When to Target Customers? Retention Management using Dynamic Off-Policy Policy Learning

Ryuya Ko[†] Kosuke Uetake[‡] Kohei Yata[§] Ryosuke Okada[¶]

Abstract

We examine how to learn personalized customer retention strategies when customers' intentions to purchase evolve over time. Working with a Japanese online platform, we first implement a large-scale randomized experiment, in which coupons are randomly sent to first-time buyers at different times. The experimental data allows us to estimate the personalized dynamic retention policy using the off-policy policy learning methods. We extend the existing methods by allowing non-Markovian strategies and by explicitly considering constraints such as budget constraints. Our results show that the optimal dynamic policy is more cost-effective than baseline policies. We also test the optimal policy online to confirm its performance.

Very Preliminary. Draft For Conference Submission. First draft: June 18, 2021. This version: October 29, 2022. We thank Tat Chan, Dean Eckles, Harikesh Nair, Shohei Sakaguchi, Jiwoong Shin, K Sudhir, Raph Thomadsen, and Duncan Simester for helpful comments. We also thank the seminar and conference participants at CODE, KDD, Marketing Science Conference, RecSys, Temple, WUSTL. The paper's results are our owns and do not represent the companies' views.

[†]University of Tokyo koryudr35@g.ecc.u-tokyo.ac.jp

[‡]Yale School of Management. kosuke.uetake@yale.edu

[§]University of Wisconsin-Madison. yata@wisc.edu

[¶]ZOZO Inc.

1 Introduction

Managing customer retention is a central part of customer relationship management (CRM). In particular, it is well-known in both academia and practice that the attrition rate at the early stage of customer life-cycle is quite high (e.g., Fader and Hardie (2007), Fader and Hardie (2010), Kim (2021)). Since the churn rate of first-time buyers tends to be much larger than other repeated customers, it is essential to increase the retention of those first-time buyers to increase the overall customer lifetime value.

To improve the retention of first-time buyers, companies usually make special treatment for those first-time buyers. For example, as many popular marketing strategy books suggest, it is common for companies to send a special thank-you message to first-time buyers. It is also commonly observed that companies send them a coupon so that they are going to make another purchase with the coupon.[1]

In the era of Big Data, data-driven CRM strategies can improve the retention of first-time buyers. E-commerce companies can collect not only basic demographic or socioeconomic variables about customers but also a large number of variables on customer behavior prior to and after the first purchase such as browsing history and installing an app. These massive, fine-grained data allow companies to design personalized targeting policies for sending messages or coupons, which enhance customer retention and facilitate overall CRM performance.

Although the availability of Big Data for personalized data-driven marketing strategies has led to a surge in interest in methods that take advantage of detailed data to help develop policy learning strategies on how and whom to target, there are some limitations in the existing methods in computer science, marketing, and economics. One potential limitation of the existing methods is that they mainly focus on a static setting where a company sees customers at one point in time and decides how to treat customers right away. However, retention management inherently involves dynamics in

[1]There are a lot of blog posts and articles online on how to send appreciation emails to first-time buyers. For example, https://www.drip.com/blog/customer-appreciation-emails.

that customers' behavior and interest may change over time. For example, it is well-known that the retention probability tends to decline as the length of time since the customer's first-time purchase increases, as known as a "recency trap" (Neslin, Taylor, Grantham, and McNeil (2013)). Hence, it is crucial to incorporate dynamics in policy learning for retention management.

Another potential limitation of the existing methods is that they do not take into account the restrictions that companies practically face. For example, an increase in the concern for customers' privacy issues leads to fairer treatment of customers by companies. Such fairness constraints restrict the way the companies personalize retention strategies.[2] Another prominent example is that typically companies' efforts of targeting are constrained by binding budget ceilings. A common rule of thumb is that the marketing budget for B2B companies is between 2 and 5% of their revenue and between 5 and 10% for B2C companies.[3] Hence, marketing managers need to efficiently select marketing strategies within the budget. When a budget is limited such that only a subset of customers can receive financial incentives, it is even more important to determine a cost-effective personalized targeting strategy.

In this paper, we aim for overcoming these challenges. Specifically, we propose an empirical framework to create dynamic personalized targeting strategies for retention management when budget constraints (or other constraints) limit what fraction of the customers can be treated with the aim of maximizing the overall retention rate. We then apply the econometric methodology to the experimental data from a leading Japanese e-commerce company, which collects detailed customer data for personalization. Since the company serves more than 100,000 first-time buyers every month, even a small improvement in the cost-effectiveness of a personalized retention strategy can have a large impact on the company's profits.

Our method follows the recent literature on the dynamic treatment regime (DTR)

[2]See, e.g., Kallus and Zhou (2021).

[3]See, e.g., https://www.bdc.ca/en/articles-tools/marketing-sales-export/marketing/what-average-marketing-budget-for-small-business

used in statistics, medicine, computer science, and economics. Dynamic treatment regimes, also called adaptive treatment strategies, are sequential decision rules that adapt over time to the changing status of each customer. A DTR determines whether or not to offer a coupon as a function of state variables such as past purchase history, past browsing history, past responses to emails, etc. This dynamic nature makes the estimation of DTRs challenging because the treatment assignment today should take both its direct effects on current outcomes and indirect effects on future outcomes and future state variables that affect future treatments. We extend the methodologies to develop DTRs by explicitly incorporating the budget constraint, which makes the dynamic optimization problem even more complicated because the current treatment assignment affects future treatment assignments and outcomes not only through dynamic customer behavior but also through an inter-temporal budget constraint. Incorporating the dynamics and the constraint, however, makes the targeting policy more practical and allows us to develop cost-effective dynamic targeting policies.

Our approach to estimating the optimal DTR given the budget constraint builds on both Q-learning (e.g., Murphy (2003)) and Backward Outcome Weighted Learning (BOWL) (e.g., Zhao, Zeng, Laber, and Kosorok (2015)). The Q-learning is an approximate dynamic programming procedure that estimates the optimal DTR by maximizing the conditional expectation of the current and future payoffs, known as a Q-function. We model the Q-function by various machine learning methods such as Random Forest, Stochastic Gradient Descent (SGD), and LASSO. Those machine learning methods are attractive as they allow one to avoid fully parameterizing underlying data-generating processes. We then maximize the Q-functions to obtain the optimal DTRs. By contrast, the BOWL approach reframes the estimation of an optimal DTR as a sequential weighted classification problem, starting from the very end period. This reformulation is helpful as one can use existing classification algorithms such as support vector machines (SVM) and Stochastic Gradient Descent Classification (SGDC), which are readily available in popular programming languages. Hence, BOWL is a direct ap-

proach to learning the optimal DTR. We provide the characterization of the optimal DTR given the budget constraint and propose an algorithm to derive the optimal policy.

To estimate the optimal DTR, we use the experimental data that comes from a large e-commerce platform in Japan. Our randomized experimental design guarantees "sequential ignorability" condition, meaning treatment assignments are independent of potential future outcomes, conditional on the history up to the current period, which allows us to infer the optimal DTR from the experimental data.

Applying the method to the experimental data, we find that the optimal DTRs can achieve significantly higher retention than the existing strategy of just sending appreciation emails, and are also more cost-effective than alternative policies. Also, due to personalization, our approach identifies that it is not always optimal to send incentives right after the first purchase. For some users, it may be more beneficial to send incentives later. In our off-line evaluation, we find that our optimal policy leads to as high as 880% of return on advertising (coupon) spending (ROAS). Lastly, the company tested our optimal DTR online and found that the optimal DTR outperforms the current algorithm.

Our approach is practically important for many marketing managers. For companies with limited resources such as start-ups, they may not have enough marketing resources to give coupons or financial incentives to all potentially loyal customers. Since we leverage one experimental data to learn optimal DTRs under different scenarios, even start-ups can set up the data and derive dynamic personalized policies. Also, our approach is useful for marketing managers in bigger institutions. That is because our cost-effective approach allows the companies to save a significant amount of resources for other marketing campaigns than retention management. Moreover, since our approach derives the optimal strategies within the budget constraint, it is much easier for the company to manage expenses. As Ascarza, Ross, and Hardie (2021) point out, many companies do not effectively use their customer data to achieve the companies'

goals within their budget. Our method can contribute to those companies' objectives.

The rest of the paper is organized as follows. In Section 2, we review the related literature in marketing, economics, and computer science. Section 3 introduces the background of the setup we study and Section 4 describes the experimental design and provides some descriptive statistics with a special emphasis on consumer heterogeneity. Section 5 explains the model and our solution approaches. In Section 6, we discuss the results of the off-policy policy evaluation and show the online evaluation results. Lastly, Section 7 concludes.

2 Related Literature

Our paper is related to the marketing literature on proactive churn management. Churn management is one of the key priorities for most businesses as customer retention is a major component of customer lifetime value (CLV) and hence a cornerstone of successful CRM (Ascarza, Neslin, and et al. (2018)). As summarized by Neslin, Taylor, Grantham, and McNeil (2013), the industry practice for data-driven retention management is to flag risky customers who are likely to churn using behavioral and demographic variables. By predicting customer attrition before they decide, firms can proactively communicate with those who are at risk of churning to convince them to stay (e.g., Neslin, Gupta, Kamakura, Lu, and Mason (2006)). Recently, a few papers go beyond estimating the churn prediction model and targeting customers with the highest risk. Ascarza (2018) points out that it is not effective to target customers with a higher chance of predicted retention as they may not be responsive to marketing interventions and propose to determine targeting based on uplift. Lemmens and Gupta (2020) note that it is crucial to take the financial impact of a retention intervention based on CLV into account. Our approach adds to the literature by developing a method that estimates targeting policies that dynamically adopt in response to evolving customer

state variables.[4]

In marketing, there is a growing number of papers that propose methods for learning optimal personalized policies.[5] Hitsch and Misra (2018) propose a policy learning method based on the estimation of conditional average treatment effect using k-nearest neighbors. Simester, Timoshenko, and Zoumpoulis (2020) consider an efficient policy evaluation method when existing policies and new policies are compared. Yoganarasimhan, Barzegary, and Pani (2020) estimate CATE with different machine learning models and compare the performance of targeting policies constructed based on these models. Yang, Eckles, Dhillon, and Aral (2020) consider how to derive targeting policies when an outcome of interest is observed only in the long-term. To do so, they use the idea of statistical surrogacy (Athey, Chetty, Imbens, and Kang (2019)) and optimal policy learning. Those papers focus on a static setting and do not consider a dynamic setting like ours. Moreover, we examine a different substantive marketing question on customer retention.

Recently, a few papers in marketing examine dynamic personalized policies. Kar, Swaminathan, and Albuquerque (2015) examine dynamic targeting in the context of advertising and show that ad quality creates inter-temporal externalities that the dynamic targeting policy should take into account. Liu (2021) develops a dynamic reinforcement learning algorithm for dynamic pricing in e-commerce which found the dynamic reference price effect plays an important role. We add to this brand-new literature by proposing a method to estimate dynamic cost-effective policies, which explicitly take constraints into account. Moreover, we differ from those papers as we

[4]For a review of the literature, see, e.g., Ascarza, Neslin, and et al. (2018).

[5]In economics, there is a strand of papers on policy learning. Athey and Wager (2021), for example, develop methods for policy learning that can work with observational data. Their method can be used to optimize various types of treatment allocation such as binary and continuous. Kitagawa and Tetenov (2018) study policy learning in a nonparametric setting and obtain regret bounds. Bhattacharya and Dupas (2012) and Sun (2021) study how to incorporate a certain policy constraint in a static setting of policy learning, while Sakaguchi (2019) proposes a way to extend it to a dynamic setting. Our method is different from Sakaguchi (2019) in that our approach transforms the constrained problem into the unconstrained problem, which allows us to use the existing approaches such as Q-learning and BOWL. Moreover, our method is computationally light and can accommodate a large number of state variables.

consider a non-Markov dynamic setup.[6]

3 Background

The company we work with is one of the largest e-commerce platforms in Japan (about $2,300 million in 2021) that mainly sells apparel products for young female customers. There are more than 1,500 retailers on the platform and more than 8 million active users. Those users purchase products from the retailers through the platform's website or through the mobile app. Those apparel brands and retailers basically delegate marketing activities to the platform company that also manages inventories at the company's warehouse and handles shipping directly to consumers. The platform charges a certain fee to retailers for each transaction, but retailers do not have to bear any costs for keeping their products in the platform's inventory warehouse.

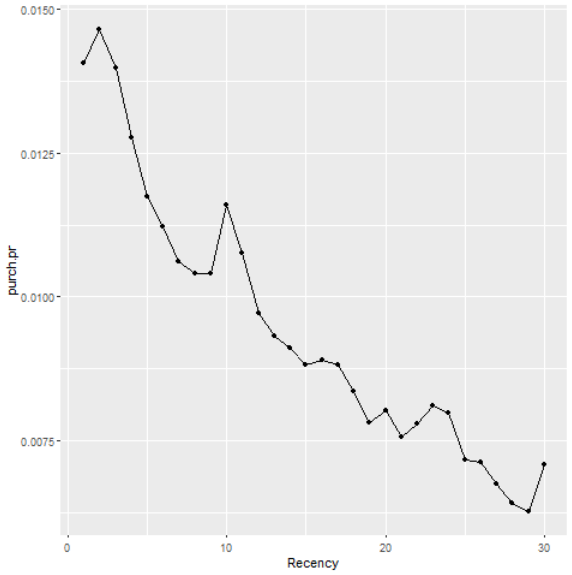
The company is interested in increasing the retention of the customers who have just made their first purchase as those first-time buyers' retention rate is lower than other customers who have experienced multiple purchases before. Also, based on the company's examination, the contribution of the second purchase to the customer lifetime value (LTV) is much larger than that of the third/fourth/fifth purchases. Hence, it is critical to have first-time buyers make a second purchase for improving the overall retention rate and LTV of customers. Before we started the project, as suggested by many popular marketing strategy books, the company sends "thank you" messages to first-time buyers to show appreciation and to provide them with some useful information about the platform such as rankings and the company's app. The purposes of such emails are trying to convince them to buy another item and also to collect customer

[6] In the literature of personalized medicine, dynamic treatment regimes are developed to adaptively select clinical treatments in response to the factors emerging over time (e.g., Murphy (2003), Murphy, Lynch, Oslin, McKay, and TenHave (2007), Zhang, Tsiatis, Laber, and Davidian (2013) to name a few). Recently, thanks to the development in Machine Learning, Q-learning, a reinforcement learning approach, has been applied to DTR to approximate the value function (Zhao, Kosorok, and Zeng (2009)). Another related paper is Nie, Brunskill, and Wager (2021), which studies when to start treatment and learn the optimal policy. Our paper extends the backward outcome weighted learning (BOWL) (Zhao, Zeng, Laber, and Kosorok (2015)) by explicitly accommodating budget constraints.

behavior information.

