Multi-channel Integrated Recommendation with Exposure Constraints

Yue Xu
Alibaba Group.
Hangzhou, China
yuexu.xy@foxmail.com

Qijie Shen
Alibaba Group.
Hangzhou, China
qijie.sqj@alibaba-inc.com

Jianwen Yin
Alibaba Group.
Hangzhou, China
yjw264077@alibaba-inc.com

Zengde Deng
Cainiao Network.
Hangzhou, China
dengzengde@gmail.com

Dimin Wang
Alibaba Group.
Hangzhou, China
dimin.wdm@alibaba-inc.com

Hao Chen
The Hong Kong Polytechnic University.
Hong Kong, China
sundaychenhao@gmail.com

Lixiang Lai
Alibaba Group.
Hangzhou, China
lixiang.llx@alibaba-inc.com

Tao Zhuang
Alibaba Group.
Hangzhou, China
zhuangtao.zt@alibaba-inc.com

Junfeng Ge
Alibaba Group.
Hangzhou, China
beili.gjf@alibaba-inc.com

ABSTRACT

Integrated recommendation, which aims at jointly recommending heterogeneous items from different channels in a main feed, has been widely applied to various online platforms. Though attractive, integrated recommendation requires the ranking methods to migrate from conventional user-item models to the new user-channel-item paradigm in order to better capture users' preferences on both item and channel levels. Moreover, practical feed recommendation systems usually impose exposure constraints on different channels to ensure user experience. This leads to greater difficulty in the joint ranking of heterogeneous items. In this paper, we investigate the integrated recommendation task with exposure constraints in practical recommender systems. Our contribution is forth-fold. First, we formulate this task as a binary online linear programming problem and propose a two-layer framework named Multi-channel Integrated Recommendation with Exposure Constraints (MIREC) to obtain the optimal solution. Second, we propose an efficient online allocation algorithm to determine the optimal exposure assignment of different channels from a global view of all user requests over the entire time horizon. We prove that this algorithm reaches the optimal point under a regret bound of $O(\sqrt{T})$ with linear complexity. Third, we propose a series of collaborative models to determine the optimal layout of heterogeneous items at each user request. The joint modeling of user interests, cross-channel correlation, and page context in our models aligns more with the browsing nature of feed products than existing models. Finally, we conduct extensive experiments on both offline datasets and online A/B tests to verify the effectiveness of MIREC. The proposed framework has now been implemented on the homepage of Taobao to serve the main traffic.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Integrated Recommendation, Feed Recommendation, Multi-channel, Exposure constraint, Context-Aware

ACM Reference Format:
Yue Xu, Qijie Shen, Jianwen Yin, Zengde Deng, Dimin Wang, Hao Chen, Lixiang Lai, Tao Zhuang, Junfeng Ge. 2023. Multi-channel Integrated Recommendation with Exposure Constraints. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599868

1 INTRODUCTION

Nowadays, the ever-expanding new breeds of content, e.g., pictures, live streams, and short videos, drive the recommender system to drift from the traditional homogeneous form into an integrated form. Integrated recommender systems (IRSs) aim to simultaneously recommend heterogeneous items from multiple channels in a row. This integrated form greatly expands users' choices on different types of content thereby satisfying users' diversified preferences on both item-level and channel-level. Therefore, IRS has nowadays been widely deployed in various online platforms such as the homepage feeds in Kuaishou [27], XiaohongShu [20], Taobao [33], and AliExpress [16]. In these products, users continuously slide down to browse and interact with heterogeneous items in a sequential manner, as shown in Figure 1.

Though attractive, integrated feed recommendation faces more challenges than conventional recommendation with homogeneous items. First, real-world applications usually impose upper or lower exposure guarantees on different channels, such as lower constraints for sponsored/new content (e.g., ads and cold-start items) or upper constraints for individual channels to ensure diversity. These constraints lead to greater difficulty in the joint ranking of heterogeneous items. Second, heterogeneous items from multiple channels usually have different features and ranking strategies. Hence, it

Yue Xu and Qijie Shen contributed equally to this work. Yue Xu is the corresponding author.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599868

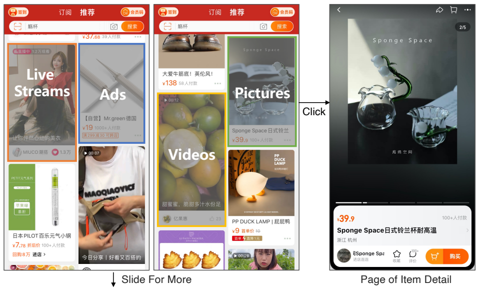


Figure 1: A snapshot of the IRS in real-world feed products. Left: the IRS presents heterogeneous items provided by multiple channels in a row, users slide down to view more items. Right: the detail page is presented after a user clicks an item.

is difficult to directly compare items from different channels for joint ranking. Third, users' interests on different channels have a great impact on their behaviors, such that traditional user-item prediction models need to evolve into user-channel-item prediction models by considering both intra-channel and inter-channel information and their correlation with user interests. Finally, in feed products, users tend to review a large number of items in a row such that the previously viewed items have a great impact on the users' behavior towards the next item [20, 27, 33]. Therefore, it is of vital importance to consider the influence from page context when determining the item order in the return list.

Although integrated feed recommendation has been widely deployed in practice, there are still few works focusing on the above challenges systematically. In this paper, we propose a general framework named Multi-channel Integrated Recommendation with Exposure Constraints (MIREC) to deal with the multi-channel integrated recommendation task under resource constraints in feed products. MIREC consists of two layers: an allocation-layer which optimizes the exposure of different channels from a global view over all user requests, and a ranking-layer which determines the optimal item layout of a given user request from a local view. These two layers operate in an iterative manner to make online decisions along with the arrival of user requests. The main contributions are as follows.

- This work formulates the integrated recommendation task with exposure constraints as a binary online linear programming problem and proposes a two-layer framework named MIREC to obtain the optimal solution. We also describe a practical system architecture for its implementation in industrial platforms.
- This work proposes an efficient multi-channel allocation algorithm to obtain the optimal exposure assignment of different channels over a fixed time horizon. The proposed algorithm is able to reach an optimal solution with linear complexity w.r.t. the number of constraints. We also prove that this algorithm admits a regret bound of $O(\sqrt{T})$ towards the global optimal point under certain assumptions.
- This work proposes a series of collaborative models to determine the optimal layout of heterogeneous items on a page, with joint modeling of user interests, cross-channel correlation, and page context. This aligns more with the browsing nature of feed products than existing models.
- This work conducts extensive experiments on both offline datasets and online A/B tests to verify the superiority of our proposed method.

MIREC has been implemented on the homepage of Taobao to serve the main traffic. It brings 3% lift on user clicks, 1.56% lift on purchase, and 1.42% lift on stay time. It now serves hundreds of millions of users towards billions of items every day.

2 RELATED WORK

Re-ranking Methods. The main objective of re-ranking methods is to consider the mutual influence among a list of items in order to refine the prediction results produced by point-wise ranking models. Three prevalent models are commonly adopted in the existing literature: RNN-based methods, attention-based methods, and evaluator-generator-based methods. The first two methods feed an initial ranking list produced by point-wise models (e.g., Wide&Deep [13], DIN [45] and SIM [32]) into RNN-based (e.g., MiDNN [47], Seq2Slate [4] and DLCM [1]) or attention-based structure (e.g., PRM [31], PFRN [19], PEAR [24], and Raiss [26]) sequentially and output the encoded vector at each subsequent layer to model the mutual influences among items. The evaluator-generator-based methods (e.g., SEG [36] and GRN [14]), use a generator to generate feasible permutations and use an evaluator to evaluate their list-wise utility to determine the optimal permutation. However, most re-ranking methods mainly focus on capturing the mutual influence among homogeneous items provided by one channel, instead of heterogeneous items provided by multiple channels. Moreover, they only optimize the item order at a single time slot, instead of considering a cumulative utility over a broad time horizon under resource constraints.

Online Allocation Methods. The online allocation problem with resource constraints has been mostly studied in online convex optimization [18]. The primal-dual methods [3, 12, 29, 38, 39] avoid taking expensive projection iterations by penalizing the violation of constraints through duality. The BwK methods [2, 22] determines an optimal action from a *finite* set of possible actions and then optimize the policy of decision-making according to the observed rewards and costs over a fixed period of time. Several recent works studied the practical performance of online allocation in advertising recommendations, including PDOA [46], MSBCB [17] and HCA2E [6]. Most related works on online convex optimization focus on theoretical analysis (e.g., regret bound) instead of real-world applications. Other related works on advertising mainly consider binary content, i.e., ads or non-ads, instead of heterogeneous content. Directly extending them to deal with multi-channel recommendations in IRSs may lead to sub-optimal results.

Integrated Recommendation. The integrated recommendation is a newly emerged but rapidly developing domain driven by practical problems [28]. Integrated recommendation methods need to consider both intra-channel and inter-channel features within the heterogeneous content and provide recommendation results continuously along with user arrivals. Recently, DHANR [16] proposed a hierarchical self-attention structure to consider the cross-channel interactions. HRL-Rec [37] decomposed the integrated re-ranking problem into two subtasks: source selection and item ranking, and use hierarchical reinforcement learning to solve the problem.

DEAR [43] proposed to interpolate ads and organic items by deep Q-networks. Cross-DQN [25] also adopt a reinforcement learning solution with a cross-channel attention unit. However, many integrated methods only focus on ranking at a single time slot instead of over a continuous time horizon. The joint consideration of both integrated ranking and online allocation of limited resources still remains to be explored.

3 PROBLEM FORMULATION

This section formulates the integrated recommendation task with exposure constraints as a binary online linear programming problem. Specifically, we consider a generic IRS setting where user requests arrive sequentially during a finite time horizon. For each request, the IRS needs to rank a list of heterogeneous items provided by multiple channels. The aim is to maximize the overall utilities (e.g., the sum of clicks and pays) of *all channels over the entire time horizon*, subject to multiple resource constraints.

