An Application of Causal Bandit to Content Optimization

Sameer Kanase, Yan Zhao, Shenghe Xu, Mitchell Goodman
Manohar Mandalapu, Benjamyn Ward, Chan Jeon, Shreya Kamath, Ben Cohen
Yujia Liu, Hengjia Zhang, Yannick Kimmel, Saad Khan, Brent Payne, Patricia Grao
{kanases,yzhaoai,shenghe,migood}@amazon.com
Amazon
USA

ABSTRACT

Amazon encompasses a large number of discrete businesses such as Retail, Advertising, Fresh, Business (B2B e-commerce), and Prime Video, most of which maintain a presence across its e-commerce website. They produce content for our customers that belong to diverse content types such as merchandising (e.g. product recommendations), product advertisements (e.g. sponsored products and display ads), program adoption banners (e.g. Amazon Fresh), and consumption (e.g. Prime Video). When customers visits a web page on the website, it triggers a content allocation process where we determine the specific content to show in regions of customer shopping experience on that web page. Content produced by the aforementioned businesses then needs to be arbitrated during this process. We present a causal bandit based framework to address the problem of content optimization in this context. The framework is responsible for fairly balancing the differing objectives and methods of these businesses, and selecting the right content to display to the customers at the right time. It does so with the goal of improving the overall site-wide customer shopping experience. In this paper, we present our content optimization framework, describe its components, demonstrate the framework's effectiveness through online randomized experiments, and share learnings from deploying and testing the framework in production.

CCS CONCEPTS

• **Computing methodologies** → **Sequential decision making**; **Batch learning**; *Learning from implicit feedback*; Causal reasoning and diagnostics; **Learning to rank**; *Supervised learning by regression*; • **Applied computing** → *Online shopping*; • **Information systems** → **Content ranking**; **Personalization**; *Top-k retrieval in databases*; *Recommender systems*; • **Mathematics of computing** → Bayesian computation.

KEYWORDS

Personalization, Recommender system, Content optimization, Content ranking, Content diversity, Causal bandit, Contextual bandit, View-through attribution, Holistic optimization

Reference Format:
Sameer Kanase, Yan Zhao, Shenghe Xu, Mitchell Goodman, Manohar Mandalapu, Benjamyn Ward, Chan Jeon, Shreya Kamath, Ben Cohen, Yujia Liu, Hengjia Zhang, Yannick Kimmel, Saad Khan, Brent Payne, Patricia Grao. 2022. An Application of Causal Bandit to Content Optimization. In *5th Workshop on Online Recommender Systems and User Modeling (ORSUM 2022), in conjunction with the 16th ACM Conference on Recommender Systems, September 23rd, 2022, Seattle, WA, USA.*

ORSUM@ACM RecSys 2022, September 23rd, 2022, Seattle, WA, USA

1 INTRODUCTION

Amazon encompasses a large number of discrete businesses such as Retail, Advertising, Fresh, Business (B2B e-commerce), and Prime Video, most of which maintain a presence across its e-commerce (or retail) website. These discrete businesses produce content for our customers that belong to diverse content types such as merchandising (e.g. product recommendations), product advertisement (e.g. sponsored products and display ads), program adoption banner (e.g. Amazon Fresh), and consumption (e.g. Prime Video). Each such content is rendered in the form of a widget within independent 'regions of customer shopping experience' on the website, also known as widget groups. For instance, widgets such as 'customers who viewed this also viewed' and 'customers who bought this also bought' are displayed on product detail pages of the website alongside other organic and advertising content. The region of customer shopping experience on the website where the collection of such widgets are displayed is an example of a widget group. We illustrate the concept of a product, widget, and widget group in (figure 1).

When customers visit a web page on the website, it triggers a content allocation process where we determine the specific content to show in the widget groups on that web page. Content produced by the aforementioned discrete businesses then needs to be arbitrated during this process. As the common integration point, Amazon's content optimization framework is responsible for this content arbitration. It accomplishes this by fairly balancing the differing objectives and methods of these businesses through optimization capabilities, and by taking into account customer, content, and shopping context. This results in the right content being shown to the customers at the right time thereby providing a consistent and personalized shopping experience. The content optimization framework is an ecosystem which enables businesses to interoperate independently by enabling content creators, customer shopping experience providers, and web page owners to efficiently construct and serve content for the retail website.

In this paper, we present a causal bandit based framework to address the problem of content optimization with the objective of improving the overall customer shopping experience on Amazon's retail website. Our contributions include:

- an application of a contextual bandit framework to enable introduction of new content through online randomized experiments (or A/B tests) and to learn the value (or benefit) of new content through exploration,