Although the company knows that the current appreciation emails to first-time buyers have some positive impacts on retention, the company would like to increase the retention rate even further by providing financial incentives.

Figure 1: Recency and Retention Probability



Note: The graph shows the time (days) since the first purchase on the horizontal axis and purchase probabilities, i.e., retention probabilities on the vertical axis. The data comes from the experimental data we use for the estimation of the optimal DTR. The company sends incentives 2 days, 10 days, and 30 days after their first purchase, giving hikes around those days.

When providing incentives, timing is crucial. It has been well-known in marketing that there is a recency trap as pointed out by Neslin, Taylor, Grantham, and McNeil (2013), i.e., the retention probability becomes smaller as the time since the last purchase increases. Figure 1 shows the relationship between the days since the last purchase and the average purchase probability in our application. As in Neslin, Taylor, Grantham, and McNeil (2013), the average purchase probability generally declines as recency increases.

The declining pattern in the figure has implications on when to target customers. It may imply that waiting too long after the first purchase may not be optimal as the

intention to buy is already too low, while sending incentives right away may not be optimal too as they may purchase for the second time even without such incentives.

Moreover, the company has a practical constraint in its marketing strategy. While the company's general objective function is to maximize retention rate (and hence LTV), in particular for first-time buyers, the company hesitates to distribute too generous incentives. That is because some customers purchase other items even without coupons and some customers may use coupons for buying low-margin items. This concern leads us to study how to design cost-effective personalized strategies.

4 Experimental Design and Data

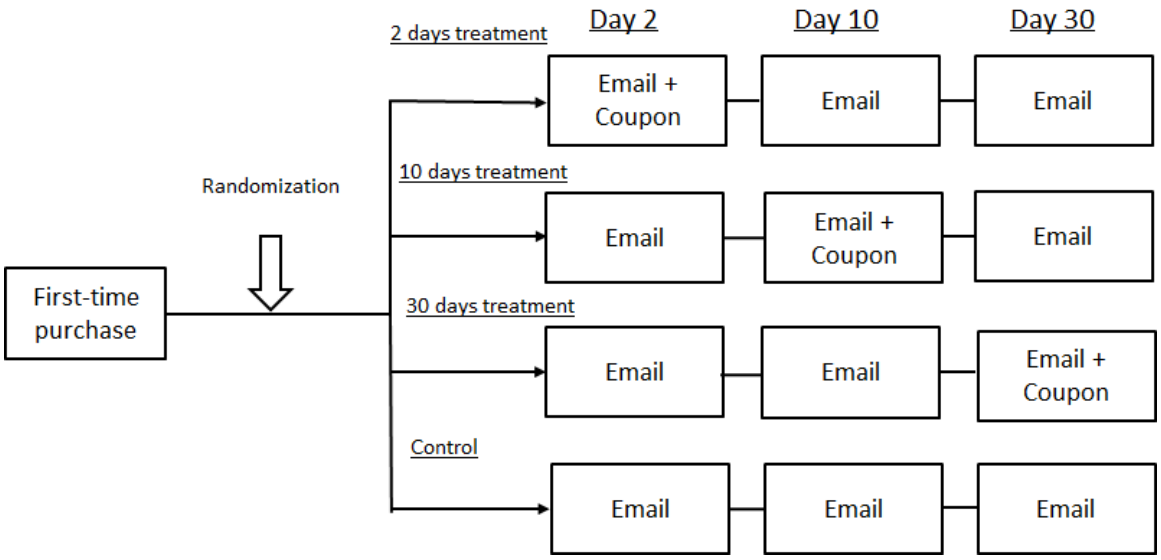
4.1 Experimental Design

In this section, we describe the design of the randomized experiment that we will use for estimating the optimal DTR later. The company conducted two experiments; one for the baseline model we will explain in Section 5.2, and another for the extension in Section 5.4. Below, we discuss the first experiment to save space. In the appendix, we discuss the experimental design of the second experiment, which is an extension of the first experiment.

The company conducted the first experiment from September 2020 to December 2020. As explained above, the company's focus is on the customers who have just made their first purchase, and there are about 150,000-200,000 first-time buyers per month during the experiment period. First-time buyers need to provide demographic information and contact information when they make a purchase. We randomly pick up those first-time buyers for the experiment.

In the first experiment, customers in the control group receive only appreciation emails, each of which contain information about how to use the platform or a generic ranking of the items sold on the platform. They receive appreciation emails at three different times: 2 days, 10 days, and 30 days after the first purchase. The customers in

Figure 2: Experimental Design



Note: The figure shows the experimental design of the first experiment where there are two actions, email only and email and coupon.

the treatment group receive the financial incentive of 1,000 points (about $10) as well as the appreciation emails.[7] There is no expiration date for those points.[8] Since providing coupons for first-time buyers was never implemented on the platform before, users did not know if they could receive coupons before they made a purchase.

In the treatment group, there are three sub-groups depending on when the customers receive the incentive. The customers in the 2-day treatment group receive the incentive 2 days after the first purchase, and they receive only the appreciation emails 10 days and 30 days after the first purchase. The 10-day treatment and the 30-day treatment groups are similarly defined. Hence, each first-time buyer in the treatment groups receives the incentive at most once during the experimental period. Note that in Section 5, we consider the optimal DTR in this class of the strategy. Hence, our experimental design covers all possible combinations of treatments to estimate the optimal DTR.

The randomization is done at the user level. For each user, we randomly assign

[7]The company was not willing to offer %-off coupons because such a coupon can be very costly for the platform if users purchase expensive items. The design of coupons is beyond our research question.

[8]For an impact of coupon expiration dates, see, e.g., Inman and McAlister (1994).

Table 1: Treatment Allocation

	Percent
Control	39.76%
2 day	19.15%
10 day	20.50%
30 day	20.59%

Table 2: Summary Statistics

Variable	**2day**		**10day**		**30 day**		**Control**	
	mean	**sd**	**mean**	**sd**	**mean**	**sd**	**mean**	**sd**
Order (2 month)	0.388	0.487	0.379	0.485	0.375	0.484	0.366	0.482
Quantity (2 month)	1.577	5.751	1.527	4.511	1.509	4.196	1.486	4.612
Amount (2 month)	6402.4	25071.9	6159.2	19625.4	6238.1	20024.5	6135.7	24013.1
Points used (2 month)	60.1	237.4	43.4	203.4	41.3	198.6	0.0	0.0
Female	0.631	0.483	0.631	0.483	0.629	0.483	0.634	0.482
Age	30.24	12.69	30.31	12.76	30.23	12.68	30.25	12.67
Quantity: first buy	1.699	1.520	1.697	1.482	1.691	1.588	1.697	1.493
Amount: first buy	8065.3	8213.6	8127.3	8405.5	8102.7	8361.3	8089.4	8157.9
Points used: first buy	557.5	756.9	556.5	760.2	556.5	762.0	560.3	860.9
# of sessions/day 1st buy	0.697	2.016	0.698	2.032	0.703	2.049	0.704	2.047
# of PV/day before delivery	15.08	31.78	15.38	33.02	15.24	32.73	15.23	32.64
# of favorites/day (2-10 day)	0.207	1.036	0.178	1.119	0.179	1.047	0.180	1.018
# of messages (10-30 day)	32.21	37.81	32.31	37.16	31.60	37.28	31.61	37.29

Note: The first two columns report the mean and standard deviation of each variable for the treatment group. The third and fourth columns are for the control group. The last column shows the t-values of the mean comparison test between the two groups. The number of observations in each treatment is not reported due to the NDA.

her right after her first purchase to one of the four groups, i.e., the control, the 2-day treatment, the 10-day treatment, and the 30-day treatment. Since we randomize the assignment right after each first-time purchase, we do not have to re-randomize each time after the purchase. Table 1 reports the fraction of first-time buyers allocated to each of the treatment conditions and the control group. We have about 40% of users for the control and 20% of users for each treatment condition.

Table 2 reports the summary statistics of the subset of the variables we use. The table includes the outcome variables that we maximize such as the order dummy, the number of items purchased (quantity), sales amount (amount), and points used. Al-

though we sometimes consider the outcome variables measured within 3 months after the delivery for some applications, we report the summary statistics of the outcome variables measured within two months after the delivery of the first item. In the first two columns, we show the mean and standard deviation of each variable for the customers in the 2-day treatment. Similarly, the third and fourth columns are for the customers in the 10-day treatment, the fifth and sixth columns for the customers in the 30-day treatment, and the seventh and eighth columns for the customers in the control group.

The table shows that the order is greater for the three treatment groups than for the control group, and it is decreasing as the user receives the incentive later. Both quantity and sales amount follow the same pattern. Although points used for purchases also follow the same pattern, the user spends significantly more points when she received points two days after the delivery.

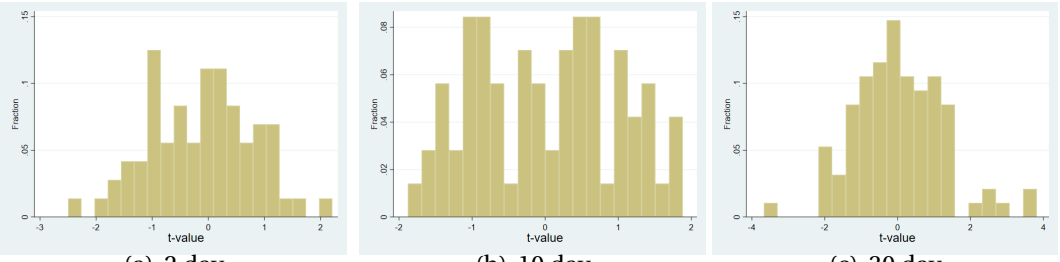
For user demographics, about 63% of users are female and the average age of users is 30. Also, we do not see any significant difference across conditions. The quantity and sales amount of the user's first purchase occasion are on average 1.7 and 8,000 across conditions. Hence, the user purchases more items and spends more for the first purchase occasion. Similarly, the user applies about 555 points for the first purchase across all conditions, which is much larger than the number of points used for the second purchase.

The last four rows in the table contain a subset of the user behavior variables. Since we use more than 100 behavioral variables based on the user access data (sessions, page views, likes, messages sent, etc.), it is not possible to report the summary statistics of all of those variables. Hence, we *randomly* pick up four variables and show their summary statistics in the table.[9]

Next, we check the balance between the treatment groups and the control group to make sure if the randomization is done accurately. Since we use more than 100 vari-

[9]Note that we do not mean those reported variables are more important than other variables. We simply choose those variables at random. This is required by the company for confidentiality issues.

Figure 3: Balance Check



(a) 2 day (b) 10 day (c) 30 day

Note: Each figure plots the histogram of t-values for a mean-comparison test between each treatment and the control group.

ables for the estimation of DTRs, we do not report the mean comparison of each variable in a table. Rather, in Figure 3, we show the histograms of the t-values for the mean comparison test between the treatment and control groups for each variable used in the estimation of DTRs. As the graph shows, there is no significant difference in mean between the control and treatment groups as the t-value is located between -2 and 2 for most of the variables.

In terms of the outcome, we consider the second-time purchase (i.e., retention) by 8 weeks after the first purchase or that by 12 weeks, and we estimate the optimal DTR to maximize them as the objective function as it is the company's main KPI. We also investigate the effects of the derived optimal DTRs on long-term outcomes to see if there are any inter-temporal substitution effects by simply shifting the timing of purchases.

4.2 Average Treatment Effect

Before we explain the estimation of DTR with the experiment data, to understand the data in more detail, we examine the average treatment effects in this section. In the appendix, we report the average treatment effects for the second experiment. In particular, we estimate the following simple linear regression:

$$y_{it} = \beta_{0t} + \beta_{1t} D_{it} + \varepsilon_{it},$$

Table 3: Average Treatment Effect

	(1) Order 1 week	(2) Order 2 week	(3) Amount 1 week	(4) Amount 2 week	(5) Quantity 1 week	(6) Quantity 2 week
Panel (A): 2 Days						
Treatment	0.018***	0.025***	96.117***	140.979***	0.030***	0.046***
	(0.001)	(0.002)	(29.437)	(40.415)	(0.005)	(0.009)
Observations	194,579	194,579	194,579	194,579	194,579	194,579
R^2	0.001	0.002	0.0002	0.0002	0.001	0.001
Panel (B): 10 Days						
Treatment	0.012***	0.016***	48.303**	55.692	0.019***	0.028***
	(0.001)	(0.002)	(21.011)	(33.279)	(0.005)	(0.007)
Observations	199,031	199,031	199,031	199,031	199,031	199,031
R^2	0.001	0.001	0.0001	0.0001	0.001	0.0005
Panel (C): 30 Days						
Treatment	0.015***	0.019***	58.146**	60.739*	0.024***	0.031***
	(0.001)	(0.001)	(19.263)	(29.387)	(0.004)	(0.006)
Observations	199,324	199,324	199,324	199,324	199,324	199,324
R^2	0.001	0.001	0.001	0.0001	0.001	0.001

Note: The first two columns report the treatment effect on whether a customer makes any purchases one week or two weeks after she received an email. The third and fourth columns report the treatment effect on total sales and the fifth and sixth columns report the treatment effect on the number of items purchased.

where y_{it} is the outcome of interest, D_{it} is the dummy for the treatment, and $t = 2, 10, 30$. We run three separate regressions for each treatment group. For each regression, we use the control group which receives no financial incentives at any t. As the outcome, we consider three variables: whether or not user i makes the second purchase (retention), the total purchase amount, and the number of items purchased.[10]

Next, we estimate the treatment effects with the outcomes measured 8 weeks and 12 weeks after the first purchase (not after each treatment). As Panel (A) of Table 4 shows, the 2 days treatment increases the retention rate by 2.4%, the 10 days treatment

[10]The estimation sample includes the users who make the second purchase before they receive incentives. We run another set of regressions where we remove all users who make a purchase before they receive coupons. The results are virtually the same as Table 3.

