Whole Page Optimization with Global Constraints

Weicong Ding
Amazon.com
Seattle, USA
weicding@amazon.com

Dinesh Govindaraj
Amazon.com
Seattle, USA
dgovind@amazon.com

S V N Vishwanathan
Amazon.com
Seattle, USA
vishy@amazon.com

ABSTRACT

The Amazon video homepage is the primary gateway for customers looking to explore the large collection of content, and finding something interesting to watch. Typically, the page is personalized for a customer, and consists of a series of widgets or carousels, with each widget containing multiple items (e.g., movies, TV shows etc). Ranking the widgets needs to maximize relevance, and maintain diversity, while simultaneously satisfying business constraints. We present the first unified framework for dealing with relevance, diversity, and business constraints simultaneously. Towards this end, we derive a novel primal-dual algorithm which incorporates local diversity constraints as well as global business constraints for whole page optimization. Through extensive offline experiments and an online A/B test, we show that our proposed method achieves significantly higher user engagement compared to existing methods, while also simultaneously satisfying business constraints. For instance, in an online A/B test, our framework improved key metrics such as customer streaming minutes by 0.77% and customer distinct streaming days by 0.32% over a state-of-the-art submodular diversity model.

ACM Reference Format:
Weicong Ding, Dinesh Govindaraj, and S V N Vishwanathan. 2019. Whole Page Optimization with Global Constraints. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330675

1 INTRODUCTION

Video services like Netflix, Amazon Video, or Youtube offer a vast and diverse selection of digital content for consumption. Users typically start from the home page to explore and find something interesting to watch. Since, different users have different tastes, the home page is usually personalized. One effective strategy for personalization is to group content logically, and present them as carousels or widgets to the user. For instance, documentaries may be grouped in a widget, and trending TV shows can appear together in another widget. Grouping content into widgets and providing interpretable labels for them is an interesting problem with existing literature [23], but is not the focus of this paper. We assume that a large selection of widgets is available for display on the homepage presented to a user, and we focus on the problem of whole page optimization by selecting and ranking of the widgets.

Most widgets to display on homepage are aimed at improving user engagement, but some widgets are used to create awareness of new type of digital content, and upcoming blockbuster shows, or to meet other business objectives such as promoting certain kinds of content. It is also well known (see e.g., [7, 14, 18, 26–28]) that user engagement increases when they are shown a diverse selection of content. In summary, widget ranking entails satisfying three, somewhat conflicting requirements: 1) **Relevance** – present relevant widgets higher in the page to maximize engagement, 2) **Diversity** – compose the page with diverse set of widgets and 3) **Constraints** – satisfy any business requirements or constraints. There is a vast literature on relevance ranking [2, 11, 12, 15, 17], and somewhat less has been written about diversity and whole-page effects [7, 10, 14, 18, 20, 26–28], and even fewer papers address enforcing business constraints [1, 8, 9, 16, 19, 29]. However, to the best of our knowledge, there is no existing work that focuses on satisfying all three constraints simultaneously within a unified framework. In this paper, we present a *unified framework* for dealing with relevance, diversity, and business constraints simultaneously. Before we proceed, we very briefly review existing literature on these three aspects and position our contributions.

The simplest way to optimize for user engagement is to predict point-wise relevance ranking scores, and sort widgets according to this score. Towards this end, much of existing research assumes that user interaction with a widget is independent of other widgets on the page [2, 11, 12, 15, 17]. Furthermore, proxies such as click-through rate (CTR) or conversion rate (CVR) are used for computing relevance scores. Unfortunately, estimating relevance scores independent of other content leads to a sub-optimal user experience where similar contents are presented to user that leads to monotony and decrease in user satisfaction [10, 14, 20, 26]. Recent studies have confirmed that the interactions between widgets impact the user conversion rate[1, 14, 18, 27]. The other end of the spectrum is to score every single page layout, and select the best [6, 18, 27] which is the globally optimal strategy. Unfortunately, this requires evaluating a combinatorially large number of layouts, which is rarely, if ever, possible for a high throughput production service. To approximate the combinatorial space, efficient diversity models have been adopted in a number of applications due to its statistical and computational efficiency [7, 10, 14, 20, 26]. Diversity models aims to select a subset of widgets that are diverse and maximize the relevance. These models capture the whole-page effect through the *number of similar widgets on the page.*

Another important problem in widget ranking is the trade-off between relevance versus discovery, which in turn manifests itself as the balance between short term engagement vs long-term satisfaction. If we display the most relevant content to the user, they will

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330675

engage with it, and this will drive up short-term metrics. However, if users do not discover new and interesting types of content, they will quickly get bored with the service, and long-term business growth will stall [1, 9, 29]. For example, consider Amazon Video, where users can watch multiple different types of video contents: a) Subscription Video On-Demand (SVOD) content that is included and has no additional cost for Prime subscribed users, b) Transactional Video On-Demand (TVOD) that is available for purchase or rental, and c) Third Party Channels (3P) that users can subscribe to, by paying a monthly fee. To facilitate discovery, we need to display all three categories to customers. However, these multiple product lines are competing for the same user budget (either money or entertainment time spent), and their success metrics are often at-odds with each other. One simple solution to the problem is to use *local* mechanisms such as *slotting*, which allocates certain fraction of slots to each category for each homepage impression. While certainly attractive, due to its simplicity and ease of implementation, slotting often leads to a sub-optimal user experience. For instance, consider a user who never purchases TVOD content and another who subscribes to many 3P channels. Slotting wastes real-estate in the case of the first user and is missing an opportunity in the case of the second, while also decreasing user satisfaction.

In this paper, we study the problem of whole page optimization with global constraints of the type that require a certain number of impressions for certain widgets across all home page impressions. Our key insight is that instead of enforcing the business constraints on a per-user basis, one can enforce them in expectation, globally across all users. To consider a grossly simplified example, suppose we have 20 users and 10 of them engage heavily with SVOD content, and 10 of them engage primarily with TVOD content. Moreover, suppose the global constraints specify 20 SVOD impressions and 20 TVOD impression. A slotting model would have displayed 1 TVOD and 1 SVOD widget per user. Instead, our algorithm will try to display 2 TVOD widgets to the TVOD users and 2 SVOD widgets to the SVOD users, thereby both increasing user satisfaction while also satisfying global constraints. Similar approaches have been developed in the special cases either when there is only one widget to display [8, 16, 19] or assuming there is no interactions among widgets [1, 29].

To formalize the above intuition, we learn the trade-off between the page reward and the regret for constraints in the dual optimization formulation [1, 19]. We develop novel prime-dual algorithm that decoupled online page-composition from offline dual-optimization. The resulting on-line composition module inherits the same sub-modular property as the constraint-free counterpart. This enables the greedy approximation suitable for production implementation. Summing up, we propose a holistic framework for *constrained whole page optimization* to tackle both whole-page diversity and global constraints. We conducted extensive offline analysis and implemented A/B testing. Our model achieved 25% higher page reward and satisfied impression constrained compared to slotting approach. Online A/B testing improved our key metrics such as minutes watched by 0.77% and customer distinct streaming days by 0.32% over state-of-the-art sub-modular diversity model.

Figure 1: Sample Amazon Video homepage with top-5 widgets. The "Prime popular movies" are SVOD content, "rent or buy new releases movies" is TVOD content, and "Starz channel movies" is a 3P channel, available for subscription. Our goal is to optimize and personalize the vertical composition of all widget.

