ($1 \leq d \leq D$) is described as follows:

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

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

3. Without causing confusion, we slightly abused the notation K to denote the topic number (i.e., the latent dimension) in topic models.

Given a document set $\boldsymbol{X} = \{\boldsymbol{x}_d\}_{d=1}^D$. Let $\boldsymbol{Z} = \{\boldsymbol{z}_d\}_{d=1}^D$ and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_d\}_{d=1}^D$ denote all the topic assignments and topic mixing vectors. LDA infers the posterior distribution $p(\boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{Z}|\boldsymbol{X}) \propto p_0(\boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{Z})p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\Phi})$ via Bayes' rule. From a variational point of view, the Bayes' post-data posterior is identical to the solution of the optimization problem:

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

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 $\boldsymbol{w}$ on the data $\boldsymbol{X} = \{\boldsymbol{x}_d\}_{d=1}^D$ and response signals $\boldsymbol{Y} = \{y_d\}_{d=1}^D$. As a regularized Bayesian (RegBayes) model (Jiang et al., 2012), MedLDA infers a distribution of the latent variables $\boldsymbol{Z}$ as well as classification weights $\boldsymbol{w}$ by solving the problem:

$$\min_{q \in \mathcal{P}} \mathcal{L}\Big(q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{Z})\Big) + 2c \sum_{d=1}^D \ell_\epsilon\Big(q(\boldsymbol{w}, \boldsymbol{z}_d); \boldsymbol{x}_d, y_d\Big),$$

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

$$f(\boldsymbol{w}, \boldsymbol{z}_d) = \boldsymbol{w}^\top \bar{\boldsymbol{z}}_d, \tag{21}$$

where $\bar{z}_{dk} = \frac{1}{n_d} \sum_i \mathbb{I}[z_{di} = k]$ 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

$$\ell_\epsilon^{\text{ave}}(q(\boldsymbol{w}, \boldsymbol{z}_d); \boldsymbol{x}_d, y_d) = \big(\epsilon - y_d \mathbb{E}_{q(\boldsymbol{w}, \boldsymbol{z}_d)}[f(\boldsymbol{w}, \boldsymbol{z}_d)]\big)_+, \tag{22}$$

and Gibbs classifiers with the expected hinge loss

$$\ell_\epsilon^{\text{gibbs}}(q(\boldsymbol{w}, \boldsymbol{z}_d); \boldsymbol{x}_d, y_d) = \mathbb{E}_{q(\boldsymbol{w}, \boldsymbol{z}_d)}\left[(\epsilon - y_d f(\boldsymbol{w}, \boldsymbol{z}_d))_+\right], \tag{23}$$

have been proposed, with extensive comparisons reported in Zhu et al. (2014a) using batch learning algorithms.

6.2 Online MedLDA

We first apply online BayesPA to MedLDA with averaging classifiers, which will be referred to as paMedLDA$^{\text{ave}}$ in the sequel. During inference, we integrate out the local variables $\boldsymbol{\Theta}_t$ 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 $\boldsymbol{\mathcal{M}} = \boldsymbol{\Phi}$ and local variables $\boldsymbol{H}_t = \boldsymbol{Z}_t$. The latent BayesPA rule (12) becomes:

$$\begin{aligned} \min_{q, \xi_d} \quad & \mathbf{KL}\Big[q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)||q_t(\boldsymbol{w}, \boldsymbol{\Phi})p_0(\boldsymbol{Z}_t)p(\boldsymbol{x}_t|\boldsymbol{\Phi}, \boldsymbol{Z}_t)\Big] + 2c \sum_{d \in B_t} \xi_d, \\ \text{s.t.:} \quad & y_d \mathbb{E}_{q(\boldsymbol{w}, \boldsymbol{z}_d)}[\boldsymbol{w}^\top \bar{\boldsymbol{z}}_d] \geq \epsilon - \xi_d, \quad \xi_d \geq 0, \quad \forall d \in B_t, \\ & q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t) \in \mathcal{P}. \end{aligned} \tag{24}$$

Since directly solving the above problem is intractable, we would impose a mild mean-field assumption $q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t) = q(\boldsymbol{w})q(\boldsymbol{\Phi})q(\boldsymbol{Z}_t)$. Now, problem (24) can be solved using an iterative procedure that alternately updates each factor distribution (Jordan et al., 1998), as detailed below:

1. **Update global $q(\mathbf{\Phi})$:** By fixing the distributions $q(\boldsymbol{Z}_t)$ and $q(\boldsymbol{w})$, we can ignore irrelevant terms and solve

$$\min_{q(\mathbf{\Phi})} \ \mathbf{KL}\Big[q(\mathbf{\Phi})q(\boldsymbol{Z}_t)||q_t(\mathbf{\Phi})p_0(\boldsymbol{Z}_t)p(\boldsymbol{x}_t|\mathbf{\Phi},\boldsymbol{Z}_t)\Big].$$

The optimal solution has the following closed form:

$$q^*(\mathbf{\Phi}_k) \propto q_t(\mathbf{\Phi}_k)\exp\Big(\mathbb{E}_{q(\boldsymbol{Z}_t)}\Big[\log p_0(\boldsymbol{Z}_t)p(\boldsymbol{X}|\boldsymbol{Z}_t,\mathbf{\Phi})\Big]\Big), \ \ k=1,2,...,K. \tag{25}$$

If initially the prior is $q_0(\mathbf{\Phi}_k) = \text{Dir}(\Delta^0_{k1},...,\Delta^0_{kW})$, then by induction the inferred distributions in each round are also in the family of Dirichlet distributions, namely, $q_t(\mathbf{\Phi}_k) = \text{Dir}(\Delta^t_{k1},...,\Delta^t_{kW})$. Using equation (25), we can derive

$$q^*(\mathbf{\Phi}_k) = \text{Dir}(\Delta^*_{k1},...,\Delta^*_{kW}), \tag{26}$$

where $\Delta^*_{kw} = \Delta^t_{kw} + \sum_{d\in B_t}\sum_{i=1}^{n_d}\gamma^k_{di}\cdot\mathbb{I}[x_{di}=w]$ for all words w ($1\le w\le W$) in the vocuabulary and $\gamma^k_{di} = \mathbb{E}_{q(\boldsymbol{z}_d)}\mathbb{I}[z_{di}=k]$ is the probability of assigning each word x_{di} to topic k.

2. **Update global weight $q(\boldsymbol{w})$:** Keeping all the other distributions fixed, $q(\boldsymbol{w})$ can be solved as

$$\begin{aligned}\min_{q(\boldsymbol{w})} \ & \mathbf{KL}\Big[q(\boldsymbol{w})||q_t(\boldsymbol{w})\Big] + 2c\sum_{d\in B_t}\xi_d,\\ \text{s.t.: } & y_d\mathbb{E}_{q(\boldsymbol{w})}[\boldsymbol{w}]^\top\widehat{\boldsymbol{z}}_d \ge \epsilon-\xi_d, \quad \xi_d\ge 0, \quad \forall d\in B_t,\end{aligned}$$

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

$$q^*(\boldsymbol{w}) = \frac{1}{Z(\tau^*_d)}q_t(\boldsymbol{w})\exp\left(\sum_{d\in B_t}\tau^*_d y_d\boldsymbol{w}^\top\widehat{\boldsymbol{z}}_d\right), \tag{27}$$

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

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

For the common spherical Gaussian prior $q_0(\boldsymbol{w}) = \mathcal{N}(0,\sigma^2 I)$, by induction the distribution $q_t(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}_t,\sigma^2 I)$ at each round. So equation (27) gives $q^*(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}^*,\sigma^2 I)$, where

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_t + \sigma^2\sum_{d\in B_t}\tau^*_d y_d\widehat{\boldsymbol{z}}_d. \tag{28}$$

Furthermore, the dual problem becomes,

$$\max_{0\le\tau_d\le 2c} \ \epsilon\sum_{d\in B_t}\tau_d - \sum_{d_1,d_2\in B_t}\frac{1}{2}\sigma^2\tau_{d_1}\tau_{d_2}\widehat{\boldsymbol{z}}^\top_{d_1}\widehat{\boldsymbol{z}}_{d_2} - \boldsymbol{\mu}_t^\top\sum_{d\in B_t}y_d\tau_d\widehat{\boldsymbol{z}}_d, \tag{29}$$

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)

$$\min_{\mu} \frac{||\boldsymbol{\mu} - \boldsymbol{\mu}_t||^2}{2\sigma^2} + 2c \sum_{d \in B_t} \Big(\epsilon - y_d \boldsymbol{\mu}^\top \widehat{\boldsymbol{z}}_d\Big)_+. \tag{30}$$

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

$$\tau_t^* = \min\left\{2c, \frac{\big(\epsilon - y_t \mu_t^\top \widehat{\boldsymbol{z}}_t\big)_+}{||\widehat{\boldsymbol{z}}_t||^2}\right\},$$

whose computation requires $O(K)$ time; for mini-batches, we could adapt methods solving linear SVM to either the dual (29) or primal (30) problem, which by state-of-the-arts require complexity $O(\text{poly}(\epsilon^{-1})K)$ per training instance in order to obtain ϵ-accurate solutions. Here we choose a gradient-based method similar to Shalev-Shwartz et al. (2011). Specifically, we first set $\boldsymbol{\mu}^1 = \boldsymbol{\mu}_t$, and then take the gradient steps $i = 2, 3, ...$ until $\boldsymbol{\mu}^i$ converges to $\boldsymbol{\mu}^*$. Let A_t^i be the set of instances in B_t that suffer non-zero loss at step i, then we use the gradients to iteratively update

$$\boldsymbol{\mu}^{i+1} \leftarrow \boldsymbol{\mu}^i - \rho_i \nabla_i, \tag{31}$$

where annealing rate $\rho_i = \sigma^2 i^{-1}$ and

$$\nabla_i = \frac{\boldsymbol{\mu}^i - \boldsymbol{\mu}_t}{\sigma^2} - 2c \sum_{d \in A_t^i} y_d \widehat{\boldsymbol{z}}_d.$$