Table 4: Average Treatment Effect (Long Term)

	Retention	Sales	Order
Panel (A): 8 Week Outcomes			
2 days	0.024***	229.6**	0.084***
	(0.002)	(101.8)	(0.022)
10 days	0.014***	-33.0	0.025
	(0.002)	(84.7)	(0.018)
30 days	0.015***	149.9	0.045**
	(0.002)	(94.7)	(0.019)
Panel (B): 12 Week Outcomes			
2 days	0.024***	385.0***	0.056***
	(0.002)	(157.0)	(0.010)
10 days	0.011***	-90.3	0.012***
	(0.002)	(102.8)	(0.009)
30 days	0.015***	144.0	0.024***
	(0.002)	(95.6)	(0.007)
Obs.	330,302	330,302	330,302

Note: The first column reports the treatment effect on whether a customer makes any purchases within 8 weeks since her first purchase. The second column reports the treatment effects on total sales and the third column reports the treatment effects on the number of items purchased. Lastly, the fourth column displays the treatment effects on page views.

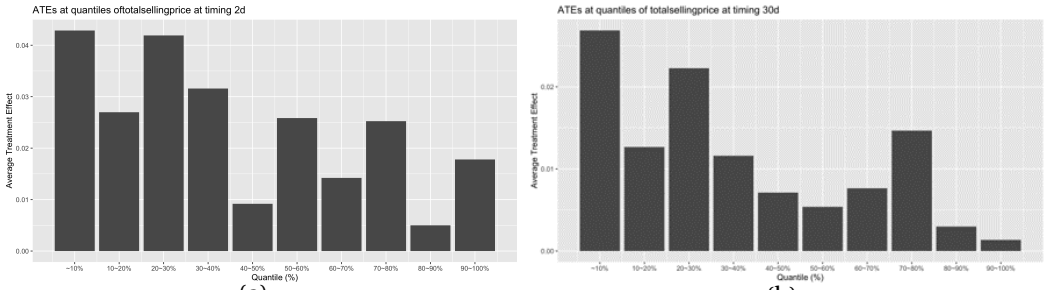
by 1.4%, and the 30 days treatment by 1.5%. Panel (B) reports similar treatment effects for the outcomes within 12 weeks. Note that the 10 days treatment effect is 1.1% for 12 weeks outcomes, which is smaller than the one for 8 weeks outcomes. This is because the users in the control group eventually increase their purchases over time and hence the treatment effect could decrease. This suggests that there is an inter-temporal substitution within a user.

In terms of total spending, the 2-day treatment increases the total spending by 1,200 JPY ($12) in 8 weeks, while the 10 days and the 30 days' treatment effects on total spending are not statistically significant. The third column reports the treatment effects of the number of items purchased. We find that the treatment effects are mostly positive for 2, 10, and 30 days treatment. Even for the 30 days treatment, the effect is positive and statistically significant. In sum, the optimal uniform strategy is to send incentives two days after the first purchase. We consider it as the baseline strategy.

Although the average treatment effects are informative to see which treatment is more effective than others on average, it may not necessarily be the case that the 2 days treatment is optimal for all users. To see how much heterogeneity exists, we now estimate the heterogeneous treatment effects by estimating interaction effects.[11]. In Figure (a) of Figure 4, we report the bar chart of the average treatment effects of the two-day treatment for each decile of the total selling price, while in Figure (b), the average treatment effects of the thirty-day treatment effects for each decile of the total selling price. For both figures, it seems the average treatment effects generally decrease as the total selling price of their first purchase increases. Although our main objective is not to identify the mechanism behind the decreasing patterns, we want to emphasize that treatment effects vary a lot across users. Significant heterogeneity of the treatment effects can be found for other conditioning variables.

[11]We report more results in the online appendix. We also estimate the causal forest (Wager and Athey (2018)) models and report the results in the appendix.

Figure 4: Heterogeneous Treatment Effects



(a) (b)

Note: For each purchase occasion, begin with a setting in which there is no product recommendation and then randomly pick up a product to recommend. Continue this process until all products in the vending machine are recommended. At each number of product recommendations, compute the machine-level sales units of products. Then, compute percentiles of the sum of inside product share for each number of recommendations (NR) across purchase occasions.

5 Dynamic Treatment Regime with Constraints

In this section, we propose our estimation strategy of the optimal dynamic treatment regime when there exist intertemporal budget constraints. We start from the discussion about the model setup and then explain how to estimate the optimal DTR.

5.1 Setup

We study the following dynamic environment. In our model, there are three periods ($t = 1, 2, 3$) and $X_t \in \mathcal{X}_t$ denotes the state variables at period t. Note that our methodology can be extended to the case with more than three periods. As we discussed above, the state variables include the user demographic information, past purchase information, browsing information, responses to past marketing activities, etc. Our methodology can easily accommodate a large number of state variables. In our application, we use more than 100 variables for X_t. The company can choose an action every period from the action set $A_t \in \mathcal{A}_t$. In our baseline application, the company either sends an incentive or not in addition to the appreciation message, and we denote $\mathcal{A}_t = \{0, 1\}$, where option 1 is sending the incentive. Later, we will extend the model to the case where the action set includes more than two options. Lastly, the final outcome is $Y \in \mathbb{R}$,

which will realize after period 3. In our case, we consider customer retention within two or three months after the first purchase.[12]

We introduce the history $H_t \in \mathscr{H}_t$ to describe the summary of the events up to period t. More precisely, we define $H_1 = X_1$, $H_2 = (X_1, A_1, X_2)$, and $H_3 = (X_1, A_1, X_2, A_2, X_3)$. Note that the history includes not only the state variables but also the actions taken up to that period. The state transition distributions are denoted by $P_{X_1}(x_1)$, $P_{X_2}(x_2|h_1, a_1)$, and $P_{X_3}(x_3|h_2, a_2)$. The final outcome is then determined by the entire history up to period 3, h_3, and the action taken in period 3, a_3, i.e., $P_Y(y|h_3, a_3)$. Thus, the model is non-Markov and hence the state transition depends not only on the previous state but also on the entire history. This implies that we may not be able to use off-the-shelf reinforcement learning algorithms that typically assume a Markov environment.

With this dynamic environment, a dynamic treatment regime (DTR) is a sequence of decision rules $\mathbf{d} = (d_1, d_2, d_3)$, where $d_t : \mathscr{H}_t \to \Delta(\mathscr{A}_t)$ is a function that maps the history up to time t into a probability distribution over actions. When a DTR $\mathbf{d}$ is applied to the above dynamic environment, the trajectory $H = (X_1, A_1, X_2, A_2, X_3, A_3, Y)$ is generated by the following process.

1. $X_1 \sim P_{X_1}$, $A_1 \sim d_1(\cdot|H_1)$
2. $X_2 \sim P_{X_2}(\cdot|H_1, A_1)$, $A_2 \sim d_2(\cdot|H_2)$
3. $X_3 \sim P_{X_3}(\cdot|H_2, A_2)$, $A_3 \sim d_3(\cdot|H_3)$
4. $Y \sim P_Y(\cdot|H_3, A_3)$

That is, the DTR generates the state variables and actions for each period, which determine the final outcome of interest Y. The resulting distribution of H is given by

$$P_{X_1}(x_1)d_1(a_1|h_1)P_{X_2}(x_2|h_1, a_1)d_2(a_2|h_2)P_{X_3}(x_3|h_2, a_2)d_3(a_3|H_3)P_Y(y|h_3, a_3).$$

[12] It is straightforward to apply our method to the case where the outcome is the cumulated rewards, e.g., $\sum_t \beta^t Y_t$.

We denote the distribution by $P_{\mathbf{d}}$ and the expectation with respect to $P_{\mathbf{d}}$ by $E_{\mathbf{d}}$.

Suppose that we observe data $\{H^{(i)}\}_{i=1}^{n}$ of n individuals, where $H^{(1)}, ..., H^{(n)}$ are independently and identically generated by the above process with some DTR $\mathbf{d}^0$. In our empirical setting, $\mathbf{d}^0$ is the random assignment policy used by the experiment and is known to us.[13] We denote the distribution of the observed trajectory $H^{(i)}$ by P and the expectation with respect to P by E.

Our learning objective is to use the data $\{H^{(i)}\}_{i=1}^{n}$ to choose a DTR that maximizes the expected value of Y over a (possibly constrained) class of DTRs $\mathscr{D}$:

$$\mathbf{d}^* \in \arg\max_{\mathbf{d}\in\mathscr{D}} E_{\mathbf{d}}[Y].$$

In order to identify the value $E_{\mathbf{d}}[Y]$ for every DTR $\mathbf{d} \in \mathscr{D}$, we assume the following overlap condition.

Assumption 1. *For all* $\mathbf{d} \in \mathscr{D}$, t, *and* $(a_t, h_t) \in \mathscr{A}_t \times \mathscr{H}_t$, *if* $d_t(a_t|h_t) > 0$, *then* $d_t^0(a_t|h_t) > 0$.

The assumption implies that one cannot estimate the optimal policy unless there is a non-zero probability of observing (a_t, h_t) in the data-generating process, $(\mathbf{d}^0)$.

5.2 Estimation Approaches without Constraints

We consider two approaches to estimating the optimal dynamic treatment strategy: Q-Learning and Backward Outcome Weighted Learning (BOWL). We first discuss two approaches without any constraints and then extend to the case with constraints in the next subsection.

[13] Note that our experiment randomly divided the customers into four groups who would receive different action profiles (A_1, A_2, A_3) prior to the first treatment. Implementing this randomization design is equivalent to implementing a sequential randomization design (or a DTR) that randomly determines each customer's treatment prior to each period based on the past treatment profile.

5.2.1 Q-Learning

The first approach to estimating the optimal DTR is to use Q-learning (e.g., Murphy, Lynch, Oslin, McKay, and TenHave (2007)). This approach first models a Q-function, which is a conditional mean of the outcome given the history, by a parametric, semi-parametric, or nonparametric function, and then derives the optimal DTR by maximizing the Q-function. In our application, we try several machine learning models such as LASSO, Random Forest, and GBM to estimate Q-functions.[14]

In the dynamic setup we consider, the Q-function is the expected outcome as a function of the state variables and actions. Thus, we can define the Q-functions sequentially as follows:

$$
\begin{aligned}
Q_3(h_3, a_3) &= E[Y|H_3 = h_3, A_3 = a_3], \\
Q_2(h_2, a_2; d_3) &= E_{d_3}[Y|H_2 = h_2, A_2 = a_2] \\
&= E[\sum_{a_3 \in \mathscr{A}_3} d_3(a_3|h_2, a_2, X_3) Q_3((h_2, a_2, X_3), a_3)|H_2 = h_2, A_2 = a_2], \\
Q_1(h_1, a_1; d_2, d_3) &= E_{d_2, d_3}[Y|H_1 = h_1, A_1 = a_1] \\
&= E[\sum_{a_2 \in \mathscr{A}_2} d_2(a_2|h_1, a_1, X_2) Q_2((h_1, a_1, X_2), a_2; d_3)|H_1 = h_1, A_1 = a_1].
\end{aligned}
$$

Letting $\mathscr{D}$ be the set of all DTRs, a (deterministic) optimal DTR $\mathbf{d}^*$ that maximizes $E_{\mathbf{d}}[Y]$ can be solved backward:

$$
\begin{aligned}
d_3^*(h_3) &\in \arg\max_{a_3 \in \mathscr{A}_3} Q_3(h_3, a_3), \\
d_2^*(h_2) &\in \arg\max_{a_2 \in \mathscr{A}_2} Q_2(h_2, a_2; d_3^*), \\
d_1^*(h_1) &\in \arg\max_{a_1 \in \mathscr{A}_1} Q_1(h_1, a_1; d_2^*, d_3^*),
\end{aligned}
$$

[14] In principle, one can apply the deep reinforcement learning (DRL) method to estimate the Q-function. There are a few challenges in our application. First, typically DRL is applied to the case with a longer time horizon. Since our application has only three periods, the benefits of using DRL may be limited. Second, typical applications of DRL consider a Markov situation, while our application considers a non-Markov situation. Hence, widely-used DRL algorithms may not work.

and the optimal Q-function (Q_1^*, Q_2^*, Q_3^*) is given by

$$Q_3^*(h_3, a_3) = Q_3(h_3, a_3), \;\; Q_2^*(h_2, a_2) = Q_2(h_2, a_2; d_3^*), \;\; Q_1^*(h_1, a_1) = Q_1(h_1, a_1; d_2^*, d_3^*).$$

We can solve this problem by the following Q-learning algorithms that estimate Q-functions backward, and obtain the period-t decision rule maximizing the period-t Q-functions. Now we write it down as an algorithm below.

Algorithm 1 (Q-Learning).

1. Conduct a (possibly nonparametric or semiparametric) regression of Y on H_3 and A_3. Let $\hat{Q}_3$ be the estimated function, and let $\hat{d}_3(h_3) \in \arg\max_{a_3 \in \mathscr{A}_3} \hat{Q}_3(h_3, a_3)$.

2. Conduct a regression of $\hat{Q}_3(H_3, \hat{d}_3(H_3))$ on H_2 and A_2. Let $\hat{Q}_2$ be the estimated function, and let $\hat{d}_2(h_2) \in \arg\max_{a_2 \in \mathscr{A}_2} \hat{Q}_2(h_2, a_2)$.

3. Conduct a regression of $\hat{Q}_2(H_2, \hat{d}_2(H_2))$ on H_1 and A_1. Let $\hat{Q}_1$ be the estimated function, and let $\hat{d}_1(h_1) \in \arg\max_{a_1 \in \mathscr{A}_1} \hat{Q}_1(h_1, a_1)$.