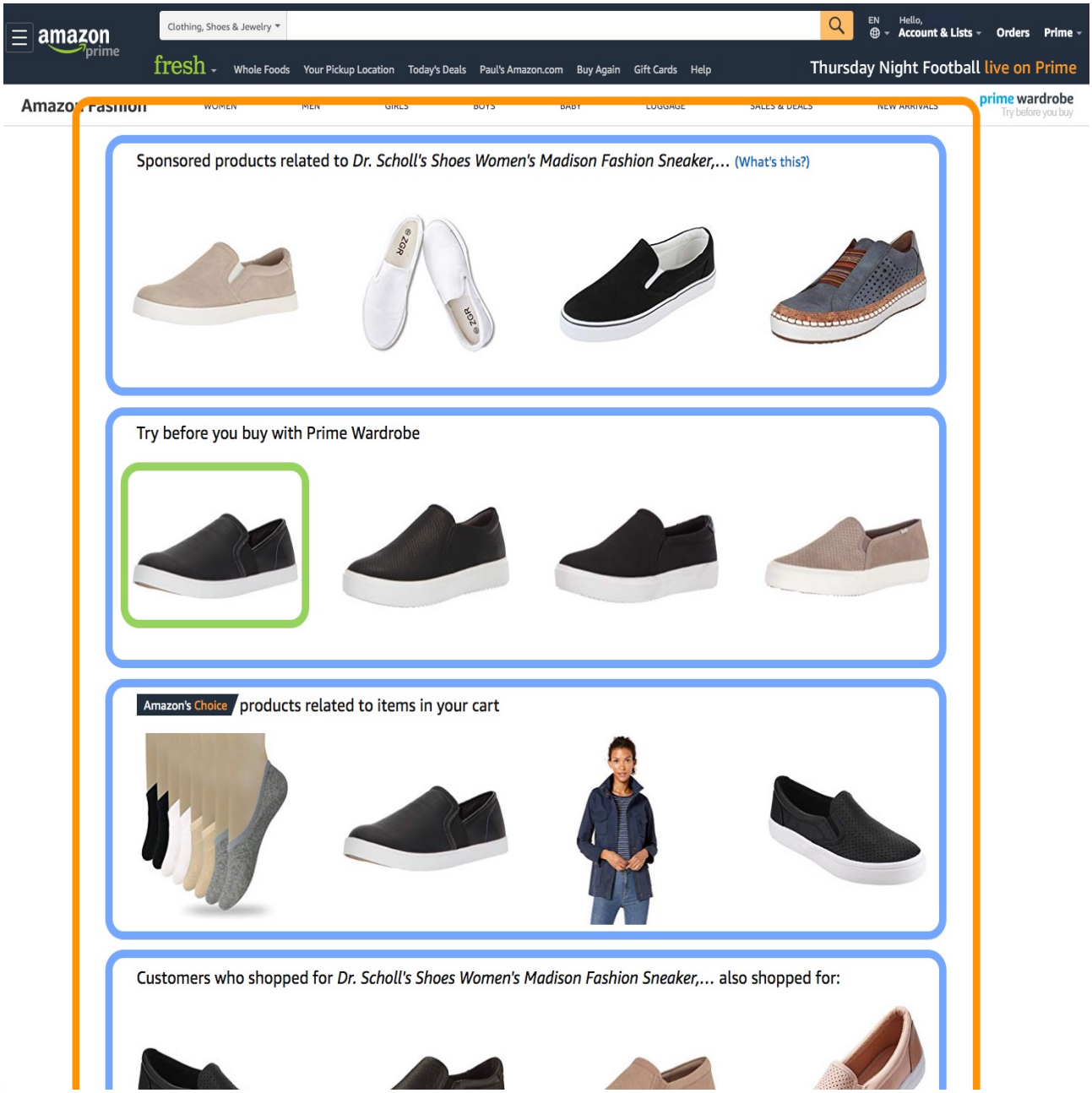

Figure 1: Below is an example of a widget group in a shopping page on Amazon's retail website. It is part of the checkout experience that surfaces after an item has been added to your cart. Here, we see carousels of products (or items) each of which is associated with a title, for instance, "Try before you buy with Prime Wardrobe". Each such carousel of products is rendered in the form of a widget which are marked by blue lines in the figure. A single product recommendation within a widget is marked by the green border while the widget group, which is a collection of widgets, is marked by the orange border. Note that widget groups are regions of customer shopping experience on the website, and each web page on the website can contain one or more widget groups. While the example in this figure illustrates a widget as a carousel of products, they are not limited to it. Widgets can also be used to render images, banners, advertisements, and other types of content. The homepage of Amazon's retail website illustrates this diversity in content type, where a widget is one of the many cards shown in the customer's feed. For reference, we have included an example of Amazon's homepage in Appendix A.

- an approach to measure the reward for actions taken by the contextual bandit framework,
- an application of view-through attribution (VTA) to attribute reward in the context of content ranking which only requires that content be impressed,
- utilization of an uplift modeling framework to augment VTA and to optimize for incremental benefit,
- a methodology to incorporate diversity in content ranking by using cross-content interactions, and
- learnings from the deployment of a low-latency learning framework in production that reduces the delay in feedback and increases the velocity of our learning loop.

Finally, we also demonstrate the effectiveness of our framework through online A/B tests, and share results and insights gathered through the same.

2 RELATED WORK

Application of exploration strategies in the context of recommender systems is an active area of research. In recent years, multiple exploration strategies have emerged and shown promising results [4, 25]. They include epsilon-greedy [18, 29], upper confidence bound (UCB) [3, 15], adding random noise to parameters [7, 8, 22], and bootstrap sampling [20, 21]. We adopt the Thompson sampling algorithm [27] to balance exploration with exploitation under the contextual bandit setting. Originally introduced in 1933 [35], Thompson sampling has been widely adopted in the context of bandit problems recently [16, 31, 32]. It has been shown to achieve state-of-the-art results on some real-world use cases and be robust to delay [1, 5].

Uplift modeling is a widely used approach to measure incremental effect [11, 12, 17, 23]. Our approach to estimate incremental effect or benefit is similar to the meta-learning approach presented in [14, 39]. In [28], the authors presented an application of a causal bandit in targeting campaigns. They estimated incremental effect to optimize for clicks in email marketing campaigns and advertisement campaigns on Amazon's mobile homepage. In this work, we explore an application of causal bandit for content optimization by estimating and optimizing for heterogeneous treatment effect [38].

3 PROBLEM DESCRIPTION

Let $\mathcal{P}$ be the set of all web pages and $\mathcal{Q}$ be the set of all widget groups on Amazon's retail website. Here, a widget group refers to real estate or region of customer shopping experience on the website which can be populated with content c in the form of a widget. Content c can belong to diverse types of content such as product advertisements (e.g. sponsored products and display ads), merchandising (e.g. product recommendations), program adoption banners (e.g. Amazon Fresh), and consumption (e.g. Prime Video). Each widget group $q \in \mathcal{Q}$ in turn can render (or display) a set of ranked content $T_q = \{c_r \mid c_r \in C_q \text{ and } r \in \{1, \ldots, k_q\}\}$ where r is the rank of content rendered in widget group q, k_q is the total number of content that can be rendered in q, and C_q is the set of all possible candidate content that is eligible to be rendered in q. Here, the cardinality of set $C_q >> k_q$. Eligibility for rendering content c in widget group q is typically determined by business rules and content creators.

