Mach Learn (2017) 106:651–670
DOI 10.1007/s10994-016-5599-z

Collaborative topic regression for online recommender systems: an online and Bayesian approach

Chenghao Liu[1] · Tao Jin[1] · Steven C. H. Hoi[2] · Peilin Zhao[3] · Jianling Sun[1]

Received: 19 April 2016 / Accepted: 18 October 2016 / Published online: 11 January 2017

Abstract Collaborative Topic Regression (CTR) combines ideas of probabilistic matrix factorization (PMF) and topic modeling (such as LDA) for recommender systems, which has gained increasing success in many applications. Despite enjoying many advantages, the existing Batch Decoupled Inference algorithm for the CTR model has some critical limitations: First of all, it is designed to work in a batch learning manner, making it unsuitable to deal with streaming data or big data in real-world recommender systems. Secondly, in the existing algorithm, the item-specific topic proportions of LDA are fed to the downstream PMF but the rating information is not exploited in discovering the low-dimensional representation of documents and this can result in a sub-optimal representation for prediction. In this paper, we propose a novel inference algorithm, called the Online Bayesian Inference algorithm for CTR model, which is efficient and scalable for learning from data streams. Furthermore, we *jointly* optimize the combined objective function of both PMF and LDA in an online learning fashion, in which both PMF and LDA tasks can reinforce each other during the

Editors: Bob Durrant, Kee-Eung Kim, Geoff Holmes, Stephen Marsland, Zhi-Hua Zhou and Masashi Sugiyama.

✉ Steven C. H. Hoi
chhoi@smu.edu.sg

Chenghao Liu
twinsken@zju.edu.cn

Tao Jin
taoj@zju.edu.cn

Peilin Zhao
zhaop@i2r.a-star.edu.sg

Jianling Sun
sunjl@zju.edu.cn

[1] School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

[2] School of Information Systems, Singapore Management University, Singapore 178902, Singapore

[3] Institute for Infocomm Research, A*STAR, Singapore 138632, Singapore

online learning process. Our encouraging experimental results on real-world data validate the effectiveness of the proposed method.

Keywords Topic modeling · Online learning · Recommender systems · Collaborative filtering · Latent structure interpretation

1 Introduction

Due to the abundance of personalized online business, Recommender Systems (RS) now play an important role to help us to make effective use of information. For example, CiteULike[1] adopts RS for article recommendations, and Movielens[2] uses RS for movie recommendations. The core technique behind RS is a personalization algorithm (Almazro et al. 2010) for predicting the preference of each individual user by making use of different sources of information with respect to users and items. The most popular algorithms adopt the collaborative filtering (CF) technique (Su and Khoshgoftaar 2009; Breese et al. 1998), which analyzes the relationship between users and interdependencies among items, in order to identify new user-item associations. In general, CF is a method of making predictions ("filtering") about the interests of a user via collecting preferences from many users ("collaborating"). One of the most successful techniques for CF methods is based on Probabilistic Matrix Factorization (PMF) (Mnih and Salakhutdinov 2007) where a partially observed user-item rating matrix is approximated by the product of two low-rank matrices (latent factors) so as to complete the rating matrix towards recommendation purposes. Despite their popularity, most of traditional CF methods only use feedback matrix which contains the ratings on the items given by users. The prediction performance often drops significantly when the feedback matrix is sparse, which occurs when most items are given feedback by few users or most users give feedback to few items, since it is susceptible to overfitting. However, in the real-world scenarios, most users provide only a little feedback especially for new users, who have yet to provide rating information. On the other hand, in addition to the feedback matrix, auxiliary information is sometimes readily available, and could provide key information for recommendation task, meanwhile many existing CF methods ignore such side information, or are not intrinsically capable of exploiting it.

To overcome the sparsity issue of CF methods, Collaborative Topic Regression (CTR) has been actively explored in recent years (Wang and Blei 2011). Instead of purely relying on CF approaches, CTR aims to leverage content-based techniques to overcome inaccurate and unreliable predictions with traditional CF methods due to data sparsity and other challenges. More specifically, CTR combines the idea of PMF for predicting ratings, and the idea of probabilistic topic modeling, such as Latent Dirchelet Allocation (LDA), for analyzing the content of items towards recommendation tasks. It is a joint probabilistic graphical model that integrates LDA model and PMF model. CTR has been shown as a promising method that produces more accurate and interpretable results and has been successfully applied in many recommender systems, such as tag recommendation (Wang et al. 2013; Lu et al. 2015), and social recommender systems (Purushotham et al. 2012; Kang and Lerman 2013).

Despite being studied actively and extended to various kinds of applications (Wang and Blei 2011; Wang et al. 2013), no attempts have been made to develop efficient and scalable approximate inference algorithms of CTR model. The existing Batch Decoupled Inference

[1] http://www.citeulike.org.

[2] http://movielens.org/.

algorithm for CTR model (bdi-CTR) suffers from several critical limitations. First of all, it is often designed to work in a batch mode learning fashion, by assuming that all text contents of items as well as the rating training data are given prior to the learning tasks. During the training process, both the inference procedure of LDA and PMF models are usually trained separately in a batch training fashion. Such an approach would suffer from a huge scalability drawback when new data (users or items) may arrive sequentially and get updated frequently in a real-world online recommender system. Second, although the graphical model of CTR is a joint model (two-way interaction exists between LDA model and PMF model), bdi-CTR only leverages the content information to improve the CF tasks, but not the rating information. It first estimates LDA model, and then feed the document-specific topic proportions of LDA to the downstream PMF part. This two-step inference procedure is inconsistent with the joint graphical model of CTR and rather suboptimal as the the rating information is not used in discovering the low-dimensional representation of documents, which is clearly not an optimal representation for prediction as the two methods are not tightly coupled to fully exploit their potential. Our work is motivated to explore more efficient, scalable, and effective techniques to maximize the potential exploiting extremes in dealing with data streams from real-world online recommender systems.

To overcome the limitations of bdi-CTR, we propose a novel approximate inference scheme, called Online Bayesian Inference algorithm for CTR model (obi-CTR), which jointly optimizes a unified objective function by combining both PMF model and LDA model in an online learning fashion. In contrast to bdi-CTR, Our novel approximate inference scheme is able to achieve a much tighter coupling of both PMF and LDA, where both LDA and PMF tasks influence each other naturally and gradually via the joint optimization in the online learning process. This interplay yields topic representations of each item that are more suitable for making accurate and reliable rating prediction tasks.

To the best of our knowledge, our novel approximate inference algorithm is the first online learning algorithm for solving CTR tasks with fully joint optimization of both LDA model and PMF model. Our encouraging results from extensive experiments on large scale real-world data show that the proposed online learning algorithms are scalable and effective, and it not only outperforms the state-of-the-art methods for rating prediction tasks but also yields more suitable latent topic proportions in topic modeling tasks. Besides, our novel approximate inference algorithm can be applied to other variants of CTR model (see Sect. 2.2 for more on related work).

In the following, we first review some important related work, then present a formal formulation of CTR model and our novel approximate inference algorithm, Online Bayesian Inference algorithm for CTR model (obi-CTR). After that, we conduct extensive empirical studies and compare the proposed algorithms with the existing techniques, and finally set out our conclusions of this work.

2 Related work

In this section, we will provide a brief review of the prior studies related to our work, and some background of CTR model.

2.1 Online learning and online Bayesian inference

Online learning has been extensively studied in the machine learning communities (Cesa-Bianchi and Lugosi 2006; Shalev-Shwartz 2011; Zhao et al. 2011; Hoi et al. 2014, 2013), mainly due to its high efficiency and scalability to large-scale learning tasks. Different from conventional batch learning methods that assume all training instances are available prior to the learning phase, online learning considers one instance each time to update the predictive models sequentially and iteratively, which is more appropriate for large-scale applications where training data often arrive sequentially. In literature, a number of online learning algorithms have been proposed. A classical online learning method is the Perceptron algorithm (Rosenblatt 1958), which adopts an additive update rule for the classifier weights when a new instance is misclassified. Recently a lot of new online learning algorithms have been proposed based on the concept of "max margin" (Crammer et al. 2006; Crammer and Singer 2003; Gentile 2002). One notable technique in this category is the online Passive-Aggressive (PA) algorithm (Crammer et al. 2006). On each round, passive-aggressive algorithms solve a constrained optimization problem which balances between two competing goals: being conservative, in order to retain information acquired on preceding rounds, and being corrective, in order to make a more accurate prediction when a new instance is misclassified or its classification score does not exceed some predefined margin. In particular, PA method considers loss functions that enforce max-margin constraints and its simple update rule enjoys a closed form solution. Motivated by PA method, Hoi et al. (2013); Wang et al. (2014) apply parameter confidence information to improve online learning performance.

