Online Bayesian Passive-Aggressive Learning

Tianlin Shi[†] STL501@GMAIL.COM
Jun Zhu[‡] DCSZJ@MAIL.TSINGHUA.EDU.CN
[†]Institute for Interdisciplinary Information Sciences, Tsinghua University, China
[‡]Dept. of Comp. Sci. & Tech., TNList Lab, State Key Lab of Intell. Tech. & Sys., Tsinghua University, China

Abstract

Online Passive-Aggressive (PA) learning is an effective framework for performing max-margin online learning. But the deterministic formulation and estimated single large-margin model could limit its capability in discovering descriptive structures underlying complex data. This paper presents online Bayesian Passive-Aggressive (BayesPA) learning, which subsumes the online PA and extends naturally to incorporate latent variables and perform nonparametric Bayesian inference, thus providing great flexibility for explorative analysis. We apply BayesPA to topic modeling and derive efficient online learning algorithms for max-margin topic models. We further develop nonparametric methods to resolve the number of topics. Experimental results show that our approaches significantly improve time efficiency while maintaining comparable results with the batch counterparts.

1. Introduction

Online learning is an effective way to deal with large-scale applications, especially applications with streaming data. Online Passive-Aggressive (PA) learning (Crammer et al., 2006) provides a generic framework for online large-margin learning, with many applications (McDonald et al., 2005; Chiang et al., 2008). Though enjoying strong discriminative ability suitable for predictive tasks, existing 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, which may have rich underlying structures.

On the other hand, Bayesian methods enjoy great flexibility in describing the possible underlying structures of complex data. Moreover, the recent progress on nonparametric Bayesian methods (Hjort, 2010; Teh et al., 2006a) further provides an increasingly important framework that allows 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 & Griffiths, 2005), and to adapt when the learning environment changes. 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 online variational Bayes (Hoffman et al., 2010; Mimno et al., 2012) and online Monte Carlo (Ahn et al., 2012) methods. However, due to the generative nature, Bayesian models are lack of the discriminative ability of large-margin methods and 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, mainly in the context of supervised learning and structured output prediction (Zhu & Xing, 2009). Recently, much attention has been focused on generalizing MED to incorporate latent variables and perform nonparametric Bayesian inference, in many contexts including topic modeling (Zhu et al., 2012), matrix factorization (Xu et al., 2012), and multi-task learning (Jebara, 2011; Zhu et al., 2011). However, posterior inference in such models remains a big challenge. It is desirable to develop efficient online algorithms for these Bayesian max-margin models.

To address the above problems of both the online PA and Bayesian max-margin models, we present online Bayesian Passive-Aggressive (BayesPA) learning, a generic 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 lin-

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

ear and the parameter prior is Gaussian. We further show that one significance of BayesPA is its natural generalization to incorporate latent variables and to perform nonparametric Bayesian inference, thus allowing 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; Wang et al., 2011) for predictive tasks. Extensive empirical results on real data sets show significant improvements on time efficiency and maintenance of comparable results with the batch counterparts.

2. Bayesian Passive-Aggressive Learning

In this section, we present a general perspective on online max-margin Bayesian inference.

2.1. Online PA Learning

The goal of online learning is to minimize the cumulative loss for a certain prediction task from the sequentially arriving training samples. Online Passive-Aggressive (PA) algorithms (Crammer et al., 2006) achieve this goal by updating some parameterized model $\boldsymbol{w}$ (e.g., the weights of a linear SVM) in an online manner with the instantaneous losses from arriving data $\{\boldsymbol{x}_t\}_{t\geq 0}$ and corresponding responses $\{y_t\}_{t\geq 0}$. The losses $\ell_\epsilon(\boldsymbol{w}; \boldsymbol{x}_t, y_t)$ could be the hinge loss $(\epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t)_+$ for binary classification or the ϵ-insensitive loss $(|y_t - \boldsymbol{w}^\top \boldsymbol{x}_t| - \epsilon)_+$ for regression, where ϵ is a hyper-parameter and $(x)_+ = \max(0, x)$. The Passive-Aggressive update rule is then derived by defining the new weight $\boldsymbol{w}_{t+1}$ as the solution to the optimization problem:

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

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. With provable bounds, (Crammer et al., 2006) shows that online PA algorithms could achieve comparable results to the optimal classifier $\boldsymbol{w}^*$. In practice, in order to account for inseparable training samples, soft margin constraints are often adopted and the resulting learning problem is

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

where c is a positive regularization parameter. For problems (1) and (2) with samples arriving one at a time, closed-form solutions can be derived (Crammer et al., 2006).

2.2. Online BayesPA Learning

Instead of updating a point estimate of $\boldsymbol{w}$, online BayesPA sequentially infers a new 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:

$$\begin{aligned} \min_{q(\boldsymbol{w})\in\mathcal{F}_t} \ & \mathrm{KL}[q(\boldsymbol{w})||q_t(\boldsymbol{w})] - \mathbb{E}_{q(\boldsymbol{w})}[\log p(\boldsymbol{x}_t|\boldsymbol{w})] \\ \text{s.t.: } \ & \ell_\epsilon[q(\boldsymbol{w}); \boldsymbol{x}_t, y_t] = 0, \end{aligned} \tag{3}$$

where $\mathcal{F}_t$ is some distribution family, e.g., the probability simplex $\mathcal{P}$. In other words, we find a posterior distribution $q_{t+1}(\boldsymbol{w})$ in the feasible zone that is not only close to $q_t(\boldsymbol{w})$ by the commonly used 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. 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. We call it *non-likelihood* BayesPA.