Formally, the request of user u triggered at time t is described as $e_t = (u, f, g, \mathcal{X}_t)$, where $f \in \mathbb{R}_+$ is a non-negative *utility function*, $g \in \mathbb{R}_+$ is a non-negative resource *consumption function*, and $\mathcal{X}_t \subset \mathbb{R}^d_+$ is a compact set denoting all possible item layouts for decision-making. For each request e_t, the IRS needs to choose a number of N heterogeneous items from a candidate set I_t and place them into N slots to form a complete page and return it to the user. This action $\boldsymbol{x}_t \in \mathcal{X}_t$ can be represented as a decision matrix $\boldsymbol{x}_t \in [0,1]^{N\times|I_t|}$, where each entry $x_{t,n,i}$ is indexed by a slot n and a card index i. Once the user finished viewing the current page, a new user request will be triggered to ask the platform to return to the next page. Therefore, this decision-making process will be performed repeatedly. Moreover, in real-world applications, the item layouts need to satisfy the following constraints:

$$\mathcal{X}_t = \begin{cases} \sum_{i\in I^t} x_{t,n,i} = 1, & \forall t \in \mathcal{T}, \forall n \in \mathcal{N} \\ \sum_n x_{t,n,i} \le 1, & \forall t \in \mathcal{T}, \forall i \in I, \end{cases} \tag{1}$$

where $\mathcal{T}$ represents time horizons ranging from 1 to T, and $\mathcal{N}$ represents the slots. The upper constraint restricts that each slot must be assigned to one item and the lower constraint restricts that each item can be assigned to at most one slot.

After executing an action $\boldsymbol{x}_t$ at request e_t, the IRS consumes a resource cost $g(\boldsymbol{x}_t)$ and obtains an utility $f(\boldsymbol{x}_t)$. In IRS, the utility function $f(\boldsymbol{x}_t)$ is defined according to the concerned metrics. For example, it can be defined as a combination of stay time, adds to cart, and favorites to encourage user engagement, or defined as a combination of clicks and purchases to encourage user conversion. On the other hand, the consumption function $g(\boldsymbol{x}_t)$ is defined based on the concerned resource constraints. For example, the platform may need to allocate a certain amount of exposure to new channels in order to support the growth of new content [15]. Meanwhile, a too large proportion of exposures on one specific channel will damage the recommendation diversity thereby harming user experience [6, 30]. In this case, the IRS needs to guarantee both a lower exposure limit and an upper exposure limit for the heterogeneous items from different channels.

In this paper, we focus on the exposure constraints in practical systems which lead to the following optimization problem

$$\mathcal{P}_0: \quad \mathrm{OPT}(\mathcal{S}) = \max_{\boldsymbol{x}_t\in\mathcal{X}_t} \sum\nolimits_{t=1}^{T} f(\boldsymbol{x}_t) \tag{2}$$

$$\text{s.t.} \quad C_1: \sum\nolimits_{t=1}^{T} g_m(\boldsymbol{x}_t) \le G^{\max}_{m,th} N(\mathcal{S}), \forall m \in \mathcal{M}, \tag{3}$$

$$C_2: \sum\nolimits_{t=1}^{T} g_m(\boldsymbol{x}_t) \ge G^{\min}_{m,th} N(\mathcal{S}), \forall m \in \mathcal{M}, \tag{4}$$

where $N(\mathcal{S})$ denotes the total available exposures to allocate over the entire time horizon, $G^{\max}_{m,th}$ and $G^{\min}_{m,th}$ denote the proportion of upper exposure limits and lower exposure limits for each channel $m \in \mathcal{M}$, respectively, and $g_m(\boldsymbol{x}_t)$ denotes the consumed exposures of cards from channel m after executing $\boldsymbol{x}_t$ at request e_t. Although this paper mainly focuses on the exposure guarantee, the above formulation is generally applicable to other problems with different resource constraints, e.g., the number of coupons to allocate.

4 METHODOLOGY

4.1 Framework Overview

Directly solving problem $\mathcal{P}_0$ is challenging. On one hand, the estimation accuracy of utility $f(\boldsymbol{x}_t)$ and consumption $g(\boldsymbol{x}_t)$ suffer influence from multi-factors, including the user's personal interest, the page context, and the cross-correlation between different channels. On the other hand, the determination of each $\boldsymbol{x}_t$ needs to consider the cumulative exposures over the entire time horizon due to the exposure guarantees. Therefore, the optimization of exposure allocation must be performed from a global view over the entire timeline instead of a single time slot.

To this end, we propose the MIREC framework which solves $\mathcal{P}_0$ through online primal-dual iterations. Specifically, MIREC consists of two layers, i.e., the allocation-layer and the ranking-layer, which correspond to the dual and primal problem of $\mathcal{P}_0$, respectively. The *allocation-layer* optimizes dual variables to control the cumulative exposure of different channels on all user requests from a global view to guarantee the exposure limits. Meanwhile, the *ranking-layer* optimizes the item layout under fixed dual variables given by the allocation-layer, with the aim to maximize the instant utility on a single user request from a local view. These two layers operate in an iterative manner along with the arrival of online user requests to determine the optimal item layout at user requests continuously. The general workflow of MIREC is shown in Figure. 2.

For the allocation-layer, we propose a simple but efficient Mirror-descent based Multi-channel Exposure Allocation (ME2A) algorithm to adaptively balance the utility gain and the exposure cost of presenting heterogeneous items from different channels. The proposed M2EA algorithm has a closed form solution that can be computed in linear time and admits a regret bound of $O(\sqrt{T})$ towards the global optimal point under certain assumptions.

For the ranking-layer, we propose a personalized cross-channel ranking (PCR) model and a context-aware reranking (CAR) model to jointly determine the optimal item layout on a given user request, with fixed dual parameters given by the allocation-layer. In particular, PCR gives point-wise estimation of the quality of candidate items by joint modeling the influence from user interests, intra-channel information, and inter-channel correlations. Afterward, CAR refines the point-wise estimation generated by PCR

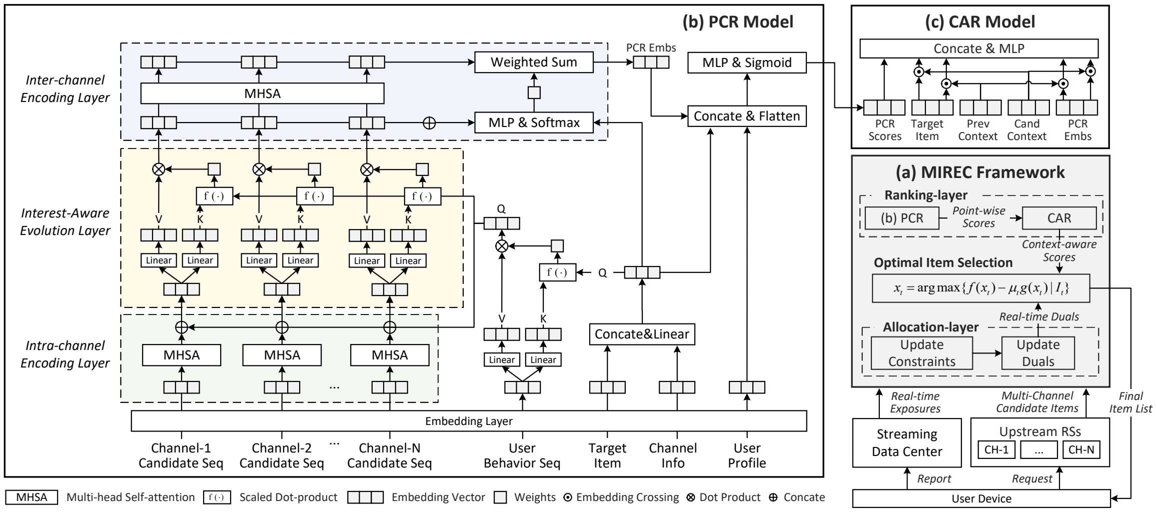


Figure 2: An overview of the MIREC framework.

into context-aware estimation by making use of both the context information and the high-level knowledge extracted from PCR.

4.2 Global: Online Exposure Allocation

In this section, we introduce the primal-dual formulation of $\mathcal{P}_0$ and propose the ME2A algorithm to obtain the optimal solution for online systems. The complete algorithm is presented in Algorithm 1.

4.2.1 Primal-dual Formulation. The Lagrangian dual function of problem $\mathcal{P}_0$ can be written as

$$\min_{\boldsymbol{\mu},\boldsymbol{\lambda}} D(\boldsymbol{\mu},\boldsymbol{\lambda}) = \sum_t f(\boldsymbol{x}_t) - \sum_m \mu_m \Big(\sum_t g_m(\boldsymbol{x}_t) - G_m^{\max}\Big) + \sum_m \lambda_m \Big(\sum_t g_m(\boldsymbol{x}_t) - G_m^{\min}\Big). \tag{5}$$

Here $\boldsymbol{\mu} \geq 0$ and $\boldsymbol{\lambda} \geq 0$ are the introduced dual parameters, $G_m^{\max}$ and $G_m^{\min}$ are short for $G_{m,th}^{\max} N(\mathcal{S})$ and $G_{m,th}^{\min} N(\mathcal{S})$, respectively. Note that the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ are related to the violation of exposure consumption over the upper bound limit and the lower bound limit, respectively, which are mutually exclusive. In particular, if one of them is positive, the other one must be zero; otherwise, both of them are zero. Hence, it is viable to only introduce one *real number* dual variable $\boldsymbol{\mu} \in \mathbb{R}^{1\times M}$ to replace $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ in the dual problem, which simplifies (5) into

$$\min_{\boldsymbol{\mu}} D(\boldsymbol{\mu}) = \sum_t f(\boldsymbol{x}_t) - \sum_m [\mu_m]_+ \Big(\sum_t g_m(\boldsymbol{x}_t) - G_m^{\max}\Big) \tag{6a}$$

$$+ \sum_m [-\mu_m]_+ \Big(\sum_t g_m(\boldsymbol{x}_t) - G_m^{\min}\Big) \tag{6b}$$

$$= \sum_t \Big(f(\boldsymbol{x}_t) - \sum_m \mu_m g_m(\boldsymbol{x}_t)\Big) + \sum_m \Big([\mu_m]_+ G_m^{\max} - [-\mu_m]_+ G_m^{\min}\Big), \tag{6c}$$

where $[\mu_m]_+ = \max\{\mu_m, 0\}$.