Although the classical research work of online learning is based on decision theory (Cesa-Bianchi and Lugosi 2006) and convex optimization (Shalev-Shwartz 2011), much progress has been made for developing online Bayesian Inference (Hoffman et al. 2010, 2013; Kingma and Welling 2013; Foulds et al. 2013). Rather than achieving a single point estimate of parameters typically in the optimization-based setting, Bayesian methods attempt to obtain the full posterior distribution over the unknown parameters and latent variables in the model, hence providing better characterizations of the uncertainties in the learning process and avoiding overfitting. There are two categories of studies on the topic of online Bayesian Inference. One direction is to extend Monte Carlo methods to the online setting. A classic approach is sequential Monte Carlo or particle filters (Robert and Casella 2013), which approximate virtually any sequence of probability distributions. Most recently, Welling and Teh (2011); Patterson and Teh (2013); Ahn et al. (2012) proposed stochastic gradient Langevin method by updating parameters according to both the stochastic gradients as well as additional noise, which asymptotically produce samples from the posterior distribution. Another direction is online variational Bayes (Hoffman et al. 2010, 2013; Kingma and Welling 2013; Foulds et al. 2013), where on each round only a mini-batch of instances is processed to give a noisy estimate of the gradient. Although these algorithms have shown impressive results most of them have adopted stochastic approximation of posterior distribution by sub-sampling a given finite data set, which is unsuitable for many applications where data size is unknown in advance.

To relax this assumption, researchers in Broderick et al. (2013); Honkela and Valpola (2003) made streaming updates to the estimated posterior. The intuition behind this idea is that we could treat the posterior after observing $T - 1$ samples as the new prior for the incoming data points. Specifically, suppose the training data $\{\mathbf{o}_t\}_{t\geq 0}$ are generated i.i.d. according to a distribution $p(\mathbf{o}|\mathbf{x})$ and the prior $p(\mathbf{x})$ is given. Bayes' theorem implies the posterior distribution of $\mathbf{x}$ given the first T samples ($T \geq 1$) satisfies

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

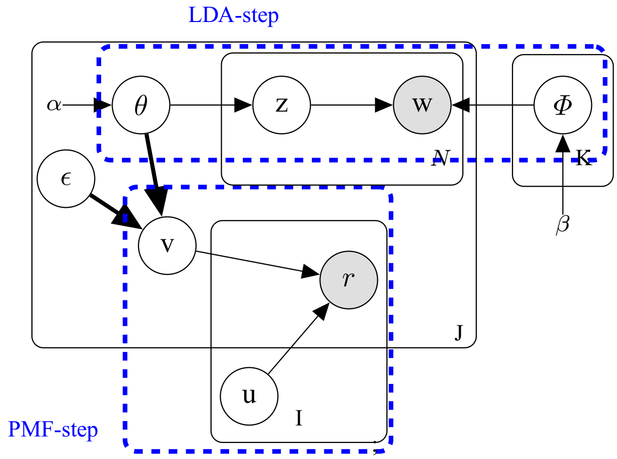

Fig. 1 The graphical model of collaborative topic regression model (Wang and Blei 2011). The approximate inference procedure consists of two steps: (i) first runs LDA-step, (ii) and then feeds topic proportions $\boldsymbol{\theta}_j$ to the PMF-step

For complex models, we can use an approximate algorithm $\mathcal{A}$ that calculates an approximate posterior q : $q(\mathbf{x}) = \mathcal{A}(\mathbf{X}, q_0(\mathbf{x}))$ where $\mathbf{X}$ is the observed data and $q_0(\mathbf{x})$ is the prior distribution. Then, we can recursively calculate an approximation to the posterior:

$$q(\mathbf{x}|\{\mathbf{o}\}_{t=0}^{T}) = \mathcal{A}\left(\mathbf{o}_T, q\left(\mathbf{x}|\{\mathbf{o}\}_{t=0}^{T-1}\right)\right)$$

In addition, McInerney et al. (2015) introduced the population Variational Bayes (PVB) method which combines traditional Bayesian inference with the frequentist idea of the population distribution for streaming inference. Shi and Zhu (2014) proposed the Online Bayesian Passive-Aggressive (BayesPA) method for max-margin Bayesian inference of online streaming data. The high scalability of the above methods motivates us to propose Online Bayesian inference for CTR model.

2.2 The graphical model of CTR and its variants

Collaborative topic regression (Wang and Blei 2011) is proposed to recommend items to users by seamlessly integrating both feedback matrix and the content of items into the same model. By combining PMF model and LDA model, CTR has gained increasing successes in many applications. Figure 1 shows the graphical model of CTR. Suppose there are I users and J items. Each data sample is a 3-tuple (i, j, r_{ij}) where $i \in \{1, 2, \ldots, I\}$ is the user index, $j \in \{1, 2, \ldots, J\}$ is the item index and $r_{ij} \in \mathbb{R}$ is the rating value assigned to item j by user i. We assume the rating data arrives sequentially in an online recommender system. Let $\mathbf{R}$ denote the whole rating samples and the collection of J items is regarded as a document set $\mathbf{W} = \{\mathbf{w}_j\}_{j=1}^{J}$. Let $\mathbf{Z} = \{\mathbf{z}_j\}_{j=1}^{J}$ and $\Theta = \{\boldsymbol{\theta}_j\}_{j=1}^{J}$ denote all the topic assignments and topic proportions of each item. We represent users and items in a shared latent low-dimensional space of dimension K, which is equal to the number of topics, user i is represented by a latent vector $\mathbf{u}_i \in \mathbb{R}^K$ and item j by a latent vector $\mathbf{v}_j \in \mathbb{R}^K$.

Basically, the CTR model assumes that each item is generated by a topic model and additionally includes a latent variable $\boldsymbol{\epsilon}_j$ which offsets the topic proportions $\boldsymbol{\theta}_j$ when modeling the user's latent vector. This offset variable $\boldsymbol{\epsilon}_j$ can capture the item preference of a particular user based on their ratings. Assume there are K topics $\Phi = \{\boldsymbol{\phi}_k\}_{k=1}^K$. The generative process of the CTR model is as follows:

1. For each user i, draw user latent vector $\mathbf{u}_i \sim \mathcal{N}(0, \frac{1}{\sigma_u^2}\mathbf{I}_K)$
2. For each item j,
   (a) Draw topic proportions $\boldsymbol{\theta}_j \sim Dirichlet(\alpha)$.
   (b) Draw item latent offset $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, \frac{1}{\sigma_\epsilon^2}\mathbf{I}_K)$ and set the item latent vector as $\mathbf{v}_j = \boldsymbol{\epsilon}_j + \boldsymbol{\theta}_j$.
   (c) For each word $w_{jn}(1 \leq n \leq N_j)$,
      (i) Draw topic assignment $z_{jn} \sim Mult(\boldsymbol{\theta}_j)$.
      (ii) Draw word $w_{jn} \sim Mult(\boldsymbol{\phi}_{z_{jn}})$.
3. For each user-item pair (i, j), draw the rating $r_{ij} \sim \mathcal{N}(\mathbf{u}_i^T\mathbf{v}_j, \frac{1}{\sigma_r^2})$.

In step 2 (c) ii. $\boldsymbol{\phi}_{z_{jn}}$ denotes the topic selected by the non-zero entry of z_{jn}. The topics are random samples drawn from a prior, e.g., $\boldsymbol{\phi}_k \sim Dirichlet(\beta)$. Note that $\mathbf{v}_j = \boldsymbol{\epsilon}_j + \boldsymbol{\theta}_j$, where $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, \frac{1}{\sigma_\epsilon^2}\mathbf{I}_K)$, is equivalent to $\mathbf{v}_j \sim \mathcal{N}(\boldsymbol{\theta}_j, \frac{1}{\sigma_\epsilon^2}\mathbf{I}_K)$. As mentioned in Wang and Blei (2011), this assumption plays a key role in CTR model, which means the item latent vector $\mathbf{v}_j$ is close to topic proportions $\boldsymbol{\theta}_j$, but can diverge from it if it has to.

Researchers have extended CTR models to different applications of recommender systems. Some researchers extended CTR models by integrating with other side information. In CTR-smf (Purushotham et al. 2012), authors integrated CTR with social matrix factorization models to take social correlation between users into account. In LA-CTR (Kang and Lerman 2013), they assumed that users divide their limited attention non-uniformly over other people. In HFT (McAuley and Leskovec 2013), they aligned hidden factors in product ratings with hidden topics in product reviews for product recommendations. In CSTR (Ding et al. 2013), authors explored how to recommend celebrities to general users in the context of social network. In CTR-SR (Wang et al. 2013), authors adapted CTR model by combining both item-tag matrix and item content information for tag recommendation tasks. There were also several works that attempted to extract latent topic proportions of text information in CTR via deep learning techniques (Wang et al. 2014, 2015; Van den Oord et al. 2013). However, all of these work followed the same approximate inference procedure as (Wang and Blei 2011) in a batch learning mode.