In practical problems, the constraints in (3) 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 (3) with a regularization term (Cortes & Vapnik, 1995):

$$q_{t+1}(\boldsymbol{w}) = \operatorname*{argmin}_{q(\boldsymbol{w})\in\mathcal{F}_t} \mathcal{L}(q(\boldsymbol{w})) + 2c\ell_\epsilon(q(\boldsymbol{w}); \boldsymbol{x}_t, y_t). \tag{4}$$

For the max-margin classifiers that we focus on in this paper, two loss functions $\ell_\epsilon(q(\boldsymbol{w}); \boldsymbol{x}_t, y_t)$ are common — the hinge loss of an *averaging classifier* that makes predictions using the rule $\hat{y}_t = \text{sign}\, \mathbb{E}_{q(\boldsymbol{w})}[\boldsymbol{w}^\top \boldsymbol{x}_t]$:

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

and the expected hinge loss of a *Gibbs classifier* that randomly draws a classifier $\boldsymbol{w} \sim q(\boldsymbol{w})$ to make predictions using the rule $\hat{y}_t = \text{sign}\, \boldsymbol{w}^\top \boldsymbol{x}_t$:

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

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

Lemma 2.1. *The expected hinge loss ℓ_ϵ^{Gibbs} is an upper bound of the hinge loss ℓ_ϵ^{Avg}, that is, $\ell_\epsilon^{Gibbs} \geq \ell_\epsilon^{Avg}$.*

Before developing BayesPA learning for practical problems, we make several observations.

Lemma 2.2. *If $q_0(\boldsymbol{w}) = \mathcal{N}(0, I)$, $\mathcal{F}_t = \mathcal{P}$ and we use ℓ_ϵ^{Avg}, the non-likelihood BayesPA subsumes the online PA.*

This can be proved by induction. See Appendix A.

Lemma 2.3. *If $\mathcal{F}_t = \mathcal{P}$ and we use ℓ_ϵ^{Gibbs}, the update rule of online BayesPA is*

$$q_{t+1}(\boldsymbol{w}) = \frac{q_t(\boldsymbol{w})p(\boldsymbol{x}_t|\boldsymbol{w})e^{-2c(\epsilon - y_t \boldsymbol{w}^\top \boldsymbol{x}_t)_+}}{\Gamma(\boldsymbol{x}_t, y_t)}, \quad (5)$$

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

Therefore, the posterior $q_t(\boldsymbol{w})$ in the previous round t becomes a prior, while the newly observed data and its loss function provide a likelihood and an unnormalized pseudo-likelihood respectively.

Mini-Batches. A useful technique to reduce the noise in data is the use of mini-batches. Suppose that we have a mini-batch of data points at time t with an index set B_t, denoted as $\boldsymbol{X}_t = \{\boldsymbol{x}_d\}_{d \in B_t}, \boldsymbol{Y}_t = \{y_d\}_{d \in B_t}$. The Bayesian PA update equation for this mini-batch is simply,

$$q_{t+1}(\boldsymbol{w}) = \underset{q \in \mathcal{F}_t}{\operatorname{argmin}} \, \mathcal{L}(q(\boldsymbol{w})) + 2c\ell_\epsilon(q(\boldsymbol{w}); \boldsymbol{X}_t, \boldsymbol{Y}_t),$$

where $\ell_\epsilon(q(\boldsymbol{w}); \boldsymbol{X}_t, \boldsymbol{Y}_t) = \sum_{d \in B_t} \ell_\epsilon(q(\boldsymbol{w}); \boldsymbol{x}_d, y_d)$.

2.3. 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 two kinds of latent variables — *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.

The goal of Bayesian PA learning with latent structures is therefore to update the distribution of $\boldsymbol{\mathcal{M}}$ as well as 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 $\boldsymbol{H}_t$, we extend BayesPA to infer the joint posterior distribution, $q_{t+1}(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t)$, as solving

$$\min_{q \in \mathcal{F}_t} \mathcal{L}(q(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t)) + 2c\ell_\epsilon(q(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t); \boldsymbol{X}_t, \boldsymbol{Y}_t), (6)$$

where $\mathcal{L}(q) = \mathrm{KL}[q||q_t(\boldsymbol{w}, \boldsymbol{\mathcal{M}})p_0(\boldsymbol{H}_t)] - \mathbb{E}_q[\log p(\boldsymbol{X}_t|\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t)]$ and $\ell_\epsilon(q; \boldsymbol{X}_t, \boldsymbol{Y}_t)$ is some cumulative margin-loss on the min-batch data induced from some classifiers defined on the latent variables $\boldsymbol{H}_t$ and/or global variables $\boldsymbol{\mathcal{M}}$. Both the averaging classifiers and Gibbs classifiers can be used as in the case without latent variables. We will present concrete examples in the next sections.

Before diving into the details, we should note that in a real online setting, only global variables are maintained in the bookkeeping, while the local information in the streaming data is forgotten. However, (6) gives us a distribution of $(\boldsymbol{w}, \boldsymbol{\mathcal{M}})$ that is coupled with the local variables $\boldsymbol{H}_t$. Although in some cases we can marginalize out the local variables $\boldsymbol{H}_t$, in general we would not obtain a closed-form posterior distribution $q_{t+1}(\boldsymbol{w}, \boldsymbol{\mathcal{M}})$ for the next optimization round, especially in dealing with some involved models like MedLDA (Zhu et al., 2012). Therefore, we resort to approximation methods, e.g., by posing additional assumptions about $q(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t)$ such as the mean-field assumption, $q(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t) = q(\boldsymbol{w})q(\boldsymbol{\mathcal{M}})q(\boldsymbol{H}_t)$. Then, we can solve the problem via an iterative procedure and use the optimal distribution $q^*(\boldsymbol{w})q^*(\boldsymbol{\mathcal{M}})$ as $q_{t+1}(\boldsymbol{w}, \boldsymbol{\mathcal{M}})$. More details will be provided in next sections.

3. Online Max-Margin Topic Models

We apply the theory of online BayesPA to topic modeling and develop online learning algorithms for max-margin topic models. We also present a nonparametric generalization to resolve the number of topics in the next section.