2 PROBLEM FORMULATION

To help customer browse and discover digital products, the Amazon Prime Video homepage consists of rows, where each row contains a collection of digital contents; the rows are called *carousels* or *widgets*. See Figure 1 for an illustration. Each row is designed based on a single strategy such as popularity, genre, theme or personalized recommendations. Moreover, the content in each widget has the same content type (e.g., TV show or movie) and the same purchasing option (e.g., prime subscription required, or transactional video on demand). We define *product category* (or simply "category") of a widget as a combination of content type and purchasing option.[1] We will assume that for a given user visiting Prime Video, we have access to several hundred widgets that are of interest, and the problem we will deal with in this paper is how to select the top-n rows of widgets for each homepage.

Notations: We will use the calligraphic script, *e.g.*, $\mathcal{A}$ to denote a set and $|\cdot|$ to indicate its size. If the set is ordered, then $\mathcal{A}_n$ is used to denote its first n elements. We will use boldface letters, *e.g.*, $\mathbf{x}$ for vectors and $\langle \cdot, \cdot \rangle$ to denote the Euclidean inner product.

Every time we need to render a homepage, we are provided with a page context $\mathbf{u}$ that encapsulates customer information, device type, and time of day. $\mathcal{U}$ denotes the universe of all page context. Every widget that needs to be ranked is also represented by a widget context $\mathbf{c}$. $\mathbf{c}$ encapsulates information such as widget-id, the content in the row, and product category meta-data. Given a page context $\mathbf{u}$, we denote by $C^{\mathbf{u}}$ the set of available widgets to display and denote by $C_n^{\mathbf{u}}$ the top-n widgets. We use $\mathbf{c}_i^{\mathbf{u}}$ to represent the widget in the i-th slot of the page. We assume a feature function $\boldsymbol{\phi}(\mathbf{c}, \mathbf{u})$ is given to us which takes the page and widget context and produces a feature vector. The design of $\boldsymbol{\phi}(\cdot, \cdot)$ is application and business

[1]For instance, the category defined by 'prime subscription' and 'movie' contains widgets like 'Prime Recommended Movies', 'Prime top-rated Movies', 'Prime Drama Movies'. In Fig 1, the 3rd row 'Prime Popular Movies' falls into this category.

specific, and our proposed approach is agnostic to the choice of the feature functions used.

We assume each widget belongs to one of the m distinct product categories and denote this mapping by an indicator function $I(\mathbf{c}) \in \{0,1\}^m$. We will slightly abuse the notation, and define $I(C_n^{\mathbf{u}}) = \sum_{i=1}^{n} I\left(\mathbf{c}_i^{\mathbf{u}}\right)$. In other words, the k-th dimension of $I(C_n^{\mathbf{u}}) \in \mathbb{R}^m$ counts how many times the k-th product category appeared in the first n-slots of the ordered set $C_n^{\mathbf{u}}$.

2.1 Joint Page Optimization with Global Constraints

Suppose we have access to a scoring function $f(C_n^{\mathbf{u}}|\mathbf{u},\mathbf{w})$ that, given an page context $\mathbf{u}$ returns the utility of any ordered subset, $C_n^{\mathbf{u}}$, of the candidate widget set $C^{\mathbf{u}}$. $\mathbf{w}$ are the model parameters of f. Also assume that we know ahead of time all the homepage contexts $\mathcal{U}^t$ to be rendered during a time period t. The unit of time period can be a hour, a day, or a week, etc. We impose *global constraints* as functions of all the $|\mathcal{U}^t|$ pages with a separable form $\sum_{\mathbf{u}\in\mathcal{U}} \mathbf{g}(C_n^{\mathbf{u}}) \leq \mathbf{0}$ for some function $\mathbf{g}(\cdot)$ to be specified later. The inequality is element-wise. With these notations in place, our goal is to jointly optimize the $|\mathcal{U}^t|$ pages,

$$\underset{\{C_n^{\mathbf{u}}:\mathbf{u}\in\mathcal{U}^t\}}{\operatorname{argmax}} \sum_{\mathbf{u}\in\mathcal{U}^t} f(C_n^{\mathbf{u}}|\mathbf{u},\mathbf{w}) \tag{1a}$$

$$\text{s.t.} \sum_{\mathbf{u}\in\mathcal{U}^t} \mathbf{g}(C_n^{\mathbf{u}}) \leq \mathbf{0} \tag{1b}$$

such that not only is the utility maximized but also the constraints are satisfied. We model the widget *relevance* and the *whole-page diversity* aspects in $f(\cdot|\mathbf{u},\mathbf{w})$, and account for the overall *business requirements* through enforcing $\mathbf{g}(\cdot)$.

2.2 Related Works

We connect two different threads of prior art to our problem statement in Eq (1a, 1b) and show how they can be viewed as special cases of Eq. (1a, 1b). In various applications such as widget selection [1, 5, 24], sponsored advertisement [8, 19, 21], information retrieval [13, 14], and multi-arm bandits [3, 22, 25, 29], global constraints have been used to model business targets, monetary budgets, risk capacities, etc. Most of the prior art optimize the *modular objective function* $f(\cdot)$ which can be viewed as assuming $n = 1$ or no widgets interactions in $f(\cdot)$ in Eq. (1a). These formulation models relevance through linear function and incorporates either local or global constraints. These prior art fails to capture the whole-page interactions or diversity of contents.

Another line of research on whole-page optimization attempts to design page objective function $f(\cdot|\mathbf{u},\mathbf{w})$ to model whole-page widgets interactions. One widely adopted approach is to incorporate widget or item diversity into $f(\cdot)$ [4, 7, 20, 26, 28]. These approaches optimize a *submodular function* that measures the diversity degree of a collection of items. Recent work in [18, 27] explicitly included pair-wise widget interactions as features in $f(\cdot)$. These approaches formulate the problem as *multivariate optimization* to compose the page holistically. All these studies, however, ignore constraints $\mathbf{g}(\cdot)$ in Eq. (1b). In this case, the problem in Eq. (1a) reduces to optimizing $|\mathcal{U}^t|$ number of pages independently.

We also note that in contrast to Eq. (1a, 1b), one can formulate the constraints as penalty term in the loss function when learning $f(\cdot)$ [9]. The key difference is that the model parameters $\mathbf{w}$ are trained to optimize a mixed objective of both predicting page scores and meeting the global constraints. Therefore, the scoring function need to be trained from scratch whenever the global targets $\mathbf{g}(\cdot)$ change. In contrast, in Eq. 1, $\mathbf{w}$ is only optimized for approximating the targeted page values. Therefore, we can set different business targets $\mathbf{g}(\cdot)$ for different time periods t without re-training the page reward model $f(\cdot)$.

2.3 Our Contribution and Approach Overview

In this paper, we model $f(\cdot|\mathbf{u},\mathbf{w})$ as a sub-modular function and developed a prime-dual algorithm for handling global constraints $\mathbf{g}(\cdot)$. We explicitly incorporate sub-modular utilities as features in $f(\cdot|\mathbf{u},\mathbf{w})$ to promote whole-page diversity. We define the diversity along the dimension of widget categories but the proposed frame is generic to the diversity dimensions. We use $f(\cdot|\mathbf{u},\mathbf{w})$ to model the value of a page and learn the model parameters from homepage and customer streaming data. To solve Eq. 1, we develop a primal-dual algorithm that decouples the online page-composition from offline dual optimization. By assuming a smooth change in the distribution of $\mathcal{U}^t$ across different t's, the optimal dual variable learned from the previous time period $t-1$, i.e., $\mathcal{U}^{t-1}$, could be used for the online composition in the upcoming time period t. As such, we approximately solved Eq. (1a, 1b) without unrealistically knowing all $\mathcal{U}^t$ ahead of time.