Correspondingly, we can derive the gradient-based update rule for the dual parameters. Imagine that we implicitly maintain the relationship $\boldsymbol{\mu} = \boldsymbol{\mu}_t + \sigma^2 \sum_{d \in B_t} \tau_d y_d \widehat{\boldsymbol{z}}_d$. Then the following update rule for τ_d $(d \in B_t)$ naturally implies the update rule (31) for $\boldsymbol{\mu}$:

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

Therefore, the gradient steps adaptively adjust the contribution of each latent $\widehat{\boldsymbol{z}}_d$ to μ based on the loss it incurs. Furthermore, the annealing makes sure that $0 \le \tau_d^i \le 2c$ for all i. Since the problem (29) is concave, it can be guaranteed that τ_d^i converges to τ_d^*. This correspondence would be used in updating $q(\boldsymbol{Z}_t)$.

3. **Update local $q(\boldsymbol{Z}_t)$:** Fixing all the other distributions, we aim to solve

$$\begin{aligned} \min_{q(\boldsymbol{Z}_t)} \quad & \mathbf{KL}\Big[q(\boldsymbol{Z}_t)||p_0(\boldsymbol{Z}_t)p(\boldsymbol{X}_t|\boldsymbol{Z}_t, \boldsymbol{\Phi})\Big] + 2c \sum_{d \in B_t} \xi_d, \\ \text{s.t.:} \quad & y_d \boldsymbol{\mu}^{*\top} \mathbb{E}_{q(\boldsymbol{z}_d)}[\bar{\boldsymbol{z}}_d] \ge \epsilon - \xi_d, \quad \xi_d \ge 0, \quad \forall d \in B_t, \end{aligned}$$

where $\boldsymbol{\mu}^* = \mathbb{E}_{q(\boldsymbol{w})}[\boldsymbol{w}]$ is the expectation of $\boldsymbol{w}$ under the fixed distribution $q(\boldsymbol{w})$. Unlike the weight $\boldsymbol{w}$, the expectation over $\boldsymbol{Z}_t$ during optimization is intractable due to combinatorial

Algorithm 1 Online MedLDA

Let $q_0(\boldsymbol{w}) = \mathcal{N}(0; \sigma^2 I), q_0(\boldsymbol{\phi}_k) = \text{Dir}(\gamma),\ \forall\ k$.
for $t = 0 \to \infty$ **do**
  Set $q(\boldsymbol{\Phi}, \boldsymbol{w}) = q_t(\boldsymbol{\Phi}) q_t(\boldsymbol{w})$. Initialize $\boldsymbol{Z}_t$.
  for $i = 1 \to \mathcal{I}$ **do**
    Draw samples $\{\boldsymbol{Z}_t^{(j)}\}_{j=1}^{\mathcal{J}}$ from Eq. (32).
    Discard the first β burn-in samples ($\beta < \mathcal{J}$).
    Use the rest $\mathcal{J} - \beta$ samples to update $q(\boldsymbol{\Phi}, \boldsymbol{w})$ following Eq.s (26, 27).
  end for
  Set $q_{t+1}(\boldsymbol{\Phi}, \boldsymbol{w}) = q(\boldsymbol{\Phi}, \boldsymbol{w})$.
end for

space of values. Instead, we adopt the same approximation strategy as MedLDA (Zhu et al., 2012): fix ξ, τ_d at the previous global step, and use the approximate solution

$$q^*(\boldsymbol{Z}_t) = p_0(\boldsymbol{Z}_t) p(\boldsymbol{X}_t | \boldsymbol{Z}_t, \boldsymbol{\Phi}) \exp\left(\sum_{d \in B_t} \tau_d^* y_d \boldsymbol{\mu}^{*\top} \bar{\boldsymbol{z}}_d \right).$$

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

$$q(z_{di} = k \mid \boldsymbol{Z}_t^{\neg di}) \propto \left(\alpha + C_{dk}^{\neg di}\right) \exp\left(\Lambda_{k, x_{di}} + \sum_{d \in B_t} n_d^{-1} y_d \tau_d^* \mu_k^* \right). \tag{32}$$

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 $\boldsymbol{C}_d^{\neg di}$ 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 $\mathcal{J}$ samples $\{\boldsymbol{Z}_t^{(j)}\}_{j=1}^{\mathcal{J}}$ using Eq. (32), discard the first β ($0 \le \beta < \mathcal{J}$) burn-in samples, and approximate $\widehat{z}_{dk}$ with the empirical sum $(J - \beta)^{-1} \sum_{j=\beta+1}^{J} \sum_{d,i} \mathbb{I}[z_{di}^{(j)} = k]$.

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

6.3 Online Gibbs MedLDA

In this section, we apply the theory of BayesPA to Gibbs MedLDA. As shown in Zhu et al. (2014a), 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

paMedLDA$^{\text{gibbs}}$. Specifically, let $\zeta_d = \epsilon - y_d f(\boldsymbol{w}, \boldsymbol{z}_d)$ and $\psi(y_d|\boldsymbol{z}_d, \boldsymbol{w}) = e^{-2c(\zeta_d)_+}$. By Lemma 4, the optimal solution to problem (12) is

$$q_{t+1}(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t) = \frac{q_t(\boldsymbol{w}, \boldsymbol{\mathcal{M}}) p_0(\boldsymbol{H}_t) p(\boldsymbol{X}_t|\boldsymbol{H}_t, \boldsymbol{\mathcal{M}}) \psi(\boldsymbol{Y}_t|\boldsymbol{H}_t, \boldsymbol{w})}{\Gamma(\boldsymbol{X}_t, \boldsymbol{Y}_t)} \ ,$$

where $\psi(\boldsymbol{Y}_t|\boldsymbol{H}_t, \boldsymbol{w}) = \prod_{d \in B_t} \psi(y_d|\boldsymbol{h}_d, \boldsymbol{w})$ and $\Gamma(\boldsymbol{X}_t, \boldsymbol{Y}_t)$ 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., 2014a):

$$\psi(y_d|\boldsymbol{z}_d, \boldsymbol{w}) = \int_0^\infty \psi(y_d, \lambda_d|\boldsymbol{z}_d, \boldsymbol{w}) d\lambda_d, \tag{33}$$

where $\psi(y_d, \lambda_d|\boldsymbol{z}_d, \boldsymbol{w}) = (2\pi\lambda_d)^{-1/2} \exp\left(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\right)$. Equality (33) essentially implies that the collapsed posterior $q_{t+1}(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)$ is a marginal distribution of

$$q_{t+1}(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t) = \frac{p_0(\boldsymbol{Z}_t) q_t(\boldsymbol{w}, \boldsymbol{\Phi}) p(\boldsymbol{X}_t|\boldsymbol{Z}_t, \boldsymbol{\Phi}) \psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t|\boldsymbol{Z}_t, \boldsymbol{w})}{\Gamma(\boldsymbol{X}_t, \boldsymbol{Y}_t)} \ ,$$

where $\psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t|\boldsymbol{Z}_t, \boldsymbol{w}) = \prod_{d \in B_t} \psi(y_d, \lambda_d|\boldsymbol{z}_d, \boldsymbol{w})$ and $\boldsymbol{\lambda}_t = \{\lambda_d\}_{d \in B_t}$ are augmented variables locally associated with individual documents. In fact, the augmented distribution $q_{t+1}(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t)$ is the solution to the problem:

$$\min_{q \in \mathcal{P}} \mathcal{L}\Big(q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t)\Big) - \mathbb{E}_q\Big[\log \psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t|\boldsymbol{Z}_t, \boldsymbol{w})\Big], \tag{34}$$

where $\mathcal{L}(q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t)) = \mathbf{KL}[q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t) \| q_t(\boldsymbol{w}, \boldsymbol{\Phi}) p_0(\boldsymbol{Z}_t)] - \mathbb{E}_q[\log p(\boldsymbol{X}_t|\boldsymbol{Z}_t, \boldsymbol{\Phi})]$. In fact, this objective is an upper bound of that in the original problem (12) (See Appendix A for details).