3.1. Batch MedLDA

A max-margin topic model consists of a latent Dirichlet allocation (LDA) (Blei et al., 2003) model for describing the underlying topic representations and a max-margin classifier for predicting responses. Specifically, LDA is a hierarchical Bayesian model that treats each document as an admixture of topics, $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_k\}_{k=1}^K$, where each topic $\boldsymbol{\phi}_k$ is a multinomial distribution over a W-word vocabulary. Let $\boldsymbol{\theta}$ denote the mixing proportions. The generative process of document d is described as

$$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha}),$$
$$z_{di} \sim \mathrm{Mult}(\boldsymbol{\theta}_d), \; x_{di} \sim \mathrm{Mult}(\boldsymbol{\phi}_{z_{di}}), \; \forall i \in [n_d]$$

where z_{di} is a topic assignment variable and $\mathrm{Mult}(\cdot)$ is a multinomial distribution. For Bayesian LDA, the topics are drawn from a Dirichlet distribution, i.e., $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\boldsymbol{\gamma})$.

Given a document set $\boldsymbol{X} = \{\boldsymbol{x}_d\}_{d=1}^D$. Let $\boldsymbol{Z} = \{\boldsymbol{z}_d\}_{d=1}^D$ and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_d\}_{d=1}^D$. 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 posterior is equivalent to the solution of the optimization problem:

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

The advantage of the variational formulation of Bayesian inference lies in the convenience of posing restrictions on the post-data distribution with a regularization term. For supervised topic models (Blei & 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:

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

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

$$f(\boldsymbol{w}, \boldsymbol{z}_d) = \boldsymbol{w}^\top \bar{\boldsymbol{z}}_d, \tag{7}$$

where $\bar{z}_{dk} = \frac{1}{n_d}\sum_i \mathbb{I}[z_{di} = k]$ are the average topic assignments of the words in document d. Based on the discriminant function, both averaging classifiers with the hinge loss

$$\ell_\epsilon^{Avg}(q(\boldsymbol{w}, \boldsymbol{z}_d); \boldsymbol{x}_d, y_d) = (\epsilon - y_d \mathbb{E}_q[f(\boldsymbol{w}, \boldsymbol{z}_d)])_+, \tag{8}$$

and Gibbs classifiers with the expected hinge loss

$$\ell_\epsilon^{Gibbs}(q(\boldsymbol{w}, \boldsymbol{z}_d); \boldsymbol{x}_d, y_d) = \mathbb{E}_q[(\epsilon - y_d f(\boldsymbol{w}, \boldsymbol{z}_d))_+], \tag{9}$$

have been proposed, with extensive comparisons reported in (Zhu et al., 2013a) using batch learning algorithms.

3.2. Online MedLDA

To apply the online BayesPA, we have the global variables $\boldsymbol{\mathcal{M}} = \boldsymbol{\Phi}$ and local variables $\boldsymbol{H}_t = (\boldsymbol{\Theta}_t, \boldsymbol{Z}_t)$. We consider Gibbs MedLDA because as shown in (Zhu et al., 2013a) it admits efficient inference algorithms by exploring data augmentation. Specifically, let $\zeta_d = \epsilon - y_d f(\boldsymbol{w}, \boldsymbol{z}_d)$ and $\psi(y_d|\boldsymbol{z}_d, \boldsymbol{w}) = e^{-2c(\zeta_d)_+}$. Then in light of Lemma 2.3, the optimal solution to problem (6), $q_{t+1}(\boldsymbol{w}, \boldsymbol{\mathcal{M}}, \boldsymbol{H}_t)$, is

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

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. To potentially improve the inference accuracy, we first integrate out the local variables $\boldsymbol{\Theta}_t$ by the conjugacy between a Dirichlet prior and a multinomial likelihood (Griffiths & Steyvers, 2004; Teh et al., 2006b). Then we have the local variables $\boldsymbol{H}_t = \boldsymbol{Z}_t$. By the equality (Zhu et al., 2013a):

$$\psi(y_d|\boldsymbol{z}_d, \boldsymbol{w}) = \int_0^\infty \psi(y_d, \lambda_d|\boldsymbol{z}_d, \boldsymbol{w}) d\lambda_d, \tag{10}$$

where $\psi(y_d, \lambda_d|\boldsymbol{z}_d, \boldsymbol{w}) = (2\pi\lambda_d)^{-1/2}\exp(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d})$, the collapsed posterior $q_{t+1}(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t)$ is a marginal distribution of $q_{t+1}(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t)$, which equals to

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

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, which are also 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:

$$\min_{q\in\mathcal{P}} \mathcal{L}(q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t)) - \mathbb{E}_q[\log \psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t|\boldsymbol{Z}_t, \boldsymbol{w})], \tag{11}$$

where $\mathcal{L}(q) = \mathrm{KL}[q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t) \| q_t(\boldsymbol{w}, \boldsymbol{\Phi}) p_0(\boldsymbol{Z}_t)] - \mathbb{E}_q[\log p(\boldsymbol{X}_t|\boldsymbol{Z}_t, \boldsymbol{\Phi})]$. We can show that this objective is an upper bound of that in the original problem (6). See Appendix B for details.

With the mild mean-field assumption that $q(\boldsymbol{w}, \boldsymbol{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t) = q(\boldsymbol{w}) q(\boldsymbol{\Phi}) q(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t)$, we can solve (11) via an iterative procedure that alternately updates each factor distribution (Jordan et al., 1998), as detailed below.

