Deviation-Based Learning[*]

Junpei Komiyama[†] Shunya Noda [‡]

September 22, 2021

Abstract

We propose *deviation-based learning*, a new approach to training recommender systems. In the beginning, the recommender and rational users have different pieces of knowledge, and the recommender needs to learn the users' knowledge to make better recommendations. The recommender learns users' knowledge by observing whether each user followed or deviated from her recommendations. We show that learning frequently stalls if the recommender always recommends a choice: users tend to follow the recommendation blindly, and their choices do not reflect their knowledge. Social welfare and the learning rate are improved drastically if the recommender abstains from recommending a choice when she predicts that multiple arms will produce a similar payoff.

[*] We are grateful to Vitor Farinha Luz for helpful comments. All remaining errors are our own.

[†] Leonard N. Stern School of Business, New York University, 44 West 4th Street, New York, NY 10012, United States. E-mail: junpei.komiyama@gmail.com.

[‡] Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Tokyo, 113-0033, Japan, and Vancouver School of Economics, University of British Columbia, 6000 Iona Dr, Vancouver, BC V6T 1L4 Canada. E-mail: shunya.noda@gmail.com. Noda has been supported by the Social Sciences and Humanities Research Council of Canada.

1 Introduction

In every day of our life, our choices rely on recommendations made by others based on their knowledge and experience. The prosperity of online platforms and artificial intelligence have enabled us to develop data-based recommendations, and many systems have been implemented in practice. Successful examples include e-commerce (Amazon), movies (Netflix), music (Spotify), restaurants (Yelp), sightseeing spots (TripAdvisor), hotels (Booking.com), classes (RateMyProfessors), hospitals (RateMD), and route directions by car navigation apps (Google Maps). These "recommender systems"[1] are helping us to make better decisions.

The advantages of the data-based recommender systems can be classified into two groups. First, the system can leverage experiences of the most knowledgeable experts. Once the system learns experts' behavior using data, the system can report what a user would do if he had experts' knowledge. Accordingly, with the help of the recommender system, all users can optimize their payoffs even when they have no experience with the problem they are facing. Second, the system can utilize information that an individual cannot access easily or quickly. For example, restaurant-reservation systems present the list of all available reservation slots at that moment, and online travel agencies provide the prices and available rooms of hotels. These conditions change over time; thus, it would be very difficult for an individual user to keep up to the minute with the latest conditions on their own. Accordingly, even experts benefit from information provided by recommender systems.

One of the largest challenges in developing a recommender system is to predict users' payoffs associated with specific alternatives. Real-world recommenders always face the problem of insufficient initial experimentation (known as the "cold start" problem). Utilization of feedback provided by users is necessary, but such data are often incomplete and insufficient. In particular, the system can rarely observe information about users' payoffs, which is crucial in many learning methods (e.g., reinforcement learning and the multi-armed bandit problem). As a proxy for payoffs, many recommender systems already implemented have adopted *rating-based learning*, which substitutes the ratings submitted by the users for the true payoffs of users. Nevertheless, a number of previous studies have reported that user-generated ratings often involve various types of biases and are not very informative signals of users' true payoffs (e.g., Salganik et al., 2006; Muchnik et al., 2013; Luca and Zervas, 2016).

In this paper, we propose *deviation-based learning*, a new approach for training recom-

[1] In a narrow sense, a "recommender system" is defined as an algorithm for predicting ratings users would enter. For example, Adomavicius and Tuzhilin (2005) state "In its most common formulation, the recommendation problem is reduced to the problem of estimating ratings for the items that have not been seen by a user" (p. 734). Our system is not a "recommender system" in this narrow sense because we do not utilize ratings. This paper adopts a broader definition of "recommender system" to denote any mechanism recommending arms (items or actions) to help users make better decisions

mender systems. In our model, a recommender (she) faces many rational users (he) sequentially. Neither users' payoffs nor ratings are available. Instead, we train a recommender system using data about past recommendations and users' final decisions. If the recommender has not yet been well-trained, expert users often deviate from her recommendations. On the flip side of the coin, upon observing expert users' deviations, the recommender system can recognize the fact that it had misestimated the underlying state. Conversely, if a user follows the recommendation even though the recommender is not completely confident in her prediction, then she can improve her confidence in the accuracy of her recommendations. Our deviation-based learning approach accumulates these observations to make better predictions of users' payoffs.