4.2.2 Dual Optimization for Exposure Constraints. The dual problem in (6) can be solved optimally via primal-dual updates. Specifically, given a user request $e_t = (u, f, b, \mathcal{X}_t)$, we assume that the utility $f(\boldsymbol{x}_t)$ and cost $\boldsymbol{\mu}_t^T g(\boldsymbol{x}_t)$ under different item layout $\boldsymbol{x}_t$ can be properly estimated by the models at the ranking layer. As such, the optimal item layout $\boldsymbol{x}_t$ under a fixed dual variable $\boldsymbol{\mu}_t$ can be obtained by solving the following primal problem:

$$\mathcal{P}_1 : \tilde{\boldsymbol{x}}_t = \arg\max_{\boldsymbol{x}_t \in \mathcal{X}} \Big\{ f(\boldsymbol{x}_t) - \boldsymbol{\mu}_t^T g(\boldsymbol{x}_t) \Big\}. \tag{7}$$

This is the focus of the ranking-layer to be discussed layer. After the optimal item layout $\boldsymbol{x}_t$ at user request e_t is properly determined, the next step is to update the dual variable $\boldsymbol{\mu}_t$ to adjust the exposure of different channels in future user requests. Specifically, the remained exposure resource of different channels after the presentation of $\boldsymbol{x}_t$ at e_t is updated by

$$G_{m,t+1}^{\max} = G_{m,t}^{\max} - g_m(\boldsymbol{x}_t), \forall m \in \mathcal{M}. \tag{8}$$

The sub-gradient of the dual function in (6) can be obtained via Danskin's theorem [5] by

$$\nabla \mu_{m,t} = -g_m(\boldsymbol{x}_t) + G_{m,t+1}^{\max} \cdot \mathbb{1}(\mu_{m,t} \geq 0) + G_m^{\min} \cdot \mathbb{1}(\mu_{m,t} \leq 0), \tag{9}$$

where $\mathbb{1}(x \in A)$ is an indicator function which equals to one if $x \in A$ otherwise zero. As such, the dual variable $\boldsymbol{\mu}_t$ can be updated based on the mirror-descent method as

$$\mu_{m,t+1} = \arg\min_{\mu_m \in \mathbb{R}} \mu_m \nabla \mu_{m,t} + \frac{1}{\eta} V_h(\mu_m, \mu_{m,t}), \tag{10}$$

where $V_h(x,y) = h(x)h(y)\nabla h(y)^T(x-y)$ is the Bregman divergence based on reference function $h(\cdot)$ and $\eta \in \mathbb{R}$ is a fixed step-size. Note that this mirror descent step can be computed in linear time since (10) admits a closed-form solution. For example, if we use $h(\boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{\mu}\|^2$ as the reference function, the dual update in (10) becomes

$$\boldsymbol{\mu}_{t+1} = [\boldsymbol{\mu}_t - \eta \nabla \boldsymbol{\mu}_t]_+, \tag{11}$$

which recovers the online projected gradient descent method. Moreover, in order to guarantee the upper exposure constraints, one needs to examine the violation of upper limits of each channel before the determination of $\boldsymbol{x}_t$ at each user request. If the sum of exposures of a specific channel exceeds its upper bound, one needs to remove all candidate items from this channel to forbid allocate

Algorithm 1 The proposed ME2A algorithm of MIREC

```
Initialization:
Initial dual solution $\mu_1$, total time periods $T$, reference function $h(\cdot)$ and step-size $\eta$.
Iteration:
for $t = 1, 2, \cdots, T$ do
    Receive request $e_t = (u, f, b, \mathcal{X}_t)$.
    Update the candidate set $I_t$ provided by multi-channels.
    Determine the optimal item list $\boldsymbol{x}_t$ by solving the primal problem in (7) at the ranking-layer.
    Update the remained exposure resource via (8).
    Obtain the sub-gradient of the dual variable via (9).
    Update the dual variable based on mirror descent via (11).
end for
```

more exposures when determining $\boldsymbol{x}_t$. We present the optimality of this proposed ME2A algorithm and its feasibility to guarantee exposure constraints of different channels as follows. Detailed proofs are deferred to the appendix.

4.2.3 Optimality. It is viable to prove that Algorithm 1 is asymptotically optimal and admits a regret bound scales as $O(\sqrt{T})$ when the user requests arrive from an *i.i.d* unknown distribution. This assumption is reasonable when the number of requests is numerous [3, 29, 46]. Specifically, we denote Algorithm 1 as π and the overall utility over all user requests in set $\mathcal{S}$ under the running of π as $R(\pi|\mathcal{S}) = \sum_{t=1}^{T} f(\boldsymbol{x}_t)$. The regret of model π is defined as the worst-case difference over $\mathcal{S}$ between the expected performance of the global optimal solution and the model π:

$$\text{Regret}(\pi|\mathcal{S}) = \sup\Big\{\mathbb{E}_{\mathcal{S}}[\text{OPT}(\mathcal{S}) - R(\pi|\mathcal{S})]\Big\}, \tag{12}$$

where $\text{OPT}(\mathcal{S})$ denotes the optimal utilities one can obtain under the request set $\mathcal{S}$. The regret bound can be given as follows.

Theorem 1. *Suppose that the requests come from an i.i.d model with unknown distribution. Then, $Regret(\pi|\mathcal{S}) \leq C_1 + C_2\eta T + \frac{C_3}{\eta}$ with $\eta > 0$ holds for any $T \geq 1$. Here C_1, C_2 and C_3 are constant values depending on the numerical bounds of the utility f, the consumption g, and terms from the dual iterates in Eq. (10).*

From Theorem 1, we obtain $\text{Regret}(\pi|\mathcal{S}) \leq O(\sqrt{T})$ when using a step-size $\eta \propto c/\sqrt{T}$ with any constant $c > 0$. We defer the proof and detailed definitions of C_1, C_2, and C_3 into the appendix.

4.2.4 Exposure Feasibility. In Algorithm 1, if the upper exposure limit of a specific channel is violated, we will forbid the exposure of any item from this channel when determining the item list $\boldsymbol{x}_t$. Therefore, the exposure can never be overspent. On the other hand, the lower exposure limits are soft-restricted by adaptively adjusting the dual variable $\boldsymbol{\mu}$. This may cause exposure underspent. However, it is viable to prove that the violation of the lower exposure limit of any channel also admits a convergence rate of $O(\sqrt{T})$. In other words, even if the violations on lower exposure limits may occur, their growth is considerably smaller than T.

Proposition 1. *Suppose the requests come from an i.i.d model with unknown distribution. Then, it holds for any $T \geq 1$ and any channel $m \in \mathcal{M}$ that $G_m^{\min} - \mathbb{E}\Big[\sum_{t=1}^{T} g_m(\boldsymbol{x}_t)\Big] \leq C_4 + \frac{C_5}{\eta}$, where C_4 and C_5 are constant values depends on the numerical bounds of utility f, consumption g, and terms from the dual iterates* (10).

Proposition 1 states that when using $\eta \propto c/\sqrt{T}$ with $c > 0$, the exposure underspend of any channel is bounded by $O(\sqrt{T})$. We defer the proof and definitions of C_4 and C_5 into the appendix.

4.3 Local: Context-Aware Integrated Ranking

Different from the allocation-layer which optimizes an objective with accumulative utilities over the entire time horizon as defined in (6), the ranking-layer focus on maximizing the utilities on a single time slot. This corresponds to the primal problem given in (7): $\mathcal{P}_1 : \tilde{\boldsymbol{x}}_t = \arg\max_{\boldsymbol{x}_t \in \mathcal{X}} \Big\{f(\boldsymbol{x}_t) - \boldsymbol{\mu}_t^T g(\boldsymbol{x}_t)\Big\}$. In other words, the allocation layer adjusts the exposure of items from different channels by optimizing the dual parameter $\boldsymbol{\mu}_t$ from a global view of all user requests. While the ranking-layer determines the optimal item list $\boldsymbol{x}_t$ under a fixed dual variable $\boldsymbol{\mu}_t$ from a local view of a given user request e_t.

There are two common characteristics that are strongly related to the estimation of $f(\boldsymbol{x}_t)$ and $g(\boldsymbol{x}_t)$ in the integrated recommendation. First, users' preference on different channels has a great impact on the utility (e.g., prefer to click or not) and exposure (e.g., prefer to view or not) estimations. Therefore, it is of vital importance to consider both intra-channel and inter-channel correlations with reference to user interests during the estimation. Second, in feed products, users tend to review a large number of items in a row such that the previously viewed items have a great impact on users' behavior towards the next item. Therefore, it is necessary to consider page context when determining the item order.

Therefore, we propose two models to deal with the above two challenges, respectively. First, we propose PCR model to deal with the joint modeling of user interests and inter/intra-channel correlation. It gives a point-wise estimation of the utility/exposure value of presenting each candidate item. Second, we propose CAR model to refine the point-wise estimation from PCR into context-aware estimation by considering both context information and the high-level knowledge obtained from PCR. It is also responsible for selecting optimal items from a set of candidate items to generate the final return list. In real-world systems, for each user request, we only need to run PCR once to get the point-wise scores, and then run CAR multiple times to generate the return list. Next, we mainly focus on the estimation of $f(x)$, the estimation of $g(x)$ can be performed in a similar way by changing the learning goals.

4.3.1 Personalized Cross-Channel Ranking Model. PCR takes four types of features as input, i.e., the user profile feature X_u, the user behavior sequences X_b, the candidate items provided by each channel X_l, and the item-level features of target item X_i. As shown in Figure 2, We use an embedding layer to transform these features into dense embedding vectors, denoted as E_u, E_b, E_l and E_i, respectively. These embedding vectors are then fed into three components, i.e., the intra-channel encoding layer, interest-aware evolution layer, and inter-channel encoding layer in order, which are described below.