Steps 2 and 3 are motivated by the following observation

$$Q_2(H_2, A_2; \hat{d}_3) = E[Q_3(H_3, \hat{d}_3(H_3))|H_2, A_2], \;\; Q_1(H_1, A_1; \hat{d}_2, \hat{d}_3) = E[Q_2(H_2, \hat{d}_2(H_2); \hat{d}_3)|H_1, A_1].$$

Since $Q_3(\cdot,\cdot)$ and $Q_2(\cdot,\cdot)$ are unobserved in data, Steps 2 and 3 use the estimated values from the previous step $\hat{Q}_3(\cdot,\cdot)$ and $\hat{Q}_2(\cdot,\cdot)$, respectively, as the outcome variable in the regressions. Since the standard Q-learning chooses the action with the largest estimated value at each step, it may result in the overestimation of the Q-value for the chosen action, i.e., $\hat{Q}_t(H_t, \hat{d}_t(H_t))$ (e.g., Lan, Pan, Fyshe, and White (2020)), which affects the optimization in the next step. To reduce the bias, we do sample splitting for Q-learning. That is, at each step we split the sample into two subsamples, use one subsample to choose the action, and use the other to estimate the Q-value for the chosen

action.[15]

5.2.2 Outcome Weighted Learning (OWL)

Another approach to estimating the optimal DTR we consider is the outcome-weighted learning algorithm (OWL) proposed by Zhao, Zeng, Rush, and Kosorok (2012) for a static setting. A key observation of Zhao, Zeng, Rush, and Kosorok (2012) is to formulate optimal policy learning as a weighted classification problem. This transformation allows one to use existing classification algorithms to learn the optimal DTR. Zhao, Zeng, Laber, and Kosorok (2015) extend the idea of OWL to dynamic settings, called Backward outcome-weighted learning algorithm (BOWL). In BOWL, a classification problem is sequentially solved from the last period going backwards towards the first period. Since BOWL directly maximizes over a class of DTRs a nonparametric estimator of the expected long-term outcome, it is different than regression-based methods such as Q-learning, which indirectly attempts such maximization and relies on the correct model specification. To see how it works, observe first that if $\mathbf{d}$ is deterministic, we can write

$$\begin{aligned} E_{d_3}[Y|H_3 = h_3] &= E\left[E[Y|H_3 = h_3, A_3 = d_3(H_3)]|H_3 = h_3\right] \\ &= E\left[\frac{Y\mathbf{1}[A_3 = d_3(H_3)]}{d_3^0(A_3|H_3)}|H_3 = h_3\right]. \end{aligned} \tag{5.1}$$

Again, $\mathbf{d^0}$ in our empirical setting is the random allocation rule of the treatments, which is known. Note that the first equality follows from the law of iterative expectation, and we fix $A_3 = d_3(H_3)$. The second equality is based on the change of the variable, fre-

[15] An alternative approach is to use the importance-weighted outcomes $\frac{Y\mathbf{1}[A_3=\hat{d}_3(H_3)]}{d_3^0(A_3|H_3)}$ and $\frac{Y\mathbf{1}[(A_2,A_3)=(\hat{d}_2(H_2),\hat{d}_3(H_3)]}{d_2^0(A_2|H_2)d_3^0(A_3|H_3)}$ instead of the estimated Q-values $\hat{Q}_3(H_3,\hat{d}_3(H_3))$ and $\hat{Q}_2(H_2,\hat{d}_2(H_2))$ in Steps 2 and 3, respectively.

quently used in Thompson sampling and propensity score matching literature.[16]

If there are no constraints in the class of DTRs, the optimal rule in period 3 then satisfies

$$d_3^*(h_3) \in \arg\max_{a_3 \in \mathscr{A}_3} E\left[\frac{Y\mathbf{1}[A_3 = a_3]}{d_3^0(A_3|H_3)}|H_3 = h_3\right]$$

for every $h_3 \in \mathscr{H}_3$. Therefore, $d_3^* \in \arg\max_{d_3} E\left[\frac{Y\mathbf{1}[A_3=d_3(H_3)]}{d_3^0(A_3|H_3)}\right]$. We can similarly define d_2^* and d_1^*. Since the action is binary, the above maximization problem is equivalent to the following *minimization* problem:

$$d_3^* \in \arg\min_{d_3} E\left[\frac{Y\mathbf{1}[A_3 \neq d_3(H_3)]}{d_3^0(A_3|H_3)}\right]. \tag{5.2}$$

The objective function in each problem can be viewed as a weighted misclassification error. We can define similar minimization problems for $t = 2$ and 1.[17] Hence, we can obtain the optimal DTR by sequentially applying any classification algorithms, which have been extensively studied in machine learning. Thus, we can easily apply existing machine learning algorithms to estimate the optimal DTR.[18]

BOWL estimates the solution by minimizing the sample analog of the objective function (plus the penalty) backward for $t = 3, 2, 1$, which is basically solving a classifi-

[16]For $t = 1$ and 2, we can write as follows:

$$E_{d_2,d_3}[Y|H_2 = h_2] = E\left[\frac{Y\mathbf{1}[(A_2,A_3) = (d_2(H_2), d_3(H_3))]}{d_2^0(A_2|H_2)d_3^0(A_3|H_3)}|H_2 = h_2\right]$$
$$E_{\mathbf{d}}[Y|H_1 = h_1] = E\left[\frac{Y\mathbf{1}[(A_1,A_2,A_3) = (d_1(H_1), d_2(H_2), d_3(H_3))]}{d_1^0(A_1|H_1)d_2^0(A_2|H_2)d_3^0(A_3|H_3)}|H_1 = h_1\right].$$

[17]The minimization problem for period 2 and period 1 is as follows:

$$d_2^* \in \arg\min_{d_2} E\left[\frac{Y\mathbf{1}[A_3 = d_3^*(H_3)]}{d_2^0(A_2|H_2)d_3^0(A_3|H_3)}\mathbf{1}[A_2 \neq d_2(H_2)]\right],$$
$$d_1^* \in \arg\min_{d_1} E\left[\frac{Y\mathbf{1}[(A_2,A_3) = (d_2^*(H_2), d_3^*(H_3))]}{d_1^0(A_1|H_1)d_2^0(A_2|H_2)d_3^0(A_3|H_3)}\mathbf{1}[A_1 \neq d_1(H_1)]\right].$$

[18]Since the function is non-convex and discontinuous, Zhao, Zeng, Laber, and Kosorok (2015) propose to replace the 0–1 loss $\mathbf{1}[A_t \neq d_t(H_t)]$ with a hinge loss $\phi(A_t f_t(H_t)) = \max(1 - A_t f_t(H_t), 0)$, where $f_t : \mathscr{H}_t \to \mathbb{R}$ is the decision function so that $d_t(h_t) = \text{sign}(f_t(h_t))$. Zhao, Zeng, Laber, and Kosorok (2015) show that the change in the loss function does not change the solution to the minimization problems.

cation problem sequentially. Zhao, Zeng, Laber, and Kosorok (2015) show that the obtained DTRs are consistent, and provide finite sample bounds for the errors using the estimated rules. Their simulation results suggest that BOWL outperforms Q-learning.

5.3 Estimation Approaches with Constraints

In this subsection, we extend the estimation of the optimal DTR to the case where constraints are imposed. For expositional simplicity, we start from the single-period OWL problem. We then extend to a multi-period case. We will discuss how to extend the Q-learning in the appendix.

5.3.1 Single-period Optimization

To get the basic idea of how we deal with constraints, we first consider the single-period constrained maximization problem:

$$\max_d E_d[Y] \text{ s.t. } E_d[C] \leq B,$$

where C is the cost variable, which is generated together with the outcome Y, and B is the per-person budget. In our empirical context, C is the coupon amount that the customer uses, and the company has a budget ceiling of B on how much can be spent for each customer. We can also consider capacity constraints. For example, if the company has a limited number of coupons, we can denote the constraint as $E_d[A] \leq B$, where A is a dummy variable indicating whether the customer receives a coupon, and $B \in [0,1]$ is the capacity on the fraction of customers who can receive a coupon.

Under certain conditions, it is straightforward to show that an optimal rule is a threshold strategy. That is, letting $\beta(h)$ and $\gamma(h)$ denote the conditional average treatment effects (CATEs) on the outcome and cost, respectively, i.e., $\beta(h) = E[Y|H=h, A=1] - E[Y|H=h, A=0]$ and $\gamma(h) = E[C|H=h, A=1] - E[C|H=h, A=0]$ for $h \in \mathscr{H}$, the

following rule is optimal:

$$d^*(h) = \mathbf{1}[\beta(h) \geq \lambda^* \gamma(h)], \tag{5.3}$$

where λ^* satisfies $E_{d^*}[C] = B$ (see Appendix A.1). If $\gamma(h) > 0$ for all h, the rule is written as $d^*(h) = \mathbf{1}[\beta(h)/\gamma(h) \geq \lambda^*]$. In other words, the optimal rule assigns the treatment to individuals from those with the highest ratios between the CATEs on the outcome and cost until the capacity is reached.

Now, we consider how to solve the constrained optimization problem. The key idea to solve the constrained optimization problem is to transform the problem into an unconstrained problem by introducing a shadow price as follows.

$$d^*_\lambda \in \arg\max_d E_d[Y - \lambda C]$$

for given $\lambda \geq 0$. It is straightforward to see that an optimal policy targets a customer if and only if $E[Y - \lambda C | H = h, A = 1] - E[Y - \lambda C | H = h, A = 0] \geq 0$. The solution is

$$\begin{aligned} d_\lambda(h) &= \mathbf{1}[E[Y - \lambda C | H = h, A = 1] - E[Y - \lambda C | H = h, A = 0] \geq 0] \\ &= \mathbf{1}[\beta(h) \geq \lambda \gamma(h)]. \end{aligned} \tag{5.4}$$

Therefore, comparing equation (5.3) and equation (5.4), observe that $d^* = d_{\lambda^*}$. In other words, the solution to the unconstrained problem indeed can find the optimal strategy that we derived above. Note that $E_{d_\lambda}[C] = E[E[C|H, A = 0] + \mathbf{1}[\beta(H) \geq \lambda \gamma(H)]\gamma(H)]$ is decreasing in λ. Using this property, it is easy to obtain the optimal constrained rule by the following modified OWL algorithm:

Algorithm 2 (Constrained OWL)

1. For given $\lambda \geq 0$, apply a single-period OWL with the outcome set to $Y - \lambda C$ to obtain d_λ.

2. Find λ^* such that $E_{d_{\lambda^*}}[C] = B$.

For the second step, we need to evaluate the expected cost under the policy d_λ. We can use existing off-policy policy evaluation (OPE) methods to estimate $E_{d_\lambda}[C]$. We will explain the evaluation method in more detail in Section 5.6.

It is straightforward to find λ^* in the second step thanks to the monotonicity. Thus, we can effectively convert the original constrained maximization problem to the loop through unconstrained maximization problems.

Although the algorithm above uses OWL for deriving the optimal strategy, we can also use Q-learning to estimate the optimal policy. To do so, we can simply use the Q-learning approach in the first step to obtaining d_λ.

5.3.2 Multi-period Optimization

We extend the above idea to the problem with multiple periods. For convenience, we describe the method in a two-period setup, but the model can be easily extended to more than two periods. Our empirical application has three periods. We consider the following constrained maximization problem:

$$\max_{(d_1,d_2)} E_{d_1,d_2}[Y] \text{ s.t. } E_{d_1,d_2}[C] \leq B, \tag{5.5}$$

where B is the per-person budget. Note that the budget constraint is not period-specific, but an inter-temporal one. Similarly to the static problem, we can show the following proposition.

Proposition 1. *Under suitable conditions, there exists* $(\lambda_1^*, \lambda_2^*)$ *such that the following threshold strategy is a solution to the problem 5.5:*

$$d_2^*(h_2) = \mathbf{1}[\beta_2(h_2) \geq \lambda_2^* \gamma_2(h_2)], \tag{5.6}$$

$$d_1^*(h_1) = \mathbf{1}[\beta_1(h_1; d_2^*) \geq \lambda_1^* \gamma_1(h_1; d_2^*)], \tag{5.7}$$

where $\beta_2(h_2) = Q_2(h_2, a_2 = 1) - Q_2(h_2, a_2 = 0)$ *and* $\beta_1(h_1; d_2) = Q_1(h_1, a_1 = 1; d_2) - Q_1(h_1, a_1 =$

$0; d_2)$ *and* $Q_t(h_t, a_t) = E[Y|H_t = h_t, A_t = a_t]$. $\gamma_2(h_2)$ *and* $\gamma_1(h_1; d_2)$ *are analogously defined for the cost* C.

Proof. See Appendix A.2. ■

Again, as in the static optimization problem, we can solve the constrained problem by transforming the problem into an inter-temporal unconstrained problem by introducing shadow values $\lambda = (\lambda_1, \lambda_2)$. To do so, we start by considering the following period-2 problem:

$$d_{2,\lambda_2} \in \arg\max_{d_2} E_{d_1^0, d_2}[Y - \lambda_2 C],$$

where d_1^0 is the data-generating rule. The solution to this problem is the same as the static one.

$$\begin{aligned} d_{2,\lambda_2}(h_2) &= \mathbf{1}[E[Y - \lambda_2 C|H_2 = h_2, A_2 = 1] \geq E[Y - \lambda_2 C|H_2 = h_2, A_2 = 0]] \\ &= \mathbf{1}[\beta_2(h_2) \geq \lambda_2 \gamma_2(h_2)]. \end{aligned} \tag{5.8}$$

Given d_{2,λ_2}, now consider the following period-1 problem:

$$d_{1,\lambda_1,\lambda_2} \in \arg\max_{d_1} E_{d_1, d_{2,\lambda_2}}[Y - \lambda_1 C].$$

Since $E_{d_1, d_{2,\lambda_2}}[Y - \lambda_1 C] = E[E_{d_{2,\lambda_2}}[Y - \lambda_1 C|H_1, A_1 = d_1(H_1)]]$, the solution to this problem is given by

$$\begin{aligned} d_{1,\lambda_1,\lambda_2}(h_1) &= \mathbf{1}[E_{d_{2,\lambda_2}}[Y - \lambda_1 C|H_1 = h_1, A_1 = 1]] \geq E_{d_{2,\lambda_2}}[Y - \lambda_1 C|H_1 = h_1, A_1 = 0]] \\ &= \mathbf{1}[\beta_1(h_1; d_{2,\lambda_2}) \geq \lambda_1 \gamma_1(h_1; d_{2,\lambda_2})] \end{aligned} \tag{5.9}$$

Therefore, comparing equations (5.6)–(5.7) and equations (5.8)–(5.9), observe that $(d_1^*, d_2^*) = (d_{1,\lambda_1^*,\lambda_2^*}, d_{2,\lambda_2^*})$. Thus, the optimal threshold strategy can be obtained by correctly specifying the shadow costs λ_1 and λ_2. This motivates us to use the following modified BOWL algorithm:

Algorithm 4. (Constrained BOWL)

1. For given $(\lambda_1, \lambda_2) \in \mathscr{R}_+^2$, apply BOWL with the outcomes for periods 1 and 2 set to $Y - \lambda_1 C$ and $Y - \lambda_2 C$ to obtain $\mathbf{d}_{\lambda_1,\lambda_2} = (d_{1,\lambda_1,\lambda_2}, d_{2,\lambda_2})$.