An illustrative example is app-based car navigation systems (e.g., Google Maps, or Waze). In recent years, such navigation apps have become extremely popular.[2] Navigation apps have an immense information advantage over individual drivers; using the app-generated data, an app can dynamically detect traffic jams and then recommend less congested routes. Accordingly, the app is useful even for local drivers who have memorized the (static) local road map and can figure out the shortest route without the recommender's help.

In the beginning, a navigation app does not have complete information about road characteristics. For example, the app may miss information about hazard conditions associated with specific roads (e.g., high-crime-rate areas, rock-fall hazard zones, accident blind spots). Such hazardous roads are often vacant because local drivers avoid them; thus, a naive recommender might consider such a route desirable and recommend it.[3] Drivers who are not familiar with this hazard information might then follow the recommendation and be exposed to danger.

The rating-based approach is not suitable for detecting hazards in the car navigation problem because (i) detailed ratings and reviews are often unavailable, and (ii) the app should not wait until it observes low payoffs because it means incidents or accidents indeed occur, and some users suffer from them. Moreover, this problem cannot be solved completely by inputting the hazard information manually because it is difficult to list all the relevant hazard conditions in advance.

Our deviation-based learning approach solves this dilemma by taking advantage of the knowledge of local drivers (the experts). When a hazardous route is recommended, a local driver ignores the recommendation and chooses a different route. Given that the app has

[2] According to Khoury (2019), Google Maps became the second app to reach the five-billion-download mark, after YouTube.

[3] For example, a blog post by Suro (2018) reports that navigation apps sometimes mistakenly recommend high-risk routes. To mitigate this problem, in Israel and Brazil, Waze provides the option of alerting about high-risk routes: https://support.google.com/waze/answer/7077122?hl=en (seen on July 22, 2021).

an information advantage about road congestion, such events would not occur unless the app misunderstands something about the static map with which the local driver is very familiar. Thus, upon observing a deviation, the app can update its knowledge about the static map. Conversely, if the app recommends a route that involves a potentially hazardous road but observes that the local driver followed the suggested route, then the app can conclude that the road is not that dangerous. In this manner, the app can obtain a better understanding of the static map and improve its recommendations. Note that the deviation-based learning approach can detect hazardous roads before any additional incident occurs because the recommender can observe that local drivers avoid such hazardous roads from the outset.

Formally, we analyze a stylized model in which each user has two arms (actions), as in seminal papers on information design theory (e.g., Kremer et al. 2014 and Che and Hörner 2017). A user's payoff from an arm is normalized to zero, and his payoff from another arm is given by $x\theta + z$. The *context* x specifies the user's problem (in the navigation problem, a context includes such elements as the origin, destination, and means of transportation). We assume each user is an expert; he knows the parameter θ and can correctly interpret his context x to predict the first term of his payoff, $x\theta$ (i.e., he knows the static map and find the shortest safe route). The recommender has additional information about the value of z (e.g., congestion), which is not observed by the user. We assume that local drivers are more knowledgeable than the recommender about the static map; the recommender does not at first know the parameter θ and only learns it over time. For each user, the recommender sends a recommendation (message) based on a precommitted information structure. Upon observing the recommendation, the user forms a belief about the unobservable payoff component z and selects either one of the two actions.

We demonstrate that the size of the message space is crucial for efficiency. The characterization of the tradeoff between the message space and the learning rate is of practical interest because the recommender often wants to minimize the message space to reduce users' cognitive costs. We show that by making the message space slightly larger than the action space, we obtain a large welfare gain. A large message space enables the recommender to send a signal that indicates the recommender is "on the fence" — i.e., the payoffs associated with two distinct actions are predicted to be similar. The availability of such messaging improves the learning rate exponentially without sacrificing the utilization of current knowledge.

We first prove that when the message space is binary (i.e., the message space and action space have the same size), then learning is very slow. In this case, the recommender recommends the better arm based on her latest information. Since the recommender has an information advantage, users often want to follow the recommendation blindly even if it is imperfect. Here, the recommender knows that no deviation will occur, and therefore, she learns nothing from the users' behaviors. Formally, we prove that the expected number of users required to improve the recommendations increases exponentially to the quality of recommendations. This effect slows down learning severely and causes a large welfare cost: while the per-round welfare loss in this situation is relatively small (since most users want to obey the recommendation blindly), the loss adds up to a large amount in the long run.