For the rest of the paper, we discuss our model-based diversity approach for $f(\cdot|\mathbf{u},\mathbf{w})$ in Section 3. We then discuss the primal-dual approach for the joint page optimization in Section 4. We present offline analysis against various baselines and other whole-page models in Section 5, and summarize online A/B testing results in Section 6.

3 MODEL-BASED SUB-MODULAR DIVERSITY

We begin by illustrating how one can design a scoring model of a page taking widget diversity into consideration, and learn the model from data. In this section, since only a single page with a fixed context $\mathbf{u}$ is considered, we drop the superscript $\mathbf{u}$ in $C_n^{\mathbf{u}}$ for simplicity and represent a page as C_n.

Given a page C_n, the simplest way to predict its value (to be defined later) is one that decomposes into *widget-independent* components: $f(C_n|\mathbf{u},\mathbf{w}) = \sum_{i=1}^{n} f(\mathbf{c}_i|\mathbf{u},\mathbf{w})$, hence the widget in each slot is scored independently of all other widgets. To take into account interactions between widgets, we model the value of the whole page as:

$$f(C_n|\mathbf{u},\mathbf{w}) = \sum_{i=1}^{n} f(\mathbf{c}_i|C_{i-1},\mathbf{u},\mathbf{w}) \tag{2}$$

In other words, the value or utility of the i-th item in the list depends on the user context and the previous $i-1$ items. By modeling the utility of $\mathbf{c}_i$ as function of the $i-1$ widget displayed above the i-th slots, we are implicitly assuming that the customers scan the home-page from top to bottom.

Now we turn our attention to the design of the scoring function $f()$. Intuitively, we want $f(\mathbf{c}_i|C_{i-1},\mathbf{u},\mathbf{w})$ to predict the utility of

a widget, for instance, by modeling conversion probability, minutes streamed, or revenue generated. For simplicity we consider the binary conversion as page/widget value in this paper, and let y_i denote a random variable that $y_i = 1$ if a customer stream any piece of content in a widget and 0 otherwise. We will define $f(\mathbf{c}_i|C_{i-1}, \mathbf{u}, \mathbf{w}) := Pr(y = 1|\mathbf{c}_i, \mathbf{u}, \mathbf{w})$ and formulate the problem in a regression framework so that $f(\mathbf{c}_i|C_{i-1}, \mathbf{u}, \mathbf{w}) \approx y_i$. We point out that during a homepage visit, customers can stream from multiple widgets so multiple y_i's from the same page can be 1.

Scoring Model with widget interaction features

We posit a simple linear model for $f(\cdot)$ where $f(\mathbf{c}_i|C_{i-1}, \mathbf{u}, \mathbf{w}) = \mathbf{x}^\top \mathbf{w}$ and design features $\mathbf{x}$ to model both widget relevance as well as widget interactions. Recall that we have access to a feature vector $\boldsymbol{\phi}(\mathbf{c}, \mathbf{u})$ for contextual widgets relevance, we will focus on the interaction/diversity features.

We model the widget interactions via promoting diversity, and define diversity with respect to the widget product categories. Recall that $I(C_n) \in \mathbb{R}^m$ counts the number of widgets in each of the m categories in the top-n rows, one can promote the diversity of C_n by maximizing non-decreasing and sub-modular [4, 7, 26, 28] function $\mathbf{1}^\top \boldsymbol{\rho}(I(C_n))$. For each widget, we consider the incremental diversity gains induced by $\boldsymbol{\rho}$ and create features $\Delta\boldsymbol{\rho}(\mathbf{c}_i, C_{i-1}) = \boldsymbol{\rho}(I(C_i)) - \boldsymbol{\rho}(I(C_{i-1}))$. We design $\boldsymbol{\rho}(\cdot)$ as a function that applies the one dimensional sub-modular function $\rho(\cdot)$ to each component of the input vector with $\rho(n+1) - \rho(n)$ diminishing as n increases. This captures the intuition that the incremental utility of displaying a widget $\mathbf{c}_i$ diminishes if more widgets of the same category have been displayed in C_{i-1}. We set $\rho(x) = \log(1 + x)$ following suggestions in [4, 7, 26].

We concatenate the widget features $\boldsymbol{\phi}$ and the diversity features $\Delta\boldsymbol{\rho}$ to form the linear features for $f(\mathbf{c}_i|C_{i-1}, \mathbf{u}, \mathbf{w})$. For the diversity features, we use their interaction with context features denoted by $\mathbf{u} \times \Delta\boldsymbol{\rho}$. Overall, our scoring model f is

$$f(\mathbf{c}_i|C_{i-1}, \mathbf{u}, \mathbf{w}) = \left[\boldsymbol{\phi}(\mathbf{c}_i, \mathbf{u})^\top, \mathbf{u} \times \Delta\boldsymbol{\rho}(\mathbf{c}_i, C_{i-1})^\top\right] \mathbf{w} \quad (3)$$

where $[\cdot, \cdot]$ represents feature concatenation. At a high level, Eq (3) models the score of a widget as function of the individual relevance (contribution from $\boldsymbol{\phi}$) as well as interactions with widgets shown above the widget (contribution from features $\mathbf{u} \times \Delta\boldsymbol{\rho}$). Note that the dependence of $\Delta\boldsymbol{\rho}$ on C_{i-1} is fully captured through $I(C_{i-1})$ in Eq (3). In another word, the widgets interactions are modeled through the accumulative counts of widgets for different categories.

Model Based Diversity: Eq. (3) can be viewed as model based approach for the standard diversity recommendation literature. To better interpret this, note that $\sum_{i=1}^{n} \Delta\boldsymbol{\rho}(\mathbf{c}_i, C_{i-1}) = \boldsymbol{\rho}(I(C_n))$, and the interaction $\mathbf{u} \times \Delta\boldsymbol{\rho}$ is linear in $\Delta\boldsymbol{\rho}$, therefore, $\sum_{i=1}^{n} \mathbf{u} \times \Delta\boldsymbol{\rho}(\mathbf{c}_i, C_{i-1}) = \mathbf{u} \times \boldsymbol{\rho}(I(C_n))$. We can sum up all the widget score model Eq. (2) and get the page level scoring model as,

$$f(C_n|\mathbf{u}, \mathbf{w}) = \left[\sum_i \boldsymbol{\phi}(\mathbf{c}_i, \mathbf{u})^\top, \mathbf{u} \times \boldsymbol{\rho}(I(C_n))^\top\right] \mathbf{w}$$

which is equivalent to the standard relevance-diversity formulation for selecting top-n items from a set [4, 7, 14, 26]. The subset of weights $\mathbf{w}$ corresponding to $\mathbf{u} \times \boldsymbol{\rho}$ can be viewed as context-aware diversity regularization weights that can be seen in [4, 7, 14, 26]. Having the context-aware $\mathbf{u} \times \boldsymbol{\rho}$ diversity features allows our model to promote different product category diversity for different customer segments and page-visit contexts. For instance, for customers who have firmly demonstrated no interest in some category, it is less appropriate to promote diversity in that dimension.

