A Unified Contextual Bandit Framework for Long- and Short-Term Recommendations

M. Tavakol[1,2] and U. Brefeld[1]

[1] Leuphana Universität Lüneburg, Germany
[2] Technische Universität Darmstadt,Germany
{tavakol, brefeld}@leuphana.de

Abstract. We present a unified contextual bandit framework for recommendation problems that is able to capture long- and short-term interests of users. The model is devised in dual space and the derivation is consequentially carried out using Fenchel-Legrende conjugates and thus leverages to a wide range of tasks and settings. We detail two instantiations for regression and classification scenarios and obtain well-known algorithms for these special cases. The resulting general and unified framework allows for quickly adapting contextual bandits to different applications at-hand. The empirical study demonstrates that the proposed long- and short-term framework outperforms both, short-term and long-term models on data. Moreover, a tweak of the combined model proves beneficial in cold start problems.

Keywords: Recommendation, contextual bandits, dual optimization, personalization

1 Introduction

Recommender systems are designed to serve user needs. While some needs arise on short notice due to weather changes, news articles, or advertisements, others manifest over a long time span and express general interest in, for example, cars, stock markets, or garments in favored colors. User needs are therefore driven by individual *long-term* and collective *short-term* interests where the latter is highly influenced by the zeitgeist and common trends.

Traditional recommender systems, however, focus on only one aspect of recommendation, that is either on a personalized long-term, or an ad-hoc short-term approach. Collaborative filtering-based methods [8, 6], for example, aim to consider long-term preferences of users, while others aim topics of user sessions and focus on short-term interests [15, 2, 13]. In general, context-aware approaches [9], and their kernelized variants [14, 4], may be leveraged to meet both aspects. On the other hand, some recent works focus on context-aware bandits for personalization purposes. Collaborative contextual bandits are introduced in [16] where the context and payoffs are shared among the neighboring users to reduce learning complexity and overall regret. In addition, contextual bandit are used to learn the latent structure of users in probabilistic settings to cope with

cold-start scenarios [17, 12]. Nevertheless, these methods are usually tailored to solve very specific recommendation tasks and may not be applicable to different scenarios. Therefore, a more flexible and comprehensive approach is required to cope with diverse facets of recommendation.

In this paper, we present a unified contextual bandit framework to capture long- and short-term interests of users. The underlying model consists of a contextual (the short-term) and an individual user-based (the long-term) part to determine the expected reward,

$$\mathbb{E}[r_{t,a_i}|u_j] = \underbrace{\boldsymbol{\theta}_i^\top \boldsymbol{x}_t}_{Short-term} + \underbrace{\boldsymbol{\beta}_j^\top \boldsymbol{z}_{a_i}}_{Long-term} + b_i.$$

In the above composition, the expected reward is computed from two distinct parts. The first term models the short-term behavior for a given context $\boldsymbol{x}_t$ at time t. The context determines the recent trend or the topical interest of the current session. In the short-term part, the outcome of choosing each arm a_i for the given context $\boldsymbol{x}_t$ is specified linearly and by its weight vector, $\boldsymbol{\theta}_i$.

The long-term model, on the other hand, allows to capture individual interests for user u_j across item features, $\boldsymbol{z}_{a_i}$ (describing item a_i). We propose to connect the short-term and long-term recommendation in one unified model. Note that b_i acts as constant term in the linear model for each arm. The optimization is performed simultaneously for all the arms so that the short-term part serves as a joint popularity-based predictor while the long-term part acts as an individual offset. All derivations are carried out in the dual space using Fenchel-Legendre conjugates of the loss functions which renders our approach as a framework for a wide range of loss functions. We obtain LinUCB [9] and LogUCB [10] as special cases for regression and classification scenarios, respectively.

The next section derives a generalized recommendation model in dual space which is followed by its instantiations for regression and classification scenarios. Section 3 contains our main contribution and presents the combination of long-term and short-term recommender systems within the unified framework with potential optimization methods. Additionally, possible extensions for our proposed approach is discussed in Section 4. We present empirical studies in Section 5 and Section 6 concludes.

2 Linear Bandits in Dual Space

In this paper, we focus on sequential recommender systems for m users, $U = \{u_1, u_2, ..., u_m\}$, and n items, $A = \{a_1, a_2, ..., a_n\}$. Every item a_i is characterized by a set of attributes given by a feature vector $\boldsymbol{z}_{a_i} \in \mathbb{R}^k$. At each time step t, the goal of the system is to recommend items for the actual context of the ongoing session, which is described by a feature vector $\boldsymbol{x}_t \in \mathbb{R}^d$. In the following, we show how to derive the general optimization framework for linear bandits in dual space considering short-term information.

2.1 General Optimization

Assume that the learning procedure for every item (arm) consists of T_i trials, and for every context $\boldsymbol{x}_t$ the reward r_t is obtained. Therefore, $\{(\boldsymbol{x}_t, r_t)\}_{t=1}^{T_i}$ is the set of T_i samples and their corresponding rewards. The reward corresponds to the user feedback w.r.t. the recommended items; its domain depends on the application at-hand; e.g., $r_t \in \{1, 0\}$ for click/no click. We deploy a contextual bandit framework with linear payoff function for arm a_i,

$$h^{(i)}_{\boldsymbol{\theta}_i, b_i}(\boldsymbol{x}_t) = \boldsymbol{\theta}_i^\top \boldsymbol{x}_t + b_i,$$

where hypothesis h predicts the expected payoff for the i-th arm, $\mathbb{E}[r_{t,a_i}]$, and $\boldsymbol{\theta}$ contains the model parameters. The bandit framework learns every hypothesis $h^{(i)}$ independently of the other arms. We therefore discard the index i in the remainder of this section for ease of notation and address the problem for a single arm.