Again, with the mild mean-field assumption that $q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t) = q(\boldsymbol{w}, \boldsymbol{\Phi}) q(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t)$, we can solve problem (34) 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, $q(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t)$, and ignoring irrelevant terms, the optimal distribution of $\boldsymbol{w}$ and $\boldsymbol{\Phi}$ can be shown to have the induced factorization form, $q(\boldsymbol{w}, \boldsymbol{\Phi}) = q(\boldsymbol{w}) q(\boldsymbol{\Phi})$. For $q(\boldsymbol{\Phi})$, the update rule is exactly (26). For $q(\boldsymbol{w})$, we have the update rule

$$q_{t+1}(\boldsymbol{w}) \propto q_t(\boldsymbol{w}) \exp\left(\mathbb{E}_{q(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t)}\Big[\log p_0(\boldsymbol{Z}_t) \psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t|\boldsymbol{Z}_t, \boldsymbol{w})\Big]\right).$$

   If the initial prior is normal $q_0(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0)$, by induction we can show that the inferred distribution in each round is also a normal distribution, namely, $q_t(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$. Indeed, the optimal solution of $q(\boldsymbol{w})$ to problem (34) is

$$q^*(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \tag{35}$$

   where the posterior paramters are computed as

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

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

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

2. **Local Update:** Given the distribution of global variables, $q(\boldsymbol{\Phi}, \boldsymbol{w})$, the mean-field update equation for $(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t)$ is

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

where Λ admits the same definition as in (32). But it is impossible to evaluate the expectation in the global update using $q(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t)$ because of the huge number of configurations for $(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t)$. 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:

- **For $\boldsymbol{Z}_t$:** By canceling out common factors, the conditional distribution of one variable z_{di} given $\boldsymbol{Z}_t^{\neg di}$ and $\boldsymbol{\lambda}_t$ is

$$\begin{aligned} q(z_{di} = k | \boldsymbol{Z}_t^{\neg di}, \boldsymbol{\lambda}_t) \propto (\alpha + C_{dk}^{\neg di}) \exp\Big(& \frac{c y_d (c\epsilon + \lambda_d) \mu_k^*}{n_d \lambda_d} \\ & + \Lambda_{k, x_{di}} - \frac{c^2 (\mu_k^{*2} + \Sigma_{kk}^* + 2(\mu_k^* \boldsymbol{\mu}^* + \boldsymbol{\Sigma}_{\cdot,k}^*)^\top \boldsymbol{C}_d^{\neg di})}{2 n_d^2 \lambda_d} \Big), \end{aligned} \tag{36}$$

  where $\boldsymbol{\Sigma}_{\cdot,k}^*$ is the k-th column of $\boldsymbol{\Sigma}^*$.

- **For $\boldsymbol{\lambda}_t$:** Let $\bar{\zeta}_d = \epsilon - y_d \bar{\boldsymbol{z}}_d^\top \boldsymbol{\mu}^*$. The conditional distribution of each variable λ_d given $\boldsymbol{Z}_t$ is

$$\begin{aligned} q(\lambda_d | \boldsymbol{Z}_t) \propto & \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{c^2 \bar{\boldsymbol{z}}_d^\top \boldsymbol{\Sigma}^* \bar{\boldsymbol{z}}_d + (\lambda_d + c\bar{\zeta}_d)^2}{2\lambda_d} \right) \\ = & \mathcal{GIG}\left(\lambda_d; \frac{1}{2}, 1, c^2 \left(\bar{\zeta}_d^2 + \bar{\boldsymbol{z}}_d^\top \boldsymbol{\Sigma}^* \bar{\boldsymbol{z}}_d \right) \right), \end{aligned} \tag{37}$$

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

$$q(\lambda_d^{-1} | \boldsymbol{Z}_t) = \mathcal{IG}\left(\lambda_d^{-1}; \frac{1}{c\sqrt{\bar{\zeta}_d^2 + \bar{\boldsymbol{z}}_d^\top \boldsymbol{\Sigma}^* \bar{\boldsymbol{z}}_d}}, 1 \right),$$

  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 $\hat{\boldsymbol{w}}$ as the classification weight and apply the prediction rule. The inference of $\bar{\boldsymbol{z}}$ for testing documents is the same as online MedLDA.

Algorithm 2 Online Gibbs MedLDA

Let $q_0(\boldsymbol{w}) = \mathcal{N}(0; \sigma^2 I), q_0(\boldsymbol{\phi}_k) = \text{Dir}(\gamma),\ \forall\ k$.
for $t = 0 \to \infty$ **do**
  Set $q(\boldsymbol{\Phi}, \boldsymbol{w}) = q_t(\boldsymbol{\Phi}) q_t(\boldsymbol{w})$. Initialize $\boldsymbol{Z}_t$.
  for $i = 1 \to \mathcal{I}$ **do**
    Draw samples $\{\boldsymbol{Z}_t^{(j)}, \boldsymbol{\lambda}_t^{(j)}\}_{j=1}^{\mathcal{J}}$ from Eq.s (36, 37).
    Discard the first β burn-in samples ($\beta < \mathcal{J}$).
    Use the rest $\mathcal{J} - \beta$ samples to update $q(\boldsymbol{\Phi}, \boldsymbol{w})$ following Eq.s (26, 35).
  end for
  Set $q_{t+1}(\boldsymbol{\Phi}, \boldsymbol{w}) = q(\boldsymbol{\Phi}, \boldsymbol{w})$.
end for

7. Extensions

In the above topic models, we assume that the number of topics (i.e., K) is pre-specified. We now present extensions of online MedLDA to automatically determine the unknown K values. We also present an extension of these models for multi-task learning.

7.1 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).

7.1.1 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 $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{\infty}$ are generated as:

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

and the topic mixing proportions are generated as $\boldsymbol{\theta}_d \sim \text{Dir}(\alpha\boldsymbol{\pi}),\ \text{for } d = 1, ..., D$. Each topic $\boldsymbol{\phi}_k$ is a sample from a Dirichlet base distribution, i.e., $\boldsymbol{\phi}_k \sim \text{Dir}(\boldsymbol{\gamma})$. After we get the topic mixing proportions $\boldsymbol{\theta}_d$, 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 $\boldsymbol{w}$ that is drawn from a Gaussian process, $\mathcal{GP}(0, \boldsymbol{\Sigma})$, where the covariance function is $\Sigma(\boldsymbol{w}, \boldsymbol{w}') = \sigma^2 \mathbb{I}[\boldsymbol{w} = \boldsymbol{w}']$. We still define the linear discriminant function in the same form as Eq. (21). Since the number of words in a document is finite, the average topic assignment vector $\bar{\boldsymbol{z}}_d$ has only a finite number of non-zero elements, and the dot product in Eq. (21) is in fact finite. Therefore, given the latent topic assignments, the conditoinal posterior of $\boldsymbol{w}$ is in fact a multivariate Gaussian distribution.

Let $\bar{\boldsymbol{\pi}} = \{\bar{\pi}_k\}_{k=1}^{\infty}$. We define maximum entropy discrimination HDP (MedHDP) topic model as solving the following RegBayes problem to infer the joint post-data posterior $q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{Z})$:[4]

$$\min_{q \in \mathcal{P}} \mathcal{L}\Big(q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{Z})\Big) + 2c \sum_{d=1}^{D} \ell_\epsilon\Big(q(\boldsymbol{w}, \boldsymbol{z}_d); \boldsymbol{x}_d, \boldsymbol{y}_d\Big), \tag{38}$$

4. Given $\bar{\boldsymbol{\pi}}$, $\boldsymbol{\pi}$ can be computed via the stick breaking process.

where $\mathcal{L}(q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{Z})) = \mathbf{KL}[q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{Z})||p(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{Z}|\boldsymbol{X})]$ is the objective corresponding to the standard Bayesian inference under the variational formulation of Bayes' rule. The loss function could be either (22) or (23). We call the resulting model with averaging classifier MedHDP$^{\text{ave}}$ and that with Gibbs classifier MedHDP$^{\text{gibbs}}$.

Since MedHDP is a new model, we would briefly discuss the corresponding inference problem. For the inference of MedLDA$^{\text{gibbs}}$, 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. (2014a). 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, but instead attain batch MedHDP algorithms from the BayesPA counterparts.

7.1.2 ONLINE MEDHDP

To apply the ideas of BayesPA to develop online MedHDP algorithms, we have the global variables $\boldsymbol{\mathcal{M}} = (\bar{\boldsymbol{\pi}}, \boldsymbol{\Phi})$, and the local variables $\boldsymbol{H}_t = (\boldsymbol{\Theta}_t, \boldsymbol{Z}_t)$. As in online MedLDA, we marginalize out $\boldsymbol{\Theta}_t$ by conjugacy. Furthermore, to simplify the sampling scheme, we introduce another set of auxiliary latent variables $\boldsymbol{S}_t = \{\boldsymbol{s}_d\}_{d \in B_t}$, where $\boldsymbol{s}_d = \{s_{dk}\}_{k=1}^{\infty}$ 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 $p(\boldsymbol{Z}_t, \boldsymbol{S}_t|\bar{\boldsymbol{\pi}}) = \prod_{d \in B_t} p(\boldsymbol{s}_d, \boldsymbol{z}_d|\bar{\boldsymbol{\pi}})$ and

$$p(\boldsymbol{s}_d, \boldsymbol{z}_d|\bar{\boldsymbol{\pi}}) \propto \prod_{k=1}^{\infty} S(n_d \bar{z}_{dk}, s_{dk})(\alpha \pi_k)^{s_{dk}}, \tag{39}$$

where $S(a, b)$ are unsigned Stirling numbers of the first kind (Antoniak, 1974). It is not hard to verify that $p(\boldsymbol{z}_d|\bar{\boldsymbol{\pi}}) = \sum_{\boldsymbol{s}_d} p(\boldsymbol{s}_d, \boldsymbol{z}_d|\bar{\boldsymbol{\pi}})$. After this "collapse-and-augment" procedure, we now have the local variables $\boldsymbol{H}_t = (\boldsymbol{Z}_t, \boldsymbol{S}_t)$. The global variables remain intact. The new BayesPA problem is now:

$$\min_{q \in \mathcal{F}_t} \mathcal{L}\Big(q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi}, \boldsymbol{H}_t)\Big) + 2c \sum_{d \in B_t} \ell_\epsilon\Big(\boldsymbol{w}; \boldsymbol{x}_t, y_t\Big), \tag{40}$$

where $\mathcal{L}(q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi}, \boldsymbol{H}_t)) = \mathbf{KL}[q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi}, \boldsymbol{H}_t)||q_t(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi}) p(\boldsymbol{Z}_t, \boldsymbol{S}_t|\bar{\boldsymbol{\pi}}) p(\boldsymbol{X}_t|\boldsymbol{Z}_t, \boldsymbol{\Phi})]$. As in online MedLDA, we adopt the mild mean field assumption $q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi}, \boldsymbol{H}_t) = q(\boldsymbol{w}) q(\bar{\boldsymbol{\pi}}) q(\boldsymbol{\Phi}) q(\boldsymbol{H}_t)$ and solve problem (40) via an iterative procedure detailed below.