2. Find (λ_1, λ_2) that maximizes $E_{\mathbf{d}_{\lambda_1,\lambda_2}}[Y]$ subject to the constraint that $E_{\mathbf{d}_{\lambda_1,\lambda_2}}[C] \leq B$.

As in the static case, for the second step, we can use existing OPE methods to estimate $E_{\mathbf{d}_{\lambda_1,\lambda_2}}[Y]$ and $E_{\mathbf{d}_{\lambda_1,\lambda_2}}[C]$ and we will explain the evaluation methods in Section 5.6.

There are a few remarks in order. First, the algorithm can also be applied to Q-learning. For the first step, we can apply Q-learning to estimate $Y - \lambda_1 C$ and $Y - \lambda_2 C$ for a given pair of λ_1 and λ_2. Then in the second step, we can search for the optimal shadow values. In the appendix, we will explain more details. Second, unlike the static model, the monotonicity of $E_{\mathbf{d}_{\lambda_1,\lambda_2}}[Y]$ and $E_{\mathbf{d}_{\lambda_1,\lambda_2}}[C]$ with respect to λ_1 and λ_2, respectively, may not hold. Hence, we have to use a grid search to find the optimal λ_1 and λ_2.

5.4 Extension: Multiple Actions

In the previous section, we consider the model where the platform's action set is a binary one, i.e., sending a message with a coupon or without it. In this section, we extend the model to the case where the action space includes more than two options.

Note that it is straightforward to extend the Q-learning model in the previous section to the case with multiple actions. Hence, we focus on how to extend the BOWL model. This extension is important for both methodological and substantive purposes. Methodologically, since BOWL is based on the idea of the sequential binary classification problem, it is not straightforward to extend the model to multiple action cases, i.e., $\mathscr{A}_t = \{0, 1, ..., A\}$. Substantively, our extension considers three options: (i) sending an appreciation email with a coupon, (ii) sending an appreciation email only, and (iii) no emails. By comparing the effect of the first option with the effect of the second option, one can decompose the effect of incentives and the mere effect of messages.

We propose a method to extend the baseline BOWL approach. Intuitively, our algorithm identifies the optimal policy by running a series of round-robin tournaments among actions to determine the optimal DTR.

Algorithm 5. (Constrained BOWL with multiple actions)
For each $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_3) \in \Lambda$ (the set of grid points on $\mathbb{R}^3_+$), repeat the following steps.

Step 1. Start from the last period $t = 3$. If the set of feasible actions is a singleton, say $a \in A$, then set $d_3(h_3; \lambda_3) = a$ for each history, h_3. If the set of feasible actions contains two or more actions, repeat the following procedure:

1. For each combination $(a, a') \in \mathscr{A}_3 \times \mathscr{A}_3$ with $a < a'$, apply the binary single-period OWL to the sub-sample with $A_3 \in \{a, a'\}$. This determines whether a is better than a'. We say this case as a winning.

2. For each history h_3, use the majority rule to choose the optimal action among $\mathscr{A}_3$. That is, we choose the action with highest winning probability.[19]

Step 2. Repeat the following procedure backward for $t = 2, 1$. If the set of feasible actions is a singleton, set that action for $d_t(h_t; \lambda_t, ..., \lambda_3)$. If the set of feasible actions contains two or more actions, repeat the following procedure:

1. For each combination a, a' with $a < a'$, apply the binary single-period OWL to the sub-sample with $A_t \in \{a, a'\}$ given $d_{t+1}(\cdot; \lambda_{t+1}, ..., \lambda_3)$. This determines whether a is better than a'. We say this case as a winning.

2. For each h_t, use the majority rule to choose the optimal action.

Once we obtain the DTR $\mathbf{d}(\boldsymbol{\lambda}) = (d_1(\cdot; \lambda_1, ..., \lambda_3), ..., d_3(\cdot; \lambda_3))$ for each $\boldsymbol{\lambda} \in \Lambda$, choose $\boldsymbol{\lambda}^*$

[19] One may ask why we do not simply apply multi-class classification for the multiple action case. First, the transformation from the maximization problem to the minimization problem does not generally work for multi-class case. Second, in many computer software packages available in R or Python, the multi-class classification is actually implemented by reducing the problem to multiple binary classification problems as we do.

such that

$$\boldsymbol{\lambda}^* \in \arg\max_{\boldsymbol{\lambda}\in\Lambda} E_{\mathbf{d}(\boldsymbol{\lambda})}[Y] \text{ s.t. } E_{\mathbf{d}(\boldsymbol{\lambda})}[C] \leq B.$$

Here, we use an OPE method to estimate $E_{\mathbf{d}(\boldsymbol{\lambda})}[Y]$ and $E_{\mathbf{d}(\boldsymbol{\lambda})}[C]$, and use the estimates in the maximization problem.

5.5 Discussion

Before we show the results in the next section, we discuss some modeling assumptions.

Comparison of Two Approaches Q-learning is an indirect approach in the sense that one needs to estimate the Q-function first and then derive the optimal strategy. By contrast, BOWL is a direct approach because one can derive the optimal strategy by solving the classification error minimization problem. Hence, there is no interim step.

The performance of the Q-learning approach depends on how well Q-functions are approximated. The approximation of the Q-function depends on how well the data covers the action-state space. In other words, if some states never arise in the data, then its value is extrapolated based on the functional form. In other words, one may need sufficiently large experimental data.

The performance of the BOWL approach relies on classification accuracy. Although non-linear classification algorithms such as deep neural networks have high classification accuracy, they may take a long time when state space is large. Hence, one may need to rely on a linear classification algorithm.

Dynamic Model One may wonder whether a dynamic model using dynamic programming is necessary given that our application is basically a three-period model. In other words, is it possible to estimate the optimal strategy in period one.

First, our proposed model can be easily extended to more than three periods. For problems with longer periods, it easily becomes untractable to solve without dynamic programming due to too many cases. Second, in our model, two components, not just

one, create dynamics. The policy in the current period affects the users' behaviors and the remaining budget, which in turn changes the optimal policy in the next period. Hence, it is appropriate to consider a dynamic model.

More precisely, a naive, static approach is to construct a policy d_t only using the current data on (H_t, A_t, Y) separately for each period. This implicitly assumes that if the estimated policy $\hat{d}_t$ is actually implemented, the actions in the future periods would be determined by the experimental policy, not by the estimated policy $(\hat{d}_{t+1}, ..., \hat{d}_T)$. That is, it ignores the policy in future periods. Furthermore, it is hard to satisfy the inter-temporal budget constraint with this approach, since the policy in each period is separately optimized.

Another static approach is to construct a policy that determines the action profile $(A_1, ..., A_T)$ based on the initial state X_1, using the data on $(X_1, A_1, ..., A_T, Y)$. This approach wastes information on the updated states $(H_2, ..., H_T)$, potentially leading to worse performance than a dynamic approach.

Deep Reinforcement Learning This paper proposes two methods of estimating DTR with intertemporal budget constraints using Q-learning and BOWL, but we do not consider deep reinforcement learning to estimate DTR. One reason that we do not do so is that there is as long as we are aware, no explicit reinforcement learning algorithm that can accommodate intertemporal constraints easily. Also, another reason is that there seem to be no established (off-policy, offline) deep reinforcement learning algorithms that can be used for learning the optimal dynamic treatment regime under a non-Markov decision process. Since our setup of retention management for first-time buyers necessitates a non-Markov strategy, it is important to consider a non-stationary dynamic programming problem.

5.6 Evaluation Methods

To see if the proposed methods work, we use the method of "off-policy policy evaluation" (OPE). OPE evaluates the performance of hypothetical policies leveraging only offline log data, i.e. the experiment data in our case. To do so, we use the inverse propensity weighting approach (IPW)(see. e.g., Precup, Sutton, and Singh (2000)).[20] Let $\mathbf{A}^* = (d_1^*(H_1), d_2^*(H_2), d_3^*(H_3))$ denote the optimal actions derived from the estimated policy in the previous section and $\pi_t = d^0(A_t|H_t)$ denote the propensity score (i.e., treatment assignment probabilities in the randomized experiment in our case). Then IPW estimates the value of the optimal DTR $\mathbf{d}^*$ as follows:

$$\hat{V}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1\{A_{i1} = A_{i1}^*\}1\{A_{i2} = A_{i2}^*\}1\{A_{i3} = A_{i3}^*\}}{\pi_{i1}\pi_{i2}\pi_{i3}} R_i, \tag{5.10}$$

where R_i is either the outcome Y_i or cost C_i. Hence, we can evaluate both $E_{d^*}[Y]$ and $E_{d^*}[C]$, which we need for Step 2 of Algorithm 2 and Algorithm 4. The IPW-based evaluation has been used in existing papers such as Hitsch and Misra (2018) and Yoganarasimhan, Barzegary, and Pani (2020). When the propensity scores are known, the IPW-based OPE is unbiased and consistent, but its variance tends to be large due to extreme weights (Hirano, Imbens, and Ridder (2003)).

6 Results

6.1 Results of the Baseline Model

We use the first experimental data to derive the optimal DTR under the budget constraint for the baseline model. Remember that the baseline model has two actions, sending a coupon with the appreciation email and sending only the email.

[20] We also tried the doubly robust approach (DR) as in Jiang and Li (2016). The results are pretty similar

6.1.1 Comparison of Classification Algorithms

For Q-learning, estimation of Q-functions can be done with various machine learning algorithms such as LASSO, Random Forest, LightGBM (Light Gradient Boosting Machine), or deep learning.[21] For BOWL, any classification algorithm can be used to solve the weighted classification error minimization problem. For example, one can use SVM (Support Vector Machine), Logistic Regression, Random Forest, and SGDC (Stochastic Gradient Descent Classifier).[22]

To see which algorithm works better, we compare the performance of different algorithms without imposing any budget constraints. Based on the results, we choose to use SGDC for BOWL, and LASSO for Q-learning as those algorithms provide sufficiently high retention rates with smaller costs per user.[23] Those algorithms are attractive also in terms of computational time.

6.1.2 Off-line Policy Evaluation Results

Table 5 summarizes the results of the OPE for BOWL, Q-learning, and the baseline policy. We consider four different levels of costs, 0, 20, 25, and 30 JPY (selected by the company), and two baseline models ("Two-days-after" policy and no-offer policy). The two-days-after policy sends coupons for all users two days after the first purchase unless the user has not made the second purchase. Note that the average treatment effects we estimate in Section 4.2 indicate that it is optimal to send incentives 2 days after the first purchase if we do not personalize. In terms of outcomes, we display the uplift probabilities of the second purchase relative to the no-incentive policy through the fifth purchase within three months after the first purchase. The table also reports

[21] LightGBM is a tree-based algorithm using distributed gradient boosting framework. LightGBM runs faster than XGBoost, a well-known gradient boosting machine while maintaining a high level of prediction accuracy.

[22] SGDC is a linear classifier including SVM, optimized by the stochastic descent gradient machine. SGDC-l2 uses L2-norm (as LASSO) for regularization, while SGDC-elastic net uses both L1 (as Ridge Regression) and L2 norms for regularization.

[23] We try LASSO, Random Forest, and LightGBM for Q-Learning and SGDC and Random Forest for BOWL. The results are available upon request.

the total cost of each policy and the return on advertising spending (ROAS). The company uses ROAS as a key KPI as it is easy to compare the performance across different settings.

The baseline policy of sending coupons two days after the purchase achieves a 1.85% of uplift for the second purchase, and the baseline policy costs 49.43 JPY. Without any budget constraint (i.e., setting cost $= \infty$), we find that the uplift of the second purchase is 1.73% for BOWL and 1.86% for Q-learning. Hence, the uplift is slightly lower for BOWL than the baseline and that of Q-learning is virtually the same as the baseline. In terms of costs, BOWL spends \$48.43 per user, which is \$1 less than Q-learning and Two-days-after. Taking both retention and cost into account, the ROAS of Q-learning and Two-days-after is 342%, while the ROAS of BOWL is 314%.

Next, when we impose budget constraints (\$20, 25 or 30), both BOWL and Q-learning have lower retention rates, but costs are much smaller. For BOWL, the uplift for the retention rate is 1.46% when the budget constraint is \$30. The actual cost of the optimal DTR is \$35.8.[24] By contrast, for Q-learning, the uplift is just 0.76% and the cost is \$28.2 when the budget is set at \$30. These lead to BOWL generating greater ROAS than Q-learning (388% vs. 213%). When there is a budget constraint, ROAS for BOWL is much greater than Q-learning and the baseline policy.

In Table 6, we show how the optimal policy allocates incentives across days for each case. First of all, the baseline Two-days-after policy sends coupons to 89% of the first-time buyers at day 2 because 11% of users have already purchased before they receive coupons. We find that the optimal DTR under BOWL and Q-learning send most of the coupons at day 2 when there is no budget ceiling. When there is a budget constraint, it is no longer optimal to send all coupons at day 2. Rather, the optimal policies, especially BOWL-based policies, send more coupons at day 30.

In sum, we find that the optimal DTR by BOWL under the budget constraint gives cost-effective personalized retention strategies compared to other approaches.

[24]The reason why the actual costs are sometimes bigger than the budget is that we use a holdout sample for OPE.

Table 5: Off-Policy Uplift Performance (Baseline Model) (Two Months)

	2nd	3rd	Cost	ROAS
BOWL: SGDC, cost=20	1.16%	0.30%	29.97	401%
BOWL: SGDC, cost=30	1.46%	0.32%	35.81	388%
BOWL: SGDC, cost=∞	1.73%	0.32%	47.96	314%
Q-learning: lasso, cost=20	0.59%	0.32%	17.97	431%
Q-learning: lasso, cost=30	0.76%	0.05%	28.24	213%
Q-learning: lasso, cost=∞	1.86%	0.40%	48.95	342%
Two days after	1.86%	0.40%	48.95	342%

Note: The table reports the uplift of the second, third, fourth, and fifth purchases within three months after their first purchase, compared to the case with no incentives. Total costs are measured in Japanese Yen and ROAS is the return on advertising spending.