Given an arbitrary loss function $V(\cdot, r_t)$, and using l_2 norm regularizer, the optimization problem can be stated as

$$\inf_{\boldsymbol{\theta}, b} \quad \frac{1}{T} \sum_{t=1}^{T} V(\boldsymbol{\theta}^\top \boldsymbol{x}_t + b, r_t) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2.$$

We rewrite the objective by incorporating y_t as shorthand for the predicted payoff. Using $C = \frac{1}{\lambda T}$ gives

$$\inf_{\boldsymbol{\theta}, b, \boldsymbol{y}} \quad C \sum_{t=1}^{T} V(y_t, r_t) + \frac{1}{2} \|\boldsymbol{\theta}\|^2 \quad s.t. \quad \forall t : \boldsymbol{\theta}^\top \boldsymbol{x}_t + b = y_t.$$

The equivalent unconstrained problem is derived by incorporating Lagrange multipliers, $\boldsymbol{\alpha} \in \mathbb{R}^T$,

$$\sup_{\boldsymbol{\alpha}} \quad \inf_{\boldsymbol{\theta}, \boldsymbol{y}, b} \quad C \sum_{t=1}^{T} V(y_t, r_t) + \frac{1}{2} \|\boldsymbol{\theta}\|^2 - \sum_{t=1}^{T} \alpha_t (\boldsymbol{\theta}^\top \boldsymbol{x}_t + b - y_t).$$

Setting the partial derivatives w.r.t. b and $\boldsymbol{\theta}$ to zero, leads to the following condition

$$\mathbf{1}^\top \boldsymbol{\alpha} = 0 \qquad \text{and} \qquad \boldsymbol{\theta} = \sum_{t=1}^{T} \alpha_t \boldsymbol{x}_t = X^\top \boldsymbol{\alpha},$$

where $X \in \mathbb{R}^{T \times d}$ is the design matrix given by the training data. Substituting the optimality conditions into the optimization function yields

$$\sup_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha} = 0} \quad \inf_{\boldsymbol{y}} \quad C \sum_{t=1}^{T} \big(V(y_t, r_t) + \frac{1}{C} \alpha_t y_t \big) - \frac{1}{2} \boldsymbol{\alpha}^\top X X^\top \boldsymbol{\alpha}.$$

Moreover, we move the infimum inside the summation as it solely depends on the first term. Using $\inf_w f(w) = -\sup_w -f(w)$, we obtain

$$\sup_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha}=0} \quad -C \sum_{t=1}^{T} \sup_{y_t} \big(-\frac{\alpha_t}{C} y_t - V(y_t, r_t) \big) - \frac{1}{2} \boldsymbol{\alpha}^\top X X^\top \boldsymbol{\alpha}.$$

Recall that the Fenchel-Legendre conjugate of a function g is defined as $g^*(\boldsymbol{u}) = \sup_{\boldsymbol{x}} \boldsymbol{u}^\top \boldsymbol{x} - g(\boldsymbol{x})$ [3]. Thus, the dual loss is given by

$$V^*\big(-\frac{\alpha_t}{C}, r_t \big) = \sup_{y_t} -\frac{\alpha_t}{C} y_t - V(y_t, r_t).$$

(for a comprehensive list of dual losses see [11]). The generalized optimization problem in dual space is therefore reduces to

$$\sup_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha}=0} \quad -C \sum_{t=1}^{T} V^*\big(-\frac{\alpha_t}{C}, r_t \big) - \frac{1}{2} \boldsymbol{\alpha}^\top X X^\top \boldsymbol{\alpha}. \tag{1}$$

2.2 Upper Confidence Bound

The challenge in bandit-based approaches is to balance exploration and exploitation to minimize the regret. Auer [1] demonstrates that confidence bounds provide useful means to balance the two oppositional strategies. The idea is to use the predicted reward together with its confidence interval to reflect the uncertainty of the model given the actual context. Thus, gathering enough information to reduce the uncertainty in a multi-armed bandit is as important as maximizing the reward.

In our contextual bandit, the expected payoff is approximated by a linear model with an arbitrary loss function where a general optimization approach is used to estimate the parameters. The uncertainty U of the obtained value for each arm is therefore proportional to the standard deviation σ of the expected payoff, $U = c\sigma$, where the variance σ^2 is estimated from training points in neighbouring contexts as well as the model parameters. The uncertainty is added as an upper bound to the prediction to produce a confidence bound for selection strategy across the arms. The computation of the confidence bound depends on the choice of the loss function. We illustrate the obtained bounds for two special cases in the remainder.

2.3 Instantiations

In the following parts, we demonstrate two well-known optimization problems which can be recovered from Equation (1) by substituting the corresponding loss functions. The instantiations illustrate how a general platform simplifies comparing and analyzing various loss functions in different situations.

Squared Loss. The first instantiation deals with regression scenarios for real-valued payoffs, $r_t \in \mathbb{R}$. The squared loss function and its dual are given by

$$V(y_t, r_t) = \frac{1}{2}(y_t - r_t)^2 \quad \text{and} \quad V^*(s_t, r_t) = \frac{1}{2}s_t^2 + s_t r_t,$$

where the latter can be rewritten as

$$V^*\big(-\frac{\alpha_t}{C}, r_t\big) = \frac{1}{2C^2}\alpha_t^2 - \frac{1}{C}\alpha_t r_t.$$

Incorporating the conjugate loss function into Equation (1) gives

$$\max_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha}=0} \quad -\frac{1}{2C}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \boldsymbol{r} - \frac{1}{2}\boldsymbol{\alpha}^\top X X^\top \boldsymbol{\alpha}, \tag{2}$$

where the supremum becomes a maximum as the loss function is continuous. The equivalent problem in the primal space corresponds to ridge regression where parameters are determined by optimizing the regularized sum of squared errors,

$$\min_{\boldsymbol{\theta}, b} \quad \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}(\boldsymbol{\theta}^\top \boldsymbol{x}_t + b - r_t)^2 + \frac{\lambda}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta}.$$

To obtain $\boldsymbol{\theta}$, we set its gradient to 0 which yields $\boldsymbol{\theta} = -\frac{1}{\lambda T}\sum_{t=1}^{T}(\boldsymbol{\theta}^\top \boldsymbol{x}_t + b - r_t)\boldsymbol{x}_t$. The relation $\alpha_t = -\frac{1}{\lambda T}(\boldsymbol{\theta}^\top \boldsymbol{x}_t + b - r_t)$ holds and we have

$$\boldsymbol{\theta} = \sum_{t=1}^{T}\alpha_t \boldsymbol{x}_t = X^\top \boldsymbol{\alpha}.$$