1. **Global Update:** By fixing the distribution of local variables, $q(\boldsymbol{H}_t)$, and ignoring the irrelevant terms, we have same mean-field update equations (26) for $\boldsymbol{\Phi}$ and (28) for $\boldsymbol{w}$ with the averaging loss. For global variable $\bar{\boldsymbol{\pi}}$, we have

$$q^*(\bar{\pi}_k) \propto q_t(\bar{\pi}_k) \prod_{d \in B_t} \exp\Big(\mathbb{E}_{q(\boldsymbol{h}_d)}\Big[\log p(\boldsymbol{s}_d, \boldsymbol{z}_d|\bar{\boldsymbol{\pi}})\Big]\Big). \tag{41}$$

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

$$q^*(\bar{\pi}_k) = \text{Beta}(u_k^*, v_k^*), \tag{42}$$

where $u_k^* = u_k^t + \sum_{d \in B_t} \mathbb{E}_{q(\boldsymbol{s}_d)}[s_{dk}]$ and $v_k^* = v_k^t + \sum_{d \in B_t} \mathbb{E}_{q(\boldsymbol{s}_d)}[\sum_{j>k} s_{dj}]$ for $k = \{1, 2, ...\}$ and $u_k^0 = 1, v_k^0 = \eta$.

Since $\boldsymbol{Z}_t$ 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 $q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi})$, we get the mean-field update equation for $(\boldsymbol{Z}_t, \boldsymbol{S}_t)$:

$$q^*(\boldsymbol{Z}_t, \boldsymbol{S}_t) \propto \tilde{q}(\boldsymbol{Z}_t, \boldsymbol{S}_t)\hat{q}(\boldsymbol{Z}_t) \tag{43}$$

where

$$\tilde{q}(\boldsymbol{Z}_t, \boldsymbol{S}_t) = \exp\left(\mathbb{E}_{q^*(\boldsymbol{\Phi})q^*(\bar{\boldsymbol{\pi}})}[\log p(\boldsymbol{X}|\boldsymbol{\Phi}, \boldsymbol{Z}_t) + \log p(\boldsymbol{Z}_t, \boldsymbol{S}_t|\bar{\boldsymbol{\pi}})]\right),$$
$$\hat{q}(\boldsymbol{Z}_t) = \exp\left(\sum_{d \in B_t} \tau_d y_d \mathbb{E}[\boldsymbol{w}]^\top \bar{\boldsymbol{z}}_d\right),$$

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 $\boldsymbol{Z}_t$ and $\boldsymbol{S}_t$. We take the ideas from (Wang and Blei, 2012b) and adopt an approximation for $\tilde{q}(\boldsymbol{Z}_t, \boldsymbol{S}_t)$:

$$\tilde{q}(\boldsymbol{Z}_t, \boldsymbol{S}_t) \approx \mathbb{E}_{q^*(\boldsymbol{\Phi})q^*(\bar{\boldsymbol{\pi}})}\left[p(\boldsymbol{X}|\boldsymbol{\Phi}, \boldsymbol{Z}_t)p(\boldsymbol{Z}_t, \boldsymbol{S}_t|\bar{\boldsymbol{\pi}})\right]. \tag{44}$$

Computing the expectation regarding $\bar{\boldsymbol{\pi}}$ in (44) turns out to be difficult. However, imagine that the expectation operator is essentially collapsing $\bar{\boldsymbol{\pi}}$ out from the joint distribution

$$\tilde{q}(\bar{\boldsymbol{\pi}}, \boldsymbol{Z}_t, \boldsymbol{S}_t) \approx \mathbb{E}_{q^*(\boldsymbol{\Phi})}\left[q^*(\bar{\boldsymbol{\pi}})p(\boldsymbol{X}|\boldsymbol{\Phi}, \boldsymbol{Z}_t)p(\boldsymbol{Z}_t, \boldsymbol{S}_t|\bar{\boldsymbol{\pi}})\right]. \tag{45}$$

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

$$q^*(\bar{\boldsymbol{\pi}}, \boldsymbol{Z}_t, \boldsymbol{S}_t) \propto \tilde{q}(\bar{\boldsymbol{\pi}}, \boldsymbol{Z}_t, \boldsymbol{S}_t)\hat{q}(\boldsymbol{Z}_t). \tag{46}$$

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

- **For $\boldsymbol{Z}_t$:** 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 (43) with $\boldsymbol{s}_d$ marginalized out for convenience.

$$q(z_{di} = k|\boldsymbol{Z}_t^{\neg di}, \bar{\boldsymbol{\pi}}) \propto \frac{(\alpha\pi_k + C_{dk}^{\neg di})(C_{kx_{di}}^{\neg di} + \Delta_{kx_{di}}^t)}{\sum_w (C_{kw}^{\neg di} + \Delta_{kw}^t)} \exp\left(\sum_{d \in B_t} n_d^{-1} y_d \tau_d \mu_k^*\right). \tag{47}$$

  Besides, for $k > K$ and symmetric Dirichlet prior $\boldsymbol{\gamma}$, (47) converge to a single rule $q(z_{di} = k|\boldsymbol{Z}_t^{\neg di}, \bar{\boldsymbol{\pi}}) \propto \alpha\pi_k/W$, and therefore the total probability of assigning a new topic is

$$q(z_{di} > K|\boldsymbol{Z}_t^{\neg di}, \bar{\boldsymbol{\pi}}) \propto \alpha\left(1 - \sum_{k=1}^{K} \pi_k\right)/W.$$

- **For $\boldsymbol{S}_t$:** The conditional distribution of s_{dk} given $(\boldsymbol{Z}_t, \bar{\boldsymbol{\pi}}, \boldsymbol{\lambda}_t)$ can be derived from the joint distribution (39):

$$q(s_{dk}|\boldsymbol{Z}_t, \bar{\boldsymbol{\pi}}) \propto S(n_d \bar{z}_{dk}, s_{dk})(\alpha \pi_k)^{s_{dk}} \tag{48}$$

- **For $\bar{\boldsymbol{\pi}}$:** It can be derived from (43) that given $(\boldsymbol{Z}_t, \boldsymbol{S}_t)$, each $\bar{\pi}_k$ follows the beta distribution,

$$\bar{\pi}_k \sim \text{Beta}(a_k, b_k), \tag{49}$$

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 data set as mini-batch ($|B| = D$), iterating the above steps till converge in fact solves the batch MedHDP problem Eq. (38). We call this batch version MedHDP$^{\text{ave}}$, and will use it as a baseline algorithm.

7.1.3 ONLINE GIBBS MEDHDP

For Gibbs MedHDP, the only difference is the loss functional ℓ_ϵ, 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 $\boldsymbol{\lambda}_t$. Then the local variables are $(\boldsymbol{Z}_t, \boldsymbol{S}_t, \boldsymbol{\lambda}_t)$ and the global variables are unchanged. We then use the mean field assumption $q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi}, \boldsymbol{H}_t) = q(\boldsymbol{w})q(\bar{\boldsymbol{\pi}})q(\boldsymbol{\Phi})q(\boldsymbol{H}_t)$ and compute the iterative steps as follows.

1. **Global Update:** The same as online MedHDP, except that the update rule for $\boldsymbol{w}$ is now (35) for the Gibbs classifier.

2. **Local Update:** This step involves drawing samples of the local variables. We develop a Gibbs sampler, which iteratively draws $\boldsymbol{S}_t$ from the local conditional in (48), draws $\bar{\boldsymbol{\pi}}$ from the conditional in (49), and draws the augmented variables $\boldsymbol{\lambda}_t$ from the conditional in (37). For $\boldsymbol{Z}_t$, we explain the sampling procedure in detail. Specifically, we infer $\boldsymbol{Z}_t$ through

$$q^*(\boldsymbol{Z}_t, \boldsymbol{S}_t, \boldsymbol{\lambda}_t, \bar{\boldsymbol{\pi}}) \propto \tilde{q}(\bar{\boldsymbol{\pi}}, \boldsymbol{Z}_t, \boldsymbol{S}_t)\hat{q}(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t) \tag{50}$$

where

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

The Gibbs sampling for each variable z_{di} is

$$q(z_{di} = k|\boldsymbol{Z}_t^{\neg di}, \boldsymbol{\lambda_t}, \bar{\boldsymbol{\pi}}) \propto \frac{(\alpha\pi_k + C_{dk}^{\neg di})(C_{kx_{di}}^{\neg di} + \Delta_{kx_{di}}^t)}{\sum_w (C_{kw}^{\neg di} + \Delta_{kw}^t)} \exp\left(\frac{cy_d(c\epsilon + \lambda_d)\mu_k^*}{n_d\lambda_d} - \frac{c^2(\mu_k^{*2} + \Sigma_{kk}^* + 2(\mu_k^*\boldsymbol{\mu}^* + \boldsymbol{\Sigma}_{\cdot,k}^*)^\top \boldsymbol{C}_d^{\neg di})}{2n_d^2\lambda_d} \right), \tag{51}$$

while the probability of sampling a new topic is

$$q(z_{di} > K | \boldsymbol{Z}_t^{\neg di}, \bar{\boldsymbol{\pi}}) \propto \alpha \left(1 - \sum_{k=1}^{K} \pi_k\right) / W.$$

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$^{\text{gibbs}}$.