Probability Model: We now turn to state our probability model for y_i given the linear model in Eq (3). We posit the following likelihood probability

$$p(y_i|\mathbf{w}, \mathbf{c}_i, C_{i-1}) = \mathcal{N}(\left[\boldsymbol{\phi}^\top, \mathbf{u} \times \Delta\boldsymbol{\rho}^\top\right] \mathbf{w}), \beta_0^2) \quad (4)$$

with Gaussian prior $p(\mathbf{w}) = \Pi_d \mathcal{N}(w_d; \mu_0, \sigma_0^2)$ for $\mathbf{w}$. β_0, μ_0, and σ_0^2 are the hyper- parameters in our Bayesian formulation.

We learn the posterior distribution of $\mathbf{w}$ given all the page impression and corresponding customer activities for $i = 1, \ldots, n, \mathbf{u} \in \mathcal{U}$. The posterior mean and variance of each dimension of $\mathbf{w}$ are learned through the online message passing update algorithm [15, 17]. We point out that Eq. (3) is sub-modular only when the subset of $\mathbf{w}$ corresponding to $\mathbf{u} \times \boldsymbol{\rho}$ is non-negative. As we posit Gaussian prior on $\mathbf{w}$ and learn the model by fitting the actual data, the non-negativity is not guaranteed. One can ideally impose a non-negative prior on the model parameter. However, we have empirically found that the non-negativity holds with Gaussian prior when model is learned from real-world home page data.

4 ONLINE PAGE OPTIMIZATION WITH GLOBAL CONSTRAINTS

Now that we have learned model $f(C_n^{\mathbf{u}}|\mathbf{u}, \mathbf{w})$ to score a page, we turn to solving the joint page optimization problem in Eq. (1a, 1b) with global constants in this section. We restrict the constraint function $\mathbf{g}(\cdot)$ to be separable in widgets hence $\mathbf{g}(C_n^{\mathbf{u}}) = \sum_{i=1}^{n} \mathbf{g}(\mathbf{c}_i^{\mathbf{u}})$. As we shall discuss later, most of the constraints we are interested in have this format.

4.1 Impression Constraints

Recall that one of the major challenge in homepage optimization is to balance the impression of widgets from different categories. While displaying the most relevant widgets with highest conversion probability drives up short-term engagement, presenting diversified categories of widgets can help customers discover new content and engage for the long-term. Simply maximizing for immediate engagement (e.g., modeled by Eq (3)) often fails to achieve proper category balance. This is because the conversion nature of different product categories are distinct. For instance, the frequency of customer starting a new TV season is much less than a movie, and it is even rare for customer to subscribe to a new thirdparty channel. We proposed to use global impression constraints to achieve balanced exposure of different categories. The key assumption is that if we achieve the same widget impression balance aggregated across all customers and contexts, the targeted downstream business targets would be properly maintained. Prior arts [1, 8, 19, 21, 24] proposed to adopt similar constraints in presence of a number of distinct categories of items. Formally, recall that $\mathcal{U}^t$ denotes the set of homepage contexts and recall that $I(C_n^{\mathbf{u}}) \in \mathbb{R}^m$ counts the number of widgets in the top n-slots in each category. We define our global

impression constraints as,

$$\frac{1}{|\mathcal{U}^t|}\sum_{u\in\mathcal{U}^t}\left(\mathbf{b}^t - I\left(C_n^{\mathbf{u}}\right)\right) = \mathbf{b}^t - \frac{1}{|\mathcal{U}^t|}\sum_{u\in\mathcal{U}^t} I\left(C_n^{\mathbf{u}}\right) \le \mathbf{0} \quad (5)$$

Here $\mathbf{b}^t$ represents the minimum number of widgets from each category to impress across all contexts. It defines the targeted balance among categories. We assumed $\mathbf{b}^t$'s are given for each time period t. Note that to satisfy this average balance, each individual pages can deviate from $\mathbf{b}^t$. In other word, our formulation allows the degree-of-freedom to distribute the desired impressions to the most relevant customers and context within $\mathcal{U}^t$.

Now we plug in Eq. (5) and Eq. (2) to the general formulation in Eq. (1), and restate our optimization problem as,

$$\min_{\mathbf{c}_i^{\mathbf{u}},\mathbf{u}\in\mathcal{U}^t,i=1,\ldots,n} \quad -\sum_{\mathbf{u}\in\mathcal{U}^t}\sum_{i=1}^{n} f\left(\mathbf{c}_i^{\mathbf{u}}|C_{i-1}^{\mathbf{u}},\mathbf{u},\mathbf{w}\right) \quad (6)$$

$$s.t., \quad \mathbf{b}^t - \frac{1}{|\mathcal{U}^t|}\sum_{u\in\mathcal{U}^t}\sum_{i=1}^{n} I\left(\mathbf{c}_i^{\mathbf{u}}\right) \le \mathbf{0} \quad (7)$$

Alternative Approaches: *Slotting* and *Calibration* are two commonly adopted alternatives to global constraints to achieve category balance in real-world applications. The slotting method dedicates certain row positions for widgets from each product category for every homepage. Standard widget relevance predictions can be used to rank widgets within each category. While this can perfectly achieve the impression balance, the one-size-fits-all approach is often sub-optimal. The Calibration method resorts to quantifying the value of a single conversion in each category using the same currency unit. One can then combine the conversion probability with the calibrated value to score each widgets. While this approach is straightforward, the values for different categories are often derived from long-term regression models and it is difficult to react to the dynamically changing content quality and customer distribution.
Other Constraints: Similar to Eq (7), we can define expected conversion constraints as $\mathbf{b} - \sum_{\mathbf{u}}\sum_n p\left(\mathbf{c}_i^{\mathbf{u}}\right) I\left(\mathbf{c}_i^{\mathbf{u}}\right) \le \mathbf{0}$ where $p(\cdot)$ is the widget conversion probability. We also note that we used the same category mapping $I\left(C_n^{\mathbf{u}}\right)$ for both diversity and constraint functions. One could use different dimensions and our proposed algorithm is agnostic to such choices as well.

We next focus on two main challenges that prevent one from directly applying Eq (6) and (7) for rendering homepages in production: 1) to decouple the joint page optimization to individual problems and 2) to optimize without knowing $\mathcal{U}^t$ ahead of time.