For the threshold parameter b, we obtain the equation $\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\theta}^\top \boldsymbol{x}_t + b - r_t) = 0$, and thus arrive at the optimality conditions

$$-\lambda \sum_{t=1}^{T}\alpha_t = 0 \quad \Rightarrow \quad \mathbf{1}^\top \boldsymbol{\alpha} = 0.$$

Expanding the terms in the summation and substituting the optimality conditions leads to the optimization problem

$$\min_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha}=0} \quad C\big(\frac{1}{2}\boldsymbol{\alpha}^\top X X^\top X X^\top \boldsymbol{\alpha} - \boldsymbol{r}^\top X X^\top \boldsymbol{\alpha}\big) + \frac{1}{2}\boldsymbol{\alpha}^\top X X^\top \boldsymbol{\alpha},$$

where $C = \frac{1}{\lambda T}$. By removing $XX^\top$ from all the terms and converting the minimization into a maximization, we have

$$\max_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha}=0} \quad -\frac{1}{2}\boldsymbol{\alpha}^\top X X^\top \boldsymbol{\alpha} + \boldsymbol{r}^\top \boldsymbol{\alpha} - \frac{1}{2C}\boldsymbol{\alpha}^\top \boldsymbol{\alpha},$$

which precisely recovers Equation (2). The confidence bound for the linear bandit with square loss is given by (cmp. also [9])

$$U = c\sqrt{\boldsymbol{x}_t^\top (X^\top X + \lambda I)^{-1}\boldsymbol{x}_t}.$$

Logistic Loss. In this section, we derive the optimization problem for the logistic loss which is defined as

$$V(y_t, r_t) = \log(1 + \exp(-y_t r_t)).$$

The conjugate of loss function is given by

$$V^*(-\frac{\alpha_t}{r_t}, r_t) = (1 - \frac{\alpha_t}{Cr_t}) \log(1 - \frac{\alpha_t}{Cr_t}) + \frac{\alpha_t}{Cr_t} \log(\frac{\alpha_t}{Cr_t}),$$

and incorporating the latter into Equation (1) leads to Equation (3)

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha}=0} & - C \sum_{t=1}^{T} [(1 - \frac{\alpha_t}{Cr_t}) \log(1 - \frac{\alpha_t}{Cr_t}) + \frac{\alpha_t}{Cr_t} \log(\frac{\alpha_t}{Cr_t})] \\ & - \frac{1}{2} \boldsymbol{\alpha}^\top X X^\top \boldsymbol{\alpha}. \end{aligned} \tag{3}$$

The analogous problem in primal space is known as a logistic regression [7] and gives

$$\min_{\hat{\boldsymbol{\alpha}}} \quad \frac{1}{2} \Big\| \sum_{t=1}^{T} \hat{\alpha}_t r_t \boldsymbol{x}_t \Big\|^2 + C \sum_{t=1}^{T} G(\frac{\hat{\alpha}_t}{C}), \quad s.t. \quad \sum_{t=1}^{T} \hat{\alpha}_t r_t = 0,$$

where $G(\delta) = \delta \log \delta + (1 - \delta) \log(1 - \delta)$. Setting $\alpha_t = \hat{\alpha}_t r_t$, and converting the minimization into a maximization recovers Equation (3).

The covariance of the parameters for the logistic regression problem is given by $\Sigma = X^T V X$, where V is diagonal matrix of $\pi(1 - \pi)$, and π is computed by the sigmoid function ρ, i.e., $\pi = \rho(X^\top \boldsymbol{\theta})$. Consequentially, the lower and upper confidence bounds are given by

$$U_{lo} = \rho\big(\hat{r_t} - c\sqrt{\boldsymbol{x}_t^\top \Sigma^{-1} \boldsymbol{x}_t}\big), \quad U_{up} = \rho\big(\hat{r_t} + c\sqrt{\boldsymbol{x}_t^\top \Sigma^{-1} \boldsymbol{x}_t}\big),$$

respectively [5]. The confidence bound for the contextual bandit is therefore $U = U_{up} - U_{lo}$. Mahajan et al. [10] introduce a variance approximation technique to obtain the confidence bound for logistic loss for probit functions.

3 A Unified Contextual Bandit

In our setting, personalized and user specific information cannot simply be incorporated into the bandit by another type of context. Instead, we suggest to incorporate a long-term model into the short-term approach of the previous section. Therefore, we are able to model the behavior of users for the recommendation process. The long-term part captures the interests of user u_t for every arm a_i. We thus assume a separate set of parameters for the personalized part of the model, given by $\boldsymbol{\beta}_j \in \{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m\}$, where $\boldsymbol{\beta}_j \in \mathbb{R}^k$. The long-term preferences of users are also modelled by a linear relationship $\boldsymbol{\beta}_j^\top \boldsymbol{z}_{a_i}$. For user $u_t \equiv u_j$, the joint long- and short-term model is

$$h^{(i)}_{\boldsymbol{\theta}_i, \boldsymbol{\beta}_t, b_i}(\boldsymbol{x}_t, \boldsymbol{z}_{a_i}) = \boldsymbol{\theta}_i^\top \boldsymbol{x}_t + \boldsymbol{\beta}_t^\top \boldsymbol{z}_{a_i} + b_i.$$