Global Update: By fixing the distribution of local variables, $q(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t)$, and ignoring irrelevant variables, we have the mean-field update equations:

$$q(\boldsymbol{\Phi}_k) \propto q_t(\boldsymbol{\Phi}_k) \exp(\mathbb{E}_{q(\boldsymbol{Z}_t)}[\log p_0(\boldsymbol{Z}_t) p(\boldsymbol{X}|\boldsymbol{Z}_t, \boldsymbol{\Phi})]),\ \forall\ k$$
$$q(\boldsymbol{w}) \propto q_t(\boldsymbol{w}) \exp(\mathbb{E}_{q(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t)}[\log p_0(\boldsymbol{Z}_t) \psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t|\boldsymbol{Z}_t, \boldsymbol{w})]).$$

If initially $q_0(\boldsymbol{\Phi}_k) = \mathrm{Dir}(\Delta_{k1}^0, ..., \Delta_{kW}^0)$ and $q_0(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0)$, by induction we can show that the inferred distributions in each round have a closed form, namely, $q_t(\boldsymbol{\Phi}_k) = \mathrm{Dir}(\Delta_{k1}^t, ..., \Delta_{kW}^t)$ and $q_t(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t)$. For the above update equations, we have

$$q(\boldsymbol{\Phi}_k) = \mathrm{Dir}(\Delta_{k1}^*, ..., \Delta_{kW}^*), \tag{12}$$

where $\Delta_{kw}^* = \Delta_{kw}^t + \sum_{d\in B_t}\sum_{i\in[n_d]} \gamma_{di}^k \cdot \mathbb{I}[x_{di} = w]$ for all words w and $\gamma_{di}^k = \mathbb{E}_{q(\boldsymbol{z}_d)}\mathbb{I}[z_{di} = k]$ is the probability of assigning word x_{di} to topic k, and

$$q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \tag{13}$$

where the posterior paramters are computed as $(\boldsymbol{\Sigma}^*)^{-1} = (\boldsymbol{\Sigma}^t)^{-1} + c^2 \sum_{d\in B_t} \mathbb{E}_{q(\boldsymbol{z}_d, \lambda_d)}[\lambda_d^{-1} \bar{\boldsymbol{z}}_d \bar{\boldsymbol{z}}_d^\top]$ and $\boldsymbol{\mu}^* = \boldsymbol{\Sigma}^*(\boldsymbol{\Sigma}^t)^{-1}\boldsymbol{\mu}^t + \boldsymbol{\Sigma}^* \cdot c \sum_{d\in B_t} \mathbb{E}_{q(\boldsymbol{z}_d, \lambda_d)}[y_d(1 + c\epsilon\lambda_d^{-1})\bar{\boldsymbol{z}}_d]$.

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

$$q(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t) \propto p_0(\boldsymbol{Z}_t) \prod_{d\in B_t} \frac{1}{\sqrt{2\pi\lambda_d}} \exp\Big(\sum_{i\in[n_d]} \Lambda_{z_{di}, x_{di}} - \mathbb{E}_{q(\boldsymbol{\Phi}, \boldsymbol{w})}[\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}]\Big), \tag{14}$$

where $\Lambda_{z_{di}, x_{di}} = \mathbb{E}_{q(\Phi)}[\log(\Phi_{z_{di}, x_{di}})] = \Psi(\Delta_{z_{di}, x_{di}}^*) - \Psi(\sum_w \Delta_{z_{di}, w}^*)$ and $\Psi(\cdot)$ is the digamma function, due to the distribution in (12). But it is impossible to evaluate the expectation in the global update using (14) 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

$$q(z_{di} = k|\boldsymbol{Z}_t^{\neg di}, \boldsymbol{\lambda}_t) \propto (\alpha + C_{dk}^{\neg di}) \exp\Big(\frac{c y_d (c\epsilon + \lambda_d)\mu_k^*}{n_d \lambda_d} + \Lambda_{k, x_{di}} - \frac{c^2(\mu_k^{*2} + \Sigma_{kk}^* + 2(\mu_k^* \boldsymbol{\mu}^* + \boldsymbol{\Sigma}_{\cdot,k}^*)^\top \boldsymbol{C}_d^{\neg di})}{2 n_d^2 \lambda_d}\Big), \tag{15}$$

Algorithm 1 Online MedLDA

```
Let q_0(w) = N(0; v^2 I), q_0(φ_k) = Dir(γ), ∀ k.
for t = 0 → ∞ do
    Set q(Φ, w) = q_t(Φ, w). Initialize Z_t.
    for i = 1 → I do
        Draw samples {Z_t^(j), λ_t^(j)}_{j=1}^J from (15, 16).
        Discard the first β burn-in samples (β < J).
        Use the rest J − β samples to update q(Φ, w)
        following (12, 13).
    end for
    Set q_{t+1}(Φ, w) = q(Φ, w).
end for
```

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

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

$$\begin{aligned} q(\lambda_d|\boldsymbol{Z}_t) &\propto \tfrac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\tfrac{c^2 \bar{\boldsymbol{z}}_d^\top \mathbf{\Sigma}^* \bar{\boldsymbol{z}}_d + (\lambda_d + c\bar{\zeta}_d)^2}{2\lambda_d}\right) \\ &= \mathcal{GIG}\left(\lambda_d; \tfrac{1}{2}, 1, c^2(\bar{\zeta}_d^2 + \bar{\boldsymbol{z}}_d^\top \mathbf{\Sigma}^* \bar{\boldsymbol{z}}_d)\right), \end{aligned} \tag{16}$$

a generalized inverse gaussian distribution (Devroye, 1986). Therefore, λ_d^{-1} follows an inverse gaussian distribution $\mathcal{IG}(\lambda_d^{-1}; \frac{1}{c\sqrt{\bar{\zeta}_d^2 + \bar{\boldsymbol{z}}_d^\top \mathbf{\Sigma}^* \bar{\boldsymbol{z}}_d}}, 1)$, 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. 1. 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 in (Zhu et al., 2013a).