Independently of our study, Gopalan et al. 2014 developed a similar graphical model (CTPF) for articles recommendations task. Unlike CTR which applies PMF for recommendation model and LDA for document model, CTPF replaces both the usual Gaussian likelihood in PMF and multinomial likelihood in LDA with Poisson likelihood. This modification of graphical model makes it become a simple conditionally conjugate model and allows it to easily enjoy scalable approximate inference by using stochastic variational inference technique. However, the graphical model of original CTR is a direct combination of PMF and LDA, which is a much more complex non-conjugate model and makes its approximate inference non-trivial and challenging. In our work, we focus on the original graphical model of CTR and jointly optimize the combined objective function by using streaming variational inference. Moreover, CTR is widely used in different applications of recommender systems and has been extended to various kind of graphical model. Our scalable approximate inference method can also be generalized to these graphical models.

3 Online Bayesian collaborative topic regression

3.1 Inference algorithm for CTR: revisited

Before introducing our novel Online Bayesian Inference algorithm for CTR model (obi-CTR). we first review the batch decoupled approximate inference method Wang and Blei (2011) proposed (bdi-CTR), which has been applied to other variants of CTR models (see Sect. 2.2 for more on related work).

Given the document set $\mathbf{W}$ and rating data $\mathbf{R}$ (observered variables), we let $\mathbf{U} = \{\mathbf{u}_i\}_{i=1}^{I}$, $\mathbf{V} = \{\mathbf{v}_j\}_{j=1}^{J}$, the goal of CTR is to infer the posterior distribution of hidden variables $\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta$,

$$\begin{aligned} &p(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta|\mathbf{W}, \mathbf{R}) \\ &\quad \propto p_0(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta) p(\mathbf{W}|\mathbf{Z}, \Phi) p(\mathbf{R}|\mathbf{U}, \mathbf{V}), \end{aligned} \tag{1}$$

where prior distribution $p_0(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta) = \prod_{i=1}^{I} p_0(\mathbf{u}_i|\sigma_u) \prod_{j=1}^{J} p_0(\mathbf{v}_j|\sigma_j) \prod_{n=1}^{N_j} p_0 (\mathbf{z}_{jn}|\boldsymbol{\theta}_j) \prod_{k=1}^{K} p_0(\Phi_k|\beta) \prod_{j=1}^{J} p_0(\boldsymbol{\theta}_j|\alpha)$ according to the generative process of CTR as shown in Fig. 1. Because computing the full posterior of $\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta$ directly is intractable, Wang and Blei (2011) proposed a heuristic two-stage method for approximate inference . It simply modifies the original posterior distribution $p(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta|\mathbf{W}, \mathbf{R})$ by separating it into two parts, posterior distribution $p(\mathbf{Z}, \Phi, \Theta|\mathbf{W})$ with respect to LDA part and posterior distribution $p(\mathbf{U}, \mathbf{V}|\mathbf{R}, \Theta)$ with respect to PMF part. First, it approximately infers posterior $p(\mathbf{Z}, \Phi, \Theta|\mathbf{W})$ of LDA part via a traditional LDA learning method (Blei et al. 2003). Then, it learns the maximum a posteriori (MAP) estimates of $\mathbf{U}, \mathbf{V}, \Theta$ with respect to $p(\mathbf{U}, \mathbf{V}|\mathbf{R}, \Theta)$ by feeding the results of Θ in the first step into the PMF part. Maximization of $p(\mathbf{U}, \mathbf{V}|\mathbf{R}, \Theta)$ is equivalent to maximizing its log likelihood as follows

$$\begin{aligned} \mathcal{L} = &-\frac{\sigma_u^2}{2}\sum_i \mathbf{u}_i^\top \mathbf{u}_i - \frac{\sigma_\epsilon^2}{2}\sum_j (\mathbf{v}_j - \boldsymbol{\theta}_j)^\top (\mathbf{v}_j - \boldsymbol{\theta}_j) \\ &+ \sum_j \sum_n \log\left(\sum_k \theta_{jk}\phi_{k,w_{jn}}\right) - \sum_{i,j}\frac{\sigma_r^2}{2}(r_{ij} - \mathbf{u}_i^T \mathbf{v}_j)^2. \end{aligned} \tag{2}$$

We can optimize this function by gradient descent method, iteratively optimizing the collaborative filtering variables $\mathbf{u}_i, \mathbf{v}_j$ and the topic proportions variables $\boldsymbol{\theta}_j$. For $\mathbf{u}_i, \mathbf{v}_j$, they follow a similar fashion as basic matrix factorization.[3]

$$\begin{aligned} &\mathbf{u}_i \leftarrow \mathbf{u}_i - \rho(\sigma_u^2 \mathbf{u}_i - (r_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)\mathbf{v}_j) \\ &\mathbf{v}_j \leftarrow \mathbf{v}_j - \rho(\sigma_\epsilon^2(\mathbf{v}_j - \boldsymbol{\theta}_j) - (r_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)\mathbf{u}_i) \end{aligned} \tag{3}$$

where ρ is the learning rate. For $\boldsymbol{\theta}_j$, projection gradient is adopted, since they cannot optimize $\boldsymbol{\theta}_j$ analytically. In addition, Wang and Blei (2011) points out that simply fixing $\boldsymbol{\theta}_j$ as the estimate from previous LDA step could give comparable performance and save computation, which is consistent with our analysis—this inference algorithm is rather suboptimal and tends to get trapped into local optimum. Finally, we summarize the bdi-CTR algorithm in Algorithm 1.

Motivated by the online LDA methods (Hoffman et al. 2010; Mimno et al. 2012), we extend bdi-CTR algorithm to an online learning version (odi-CTR) by incorporating the

[3] Wang and Blei (2011) adopts the ALS algorithm (Hu et al. 2008) to solve an implicit feedback problem. In our context, we use the SGD algorithm (Koren et al. 2009) since ratings data are explicit.

Algorithm 1 The Batch Decoupled Inference algorithm for CTR model (**bdi-CTR**)

```
Initialize U, V, Z randomly
Input: rating data R and document set W
LDA-step: Compute $\Theta$, $\Phi$, Z by traditional LDA inference method (Blei et al. 2003)
PMF-step: Given $\Theta$ (the result of LDA-step) and R, compute U, V by traditional gradient descent method
Output: U, V and Z
```

online LDA method (Hoffman et al. 2010). Online LDA is an EM style method. In the E-step, it approximately finds locally optimal values of $\boldsymbol{\theta}_j$ via an iterative method, holding Φ fixed. And then, in the M-step, online LDA updates Φ using a weighted average of its previous value and noisy estimate corresponding to $\boldsymbol{\theta}_j$. If we control the learning rate such that old values are forgotten gradually, the objective with respect to posterior distribution $p(\mathbf{Z}, \Phi, \Theta|\mathbf{W})$ converges to a stationary point [more details can be found in Hoffman et al. (2010)].

The odi-CTR algorithm follows the same strategy as bdi-CTR algorithm, separating the posterior distribution $p(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta|\mathbf{W}, \mathbf{R})$ into LDA part and PMF part. At each round, the estimations of LDA part and PMF part are conducted simultaneously. The algorithm is described in Algorithm 2. It is obvious that a significant disadvantage of Algorithm bdi-CTR and odi-CTR is that both of them follow a two-step inference procedure which is inconsistent with the joint graphical model of CTR and rather suboptimal as the rating information is not used in discovering the low-dimensional representation of documents.

Algorithm 2 The Online Decoupled Inference algorithm for CTR model (**odi-CTR**)

```
Initialize U, V, Z randomly
for t = 1 to $\infty$ do
  Receive data sample $(i, j, r_{ij}, \mathbf{w}_j)$
  Update $\boldsymbol{\theta}_j$ as the E-step in the online LDA method (Algorithm 2 in Hoffman et al. (2010))
  Update $\Phi$ as the M-step in the online LDA method
  Update $\mathbf{u}_i$, $\mathbf{v}_j$ using online gradient descent by Eq. (3)
end for
Output: U, V and Z
```

The main challenge is the joint optimization of CTR model in an online learning fashion. To start off, we first present inefficient (baseline) approach, bdi-CTR and odi-CTR, and later shows our novel Online Bayesian Inference algorithm for CTR model (obi-CTR).