3.1 The Objective Function

As in Section 2, all the parameters of the short-term model are still independent from every other item as well as the user parameters among themselves. However, user parameters $\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m\}$ are shared across the arms and that makes the objective function to be connected for all the arms and users. Hence, the general optimization problem with arbitrary loss function, $V(\cdot, r_t)$ becomes

$$\inf_{\substack{\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_n \\ \boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_m \\ \boldsymbol{b}}} \quad \frac{1}{T}\sum_{t=1}^{T} V(\boldsymbol{\theta}_t^\top \boldsymbol{x}_t + \boldsymbol{\beta}_t^\top \boldsymbol{z}_t + b_t, r_t) + \frac{\lambda}{2}\sum_i \|\boldsymbol{\theta}_i\|^2 + \frac{\hat{\mu}}{2}\sum_j \|\boldsymbol{\beta}_j\|^2$$

where λ and $\hat{\mu}$ are the regularization parameters for the item and user weights, respectively. Let $C = \frac{1}{\lambda T}$, $\mu = \frac{\hat{\mu}}{\lambda}$, and $\boldsymbol{y} = (\dots, y_t, \dots)^\top$, we have

$$\inf_{\substack{\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_n \\ \boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_m \\ \boldsymbol{b},\boldsymbol{y}}} \quad C\sum_{t=1}^{T} V(y_t, r_t) + \frac{1}{2}\sum_i \|\boldsymbol{\theta}_i\|^2 + \frac{\mu}{2}\sum_j \|\boldsymbol{\beta}_j\|^2$$

$$s.t. \quad \forall t: \quad \boldsymbol{\theta}_t^\top \boldsymbol{x}_t + \boldsymbol{\beta}_t^\top \boldsymbol{z}_t + b_t = y_t,$$

which results in the Lagrange function

$$\sup_{\boldsymbol{\alpha}} \quad \inf_{\substack{\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_n \\ \boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_m \\ \boldsymbol{b},\boldsymbol{y}}} \quad C\sum_{t=1}^{T} V(y_t, r_t) + \frac{1}{2}\sum_i \|\boldsymbol{\theta}_i\|^2 + \frac{\mu}{2}\sum_j \|\boldsymbol{\beta}_j\|^2$$
$$- \sum_{t=1}^{T} \alpha_t(\boldsymbol{\theta}_t^\top \boldsymbol{x}_t + \boldsymbol{\beta}_t^\top \boldsymbol{z}_t + b_t - y_t).$$

Note that $\{\boldsymbol{\theta}_t, \boldsymbol{z}_t\} \in \{\{\boldsymbol{\theta}_1, \boldsymbol{z}_{a_1}\}, \dots, \{\boldsymbol{\theta}_n, \boldsymbol{z}_{a_n}\}\}$, $\boldsymbol{\beta}_t \in \{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m\}$, and $\boldsymbol{b_t} \in \{\boldsymbol{b}_1, \dots, \boldsymbol{b}_n\}$. The derivatives with respect to $\boldsymbol{\theta}_i$ generate

$$\boldsymbol{\theta}_i = \sum_{\substack{t \\ \boldsymbol{\theta}_t = \boldsymbol{\theta}_i}} \alpha_t \mathbf{x}_t = \sum_t \delta_{it}\alpha_t \mathbf{x}_t = (X \circ \boldsymbol{\delta_i})^\top \boldsymbol{\alpha}.$$

In the above equation, $\boldsymbol{\delta_i} \in \mathbb{R}^T$ is a binary vector which is 1 when $\boldsymbol{\theta}_t = \boldsymbol{\theta}_i$, and zero otherwise. $X \in \mathbb{R}^{T\times d}$ is the design matrix of input vectors, and $\circ$ is element-wise product (each element in the vector multiplies by a row in the matrix). We compute the derivations for $\boldsymbol{\beta}_j$,

$$\boldsymbol{\beta}_j = \frac{1}{\mu}\sum_{\substack{t \\ \boldsymbol{\beta}_t = \boldsymbol{\beta}_j}} \alpha_t \mathbf{z}_t = \frac{1}{\mu}\sum_t \phi_{jt}\alpha_t \mathbf{z}_t = \frac{1}{\mu}(Z \circ \boldsymbol{\phi_j})^\top \boldsymbol{\alpha},$$

where again $\boldsymbol{\phi}_j \in \mathbb{R}^T$ is the indicator vector for the corresponding user and Z is the design matrix for the items features. Additionally, the derivatives w.r.t. b_i gives

$$\forall i, \quad \sum_{t:b_t=b_i} \alpha_t = 0 \quad \rightarrow \mathbf{1}^\top \boldsymbol{\alpha} = 0.$$

Substituting the obtained conditions in the original problem leads to

$$\begin{aligned} \sup_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha}=0} \quad \inf_{\boldsymbol{y}} \quad & C \sum_{t=1}^{T} [V(y_t, r_t) + \frac{1}{C} \alpha_t y_t] \\ & - \frac{1}{2} \sum_i \boldsymbol{\alpha}^\top (X \circ \boldsymbol{\delta_i})(X \circ \boldsymbol{\delta_i})^\top \boldsymbol{\alpha} - \frac{1}{2\mu} \sum_j \boldsymbol{\alpha}^\top (Z \circ \boldsymbol{\phi_j})(Z \circ \boldsymbol{\phi_j})^\top \boldsymbol{\alpha}, \end{aligned}$$

which can be written as

$$\begin{aligned} \sup_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha}=0} \quad & -C \sum_{t=1}^{T} \sup_{y_t} (-\frac{\alpha_t}{C} y_t - V(y_t, r_t)) \\ & - \frac{1}{2} \sum_i \boldsymbol{\alpha}^\top (X \circ \boldsymbol{\delta_i})(X \circ \boldsymbol{\delta_i})^\top \boldsymbol{\alpha} - \frac{1}{2\mu} \sum_j \boldsymbol{\alpha}^\top (Z \circ \boldsymbol{\phi_j})(Z \circ \boldsymbol{\phi_j})^\top \boldsymbol{\alpha}. \end{aligned}$$

Finally, by converting the first term to the conjugate of the loss function using Fenchel-Legendre conjugates, we obtain

$$\begin{aligned} \sup_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha}=0} \quad & - C \sum_{t=1}^{T} V^*(-\frac{\alpha_t}{C}, r_t) - \frac{1}{2} \sum_i \boldsymbol{\alpha}^\top (X \circ \boldsymbol{\delta_i})(X \circ \boldsymbol{\delta_i})^\top \boldsymbol{\alpha} \\ & - \frac{1}{2\mu} \sum_j \boldsymbol{\alpha}^\top (Z \circ \boldsymbol{\phi_j})(Z \circ \boldsymbol{\phi_j})^\top \boldsymbol{\alpha}. \end{aligned} \tag{4}$$

Equation (4) constitutes a generalized optimization problem for contextual bandits with arbitrary loss function. It contains the short-term model in Equation (1) as a special case when no personal long-term interests need to be captured.