4.2 Primal-Dual Algorithm

Since the algorithm needs to render a homepage whenever a page context $\mathbf{u}$ is requested, we need to decouple the proposed joint optimization problem into individual page optimization tasks. Primal-dual formulation provides a tool for such decoupling. Note that Eq (6) and (7) is equivalent to its dual-form,

$$\max_{\boldsymbol{\xi}\ge\mathbf{0}}\min_{\{\mathbf{c}_i^{\mathbf{u}}\}} \quad -\sum_{\mathbf{u}\in\mathcal{U}^t}\left\{\sum_i f(\mathbf{c}_i^{\mathbf{u}}|C_{i-1},\mathbf{u},\mathbf{w}) + \boldsymbol{\xi}^\top(\mathbf{b} - \sum_i I\left(\mathbf{c}_i^{\mathbf{u}}\right))\right\} \quad (8)$$

Here $\boldsymbol{\xi}\in\mathbb{R}^m$ are the dual variables. The primal-dual approach solves this min-max dual problem by alternating between: a) assuming dual variables $\boldsymbol{\xi}$ are fixed, update the optimal pages $C_n^{\mathbf{u}}$ to minimize the inner layer problem; b) fix the pages $C_n^{\mathbf{u}}$, find the best dual variables $\boldsymbol{\xi}$ to maximize the outer layer optimization task. The key benefit of adopting this framework is the fact that given dual variables $\boldsymbol{\xi}\ge\mathbf{0}$, each of the pages $C_n^{\mathbf{u}}$ can be optimized *independently* as,

$$\arg\max_{\mathbf{c}_i^{\mathbf{u}}}\sum_{i=1}^{n} f(\mathbf{c}_i^{\mathbf{u}}|C_{i-1},\mathbf{u},\mathbf{w}) + \boldsymbol{\xi}^\top I(\mathbf{c}_i^{\mathbf{u}}) \quad (9)$$

One can view Eq (9) as regularized page value of the non-constrained version in Eq (2) where the dual variables $\boldsymbol{\xi}$'s control the trade-off between maximizing the expected value of a page and the need to meet the global balance. We will discuss how Eq (9) can be solved efficiently in Section 4.2 with greedy approximation thanks to the sub-modular property of $f(\cdot)$.

On the other hand, optimizing $\boldsymbol{\xi}$ given $C_n^{\mathbf{u}}$'s in the dual formulation Eq. (8) can be solved by gradient updates. We add a ℓ_2 regularization term $\gamma\|\boldsymbol{\xi}\|^2$ for some γ_j in Eq (8) to prevent $\boldsymbol{\xi}$ from being too large. This technical modification has been adopted in similar works in [8, 16, 19]. The gradient update for $\boldsymbol{\xi}$ is given in Algorithm 1 (line 6) with gradient stepsize $\eta\ge 0$ and $[x]_+ = \max(x, 0)$ being applied element-wise. Recall that $\sum_{\mathbf{u}\in\mathcal{U}^t}(\mathbf{b}^t - \sum_i I(\mathbf{c}_i^{\mathbf{u}}))$ measures the miss in satisfying the impression target, the gradient update can be interpreted as attempting to increase the corresponding weight in $\boldsymbol{\xi}$ if some categories fail to meet the global targets.

Algorithm 1 Primal-dual algorithm

Require: Page model $f(\cdot|\cdot,\mathbf{u},\mathbf{w})$; Page context $\mathcal{U}_n^t$; Impression target $\mathbf{b}^t$; Stepsize η_j; Stopping threshold ϵ; Max Iteration J; Initial duals $\boldsymbol{\xi}\ge\mathbf{0}$;
Ensure: Dual variables $\boldsymbol{\xi}\in\mathbb{R}^m$, Optimal Pages $C_n^{\mathbf{u}}, \mathbf{u}\in\mathcal{U}^t$;
1: **for** each iteration $j = 1,\ldots,J$ **do**
2: **for** each page i **do**
3: $C_n^{\mathbf{u}} \leftarrow \arg\max_{\mathbf{c}_i^{\mathbf{u}}}\sum_{i=1}^{n} f(\mathbf{c}_i^{\mathbf{u}}|C_{i-1}^{\mathbf{u}},\mathbf{u},\mathbf{w}) + \boldsymbol{\xi}^\top I(\mathbf{c}_i^{\mathbf{u}})$ (see Algorithm 2)
4: **end for**
5: **for** m=1,..., M **do**
6: $\boldsymbol{\xi} \leftarrow \left[\boldsymbol{\xi} + \eta_j\frac{1}{|\mathcal{U}|}\sum_{\mathbf{u}\in\mathcal{U}}(\mathbf{b} - \sum_i I(\mathbf{c}_i^{\mathbf{u}})) - \eta_j\gamma_j\boldsymbol{\xi}\right]_+$
7: **end for**
8: If $\frac{1}{|\mathcal{U}^t|}\sum_{\mathbf{u}}\mathbf{1}^\top\left[\mathbf{b} - \sum_i I(\mathbf{c}_i^{\mathbf{u}})\right]_+ \le \epsilon$, Break;
9: **end for**

We summarize the primal-dual approach in Algorithm 1. For each iteration, we first find the best pages and then apply gradient update to the dual variables. We use the miss in global constraints as stopping criteria. Since $|\mathcal{U}^t|$ is typically very large, for numerical stability we re-normalized η and γ by $|\mathcal{U}^t|$ in Alg. 1. To be more specific, we use diminishing step-size η_j and γ_j in the algorithm. We will discuss these details later in the offline experiment section. We point out Alg. 1 does not solve the challenge of rendering homepages online since it assumes knowing all the $\mathcal{U}^t$ ahead of time period t. The decoupling of the page optimization task still requires the knowledge of $\boldsymbol{\xi}$. We next focus on how we approximate the dual variable from historical data. **On the convergence of Algorithm 1**: When $n = 1$ and $f()$ is linear, Algorithm 1 reduces to the algorithm in [1, 8, 13, 16, 19] with convergence and regret bounds established. With empirical evaluations, we found Algorithm 1 can converge to desired stationary point (see Section 5) but defer the analysis to future work.

We next focus on the challenge that the universe of contexts $\mathcal{U}^t$ is not known to us ahead of time. To address this, we used historical data to approximate the optimization of dual variables ξ. Let ξ^t be the optimal dual variable estimated from Algorithm 1 using $\mathcal{U}^t$. Since we can calculate the optimal ξ^{t-1} from $\mathcal{U}^{t-1}$ before the t-th time period, we can assume $\xi^t \approx \xi^{t-1}$. During the time period t, we could compose page for each context $\mathbf{u}$ independently by optimize Eq. (9) with $\xi = \xi^{t-1}$. Our approximation $\xi^t \approx \xi^{t-1}$ implicitly assumed a smooth transition between the context distribution of $\mathcal{U}^{t-1}$ and $\mathcal{U}^t$, and $\mathbf{b}^{t-1} \approx \mathbf{b}^t$. Similar strategies have been suggested in [1, 13, 16]. In Section 5, we empirically validated this assumption and showed that offline dual approximation can achieve the expected target balances with very small deviation.

4.3 Greedy Page Composition

We next discuss how to solve the regularized page optimization problem in Eq (9). Note that $f(C_n)$ is submodular and $\xi^\top I(C_n)$ is linear, the overall objective in Eq (9) is sub-modular. We therefore apply the greedy algorithm to approximately solve it. The detail of the greedy steps are summarized in Algorithm 2. The greedy

Algorithm 2 Greedy Page Composition with Dual Variables

Require: Page size n; Candidate Widgets C; Page context $\mathbf{u}$; Constraints $\mathbf{b}$. Pre-estimated Dual Variables ξ;
Ensure: Page $C_n^{\mathbf{u}}$ (i.e., $\mathbf{c}_i^{\mathbf{u}}, i = 1, \ldots, n$)
1: Sample page scoring model $\mathbf{w}$ from learned posterior distribution;
2: $C_0^{\mathbf{u}} = \emptyset$;
3: **for** $i = 1, \ldots, n$ **do**
4: bestscore $= 0, a^* = null$
5: **for** $a \in C$ **do**
6: $s_a = f(a|C_{i-1}^{\mathbf{u}}, \mathbf{u}, \mathbf{w}) + \xi^\top I(a)$
7: If $s_a >$ bestscore, $bestscore \leftarrow s_a$, $a^* \leftarrow a$
8: **end for**
9: $\mathbf{c}_i^{\mathbf{u}} \leftarrow a^*$
10: $C \leftarrow C \setminus \{a^*\}$
11: **end for**

solution of Algorithm 2 is no less than $1-1/e$ of the optimal solution Eq (9). We note that more general constraints can be solve similarly if $\mathbf{g}(C_n)$ is submodular. We also note that when $\xi = \mathbf{0}$, Algorithm 2 reduces to the constraint-free special case which has been widely used for diversified recommendation tasks [4, 7, 26, 28]. To conclude this section, there are overall three modules of our system: 1) a training module that learns the page scoring function $f(\cdot|\cdot, \mathbf{w})$; 2) a dual-optimal module that estimate the optimal dual-variables (Algorithm. 1) from data $\mathcal{U}^{t-1}$ and $\mathbf{b}^{t-1}$; and 3) the online page optimization module (Algorithm. 2) that renders the homepage requests in $\mathcal{U}^t$ with dual variables ξ^{t-1}.