4. Online Nonparametric MedLDA

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

4.1. Batch MedHDP

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 & Blei, 2012), where the stick lengths $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^\infty$ are generated as:

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

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

To augment the HDP topic model for predictive tasks, we introduce a classifier $\boldsymbol{w}$ and define the linear discriminant function in the same form as (7), where we should note that 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. Therefore, the dot product in (7) is in fact finite. Let $\bar{\boldsymbol{\pi}} = \{\bar{\pi}_k\}_{k=1}^\infty$. We define MedHDP as solving the following problem to infer the joint posterior $q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \mathbf{\Phi}, \mathbf{\Theta}, \boldsymbol{Z})$[1]:

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

where $\mathcal{L}(q(\boldsymbol{w}, \mathbf{\Phi}, \bar{\boldsymbol{\pi}}, \mathbf{\Theta}, \boldsymbol{Z})) = \text{KL}[q(\boldsymbol{w}, \mathbf{\Phi}, \bar{\boldsymbol{\pi}}, \mathbf{\Theta}, \boldsymbol{Z})||p(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \mathbf{\Phi}, \mathbf{\Theta}, \boldsymbol{Z}|\boldsymbol{X})]$, and the loss function could be either (8) or (9), leading to the MedHDP topic models with either averaging or Gibbs classifiers.

4.2. Online MedHDP

To apply the online BayesPA, we have the global variables $\boldsymbol{\mathcal{M}} = (\bar{\boldsymbol{\pi}}, \mathbf{\Phi})$, and the local variables $\boldsymbol{H}_t = (\mathbf{\Theta}_t, \boldsymbol{Z}_t)$. We again focus on the expected hinge loss (9) in this paper. As in online MedLDA, we marginalize out $\mathbf{\Theta}_t$ and adopt the same data augmentation technique with the augmented variables $\boldsymbol{\lambda}_t$. Furthermore, to simplify the sampling scheme, we introduce auxiliary latent variables $\boldsymbol{S}_t = \{\boldsymbol{s}_d\}_{d\in B_t}$, where s_{dk} represents the number of occupied tables serving dish k in a Chinese Restaurant Process (Teh et al., 2006a; Wang & Blei, 2012). 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

$$p(\boldsymbol{s}_d, \boldsymbol{z}_d|\bar{\boldsymbol{\pi}}) \propto \prod_{k=1}^\infty S(n_d\bar{z}_{dk}, s_{dk})(\alpha\pi_k)^{s_{dk}}, \tag{17}$$

where $S(a,b)$ are unsigned Stirling numbers of the first kind (Antoniak, 1974). It is not hard to verify that $p(\boldsymbol{z}_d|\bar{\boldsymbol{\pi}}) = \sum_{\boldsymbol{s}_d} p(\boldsymbol{s}_d, \boldsymbol{z}_d|\bar{\boldsymbol{\pi}})$. Therefore, we have local variables $\boldsymbol{H}_t = (\boldsymbol{Z}_t, \boldsymbol{S}_t, \boldsymbol{\lambda}_t)$, and the target collapsed posterior $q_{t+1}(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \mathbf{\Phi}, \boldsymbol{Z}_t, \boldsymbol{\lambda}_t)$ is the marginal distribution of $q_{t+1}(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \mathbf{\Phi}, \boldsymbol{H}_t)$, which is the solution of the problem:

$$\min_{q\in\mathcal{F}_t} \mathcal{L}(q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \mathbf{\Phi}, \boldsymbol{H}_t)) - \mathbb{E}_q[\log\psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t|\boldsymbol{Z}_t, \boldsymbol{w})], \tag{18}$$

where $\mathcal{L}(q) = \text{KL}[q||q_t(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \mathbf{\Phi})p(\boldsymbol{Z}_t, \boldsymbol{S}_t|\bar{\boldsymbol{\pi}})p(\boldsymbol{X}_t|\boldsymbol{Z}_t, \mathbf{\Phi})]$. As in online MedLDA, we solve (18) via an iterative procedure detailed below.

Global Update: By fixing the distribution of local variables, $q(\boldsymbol{Z}_t, \boldsymbol{S}_t, \boldsymbol{\lambda}_t)$, and ignoring the irrelevant terms, we have the mean-field update equations for $\mathbf{\Phi}$ and $\boldsymbol{w}$, the same as in (12) and (13), while for $\bar{\boldsymbol{\pi}}$, we have

$$q(\bar{\pi}_k) \propto q_t(\bar{\pi}_k) \prod_{d\in B_t} \exp(\mathbb{E}_{q(\boldsymbol{h}_d)}[\log p(\boldsymbol{s}_d, \boldsymbol{z}_d|\bar{\boldsymbol{\pi}})]). \tag{19}$$

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

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

$$q(\bar{\pi}_k) = \text{Beta}(u_k^*, v_k^*), \tag{20}$$

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 $k = \{1, 2, ...\}$. Since $\boldsymbol{Z}_t$ contains only finite number of discrete variables, we only need to maintain and update the global distribution for a finite number of topics.

Local Update: Fixing the global distribution $q(\boldsymbol{w}, \bar{\boldsymbol{\pi}}, \boldsymbol{\Phi})$, we get the mean-field update equation for $(\boldsymbol{Z}_t, \boldsymbol{S}_t, \boldsymbol{\lambda}_t)$:

$$q(\boldsymbol{Z}_t, \boldsymbol{S}_t, \boldsymbol{\lambda}_t) \propto \tilde{q}(\boldsymbol{Z}_t, \boldsymbol{S}_t)\tilde{q}(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t) \tag{21}$$

where $\tilde{q}(\boldsymbol{Z}_t, \boldsymbol{S}_t) = \exp(\mathbb{E}_{q(\boldsymbol{\Phi})q(\bar{\boldsymbol{\pi}})}[\log p(\boldsymbol{X}_t|\boldsymbol{\Phi}, \boldsymbol{Z}_t) + \log p(\boldsymbol{Z}_t, \boldsymbol{S}_t|\bar{\boldsymbol{\pi}})])$ and $\tilde{q}(\boldsymbol{Z}_t, \boldsymbol{\lambda}_t) = \exp(\mathbb{E}_{q(\boldsymbol{w})}[\log \psi(\boldsymbol{Y}_t, \boldsymbol{\lambda}_t|\boldsymbol{w}, \boldsymbol{Z}_t)])$. To overcome the the potentially unbounded latent space, we take the ideas from (Wang & Blei, 2012) and adopt an approximation for $\tilde{q}(\boldsymbol{Z}_t, \boldsymbol{S}_t)$:

$$\tilde{q}(\boldsymbol{Z}_t, \boldsymbol{S}_t) \approx \mathbb{E}_{q(\boldsymbol{\Phi})q(\bar{\boldsymbol{\pi}})}[p(\boldsymbol{X}|\boldsymbol{\Phi}, \boldsymbol{Z}_t)p(\boldsymbol{Z}_t, \boldsymbol{S}_t|\bar{\boldsymbol{\pi}})]. \tag{22}$$