When a customer visits web page $p \in \mathcal{P}$ on Amazon's retail website, a request is generated with customer and shopping context X to optimize and display content for widget group q on page p. Context for candidate content $c \in C_q$ can be constructed and is denoted by Z. We now formally define the problem we address as determining ranked set T_q from C_q given contexts X and Z so as to maximize the expected reward $\mathcal{R}$. Here, reward $\mathcal{R}$ is a measure of improved customer shopping experience on the retail website. We denote the metric for measuring reward $\mathcal{R}$ by *MOI*, short for 'metric of interest'. In our setting, *MOI* takes into account the short-term as well as long-term impact to the customer's shopping experience, and helps us to fairly balance multiple and differing objectives of various stakeholders. It is computed using actions taken by the customer after interacting with content such as impressions, clicks, purchases and other high-value actions. Note that our problem is different from that of ranking products (or items) within a single widget for a particular recommender system.

A key challenge we face in predicting *MOI* using X and Z is that of the estimate being biased due to the cold-start problem. New content gets continually introduced to be shown on Amazon's retail website while existing content can be sunsetted at any point of time. Empirically, we observe a propensity in customers to interact more with content displayed higher up in the widget group and on the web page. Furthermore, we only observe reward for content that was shown to customers before, but we only show content to customers for which we predict there will be sufficient reward. Consequently, content with few or no prior observations is unlikely to be ranked higher or chosen to be shown to the customers even if it could generate a high-reward in the counterfactual event where a customer were to interact with it. Here, we could use aggregate-level features to partially address the cold-start problem but cannot fully solve it. Moreover, we observe that customer preferences and their interactions with content change over time. To address these challenges, we use a contextual bandit based framework to create a learning loop for new content that has never been shown before to the customers and to dynamically adapt to changing customer preferences.

4 METHODOLOGY

In this section, we present a causal bandit based framework to address the problem of content optimization.

4.1 Features

When a customer visits web page $p \in \mathcal{P}$ on Amazon's retail website, we receive customer and shopping context X. Context Z corresponding to each candidate content can also be generated separately. We then combine contexts X and Z non-linearly to form a single d-dimensional vector $B \in \mathbb{R}^d$. We also include second- and third-order interaction terms between the explanatory variables observed in the context. For reference, we include a few examples of context below:

- Shopping context: region, web page type, widget group id, page item, metadata of page item, and search query
- Customer context: recent interaction events, customer signed-in status, and prime membership status

- Content context: widget id, widget meta information, and content attributes

4.2 Ranking Model

We formulate the problem of content optimization as that of learning to rank the set of eligible content C_q. Our aim is to determine the rank of each eligible candidate content $c \in C_q$ and return the $top-K$ ranked content T_q so as to render them in widget group q. To do so, we need a utility function using which we can evaluate eligible content and rank them. We propose using reward $\mathcal{R}$ to be generated over a subsequent time horizon in the event content were to be shown to a customer, $\mathcal{S} \in \{0, 1\}$, as our utility function. We model it using a generalized linear model,

$$E(R|S, B) = g(B^\top W) \tag{1}$$

where, g is the link function. Since reward $\mathcal{R}$ takes continuous values in our problem setting, we choose an identity link function. We use the set of past observations H made up of triplets of context, action and reward $\{(X_\tau, A_\tau, R_\tau), \tau = 1, \ldots, t-1\}$ to train the ranking model and estimate regression parameters using a Bayesian framework.

4.3 Reward

A fundamental challenge in our problem setting is that of defining and measuring reward so as to evaluate diverse types of content together on an equal footing [36]. When content optimization systems seek to maximize the attributed value (or reward) to individual content, we observe that it leads to development and launch of bespoke recommender systems that optimize for individual objectives and cater to page specific use cases. For instance, recommender systems displayed on different web pages can optimize for increasing customer interactions with themselves through view, clicks, purchases and other high-value actions without being complementary (or incremental) to the customer's current shopping intent. This often results in a poor customer shopping experience which in turn leads to a negative impact to business metrics such as revenue. An alternative here is to attribute value to individual content only if interaction with it is in addition to purchase of the page item (or product) wherever applicable. In empirical evaluation, we observe that results from such an approach can be mixed in that it may improve customer shopping experience on some web pages of the website but not all of them.

To address these challenges, we propose optimizing directly for overall down-session value generated after customer has interacted with content. In this approach, once content has been ranked and rendered, we record customer's interaction events with it such as impressions, clicks, purchases and other high-value actions. Thereafter, we measure the aggregate value generated from these events over a subsequent time horizon to compute our metric of interest MOI. The measured value is attributed to content as reward, if it meets a predefined criteria. Content ranking models then learn to predict for this down-session value of showing content to customers given a context, and make ranking decisions based on the predicted value. This approach enables us to measure and attribute site-wide impact across all devices, apps, widget groups, and web pages from the moment a customer has interacted with content. We call this approach to define reward and rank diverse type of content using aggregate down-session value as holistic optimization.

4.4 Attribution

The predefined criteria used to attribute aggregate down-session value as reward also defines the form of attribution such as click-through attribution (CTA) or view-through attribution (VTA). The distinction between these two forms of attributions is the customer interaction event that triggers the measurement of reward. In CTA, reward is measured after a click event with content occurs while in VTA reward is measured after a view event with content occurs. Note that both VTA and CTA are a form of equal credit attribution model. Likewise, the time horizon over which the reward is measured is called an attribution window. The window is triggered after a customer interaction event with content occurs. We determine attribution windows by performing exploratory data analysis of the length of customer shopping sessions and use multiple windows in practice to cater to varied use cases. We illustrate the concept of VTA and CTA with an attribution window using the example in (figure 2).