3.2 The online Bayesian inference algorithm for CTR model (obi-CTR)

Instead of learning two point estimates of coefficients $\mathbf{u}_i$, $\mathbf{v}_j$, we take a more general Bayesian-style approach and learn the posterior distribution $q(\mathbf{u}_i, \mathbf{v}_j)$ in an online method. For rating prediction, we take a weighted average over all the possible latent vectors $\mathbf{u}_i$ and $\mathbf{v}_j$, or more precisely, an expectation of the prediction over $q(\mathbf{u}_i, \mathbf{v}_j)$ which is defined as

$$\hat{r}_{ij} \triangleq \mathbb{E}[\mathbf{u}_i^\top \mathbf{v}_j].$$

In addition, we set $\mathbf{v}_j = \boldsymbol{\epsilon}_j + \bar{\mathbf{z}}_j$, which means the item latent vector $\mathbf{v}_j$ is directly close to $\bar{\mathbf{z}}_j$, where $\bar{\mathbf{z}}_j$ is a vector with element $\bar{\mathbf{z}}_j = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}(z_n^k = 1)$ and $\mathbb{I}$ is the indicator function that equals to 1 if predicate holds otherwise 0. This setting is widely used in supervised topic model (Mcauliffe and Blei 2008; Zhu et al. 2012; Agarwal and Chen 2010), and could simplify our following inference procedure.

Finally, Algorithm 3 summarizes the detailed framework of the proposed obi-CTR algorithm. At each round t, we receive data sample and update both the parameters of LDA part and PMF part. The following discusses the optimization and each step of the algorithm in detail.

Algorithm 3 Online Bayesian Inference algorithm for CTR model (**obi-CTR**)

```
Initialize $\mathbf{U}, \mathbf{V}, \mathbf{Z}$ randomly.
for t = 1 to $\infty$ do
    Receive data sample $(i, j, r_{ij}, \mathbf{w}_j)$
    Draw samples $\mathbf{z}_j^t$ from Eq. (12)
    Discard $B$ burn-in sweeps, use the rest samples to update $\mathbf{u}_i, \mathbf{v}_j, \Phi$ following Eq. (6),(8),(10)
end for
Output: $\mathbf{U}, \mathbf{V}$ and $\mathbf{Z}$
```

Now, we propose our novel Online Bayesian Inference algorithm for CTR model (obi-CTR) which is efficient and scalable for learning from data streams. Instead of separate CTR into LDA step and PMF step, we consider to jointly optimize the unified objective function. Let us first review the objective function of CTR defined in (1), from a variational point of view, this posterior is identical to the solution of the following optimization problem:

$$\begin{aligned}\min_{q(\mathbf{U},\mathbf{V},\mathbf{Z},\Phi,\Theta)} \; & KL[q(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta)\|p_0(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta))] \\ & - \mathbb{E}_q[\log p(\mathbf{W}|\mathbf{Z}, \Phi)p(\mathbf{R}|\mathbf{U}, \mathbf{V})] \\ & s.t. \quad q(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta) \in \mathcal{P},\end{aligned} \tag{4}$$

where $KL(q\|p)$ is the Kullback-Leibler divergence, and $\mathcal{P}$ is the space of probability distributions. Specifically, we find a posterior distribution $q(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta)$ that is not only close to the prior distribution $p_0(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta)$ in terms of KL-divergence, which implicitly express the relationship between $\mathbf{V}$ and $\mathbf{Z}$ (this is the key to CTR model which makes the item vector $\mathbf{v}_j$ close enough to the topic proportions $\bar{\mathbf{z}}_j$ and diverge from it if necessary) but also has a high likelihood of explaining the observed data $\mathbf{R}, \mathbf{W}$. If we add the constant $\log p(\mathbf{W})p(\mathbf{R})$ to the above objective function, it is the minimization of $KL(q(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta)\|p(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \Phi, \Theta|\mathbf{W}, \mathbf{R}))$. We can use mean-field variational approximate inference which is a popular method for approximate posteriors (Blei et al. 2003; Hoffman et al. 2010). Inspired by streaming Bayesian inference (Broderick et al. 2013; Honkela and Valpola 2003), on the arrival of new data $(i, j, r_{ij}, \mathbf{w}_j)$, if we treat the posterior after observing $t-1$ samples as the new prior, the post-data posterior distribution $q_{t+1}(\mathbf{u}_i, \mathbf{v}_j, \mathbf{z}_j, \Phi, \Theta)$ is equivalent to the solution of the following optimization problem:

$$\begin{aligned}\min_q \; & KL[q(\mathbf{u}_i, \mathbf{v}_j, \mathbf{z}_j, \Phi, \Theta)\|q_t(\mathbf{u}_i, \mathbf{v}_j, \mathbf{z}_j, \Phi, \Theta))] \\ & - \mathbb{E}_q[\log p(\mathbf{w}_j|\mathbf{z}_j, \Phi)p(r_{ij}|\mathbf{u}_i^\top \mathbf{v}_j)] \\ & s.t. \quad q(\mathbf{u}_i, \mathbf{v}_j, \mathbf{z}_j, \Phi, \Theta) \in \mathcal{P}.\end{aligned} \tag{5}$$

This problem is intractable to compute. Here, we use mean field methods (Jordan et al. 1999) widely employed in fitting topic model to efficiently obtain an approximation for the above problem. Specifically, we assume that $q(\mathbf{u}_i, \mathbf{v}_j, \mathbf{z}_j) = q(\mathbf{u}_i)q(\mathbf{v}_j)q(\mathbf{z}_j)$. Therefore, we can solve this problem via an iterative procedure that alternatively updates each factor distribution as follows in detail.

For $\mathbf{u}_i$: By fixing the distribution $q(\mathbf{v}_j)$, we can ignore irrelevant terms and solve

$$\min_{q(\mathbf{u}_i)} KL[q(\mathbf{u}_i)q(\mathbf{v}_j)\|q_t(\mathbf{u}_i)p(r_{ij}|\mathbf{u}_i^\top \mathbf{v}_j)].$$

The optimal solution has the following closed form solution:

$$q_{t+1}(\mathbf{u}_i) \propto q_t(\mathbf{u}_i)\exp(\mathbb{E}_{q(\mathbf{v}_j)}[\log p(r_{ij}|\mathbf{u}_i^\top \mathbf{v}_j)]).$$

If initial prior is normal $q_0(\mathbf{u}_i) = \mathcal{N}(\mathbf{u}_i; \mathbf{m}_{ui}^0, \Sigma_{ui}^0)$, by induction we can show that the inferred distribution at each round is also a normal distribution. Let us assume $q_t(\mathbf{u}_i) = \mathcal{N}(\mathbf{u}_i; \mathbf{m}_{ui}^t, \Sigma_{ui}^t)$. Then, we have

$$\begin{aligned} q_{t+1}(\mathbf{u}_i) \propto\ & \exp\left(-\frac{1}{2}(\mathbf{u}_i - \mathbf{m}_{ui}^t)^\top \left(\Sigma_{ui}^t\right)^{-1}(\mathbf{u}_i - \mathbf{m}_{ui}^t)\right. \\ & \left. + \mathbb{E}_{q(\mathbf{v}_j)}\left[-\frac{(r_{i,j} - \mathbf{u}_i^\top \mathbf{v}_j)^2}{2\sigma_r^2}\right]\right) \\ & = \mathcal{N}(\mathbf{u}_i; \mathbf{m}_{ui}^*, \Sigma_{ui}^*), \end{aligned}$$

where the posterior parameters are computed as

$$\begin{aligned} \Sigma_{ui}^* &= \left(\left(\Sigma_{ui}^t\right)^{-1} + \frac{\mathbf{m}_{vj}\mathbf{m}_{vj}^\top}{\sigma_r^2 \mathbf{I}_K}\right)^{-1}, \\ \mathbf{m}_{ui}^* &= \mathbf{m}_{ui}^t + \frac{r_{i,j} - \mathbf{m}_{vj}^\top \mathbf{m}_{ui}^t}{\sigma_r^2 + \mathbf{m}_{vj}^\top \Sigma_{ui}^t \mathbf{m}_{vj}} \Sigma_{ui}^t \mathbf{m}_{vj}. \end{aligned} \tag{6}$$

Computing the full matrix Σ_{ui}^* could be computationally expensive, particularly when k is large. To reduce computational cost, we only update the diagonals of covariance matrix Σ_{ui}^*, which is equivalent to the assumption of Gaussian distribution $q(\mathbf{u})$ with diagonal covariance matrix.