3.2 Optimization

Equation (4) can be optimized with various optimization methods depending on the loss function as well as standard techniques such as gradient-based approaches. For real-time applications and online scenarios, model updates can be performed using (mini-) batches at regular intervals as well, for efficiency. The objective function needs to be maximized w.r.t. the dual parameters $\boldsymbol{\alpha}$ and is given by

$$\begin{aligned} \sup_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha}=0} \quad & - C \mathbb{I}^\top V^*(-\frac{\boldsymbol{\alpha}}{C}, \boldsymbol{r}) - \frac{1}{2} \sum_i \boldsymbol{\alpha}^\top (X \circ \boldsymbol{\delta_i})(X \circ \boldsymbol{\delta_i})^\top \boldsymbol{\alpha} \\ & - \frac{1}{2\mu} \sum_j \boldsymbol{\alpha}^\top (Z \circ \boldsymbol{\phi_j})(Z \circ \boldsymbol{\phi_j})^\top \boldsymbol{\alpha}. \end{aligned}$$

The gradient w.r.t. $\boldsymbol{\alpha}$ is obtained by computing the derivatives

$$-C\frac{\partial V^*(-\frac{\boldsymbol{\alpha}}{C}, \boldsymbol{r})}{\partial \boldsymbol{\alpha}} - [\sum_i (X \circ \boldsymbol{\delta_i})(X \circ \boldsymbol{\delta_i})^\top]\boldsymbol{\alpha} - \frac{1}{\mu}[\sum_j (Z \circ \boldsymbol{\phi_j})(Z \circ \boldsymbol{\phi_j})^\top]\boldsymbol{\alpha} - \gamma\mathbb{I} = 0.$$

The actual form of the gradient depends on the dual loss V^* and further derivations are omitted accordingly. Note that instantiations often give rise to more sophisticated and efficient optimization techniques than the general form in Equation (4) allows, see also Sections 2.3 and 2.3. Nevertheless, the sketched gradient-based approach will always work in case a general optimizer is needed, e.g., in cases where several loss functions should be tried out. Once the optimal parameters, $\boldsymbol{\alpha}^{opt}$, have been found, they can be used to compute the primal parameters

$$\boldsymbol{\theta}_i = (X \circ \boldsymbol{\delta_i})^\top \boldsymbol{\alpha}^{opt}, \quad \boldsymbol{\beta}_j = \frac{1}{\mu}(Z \circ \boldsymbol{\phi_j})^\top \boldsymbol{\alpha}^{opt}.$$

Alternatively, kernels $K_X = \phi_X(X, X)$ and $K_Z = \phi_Z(Z, Z)$ could be deployed in the dual representation to allow for non-linear transformations and convolutions of the feature space.

Once the required parameters are found, the payoff estimates are used together with the respective confidence interval U of the arm to choose the arm with the maximum upper confidence value according to

$$a_t = \arg\max_{a_i \in A} \quad \boldsymbol{\theta}_i^\top \boldsymbol{x}_t + \boldsymbol{\beta}_t^\top \boldsymbol{z}_{a_i} + b_i + U_{i,t}.$$

Learning with Squared Loss In this section, we present the optimization algorithm for a special case of unified contextual bandit framework with squared loss. As it is mentioned in Section 2.3, the conjugate of squared loss is given by

$$V^*(-\frac{\alpha_t}{C}, r_t) = \frac{1}{2C^2}\alpha_t^2 - \frac{1}{C}\alpha_t r_t,$$

which leads to the following objective

$$\begin{aligned}\max_{\boldsymbol{\alpha}, \mathbf{1}^\top\boldsymbol{\alpha}=0} \quad & -\frac{1}{2C}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + \boldsymbol{r}^\top\boldsymbol{\alpha} - \frac{1}{2}\sum_i \boldsymbol{\alpha}^\top(X \circ \boldsymbol{\delta_i})(X \circ \boldsymbol{\delta_i})^\top\boldsymbol{\alpha} \\ & - \frac{1}{2\mu}\sum_j \boldsymbol{\alpha}^\top(Z \circ \boldsymbol{\phi_j})(Z \circ \boldsymbol{\phi_j})^\top\boldsymbol{\alpha}.\end{aligned}$$

The summation $\sum_i (X \circ \boldsymbol{\delta_i})(X \circ \boldsymbol{\delta_i})^\top$ is equivalent to $(\sum_i \boldsymbol{\delta}_i \otimes \boldsymbol{\delta}_i^\top) \circ XX^\top$, where $\otimes$ stands for the vector outer product. Considering the same equivalency for the last term as well, we rewrite the equation as follows

$$\begin{aligned}\max_{\boldsymbol{\alpha}, \mathbf{1}^\top\boldsymbol{\alpha}=0} \quad & -\frac{1}{2C}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + \boldsymbol{r}^\top\boldsymbol{\alpha} \\ & - \frac{1}{2}\boldsymbol{\alpha}^\top[(\sum_i \boldsymbol{\delta}_i \otimes \boldsymbol{\delta}_i^\top) \circ XX^\top + \frac{1}{\mu}(\sum_i \boldsymbol{\phi}_i \otimes \boldsymbol{\phi}_i^\top) \circ ZZ^\top]\boldsymbol{\alpha}.\end{aligned}$$

By using *min* instead of *max*, setting $P = \frac{1}{C}\mathbb{I} + (\sum_i \boldsymbol{\delta}_i \otimes \boldsymbol{\delta}_i^\top) \circ XX^\top + \frac{1}{\mu}(\sum_i \boldsymbol{\phi}_i \otimes \boldsymbol{\phi}_i^\top) \circ ZZ^\top$, and $\mathbf{q} = -\mathbf{r}$, the problem becomes a standard quadratic optimization with a constraint,

$$\min_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha}=0} \quad \frac{1}{2}\boldsymbol{\alpha}^\top P \boldsymbol{\alpha} + \mathbf{q}^\top \boldsymbol{\alpha}. \tag{5}$$

Algorithm 1 summarizes the procedure of optimizing for the squared loss. In each iteration, the algorithm computes the UCB value of all arms for the observed user, and in line 14 chooses the arm with the highest value. The required parameters for the quadratic optimization are updated from line 15 to 22 which leads to optimizing $\boldsymbol{\alpha}$. The obtained vector is used to update the model parameters. Note that the objective function is optimized for all the parameters, therefore, it affects them all and not just one user and one item. In this algorithm, we assume that the covariance matrices of item and user parameters are independent from each other. Hence, we discard the correlation between them and obtain the variance by summing them as $\mathbf{z}_a^\top A_{u_t}^{-1} \mathbf{z}_a + \mathbf{x}_t^\top A_a^{-1} \mathbf{x}_t$ (line 11) in order to compute the confidence bound.