The learning rate is improved drastically when a ternary message space is allowed. We focus on a simple policy that recommends a particular arm only if the recommender is confident in her prediction. Otherwise, the recommender honestly discloses the fact that the two actions are predicted to produce similar payoffs according to the recommender's current information. When the recommender is confident about her prediction (which is almost always the case after the quality of her recommendation has become high), the user also confidently follows the recommendation, which maximizes the true payoff with high probability. Furthermore, when the recommender admits that she is on the fence, the user's choice is very useful in updating the recommender's belief. With the ternary message space, the total welfare loss is bounded by a constant (independent of the number of users).

2 Related Literature

Information Design This paper elucidates how the recommender learns about experts' knowledge through users' actions. This contrasts with previous studies on information design (e.g., Kamenica and Gentzkow, 2011; Bergemann and Morris, 2016a,b) and strategic experimentation (e.g., Kremer et al., 2014; Che and Hörner, 2017) that have explored how to incentivize agents to obey recommendations. Indeed, when either (i) the recommender (sender) has complete information about the underlying parameter (as in information design models) or (ii) payoffs (or signals about them) are observable (as in strategic experimentation models), a version of the "revelation principle" (originally introduced by Myerson, 1982) holds. In these cases, without loss of generality, we can focus on policies that always recommend actions from which no user has an incentive to deviate. In contrast, we demonstrate that when the recommender learns about underlying parameters by observing what users will do after knowing her recommendation, recommending just one arm is often inefficient.

Recommender System While the recommender system has mostly focused on predicting ratings, the vulnerability of rating-based learning has been widely recognized. Salganik et al. (2006) and Muchnik et al. (2013) show that prior ratings bias the evaluations of subsequent reviewers. Marlin and Zemel (2009) show that the rating often involves nonrandom missing data because users choose which item to rate. Mayzlin et al. (2014) and Luca and Zervas (2016) report that firms attempt to manipulate their online reputations strategically. While the literature has proposed several methods to address these issues (for example, Sinha et al. (2016) propose a way to correct bias by formulating recommendations as a control problem), the solutions proposed thus far are somewhat heuristic in the sense that their authors have not identified the fundamental source of the biases in rating systems using a model with rational agents.[4] In contrast, our deviation-based approach does not suffer from these biases because our approach does not rely on ratings.

Learning from Observed Behaviors In the literature of economic theory, inferring a rational agent's preferences given their observed choices is rather a classic question (*revealed preference theory*, pioneered by Samuelson, 1938). Furthermore, recent studies on machine learning and operations research, such as inverse reinforcement learning (Ng and Russell, 2000) and contextual inverse optimization (Ahuja and Orlin, 2001; Besbes et al., 2021) have also proposed learning methods to recover a decision-maker's objective function from his behavior.[5] These methods are useful for extracting experts' knowledge to make a better prediction about users' payoffs.

Our contribution to this literature can be summarized as follows. First, we elucidate the effect of the recommender's information advantage. In many real-world problems (e.g., navigation), the recommender is not informationally dominated by expert users; thus, decisions made by experts who are not informed of the recommender's information are typically suboptimal. This paper proposes a method to efficiently extract experts' knowledge and combine it with the recommender's own information. Second, we articulate the role of users' beliefs about the accuracy of the recommender's predictions. When the recommendation is accurate, users tend to follow recommendations blindly, and therefore, learning stalls under a naive policy (the binary recommendation policy). The extant studies have overlooked this effect because their learning models do not take into account interactions between the learner (the recommender) and the decision-maker (the users). Third, we demonstrate that the recommender can improve her learning rate significantly by intervening in the data generation process through information design. In our environment, learning under the ternary recommendation policy is exponentially faster than learning under the binary recommendation policy. The difference in social welfare achieved is also large.

In the marketing science literature, adaptive conjoint analysis has been proposed as a

[4] See the survey of the biases in rating systems by Chen et al. (2020).

[5] Classical learning methods, such as reinforcement learning (Sutton and Barto 2018, a standard textbook on this subject) and multi-armed bandit learning (Thompson, 1933; Lai and Robbins, 1985), assume that the learner can directly observe realized payoffs.

method to pose questions to estimate users' preference parameters in an adaptive manner. Several studies, such as Toubia et al. (2007) and Sauré and Vielma (2019), have considered adaptive *choice-based* conjoint analysis, which regards choice sets as questions and actual choices as answers for them. This strand of the literature has also developed efficient methods for intervening in the data generation process to extract users' knowledge. However, in the recommender problem, the recommender is not allowed to select users' choice sets to elicit their preferences. Instead, the recommender needs to design information.