7.2 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 an one 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 $\boldsymbol{x}_d$ belongs to a specific category $\boldsymbol{y}_d^\tau \in \{+1, -1\}$. These binary classifiers are allowed to share common latent representations and therefore could be attained via a modified BayesPA update equation:

$$\min_{q \in \mathcal{F}_t} \mathcal{L}\Big(q(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t)\Big) + 2c \sum_{\tau=1}^{\mathcal{T}} \ell_\epsilon\Big(q(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t); \boldsymbol{X}_t, \boldsymbol{Y}_t^\tau\Big)$$

where $\mathcal{T}$ is the total number of tasks. We can then derive the multi-task version of Passive-Aggressive topic models, denoted by paMedLDAmt$^{\text{ave}}$ and paMedLDAmt$^{\text{gibbs}}$, in a way similar as in Section 6. We can further develop the nonparametric multi-task MedLDA topic models in a way similar as in Section 7.1 and the online PA learning algorithms. We denote the nonparametric online models by paMedHDP$^{\text{ave}}$ and paMedHDP$^{\text{gibbs}}$, according to whether the task-specific classifier is averaging or Gibbs.

8. 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 data set. 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.

8.1 Classification on 20Newsgroup

We perform multi-class classification on the entire 20NG data set 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 paMedLDA$^{\text{ave}}$ and paMedLDA$^{\text{gibbs}}$ for online MedLDA and online Gibbs MedLDA respectively. The batch counterparts are MedLDA (Zhu et al., 2012) and Gibbs

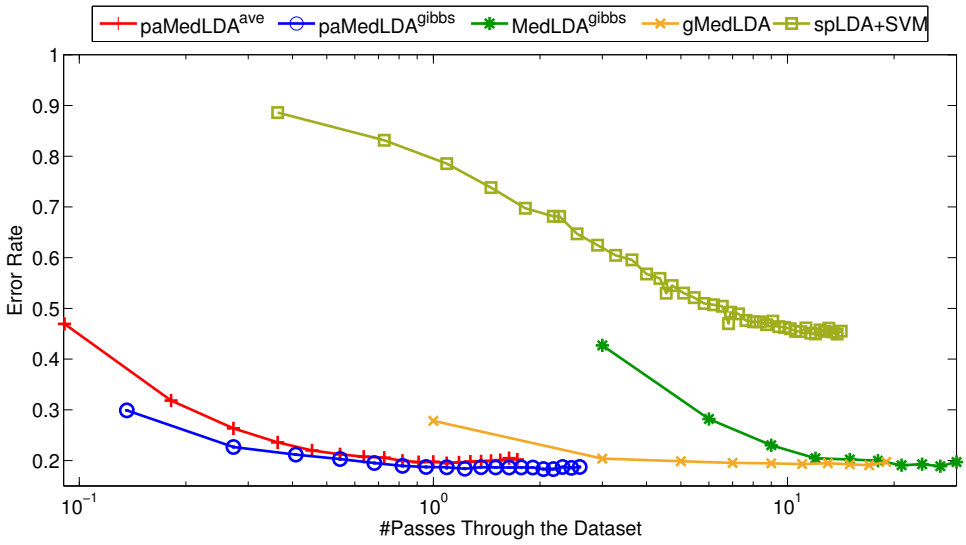

Figure 3: Test errors of different models with respect to the number of passes through the 20NG training data set.

MedLDA (MedLDA$^{\text{gibbs}}$) (Zhu et al., 2014a), 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). The performances of the other batch-learning-based supervised topic models, such as sLDA (Blei and McAuliffe, 2010) and DiscLDA (Lacoste-Julien et al., 2008), were reported in Zhu et al. (2012). For all the LDA-based topic models, we use symmetric Dirichlet priors $\boldsymbol{\alpha} = 1/K \cdot \mathbf{1}$ and $\boldsymbol{\gamma} = 0.45 \cdot \mathbf{1}$. 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. (2014a). 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 8.3, 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 $(\mathcal{I}, \mathcal{J}, \beta) = (1, 2, 0)$. As we can observe, by solving a series of latent BayesPA problems, both paMedLDA$^{\text{ave}}$ and paMedLDA$^{\text{gibbs}}$ fully explore the redundancy of documents and converge in less than one pass, while their batch counterparts (i.e., MedLDA and MedLDA$^{\text{gibbs}}$) 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

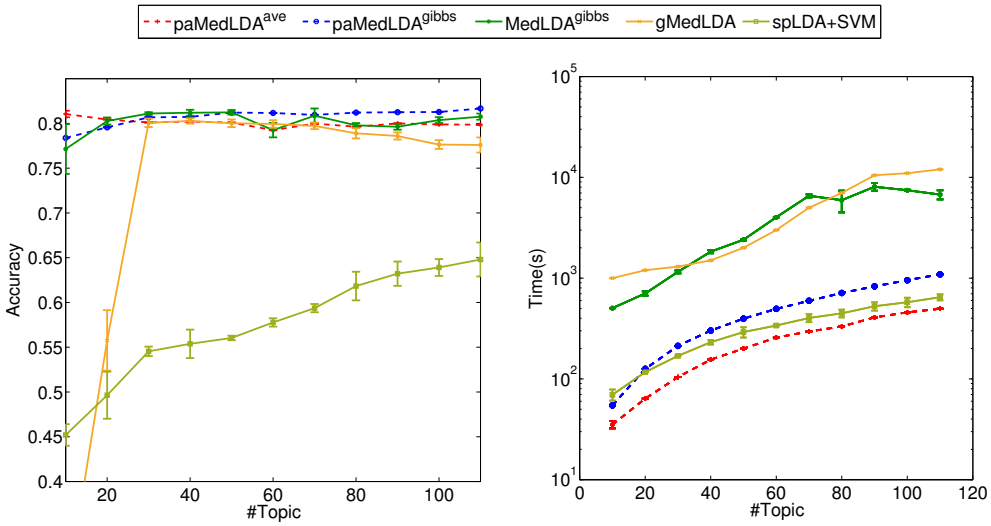

Figure 4: Classification accuracy and running time of various models with respect to the number of topics on the 20NG data set.

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 data set with varying numbers of topics. As we can see, BayesPA topic models, at the power of online learning, are more than order of magnitude faster than their batch counterparts in training time. paMedLDA$^{\text{ave}}$ is faster than paMedLDA$^{\text{gibbs}}$, because it does not need to update the covariance matrix of classifier weights $\boldsymbol{w}$. But the tradeoff is that for averaging models, they are more sensitive to the initial choice of σ^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.

Table 2 visualizes the learnt topic representation by paMedLDA$^{\text{ave}}$ and paMedLDA$^{\text{gibbs}}$. 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 $1\sim2$ 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 discriminatrive patterns are similar to what have been shown in batch settings (Zhu et al., 2012, 2014a).

Results by paMedLDA^{ave}							
Category	Visualization						
alt.altheism	#Observation	1	64	512	4096	11269	
	Topic distribution						
	Top words	T8	writes	article	don	people	time	good
		T11	god	people	writes	don	article	atheism
comp.graphics	#Observation	1	64	512	4096	11269	
	Topic distribution						
	Top words	T19	image	graphics	file	jpeg	files	software
		T13	writes	article	people	don	time	good
misc.forsale	#Observation	1	64	512	4096	11269	
	Topic distribution						
	Top words	T4	sale	offer	shipping	mail	dos	price
		T8	writes	article	people	don	time	good
Results by paMedLDA^{gibbs}							
Category	Visualization						
alt.altheism	#Observation	1	64	512	4096	11269	
	Topic distribution						
	Top words	T11	god	people	writes	article	don	time
		T16	god	people	don	writes	jesus	time
comp.graphics	#Observation	1	64	512	4096	11269	
	Topic distribution						
	Top words	T14	graphics	image	file	program	images	writes
		T13	entry	space	don	people	system	writes
misc.forsale	#Observation	1	64	512	4096	11269	
	Topic distribution						
	Top words	T17	sale	mail	dos	good	offer	shipping
		T18	don	writes	article	time	good	work

Table 2: Visualization of the learnt topics by paMedLDA^{ave} and paMedLDA^{gibbs} . See text for details.