Algorithm 1 Short- and long-term regression UCB

1: Inputs: c, C, and μ
2: Initialize $X \leftarrow \emptyset_{0\times d}$, $Z \leftarrow \emptyset_{0\times k}$, $\mathbf{r} \leftarrow \emptyset$
3: **for** $t = 1, 2, ..., T$ **do**
4: **if** u_t is new **then** (Observe the user u_t and context $\mathbf{x}_t \in \mathbb{R}^{d\times 1}$)
5: $A_{u_t} \leftarrow \mathbb{I}_k \cdot \mu$, $\boldsymbol{\beta}_{u_t} \leftarrow \mathbf{0}_{k\times 1}$, $\boldsymbol{\phi}_{u_t} \leftarrow \mathbf{0}_{t\times 1}$
6: **end if**
7: **for all** $a \in A_t$ **do**
8: **if** a is new **then** (Observe the features of arm $\mathbf{z}_a \in \mathbb{R}^{k\times 1}$)
9: $A_a \leftarrow \mathbb{I}_d$, $\boldsymbol{\theta}_a \leftarrow \mathbf{0}_{d\times 1}$, $\boldsymbol{\delta}_a \leftarrow \mathbf{0}_{t\times 1}$
10: **end if**
11: $s_{t,a} = \mathbf{z}_a^\top A_{u_t}^{-1} \mathbf{z}_a + \mathbf{x}_t^\top A_a^{-1} \mathbf{x}_t$
12: $p_{t,a} = \boldsymbol{\theta}_a^\top \mathbf{x}_t + \boldsymbol{\beta}_{u_t}^\top \mathbf{z}_a + c\sqrt{s_{t,a}}$
13: **end for**
14: Choose arm $a_t = \arg\max_a p_{t,a}$ with tie broken randomly, and observe payoff r_t
15: $A_{a_t} = A_{a_t} + \mathbf{x}_t \mathbf{x}_t^\top$
16: $A_{u_t} = A_{u_t} + \mathbf{z}_{a_t} \mathbf{z}_{a_t}^\top$
17: $X \leftarrow [X; \mathbf{x}_t^\top]$ (Append vertically)
18: $Z \leftarrow [Z; \mathbf{z}_{a_t}^\top]$ (Append vertically)
19: $\mathbf{r} \leftarrow [\mathbf{r}, r_t]$
20: **for all** $a \in A_t$ and $u \in U_t$ **do**
21: Update $\boldsymbol{\delta}_a$ and $\boldsymbol{\phi}_u$
22: **end for**
23: **for all** $a \in A_t$ and $u \in U_t$ **do** (Obtain $\boldsymbol{\alpha}$ by optimizing equation 5)
24: $\boldsymbol{\theta}_a = (X \circ \boldsymbol{\delta_a})^\top \boldsymbol{\alpha}$
25: $\boldsymbol{\beta}_u = (Z \circ \boldsymbol{\phi_u})^\top \boldsymbol{\alpha}$
26: **end for**
27: **end for**

Learning with Logistic Loss Another special case of our unified framework is to apply the logistic loss for the optimization process. As we introduced in Section 2.3, the conjugate of logistic loss is as follows

$$V^*(-\frac{\alpha_t}{r_t}, r_t) = (1 - \frac{\alpha_t}{Cr_t})\log(1 - \frac{\alpha_t}{Cr_t}) + \frac{\alpha_t}{Cr_t}\log(\frac{\alpha_t}{Cr_t}).$$

Employing the above conjugate into the Equation (4) leads to

$$\min_{\boldsymbol{\alpha}, \mathbf{1}^\top \boldsymbol{\alpha}=0} \quad C\sum_{t=1}^{T}\left[(1 - \frac{\alpha_t}{Cr_t})\log(1 - \frac{\alpha_t}{Cr_t}) + \frac{\alpha_t}{Cr_t}\log(\frac{\alpha_t}{Cr_t})\right]$$
$$+ \frac{1}{2}\boldsymbol{\alpha}^\top[(\sum_i \boldsymbol{\delta}_i \otimes \boldsymbol{\delta}_i^\top) \circ XX^\top + \frac{1}{\mu}(\sum_i \boldsymbol{\phi}_i \otimes \boldsymbol{\phi}_i^\top) \circ ZZ^\top]\boldsymbol{\alpha}.$$

The procedure for learning the model is similar to Algorithm 1 in the previous section. Nevertheless, the objective function in line 29 needs to be optimized differently, and also computing $s_{t,a}$ in line 17. In the latter, the covariance matrix is computed for both set of parameters, $\Sigma_a = X^T V_a X$ and $\Sigma_{ut} = Z^T V_{ut} Z$, respectively. Therefore, $\boldsymbol{x}_t^\top \Sigma_a^{-1} \boldsymbol{x}_t + \boldsymbol{z}_t^\top \Sigma_{ut}^{-1} \boldsymbol{z}_t$ is used as the variance in computing the lower and upper confidence bounds (see Section 2.3). Note that gradient based methods are still applicable in the optimization part.

4 Discussion

In the following, we discuss some potential alternatives of our proposed approach which are suitable for particular circumstances.

4.1 Complexity of the Model

The presented unified model in Section 3 combines the contextual item model with the user interest in one framework. The model is therefore more than the vanilla bandit-based approaches that only model one of those. However, the model contains many parameters and the optimization part becomes more and more complex as the system size (both the number of items and users) grows. We propose to simplify the approach in two different directions; relaxing the item model or discarding the personalized term. Hence, we introduce four simplified cases of the combined approach as follows.