For $\mathbf{v}_j$: The update rule of $\mathbf{v}_j$ is similar to $\mathbf{u}_i$ except adding a Gaussian distribution $p(\boldsymbol{\epsilon}_j|\bar{\mathbf{z}}_j, \mathbf{v}_j)$, a constraint about the distance between $\mathbf{v}_j$ and $\bar{\mathbf{z}}_j$, that explains the difference between topic assignments in content and item preference based on ratings. By fixing the distribution of $q(\mathbf{u}_i)$ and $q(\mathbf{z}_j)$, we have the update rule

$$\begin{aligned} q_{t+1}(\mathbf{v}_j) &\propto q_t(\mathbf{v}_j)\exp\left(\mathbb{E}_{q(\mathbf{u}_i,\mathbf{z}_j)}[\log p(r_{ij}|\mathbf{u}_i^\top \mathbf{v}_j)p(\boldsymbol{\epsilon}_j|\bar{\mathbf{z}}_j, \mathbf{v}_j)]\right) \\ &\propto \exp(-\frac{1}{2}(\mathbf{v}_j - \mathbf{m}_{vj}^t)^\top(\Sigma_{vj}^t)^{-1}(\mathbf{v}_j - \mathbf{m}_{vj}^t) \\ &\quad + \mathbb{E}_{q(\mathbf{u}_i)q(\mathbf{z}_j)}\left[-\frac{(r_{i,j} - \mathbf{u}_i^\top \mathbf{v}_j)^2}{2\sigma_r^2} - \frac{(\bar{\mathbf{z}}_j - \mathbf{v}_j)^\top(\bar{\mathbf{z}}_j - \mathbf{v}_j)}{2\sigma_\epsilon^2 \mathbf{I}_K}\right]\Bigg) \\ &= \mathcal{N}(\mathbf{v}_j; \mathbf{m}_{vj}^*, \Sigma_{vj}^*), \end{aligned} \tag{7}$$

where the posterior parameters are computed as

$$\begin{aligned}\Sigma_{mix} &= (\Sigma_{vj}^{-1} + \frac{1}{\sigma_\epsilon^2})^{-1},\\ \Sigma_{vj}^{*} &= \left((\Sigma_{vj}^{t})^{-1} + \frac{1}{\sigma_\epsilon^2 \mathbf{I}_K} + \frac{\mathbf{m}_{ui}\mathbf{m}_{ui}^\top}{\sigma_r^2 \mathbf{I}_K}\right)^{-1},\\ \mathbf{m}_{vj}^{*} &= \Sigma_{mix}\Sigma_{vj}^{-1}\mathbf{m}_{vj}^{t} + \Sigma_{mix}\frac{1}{\sigma_\epsilon^2}\bar{\mathbf{z}}_j - \Sigma_{mix}\frac{1}{\sigma_r^2}\mathbf{m}_{ui}\\ &\quad \left(\frac{\mathbf{m}_{ui}^\top \Sigma_{mix}\Sigma_{vj}^{-1}\mathbf{m}_{vj}^{t} + \mathbf{m}_{ui}^\top \Sigma_{mix}\frac{1}{\sigma_\epsilon^2}\bar{\mathbf{z}}_j - r_{ij}}{1 + \mathbf{m}_{ui}^\top \Sigma_{mix}\frac{1}{\sigma_r^2}\mathbf{m}_{ui}}\right). \end{aligned} \tag{8}$$

Besides, we adopt the same strategy that only updating the diagonals of covariance matrix Σ_{vj}^{*}.

For Φ By fixing the distribution $q(\mathbf{Z})$ and $q(\mathbf{W})$, $q(\Phi)$ can be solved as,

$$q_{t+1}(\Phi_k) \propto q_t(\Phi_k)\exp\left(\mathbb{E}_{q(\mathbf{Z}_t)}\left[\log p_0(\mathbf{Z}_t)p(\mathbf{X}|\mathbf{Z}_t,\Phi)\right]\right), \quad k = 1, 2, \ldots, K. \tag{9}$$

If the prior distribution $q_0(\Phi_k)$ satisfy a Dirichlet distribution $\Phi_k = Dir(\Delta_{k1}^0, \ldots, \Delta_{kW}^0)$, then by induction the inferred distributions are also in the family of Dirichlet distributions. We denote that $q_t(\Phi_k) = Dir(\Delta_{k1}^t, \ldots, \Delta_{kW}^t)$, then we can derive

$$q^\star(\Phi_k) = Dir(\Delta_{k1}^\star, \ldots, \Delta_{kW}^\star), \tag{10}$$

where $\Delta_{kw}^\star = \Delta_{kw}^t + \sum_{n=1}^{N_j}\gamma_{jn}^k \mathbb{I}[w_{jn} = w_{voc}]$ for all words w_{voc} $(1 \leq w_{voc} \leq D)$ in the vocabulary (D is the vocabulary size) and $\gamma_{jn}^k = \mathbb{E}_{q(z_j)}\mathbb{I}[z_{jn} = k]$ is the probability of assigning each word w_{jn} to topic k.

For $\mathbf{z}_j$: Given the distribution of other variables, the conditional distribution of $\mathbf{z}_j$ is:

$$\begin{aligned}&q(\mathbf{z}_j|\mathbf{v}_j, \Phi, \mathbf{w}_j)\\ &\quad\propto p_0(\mathbf{z}_j)\exp\left(\mathbb{E}_{q(\Phi)q(\mathbf{v}_j)}\left[\log p(\mathbf{w}_j|\mathbf{z}_j,\Phi)p(\boldsymbol{\epsilon}_j|\bar{\mathbf{z}}_j,\mathbf{v}_j)\right]\right)\\ &\quad\propto p_0(\mathbf{z}_j)\exp\left(\sum_{n\in[N_j]}\varLambda_{z_{jn},w_{jn}} - \mathbb{E}_{q(\mathbf{v}_j)}\left[\frac{(\mathbf{v}_j - \bar{\mathbf{z}}_j)^\top(\mathbf{v}_j - \bar{\mathbf{z}}_j)}{2\sigma_\epsilon^2 \mathbf{I}_K}\right]\right)\end{aligned} \tag{11}$$

where $\varLambda_{z_{jn},w_{jn}} = \mathbb{E}_{q(\Phi)}\left[\log(\Phi_{z_{jn},w_{jn}})\right] = \varPsi(\Delta^\star_{z_{jn},w_{jn}}) - \varPsi(\sum_{w_{voc}}\Delta^\star_{z_{jn},w_{voc}})$ (note that $\varPsi(\cdot)$ is the digamma function). It is difficult to directly update Φ and $\mathbf{v}_j$ by using $q(\mathbf{z}_j)$ due to the huge number of configurations. Therefore, we can do Gibbs sampling to infer $q(\mathbf{z}_j)$ by canceling out common factors and estimate the required expectations with multiple empirical samples. This hybrid strategy has shown promising performance for LDA (Mimno et al. 2012; Shi and Zhu 2014). Specifically, the conditional distribution of one variable z_{jn} (the topic assignment of the n-th word in item j) given others $\mathbf{z}_{j\neg n}$ is

$$\begin{aligned}&q(z_{jn} = k|\mathbf{z}_{j\neg n}, \mathbf{v}_j, \Phi, w_{jn} = w_{voc})\\ &\quad\propto \underbrace{(\alpha + C_{j\neg n}^k)\exp(\varLambda_{k,w_{jn}}}_{(i)} + \underbrace{\frac{1}{2\sigma_\epsilon^2 N_j}(2m_{vjk} - \frac{1 + 2C_{j\neg n}^k}{N_j}))}_{(ii)},\end{aligned} \tag{12}$$

where $\mathbf{z}_{j\neg n}$ is the topic assignments in item j (except the n-th word) and $C_{j\neg n}^k$ is the number of words in item j (except the n-th word) that are assigned to topic k. We can see that term

(i) is from the LDA model for observed word counts and the term (ii) is from the PMF model and the relationship between $\mathbf{v}_j$ and $\bar{\mathbf{z}}_j$.

4 Experimental results

4.1 Dataset

Our experiments were conducted on an extended MovieLens dataset, named as "MovieLens-10M-Plot" and "MovieLens-1M-Plot",[4] which was originated from the MovieLens.[5] Specifically, the original MovieLens 10M dataset provides a total of 10,000,053 rating records for 10,681 movies (items) by 69,878 users. However, the original dataset has very limited *text* content information. We enrich the dataset by collecting additional text contents for each of the movie items. Specifically, for each movie item, we first used its identifier number to find the movie listed in the IMDb[6] website, and then collected its related text of "plot summary". We then combine the "plot summary" text together with each movie's title and category text given in the MovieLens-10M dataset as a text document to represent each movie. For detailed text preprocessing, we follow the same procedure as the one described in Wang and Blei (2011) to process text information. Finally, we form a vocabulary with 7,689 distinct words. We then randomly select 1 million rating records to form a small dataset named "MovieLens-1M-Plot". Note that we did not consider the CiteUlike dataset[7] which was used in the previous study (Wang and Blei 2011), because their dataset only provides "like" and "dislike" preference, which is a kind of implicit feedback and thus unsuitable for our regression task. By contrast, the MovieLens-10M dataset has explicit feedback with ratings ranging from 1 to 5.