A key drawback of CTA is that it cannot attribute reward to content that cannot be clicked or where clicking on content does not necessarily indicate a positive customer shopping experience. We observe that CTA also leads ranking models to favor content that has a high click propensity. Consequently, such models promote content which at times is not relevant to customer's ongoing shopping mission. This distracts the customer from their mission which in turn results in a negative impact to their shopping experience. VTA on the other hand allows us to capture both the positive and negative impact of presenting content to customers. It enables us to capture the value of showing content which inspires customer shopping missions including scenarios where customers can compare selection without requiring direct interaction with content. Furthermore, it is closer in alignment with how an experimentation framework for conducting online A/B tests may measure and attribute aggregate downstream impact after a customer has been exposed to a new shopping experience (or treatment). Thus, it can enable parity in the methodology used to attribute reward in the content optimization and online experimentation systems.

4.5 Uplift Modeling Framework

Both VTA and CTA assume a causal relationship between customers interacting with shown content through views and clicks (cause), and observed reward (effect). In the case of CTA, there is a strong connection between the cause and effect as often times a click is an intentional action on the part of a customer. However, with VTA we cannot establish this direct connection between a customer viewing content and the observed reward. As such, we assume a causal relationship which introduces noise in our observations. Models incrementally trained using such observations are likely to have a high variance in the predictions.

In addition, the attribution model described in the previous section does not capture the incremental value of showing content. Customers can have an underlying propensity to shop products or consume content based on prior exposure or affinity. As a result, attributing observed down-session value as reward without

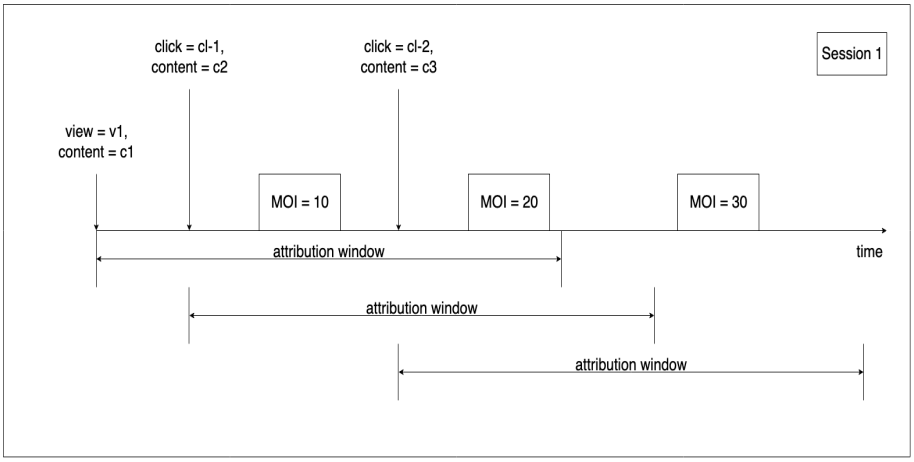

Figure 2: Here, content (c1) corresponding to view (v1) will be attributed with MOI of 30 (= 10 + 20) under view-through attribution. Likewise, under click-through attribution content (c2) corresponding to click (cl-1) will be attributed with MOI of 30 (= 10 + 20) while content (c3) corresponding to click (cl-2) will be attributed with MOI of 50 (= 20 + 30). Note that content (c1) corresponding to view (v1) will be attributed MOI=0 under click-through attribution.

accounting for the counterfactual outcome could lead to models overestimating the predicted benefit at inference time. To address these challenges, we use an uplift modeling framework. It estimates Conditional Average Treatment Effect (CATE) [2] between exposure and non-exposure of content to customers using observational data. We assume conditional unconfoundedness in our problem setting [13, 26]. Then,

$$\begin{aligned} CATE &\equiv E[R(1) - R(0)|B = b] \\ &= E[R(1)|B = b] - E[R(0)|B = b] \end{aligned} \tag{2}$$

where, B is the d-dimensional feature vector. Here, $E[R(1)|B = b]$ is the mean of the treated group where content c is shown in the shopping session, and $E[R(0)|B = b]$ is the mean of the untreated group where content c is not shown in the shopping session. We have explored two approaches to estimate the latter: i) using the mean of untreated group calculated from our ranking logs as a biased estimate for $E[R(0)|B = b]$, and ii) by estimating $E[R(0)|B = b]$ from randomized controlled trials.

The uplift modeling framework is then defined using a two-part model. First, a baseline model estimates the expected counterfactual reward when content c is ranked but not shown in the shopping session. We illustrate the underlying theory using a linear regression model:

$$\mu_0 = B^\top \beta + \epsilon \tag{3}$$

Treatment or incremental effect for each observation in the treated group where content c is ranked and shown in the shopping session is then estimated as:

$$D_i^{(1)} = R_i - \hat{\mu}_0(b) \tag{4}$$

where, R_i is the observed down-session reward for observation i, and $D_i^{(1)}$ is the imputed incremental effect for observation i in the treated group. In the second part, pseudo-effect D is used as the target variable in our ranking model, described in (eqn. 1), to predict the incremental benefit of showing content c to a customer in widget group q.

4.6 Exploration Strategy

The exploration component of our content optimization framework explores content with few observations from the past. To do so, it aims at solving a contextual bandit problem. Here, we use Thompson sampling, an algorithm widely used to balance exploration and exploitation. It suggests to randomly play each arm according to its probability of being optimal. In our problem setting, it means choosing content proportional to the probability of it being optimal. This implies we won't be necessarily choosing content with the highest expected incremental benefit at each time step. It is a trade-off we make to explore content with few observations from the past which have high uncertainty but ultimately may drive a higher reward. In practice, we apply the Thompson sampling algorithm by sampling model parameters $\hat{W}_t$ from their posterior distributions followed by choosing content that maximizes the reward.