1. **Short-Term:** To model the payoff function only for the items, no personalization (aka. LinUCB [9]): $\mathbb{E}[r_{t,a_i}] = \boldsymbol{\theta}_i^\top \boldsymbol{x}_t$.
2. **Short-Term+Average:** Considering an average term for all the items, no personalization (resembling HybridUCB [9]): $\mathbb{E}[r_{t,a_i}] = \boldsymbol{\theta}_i^\top \boldsymbol{x}_t + \boldsymbol{\beta}^\top \boldsymbol{z}_{a_i}$.
3. **Long-Term:** Only personalized model: $\mathbb{E}[r_{t,a_i}|u_j] = \boldsymbol{\beta}_j^\top \boldsymbol{z}_{a_i}$.
4. **Long-Term+Average:** Incorporating the average term into the personalized model: $\mathbb{E}[r_{t,a_i}|u_j] = \boldsymbol{\beta}_j^\top \boldsymbol{z}_{a_i} + \boldsymbol{\theta}^\top \boldsymbol{z}_{a_i}$.

These cases are easily derivable from equations in Section 3. Note that the average part in case 2 and 4 depicts the item popularity in the recommender systems. We further examine the benefits of average models in Section 5.

4.2 Preference Based Bandits

One natural extension of our approach is to characterize the model in the preference-based setting. There are many systems with no available quantitative feedback, whereas the feedback is provided in terms of pairwise comparison between items. In such cases, the preferences are used in the learning process and the rankings are predicted directly from the model. In this section, we discuss how to phrase our bandit framework in a preference-based context.

We consider the contextual bandit problem in a way that the context is specified by the features of items to recommend. The model is thus defined by a single bandit which learns the preferences between items for all the users. Assume that $\mathbf{z}_i$ and $\mathbf{z}_k$ are the features of items a_i and a_k, respectively, and we assign $\mathbf{z}_{i \succ k} := \mathbf{z}_i - \mathbf{z}_k$ to show the preference of item a_i over a_k. The payoff is therefore determined as a linear model of the preference,

$$\mathbb{E}[r_{t,i \succ k} | u_t = u_j] = \boldsymbol{\theta}^\top \mathbf{z}_{i \succ k} + \boldsymbol{\beta}_t^\top \mathbf{z}_{i \succ k},$$

where $\boldsymbol{\theta}$ is the weight vector for the average model, while $\boldsymbol{\beta}_t = \boldsymbol{\beta}_j$ is the individual parameter for user j which acts as a personal offset. The above equation is theoretically analogous to the case number 4 in the previous section.

5 Empirical Study

The purpose of this section is to evaluate the performance of our combined contextual bandit approach compare to either short-term or long-term models. We use the squared loss in our experiments as in Algorithm 1. The quality of recommendation is measured via normalized average rank. For every test instance, a ranking of all items is inferred by the model. The position of the actually clicked item in the ranking is then normalized (divided by the total number of items) and averaged over all test samples. The empirical study illustrates that adding a long-term model describing the user preferences improves the short-term recommendation. Additionally, we show that the simplified average models are beneficial in cold start scenarios.

5.1 Data

The experiments are conducted on a real-world dataset from Zalando, a large European online fashion retailer, with anonymized click history of various users. The data is collected over time and bucketized into consecutive sessions. Each user interacts with the system in different sessions, and each session contains a sequence of products views. Products are described with some categorical attributes, such as *category*, *brand*, *color*, *gender*, *price level*, and *action*. We apply a one-hot encoding of the categorical features and enrich the representation by three additional features: the item popularity for each item, and "sale to view" as well as "view to action" ratios per user.The augmented dataset encompasses users with at least 5 sessions, where all sessions with more than one click are considered a valid session.

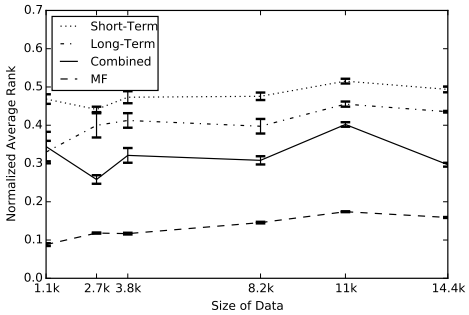


Fig. 1: Normalized average rank for different data sizes.

5.2 Overall Performance

In the first experiment, we examine how the combined approach performs on data sets of different sizes compared to the long- and short-term models and a matrix factorization baseline [8]. The parameters of the latter are optimized by model selection (200 factors, regularization constant 0.1). We thus generate several subsets of data by randomly sampling different numbers of users to obtain sets with about 1k user transactions to 15k. We split each set into training and test sets by reserving 80% of sessions for the former and assigning the rest to the test set. Note that there is no new user or new item in the test data.

The context in our setup is the feature vector of the previously viewed product. Therefore, the first click of each session is discarded and kept as the context for the next click. The reward value for each action is either 1 for the correct arm or -1 otherwise. We consider a fixed $c = 2.36$, and set regularization parameters $\lambda = \mu = 1$ for simplicity. Figure 1 depicts the results for our approach as well as the long- and short-term models averaged over several runs.

The figure shows that the combined approach outperforms both the long- and short-term methods in terms of average rank (lower is better). The short-term approach performs worse than the other two, since the data is obtained by sampling users, and there are many more items than users. However, the size of data does not change the behavior of the tested methods significantly apart from the combined model that improves performance with increasing data sizes; an indicator for the necessity of experiments at even larger scales. The matrix factorization baseline performs best when all users and items are known.

5.3 Cold Start

One of the main contributions of our proposed approach in the contextual setting is the ability to generalize over different items for individual users. This advantage suits well in cold start situations where content is highly dynamic and item and user sets change frequently. First, we demonstrate the behavior of the combined approach when new users and items appear in the test set.