Table 6: Offline Optimal Policies (Baseline Model)

	2 day	10 day	30 day
BOWL: SGDC, cost= ∞	87.80%	0.01%	0.74%
BOWL: SGDC, cost=20	30.61%	0.00%	46.23%
BOWL: SGDC, cost=30	47.74%	0.01%	32.36%
Q-learning: lasso, cost=∞	88.99%	0.00%	0.00%
Q-learning: lasso, cost=20	29.46%	17.05%	14.07%
Q-learning: lasso, cost=30	36.44%	6.81%	20.63%
Two days after	88.99%	0.00%	0.00%

Note: The table reports the fraction of customers who receive the incentive at 2 days, 10 days or 30 days after their first purchase.

Table 7: Off-Policy Uplift Performance (Extension) Three Months

	2nd	3rd	Total Cost	ROAS
Constrained Dynamic BOWL	8.04%	4.83%	116.21	877%
Two Days After	7.40%	4.61%	129.27	741%

Note: The table reports the uplift of the optimal policy on their second and third purchases within three months after their first purchase. Total costs are measured in Japanese Yen and ROAS is the return on advertising spending.

6.2 Results of Extension

Next, we report the results of the offline policy evaluation for the dynamic model with the option of not sending the message. This extension allows us to decompose the effect of sending appreciation emails and the effect of providing incentives on top of the email.

Since the extension of Q-learning to a multi-action setup is straightforward, we compare the constrained BOWL with the baseline Two-days-after policy. The budget for the constrained BOWL is set at 130 JPY.[25]

Table 7 summarizes the OPE results. The first four columns report the fraction of consumers who make the second, third, fourth, and fifth purchases within the three months after their first purchase. The fifth column reports the per-consumer marketing cost including the cost of coupons and the cost of sending messages. The last column reports ROAS.

We find that the constrained dynamic BOWL achieves a greater chance of retention. In terms of the second purchase, the constrained dynamic BOWL leads to an uplift of 8.04% of the retention rate, which is higher than the baseline two-day policy. We also investigate longer-term effects and find that even for the fifth purchase, the constrained dynamic BOWL achieves higher retention.

For the financial implications, the dynamic BOWL leads to better performance. The total marketing cost of the dynamic BOWL is 116.21 JPY, while that of the two-day pol-

[25]The budget for the second experiment is bigger than the first one because, after the first experiment, the platform tried to encourage users to spend more points.

Table 8: Off-Policy Evaluation: Policies (Extension)

	2 days	10 days	30 days
	Email with incentives		
Constrained Dynamic BOWL	62.40%	17.34%	6.28%
Two Days After	90.54%	-	-
	Email without incentives		
Constrained Dynamic BOWL	20.29%	50.60%	49.36%
Two Days After	9.46%	100.00%	100.00%

Note: The table reports the fraction of customers who receive the incentive at 2 days, 10 days or 30 days after their first purchase.

icy is 129.27 JPY, 11.2% higher. The cost difference translates into a huge difference in ROAS; ROAS for the dynamic constrained BOWL is 136% greater than ROAS for the two-day policy. Thus, we show that our approach is a cost-effective strategy to improve customer retention not only in the short-run but also in the long run.

In Table 8, we describe how the optimal policy sends incentives at different timing. The constrained dynamic BOWL sends the incentives for 62.40% of consumers at 2 days after the first purchase, 17.34% at 10 days after, and 6.28%. Hence, the optimal strategy assigns incentives to later days. For emails without incentives, the constrained dynamic BOWL send 50% and 49.4% of users 10 days and 30 days after. Hence, it is not necessary to send emails to all users, which can save the cost of targeting through apps. Moreover, the optimal policy can save money because it does not send incentives for those who would have purchased even without incentives.[26]

6.3 Online Evaluation

Finally, the company tested some policies based on the baseline model (i.e., two actions) that we find perform well using the OPE methods. The company runs an A/B test by randomly allocating first-time buyers to each candidate policy to see its effects

[26]Note that the two-days policy does not send a coupon to all users. That is because some customers make their second purchase before receiving the coupon. The two-day policy does not send coupons to those customers.

Table 9: Online Evaluation (Uplift): Baseline Model

	2nd	3rd	4th	5th	Cost	ROAS
Dynamic BOWL	5.87%	2.23%	0.66%	0.28%	120.9	1845%
Dynamic Q-Learning	5.16%	1.84%	0.45%	0.21%	107.2	1428%
Baseline	5.97%	2.16%	0.76%	0.44%	122.8	1885%
Constrained Dynamic BOWL	4.55%	1.76%	0.43%	0.07%	94.4	2528%

Note: The table reports the results of the online evaluation.

Table 10: Online Evaluation (Uplift): Extended Model

	2nd	3rd	4th	5th	Cost	ROAS
Constrained Dynamic BOWL	3.41%	2.00%	0.83%	0.14%	–	1405%

Note: The table reports the results of the online evaluation of the extension model.

on retention. More precisely, the test considers four candidates: BOWL (without constraints), Q-learning (without constraints), static OWL, and BOWL with constraints as well as the control group with no offers.

Table 9 reports the retention rate and ROAS for each policy. Overall, the online test results show that all targeting policies perform better than the offline test results even though the control group (no offer) sees slightly lower retention rates during the online test period. Among the policies without any budget constraints, BOWL-based policies achieve higher retention rates and better ROAS than Q-learning. We find that static BOWL performs better than dynamic BOWL, which looks a bit strange, which we guess is due to the insufficient tuning of hyperparameters for dynamic BOWL. Our proposed dynamic-constrained BOWL can also achieve a high retention rate, and importantly, much better ROAS than any other strategies. Therefore, we confirm that our approach works for both offline and online settings.

Lastly, Table 10 reports the online test results of the extension model. For this online test, the company was willing to test only the constrained dynamic BOWL against the control group who received neither emails nor incentives. The online test was conducted from July 2022 to October 2022. The treatment effect on the second retention

is 3.41% and ROAS is 1405%, which are smaller than the offline evaluation results. We suspect this is because the Japanese economy went back to normal during this period and hence users spend more offline rather than online.

7 Conclusion

This paper proposes a method to infer the optimal dynamic targeting policy for customer retention management when there is a budget constraint. Dynamic policies are crucial for customer retention management as the states of consumers such as an intention for future purchases inherently evolve over time. Moreover, since most marketing campaigns have certain budget ceilings, it is practically important to consider a cost-efficient way to target consumers.

To do so, we extend the existing methods of estimating dynamic treatment regimes to account for inter-temporal budget constraints. In particular, we examine Q-learning and Backward Outcome Weighted Learning methods, which we incorporate into constrained optimization problems. We provide the algorithms to find the optimal DTR under budget constraints for both Q-learning and BOWL.

Our empirical application is a large online E-commerce platform in Japan. The company sends "thank you" messages to those who make their first purchase to urge second purchases, and we personalize it by adding coupons. The company runs a series of large-scale randomized experiments with more than 100,000 monthly new buyers. The experimental data allows us to estimate the optimal DTR in the offline setting before the company actually implements the dynamic personalized policies for the entire population.

The results show that the estimated DTRs are highly effective. With budget constraints, we can derive cost-effective optimal policies with almost the same level of customer retention relative to non-constrained policies. Hence, the return on advertising spending, the company's main KPI, can be as high as 800%, which is much higher

than typical marketing campaigns.

References

ASCARZA, E. (2018): "Retention Futility: Targeting High-Risk Customers Might be Ineffective," *J. Mark. Res.*, 55(1), 80–98.

ASCARZA, E., S. NESLIN, AND O. N. ET AL. (2018): "In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions," *Customer Needs and Solution*, 5, 65–81.

ASCARZA, E., M. ROSS, AND B. HARDIE (2021): "Why You Aren't Getting More from Your Marketing AI," *Harvard Business Review*.

ATHEY, S., R. CHETTY, G. W. IMBENS, AND H. KANG (2019): "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," *NBER Working Papers*.

ATHEY, S., AND S. WAGER (2021): "Policy learning with observational data," *Econometrica*, 89(1), 133–161.

BHATTACHARYA, D., AND P. DUPAS (2012): "Inferring welfare maximizing treatment assignment under budget constraints," *J. Econom.*, 167(1), 168–196.

FADER, P. S., AND B. G. HARDIE (2007): "How to project customer retention," *Journal of Interactive Marketing*, 21(1), 76–90.

FADER, P. S., AND B. G. S. HARDIE (2010): "Customer-Base Valuation in a Contractual Setting: The Perils of Ignoring Heterogeneity," *Marketing Science*, 29(1), 85–93.

HIRANO, K., G. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

HITSCH, G. J., AND S. MISRA (2018): "Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation," *Available at SSRN*.

INMAN, J. J., AND L. MCALISTER (1994): "Do coupon expiration dates affect consumer behavior?," *Journal of Marketing Research*.

JIANG, N., AND L. LI (2016): "Doubly Robust Off-policy Value Evaluation for Reinforcement Learning," in *Proceedings of The 33rd International Conference on Machine Learning*, ed. by M. F. Balcan, and K. Q. Weinberger, vol. 48 of *Proceedings of Machine Learning Research*, pp. 652–661, New York, New York, USA. PMLR.

KALLUS, N., AND A. ZHOU (2021): "Fairness, Welfare, and Equity in Personalized Pricing," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 296–314, New York, NY, USA. Association for Computing Machinery.

KAR, W., V. SWAMINATHAN, AND P. ALBUQUERQUE (2015): "Selection and Ordering of Linear Online Video Ads," in *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, pp. 203–210, New York, NY, USA. Association for Computing Machinery.

KIM, Y. (2021): "Customer Retention under Imperfect Information," *mimeo*.

KITAGAWA, T., AND A. TETENOV (2018): "Who should be treated? Empirical welfare maximization methods for treatment choice," *Econometrica*, 86(2), 591–616.

LAN, Q., Y. PAN, A. FYSHE, AND M. WHITE (2020): "Maxmin Q-learning: Controlling the Estimation Bias of Q-learning," *The International Conference on Learning Representations (ICLR)*.

LEMMENS, A., AND S. GUPTA (2020): "Managing Churn to Maximize Profits," *Marketing Science*, 39(5), 956–973.

LIU, X. (2021): "Dynamic Coupon Targeting Using Batch Deep Reinforcement Learning: An Application to LiveStream Shopping," *mimeo*.

MURPHY, S. A. (2003): "Optimal dynamic treatment regimes," *J. R. Stat. Soc. Series B Stat. Methodol.*, 65(2), 331–355.

MURPHY, S. A., K. G. LYNCH, D. OSLIN, J. R. MCKAY, AND T. TENHAVE (2007): "Developing adaptive treatment strategies in substance abuse research," *Drug Alcohol Depend.*, 88 Suppl 2, S24–30.

NESLIN, S. A., S. GUPTA, W. KAMAKURA, J. LU, AND C. H. MASON (2006): "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models," *Journal of Marketing Research*, 43(2), 204–211.

NESLIN, S. A., G. A. TAYLOR, K. D. GRANTHAM, AND K. R. MCNEIL (2013): "Overcoming the "recency trap" in customer relationship management," *Journal of the Academy of Marketing Science*, 41(3).

NIE, X., E. BRUNSKILL, AND S. WAGER (2021): "Learning When-to-Treat Policies," *Journal American Statistical Association*, 116(533), 392–409.

PRECUP, D., R. S. SUTTON, AND S. P. SINGH (2000): "Eligibility Traces for Off-Policy Policy Evaluation," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 759–766, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

SAKAGUCHI, S. (2019): "Estimating optimal dynamic treatment assignment rules under intertemp oral budget constraints," https://www.ifs.org.uk/uploads/papers.pdf, Accessed: 2021-5-6.

SIMESTER, D., A. TIMOSHENKO, AND S. I. ZOUMPOULIS (2020): "Efficiently Evaluating Targeting Policies: Improving on Champion vs. Challenger Experiments," *Management Science*, 66(8), 3412–3424.

SUN, L. (2021): "Empirical Welfare Maximization with Constraints," .

WAGER, S., AND S. ATHEY (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *J. Am. Stat. Assoc.*, 113(523), 1228–1242.

YANG, J., D. ECKLES, P. S. DHILLON, AND S. ARAL (2020): "Targeting for long-term outcomes," *ArXiv*.

YOGANARASIMHAN, H., E. BARZEGARY, AND A. PANI (2020): "Design and Evaluation of Personalized Free Trials," .

ZHANG, B., A. A. TSIATIS, E. B. LABER, AND M. DAVIDIAN (2013): "Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions," *Biometrika*, 100(3), 681–694.

ZHAO, Y., M. R. KOSOROK, AND D. ZENG (2009): "Reinforcement learning design for cancer clinical trials," *Stat. Med.*, 28(26), 3294–3315.

ZHAO, Y., D. ZENG, A. J. RUSH, AND M. R. KOSOROK (2012): "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *J. Am. Stat. Assoc.*, 107(449), 1106–1118.

ZHAO, Y.-Q., D. ZENG, E. B. LABER, AND M. R. KOSOROK (2015): "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *J. Am. Stat. Assoc.*, 110(510), 583–598.

A Mathematical Appendix

A.1 Single-Period Optimization

Consider the single-period constrained maximization problem:

$$\max_d E_d[Y] \text{ s.t. } E_d[C] \leq B, \tag{A.1}$$

where C is the cost variable and B is the budget per person.