3 Model

We consider a sequential game that involves a long-lived *recommender* and T short-lived *user*s. At the beginning of the game, the state of the world $\theta \sim \mathrm{Unif}[-1, 1]$ is drawn. We assume that all the users are experts and more knowledgeable than the recommender is about the state θ initially.[6] Formally, we assume that while users know the realization of θ, the recommender knows only the distribution of θ. Accordingly, the recommender learns about θ through the data obtained in the game.

Users arrive sequentially. At the beginning of round $t \in [T] \coloneqq \{1, \dots, T\}$, user t arrives with the shared context $x_t \sim \mathcal{N}$, where $\mathcal{N}$ is the standard (i.e., with a zero mean and unit variance) normal distribution. The context x_t is public information and observed by both user t and the recommender. The recommender also observes her private information $z_t \sim \mathcal{N}$, whose realization is not disclosed to user t. Each user has binary actions: arm -1 and arm 1.[7] Without loss of generality, the user's payoff from choosing arm -1 is normalized to zero: $r_t(-1) = 0$. The payoff from choosing arm 1 is given by

$$r_t(1) = x_t\theta + z_t.$$

We refer to $x_t\theta$ as the *static payoff* and to z_t as the *dynamic payoff*. These names come from the navigation problem presented as an illustrative example, in which users are assumed to be familiar with the static road map but do not observe dynamic congestion information until they actually select the route.

In round t, the recommender first selects a *recommendation* $a_t \in A$, where A is the *message space*. For example, if the recommender just wants to recommends an arm, then

[6] As long as the recommender can identify the set of expert users, she can exclude nonexpert users from the model. In the navigation app example, it should not be difficult for the app to identify the set of local residents who drive cars frequently. Once the recommender trains the system using the data of the experts' decisions, then she can use it to make recommendations to nonexpert users.

[7] Alternatively, we can assume that each user has many actions but all but two are obviously undesirable in each round. We assign ±1 to index the arms for the sake of mathematical clarity.

the message space is equal to the action space: $A = \{-1, 1\}$. Observing the recommendation a_t, user t chooses an action $b_t \in B = \{-1, 1\}$. User t receives a payoff of $r_t(b_t)$ and leaves the market. The recommender cannot observe users' payoffs.

As in the literature of information design, we assume that the recommender commits to an information structure that mechanically decides which message to submit. When the recommender makes her round-t recommendation, she can observe the sequences of all past recommendations, $(a_s)_{s=1}^{t-1}$, and all past actions that users took, $(b_s)_{s=1}^{t-1}$. Using these pieces of information, the recommender computes her belief about parameter θ using Bayes' rule. The recommender's belief at the beginning of round 1 is the same as the prior belief: $\mathrm{Unif}[-1, 1]$. Due to the property of uniform distributions, the posterior distribution of θ always belongs to the class of uniform distributions. The posterior distribution at the beginning of round t is specified by $\mathrm{Unif}[l_t, u_t]$, where l_t and u_t are the lower and upper bounds, respectively, of the confidence region in the beginning of round t. Note that the *confidence region* $[l_t, u_t]$ shrinks over time:

$$-1 =: l_1 \leq l_2 \leq \cdots \leq l_{T-1} \leq l_T \leq \theta \leq u_T \leq u_{T-1} \leq \cdots \leq u_2 \leq u_1 := 1,$$

and thus the *width of the confidence region* $w_t := u_t - l_t$ is monotonically decreasing. We assume users are informed of the recommender's current confidence region $[l_t, u_t]$. From the perspective of the recommender in round t, the predicted payoff from arm 1 is

$$\hat{r}_t(1) := \mathbb{E}_{\hat{\theta}_t \sim \mathrm{Unif}[l_t, u_t]}[x_t \hat{\theta}_t] + z_t = x_t m_t + z_t,$$

where $m_t := (l_t + u_t)/2 = \mathbb{E}_{\hat{\theta}_t \sim \mathrm{Unif}[l_t, u_t]}[\hat{\theta}_t]$.

Upon observing recommendation a_t, user t updates his belief about the dynamic payoff z_t. User t computes the conditional expected payoff from choosing arm 1, $\mathbb{E}[r_t(1)|m_t, a_t]$. Then, user t selects an arm $b_t \in B := \{-1, 1\}$ that provides a larger payoff in expectation: $b_t = 1$ if $\mathbb{E}[r_t(1)|m_t, a_t] > 0$ and $b_t = -1$ otherwise. Note that this is the strategy users would take in a perfect Bayesian equilibrium of this sequential game.