Algorithm 1 Thompson Sampling Algorithm for Content Optimization

for $t = 1, \ldots, T$ **do**
  for all $rank = 1, \ldots, k_q$ **do** ▷ k_q is the value of K corresponding to widget group q
    Receive context X_t
    Sample $\hat{W}_t$ from the posterior distribution $\Pr(W|H_t)$
    Select $A_{t_{rank}} = \text{argmax}_A \, B^\top \hat{W}_t$
  end for
  Choose $top - K$ arms and observe reward R
  $H_t = H_{t-1} \cup \{(x_{t_{rank}}, a_{t_{rank}}, r_{t_{rank}}), rank = 1, \ldots, k_q\}$
end for

Candidates that are ranked and chosen to be displayed by the ranking model are then logged along with their observed reward in the form of triplets (x_t, a_t, r_t). Thereafter, we estimate the incremental effect for each observation in the logged feedback using

our uplift modeling framework. The incremental effects are subsequently used as target variables to incrementally train our ranking model using a batch update under the Thompson Sampling framework. While doing so, we decay the model parameters by a small fraction of their existing value to account for changes in the environment [9]. This completes the feedback loop which allows us to continuously explore actions and expand our knowledge for making better decisions in the future. This in turn enables us to support running online A/B tests using which content creators can introduce new content across Amazon's retail website with the goal of improving customer shopping experience and measure its benefit while doing so.

4.7 Incorporating Diversity in Ranking

Showing high-relevance content without taking content diversity into account leads to monotony and tends to make the holistic shopping experience less meaningful for the customers. Optimizing for the whole widget group involves balancing relevance and diversity of the content therein, where the whole-widget group effect is represented using the amount of similar content displayed in it. One approach followed here is to model this as a submodular optimization problem [37]. In [19, 34], the authors propose using submodular functions which have a diminishing returns property. In their approach, the total score for a content is derived from its relevance while also accounting for the decreasing utility of showing multiple content of the same type. As a result, the value of selecting content from a given category or type decreases as a function of the number of content belonging to that type already selected. A key shortcoming of this approach is that it lacks a feedback loop and parameters of the diversity scoring function aren't learned to optimize for the same objective as the relevance scoring function.

Instead we propose a two-stage model for incorporating diversity into content. We first rank all the eligible content $c \in C_q$ to be shown in widget group q using our underlying ranking model (eqn. 1). Thereafter, we iteratively re-rank content at each position r in the widget group by taking into account the content that is already ranked in the previous $r-1$ positions. This is accomplished by using a second ranking model which includes cross-content interaction features. To capture these interactions, we categorize each content c in to one of m distinct categories or types. The goal is to then select an optimal number of highly relevant widgets in each category. For a given widget group $q \in Q$, the overall value of widget group q is represented as:

$$f(c_k|q, X) = \sum_{r=1}^{k} f(c_r|T_{q_{(r-1)}}, X, W) \tag{5}$$

$$f(c_r|T_{q_{(r-1)}}, X, W) = g(B^\top W) \tag{6}$$

where, c_r is the content at rank r in widget group q, $T_{q_{(r-1)}}$ is the set of content allocated in the top $r-1$ positions, X is the request and customer context received as before, W are the model parameters, and g is a generalized linear model.

5 LOW LATENCY LEARNING FRAMEWORK

In production, we observed that our content optimization framework suffered from a delay in feedback as it needed multiple days on average to complete the learning loop. This involves logging the feedback after customers are shown with content on the website, measuring and attributing reward in our data pipelines, incrementally training the models at a daily cadence, and deploying the retrained models in production. A delay in feedback has the following consequences in production for our content optimization framework:

- When new content is introduced into the ecosystem, the optimization framework is not able to effectively estimate its potential benefit, and the content is subject to exploration as expected. Due to the delay in feedback, this can result in new content being explored at a higher show-rate for the duration of the delay during the initial learning period before sufficient observations are logged for the model to learn from its own feedback. This in turn can result in sub-optimal decision making and introduction of poor customer shopping experience during the learning period.
- While the cost of exploration may be amortized over long running content campaigns, a longer feedback loop induces limitations in realizing benefit during high-value events such as Cyber Monday where new content may be introduced for a short period of time. In such cases, some content promoting sales or other events will be turned off even before their benefit is effectively learned by the model.
- A longer feedback loop also decreases the velocity of running online A/B tests where new content and improvements to optimization framework may be introduced with the aim of improving customer shopping experience.

To address these challenges, we developed a Low Latency Learning (L3) framework which has reduced the learning loop for our content optimization framework by 90% from multiple days to a couple of hours.

5.1 Bayesian Regularization

A key challenge we encountered in the development of L3 framework was that of the number of samples available for incremental training being significantly lower than before as we retrained the models at a faster cadence. This impacts the model's learning process in two ways: i) outliers in the dataset can cause the model to incorrectly associate higher potential reward for some content despite winsorization techniques, and ii) insufficient data can limit the model's ability to learn about content's reward distribution especially in regions with small amount of traffic. Empirically, we observe that both of these scenarios lead to over-exploration of content.

We address this challenge using Bayesian Regularization. The use of Gaussian priors has been established in [6] as a form of L2 regularization wherein the following equivalence is explored:

$$(BB^T + \alpha I)^{-1} B^T R = \mathbb{E}\Big[\frac{P(B|W)P(W)}{P(B)}\Big] \tag{7}$$

Here, instead of initiating the feature weights from a static prior (i.e. mean 0 and variance 1), we derive a prior distribution from previously learned weight distributions in order to allow for a more pessimistic exploration regime. For instance, by using a prior representing 20th percentile mean of all features and 75th percentile variance of all features, we can reduce the chances of over-exposure

Table 1: Online experiment results.

Experiment	Incremental Impact (% improvement)	p-value
EXP-1	+0.16%	0.02
EXP-2A	+0.01%	0.12
EXP-2B	+0.01%	0.09
EXP-3	+0.09%	0.02
EXP-4	+0.05%	0.19
EXP-5	+0.11%	0.00

for new content during the learning period. This improvement in turn has enabled us to launch the L3 framework in production and reduced the delay in feedback by 90%.