Instead of marginalizing out $\bar{\boldsymbol{\pi}}$ in (22), which is analytically difficult, we sample $\bar{\boldsymbol{\pi}}$ jointly with $(\boldsymbol{Z}_t, \boldsymbol{S}_t, \boldsymbol{\lambda}_t)$. This leads to the following Gibbs sampling scheme:

For $\boldsymbol{Z}_t$: Let K be the current inferred number of topics. The conditional distribution of one variable z_{di} given $\boldsymbol{Z}_t^{\neg di}$, $\boldsymbol{\lambda}_t$ and $\bar{\boldsymbol{\pi}}$ can be derived from (21) with $\boldsymbol{s}_d$ marginalized out for convenience:

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

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

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

For $\boldsymbol{\lambda}_t$: The conditional distribution $q(\lambda_d|\boldsymbol{Z}_t, \boldsymbol{S}_t, \bar{\boldsymbol{\pi}})$ is the same as (16).

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 (17):

$$q(s_{dk}|\boldsymbol{Z}_t, \boldsymbol{\lambda}_t, \bar{\boldsymbol{\pi}}) \propto S(n_d\bar{z}_{dk}, s_{dk})(\alpha\pi_k)^{s_{dk}} \tag{23}$$

For $\bar{\boldsymbol{\pi}}$: It can be derived from (21) that given $(\boldsymbol{Z}_t, \boldsymbol{S}_t, \boldsymbol{\lambda}_t)$, each $\bar{\pi}_k$ follows the beta distribution, $\bar{\pi}_k \sim \text{Beta}(a_k, b_k)$, 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.

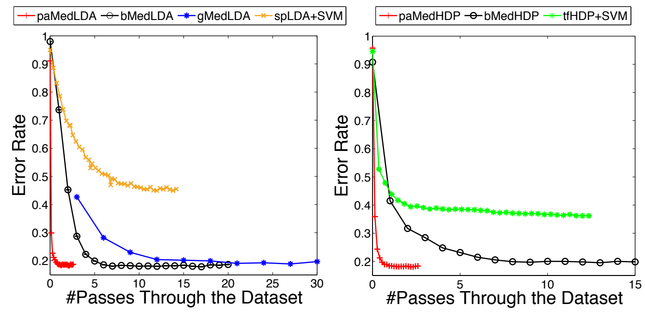

Figure 1. Test errors with different number of passes through the 20NG training dataset. **Left:** LDA-based models. **Right:** HDP-based models.

5. Experiments

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

5.1. Classification on 20Newsgroup

We perform multi-class classification on the 20NG dataset with all the 20 categories. The training set contains 11,269 documents, with the smallest category having 376 documents and the biggest category having 599 documents. The test set contains 7,505 documents, with the smallest and biggest categories having 259 and 399 documents respectively. We use the "one-vs-all" strategy (Rifkin & Klautau, 2004) for multi-class classification.

We compare paMedLDA and paMedHDP with their batch counterparts, denoted as *bMedLDA* and *bMedHDP*, which are obtained by letting the batch size $|B|$ be equal to the dataset size D, and Gibbs MedLDA, denoted as *gMedLDA*, (Zhu et al., 2013a), which performs Gibbs sampling in the batch manner. We also consider online unsupervised topic models as baselines, including sparse inference for LDA (*spLDA*) (Mimno et al., 2012), which has been demonstrated to be superior than online variational LDA (Hoffman et al., 2010) in performance, and truncation-free online variational HDP (*tfHDP*) (Wang & Blei, 2012), which has been shown to be promising in nonparametric topic modeling. For both of them, we learn a linear SVM with the topic representations using LIBSVM (Chang & Lin, 2011). The performances of other batch supervised topic models, such as sLDA (Blei & McAuliffe, 2010) and DiscLDA (Lacoste-Julien et al., 2008), are reported in (Zhu et al., 2012). For all LDA-based topic models, we use symmetric Dirichlet priors $\boldsymbol{\alpha} = 1/K \cdot \mathbf{1}, \boldsymbol{\gamma} = 0.5 \cdot \mathbf{1}$; for all HDP-based topic models, we use $\alpha = 5, \gamma = 1, \boldsymbol{\eta} = 0.45 \cdot$

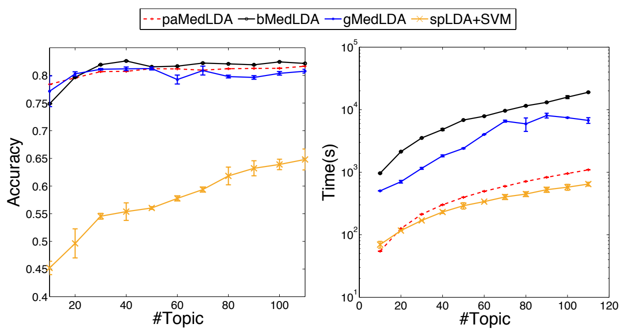

Figure 2. Classification accuracy and running time of paMedLDA and comparison models on the 20NG dataset.

1; for all MED topic models, we use $\epsilon = 164, c = 1, v = 1$, the choice of which is not crucial to the models' performance as shown in (Zhu et al., 2013a).