4.2 Experimental setup and metric

For each experiment, we randomly shuffle the rating records, and then divide them into two parts: the first 90 % of the shuffled rating records are used as the training data, and the rest 10 % rating data are used as test set. We also randomly draw 5 % out of the training data as the validation set for parameter selection. To make fair comparisons, all the algorithms are conducted over 5 experimental runs of different random permutations. For performance metric, we evaluate the performance of our proposed method for prediction task by measuring Root Mean Square Error (RMSE) defined as:

$$RMSE = \sqrt{\sum(\hat{r} - r_{i,j})^2/N}$$

In the online learning experiments, we evaluate the RMSE performance on the test set after every 50,000 online iterations. In addition, we also evaluate the performance of topic modeling via the log-likelihood of each word in text collection (Hoffman et al. 2010), defined as,

$$perplexity(\mathbf{w}^{test}|\Phi, \Theta) = exp\left\{-\frac{\sum_d \log p(\mathbf{w}_d^{test}|\Phi, \Theta)}{\sum_{d,w} n_{dw}^{test}}\right\},$$

[4] http://ctr.stevenhoi.org/.

[5] http://grouplens.org/datasets/movielens/.

[6] http://www.imdb.com.

[7] http://www.citeulike.org/faq/data.adp.

where n_{dw}^{test} is the word count for word w in the d-th document.

4.3 Baselines for comparison and experimental settings

In our experiments, we evaluate the proposed obi-CTR algorithms for rating predictions by comparing with some important baselines as follows:

- **PA-I**: An online learning algorithm for solving online collaborative filtering tasks by applying the popular online Passive-Aggressive (PA) algorithm (Blondel et al. 2014);
- **bdi-CTR**: the existing Collaborative Topic Regression (Wang and Blei 2011) . In our context, we replace the ALS algorithm (Hu et al. 2008) with SGD algorithm (Koren et al. 2009) since ratings data are explicit, and keep the rest same as the original CTR (note that the LDA step is still performed in a batch manner);
- **odi-CTR**: The proposed Online Decoupled Inference algorithm for CTR model in Algorithm 2;
- **obi-CTR**: The proposed Online Bayesian Inference algorithm for CTR model in Algorithm 3.

Besides, to evaluate the topic modeling performance, we also compare our method with the typical Online LDA method:

- **Online-LDA**: an online Bayesian variational inference algorithm for LDA model (Hoffman et al. 2010). We take it as a baseline to evaluate how well the model fits the data with the predictive distribution.

For parameter settings, we find the optimal parameters for different algorithms (PA-I, bdi-CTR, odi-CTR and obi-CTR). Specifically, the parameters including c in PA-I, σ_u, σ_v and ρ in bdi-CTR and odi-CTR, and σ_ϵ and σ_r in obi-CTR. All of these parameters are found by performing a grid search as follows: $\sigma_\epsilon, \sigma_r \in \{0.5, 1, 2, 4, 8, 16, 32\}, c \in \{0.01, 0.1, 0.2, 0.5, 1\}$, $\rho \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$, $\sigma_u, \sigma_v \in \{0.01, 0.02, 0.04, 0.08, 0.16, 0.32\}$ and $K \in \{5, 10, 20\}$.

4.4 Evaluation of online rating prediction tasks

Figure 2a–c compares the online performance of the above methods in $K = 5$, $K = 10$ and $K = 20$ on the MovieLens-10M-Plot dataset and MovieLens-1M-Plot dataset. Note that the bdi-CTR method needs to precompute the parameters Θ and Φ by a batch variational inference algorithm.[8] Figure 2 shows only its performance in the downstream collaborative filtering phase.

As we can see from Fig. 2a–c, the CTR-based approaches outperform the online CF algorithm (PA-I) for most cases, which is in line with the experiments in Wang and Blei (2011) and validates the efficacy of leveraging additional text information to improve the performance of PMF for online rating prediction tasks. Second, among different CTR-based approaches, the proposed obi-CTR consistently outperforms the other algorithms for most cases. This validates the importance of jointly optimizing both online PMF and online LDA to achieve tight coupling of the two techniques. Moreover, it is interesting to find that the gap between the proposed odi-CTR variant and obi-CTR tends to become more significant when K is smaller. We conjecture that this is because when K is small, the PMF performance is relatively inaccurate and thus including the joint optimization becomes more critical for enhancing the unreliable PMF prediction performance. Finally, Table 1 summarizes the final

[8] For the vanilla LDA inference method, a larger K value often needs more time for computation.

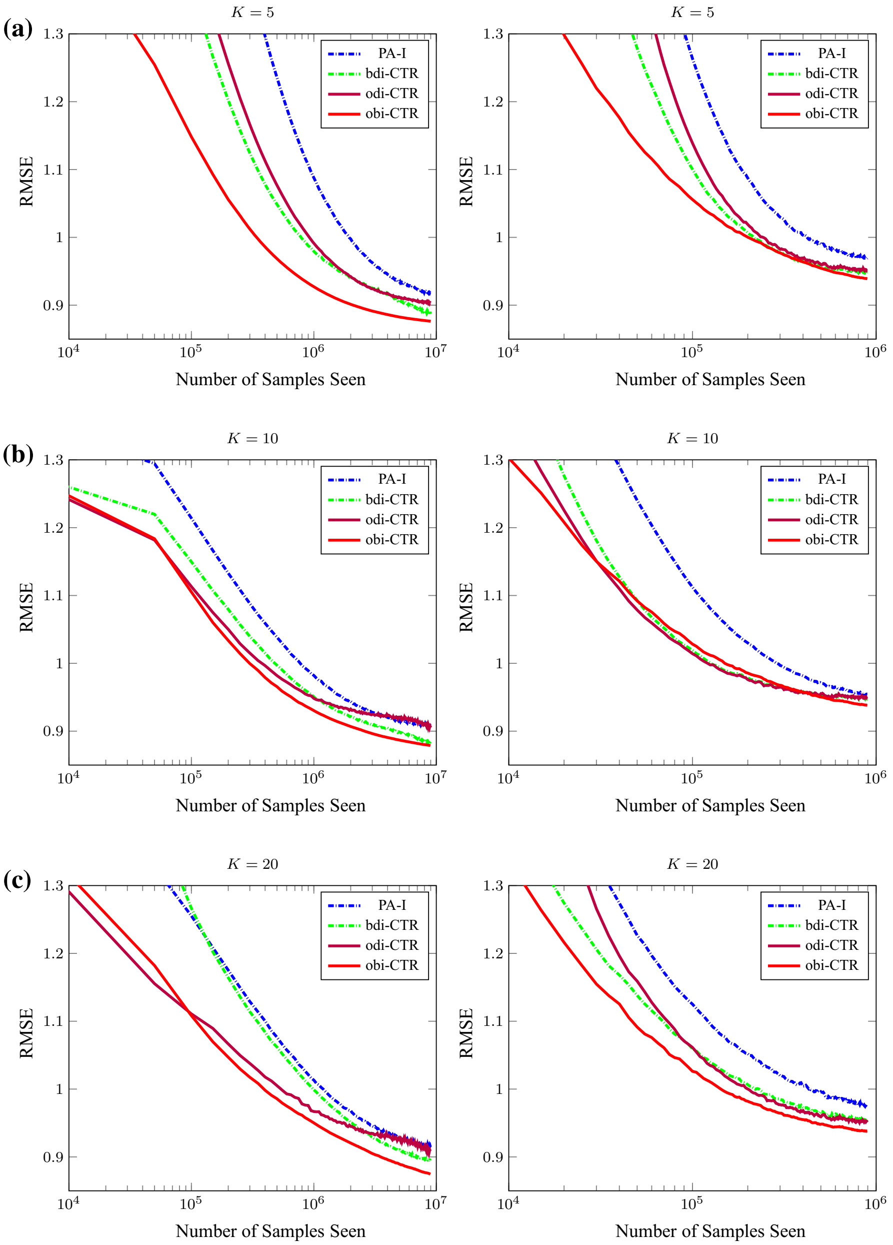

Fig. 2 Figure (**a**–**c**) shows the evaluation of RMSE performance by different online algorithms (*left column* for the MovieLens-10M-Plot dataset, *right colum* for the MovieLens-1M-Plot dataset)

test-set RMSE results after finishing the whole online learning tasks (by a single pass over the training set). Similar observations can be found , in which obi-CTR achieves the lowest RMSE result on the test set for rating prediction among all the algorithms. In addition, bdi-CTR has better performance than odi-CTR. This is because bdi-CTR directly takes the batch

Table 1 RMSE results after a single pass over MovieLens-10M-Plot and MovieLens-1M-Plot dataset

MovieLens-10M-Plot	k = 5	k = 10	k = 20
PA-I	0.9176 ± 0.0004	0.9085 ± 0.0002	0.9148 ± 0.0003
bdi-CTR	0.8874 ± 0.0003	0.8812 ± 0.0005	0.8947 ± 0.0007
odi-CTR	0.9034 ± 0.0006	0.9054 ± 0.0008	0.9085 ± 0.0002
obi-CTR	**0.8763** ± 0.0006	**0.8788** ± 0.0001	**0.8747** ± 0.0006
MovieLens-1M-Plot	k = 5	k = 10	k = 20
PA-I	0.9692 ± 0.0007	0.9547 ± 0.0008	0.9775 ± 0.0000
bdi-CTR	0.9488 ± 0.0004	0.9488 ± 0.0003	0.9548 ± 0.0007
odi-CTR	0.9805 ± 0.0004	0.9809 ± 0.0003	0.9826 ± 0.0003
obi-CTR	**0.9390** ± 0.0001	**0.9393** ± 0.0006	**0.9392** ± 0.0006

Bold values indicate the best result compared with other baselines

LDA results (pre-computed Θ and Φ) as input for leveraging online PMF task, while odi-CTR may converge relatively slowly (without the tight coupling). This again shows that it is crucial for the joint optimization in obi-CTR.