6 EXPERIMENTS

We first evaluate our content optimization framework using both traditional offline evaluation and off-policy evaluation methodologies [30, 33]. This allows us to evaluate and prune alternative treatment policies before introducing them in online randomized experiments [10, 24]. Here, we use regression and ranking metrics to evaluate the framework quantitatively, and content's share-of-voice and ranking distributions to evaluate it qualitatively using domain knowledge. Subsequently, we demonstrate the effectiveness of our framework through five online A/B tests.

6.1 Online Experiment Setting

In our online experimentation setting, observational units (or shopping sessions) are randomly exposed to either the baseline control policy or the alternative treatment policies. Here, we track the impact to our metric of interest *MOI*, which is a measure of improved site-wide customer shopping experience. In the results, we include the causal effect w.r.t percentage improvement in this metric at Amazon's scale. The experiments are conducted across all of Amazon's world-wide marketplaces and product categories. Level of significance α for these experiments was determined by Amazon's business objectives and was set to 0.10. Duration for these experiments was estimated from statistical power analysis. We allocated equal traffic to both the control and treatment groups. During the course of the experiment, the models were incrementally trained using their own set of logged feedback.

6.2 Experiment 1: Application of Holistic Optimization Framework

We first test the effectiveness of our holistic optimization framework to rank content in a widget group on product detail pages of Amazon's retail website. This is a region of customer shopping experience on the website where we usually see organic content such as 'customers who viewed this also viewed' and 'customers who bought this also bought' widgets being displayed alongside advertising content. A key challenge in dynamically ranking content in this setting was that of attribution of reward to diverse type of content which were generated by content creators who optimized for differing business objectives. As such, our framework needed to arbitrate content during the content allocation process and fairly balance the differing objectives. Since holistic optimization framework measures reward using the aggregate down-session value after customer has interacted with content, we wanted to test its effectiveness in addressing this problem. In the control group, content was statically ranked by a rule-based system while in the treatment group, our framework dynamically ranked content using the holistic optimization framework. In the results (EXP-1), we observe a practically and statistically significant improvement in the *MOI* metric which is a measure of site-wide improvement in customer shopping experience.

6.3 Experiment 2: Application of View-through Attribution

In this experiment, we applied our content optimization framework to the image size selection problem. Usually, product display images on Amazon's detail page exist in three sizes – small, medium and large. Here, the size of a rendered image can influence the customer's understanding of the product. Hence, we want to select and render an optimal size of the same product image so as to help the customers evaluate products better especially for high consideration purchases. This is a use case where click-through attribution cannot be used as clicking on the content does not necessarily indicate a positive customer shopping experience. We formulate the task of optimal image size selection as a learning to rank problem, and use view-through attribution to measure and attribute reward to the rendered image size. To demonstrate the effectiveness of this approach, we ran two experiments – one each for desktop and mobile surfaces. In the control group, image size was selected by a rule-based system while in the treatment group, our framework ranked the image size variations and chose the top ranked variation to render. In both the experiments (EXP-2A and EXP-2B), we observe an improvement in the *MOI* metric which is practically significant at Amazon's scale.

6.4 Experiment 3: Application of Causal Bandit

After demonstrating the effectiveness of VTA, we tested the utility of the uplift modeling framework. The framework allows us to measure and optimize for incremental value generated by content, and reduces the observational bias in data. We conducted an experiment in a widget group which is located at the bottom of product detail pages on the desktop retail website where personalized content that is usually generated by taking recent browsing history into account is shown. This in turn allowed us to test our hypothesis that customers can have an underlying propensity to shop products or consume content based on prior exposure or affinity, and optimizing for incremental benefit can result in a positive customer shopping experience. In the control group, content was ranked by a linear bandit without using the uplift modeling framework while rewards were measured and attributed using CTA. In the treatment group, content was ranked using a linear causal bandit with VTA. In the results (EXP-3), we observe that the linear causal bandit using VTA performed better than the linear bandit which did not use uplift modeling framework. The improvement in *MOI* metric was both practically and statistically significant.

6.5 Experiment 4: Application of Incorporating Diversity in Ranking

Subsequently, we ran an experiment (EXP-4) on the desktop homepage of Amazon's retail website to test the impact of incorporating diversity in content. In the control group, content was ranked using just the single baseline ranking model, while in the treatment group, content was ranked using the two-stage ranking model – first using the baseline model followed by a re-ranking model which incorporates diversity using cross-content interaction features. Here, we observe a practically significant improvement in the *MOI* metric. Based on the results, we infer that incorporating diversity into content ranking can lead to a better customer shopping experience.

6.6 Experiment 5: Application of Low Latency Learning Framework

L3 pipeline has shortened the delay in feedback for our contextual bandit based framework by 90%. As a result, we expect the bandit retrained at hourly cadence to converge sooner and perform better than the one retrained at a daily cadence. To test the benefit and measure the impact of low latency learning, we ran an experiment on the mobile homepage of Amazon's retail website. In the control group, content was ranked by a linear bandit incrementally trained at a slower cadence with a learning loop of multiple days, while in the treatment group, content was ranked by a linear bandit incrementally trained at a faster cadence with a learning loop of a few hours. In the results (EXP-5), we observe that the bandit with a shorter delay in feedback performed better w.r.t our metric of interest *MOI* where the improvement was both practically and statistically significant. Based on the results, we infer that reducing the delay in feedback and increasing the velocity of learning loop has a positive impact on customer shopping experience.