We first analyze how many processed documents are sufficient for each model to converge. Figure 1 shows the prediction accuracy with the number of passes through the entire 20NG dataset, where $K = 80$ for parametric models and $(\mathcal{I}, \mathcal{J}, \beta) = (1, 2, 0)$ for BayesPA. As we could observe, by solving a series of latent BayesPA learning problems, paMedLDA and paMedHDP fully explore the redundancy of documents and converge in one pass, while their batch counterparts need many passes as burn-in steps. Besides, compared with the online unsupervised learning algorithms, BayesPA topic models utilize supervising-side information from each mini-batch, and therefore exhibit a faster convergence rate in discrimination ability.

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 2 shows the accuracy and training time of LDA-based models on the whole dataset with varying numbers of topics. Similarly, Figure 3 shows the accuracy and training time of HDP-based models, where the dots stand for the mean inferred numbers of topics, and the lengths of the horizontal bars represent their standard deviations. As we can see, BayesPA topic models, at the power of online learning, are about 1 order of magnitude faster than their batch counterparts in training time. Furthermore, thanks to the merits of Gibbs sampling, which does not pose strict mean-field assumptions about the independence of latent variables, BayesPA topic models parallel their batch alternatives in accuracy.

5.2. Further Discussions

We provide further discussions on BayesPA learning for topic models. First, we analyze the models' sensitivity to some key parameters. Second, we illustrate an application with a large Wikipedia dataset containing 1.1 million documents, where class labels are not exclusive.

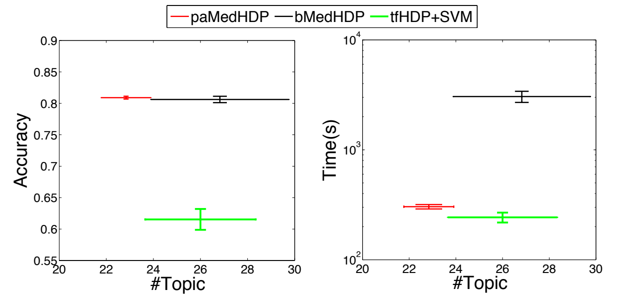

Figure 3. Classification accuracy and running time of paMedHDP and comparison models on the 20NG dataset.

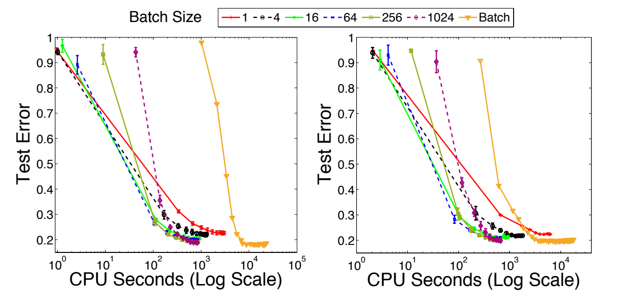

Figure 4. Test errors of paMedLDA (left) and paMedHDP (right) with different batch sizes on the 20NG dataset.

5.2.1. Sensitivity Analysis

Batch Size $|B|$: Figure 4 presents the test errors of BayesPA topic models as a function of training time on the entire 20NG dataset with various batch sizes, where $K = 40$. We can see that the convergence speeds of different algorithms vary. First of all, the batch algorithms suffer from multiple passes through the dataset and therefore are much slower than the online alternatives. Second, we could observe that algorithms with medium batch sizes ($|B| = 64, 256$) 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 reduces.

Number of iterations $\mathcal{I}$ and samples $\mathcal{J}$: Since the time complexity of Algorithm 1 is linear in both $\mathcal{I}$ and $\mathcal{J}$, we would like to know how these parameters influence the quality of the trained model. First, 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 1 for $K = 40, |B| = 512$, as shown in Table 1.

We can see that accuracies 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. The number of kept samples exhibits a more significant role in the performance of BayesPA topic models than the burn-in steps.

Table 1. Effect of the number of samples and burn-in steps.

$\mathcal{J}$ \ β	0	2	4	6	8
1	0.783				
3	0.803	0.799			
5	0.808	0.803	0.792		
9	0.806	0.806	0.806	0.804	0.796

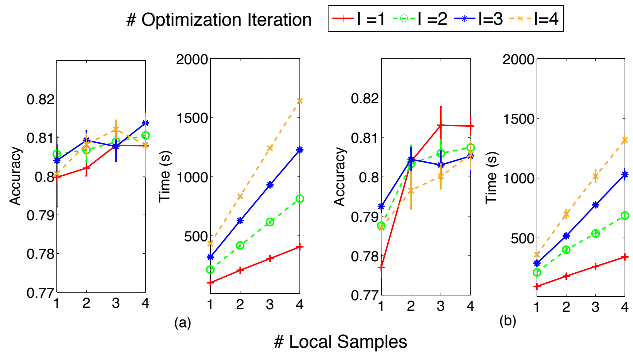

Figure 5. Classification accuracies and training time of **(a)**: paMedLDA, **(b)**: paMedHDP, with different combinations of $(\mathcal{I}, \mathcal{J})$ on the 20NG dataset.

Next, we analyze which setting of $(\mathcal{I}, \mathcal{J})$ guarantees good performance. Figure 5 presents the results. As we can see, for $\mathcal{J} = 1$, the algorithms suffer from the noisy approximation and therefore sacrifices prediction accuracy. But for larger $\mathcal{J}$, simply $\mathcal{I} = 1$ is promising, possibly due to the redundancy among mini-batches.

5.2.2. Multi-Task Classification

For multi-task classification, a set of binary classifiers are trained, each of which identifies whether a document $\boldsymbol{x}_d$ belongs to a specific task/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:

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

where $\mathcal{T}$ is the total number of tasks. We can then derive the multi-task version of Passive-Aggressive topic models, denoted by *paMedLDA-mt* and *paMedHDP-mt*, in a way similar to Section 3.2 and 4.2.