After observing user t's choice b_t, the recommender updates the posterior distribution about θ, characterized by (l_t, u_t), according to Bayes' rule. The choice b_t and the update rule of (l_t, u_t) will be explained in detail in Section 4. The (utilitarian) *social welfare* is defined as the sum of all users' payoffs: $\sum_{t=1}^{T} r_t(b_t)$. We quantify the welfare loss compared to the first-best scenario using *regret*. *Regret* is defined as follows:

$$\mathrm{reg}(t) := r_t(b_t^*) - r_t(b_t);$$

$$\mathrm{Reg}(T) \coloneqq \sum_{t=1}^{T} \mathrm{reg}(t),$$

where $b_t^* \coloneqq \max_{b\in\{-1,1\}} r_t(b)$ is the superior arm with respect to true payoffs. The value $\mathrm{reg}(t)$ indicates the loss of welfare caused by the suboptimal decision-making in round t; thus, the maximization of the social welfare is equivalent to the minimization of the regret. If the recommender already knows (or has learned accurately) the state θ, then the recommender would always inform user t of the superior arm, and the user would always obey the recommendation. Therefore, $b_t = b_t^*$ and $\mathrm{reg}(t) = 0$ would be achieved. Conversely, if the recommender's belief about the state θ is not accurate, then users cannot always select the superior arm. Therefore, the regret also measures the progress of the recommender's learning of θ.

In this paper, we characterize the relationship between the size of the message space $|A|$ and the order of regret $\mathrm{Reg}(T)$. It is easy to observe that the regret achievable is closely related to the size of the message space. When the message space is a singleton (i.e., $|A| = 1$), the recommender can convey no information about the dynamic payoff component z_t. As a result, users suffer from constant welfare loss for each round; therefore, the regret grows linearly in T: $\mathrm{Reg}(T) = \Theta(T)$. In contrast, if the message space is a continuum (i.e., $A = \mathbb{R}$), the recommender can inform each user t of the "raw data" about the dynamic payoff z_t, i.e., she can send $a_t = z_t$ as a message. In this case, users can recover true payoffs $r_t(1)$ and select the superior arms for every round. There is no need for the recommender to learn, and the regret of exactly zero is achieved: $\mathrm{Reg}(T) = 0$ for all T. Nevertheless, an infinite message space incurs a large cognitive cost, and is therefore inconvenient. Indeed, it is infeasible for real-world recommender systems to disclose all the information about current congestion. In the following, we evaluate the regret incurred with small finite message spaces, namely, the case of binary and ternary message spaces ($|A| = 2, 3$).

4 Binary Recommendations

4.1 Policy and Belief Updates

First, we consider the case of the binary message space, i.e., $A = \{-1, 1\} = B$. We consider a natural recommendation policy that simply discloses an arm predicted to be superior; it

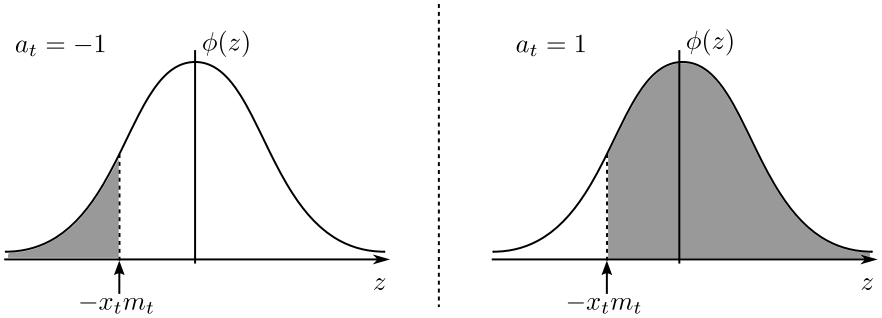

Figure 1: Region of possible z_t for $a_t = -1$ (left) and $a_t = 1$ (right) when $x_t m_t > 0$. The quantity Z_t is the expected value of the gray region.

recommends arm 1 if and only if $\hat{r}_t(1) = x_t m_t + z_t > 0 = \hat{r}_t(0)$.[8] That is,

$$a_t = \begin{cases} 1 & \text{if } x_t m_t + z_t > 0; \\ -1 & \text{otherwise.} \end{cases}$$

Since user t observes the quality of the recommender $[l_t, u_t]$, he also knows the value of $m_t \coloneqq (l_t + u_t)/2$. In addition, he observes the recommendation a_t before choosing an arm. User t's conditional expected payoff from choosing arm 1 is given by

$$\mathbb{E}[r_t(1)|m_t, a_t] = x_t\theta + Z_t,$$

where $Z_t \coloneqq \mathbb{E}[z_t|m_t, a_t]$.