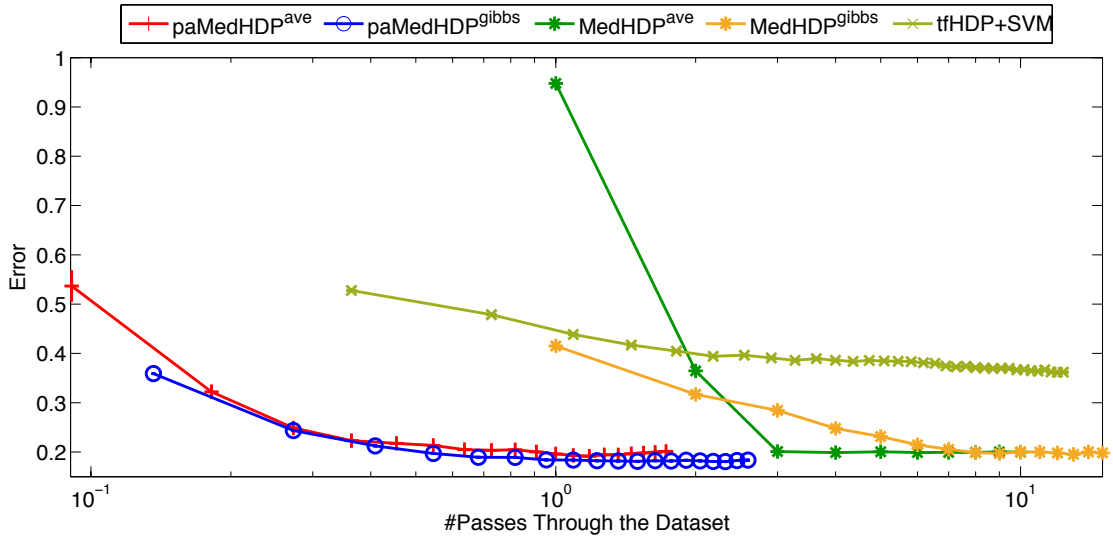

Figure 5: Test errors of different nonparametric models with respect to the number of passes through the 20NG training data set.

8.2 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 data set. Then we demonstrate multi-task learning on a large Wikipedia data set with more than 1 million documents and about 1 million unique terms.

8.2.1 NONPARAMETRIC TOPIC MODELING

Recall the nonparametric extensions of BayesPA topic models paMedHDP$^{\text{ave}}$ and paMedHDP$^{\text{gibbs}}$. To validate the advantage of online learning, we test them against their batch counterparts (i.e., the models MedHDP$^{\text{ave}}$ and MedHDP$^{\text{gibbs}}$) on the 20NG corpus. We also include an unsupervised baseline as 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: $\boldsymbol{\alpha} = 5 \cdot \mathbf{1}, \boldsymbol{\gamma} = 0.5 \cdot \mathbf{1}$ and $\eta = 1$. As an initial number of topic numbers 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 5 shows the convergence of paMedHDP$^{\text{ave}}$ and paMedHDP$^{\text{gibbs}}$, and Figure 6 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 their batch counterparts 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.

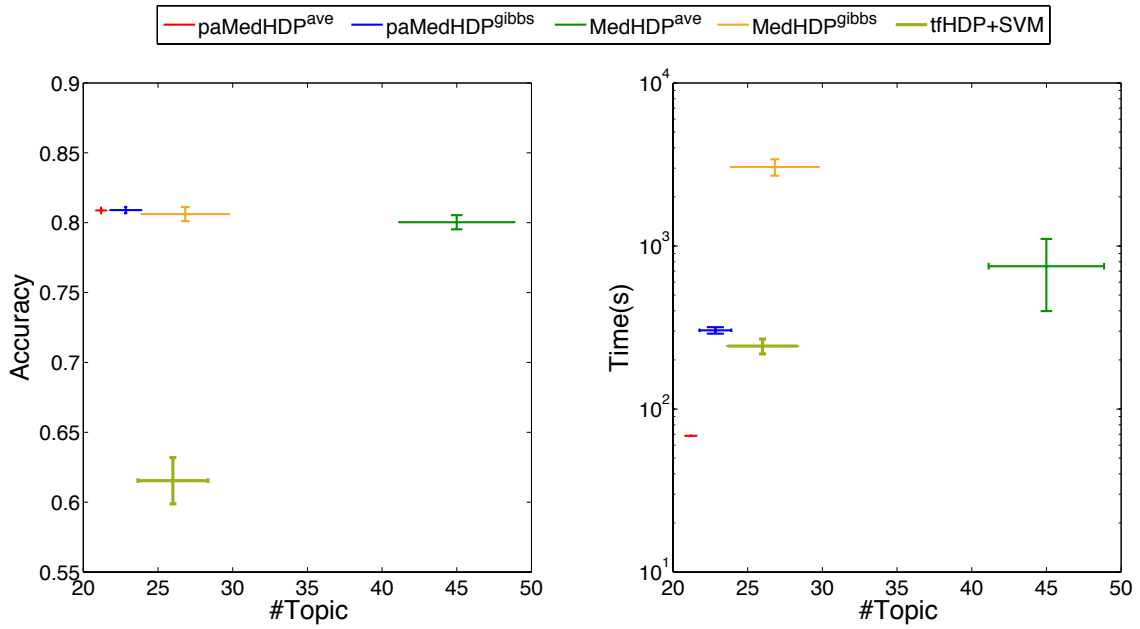

Figure 6: Classification accuracy and running time of the nonparametric paMedHDP and its batch counterpart models on the 20NG data set.

8.2.2 MULTI-TASK CLASSIFICATION

We test paMedLDAmt$^{\text{ave}}$, paMedLDAmt$^{\text{gibbs}}$ and their nonparametric exentions on a large Wiki data set. The Wiki data set is built from the Wikipedia set used in PASCAL LSHC challenge 2012 [5]. The Wiki data set 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 extention of Gibbs MedLDA (Zhu et al., 2013). Since MedHDP is a new model, there is no exising 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$^{\text{ave}}$; We choose $|B| = 512, c = 1, \sigma^2 = 1$ for paMedLDAmt$^{\text{gibbs}}$. For both models, the Dirichlet parameters are $\boldsymbol{\alpha} = 0.8 \cdot \mathbf{1}, \boldsymbol{\gamma} = 0.5 \cdot \mathbf{1}$, and $\epsilon = 1$. The nonparametric extensions use exactly the same parameter settings except that $\boldsymbol{\alpha} = 5 \cdot \mathbf{1}, \eta = 1$ and we do not need to specify the topic number K.

Figure 7 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

5. See http://lshtc.iit.demokritos.gr/.

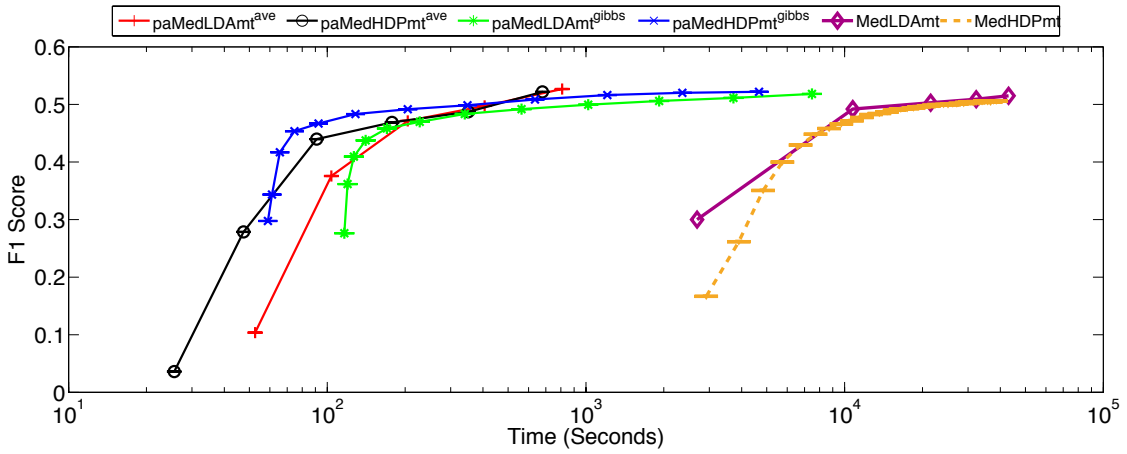

Figure 7: F1 scores of various multi-task topic models with respect to the running time on the 1.1M wikipedia data set.

magnitude faster than their batch counterparts. Therefore, BayesPA topic models could potentially be applied to large-scale multi-class settings.