Intra-Channel Encoding layer. This layer aims at extracting the mutual influence of item pairs and other extra information within the channel. We adopt the well-known multi-head attention [35] as the basic learning unit for intra-channel encoding. This is due to that the self-attention mechanism is able to directly capture the mutual influences between any two items, and is robust to far distance within the sequence. Formally, the formulation of this attention-based encoding can be written as

$$V_l^m = [head_1, head_2, ..., head_h]W^O, \tag{13a}$$

$$head_i = \text{Softmax}\left(\frac{(E_b W_Q)(E_b W_K)^T}{\sqrt{d_h/h}}\right)(E_b W_V), \tag{13b}$$

where $W^O \in \mathbb{R}^{d_h \times d_h}$ denotes the learnable parameters for each head with d_h being the length of projected embedding vector after attention, $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_h/h}$ are the vectors of query, key and value with d being the length of original embedding vector and h being the number of heads, V_l^m represents the encoded candidate items of each channel $m \in \mathcal{M}$.

Interest-Aware Evolution Layer. User interest modeling based on user-item relationship attracts much research attention in recommendation [7–9, 21, 40, 42]. Existing works such as PRM [31] and DHANR [16] directly apply the self-attention mechanism to model the inter-dependencies among items and channels without considering user's recent behavior. However, the interests hidden in user's behavior items usually have a great impact on the prediction accuracy in recommendation tasks [10, 32, 44, 45]. The recently proposed PEAR [24] firstly models the dependency between the candidate item list and the user's historical behaviors based on a transformer-like structure, which, however, suffers from two limitations. First, directly mixing the raw item-level features from user behavior items may introduce redundant or noisy information to degrade the learning performance. Second, each user may exhibit multiple interest points, such that it is beneficial to reinforce the interest related to the target item before feature-crossing to avoid drifting. Therefore, we first reinforce the interest vector according to the correlation between behavior items and the target item as

$$V_U = f(E_b; E_i) = \sum_{i=1}^{B} A(b_i, E_i) b_i = \sum_{i=1}^{B} w_i b_i, \tag{14}$$

where B is the length of behavior sequence, b_i is the behavior item, V_U denotes the user representation feature with respect to E_i, and $A(\cdot)$ is a feed-forward network whose output is the activation weight w_i. Then, we make use of this reinforced interest vector to extract useful information from the candidate items of different channels. Formally, for each channel m, given V_l^m and V_U as inputs, we use scaled dot-product attention formulated as follows:

$$H_s^i = \text{Softmax}\left(\frac{(V_l^m W_q)[V_U W_{k1}, V_l^m W_{k2}]^T}{\sqrt{d_h}}\right)[V_U W_{v1}, V_l^m W_{v2}], \tag{15}$$

where $W_{k1}, W_{v1} \in \mathbb{R}^{d \times d_h}$ and $W_{k2}, W_q, W_{v2} \in \mathbb{R}^{d_h \times d_h}$ are all learnable parameters, $[\cdot]$ denotes the concatenation operation. After the above operations, we successfully merged the information from the candidate item lists and the user's historical behaviors into a series of evolved embedding vectors for further processing.

Inter-Channel Encoding Layer. Previous layers mainly extract intra-channel correlations. We now focus on the modeling of inter-channel correlation. First, we feed the embedding vector of each channel and the target item embedding into the MLP layer with softmax function to obtain the importance weights on each channel that is related to the target item: $W_{CH} = \text{Softmax}(\text{MLP}[H_s^m, E_i])$, where $W_{CH} \in \mathbb{R}^{1 \times m}$ is the importance weights, and H_s^m denotes the concatenation of all channels' output from the Interest-Aware Evolution Layer. Second, we perform multi-head self-attention on the evolved embedding vector H_s^m of each channel $m \in \mathcal{M}$ to obtain the mixed embedding $\tilde{H}_s^m$ which contains rich inter-channel information. Then, we perform the weighted sum on the mixed embedding of all channels based on W_{CH} to get the final representation of multi-channel modeling: $V_L = W_{CH} \cdot [\tilde{H}_s^m]^T, m \in \mathcal{M}$, where $[\tilde{H}_s^m]$ represents the concatenation of the mixed embeddings of all channels.

Finally, we concatenate all vectors as input and feed it into the MLP layers with a sigmoid function to predict the utility of presenting a given target item to a given target user as $Y_{PCR} = \text{Sigmoid}(\text{Concat}(E_u, E_i, V_L))$.

4.3.2 Context-Aware Refinement Model.

In this section, we propose the CAR model to refine the point-wise utility scores estimated by PCR into context-aware utility scores. Given a candidate item set I_{cand} with size N, the aim of CAR is to optimally choose K items from I_{cand} and allocate them to the K slots in a page based on the learning results from PCR.

We maintain two types of context information in CAR, i.e., the context of previous items and the context of remaining candidate items. Specifically, when selecting the k-th item in a page, we represent the context of previously presented $k-1$ items h_{pre} by mean-pooling over their embeddings. Meanwhile, we represent the context of all candidate items h_{can} by mean-pooling over the embeddings of all remained candidates. These two context vectors are updated and repeated along with the item selection process. Furthermore, we perform a series of embedding crossing operations between the target item embedding e_i and the context embeddings to model the influence from page context. In specific, the operations in the context of previous items can be formulated as follows:

$$H_{pt} = \text{Concat}(h_{pre} \oplus e_i, h_{pre} \otimes e_i, h_{pre} \ominus e_i), \tag{16}$$

where $\oplus$, $\otimes$, and $\ominus$ denote the addition, subtraction, and dot product between embedding vectors, respectively. The same goes for H_{ct} by replacing h_{pre} in (16) with the context of candidate items h_{can}. Additionally, we also perform embedding-crossing between the context embeddings and the high-level knowledge V_L from PCR to obtain another two vectors, i.e., H_{ct}^v and H_{pre}^v.

Finally, for each candidate item $i \in I_{cand}$, we predict the context-aware utility score by feeding these embedding vectors along with the point-wise utility score Y_{PCR} from PCR and user profile features E_u into one MLP layer as

$$H_{\text{all}} = \text{Concat}(H_{pt}, H_{ct}, H_{ct}^v, H_{pre}^v, h_{pre}, h_{can}, E_u, Y_{PCR}), \tag{17a}$$

$$Y_{CAR} = \sigma(\text{MLP}(H_{\text{all}})), \tag{17b}$$

where σ represents the sigmoid activation function. After scoring all candidate items, we choose the item with the highest score as the optimal item for slot k, and update the context vectors and the remained candidate items accordingly. This process will be repeated K times to generate a return item list of length K. Note that the

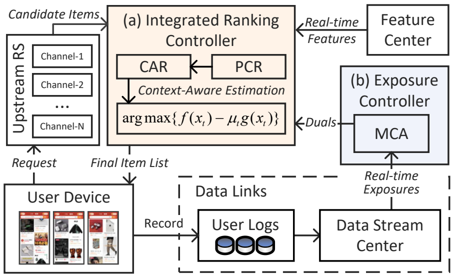


Figure 3: Online system architecture.

above operations in CAR only involve linear computations, such that this item selection process is cost-efficient in online systems.

Both PCR and CAR can be trained with the commonly used cross-entropy loss as in other ranking models, the learning objective can be given as $J = \sum_{e_t \in \mathcal{D}} \left(y^{e_t}_{u,i} \log \hat{y}^{e_t}_{u,i} + (1 - y^{e_t}_{u,i}) \log(1 - \hat{y}^{e_t}_{u,i}) \right)$, where $\mathcal{D}$ denotes the training dataset, $y^{e_t}_{u,i}$ is the real user-item recommendation label (equals 1 or 0) between user u and item i at request e_t, and $\hat{y}^{e_t}_{u,i}$ is the predicted label given by Y_{PCR} or Y_{CAR}. In our experiments, when predicting the utility function f with user clicks, $y^{e_t}_{u,i}$ refers to the click label; when predicting the cost function g with exposure constraints, $y^{e_t}_{u,i}$ refers to the exposure label between user u and item i at request e_t, i.e., whether user u has seen item i at request e_t. One can readily change the learning objective according to actual demands.

4.4 Online Implementation

In this section, we introduce the online implementation of our proposed MIREC model in the homepage feed of Taobao. It now serves the main traffic of Taobao to provide services to hundreds of millions of users towards billions of items in Taobao every day.

Figure. 3 gives a general architecture to implement our proposed MIREC model in real-world IRS. Each time a user request is triggered from the device, the upstream RS of each channel will run its own recommendation models to determine the top items to return. The Integrated Recommendation Controller uses the top items from all channels as the candidates. It retrieves user/item features from a feature center in real-time and ranks candidate items by solving based on our proposed CAR and PCR model. Meanwhile, the dual variable μ is estimated by an Exposure Controller. This module monitors the completeness of exposure guarantees based on the real-time exposures collected from user logs and updates the dual variable to adjust the exposures on different channels periodically.

For online complexity, the Exposure Controller utilizes a projected gradient descent algorithm requiring only linear computation with little overhead. CAR is a linear model requiring little overhead, while PCR is a point-wise ranking model involving attention-based computation that entails greater cost. However, the system only needs to perform inference on PCR once per user request, resulting in computational complexity comparable to that of common point-wise ranking models. In fact, the presented system architecture is able to handle 120, 000 QPS at traffic peak and exhibits only a 10 − 20 ms increase in online time delay and limited increase in machine overhead after replacing the original point-wise ranking model with MIREC. Researchers may choose to simplify the network structure of PCR to further reduce the online machine overhead in practical systems. However, such simplification may come at a cost of estimation accuracy, such that one may need to control the trade-off based on actual performance carefully.

5 EXPERIMENTAL RESULTS

This section conducts extensive experiments on both offline datasets and real-world applications with the goal of answering the following research questions: **Q1:** Does our proposed PCR and CAR outperform other baseline models in integrated ranking tasks? **Q2:** Does our proposed MIREC framework outperform other methods in integrated recommendation tasks with exposure constraints? **Q3:** How does MIREC perform in real-world applications? Our source codes and hyper-parameter settings have been made public to ensure reproducibility[1].