5 OFFLINE EXPERIMENTS

We conducted a number of offline experiments to demonstrate the key benefits of proposed framework. We collect two weeks' of Amazon Prime Video homepage request data in August 2018 for all the offline analysis. We consider "per day" as the unit for time periods $\mathcal{U}^t$. We incrementally update the page scoring model (Eq (3)) after collecting page impressions and customer interactions from the current day $\mathcal{U}^t$, and test the performance on the pages in next day $\mathcal{U}^{t+1}$. We aggregated metrics from only the second week. Dual variables ξ^t were estimated and applied in similar fashion (see Section 4.2). The whole page model Eq (3) was trained using a distributed variation of the online message passing algorithm [15]. We set the Bayesian hyper-parameters as $\beta_0 = 1.0, \mu_0 = 0.001$, and $\sigma_0^2 = 1.0$ in Eq (4).

Impression Constraints: We set the target impression balance $\mathbf{b}^t$ in Eq. (7) to replicate the average balance in current production model. It also ensures the feasibility of $\mathbf{b}^t$. Empirically, the variation of $\mathbf{b}^t$ for different days is insignificant and is mostly due to the difference between weekdays and weekend.

Evaluation Metrics: We measure the prediction accuracy and the homepage quality of whole page models using standard AUC and Precision@K. Precision@K is defined as the number of positive widgets in top-K slot of each homepage ($K \leq n$) and is then normalized by K. We measure the diversity degree of composed homepages using the standard pair-wise similarity (Div-pair@n) metrics [7]. Recall that $I(C_n) \in \mathbb{R}^m$ counts the number of widgets from each of the m-categories on each page C_n and let $\hat{I}(C_n)$ be its ℓ_1 normalized version. The diversity metrics is formally defined as, Div-pair@n $= 1 - \sum_{1 \leq s < t \leq N} I_s(C_n) I_t(C_n) / 0.5n(n-1)$ where I_s is the $s-th$ dimension of I. Intuitively, Div-pair@n measures the average pair-wise dis-similarity between all widgets on the page, higher value indicates more diverse homepages. To quantify the violation of the imposed constraints, i.e., the derivation from targeted product category balance $\mathbf{b}^t$, we calculate the percentage miss of all the constraints as

$$\text{miss@n} = \frac{1}{m} \sum_{s=1}^{m} \left[1 - \sum_{\mathbf{u} \in \mathcal{U}^t} I_s(C_n^{\mathbf{u}}) / b_s^t |\mathcal{U}^t| \right]_+ \tag{10}$$

We note that this has been used in Algorithm 1 as the stopping condition for the primal-dual optimization. We point out that we focused on the performance of the models and the quality of the homepages composed. We did not include standard multi-arm bandit feedback metrics such as regrets in this paper.

5.1 Validating the Primal-Dual Algorithm 1

Convergence: We first validate that the proposed Algorithm 1 can empirically converge to a stationary point as expected. To show this we take $|\mathcal{U}^t| = 1MM$ homepages randomly sampled from one day production pages and use the whole page scoring model parameters $\mathbf{w}$ after training with the same day's data. We then applied Alg. 1 on this collections of pages $|\mathcal{U}^t|$ and focused on the convergence of dual variables ξ^t. We initialize $\xi = \mathbf{0}$ and set a diminishing learning rate $\eta_j = 0.01/j, \gamma_j = 0.1/\sqrt{j}$. We set $\epsilon = 0.05$ and $J = 50$ as stopping conditions. To monitor the convergence, we report *miss@n* in Figure 2. We also illustrate the percentage loss and the dual variable value ξ_s of a particular constraint dimension s in Figure 2. As expected, *miss@n* converged to 0 and all the constraints are approximately satisfied after a number of iterations. The particular percentage loss of constraint dimension s also diminished and the corresponding dual variable ξ_s reached a stationary point. In sum, our alternative optimization approach in Algorithm 1 does converge to a feasible stationary point. In most

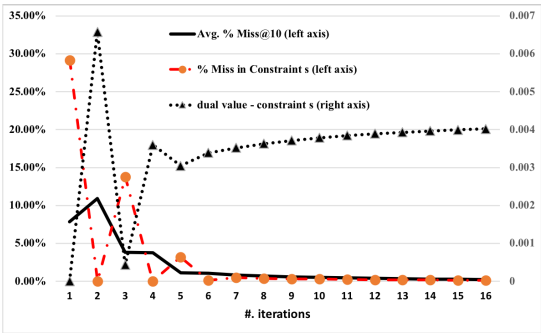

Figure 2: Convergence of Alternating Optimization Algorithm 1. The average percent miss of constraints after each iteration are reported as well as the percent miss and dual variable value of an example constraint (corresponding to category s). The figure is best visualized in color.

online and offline analysis, we found that our algorithm meets the stopping condition after 30 iterations.

Recall that we propose to learn ξ^{t-1} from $\mathcal{U}^{t-1}, \mathbf{b}^{t-1}$ and use that as approximation for composing pages from $\mathcal{U}^t$, we next validate how this approximation works in practice. Let each t be a day and scoring models are updated daily. We simulate this procedure on one week's evaluation data. We calculate the *miss@n* using ξ^{t-1} to predict and compose pages on $\mathcal{U}^t$ (will refer to this as testing miss@n) which is different from the curves shown in Figure 2. We report the testing *miss@n* in Figure 3 as a function of t (the solid-line curve). We also include the testing miss for the category with highest overall miss in the same Figure (the dash-line curve in Fig 3). We note that the average testing *miss@n* is very small and is suitable for online homepage composition task. This shows that the underlying distribution of $\mathcal{U}^t$ does change slowly w.r.t t. In Figure 3, day-4 and 5 are weekends, and the underlying customer/context distributions $\mathcal{U}^t$ are slightly different. Hence we observe an increase in the testing *miss@n* transiting from day-3 to 4 and from day-5 to 6. We note that the constraint with maximum testing deviation, as indicated in Figure 2, has a high deviation. This is due to the very small absolute average impression for that category.

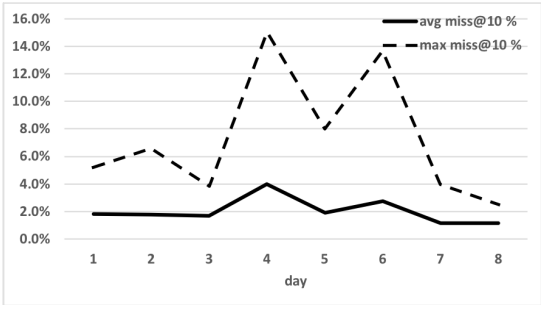

Figure 3: The average testing miss@N per hour and testing miss@n for the constraints with maximum miss. Day-4 and 5 are weekend.