8.3 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 8 presents the test errors of BayesPA topic models (paMedLDA$^{\text{ave}}$, paMedLDA$^{\text{gibbs}}$) as a function of training time on the entire 20NG data set 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 data set 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$^{\text{ave}}$ runs faster than paMedLDA$^{\text{gibbs}}$. This is because for averaging classifers, we do not update the covaraince of the classifier weights, which requires frequent matrix inverse operations. Furthermore, paMedLDA$^{\text{ave}}$ appears to be more robust against change in batchsize. Similarly, Figure 9 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 $\mathcal{I}$ and samples $\mathcal{J}$: Since the time complexity of Algorithm 2 is linear in both $\mathcal{I}$ and $\mathcal{J}$, we would like to know how these parameters influence the quality of the trained models. First, we analyze which setting of $(\mathcal{I}, \mathcal{J})$ guarantees good performance. Fix $\beta = 0, K = 40$. Figure 10 presents the results. First, the number of samples $\mathcal{J}$ does not have a large effect on the accuracy. Second, the peformance of paMedLDA$^{\text{gibbs}}$ and paMedHDP$^{\text{gibbs}}$ is not sensitive the

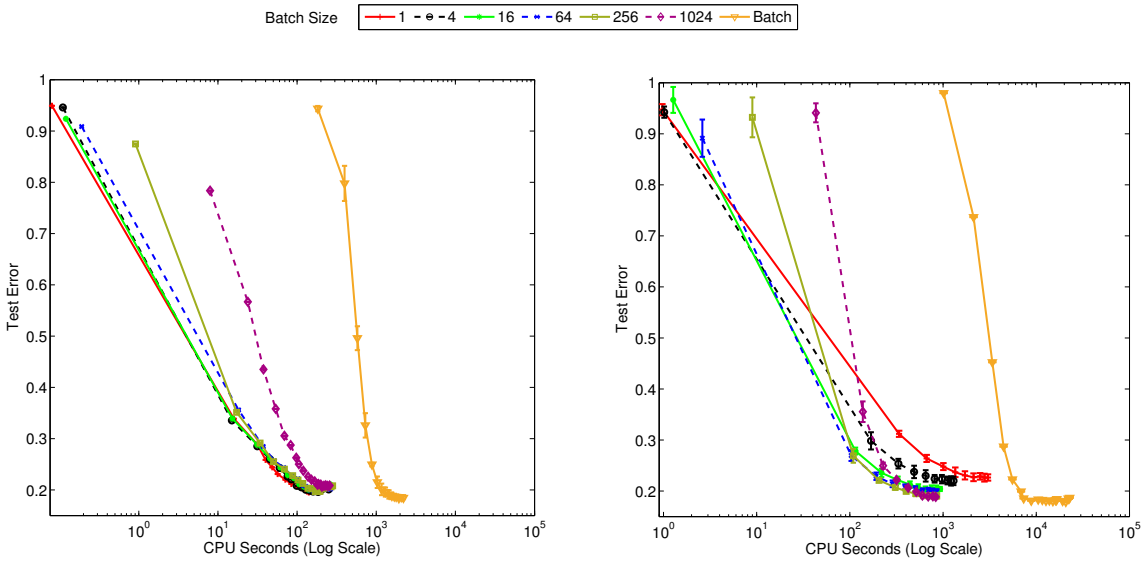

Figure 8: Test errors of paMedLDA$^{\text{ave}}$ (left) and paMedLDA$^{\text{gibbs}}$ (right) with different batch sizes on the 20NG data set.

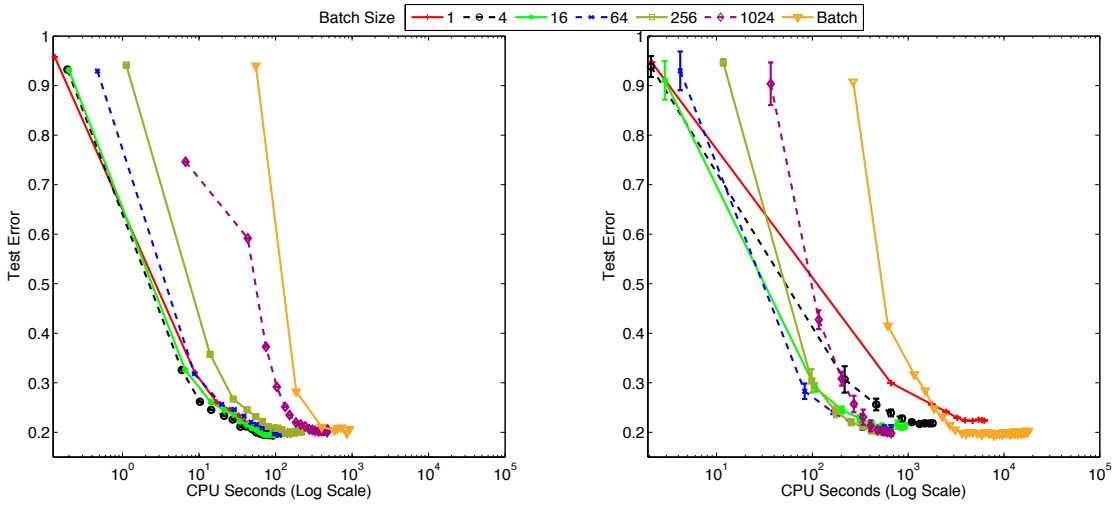

Figure 9: Test errors of paMedHDP$^{\text{ave}}$ (left) and paMedHDP$^{\text{gibbs}}$ (right) with different batch sizes on the 20NG data set.

number of optimization round, but paMedLDA$^{\text{ave}}$ and paMedHDP$^{\text{ave}}$ suffers largely if $\mathcal{I} = 1$. This is because with averaging classifiers, the sampling of latent variable $\boldsymbol{Z}$ relies not only on global parameters, but also on a local variable τ, so more optimization rounds are needed. The training time of all models scale linearly in terms of $\mathcal{I}$ and $\mathcal{J}$.

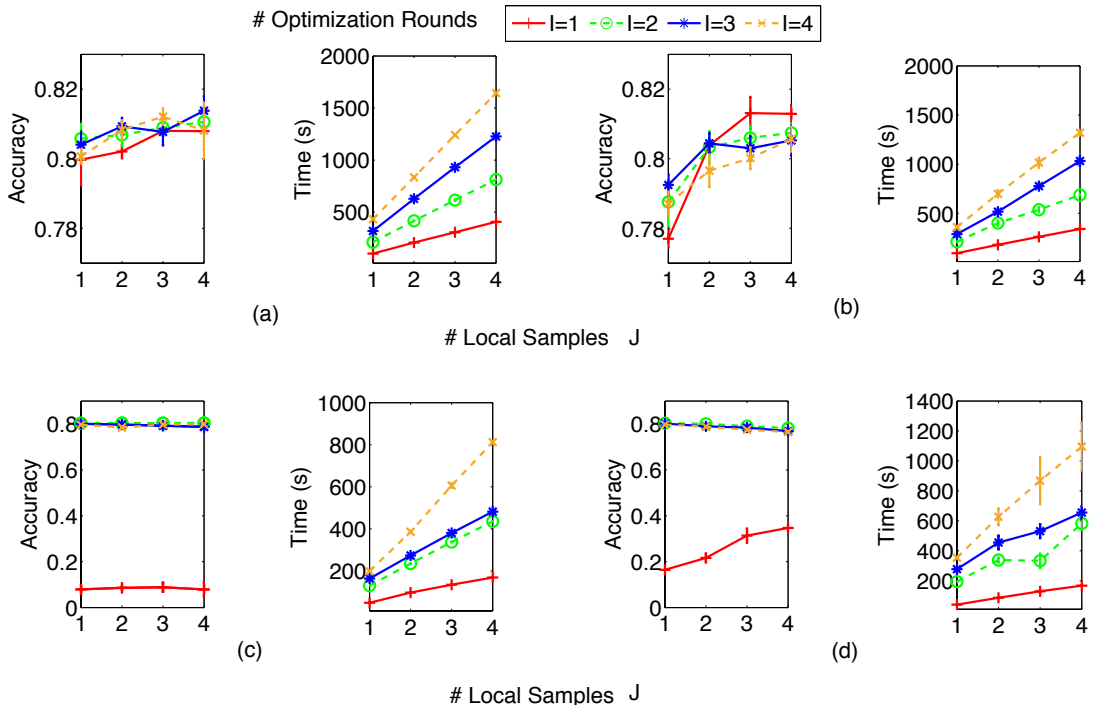

Figure 10: Classification accuracies and training time of **(a)**: paMedLDA$^{\text{gibbs}}$, **(b):** paMedHDP$^{\text{gibbs}}$, **(c)**: paMedLDA$^{\text{ave}}$, and **(d):** paMedHDP$^{\text{ave}}$, with different combinations of $(\mathcal{I}, \mathcal{J})$ on the 20NG data set.

$\mathcal{J}$ \ β	0	2	4
1	0.802		
3	0.803	0.803	
5	0.805	0.805	0.803

(a)

$\mathcal{J}$ \ β	0	2	4
1	0.783		
3	0.803	0.799	
5	0.808	0.803	0.792

(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 β samples are discarded as burn-in steps. To understand how large β is sufficient, we consider the settings of the pairs $(\mathcal{J}, \beta)$ and check the prediction accuracy of Algorithm 2 for $K = 40$. Based on the sensitivity analysis of $\mathcal{I}$ and $\mathcal{J}$, we fix $\mathcal{I} = 1$ for paMedLDA$^{\text{gibbs}}$, paMedHDP$^{\text{gibbs}}$ and $\mathcal{I} = 2$ 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. $(\mathcal{J}, \beta) = (3, 0), (5, 2), (9, 6)$, 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.

9. 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 data sets, 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. We provide provable regret bounds for the BayesPA models using either an averaging classifier or a Gibbs classifier. 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 data set 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).

Acknowledgements

This work is supported by National Key Foundation R&D Projects (No. 2013CB329403), National Natural Science Foundation of China (Nos. 61322308, 61332007, 61305066), Tsinghua University Initiative Scientific Research Program (No. 20121088071), and a Microsoft Research Asia Research Fund (No. 20123000007).

Appendix A.