5.1 Experimental Setup

5.1.1 *Datasets.* We use one public dataset named MicroVideo-1.7M and one industrial dataset named Taobao for experiments. The public available MicroVideo-1.7M dataset[2] released by [11] contains 12, 737, 619 interactions that 10, 986 users have made on 1, 704, 880 micro-videos. This dataset provides rich user behavior data and timestamps to evaluate the performance on both interest modeling and context-aware reranking. The Taobao dataset is an industrial private dataset that contains users' behaviors and feedback logs from multiple channels in the homepage feed of Taobao Mobile App. It is one of the largest feed scenarios for online merchandise in China. The feed provides items in form of the streams, videos, pictures, etc, from various channels. Users can slide to view more items in a row. This dataset contains about ten billion interactions that one hundred million of users have made on sixty million items. We also conduct online A/B tests on the platform Taobao to examine the performance of MIREC in real-world applications.

5.1.2 *Comparing Methods.* We compare MIREC with two mainstreams of baselines. The first steam of baselines are the methods for ranking tasks with different goals on user interest modeling (i.e., DIN and DIEN), re-ranking (i.e., DLCM, PRM, and PEAR), or multi-channel recommendation (i.e., STAR and DHANR). Specifically, **DIN** [45] is a widely used benchmark for sequential user data modeling in point-wise CTR predictions, which models short behavior sequences with target attention. **DIEN** [45] combines GRUs and attention to capture temporal interests from users' historical behaviors with respect to the target item. **DLCM** [1] uses gated recurrent units (GRU) to sequentially encode the top-ranked items with their feature vectors. **PRM** [31]: directly optimizes the whole recommendation list by employing a Transformer structure to efficiently encode the information of all items in the list. **PEAR** [24] not only captures feature-level and item-level interactions but also models item contexts from both the candidate list and the historical clicked item list. **STAR** [34] trains a unified model to serve all channels simultaneously, which consists of shared centered parameters and channel-specific parameters. **DHANR** [16] proposes a hierarchical self-attention structure to consider cross-channel interactions.

[1] https://github.com/EzailShen/MIREC
[2] https://github.com/Ocxs/THACIL

Table 1: Comparison of ranking performance (bold: best; underline: runner-up).

Dataset	Method	AUC	Logloss	NDCG@20	NDCG@30
MicroVideo[3]	DIN	0.6831	0.5922	0.5403	0.6535
	DIEN	0.6842	0.5909	0.5408	0.6537
	DLCM	0.6872	0.5898	0.5582	0.6698
	PRM	0.6979	0.5872	0.5591	0.6708
	PEAR	0.7021	0.5821	0.5632	0.6745
	Ours	**0.7084**	**0.5787**	**0.5667**	**0.6826**
Taobao	DIN	0.7681	0.4982	0.5203	0.6481
	DIEN	0.7692	0.4971	0.5202	0.6479
	DLCM	0.7699	0.4965	0.5209	0.6482
	PRM	0.7722	0.4941	0.5211	0.6489
	PEAR	0.7748	0.4919	0.5232	0.6511
	STAR	0.7738	0.4931	0.5219	0.6492
	DHANR	0.7753	0.4923	0.5243	0.6513
	Ours	**0.7791**	**0.4899**	**0.5275**	**0.6545**

The second stream of baselines is the online allocation methods which have been successfully applied in industrial applications for online resource allocation.These methods are similar to the previous solutions used in our system, such that they are all comparable and competitive baselines. **Fixed** is the fixed-positions strategy, where the positions of recommended items and ads are manually pre-determined for every request. β-**WPO**: is based on the Whole-Page Optimization (WPO) [41]. WPO ranks recommended and ad candidates jointly according to the predefined ranking scores. Similar to [6], we introduce an adjustable variable β to control the proportion of different channels on each request to satisfy the resource constraint. In general, β-WPO can be regarded as a heuristic list merging algorithm. Each list from one channel is assigned a priority weight. The algorithm merges the top items of each list based on both their ranking scores and the priority weights into a final return list. **HCA2E** [6]: proposed a two-level optimization framework based on BwK methods. The high-level determines whether to present ads on the page while the low-level searches the optimal position to insert ads heuristically.

5.1.3 Metrics. For offline experiments, we use user clicks to measure the utility function f. We compare the performance using the widely used Area Under ROC (AUC) and normalized discounted cumulative gain (nDCG) [23], where nDCG@K refers to the performance of top-k recommended items. For online experiments, we consider a joint measurement of user click, purchase, and stay-time for utility function f. The metrics are CLICK, Click-Through-Rate (CTR), Gross Merchandise Volume (GMV), and Stay Time. Here, CLICK refers to the total number of clicked items. CTR is defined as CLICK/PV with PV denoting the total number of impressed items. CTR measures users' willingness to click and is therefore a widely used metric in practical applications. GMV is a term used in online retailing to indicate a total sales monetary-value for merchandise sold over a certain period of time. Stay Time denotes the time period of users' average stay time in the product, averaged on all users. For all experiments, we use the exposure of items to measure cost function g.

Table 2: Ablation study of the ranking components.

	AUC	Logloss	NDCG@20	NDCG@30
PCR*	0.7758	0.4933	0.5222	0.6511
PCR†	0.7761	0.4932	0.5246	0.6513
PCR w/o IntraCE	0.7763	0.4928	0.5247	0.6516
PCR w/o InterCE	0.7756	0.4935	0.5239	0.6509
PCR	0.7778	0.4913	0.5262	0.6531
PCR+CAR (propsed)	0.7791	0.4899	0.5275	0.6545

5.2 Offline Evaluation

5.2.1 Q1: Performance on Integrated Ranking. We first compare the performance with the first stream of baselines on item ranking. The results are shown in Table 1, which leads to the following findings. First, the re-ranking methods perform generally better than the point-wise user interest methods, indicating that modeling mutual influence among the input ranking list is of vital importance for the ranking. Therefore, it is essential to consider the influence from page context in feed recommendations. Second, the multi-channel methods perform better than the reranking methods, which verifies that exploiting the distinction and mutual influence among different channels has a great impact on integrated recommendations. Besides, we also notice that DHANR performs better than STAR, which may be due to that DHANR considers both the correlation among different channels and the influence from the candidate list. Finally, our proposed MIREC model achieves superior performance than other competitors on all datasets, verifying the effectiveness of joint modeling the cross-channel information, user interest, context information, and candidate list.

5.2.2 Ablation Study. The results in Table. 2 investigates the impact of each component of MIREC on item quality estimation. Specifically, PCR* replaces the attention mechanism for user behaviors in the Merged-Sequence Evolution layer with a self-attention mechanism which is in accordance with PEAR [24]. PCR outperforms PCR* indicates that the tailored attention mechanism in PCR can filter out noisy or redundant information from historical behaviors to benefit the subsequent modeling of bi-sequence interaction. PCR† removes the scaled dot-product attention mechanism (i.e. there is no explicit interaction between initial lists and user behaviors) and achieved a worse performance. This demonstrates the necessity of this direct modeling between sequences, directly guiding the reordering of the initial lists. PCR w/o IntraCE removes the Intra-Channel Encoding module, i.e., directly feeding the embeddings of the initial item lists into subsequent layers for learning. The result shows that PCR achieves superior performance than PCR w/o IntraCE, verifying that it is of vital importance to model the mutual information inside each channel for final prediction. PCR w/o InterCE removes the Inter-Channel Encoding, which also leads to worse performance. It verifies that without considering the relationship and distinction between different channels will degrade the model performance considerably. Moreover, the joint learning of PCR and CAR performs better than only using PCR. This verifies that the modeling of page-wise context information can improve prediction accuracy effectively.

5.2.3 Q2: Performance with Exposure Constraints. To the best of our knowledge, there does not exist publicly available datasets which has rich user logs and multi-channel features to examine the

[3] Note that MicroVideo is a public dataset with a single channel, such that we omit the comparison with STAR and DHANR which are proposed for multi-channel modeling.

Table 3: Joint performance of allocation and ranking (bold: best; dagger: baseline).

Exp. Settings	Method	Exposure completeness				CTR	CTR Lift
		CH1	CH2	CH3	CH4		
Setting-1	Fixed	0.44%	1.35%	0.20%	0.50%	5.54%†	-
	WPO	0.15%	0.15%	0.13%	0.30%	6.09%	+9.93%
	HCA2E	0.24%	0.95%	0.33%	0.10%	6.34%	+14.44%
	Ours	0.02%	0.65%	0.67%	0.20%	**6.56%***	**+18.41%***
Setting-2	Fixed	0.10%	0.53%	0.10%	0.20%	5.91%†	-
	WPO	0.16%	0.27%	1.40%	0.20%	6.28%	+6.26%
	HCA2E	0.19%	0.53%	0.30%	0.40%	6.53%	+10.49%
	Ours	0.17%	0.53%	0.40%	0.20%	**6.76%***	**+14.38%***

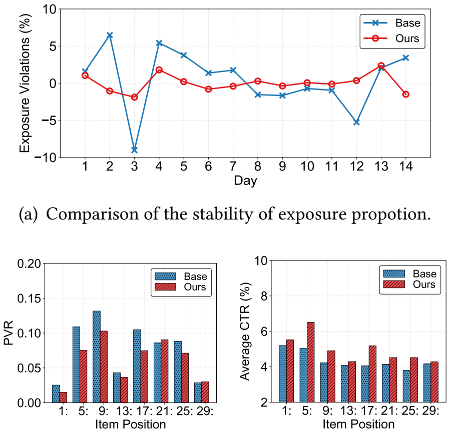
Table 4: Results of online A/B tests.