4.5 Performance on online topic modeling tasks

Figure 3 shows the results about online average predictive log likelihood for obi-CTR and Online LDA. Online learning allows us to conduct a large-scale comparison. We can see that obi-CTR exhibits consistently better performance than Online LDA, which ignores ratings information, regardless of how many topics we use. That is due to the utilization of rating information to discover the low-dimensional topic proportions, where obi-CTR yields additional benefit on this task.

4.6 Case study

To gain a deeper insight into the difference between bdi-CTR and obi-CTR, we choose one example user to conduct a case study. One advantage of the obi-CTR model is that it can explain the user latent space better than bdi-CTR model. In Table 2, we list the top 2 topic of this user and randomly select 10 movies he has rated before. obi-CTR gives a more accurate prediction than bdi-CTR. When digging into the data, we find that the top topic of obi-CTR contains words like “children”, “comedy”, but the top topic of bdi-CTR contains word like “adventure”, “story”. Thus, obi-CTR gives a higher rating for movie “Finding Nemo” which is more close to the true rating.

4.7 Evaluation of parameter sensitivity

Figure 4 shows how RMSE is affected by the choice of two key parameters σ_ϵ and σ_r in obi-CTR. As observed from Fig. 4, at the beginning, increasing σ_ϵ leads to decrease the RMSE quickly. After arriving some optimal value, increasing σ_ϵ further may increase the RMSE gradually. Second, we found the optimal value of σ_ϵ also largely depends on the setting of the parameter σ_r. When σ_r is smaller, the optimal value of σ_ϵ is relatively smaller. However, after reaching the optimal value, the further performance changing becomes limited. This

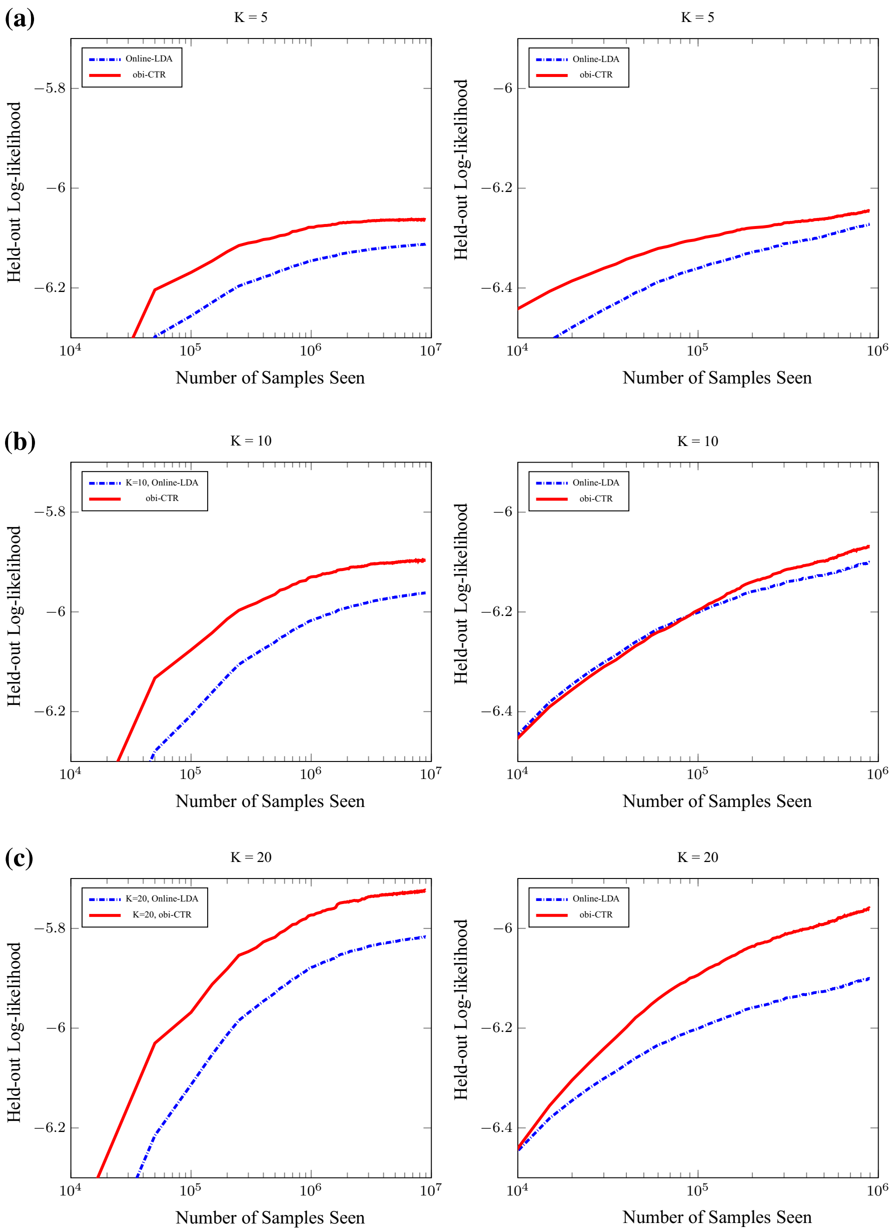

Fig. 3 Figure (**a**–**c**) demonstrate the online per-word predictive log likelihood comparisons between obi-CTR and Online LDA (*left column* for the MovieLens-10M-Plot dataset, *right colum* for the MovieLens-1M-Plot dataset)

indicates that overall, it is relatively easy to choose a good value of σ_ϵ given a fixed σ_r setting due to its less sensitivity in the range of optimal values. Our results are consistent with similar phenomena observed in Wang and Blei (2011).

Table 2 Interpretability of the latent structures learned

Top topic by obi-CTR

1. comedy, children, romance, animal, music, fantasy, drama, friend, family
2. work, find, die, life, only, time, kill, event, end, plan, final

Top topic by bdi-CTR

1. adventure, story, young, ring, king, prince, come, toy, music, world, begin, place
2. thriller, help, kill, mission, murder, lawyer, harry, evil, want, live, discover

In user's ratings	r	$\hat{r}_{obi-CTR}$	$\hat{r}_{bdi-CTR}$
Sound of music	4.5	**4.4**	4.7
1984 (Nineteen eighty-four)	4	**3.9**	4.4
Fantasia	5	4.2	4.2
Finding nemo	5	**4.5**	4.2
Schindler's list	5	4.8	**5**
Memento	5	4.7	**5**
Star wars: Episode IV	4.5	**4.6**	4.8
Matrix reloaded, The	3	**3.5**	3.9
Life is beautiful	5	**4.9**	4.6
City of God	4.5	**4.7**	4.8

Bold values indicate the best result compared with other baselines

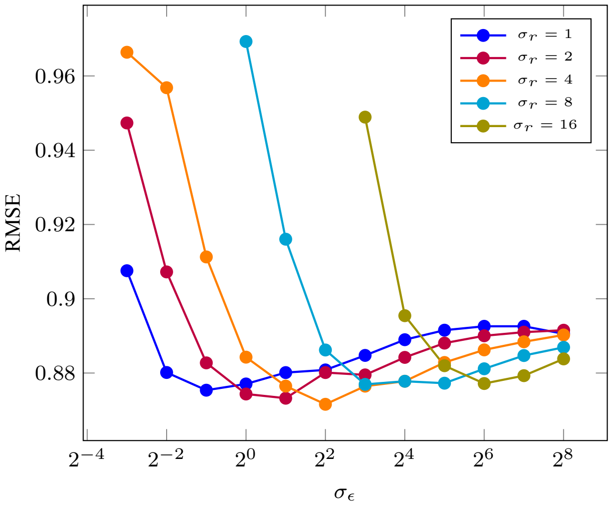

Fig. 4 This figure shows the evaluation of parameter influences (σ_r and σ_ϵ)

Table 3 shows the computation cost for training Online-LDA, bdi-CTR, odi-CTR and obi-CTR. Figure 5 demonstrates the effect of increasing model complexity K. This investigation is done by selecting the best achievable RMSE and log-likelihood during the grid parameter search process. As shown in the diagram, increasing the complexity of models (higher K values) leads to improvement of both RMSE and log-likelihood results. However, the gain of predictive performance is paid by a significant computational overhead for more complex