7 CONCLUSION

In this paper, we presented a causal bandit framework to address the problem of content optimization with the objective of improving the overall customer shopping experience on Amazon's e-commerce (or retail) website. Therein, we introduced a holistic optimization framework that enables us to define reward and rank diverse types of content using aggregate down-session value; presented the concept of view-through attribution; discussed how it addresses some of the shortcomings of click-through attribution; and presented applications of VTA in ranking content belonging to diverse type. To address the shortcomings of view-through attribution, we used an Uplift modeling framework which has enabled us to rank content using incremental or causal benefit instead of overall value. Subsequently, we proposed a two-stage model to incorporate diversity in content ranking by using cross-content interaction features. It helps us to balance relevance with diversity in content shown on Amazon's retail website and provide a meaningful experience to our customers. Thereafter, we shared learnings from the deployment of a low-latency learning framework in production that has reduced the delay in feedback and shortened the learning loop by 90%. Here, we described our application of Gaussian prior as a form of L2 regularization which in turn enabled the launch of the L3 framework. We then demonstrated the effectiveness of our methodology through multiple online experiments, and shared results and insights gathered through the same. Finally, we believe our methodology and learnings are generic and can be extended to content optimization problems in other domains. It can also be extended to rank items (or products) within a single widget for a product recommendation system.

REFERENCES

[1] Shipra Agrawal and Navin Goyal. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*. 127–135.

[2] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360. https://doi.org/10.1073/pnas.1510489113

[3] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. Finite-Time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.* 47, 2–3 (may 2002), 235–256. https://doi.org/10.1023/A:1013689704352

[4] Alberto Bietti, Alekh Agarwal, and John Langford. 2018. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064* (2018).

[5] Olivier Chapelle and Lihong Li. 2011. An Empirical Evaluation of Thompson Sampling. In *Proceedings of the 24th International Conference on Neural Information Processing Systems* (Granada, Spain) *(NIPS'11)*. Curran Associates Inc., Red Hook, NY, USA, 2249–2257.

[6] Mário AT Figueiredo. 2003. Adaptive sparseness for supervised learning. *IEEE transactions on pattern analysis and machine intelligence* 25, 9 (2003), 1150–1159.

[7] Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. 2018. Noisy Networks For Exploration. In *International Conference on Learning Representations*.

[8] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (New York, NY, USA) *(ICML'16)*. JMLR.org, 1050–1059.

[9] Thore Graepel, Joaquin Quiñonero 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 the 27th International Conference on International Conference on Machine Learning* (Haifa, Israel) *(ICML'10)*. Omnipress, Madison, WI, USA, 13–20.

[10] Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, Mike Curtis, Alex Deng, Weitao Duan, Peter Forbes, Brian Frasca, Tommy Guy, Guido W. Imbens, Guillaume Saint Jacques, Pranav Kantawala, Ilya Katsev, Moshe Katzwer, Mikael Konutgan, Elena Kunakova, Minyong Lee, MJ Lee, Joseph Liu, James McQueen, Amir Najmi, Brent Smith, Vivek Trehan, Lukas Vermeer, Toby Walker, Jeffrey Wong, and Igor Yashkov. 2019. Top Challenges from the First Practical Online Controlled Experiments Summit. *SIGKDD Explor. Newsl.* 21, 1 (may 2019), 20–35. https://doi.org/10.1145/3331651.3331655

[11] Pierre Gutierrez and Jean-Yves Gérardy. 2017. Causal inference and uplift modelling: A review of the literature. In *International Conference on Predictive Applications and APIs*. PMLR, 1–13.

[12] Behram Hansotia and Brad Rukstales. 2002. Incremental value modeling. *Journal of Interactive Marketing* 16, 3 (2002), 35.

[13] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

[14] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116, 10 (2019), 4156–4165.

[15] T.L Lai and Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6, 1 (1985), 4–22. https://doi.org/10.1016/0196-8858(85)90002-8

[16] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New York, NY, USA, 661–670. https://doi.org/10.1145/1772690.1772758

[17] Victor S. Y. Lo. 2002. The True Lift Model: A Novel Data Mining Approach to Response Modeling in Database Marketing. *SIGKDD Explor. Newsl.* 4, 2 (Dec. 2002), 78–86. https://doi.org/10.1145/772862.772872

[18] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533. https://doi.org/10.1038/nature14236

[19] Houssam Nassif, Kemal Oral Cansizlar, Mitchell Goodman, and S. V. N. Vishwanathan. 2016. Diversifying music recommendations. In *ICML 2016*.
[20] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. 2016. Deep Exploration via Bootstrapped DQN. In *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 4026–4034.
[21] Ian Osband and Benjamin Van Roy. 2015. Bootstrapped Thompson Sampling and Deep Exploration. https://doi.org/10.48550/ARXIV.1507.00300
[22] Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y. Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. 2018. Parameter Space Noise for Exploration. In *International Conference on Learning Representations*.
[23] Nicholas Radcliffe. 2007. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal* (2007), 14–21.
[24] Thomas S. Richardson, Yu Liu, James Mcqueen, and Doug Hains. 2022. A Bayesian Model for Online Activity Sample Sizes. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 151)*. PMLR, 1775–1785.
[25] Carlos Riquelme, George Tucker, and Jasper Snoek. 2018. Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling. In *International Conference on Learning Representations*.
[26] D. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66 (1974), 688–701.
[27] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, and Ian Osband. 2017. A Tutorial on Thompson Sampling. *CoRR* abs/1707.02038 (2017). arXiv:1707.02038
[28] Neela Sawant, Chitti Babu Namballa, Narayanan Sadagopan, and Houssam Nassif. 2018. Contextual multi-armed bandits for causal marketing. In *ICML 2018*. https://www.amazon.science/publications/contextual-multi-armed-bandits-for-causal-marketing
[29] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized Experience Replay. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
[30] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (New York, NY, USA) *(ICML'16)*. JMLR.org, 1670–1679.
[31] Steven L Scott. 2010. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* 26, 6 (2010), 639–658.
[32] Malcolm Strens. 2000. A Bayesian framework for reinforcement learning. In *ICML*, Vol. 2000. 943–950.
[33] Adith Swaminathan and Thorsten Joachims. 2015. The Self-Normalized Estimator for Counterfactual Learning. In *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc.
[34] Choon Hui Teo, Houssam Nassif, Daniel Hill, Sriram Srinivasan, Mitchell Goodman, Vijai Mohan, and S.V.N. Vishwanathan. 2016. Adaptive, Personalized Diversity for Visual Discovery. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) *(RecSys '16)*. Association for Computing Machinery, New York, NY, USA, 35–38.
[35] William R Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3/4 (1933), 285–294.
[36] Shenghe Xu, Yan Zhao, Sameer Kanase, Mitchell Goodman, Saad Khan, Brent Payne, and Patricia Grao. 2022. Machine learning attribution: Inferring item-level impact from slate recommendation in e-commerce. In *KDD 2022 Workshop on First Content Understanding and Generation for e-Commerce*. https://www.amazon.science/publications/machine-learning-attribution-inferring-item-level-impact-from-slate-recommendation-in-e-commerce
[37] Yisong Yue and Carlos Guestrin. 2011. Linear Submodular Bandits and their Application to Diversified Retrieval. In *Advances in Neural Information Processing Systems*, Vol. 24. Curran Associates, Inc.
[38] Yan Zhao, Mitchell Goodman, Sameer Kanase, Shenghe Xu, Yannick Kimmel, Brent Payne, Saad Khan, and Patricia Grao. 2022. Mitigating Targeting Bias in Content Recommendation with Causal Bandits. In *Proceedings of the 2nd Workshop on Multi-Objective Recommender Systems (MORS 2022), in conjunction with the 16th ACM Conference on Recommender Systems (RecSys 2022), Seattle, WA, USA*.
[39] Zhenyu Zhao and Totte Harinen. 2019. Uplift modeling for multiple treatments with cost optimization. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 422–431.