	CLICK	CTR	GMV	Stay Time
Ours vs Fixed	+4.02%	+2.15%	+1.98%	+2.01%
Ours vs Baseline	+3.00%	+1.75%	+1.56%	+1.42%

joint performance of integrated recommendation and exposure allocation. Therefore, we only perform experiments on Taobao dataset, using the complete platform logs. In this experiment, we assume that the IRS needs to allocate exposures to satisfy the exposure guarantees of four distinct channels. The aim is to maximize the overall user-click utility of all channels. The compared fixed, WPO, and HCA2E methods all use point-wise scores to be consistent with their original proposals. For HCA2E, we use their proposed heuristic search method to determine the final order of the item list. The results are shown in Table 3, which are averaged on multiple runs to give a fair comparison. The simulated time horizon is one complete day with more than one billion user requests from real productive environment. We evaluate the performance using two different sets of lower bounds: 1) Channel 1=55%, Channel 2 = 20%, Channel 3 = 15%, Channel 4 = 10%; 2) Channel 1=70%, Channel 2 = 15%, Channel 3 = 10%, Channel 4 = 5%. The parameters of all comparing methods are carefully tuned to satisfy the exposure constraints. The completeness in Table 3 shows that all methods can control the violation of constraints to a low-level. HCA2E and our proposed MIREC perform slightly better than the fixed method and the WPO method. Noticeably, our proposed method outperforms other comparing methods considerably in terms of CTR enhancement, which verifies that the joint use of the allocation and estimation algorithm can bring a remarkable improvement in practical environments.

5.3 Online Evaluation

MIREC has been fully deployed in the homepage feed of Taobao named *Guess-you-like* to serve the main traffic. Guess-you-like is one of the largest merchandise feed recommendation platform in China, which serves more than hundreds of millions of users toward billions of items every day. We deploy MIREC at the integrated recommendation stage in Guess-you-like platform, which takes hundreds of candidate items provided by multiple channels as input and outputs the final item list to return to the user. The online performance is compared against our previous baseline which is similar as a combination of β-WPO and HCA2E. In particular, the baseline uses a point-wise model for item quality estimation and uses a PID-based feedback control to automatically adjust parameter β to guarantee the exposures for different channels. For each user request, the baseline also runs an MDP-based search method to determine the optimal card layout based on the estimated scores, which is similar as the heuristic search method in HCA2E.



(a) Comparison of the stability of exposure propotion.

(b) Exposure distribution.

(c) CTR on different positions.

Figure 4: Online Performance Analysis.

The overall performance in Table 4 is averaged over two consecutive weeks. The results show that compared with the baseline method, MIREC brings an improvement of 3.00% for CLICK, 1.75% for CTR, 1.56% for GMV, and 1.42% for stay time. Compared with the fixed method, MIREC brings an improvement of 3.00% for CLICK, 1.75% for CTR, 1.56% for GMV, and 1.42% for stay time. These improvements indicate that our framework is able to increase user’s willingness to stay and interact with the recommended items in practical applications. It is noteworthy that 1% improvement on CLICK in Guess-you-like brings millions of clicks every day. Figure. 4 shows a detailed comparison of the exposure allocation results of a specific channel, where the items from this channel have a generally lower CTR than others. Each line in Figure. 4(a) represents the robustness of long-term exposure guarantee of this channel within two consecutive weeks. It is clear that compared with the baseline, our proposed MIREC is more robust to alleviate daily exposure fluctuations. The distribution of exposures on different positions in the feed is given in Figure. 4(b). The result shows that our proposed framework tends to put lower-quality items backward to increase the overall utilities of all channels. Consequently, as shown in Figure. 4(c), the averaged CTR of all channels on each position can be improved remarkably. This verifies that MIREC is superior in optimizing the item layout from a global perspective.

6 CONCLUSION

In this paper, we propose a two-layer framework named MIREC for the integrated recommendation task with exposure constraints. Extensive experiments verified the effectiveness of our proposed framework. MIREC has been implemented on the homepage feed in Taobao to serve the main traffic.

ACKNOWLEDGMENTS

We express our sincere gratitude to Deqiang Kong and Tao Wang for their support at the exploratory stage of this project and the implementation in online systems.

REFERENCES

[1] Qingyao Ai, Keping Bi, Jiafeng Guo, and W Bruce Croft. Learning a deep listwise context model for ranking refinement. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 135–144, 2018.

[2] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

[3] Santiago R Balseiro, Haihao Lu, and Vahab Mirrokni. The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research*, 2022.

[4] Irwan Bello, Sayali Kulkarni, Sagar Jain, Craig Boutilier, Ed Chi, Elad Eban, Xiyang Luo, Alan Mackey, and Ofer Meshi. Seq2slate: Re-ranking and slate optimization with rnns. *arXiv preprint arXiv:1810.02019*, 2018.

[5] Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

[6] Dagui Chen, Qi Yan, Chunjie Chen, Zhenzhe Zheng, Yangsu Liu, Zhenjia Ma, Chuan Yu, Jian Xu, and Bo Zheng. Hierarchically constrained adaptive ad exposure in feeds. *arXiv preprint arXiv:2205.15759*, 2022.

[7] Hao Chen, Zhong Huang, Yue Xu, Zengde Deng, Feiran Huang, Peng He, and Zhoujun Li. Neighbor enhanced graph convolutional networks for node classification and recommendation. *Knowledge-Based Systems*, 246:108594, 2022.

[8] Hao Chen, Zefan Wang, Feiran Huang, Xiao Huang, Yue Xu, Yishi Lin, Peng He, and Zhoujun Li. Generative adversarial framework for cold-start item recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2565–2571, 2022.

[9] Hao Chen, Yue Xu, Feiran Huang, Zengde Deng, Wenbing Huang, Senzhang Wang, Peng He, and Zhoujun Li. Label-aware graph convolutional networks. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1977–1980, 2020.

[10] Qiwei Chen, Yue Xu, Changhua Pei, Shanshan Lv, Tao Zhuang, and Junfeng Ge. Efficient long sequential user data modeling for click-through rate prediction. *arXiv preprint arXiv:2209.12212*, 2022.

[11] Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, and Yan Li. Temporal hierarchical attention at category- and item-level for micro-video click-through prediction. In *MM*, pages 1146–1153, 2018.

[12] Yuwei Chen, Zengde Deng, Yinzhi Zhou, Zaiyi Chen, Yujie Chen, and Haoyuan Hu. An online algorithm for chance constrained resource allocation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[13] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.

[14] Yufei Feng, Binbin Hu, Yu Gong, Fei Sun, Qingwen Liu, and Wenwu Ou. GRN: Generative rerank network for context-wise recommendation. *arXiv preprint arXiv:2104.00860*, 2021.

[15] Jyotirmoy Gope and Sanjay Kumar Jain. A survey on solving cold start problem in recommender systems. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 133–138. IEEE, 2017.

[16] Qi Hao, Tianze Luo, and Guangda Huzhang. Re-ranking with constraints on diversified exposures for homepage recommender system. *arXiv preprint arXiv:2112.07621*, 2021.

[17] Xiaotian Hao, Zhaoqing Peng, Yi Ma, Guan Wang, Junqi Jin, Jianye Hao, Shan Chen, Rongquan Bai, Mingzhou Xie, Miao Xu, et al. Dynamic knapsack optimization towards efficient multi-channel sequential advertising. In *International Conference on Machine Learning*, pages 4060–4070, 2020.

[18] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[19] Jinhong Huang, Yang Li, Shan Sun, Bufeng Zhang, and Jin Huang. Personalized flight itinerary ranking at fliggy. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2541–2548, 2020.

[20] Yanhua Huang, Weikun Wang, Lei Zhang, and Ruiwen Xu. Sliding spectrum decomposition for diversified recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3041–3049, 2021.

[21] Zhongyu Huang, Yingheng Wang, Chaozhuo Li, and Huiguang He. Going deeper into permutation-sensitive graph neural networks. In *International Conference on Machine Learning*, pages 9377–9409. PMLR, 2022.

[22] Nicole Immorlica, Karthik Abinav Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 202–219, 2019.

[23] Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, pages 243–250, 2017.

[24] Yi Li, Jieming Zhu, Weiwen Liu, Liangcai Su, Guohao Cai, Qi Zhang, Ruiming Tang, Xi Xiao, and Xiuqiang He. Pear: Personalized re-ranking with contextualized transformer for recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 62–69, 2022.

[25] Guogang Liao, Ze Wang, Xiaoxu Wu, Xiaowen Shi, Chuheng Zhang, Yongkang Wang, Xingxing Wang, and Dong Wang. Cross DQN: Cross deep Q network for ads allocation in feed. In *Proceedings of the ACM Web Conference 2022*, pages 401–409, 2022.

[26] Zhuoyi Lin, Sheng Zang, Rundong Wang, Zhu Sun, Chi Xu, and Chee-Keong Kwoh. Attention over self-attention: Intention-aware re-ranking with dynamic transformer encoders for recommendation. *arXiv preprint arXiv:2201.05333*, 2022.

[27] Zihan Lin, Hui Wang, Jingshu Mao, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and Ji-Rong Wen. Feature-aware diversified re-ranking with disentangled representations for relevant recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3327–3335, 2022.

[28] Weiwen Liu, Yunjia Xi, Jiarui Qin, Fei Sun, Bo Chen, Weinan Zhang, Rui Zhang, and Ruiming Tang. Neural re-ranking in multi-stage recommender systems: A review. In *IJCAI*, pages 5512–5520, 2020.

[29] Alfonso Lobos, Paul Grigas, and Zheng Wen. Joint online learning and decision-making via dual mirror descent. In *International Conference on Machine Learning*, pages 7080–7089, 2021.

[30] Xingyu Lu, Qintong Wu, and Wenliang Zhong. Multi-slots online matching with high entropy. In *International Conference on Machine Learning*, pages 14412–14428. PMLR, 2022.

[31] Changhua Pei, Yi Zhang, Yongfeng Zhang, Fei Sun, Xiao Lin, Hanxiao Sun, Jian Wu, Peng Jiang, Junfeng Ge, Wenwu Ou, et al. Personalized re-ranking for recommendation. In *Proceedings of the 13th ACM conference on recommender systems*, pages 3–11, 2019.

[32] Qi Pi, Weijie Bian, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. Practice on long sequential user behavior modeling for click-through rate prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2671–2679, 2019.

[33] Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang, Ziwen Jiang, Qingwen Liu, Xiaoyi Zeng, Tat-Seng Chua, and Fei Wu. Intelligent request strategy design in recommender system. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3772–3782, 2022.