Let $\beta(h) = E[Y|H = h, A = 1] - E[Y|H = h, A = 0]$ and $\gamma(h) = E[C|H = h, A = 1] - E[C|H = h, A = 0]$. Define DTR $d(\cdot;\lambda)$ by

$$d(h;\lambda) = \mathbf{1}[\beta(h) \geq \lambda\gamma(h)], \quad h \in \mathscr{H}.$$

Proposition 2. *Suppose that a deterministic solution to the maximization problem (A.1) exists. Suppose also that there exists $\lambda^* \geq 0$ such that $E_{d(\cdot;\lambda^*)}[C] = B$. Then $d(\cdot;\lambda^*)$ solves (A.1).*

Proof. Observe that for any DTR d,

$$E_d[Y] = E[E[Y|H, A = 0]] + E[d(H)\beta(H)], \;\; E_d[C] = E[E[C|H, A = 0]] + E[d(H)\gamma(H)].$$

The problem (A.1) is then equivalent to

$$\max_d E[d(H)\beta(H)] \text{ s.t. } E[d(H)\gamma(H)] \leq \bar{B},$$

where $\bar{B} = B - E[E[C|H, A = 0]]$. Let d^* be a deterministic solution to (A.1). Also, let

$$S_1 = \{h \in \mathscr{H} : d(h;\lambda^*) = 0, d^*(h) = 1\}, \;\; S_0 = \{h \in \mathscr{H} : d(h;\lambda^*) = 1, d^*(h) = 0\}.$$

Observe that

$$E[d^*(H)\gamma(H)] = E[d(H;\lambda^*)\gamma(H)] + E[\mathbf{1}[H \in S_1]\gamma(H)] - E[\mathbf{1}[H \in S_0]\gamma(H)].$$

Since $E[d^*(H)\gamma(H)] \leq \bar{B}$ and $E[d(H;\lambda^*)\gamma(H)] = \bar{B}$, it follows that $E[\mathbf{1}[H \in S_1]\gamma(H)] \leq E[\mathbf{1}[H \in S_0]\gamma(H)]$. We then obtain that

$$\begin{aligned} E[d^*(H)\beta(H)] &= E[d(H;\lambda^*)\beta(H)] + E[\mathbf{1}[H \in S_1]\beta(H)] - E[\mathbf{1}[H \in S_0]\beta(H)] \\ &\leq E[d(H;\lambda^*)\beta(H)] + \lambda^*(E[\mathbf{1}[H \in S_1]\gamma(H)] - E[\mathbf{1}[H \in S_0]\gamma(H)]) \\ &\leq E[d(H;\lambda^*)\beta(H)], \end{aligned}$$

where the second line holds since $d(H;\lambda^*) = \mathbf{1}[\beta(H) \geq \lambda^*\gamma(H)]$. Therefore, $d(\cdot;\lambda^*)$ is a solution to (A.1). ■

A.2 Multi-Period Optimization

Consider the T-period constrained maximization problem:

$$\max_{\mathbf{d}} E_{\mathbf{d}}[Y] \text{ s.t. } E_{\mathbf{d}}[C] \leq B, \tag{A.2}$$

where $\mathbf{d} = (d_1, ..., d_T)$, C is the cost variable and B is the budget per person.

We introduce some notation. Given a DTR $\mathbf{d}$, let $\underline{\mathbf{d}}_t = (d_1, ..., d_t)$ and $\bar{\mathbf{d}}_t = (d_t, ..., d_T)$. Also, let $Q_T(h_T, a_T) = E[Y|H_T = h_T, A_T = a_T]$, $Q_T^C(h_T, a_T) = E[C|H_T = h = T, A_T = a_T]$, $\beta_T(h_T) = Q_T(h_T, 1) - Q_T(h_T, 0)$, and $\gamma_T(h_T) = Q_T^C(h_T, 1) - Q_T^C(h_T, 0)$. In addition, for $t = 1, ..., T-1$, let $Q_t(h_t, a_t; \bar{\mathbf{d}}_{t+1}) = E_{\bar{\mathbf{d}}_{t+1}}[Y|H_t = h_t, A_t = a_t]$, $Q_t^C(h_t, a_t; \bar{\mathbf{d}}_{t+1}) = E_{\bar{\mathbf{d}}_{t+1}}[C|H_t = h_t, A_t = a_t]$, $\beta_t(h_t; \bar{\mathbf{d}}_{t+1}) = Q_t(h_t, 1; \bar{\mathbf{d}}_{t+1}) - Q_t(h_t, 0; \bar{\mathbf{d}}_{t+1})$, and $\gamma_t(h_t; \bar{\mathbf{d}}_{t+1}) = Q_t^C(h_t, 1; \bar{\mathbf{d}}_{t+1}) - Q_t^C(h_t, 0; \bar{\mathbf{d}}_{t+1})$. Lastly, for each $\boldsymbol{\lambda} \in \mathbb{R}_+^T$, let $\bar{\boldsymbol{\lambda}}_t = (\lambda_t, ..., \lambda_T)$ and define DTR $\mathbf{d}(\boldsymbol{\lambda}) = (d_t(\cdot; \bar{\boldsymbol{\lambda}}_t))_{t=1}^T$ recursively as

$$d_T(h_T; \lambda_T) = \mathbf{1}\big[\beta_T(h_T) \geq \lambda_T \gamma_T(h_T)\big], \quad h_T \in \mathscr{H}_T$$

and for $t = T-1, ..., 1$,

$$d_t(h_t; \bar{\boldsymbol{\lambda}}_t) = \mathbf{1}\left[\beta_t(h_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1})) \geq \lambda_t \gamma_t(h_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}))\right], \quad h_t \in \mathscr{H}_t.$$

Proposition 3. *Suppose that a deterministic solution to the maximization problem (A.2) exists, and let* $\mathbf{d}^*$ *denote a solution. Suppose also that there exists* $\boldsymbol{\lambda}^* \geq 0$ *such that* $E_{\mathbf{d}(\lambda^*)}[C] = B$ *and that* $E_{\underline{\mathbf{d}}_t^*, \bar{\mathbf{d}}_{t+1}(\bar{\lambda}_{t+1}^*)}[C] = B$ *for all* $t = 1, ..., T-1$. *Then* $\mathbf{d}(\boldsymbol{\lambda}^*)$ *solves (A.2).*

Proof. Given the optimal DTR $\mathbf{d}^*$, we show by induction that $\mathbf{d}(\boldsymbol{\lambda}^*)$ solves (A.2).

First, consider period T. Since $\mathbf{d}^*$ is optimal, d_T^* solves

$$\max_{d_T} E_{\underline{\mathbf{d}}_{T-1}^*, d_T}[Y] \;\; s.t. \;\; E_{\underline{\mathbf{d}}_{T-1}^*, d_T}[C] \leq B. \tag{A.3}$$

Since $E_{\underline{\mathbf{d}}_{T-1}^*, d_T(\cdot; \lambda_T^*)}[C] = B$, using the argument in the proof of Proposition 2 shows that $d_T(\cdot; \lambda_T^*)$ solves (A.3). Therefore, $(\underline{\mathbf{d}}_{T-1}^*, d_T(\cdot; \lambda_T^*))$ solves (A.2).

Now consider period $t \leq T-1$, and suppose that $(\underline{\mathbf{d}}_t^*, \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}^*))$ solves the problem (A.2). We show that $(\underline{\mathbf{d}}_{t-1}^*, \bar{\mathbf{d}}_t(\bar{\boldsymbol{\lambda}}_t^*))$ solves the problem (A.2), where we interpret $(\underline{\mathbf{d}}_{t-1}, \bar{\mathbf{d}}_t)$ as $\mathbf{d}$ when $t = 1$. Since $(\underline{\mathbf{d}}_t^*, \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}^*))$ is optimal, d_t^* solves

$$\max_{d_t} E_{\underline{\mathbf{d}}_{t-1}^*, d_t, \bar{\mathbf{d}}_{t+1}(\bar{\lambda}_{t+1}^*)}[Y] \;\; s.t. \;\; E_{\underline{\mathbf{d}}_{t-1}^*, d_t, \bar{\mathbf{d}}_{t+1}(\bar{\lambda}_{t+1}^*)}[C] \leq B. \tag{A.4}$$

Observe that for any d_t,

$$E_{\underline{\mathbf{d}}_{t-1}^*, d_t, \bar{\mathbf{d}}_{t+1}(\bar{\lambda}_{t+1}^*)}[Y] = E_{\underline{\mathbf{d}}_{t-1}^*}[Q_t(H_t, 0; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}^*)) + d_t(H_t)\beta_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}^*))],$$
$$E_{\underline{\mathbf{d}}_{t-1}^*, d_t, \bar{\mathbf{d}}_{t+1}(\bar{\lambda}_{t+1}^*)}[C] = E_{\underline{\mathbf{d}}_{t-1}^*}[Q_t^C(H_t, 0; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}^*)) + d_t(H_t)\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}^*))].$$

The problem (A.4) is then equivalent to

$$\max_{d_t} E_{\underline{\mathbf{d}}_{t-1}^*}[d_t(H_t)\beta_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}^*))] \;\; s.t. \;\; E_{\underline{\mathbf{d}}_{t-1}^*}[d_t(H_t)\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}^*))] \leq \bar{B},$$

where $\bar{B} = B - E_{\underline{\mathbf{d}}^*_{t-1}}[Q_t^C(H_t, 0; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))]$. Let

$$S_1 = \{h_t \in \mathscr{H}_t : d_t(h_t; \bar{\boldsymbol{\lambda}}^*_t) = 0, d_t^*(h_t) = 1\},$$
$$S_0 = \{h_t \in \mathscr{H}_t : d_t(h_t; \bar{\boldsymbol{\lambda}}^*_t) = 1, d_t^*(h_t) = 0\}.$$

Observe that

$$\begin{aligned}
&E_{\underline{\mathbf{d}}^*_{t-1}}[d_t^*(H_t)\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1})] \\
&= E_{\underline{\mathbf{d}}^*_{t-1}}[d_t(H_t; \bar{\boldsymbol{\lambda}}^*_t)\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))] + E_{\underline{\mathbf{d}}^*_{t-1}}[\mathbf{1}[H_t \in S_1]\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))] \\
&\quad - E_{\underline{\mathbf{d}}^*_{t-1}}[\mathbf{1}[H_t \in S_0]\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))].
\end{aligned}$$

Since $E_{\underline{\mathbf{d}}^*_{t-1}}[d_t^*(H_t)\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1})] \le \bar{B}$ and $E_{\underline{\mathbf{d}}^*_{t-1}}[d_t(H_t; \bar{\boldsymbol{\lambda}}^*_t)\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))] = \bar{B}$, it follows that $E_{\underline{\mathbf{d}}^*_{t-1}}[\mathbf{1}[H_t \in S_1]\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))] \le E_{\underline{\mathbf{d}}^*_{t-1}}[\mathbf{1}[H_t \in S_0]\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))]$. We then obtain that

$$\begin{aligned}
&E_{\underline{\mathbf{d}}^*_{t-1}}[d_t^*(H_t)\beta_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1})] \\
&= E_{\underline{\mathbf{d}}^*_{t-1}}[d_t(H_t; \bar{\boldsymbol{\lambda}}^*_t)\beta_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))] + E_{\underline{\mathbf{d}}^*_{t-1}}[\mathbf{1}[H_t \in S_1]\beta_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))] \\
&\quad - E_{\underline{\mathbf{d}}^*_{t-1}}[\mathbf{1}[H_t \in S_0]\beta_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))] \\
&\le E_{\underline{\mathbf{d}}^*_{t-1}}[d_t(H_t; \bar{\boldsymbol{\lambda}}^*_t)\beta_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))] + \lambda_t^*(E_{\underline{\mathbf{d}}^*_{t-1}}[\mathbf{1}[H \in S_1]\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))] \\
&\quad - E_{\underline{\mathbf{d}}^*_{t-1}}[\mathbf{1}[H \in S_0]\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))]) \\
&\le E_{\underline{\mathbf{d}}^*_{t-1}}[d_t(H_t; \bar{\boldsymbol{\lambda}}^*_t)\beta_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1}))],
\end{aligned}$$

where the second line holds since $d_t(H_t; \bar{\boldsymbol{\lambda}}^*_t) = \mathbf{1}\left[\beta_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}^*_{t+1})) \ge \lambda_t^*\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1})^*)\right]$. Therefore, $(\underline{\mathbf{d}}^*_{t-1}, \bar{\mathbf{d}}_t(\bar{\boldsymbol{\lambda}}^*_t))$ solves the problem (A.2).

By induction, $\mathbf{d}(\boldsymbol{\lambda}^*)$ solves (A.2). ■

A.3 Fisher consistency

Recall our optimization problem is given by

$$\max_{\mathbf{d}} E_{\mathbf{d}}[Y] \tag{A.5}$$
$$s.t.\ E_{\mathbf{d}}\left[\sum_{t=1}^{T}\mathbf{1}[A_t=m]\right]\leq p$$
$$d_t(h_t)\in\mathscr{A}_t(h_t),\ h_t\in\mathscr{H}_t,\ t=1,...,T.$$

Let

- $\underline{\mathbf{d}}_t=(d_1,...,d_t)$, $\bar{\mathbf{d}}_t=(d_t,...,d_T)$
- $Q_t(h_t,a;\bar{\mathbf{d}}_{t+1})=E_{\bar{\mathbf{d}}_{t+1}}[Y|H_t=h_t,A_t=a]$
- $Q_t^C(h_t,a;\bar{\mathbf{d}}_{t+1})=E_{\bar{\mathbf{d}}_{t+1}}[\sum_{j=t}^{T}\mathbf{1}[A_t=m]|H_t=h_t,A_t=a]$
- $\mathscr{D}_t=\{d_t:d_t(h_t)\in\mathscr{A}_t(h_t),\ h_t\in\mathscr{H}_t\}$

For each $\boldsymbol{\lambda}\in\mathbb{R}_+^T$, let $\mathbf{d}(\boldsymbol{\lambda})=(d_t(\cdot;\bar{\boldsymbol{\lambda}}_t))_{t=1}^T$ be the DTR defined recursively from $t=T$ to $t=1$ as

$$d_t(h_t;\bar{\boldsymbol{\lambda}}_t)\in_{a\in\mathscr{A}_t(h_t)}\left\{Q_t(h_t,a;\bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}))-\lambda_t Q_t^C(h_t,a;\bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}))\right\}.$$

The following assumption implies that the probability that a tie occurs in the above maximization is zero so that $\mathbf{d}(\boldsymbol{\lambda})$ is unique almost everywhere.

Assumption 2. *For all* $t=1,...,T$, $\Pr_{\underline{\mathbf{d}}_{t-1}}(Q_t(H_t,a;\bar{\mathbf{d}}_{t+1})-\lambda Q_t^C(H_t,a;\bar{\mathbf{d}}_{t+1})=Q_t(H_t,a';\bar{\mathbf{d}}_{t+1})-\lambda Q_t(H_t,a';\bar{\mathbf{d}}_{t+1})|\{a,a'\}\in\mathscr{A}_t(H_t))=0$ *for any* a,a' *with* $a\neq a'$, *any* $\lambda\geq 0$, *any* $\underline{\mathbf{d}}_{t-1}\in\prod_{j=1}^{t-1}\mathscr{D}_j$, *and any* $\bar{\mathbf{d}}_{t+1}\in\prod_{j=t+1}^{T}\mathscr{D}_j$.