We test paMedLDA-mt and paMedHDP-mt as well as comparison models, including bMedLDA-mt, bMedHDP-mt and gMedLDA-mt (Zhu et al., 2013b) on a large Wiki dataset built from the Wikipedia set used in PASCAL LSHC challenge 2012 [2]. The Wiki dataset is a collection of documents with labels up to 20 different kinds, while the data distribution among the labels is balanced. The training/testing split is 1.1 million / 5 thousand. To measure performance, we use F1 score, the harmonic mean of precision and recall.

[2] See http://lshtc.iit.demokritos.gr/.

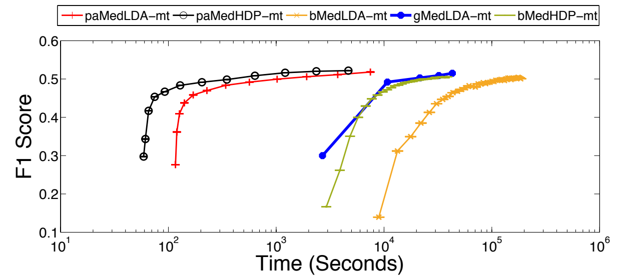

Figure 6. F1 scores of various models on the 1.1M wikipedia dataset.

Figure 6 shows the F1 scores of various models as a function of training time. We can see that BayesPA topic models are again about 1 order of magnitude faster than the batch alternatives and yet produce comparable results. Therefore, BayesPA topic models are potentially extendable to large-scale multi-class settings.

6. Conclusions and Future Work

We present online Bayesian Passive-Aggressive (BayesPA) learning as a new framework for max-margin Bayesian inference of online streaming data. We show that BayesPA subsumes the online PA, and more significantly, generalizes naturally to incorporate latent variables and to perform nonparametric Bayesian inference, therefore providing great flexibility for explorative analysis. Based on the ideas of BayesPA, we develop efficient online learning algorithms for max-margin topic models as well as their nonparametric extensions. Empirical experiments on several real datasets demonstrate significant improvements on time efficiency, while maintaining comparable results.

As future work, we are interested in showing provable bounds in its convergence and limitations. Furthermore, better understanding its mathematical structure would allow one to design more involved BayesPA algorithms for various models. We are also 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.

Acknowledgments

The work is supported by the National Basic Research Program of China (No. 2013CB329403), National Natural Science Foundation of China (Nos. 61322308, 61332007), and Tsinghua University Initiative Scientific Research Program (No. 20121088071).

References

Ahn, S., Korattikara, A., and Welling, M. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *ICML*, 2012.

Antoniak, C.E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, pp. 1152–1174, 1974.

Blei, D.M. and McAuliffe, J.D. Supervised topic models. In *NIPS*, 2010.

Blei, D.M., Ng, A., and Jordan, M.I. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

Broderick, T., Boyd, N., Wibisono, A., Wilson, A.C., and Jordan, M.I. Streaming variational Bayes. *arXiv preprint arXiv:1307.6769*, 2013.

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

Chiang, D., Marton, Y., and Resnik, P. Online large-margin training of syntactic and structural translation features. In *EMNLP*, 2008.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Crammer, K., Dekel, O., Keshet, J., Shalel-Shwartz, S., and Singer, Y. Online Passive-Aggressive learning. *JMLR*, 7:551–585, 2006.

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

Ghahramani, Z. and Griffiths, T.L. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2005.

Griffiths, T.L. and Steyvers, M. Finding scientific topics. *PNAS*, 101:5228–5235, 2004.

Hjort, N.L. *Bayesian Nonparametrics*. Cambridge University Press, 2010.

Hoffman, M., Bach, F.R., and Blei, D.M. Online learning for latent Dirichlet allocation. In *NIPS*, 2010.

Jaakkola, T., Meila, M., and Jebara, T. Maximum entropy discrimination. In *NIPS*, 1999.

Jebara, T. Multitask sparsity via maximum entropy discrimination. *JMLR*, 12:75–110, 2011.

Jiang, Q., Zhu, J., Sun, M., and Xing, E.P. Monte Carlo Methods for Maximum Margin Supervised Topic Models. In *NIPS*, 2012.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., and Saul, L.K. *An Introduction to Variational Methods for Graphical Models*. Springer, 1998.

Lacoste-Julien, S., Sha, F., and Jordan, M.I. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2008.

McDonald, R., Crammer, K., and Pereira, F. Online large-margin training of dependency parsers. In *ACL*, 2005.

Michael, J.R., Schucany, W.R., and Haas, R.W. Generating random variates using transformations with multiple roots. *The American Statistician*, 30(2):88–90, 1976.

Mimno, D., Hoffman, M., and Blei, D.M. Sparse stochastic inference for latent dirichlet allocation. *ICML*, 2012.

Rifkin, R. and Klautau, A. In defense of one-vs-all classification. *JMLR*, 5:101–141, 2004.

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

Teh, Y.W., Newman, D., and Welling, M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *NIPS*, 2006b.

Wang, C. and Blei, D.M. Truncation-free online variational inference for Bayesian nonparametric models. In *NIPS*, 2012.

Wang, C., Paisley, J.W., and Blei, D.M. Online variational inference for the hierarchical Dirichlet process. In *AISTATS*, 2011.

Xu, M., Zhu, J., and Zhang, B. Bayesian nonparametric maximum margin matrix factorization for collaborative prediction. In *NIPS*, 2012.

Zhu, J. and Xing, E.P. Maximum entropy discrimination Markov networks. *JMLR*, 10:2531–2569, 2009.

Zhu, J., Chen, N., and Xing, E.P. Infinite SVM: a Dirichlet process mixture of large-margin kernel machines. In *ICML*, 2011.

Zhu, J., Ahmed, A., and Xing, E.P. MedLDA: maximum margin supervised topic models. *JMLR*, 13:2237–2278, 2012.

Zhu, J., Chen, N., Perkins, H., and Zhang, B. Gibbs max-margin topic models with fast sampling algorithms. *ICML*, 2013a.

Zhu, J., Zheng, X., Zhou, L., and Zhang, B. Scalable inference in max-margin topic models. In *SIGKDD*, 2013b.