[34] Xiang-Rong Sheng, Liqin Zhao, Guorui Zhou, Xinyao Ding, Binding Dai, Qiang Luo, Siran Yang, Jingshan Lv, Chi Zhang, Hongbo Deng, et al. One model to serve all: Star topology adaptive recommender for multi-domain ctr prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4104–4113, 2021.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] Fan Wang, Xiaomin Fang, Lihang Liu, Yaxue Chen, Jiucheng Tao, Zhiming Peng, Cihang Jin, and Hao Tian. Sequential evaluation and generation framework for combinatorial recommender system. *arXiv preprint arXiv:1902.00245*, 2019.

[37] Ruobing Xie, Shaoliang Zhang, Rui Wang, Feng Xia, and Leyu Lin. Hierarchical reinforcement learning for integrated recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4521–4528, 2021.

[38] Biao Yuan, Zengde Deng, Na Geng, Yujie Chen, and Haoyuan Hu. Practice summary: Cainiao optimizes the fulfillment routes of parcels. *INFORMS Journal on Applied Analytics*, 2023.

[39] Jianjun Yuan and Andrew Lamperski. Online convex optimization for cumulative constraints. pages 6140–6149, 2018.

[40] Peiyan Zhang, Jiayan Guo, Chaozhuo Li, Yueqi Xie, Jae Boum Kim, Yan Zhang, Xing Xie, Haohan Wang, and Sunghun Kim. Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 168–176, 2023.

[41] Weiru Zhang, Chao Wei, Xiaonan Meng, Yi Hu, and Hao Wang. The whole-page optimization via dynamic ad allocation. In *Companion Proceedings of the The Web Conference 2018*, pages 1407–1411, 2018.

[42] Yiding Zhang, Chaozhuo Li, Xing Xie, Xiao Wang, Chuan Shi, Yuming Liu, Hao Sun, Liangjie Zhang, Weiwei Deng, and Qi Zhang. Geometric disentangled collaborative filtering. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 80–90, 2022.

[43] Xiangyu Zhao, Changsheng Gu, Haoshenglun Zhang, Xiwang Yang, Xiaobing Liu, Jiliang Tang, and Hui Liu. Dear: Deep reinforcement learning for online advertising impression in recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 750–758, 2021.

[44] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5941–5948, 2019.

[45] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1059–1068, 2018.

[46] Yu-Hang Zhou, Peng Hu, Chen Liang, Huan Xu, Guangda Huzhang, Yinfu Feng, Qing Da, Xinshang Wang, and An-Xiang Zeng. A primal-dual online algorithm for online matching problem in dynamic environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11160–11167, 2021.

[47] Tao Zhuang, Wenwu Ou, and Zhirong Wang. Globally optimized mutual influence aware ranking in e-commerce search. In *IJCAI*, pages 3725–3731, 2018.

A PROOF OF REGRET BOUND

Our proof shares the same spirit as that of Theorem 1 in [3, 29]. The difference is that [3] does not consider a lower resource limit while [29] develops proof with an additional learnable parameter within $f(\boldsymbol{x}_t)$ and $g(\boldsymbol{x}_t)$. Therefore, we here develop a separate proof that is consistent with our formulation. We directly refer to a few propositions in [3, 29] as preliminaries for simplicity. It is noteworthy that developing new proof of the online revenue maximization problem is not the main focus of this paper.

Recall that the integrated recommendation problem is

$$\mathcal{P}_0: \quad \mathrm{OPT}(\mathcal{S}) = \max_{\boldsymbol{x}_t \in \mathcal{X}} \sum\nolimits_{t=1}^{T} f(\boldsymbol{x}_t) \tag{18}$$

$$\text{s.t.} \quad C_1: \sum\nolimits_{t=1}^{T} g_m(\boldsymbol{x}_t) \leq G_{m,th}^{\max} N(\mathcal{S}), \forall m \in \mathcal{M}, \tag{19}$$

$$C_2: \sum\nolimits_{t=1}^{T} g_m(\boldsymbol{x}_t) \geq G_{m,th}^{\min} N(\mathcal{S}), \forall m \in \mathcal{M}. \tag{20}$$

Since $N(\mathcal{S})$ denotes the sum exposures over the entire time horizon from $t = 1$ to T, we can replace the the upper exposure limit $G_{m,th}^{\max} N(\mathcal{S})$ and lower exposure limit $G_{m,th}^{\min} N(\mathcal{S})$ with TG_m and αTG_m for simplicity, respectively, where $G_m, \alpha \in [0,1]$ are constants. As such, problem $\mathcal{P}_0$ can be reformulated as

$$\mathcal{P}_1: \quad \mathrm{OPT}(\mathcal{S}) = \max_{\boldsymbol{x}_t \in \mathcal{X}} \sum\nolimits_{t=1}^{T} f(\boldsymbol{x}_t) \tag{21}$$

$$\text{s.t.} \quad \alpha TG_m \leq \sum\nolimits_{t=1}^{T} g_m(\boldsymbol{x}_t) \leq TG_m, \forall m \in \mathcal{M}. \tag{22}$$

Before our analysis, we define constants $\bar{f}, \bar{g} > 0, \underline{G} > 0$ and $\bar{G} > 0$ such that $\sup_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) \leq \bar{f}, \sup_{\boldsymbol{x} \in \mathcal{X}} g(\boldsymbol{x}) \leq \bar{g}, \underline{G} := \min_{m \in \mathcal{M}} G_m$ and $\bar{G} := \max_{m \in \mathcal{M}} G_m$. Also, θ refers to the strongly-convexity parameter of the reference function $h(\cdot)$.

First, we bound the dual iterates as follows.

Assumption A.1. *There exists a constant $C_h > 0$ such that the dual iterates μ^t satisfy $\mathbb{E}[||\nabla h(\mu^t)||_\infty] \leq C_h, \forall t \in [T]$.*

Remark 1. *Note that, when choosing the reference function $h(\lambda) := \frac{1}{2}\|\lambda\|^2$, Assumption A.1 can be omited according to Proposition 3 in [29].*

Denote the online Algorithm 1 as π which makes a real-time decision $\boldsymbol{x}_t$ at time t. Define the stopping time $\tau_\pi \leq T$ as the minimum between T and the smallest time t such that there exists $m \in \mathcal{M}$ with $\sum_{t=1}^{\tau_\pi} g_m(\boldsymbol{x}_t) + \bar{g} > TG_k$. In other words, τ_π refers to the first time the violation of one resource constraint happens. We bound the averaged gap between T and τ_π as follows.

Proposition 2. *Suppose that Assumption A.1 holds, using a constant step-size $\eta > 0$ in Algorithm 1 yields*

$$\mathbb{E}\left[T - \tau_\pi\right] \leq \frac{\bar{g}}{\underline{G}} + \frac{C_h + \|\nabla h(\lambda^1)\|_\infty}{\eta \underline{G}}. \tag{23}$$

Proof. According to Step. 9 in Algorithm 1 we have

$$\begin{aligned} \nabla \mu_{k,t} &= -g_k(\boldsymbol{x}_t) + G_k\left(\mathbb{1}(\boldsymbol{\mu}_k \geq 0) + \alpha_k \mathbb{1}(\boldsymbol{\mu}_k < 0)\right), \\ &\leq -g_k(\boldsymbol{x}_t) + G_k, \quad \forall k \in [m]. \end{aligned} \tag{24}$$

Assume that the stopping time τ_π is activated due to the violation of constraint on k-th channel, we have

$$\sum_{t=1}^{\tau_\pi} \nabla \mu_{k,t} \leq G_k \tau_\pi - \sum_{t=1}^{\tau_\pi} g_k(\boldsymbol{x}_t) \leq G_k \tau_\pi - TG_k + \bar{g}, \tag{25}$$

which leads to

$$T - \tau_\pi \leq \frac{1}{G_k}\left(\bar{g} - \sum_{t=1}^{\tau_\pi} \nabla \mu_{k,t}\right). \tag{26}$$

According to Proposition 6 in [29], the gradients of mirror descent satisfy $\nabla h_k(\mu_k^{t+1}) \geq \nabla h_k(\mu_k^t) - \eta \nabla \mu_{k,t}^t, \forall t \leq \tau_\pi$, such that $-\sum_{t=1}^{\tau_\pi} \nabla \mu_{k,t} \leq \frac{1}{\eta}\left(\nabla h_k(\mu_k^{\tau_\pi+1}) - \nabla h_k(\mu_k^1)\right)$. Combing with the inequality in (26), we obtain

$$\mathbb{E}\left[T - \tau_\pi\right] \leq \frac{\bar{g}}{G_k} + \mathbb{E}\left[\frac{\nabla h_k(\mu_k^{\tau_\pi+1}) - \nabla h_k(\mu_k^1)}{\eta G_k}\right] \tag{27}$$

$$\leq \frac{\bar{g}}{\underline{G}} + \frac{C_h + \|\nabla h(\lambda^1)\|_\infty}{\eta \underline{G}}, \tag{28}$$

as required. ■

Let us denote the random variable γ_t to be the type of the request at period t, which can determine the sample of the request.

Proposition 3. *Using a constant step-size rule $\eta > 0$ for $t > 1$ in Algorithm 1, it holds*

$$\mathbb{E}\left[\tau_\pi D(\bar{\boldsymbol{\mu}}^{\tau_\pi}) - \sum_{t=1}^{\tau_\pi} f(\boldsymbol{x}_t)\right] \leq \frac{2(\bar{g}^2 + \bar{G}^2)}{\theta} \eta \mathbb{E}[\tau_\pi] + \frac{1}{\eta} V_h(\boldsymbol{\mu}, \boldsymbol{\mu}^1), \tag{29}$$

where $\bar{\boldsymbol{\mu}}^{\tau_\pi} = \frac{\sum_{t=1}^{\tau_\pi} \boldsymbol{\mu}^t}{\tau_\pi}$.