Below we show

1. The DTR obtained from Algorithm 5 for any fixed $\boldsymbol{\lambda}$ is $\mathbf{d}(\boldsymbol{\lambda})$ defined above.
2. There exists $\boldsymbol{\lambda}$ such that $\mathbf{d}(\boldsymbol{\lambda})$ solves the problem (A.5). (only for the binary action case)

Thus, it is possible to obtain the optimal DTR by implementing Algorithm 5 with a large number of $\boldsymbol{\lambda}$'s.

Under Assumption 2, $\mathbf{d}(\boldsymbol{\lambda})$ is obtained from Algorithm 5 for any given $\boldsymbol{\lambda} \in \mathbb{R}_+^T$.

Now consider the binary action case with $\mathscr{A} = \{0,1\}$. In this case, $d_t(h_t; \bar{\boldsymbol{\lambda}}_t) \in \mathscr{A}_t(h_t)$ if $\mathscr{A}_t(h_t)$ is a singleton and $d_t(h_t; \bar{\boldsymbol{\lambda}}_t) = \mathbf{1}[\beta_t(h_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1})) - \lambda_t \gamma_t(h_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1})) \geq 0]$ if $\mathscr{A}_t(h_t) = \{0,1\}$, where $\beta_t(h_t; \bar{\mathbf{d}}_{t+1}) = Q_t(h_t, 1; \bar{\mathbf{d}}_{t+1}) - Q_t(h_t, 0; \bar{\mathbf{d}}_{t+1})$ and $\gamma_t(h_t; \bar{\mathbf{d}}_{t+1}) = Q_t^C(h_t, 1; \bar{\mathbf{d}}_{t+1}) - Q_t^C(h_t, 0; \bar{\mathbf{d}}_{t+1})$.

Assumption 3. *For all $t = 1, ..., T$ and $\bar{\mathbf{d}}_{t+1} \in \prod_{j=t+1}^T \mathscr{D}_j$, the following holds.*

1. *$\gamma_t(h_t; \bar{\mathbf{d}}_{t+1}) > 0$ for all $h_t \in \mathscr{H}_t$;*

2. *The distribution of $\beta_t(H_t; \bar{\mathbf{d}}_{t+1}) / \gamma_t(H_t; \bar{\mathbf{d}}_{t+1})$ conditional on $\mathscr{A}_t(H_t) = \{0,1\}$ is absolutely continuous with respect to the Lebesgue measure and has a strictly increasing cumulative distribution function on its support under any distribution of H_t generated by $\underline{\mathbf{d}}_{t-1} \in \prod_{j=1}^{t-1} \mathscr{D}_j$.*

Assumption 3 makes it easier to characterize the optimal DTR. When $\mathscr{A} = \{0,1\}$, Assumption 3 implies Assumption 2.

- Part 1 says that the expected number of times action 1 is chosen in the current and future periods (the Q-value for the cost) is larger if we take the action now than if we do not.

- Part 2 rules out the possibility that there is a mass of individuals with the same benefit-to-cost ratio.

$$d_t(h_t; \bar{\boldsymbol{\lambda}}_t) \in \begin{cases} \arg\max_{a \in \mathscr{A}_t(h_t)} Q_t(h_t, a; \bar{\mathbf{d}}_{t+1}) & \text{if } m \notin \mathscr{A}_t(h_t) \text{ or } |\mathscr{A}_t(h_t)| = 1 \\ \arg\max_{a \in \mathscr{A}_t(h_t) \setminus \{m\}} Q_t(h_t, a; \bar{\mathbf{d}}_{t+1}) & \text{if } m \in \mathscr{A}_t(h_t), |\mathscr{A}_t(h_t)| \geq 2, \text{ and } \frac{\beta_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}))}{\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}))} < \lambda_t \\ \{m\} & \text{if } m \in \mathscr{A}_t(h_t), |\mathscr{A}_t(h_t)| \geq 2, \text{ and } \frac{\beta_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}))}{\gamma_t(H_t; \bar{\mathbf{d}}_{t+1}(\bar{\boldsymbol{\lambda}}_{t+1}))} \geq \lambda_t \end{cases}$$

Let $\mathscr{A} = \{0,1\}$, and suppose that a (deterministic) solution to the maximization problem (A.5) exists. Under Assumption 3, there exists $\boldsymbol{\lambda} \in (\mathbb{R}_+ \cup \{\infty\})^T$ such that $\mathbf{d}(\boldsymbol{\lambda})$ solves (A.5).

A.4 Q-Learning in a general setup

Algorithm 6. (Constrained Q-Learning)

For each $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_T) \in \Lambda$ (the set of grid points on $\mathbb{R}_+^T$), repeat the following steps.

Step 1.

1. Conduct a regression of Y on H_T and A_T. Let $\hat{Q}_T$ be the estimated function.
2. Let $Q_T^C : \mathscr{H}_T \times \mathscr{A}_T \to \mathbb{R}$ be a function such that $Q_T^C(h_T, a) = 0$ for all $h_T \in \mathscr{H}_T$ if $a \neq m$ and $Q_T^C(h_T, m) = 1$ for all $h_T \in \mathscr{H}_T$.
3. Let $d_T(h_T; \lambda_T) \in \arg\max_{a \in \mathscr{A}_T(h_T)} \big(Q_T(h_T, a) - \lambda_T Q_T^C(h_T, a)\big)$.

Step 2. Do the following backward for $t = T-1, T-2, ..., 1$.

1. Conduct a regression of $\hat{Q}_{t+1}(H_{t+1}, d_{t+1}(H_{t+1}; \lambda_{t+1}, ..., \lambda_T))$ or $\frac{Y \prod_{j=t+1}^T \mathbf{1}[A_j = d_j(H_j; \lambda_j, ..., \lambda_T)]}{\prod_{j=t+1}^T d_j^0(A_j|H_j)}$ on H_t and A_t. Let $\hat{Q}_t$ be the estimated function.
2. Conduct a regression of $\hat{Q}_{t+1}^C(H_{t+1}, d_{t+1}(H_{t+1}; \lambda_{t+1}, ..., \lambda_T))$ or $\frac{(\sum_{j=t}^T \mathbf{1}[A_j = m]) \prod_{j=t+1}^T \mathbf{1}[A_j = d_j(H_j; \lambda_j, ..., \lambda_T)]}{\prod_{j=t+1}^T d_j^0(A_j|H_j)}$ on H_t and A_t. Let $\hat{Q}_t^C$ be the estimated function.
3. Let $d_t(h_t; \lambda_t, ..., \lambda_T) \in \arg\max_{a \in \mathscr{A}_t(h_t)} \big(Q_t(h_t, a) - \lambda_t Q_t^C(h_t, a)\big)$.

(In each of Steps 1 and 2, we can instead set the outcome to $Y - \lambda_t \sum_{j=t}^T \mathbf{1}[A_j = m]$ and omit estimating Q_t^C.)

Once we obtain the DTR $\mathbf{d}(\boldsymbol{\lambda}) = (d_1(\cdot;\lambda_1,...,\lambda_T),...,d_T(\cdot;\lambda_T))$ for each $\boldsymbol{\lambda} \in \Lambda$, choose $\boldsymbol{\lambda}^*$ such that

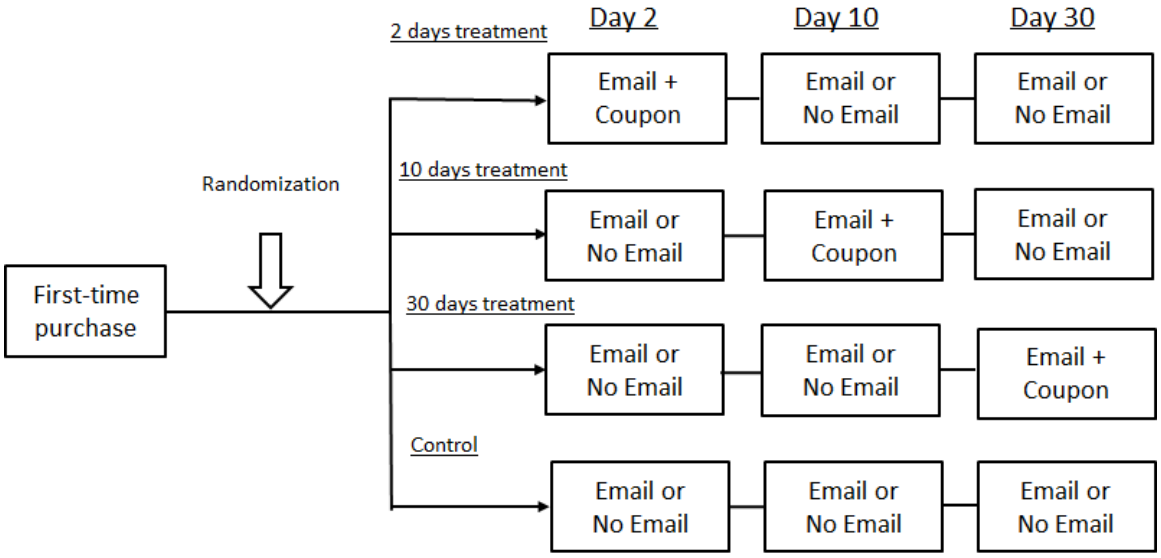
$$\boldsymbol{\lambda}^* \in \arg\max_{\boldsymbol{\lambda}\in\Lambda} E_{\mathbf{d}(\boldsymbol{\lambda})}[Y] \text{ s.t. } E_{\mathbf{d}(\boldsymbol{\lambda})}\left[\sum_{t=1}^{T} \mathbf{1}[A_t = m]\right] \leq p.$$

B Details of the Second Experiment

B.1 Experimental Design

In addition to the experiment we discuss in the main text, the company conducted another experiment to learn the optimal DTR of the model in Section 5.4, which includes not to send incentives. The experimental design is similar to the original experiment and looks like the figure below. Other features of the experimental design, including the timing and the number of treatment, the amount of financial incentive, and the user-level randomization, is the same as the first experiment.

Figure 5: Experimental Design: Extension



Note: The figure shows the experimental design of the second experiment where there are three actions, no email, email only and email and coupon.

Table 11 reports the summary statistics of the subset of the variables we use, com-

Table 11: Summary Statistics (Second Experiment)

Variable	**2day**		**10day**		**30 day**		**Control**	
	mean	**sd**	**mean**	**sd**	**mean**	**sd**	**mean**	**sd**
Order (2 month)	0.838	1.689	0.804	1.705	0.801	1.721	0.727	1.685
Quantity (2 month)	1.671	6.950	1.598	5.410	1.603	4.756	1.419	3.556
Amount (2 month)	5141	21415	4934	16435	4948	15571	4682	16336
Points used (2 month)	146.9	353.1	125.2	330.1	98.6	297.4	0.0	0.0
Female	0.632	0.482	0.631	0.483	0.628	0.483	0.632	0.482
Age	32.18	18.23	32.21	18.25	32.05	18.21	31.97	17.98
Quantity: first buy	1.013	0.179	1.013	0.172	1.014	0.221	1.015	0.252
Amount: first buy	4057	4187	4090	4298	4063	4082	4070	4090
# of sessions/day 1st buy	0.777	2.108	0.769	2.067	0.779	2.129	0.759	2.032
# of PV/day before delivery	12.81	28.34	12.75	29.30	12.96	29.38	12.60	27.27
# of favorites/day (2-10 day)	0.155	0.805	0.128	0.911	0.135	0.824	0.119	0.640
# of messages sent (10-30 day)	5.145	17.389	5.167	17.941	5.192	18.177	5.260	19.651
Observations	58739		58765		58506		14801	

Note: The first two columns report the mean and standard deviation of each variable for the treatment group. The third and fourth columns for the control group. The last column shows the t-values of the mean comparison test between two groups.

puted from the data of the second experiment.

Note that there are two types of incentives: coupon and appreciation email. To save the space, we do not report the summary statistics for all possible treatment conditions, but we report the summary statistics only for the case where coupons are sent. The control group includes the users who do not receive incentives nor emails.

Again, we find that the second purchase incidence is decreasing as users receive incentives later. For the amount of items purchased and sales, those who receive incentives 2 days after the first purchase have greater quantities and sales, while there are no significant difference between those who receive coupons 10 days after the first purchase and 30 days. For other state variables including both demographic and behavioral variables, we do not find any significant differences across conditions.

Table 12: Average Treatment Effect (Long Term)

	Retention	Sales	Order
Panel (A): Financial incentive			
2 days	0.066***	364.865**	0.160***
	(0.002)	(57.363)	(0.017)
10 days	0.055***	255.669	0.117***
	(0.002)	(42.484)	(0.013)
30 days	0.046***	204.958	0.095**
	(0.002)	(28.086)	(0.011)
Panel (B): Only appreciation email			
2 days	0.022***	260.748**	0.108***
	(0.004)	(106.041)	(0.033)
10 days	0.024***	110.355	0.076***
	(0.004)	(72.875)	(0.025)
30 days	0.003	45.196	0.011
	(0.003)	(48.195)	(0.014)

Note: The first column reports the treatment effect on whether a customer makes any purchases within 8 weeks since her first purchase. The second column reports the treatment effects on total sales and the third column reports the treatment effects on the number of items purchased. The data is from the second stage experiment.

B.2 Average Treatment Effects

Next, we estimate the average treatment effects with the second experiment data. Notice that there are two treatments, i.e., coupons and emails. To save the space, we report the results of the regression with the outcome measured for 8 weeks.

Table 12 reports the treatment effects measured for a longer period, 8 weeks. Again, we specify two types of treatment: one as email with financial incentive and the other as only appreciation emails.

The estimation results show that the appreciation email without coupons can increase retention. The two-days treatment effect of emails is 2.2% and the ten-days treatment effect is 2.4%. Sales also increase by 260 JPY by emails for the two-days treatment, but not for 10 days nor 30 days.

The results also reveal that the coupons in addition to the appreciation emails further increase retention, sales and the number of items purchased. Hence, by comparing the treatment effects in Panel (A) and ones in Panel (B), we can back out the pure effect of coupons. For the effect on retention, the coupons itself increase the retention by 4.4$ if users receive them after two days (0.066-0.022) and sales increase by 100 JPY.