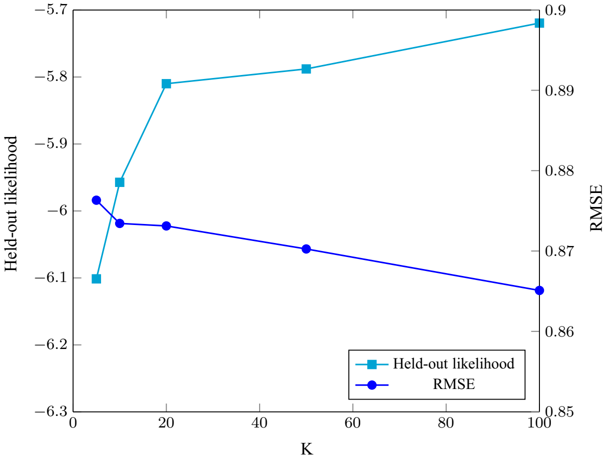

Fig. 5 This figure demonstrates the evaluation of obi-CTR result by varying K

Table 3 Running time measured in seconds consumed for each model size (K)

	k = 5	k = 10	k = 20	k = 50	k = 100
Online-LDA	725	892	1790	4004	7901
bdi-CTR	5338	6407	14456	31125	63048
odi-CTR	1177	1571	3085	6256	12816
obi-CTR	1839	2372	4817	10372	21013

models (as shown in Table 3). In a practical online recommender system, one may want to choose a proper value of K to balance the tradeoff between accuracy and computational efficiency.

5 Conclusion

This paper investigated online learning algorithms for making inference algorithm for Collaborative Topic Regression (CTR) model practical for real-world online recommender systems. Specifically, unlike bdi-CTR that loosely combines LDA and PMF, we propose a novel Online Bayesian Inference algorithm for CTR model (obi-CTR) which performs a joint optimization of both LDA and PMF to achieve a tight coupling. Our encouraging results showed that obi-CTR converges much faster than the other competing algorithms in the online learning, and thus achieved the best prediction performance among all the compared algorithms. Our future work will analyze model interpretability and theoretical performance of the proposed algorithms.

Acknowledgements This research is supported in part by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative, and the China Knowledge Centre for Engineering Sciences and Technology (No. CKCEST-2014-1-5). This work was done when the first two authors visited Prof Hoi's research group at Singapore Management University.

References

Agarwal, D. & Chen, B.-C. (2010). fLDA: Matrix factorization through latent Dirichlet allocation. *Proceedings of the third ACM international conference on web search and data mining*, (pp. 91–100). ACMM.

Ahn, S., Korattikara, A. & Welling, M. (2012). Bayesian posterior sampling via stochastic gradient fisher scoring. arXiv preprint arXiv:1206.6380.

Almazro, D., Shahatah, G., Albdulkarim, L., Kherees, M., Martinez, R. & Nzoukou, W. (2010). A survey paper on recommender systems. arXiv preprint arXiv:1006.5278 .

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of machine Learning research*, *3*, 993–1022.

Blondel, M., Kubo, Y. & Ueda, N. (2014). Online passive-aggressive algorithms for non-negative matrix factorization and completion. *Proceedings of the seventeenth international conference on artificial intelligence and statistics*, (pp. 96–104).

Breese, J. S., Heckerman, D. & Kadie, C. (1998). Empirical analysis of predictive Aagorithms for collaborative filtering. In *Proceedings of the fourteenth conference on uncertainty in artificial intelligence*, (pp. 43–52). Morgan Kaufmann Publishers Inc.

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., & Jordan, M. I. (2013). Streaming variational Bayes. *Advances in Neural Information Processing Systems*, *26*, 1727–1735.

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge: Cambridge University Press.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006). Online passive-aggressive algorithms. *The Journal of Machine Learning Research*, *7*, 551–585.

Crammer, K., & Singer, Y. (2003). Ultraconservative online algorithms for multiclass problems. *The Journal of Machine Learning Research*, *3*, 951–991.

Ding, X., Jin, X., Li, Y., & Li, L. (2013). Celebrity recommendation with collaborative social topic regression. In *Proceedings of the twenty-third international joint conference on artificial intelligence*, (pp. 2612–2618). AAAI Press.

Foulds, J., Boyles, L., DuBois, C., Smyth, P. & Welling, M. (2013). Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 446–454). ACM.

Gentile, C. (2002). A new approximate maximal margin classification algorithm. *The Journal of Machine Learning Research*, *2*, 213–242.

Gopalan, P. K., Charlin, L., & Blei, D. (2014). Content-based recommendations with Poisson factorization. *Advances in Neural Information Processing Systems*, *27*, 3176–3184.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, *23*, 856–864.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, *14*(1), 1303–1347.

Hoi, S. C., Jin, R., Zhao, P., & Yang, T. (2013). Online multiple kernel classification. *Machine Learning*, *90*(2), 289–316.

Hoi, S. C., Wang, J., & Zhao, P. (2014). Libol: A library for online learning algorithms. *The Journal of Machine Learning Research*, *15*(1), 495–499.

Hoi, S. C., Zhao, P., Zhao, P. & Hoi, S. C. (2013). Cost-sensitive double updating online learning and its application to online anomaly detection. *SDM*, SIAM, pp. 207–215.

Honkela, A., & Valpola, H. (2003). Online variational Bayesian learning. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation* (pp. 803–808).

Hu, Y., Koren, Y. & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *Eighth IEEE international conference on data mining*, (pp. 263–272). IEEE.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*(2), 183–233.

Kang, J.-H. & Lerman, K. (2013). LA-CTR: A limited attention collaborative topic regression for social media, arXiv preprint arXiv:1311.1247 .

Kingma, D. P. & Welling, M. (2013). Auto-encoding variational Bayes, arXiv preprint arXiv:1312.6114.

Koren, Y., Bell, R., Volinsky, C., et al. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30–37.

Lu, Z., Dou, Z., Lian, J., Xie, X. & Yang, Q. (2015). Content-based collaborative filtering for news topic recommendation, *Twenty-ninth AAAI conference on artificial intelligence*.

McAuley, J. & Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text, *Proceedings of the 7th ACM conference on recommender systems*, (pp. 165–172). ACM.

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. *Advances in Neural Information Processing Systems*, *21*, 121–128.

McInerney, J., Ranganath, R., & Blei, D. (2015). The population posterior and Bayesian modeling on streams. *Advances in Neural Information Processing Systems*, *28*, 1153–1161.

Mimno, D., Hoffman, M. & Blei, D. (2012). Sparse stochastic inference for latent Dirichlet allocation, arXiv preprint arXiv:1206.6425.

Mnih, A., & Salakhutdinov, R. (2007). Probabilistic matrix factorization. *Advances in Neural Information Processing Systems*, *20*, 1257–1264.

Patterson, S., & Teh, Y. W. (2013). Stochastic gradient Riemannian Langevin dynamics on the probability simplex. *Advances in Neural Information Processing Systems*, *26*, 3102–3110.

Purushotham, S., Liu, Y. & Kuo, C.-C. J. (2012). Collaborative topic regression with social matrix factorization for recommendation systems, arXiv preprint arXiv:1206.4684.

Robert, C., & Casella, G. (2013). *Monte Carlo statistical methods*. Berlin: Springer-Verlag New York, Inc.

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

Shalev-Shwartz, S. (2011). Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, *4*(2), 107–194.

Shi, T. & Zhu, J. (2014). Online Bayesian passive-aggressive learning. In *Proceedings of the 31st international conference on machine learning*, (pp. 378–386).

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, *2009*, 4.

Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. *Advances in Neural Information Processing Systems*, *26*, 2643–2651.

Wang, C. & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, (pp. 448–456). ACM.

Wang, H., Chen, B., & Li, W.-J. (2013). *Collaborative topic regression with social regularization for tag recommendation*. In Proceedings of the twenty-third international joint conference on artificial intelligence, (pp. 2719–2725). AAAI Press.

Wang, H., Shi, X. & Yeung, D.-Y. (2015). Relational stacked denoising autoencoder for tag recommendation. In *Twenty-ninth AAAI conference on artificial intelligence*.

Wang, H., Wang, N. & Yeung, D.-Y. (2014). Collaborative deep learning for recommender systems, arXiv preprint arXiv:1409.2944.

Wang, J., Zhao, P., & Hoi, S. C. (2014). Cost-sensitive online classification. *IEEE Transactions on Knowledge and Data Engineering*, *26*(10), 2425–2438.

Welling, M. & Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, (pp. 681–688).

Zhao, P., Hoi, S. C., & Jin, R. (2011). Double updating online learning. *The Journal of Machine Learning Research*, *12*, 1587–1615.

Zhu, J., Ahmed, A., & Xing, E. P. (2012). MedLDA: Maximum margin supervised topic models. *The Journal of Machine Learning Research*, *13*(1), 2237–2278.