APPENDIX

A An Illustration of Diverse Type of Content on Amazon's Homepage

Below, (figure 3) illustrates diverse type of content being shown on the homepage of Amazon's retail website.

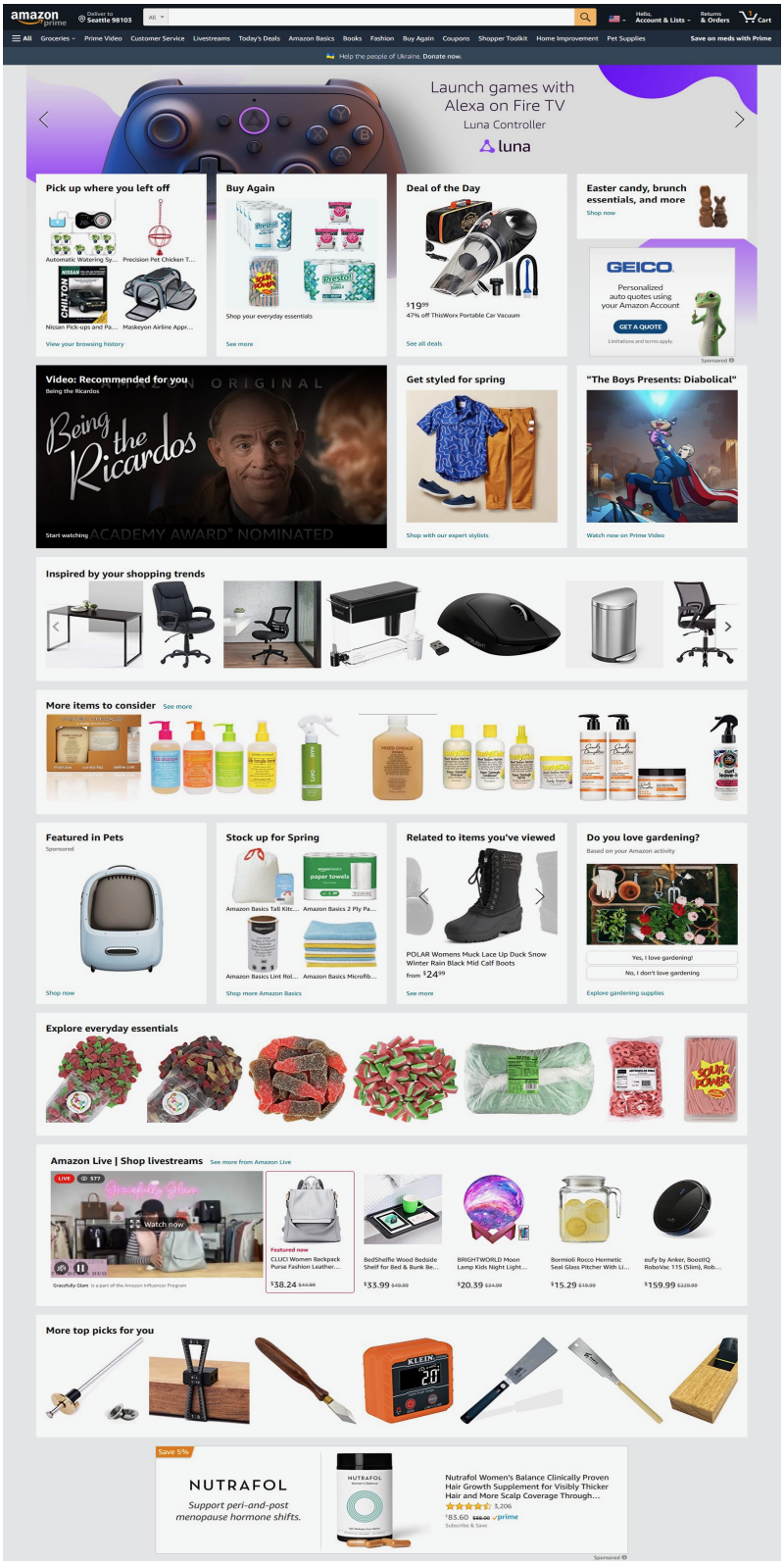

Figure 3: Homepage of Amazon's Retail Website.