5.2 Benefit of Incorporating Whole-Page Diversity Features

We next study the empirical impact of having the whole-page view in our model-based diversity approach. We ignore all the constraints in this section. Our primary baseline (CVR) is to compose homepage by ranking based on the CVR of each individual widget where CVR is predicted without the whole-page view. Baseline CVR model has no page-level features (only $\boldsymbol{\phi}$ in Eq (3)) and uses the Bayesian Linear Probit model [15, 17, 26]. The same customer, context and widget meta-data are used for both CVR and our model to ensure that the *only* difference between them is the context-aware diversity features (the additional $\mathbf{u} \times \boldsymbol{\rho}$ in Eq (3)). For our model, we do not consider any constraints in this experiment, we simply set $\xi^t = \mathbf{0}$ in Algorithm 2. We report the quality of models using AUC, Precision@K, as well as the diversity entropy metrics to quantify the page diversity degree. We also report *miss@n* metrics to quantify the deviation for constraint-free models. All metrics are reported in percentage difference compared to a baseline indicated by (B) in Table 1. As summarized in Table 1, AUC and Precision@K metrics

Table 1: Performance metric of our model without any global constraints (WPO-free) against various baselines. (B) indicates the baseline for percentage calculation. All numbers are reported in percentage lift w.r.t. baseline(s).

	CVR	WPO-Free	DIV	MVT
AUC	(B)	4.03%	0%	-
Precision@10	(B)	2.22%	-1.06%	0.56%
Precision@20	(B)	1.56%	1.26%	0.12%
Div-pair@20	(B)	5.91%	34.4%	17.3%
miss@10	10%	8.9%	0%	4.5%
miss@20	9.4%	10.0%	0%	5.5%

of our model outperformed the CVR baseline. This demonstrates that having a whole-page view can improve the prediction accuracy for each individual widget. We also observed that the entropy metric improves over the CVR baseline. Given a context, there is always a particular type of widget with higher conversion probability so simply ranking widgets leads to less diverse pages. In contrast, our model learns positive contribution from the diversity features which forces the greedy approach to compose diverse pages. But percentage constraint miss metrics is significantly large. Therefore **both** CVR and our constraint-free model requires additional treatment to meet the desired impression targets. To further explore the impact of other whole-page modeling choices, we compared our approach against the diversity based algorithm [26] (DIV) where the sub-modular diversity contribution of each dimension are manually tuned to meet constraints. We also compare against the state-of-the-art whole page model [18, 27] (MVT) where pair-wise widget interaction features are constructed for predicting conversion. The CVR probabilities are consumed by DIV algorithm for the joint relevance and diversity optimization. The AUC for DIV is defined based on relevance probabilities hence it is the same as CVR. For MVT, we only consider the interactions between slots within a window of 2 (i.e., the two widgets impressed before and below), and run the hill-climbing algorithm for 20 epochs for each page optimization as suggested in [18]. CVR is also used as features for

MVT. Note that MVT model is on page-level, so we did not report AUC since it is not comparable with other models whose AUC is calculated at widget level. As shown in Table 1, DIV achieved slightly lower Precision and the highest diversity degree. This is expected as DIV enforces high diversity on top of the conversion probabilities with the manually tuning of regularization weights. The precision metrics for MVT are similar to the CVR baseline. This could be attributed to the offline data collection bias since MVT tends to compose novel combinations of widgets that are much different from the other approaches. We note that MVT model also deviate from the targeted impression distribution (i.e., % difference 10 and 20). We did not select MVT for further online experiment since it is not straightforward to incorporate global constraints in the hill-climbing based optimization framework. Overall, our model achieved the best prediction accuracy among all the whole-page view models. We conclude this section by pointing out that the constraint-free model fails to achieve the global constraints. It is necessary to have constrained optimization to avoid changing the constrained overall balance. We next discuss the effect of our global constraints and its impact on the individual pages.

5.3 Performance of Global Constraints

We compared our global constraint approach (Global) against pages without any constraints (Free), and the slotting approach (Slot) which enforces the impression proportion at every page. To avoid the problem of determining the best slotting assignments on the page, we took the set of actual slotting assignments based on our production setting and global constraints are estimated accordingly. We report the Precision@K and miss@n for all the approaches. We also report the overall page reward value (the average of Eq (6) across all the pages) with the learned model parameters $\mathbf{w}$. We use this page-reward to quantify the quality of a page and the percentage of relevance loss by introducing constraints. [2] All the metrics are summarized in Table 2. The results are reported in percentage difference and baselines are indicated as (B) in table 2.

Table 2: Performance metric of global constraints optimization (WPO-Global) against pure model-based diversity (WPO-Free) and slotting/pinning approach (Slot). All metrics are reported in percentage lift w.r.t baseline(s) indicated as (B).

	WPO-Free	WPO-Global	Slot	DIV
Precision@10	(B)	-0.07%	-9.08%	-2.36%
Precision@20	(B)	-1.47%	0.77%	0.3%
Div-pair@20	(B)	17.6%	22.1%	19.3%
miss@10	15.9%	1.5%	0% (B)	0.4%
Page reward	(B)	-1.31%	-25.6%	-4.9%

As shown in Table 2, our Global constraint approach can achieve the average constraints closely (1.5% miss) with minimum impacts on the value of the each page (-1.31% decrease of the average page rewards and less than 1% drop in Precision@K's). In other words, the global constraints could achieve *most* of the page values of the optimal constraint-free page. In contrast, while the slotting based approach can perfectly achieve the product balance, it significantly decreased the overall page value (-25.6% decrease in page reward). This validates that the local constraints are sub-optimal. We note that page-level slotting has the highest pair-wise and logarithm entropy metrics since diversity is enforced on every page. To compare our approach against other whole-page models in the presence of global constraints, we choosed the DIV [26] approach since one can manually adjust the diversity weights for different categories to achieve the target impression balance $\mathbf{b}^t$ .Results are also listed in Table 2. We note that DIV with specifically calibrated diversity weights could achieve the target impression balance (with only 0.4% miss). Compared to the Slotting approach it achieves a higher page reward value and precision metrics. Compared to our approach, DIV is still sub-optimal in achieving page-level personalization with a lower page reward and precision metrics. We hypothesize this is due to the one-size-fits-all diversity regularization weights in the DIV weights hence the contextual difference is not taken into account. We did not include MVT in this experiment since it is not straightforward to enforce constraints within the hill-climbing based optimization approach.

Table 3: Coverage of a product category X by the customer streaming propensity in product category X.

Customer propensity for X	DIV	WPO-Global
0	-3.0%	-15.2%
low	1.0%	7.9%
moderate	15.1%	32.7%
high	18.1%	59.2%
extremely high	27.5%	67.7%

To further illustrate the impact of our global constraints, we report in Table 3 how we distribute the overall targeted impression of a particular category (X) for different customer segments. The percentage difference compared to the overall category (X) impressions (i.e., b_X) are reported in Table 3. As indicated in Table 3, our approach distributed the impression of the category (X) correctly to the most relevant customer segment for e.g., fewer category (X) widgets are shown to customers with 0 propensities. Note that customer propensity for category X is calculated as number of video streams by the customer in category X in last 28 days.