We show the objective in (34) is an upper bound of that in (12), that is,

$$\begin{aligned} &\mathcal{L}\Big(q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t)\Big) - \mathbb{E}_q\Big[\log(\psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t | \boldsymbol{Z}_t, \boldsymbol{w}))\Big] \\ &\geq \mathcal{L}\Big(q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)\Big) + 2c \sum_{d \in B_t} \mathbb{E}_q\Big[(\xi_d)_+\Big], \end{aligned} \tag{52}$$

where $\mathcal{L}(q) = \mathbf{KL}[q||q_t(\boldsymbol{w}, \boldsymbol{\Phi})q_0(\boldsymbol{Z}_t)]$.

Proof We first have

$$\mathcal{L}\Big(q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t)\Big) = \mathbb{E}_q\left[\log \frac{q(\boldsymbol{\lambda}_t \mid \boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t) q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)}{q_t(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)}\right],$$

and

$$\mathcal{L}\Big(q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)\Big) = \mathbb{E}_q\left[\log \frac{q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)}{q_t(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)}\right].$$

Comparing these two equations and canceling out common factors, we know that in order for (52) to make sense, it suffices to prove

$$\mathbb{H}[q'] - \mathbb{E}_{q'}[\log(\psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t | \boldsymbol{Z}_t, \boldsymbol{w})] \geq 2c \sum_{d \in B_t} \mathbb{E}_{q'}[(\xi_d)_+] \tag{53}$$

is uniformly true for any given $(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)$, where $\mathbb{H}(\cdot)$ is the entropy operator and $q' = q(\boldsymbol{\lambda}_t \mid \boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)$. The inequality (53) can be reformulated as

$$\mathbb{E}_{q'}\left[\log \frac{q'}{\psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t | \boldsymbol{Z}_t, \boldsymbol{w})}\right] \geq 2c \sum_{d \in B_t} \mathbb{E}_{q'}[(\xi_d)_+] \tag{54}$$

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

$$-\mathbb{E}_{q'}\left[\log \frac{\psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t | \boldsymbol{Z}_t, \boldsymbol{w})}{q'}\right] \geq -\log \int_{\boldsymbol{\lambda}_t} \psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t | \boldsymbol{Z}_t, \boldsymbol{w})\, d\boldsymbol{\lambda}_t,$$

and utilizing the equality (33), we then have (54) and therefore prove (52). ■

References

Ryan Prescott Adams and David MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *International Conference on Machine Learning (ICML)*, 2012.

Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, pages 1152–1174, 1974.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *International Conference on Machine Learning (ICML)*, 2014.

David M. Blei and Jon D. McAuliffe. Supervised topic models. In *Neural Information Processing Systems (NIPS)*, 2010.

David M. Blei, Andrew Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael Jordan. Streaming variational bayes. In *Neural Information Processing Systems (NIPS)*, pages 1727–1735, 2013.

Kevin R. Canini, Lei Shi, and Thomas Griffiths. Online inference of topics with latent Dirichlet allocation. In *International Conference on Artificial Intelligence and Statistics*, pages 65–72, 2009.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

Ning Chen, Jun Zhu, Fei Xia, and Bo Zhang. Generalized relational topic models with data augmentation. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1273–1279. AAAI Press, 2013.

David Chiang, Yuval Marton, and Philip Resnik. Online large-margin training of syntactic and structural translation features. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2008.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalel-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research (JMLR)*, 7:551–585, 2006.

Koby Crammer, Mark Dredze, and Fernando Pereira. Exact convex confidence-weighted learning. In *Neural Information Processing Systems (NIPS)*, 2008.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977.

Luc Devroye. *Non-Uniform Random Variate Generation*. Springer, 1986.

Arnaud Doucet and Adam M Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12:656–704, 2009.

Arnaud Doucet, Nando De Freitas, Kevin Murphy, and Stuart Russell. Rao-Blackwellised particle filtering for dynamic Bayesian networks. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 176–183. Morgan Kaufmann Publishers Inc., 2000.

Mark Dredze, Koby Crammer, and Fernando Pereira. Confidence-weighted linear classification. In *International Conference on Machine Learning (ICML)*, pages 264–271, 2008.

Rick Durret. *Probabiilty: Theory and Examples*. Cambridge University Press, 2010.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Zoubin Ghahramani and Thomas Griffiths. Infinite latent feature models and the Indian buffet process. In *Neural Information Processing Systems (NIPS)*, 2005.

Thomas Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228–5235, 2004.

James Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

Nils Lid Hjort. *Bayesian Nonparametrics*. Cambridge University Press, 2010.

Matthew Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research (JMLR)*, 14(1):1303–1347, 2013.

Antti Honkela and Harri Valpola. Online variational Bayesian learning. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 803–808, 2003.

Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum entropy discrimination. In *Neural Information Processing Systems (NIPS)*, 1999.

Tony Jebara. *Discriminative, Generative and Imitative learning*. PhD thesis, Massachusetts Institute of Technology, 2001.

Tony Jebara. Multitask sparsity via maximum entropy discrimination. *Journal of Machine Learning Research (JMLR)*, 12:75–110, 2011.

Qixia Jiang, Jun Zhu, Maosong Sun, and Eric Xing. Monte Carlo methods for maximum margin supervised topic models. In *Neural Information Processing Systems (NIPS)*, 2012.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. *An Introduction to Variational Methods for Graphical Models*. Springer, 1998.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *International Conference on Machine Learning (ICML)*, 2014.

Simon Lacoste-Julien, Fei Sha, and Michael I Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems (NIPS)*, 2008.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

Jan Luts, Tamara Broderick, and Matt Wand. Real-time semiparametric regression. *arXiv:1209.3550*, 2013.

Ryan McDonald, Koby Crammer, and Fernando Pereira. Online large-margin training of dependency parsers. In *Annual Meeting of the Association for Computational Linguistics*, 2005.

John R. Michael, William R. Schucany, and Roy W. Haas. Generating random variates using transformations with multiple roots. *Journal of the American Statistician*, 30(2):88–90, 1976.

David Mimno, Matthew Hoffman, and David M. Blei. Sparse stochastic inference for latent Dirichlet allocation. *International Conference on Machine Learning (ICML)*, 2012.

Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence (UAI)*. Morgan Kaufmann Publishers Inc., 2001.

Kevin P. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012.

Sam Patterson and Yee Whye Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Neural Information Processing Systems (NIPS)*, pages 3102–3110, 2013.

Nicholas G. Polson and Steven L. Scott. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research (JMLR)*, 5:101–141, 2004.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

Shai Shalev-Shwartz and Yoram Singer. Convex repeated games and Fenchel duality. In *Neural Information Processing Systems (NIPS)*, pages 1265–1272, 2006.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical Programming*, 127(1):3–30, 2011.

Tianlin Shi and Jun Zhu. Online Bayesian passive-aggressive learning. In *International Conference on Machine Learning (ICML)*, 2014.

Jacob Steinhardt and Percy Liang. Filtering with abstract particles. In *International Conference on Machine Learning (ICML)*, 2014.

Robert H. Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58(2):86–88, 1987.

Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin Markov networks. In *Neural Information Processing Systems (NIPS)*, 2003.

Yee Whye Teh, M.I. Jordan, M.J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006a.

Yee Whye Teh, David Newman, and Max Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Neural Information Processing Systems (NIPS)*, 2006b.

David Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1), 2001.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Chong Wang and David M. Blei. A split-merge MCMC algorithm for the hierarchical Dirichlet process. *arXiv preprint arXiv:1201.1657*, 2012a.

Chong Wang and David M. Blei. Truncation-free online variational inference for Bayesian nonparametric models. In *Neural Information Processing Systems (NIPS)*, 2012b.

Chong Wang, John Paisley, and David M. Blei. Online variational inference for the hierarchical Dirichlet process. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning (ICML)*, pages 681–688, 2011.

Max Welling, Yee Whye Teh, and Hilbert J. Kappen. Hybrid variational/Gibbs collapsed inference in topic models. In *Uncertainty in Artificial Intelligence (UAI)*, 2008.

Minjie Xu, Jun Zhu, and Bo Zhang. Bayesian nonparametric maximum margin matrix factorization for collaborative prediction. In *Neural Information Processing Systems (NIPS)*, 2012.

Minjie Xu, Jun Zhu, and Bo Zhang. Fast max-margin matrix factorization with data augmentation. In *International Conference on Machine Learning (ICML)*, pages 978–986, 2013.

Jun Zhu. Max-margin nonparametric latent feature models for link prediction. In *International Conference on Machine Learning (ICML)*, 2012.

Jun Zhu and Eric P. Xing. Maximum entropy discrimination Markov networks. *Journal of Machine Learning Research (JMLR)*, 10:2531–2569, 2009.

Jun Zhu, Ning Chen, and Eric P Xing. Infinite SVM: a Dirichlet process mixture of large-margin kernel machines. In *International Conference on Machine Learning (ICML)*, pages 617–624, 2011.

Jun Zhu, A. Ahmed, and E.P Xing. MedLDA: maximum margin supervised topic models. *Journal of Machine Learning Research (JMLR)*, 13:2237–2278, 2012.

Jun Zhu, Xun Zheng, Li Zhou, and Bo Zhang. Scalable inference in max-margin topic models. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2013.

Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. Gibbs max-margin topic models with data augmentation. *Journal of Machine Learning Research (JMLR)*, 15:1073–1110, 2014a.

Jun Zhu, Ning Chen, and Eric P Xing. Bayesian inference with posterior regularization and applications to infinite latent SVMs. *Journal of Machine Learning Research (JMLR, in press; arXiv preprint, arXiv:1210.1766v3)*, 2014b.