Proof. According to the definition of $\nabla \boldsymbol{\mu}_t$ and the subgradient inequality, we have

$$\begin{aligned} &(\nabla \boldsymbol{\mu}^t)^T(\boldsymbol{\mu}^t - \boldsymbol{\mu}) \geq D(\boldsymbol{\mu}^t) - D(\boldsymbol{\mu}) \\ &\geq D(\boldsymbol{\mu}^t) - \left(\mathbb{E}_{\gamma_t}[\varphi(\boldsymbol{\mu})] + \sum_{k \in [K]} G_k([\boldsymbol{\mu}_k]_+ - \alpha_k[-\boldsymbol{\mu}_k]_+)\right), \end{aligned} \tag{30}$$

where $\varphi(\boldsymbol{\mu}) = f^*(\boldsymbol{\mu}) - \boldsymbol{\mu}^T g(\boldsymbol{x}_t)$. Considering that $\boldsymbol{x}_t$ is an optimal solution of $\varphi(\boldsymbol{\mu}_t)$ not of $\varphi(\boldsymbol{\mu})$, we have $f(\boldsymbol{x}_t) - \boldsymbol{\mu}^T g(\boldsymbol{x}_t) \leq \varphi(\boldsymbol{\mu})$. Then, by taking $\boldsymbol{\mu} = [0, 0, \ldots, 0]$, and summing from one to τ_π, we obtain

$$\begin{aligned} &\sum_{t=1}^{\tau_\pi} (\nabla \boldsymbol{\mu}^t)^T(\boldsymbol{\mu}^t - 0) \geq \sum_{t=1}^{\tau_\pi} D(\boldsymbol{\mu}^t) - \sum_{t=1}^{\tau_\pi} \mathbb{E}_{\gamma_t}[f(\boldsymbol{x}_t)] \\ &\geq \tau_\pi D(\bar{\boldsymbol{\mu}}^{\tau_\pi}) - \sum_{t=1}^{\tau_\pi} \mathbb{E}_{\gamma_t}[f(\boldsymbol{x}_t))], \end{aligned} \tag{31}$$

where $\bar{\boldsymbol{\mu}}^{\tau_\pi} = \frac{\sum_{t=1}^{\tau_\pi} \boldsymbol{\mu}^t}{\tau_\pi}$ and the inequality is based on the fact that the dual function is convex. In this paper, we adopt θ-strongly convex function as the relation function $h(\cdot)$ in mirror descents. According to Step. 2 of Proposition 8 in [29], we have

$$\mathbb{E}\left[\sum_{t=1}^{\tau_\pi} (\nabla \boldsymbol{\mu}^t)^T(\boldsymbol{\mu}^t - \boldsymbol{\mu})\right] \leq \frac{2(\bar{g}^2 + \bar{G}^2)}{\theta} \eta \mathbb{E}[\tau_A] + \frac{V_h(\lambda, \lambda^1)}{\eta}. \tag{32}$$

Combining (32) with (31), we get

$$\mathbb{E}\left[\tau_\pi D(\bar{\boldsymbol{\mu}}^{\tau_\pi}) - \sum_{t=1}^{\tau_\pi} \mathbb{E}_{\gamma_t}[f(\boldsymbol{x}_t)]\right] \leq \frac{2(\bar{g}^2 + \bar{G}^2)}{\theta} \eta \mathbb{E}[\tau_\pi] + \frac{1}{\eta} V_h(\boldsymbol{\mu}, \boldsymbol{\mu}^1). \tag{33}$$

According to Step. 3 of Proposition 8 in [29], we have

$$\mathbb{E}\left[\sum_{t=1}^{\tau_\pi} \mathbb{E}_{\gamma_t}[f(\boldsymbol{x}_t)]\right] = \mathbb{E}\left[\sum_{t=1}^{\tau_\pi} f(\boldsymbol{x}_t)\right]. \tag{34}$$

Combining (33) and (34), we complete the proof of Proposition 3. ■

Before providing more details on the proof of regret bound, we introduce a new benchmark of problem $\mathcal{P}_0$ as in [29] due to that problem $\mathcal{P}_0$ may be infeasible due the presence of both lower and upper exposure constraints. Specifically, we define

$$\mathrm{F}(\boldsymbol{x}_t, \lambda) = (1-\lambda) f(\boldsymbol{x}_t) + \lambda \mathbb{E}_{\mathcal{S}}[f(\boldsymbol{x}_t)]$$
$$\mathrm{G}(\boldsymbol{x}_t, \lambda) = (1-\lambda) g(\boldsymbol{x}_t) + \lambda \mathbb{E}_{\mathcal{S}}[g(\boldsymbol{x}_t)],$$

where $\lambda \in [0,1]$ is the interpolation parameter. We define

$$\mathrm{OPT}(\mathcal{S}, \lambda) = \mathbb{E}_{\mathcal{S}^T}\left[\begin{array}{cc} \max\limits_{\boldsymbol{x}^t, t\in[T]} & \sum_{t=1}^T F(\boldsymbol{x}_t, \lambda) \\ \text{s.t. } T\alpha \odot G \le \sum_{t=1}^T G(\boldsymbol{x}_t, \lambda) \le TG & \end{array}\right]$$

where $\mathcal{S}^T := \mathcal{S} \times \cdots \times \mathcal{S}$ is a product distribution of length T. Now we give the definition of the new benchmark as

$$\mathrm{OPT}(\mathcal{S}) := \max_{\lambda \in [0,1]} \mathrm{OPT}(\mathcal{S}, \lambda). \tag{35}$$

This benchmark is an interpolate value between the expected optimal value of problem $\mathcal{P}_0$ and a deterministic problem which replaces the varying utility values $f(\boldsymbol{x}_t)$ and cost values $g(\boldsymbol{x}_t)$ with their expected values. For this benchmark, we have

$$\begin{aligned} \mathbb{E}_{\mathcal{S}}[\mathrm{OPT}(\mathcal{S})] &= \frac{\tau_\pi}{T}\mathbb{E}_{\mathcal{S}}[\mathrm{OPT}(\mathcal{S})] + \frac{T-\tau_\pi}{T}\mathbb{E}_{\mathcal{S}}[\mathrm{OPT}(\mathcal{S})] \\ &\le \tau_\pi \bar{D}(\boldsymbol{\mu}_{\tau_\pi}|\mathcal{S}) + (T-\tau_\pi)\bar{f}, \end{aligned} \tag{36}$$

where the inequality uses the fact that $\mathrm{OPT}(\mathcal{S}) \le D(\boldsymbol{\mu}|\mathcal{S})$ according to Proposition 1 in [29]and that $\mathrm{OPT}(\mathcal{S}) \le T\bar{f}$. Combining all findings together, we have

$$\mathrm{Regret}(\pi|\mathcal{S}) = \mathbb{E}_{\mathcal{S}}[\mathrm{OPT}(\mathcal{S}) - R(\pi|\mathcal{S})] \tag{37a}$$

$$\le \mathbb{E}_{\mathcal{S}}\left[\tau_\pi \bar{D}(\boldsymbol{\mu}_{\tau_\pi}|\mathcal{S}) + (T-\tau_\pi)\bar{f} - \sum_{t=1}^{\tau_\pi} f(\boldsymbol{x}_t)\right] \tag{37b}$$

$$= \mathbb{E}_{\mathcal{S}}\left[\tau_\pi \bar{D}(\boldsymbol{\mu}_{\tau_\pi}|\mathcal{S}) - \sum_{t=1}^{\tau_\pi} f(\boldsymbol{x}_t)\right] + \mathbb{E}_{\mathcal{S}}[T-\tau_\pi]\bar{f} \tag{37c}$$

$$\le \frac{2(\bar{g}^2 + \bar{G}^2)}{\theta}\eta\mathbb{E}[\tau_\pi] + \frac{1}{\eta}V_h(\boldsymbol{\mu}, \boldsymbol{\mu}^1) + \frac{\bar{g}}{\underline{G}} + \frac{C_h + \|\nabla h(\lambda^1)\|_\infty}{\eta\underline{G}}, \tag{37d}$$

where the first inequality is from (36) and the second inequality is from Proposition 2 and Proposition 3. Therefore, the constants in Theorem 1 are $C_1 = \frac{\bar{g}}{\underline{G}}$, $C_2 = \frac{2(\bar{C}^2+\bar{b}^2)}{\theta}\eta$, and $C_3 = V_h(\boldsymbol{\mu}, \boldsymbol{\mu}^1) + \frac{\bar{g}}{\underline{G}} + \frac{C_h + \|\nabla h(\lambda^1)\|_\infty}{\underline{G}}$, respectively. Moreover, recall Remark 1, we choose $h(\lambda) := \frac{1}{2}\|\lambda\|^2$ and dual iterates are bounded. Hence, we complete the proof of Theorem 1. ■

B PROOF OF COST FEASIBILITY

Proposition 1 shows that a solution obtained using Algorithm 1 can not overspend, but may underspend. Based on the definition of subgradient $\nabla\mu_k^t$, we have

$$\frac{\nabla h_k(\mu^1) - \nabla h_k(\mu^{\tau_\pi+1})}{\eta} = \sum_{t=1}^{\tau_A}(G_k(\mathbb{1}(\mu_k \ge 0) + \alpha_k\mathbb{1}(\mu_k < 0)) - g_k(\boldsymbol{x}_t)) \tag{38}$$

Now, given that $\mathbb{1}(\mu \ge 0) + \alpha_k\mathbb{1}(\mu < 0) \ge \alpha_k$ for any $\mu \in \mathbb{R}$ and that $\tau_\pi \le T$ by definition, we have

$$\sum_{t=1}^{\tau_A} G_k(\mathbb{1}(\mu_k \ge 0) + \alpha_k\mathbb{1}(\mu_k < 0)) + (T-\tau_A)\alpha_k b_k \ge T\alpha_k b_k. \tag{39}$$

Combining (38) and (39) and taking expectation, we get

$$T\alpha_k G_k - \mathbb{E}[\sum_{t=1}^{\tau_\pi} g_k(\boldsymbol{x}_t)] \tag{40a}$$

$$\le \frac{\nabla h_k(\mu^1) - \mathbb{E}[\nabla h_k(\mu^{\tau_\pi+1})]}{\eta} + \mathbb{E}[T-\tau_A]\alpha_k b_k \tag{40b}$$

$$\le \left(\frac{\|\nabla h(\lambda^1)\|_\infty + C_h}{\eta}\right)\frac{\underline{G} + \alpha_k b_k}{\underline{G}} + \frac{\alpha_k b_k \bar{g}}{\underline{G}}, \tag{40c}$$

where the second inequality comes from Proposition 2.