6 ONLINE EXPERIMENTS

We have implemented both submodular diversity algorithm (DIV) [26] and our whole-page model with global constraints (WPO-Global) in production and conducted A/B testing experiments. We manually tuned diversity model (DIV) to achieve the same product category coverage as existing production model. We derived impression constraints for WPO-Global from the same category balance. We evaluated improvement of DIV and WPO-Global on our customer engagement metrics: 1) overall streaming minutes and 2) Distinct Streaming Days (DSDs). The total streaming minutes reflects short-term customer activeness while DSDs represents the customer stickiness with the platform. DIV improved minutes by 0.44% and DSDs by 0.12% over the existing production model with

[2]The pinning assignment specified all the purchasing type requirement for each of the top-10 slots on the page. The CVR probabilities are used to select actual widgets. As an illustrating example, one can specify that all odd-number slots to display TV widgets and ranking algorithm should select the most relevance TV widget to display in the 1st row, the second relevance TV widget in the 3rd row, etc.

constraints were closely matched with no impact to any particular product categories. Our proposed WPO-Global approach further improved minutes by 0.77% and DSDs by 0.32% metrics on top of DIV model with a significant margin. At the same time, the global constraints have also achieved the business targets.

7 CONCLUSION

We present a primal-dual framework that encompasses relevance, diversity and global constraints for ranking widgets. We have empirical shown that our framework increases diversity of widgets and satisfy constraints but at the same time improve relevance. Extensive online A/B testing showed that our framework can indeed improve customer engagement without affecting any product categories.

REFERENCES

[1] Deepak Agarwal, Shaunak Chatterjee, Yang Yang, and Liang Zhang. 2015. Constrained optimization for homepage relevance. In *Proceedings of the 24th International Conference on World Wide Web*. 375–384.

[2] Deepak Agarwal, Bee-Chung Chen, and Pradheep Elango. 2009. Spatio-temporal models for estimating click-through rate. In *Proceedings of the 18th international conference on World wide web*. ACM, 21–30.

[3] Shipra Agrawal and Nikhil Devanur. 2016. Linear contextual bandits with knapsacks. In *Advances in Neural Information Processing Systems*. 3450–3458.

[4] Amr Ahmed, Choon Hui Teo, S.V.N. Vishwanathan, and Alex Smola. 2012. Fair and Balanced: Learning to Present News Stories. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. 333–342.

[5] K. Basu, S. Chatterjee, and A. Saha. 2016. Constrained Multi-Slot Optimization for Ranking Recommendations. *arXiv preprint arXiv:1602.04391* (2016).

[6] Kinjal Basu, Ankan Saha, and Shaunak Chatterjee. 2017. Large-Scale Quadratically Constrained Quadratic Program via Low-Discrepancy Sequences. In *Advances in Neural Information Processing Systems*. 2297–2307.

[7] Laming Chen, Guoxin Zhang, and Hanning Zhou. 2018. Improving the Diversity of Top-N Recommendation via Determinantal Point Process. In *Advances in Neural Information Processing Systems*.

[8] Ye Chen, Weiguo Liu, Jeonghee Yi, Anton Schwaighofer, and Tak W Yan. 2013. Query clustering based on bid landscape for sponsored search auction optimization. In *Proc.of the 19th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*. ACM, 1150–1158.

[9] K. Christakopoulou, J. Kawale, and A. Banerjee. 2017. Recommendation with Capacity Constraints. In *Proceedings of the 2017 ACM International Conference on Information and Knowledge Management*. 1439–1448.

[10] C. Clarke, M. Kolla, G. V Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st ACM SIGIR conference on Research and development in information retrieval*. ACM, 659–666.

[11] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 191–198.

[12] A. Das, M. Datar, A. Garg, and S. Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 271–280.

[13] Nikhil R Devanur, Jugal Garg, Ruta Mehta, Vijay V Vaziranb, and Sadra Yazdanbod. 2018. A New Class of Combinatorial Markets with Covering Constraints: Algorithms and Applications. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2311–2325.

[14] Nikhil R Devanur, Zhiyi Huang, Nitish Korula, Vahab S Mirrokni, and Qiqi Yan. 2016. Whole-page optimization and submodular welfare maximization with online bidders. *ACM Transactions on Economics and Computation (TEAC)* 4, 3 (2016), 14.

[15] Thore Graepel, Joaquin Quinonero Candela, Thomas Borchert, and Ralf Herbrich. 2010. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of International Conference on Machine Learning 2011*.

[16] R. Gupta, G. Liang, H. Tseng, Ravi K. Holur V., X. Chen, and R. Rosales. 2016. Email volume optimization at LinkedIn. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 97–106.

[17] X. He, J. Pan, Ou Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, and S. Bowers. 2014. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*. ACM, 1–9.

[18] Daniel N Hill, Houssam Nassif, Yi Liu, Anand Iyer, and SVN Vishwanathan. 2017. An Efficient Bandit Algorithm for Realtime Multivariate Optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1813–1821.

[19] Jim C Huang, Rodolphe Jenatton, and Cédric Archambeau. 2016. Online Dual Decomposition for Performance and Delivery-Based Distributed Ad Allocation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 117–126.

[20] Neil Hurley and Mi Zhang. 2011. Novelty and diversity in top-n recommendation–analysis and evaluation. *ACM Transactions on Internet Technology (TOIT)* 10, 4 (2011), 14.

[21] Rodolphe Jenatton, Jim Huang, and Cedric Archambeau. 2016. Adaptive Algorithms for Online Convex Optimization with Long-term Constraints. In *International Conference on Machine Learning*. 402–411.

[22] Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. 2012. Trading regret for efficiency: online convex optimization with long term constraints. *Journal of Machine Learning Research* 13, Sep (2012), 2503–2528.

[23] Georgina Peake and Jun Wang. 2018. Explanation Mining: Post Hoc Interpretability of Latent Factor Models for Recommendation Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. 2060–2069.

[24] Parikshit Shah, Akshay Soni, and Troy Chevalier. 2017. Online Ranking with Constraints: A Primal-Dual Algorithm and Applications to Web Traffic-Shaping. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 405–414.

[25] Wen Sun, Debadeepta Dey, and Ashish Kapoor. 2017. Safety-Aware Algorithms for Adversarial Contextual Bandit. In *International Conference on Machine Learning*. 3280–3288.

[26] Choon Hui Teo, Houssam Nassif, Daniel Hill, Sriram Srinivasan, Mitchell Goodman, Vijai Mohan, and SVN Vishwanathan. 2016. Adaptive, personalized diversity for visual discovery. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 35–38.

[27] Yue Wang, Dawei Yin, Luo Jie, Pengyuan Wang, Makoto Yamada, Yi Chang, and Qiaozhu Mei. 2016. Beyond ranking: Optimizing whole-page presentation. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*. ACM, 103–112.

[28] M. Wilhelm, A. Ramanathan, A. Bonomo, S. Jain, E. Chi, and J. Gillenwater. 2018. Practical Diversified Recommendations on YouTube with Determinantal Point Processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2165–2173.

[29] Hao Yu, Michael Neely, and Xiaohan Wei. 2017. Online Convex Optimization with Stochastic Constraints. In *Advances in Neural Information Processing Systems*. 1427–1437.