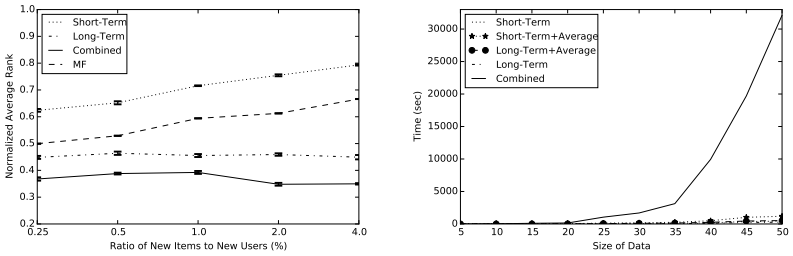
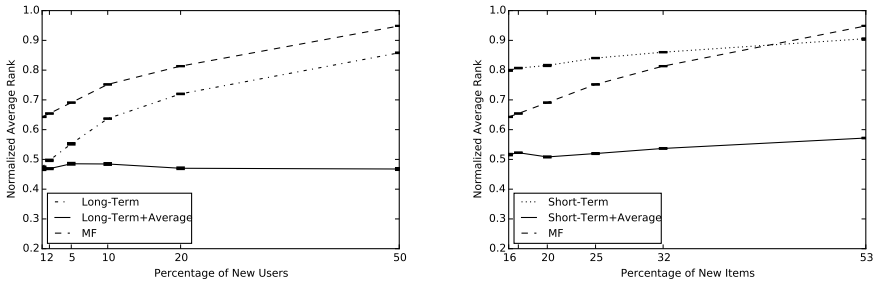


Fig. 2: Left: Normalized average rank for different ratios of new items to new users. Right: Execution time for different data sizes.

We create a subset of data from all the sessions of 100 randomly selected users. The data contains 1,295 sessions which gives an average of 13 sessions per user, and about 8,000 products. We split the data into training and test sets with different ratios for the percentage of new items and new users in the test data. To this purpose, we leave $d\%$ of the users and $e\%$ of the products to appear only in the test set such that it realizes a ratio of $\frac{e}{d}$ for the new items over new users. We train our combined approach as well as both long- and short-term models, where the context and reward setup is as in the previous section. Figure 2 (left) shows the behavior of different approaches; results are averaged over multiple runs.

The first impression from Figure 2 (left) suggests that although the combined method still outperforms the baselines, its performance declines a bit near the ratio of 1 when many new users and items are available. Unsurprisingly, the short-term model performs better for scenarios with only a few new items. This holds vice versa also for the long-term model that performs better for scenarios with almost constant sets of users. The performance of matrix factorization degrades significantly in the new setting which confirms the robustness of our combined method in real scenarios. However, the robustness comes at the cost of run-time: the combined approach is computationally expensive because of the involved convex optimization. The run-time analysis in Figure 2 (right) displays the exponential growth in execution time in comparison to the other approaches discussed in Section 4.1.

In this section, we focus on the evaluation of adding average models to the long- and short-term approaches. We conduct the experiments on a medium sized dataset to evaluate their performance. The dataset in this experiment contains all transactions of 500 random users. We split the data by modifying the percentage of new users and new items in the test set and analyze two cases. Figure 3 shows how adding the average term significantly improves the performance of both long-term and short-term models, respectively. As in the previous experiment, in Figure 3a, the performance of the long-term model decreases for increasing numbers of new users. By contrast, extending the long-term model by an average model remedies this effect and the extended model is able to cope with the



(a) Performance in terms of new users (b) Performance in terms of new items

Fig. 3: Normalized average rank for the data with new items and users.

challenging scenario and even improves performance. Similar behavior is shown in Figure 3b where short-term model, augmented by an average model, eliminates the shortcomings of the short-term model in dealing with new items. By contrast, collaborative filtering fails to catch up and performs poorly in both scenarios. As a result, maintaining additional average models is an effective and efficient means in cold start situations. The experiments however also show that there is no one model that rules them all; instead, the model of choice depends clearly on the intrinsic dynamics of the applications.

6 Conclusion

In this paper, we presented a unified model for short-term and long-term recommendation in a multi-armed bandit framework. The model incorporated the information from the actual context as well as the long-term preferences of the users into a single contextual bandit. We transformed the optimization problem of our bandits into the dual space considering a linear payoff model for the arms.

Addressing the problem in dual space led to a generalized optimization problem where arbitrary loss functions could be used to reshape the payoff function according to the application at-hand. As a result, applying contextual bandits for long- and short-term recommendations is considerably simplified. The experiments show that adding an average model to short- and long-term models leads to robust methods that clearly outperform their vanilla peers in terms of normalized average rank.

References

1. P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2003.
2. N. Barbieri, G. Manco, E. Ritacco, M. Carnuccio, and A. Bevacqua. Probabilistic topic models for sequence data. *Machine Learning*, 93(1):5–29, 2013.

3. S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.
4. A. A. Deshmukh, U. Dogan, and C. Scott. Multi-task learning for contextual bandits. *arXiv preprint arXiv:1705.08618*, 2017.
5. R. Dybowski and S. Roberts. Confidence intervals and prediction intervals for feed-forward neural networks. *Clinical applications of artificial neural networks*, pages 298–326, 2001.
6. Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 263–272. IEEE, 2008.
7. S. S. Keerthi, K. B. Duan, S. K. Shevade, and A. N. Poo. A fast dual algorithm for kernel logistic regression. *Machine Learning*, 61(1-3):151–165, 2005.
8. Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 8:30–37, 2009.
9. L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the International World Wide Web Conference*, 2010.
10. D. K. Mahajan, R. Rastogi, C. Tiwari, and A. Mitra. LogUCB: an explore-exploit algorithm for comments recommendation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 6–15. ACM, 2012.
11. R. M. Rifkin and R. A. Lippert. Value regularization and fenchel duality. *Journal of Machine Learning Research*, 8:441–479, 2007.
12. L. Tang, Y. Jiang, L. Li, C. Zeng, and T. Li. Personalized recommendation via parameter-free contextual bandits. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 323–332. ACM, 2015.
13. M. Tavakol and U. Brefeld. Factored mdps for detecting topics of user sessions. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 33–40. ACM, 2014.
14. M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*, 2013.
15. C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 448–456. ACM, 2011.
16. Q. Wu, H. Wang, Q. Gu, and H. Wang. Contextual bandits in a collaborative environment. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 529–538. ACM, 2016.
17. L. Zhou and E. Brunskill. Latent contextual bandits and their application to personalized recommendations for new users. *arXiv preprint arXiv:1604.06743*, 2016.