We evaluate Z_t to identify the user's equilibrium behavior. The prior distribution of z_t is the standard normal distribution, $\mathcal{N}$. In addition, $a_t = 1$ implies $z_t > -x_t m_t$, whereas $a_t = -1$ implies $z_t < -x_t m_t$. Accordingly, the posterior distribution of z_t is always a truncated standard normal distribution. Let $\mathcal{N}^{\text{tr}}(\alpha, \beta)$ be the truncated standard normal distribution with support (α, β). Then, the posterior distribution of z_t after $a_t = 1$ and $a_t = -1$ are $\mathcal{N}^{\text{tr}}(-x_t m_t, \infty)$ and $\mathcal{N}^{\text{tr}}(-\infty, -x_t m_t)$, respectively. These distributions are illustrated as Figure 1. To summarize, we have

$$Z_t \coloneqq \mathbb{E}[z_t|m_t, a_t] = \begin{cases} \mathbb{E}_{z\sim\mathcal{N}^{\text{tr}}(-\infty, -x_t m_t)}[z] & \text{if } a_t = -1; \\ \mathbb{E}_{z\sim\mathcal{N}^{\text{tr}}(-x_t m_t, \infty)}[z] & \text{if } a_t = 1. \end{cases}$$

The arm that user t will choose is as follows:

$$b_t = \begin{cases} 1 & \text{if } x_t\theta + Z_t > 0; \\ -1 & \text{otherwise.} \end{cases} \tag{1}$$

[8] We ignore equalities of continuous variables that are of measure zero, such as $\hat{r}_t(1) = 0$.

Upon observing the user's decision b_t, the recommender updates her confidence region, $[l_t, u_t]$. When the user chooses $b_t = 1$, the recommender can recognize that $x_t\theta + Z_t > 0$. If $x_t > 0$, this is equivalent to $\theta > -Z_t/x_t$; and if $x_t < 0$, this is equivalent to $\theta < -Z_t/x_t$. Using this information, the recommender may be able to shrink the support of the posterior distribution about θ. We can analyze the case of $b_t = -1$ in a similar manner. The belief update rule is as follows:

$$l_{t+1} = \begin{cases} l_t & \text{if } b_t \cdot \mathrm{sgn}(x_t) < 0; \\ \max\{l_t, -Z_t/x_t\} & \text{if } b_t \cdot \mathrm{sgn}(x_t) > 0, \end{cases} \tag{2}$$

$$u_{t+1} = \begin{cases} \min\{u_t, -Z_t/x_t\} & \text{if } b_t \cdot \mathrm{sgn}(x_t) < 0; \\ u_t & \text{if } b_t \cdot \mathrm{sgn}(x_t) > 0, \end{cases} \tag{3}$$

where sgn is the following signum function:[9]

$$\mathrm{sgn}(x) := \begin{cases} 1 & \text{if } x > 0; \\ -1 & \text{if } x < 0. \end{cases}$$

4.2 Informativeness of Deviations

First, we present a fundamental theorem that constitutes the basis for the concept of deviation-based learning.

Theorem 1 (Informativeness)**.** If $a_t \neq b_t$, then

$$w_{t+1} < \frac{1}{2} w_t. \tag{4}$$

Conversely, if $a_t = b_t$, then

$$w_{t+1} > \frac{1}{2} w_t. \tag{5}$$

All the formal proofs are relegated to the appendix.

Theorem 1 elucidates the informativeness of deviations. When a user deviates from the recommendation (i.e., when $a_t \neq b_t$), the width of the recommender's confidence region will be at least halved. Since the recommender has an information advantage about z_t, a deviation occurs only if the recommender significantly misestimates the static payoff component. Accordingly, upon observing a deviation, the recommender can update her belief about θ by a large amount. In contrast, when a user obeys the recommendation (i.e., when $a_t = b_t$), the decrease of w_t is bounded. Note that, when users are obedient, it is frequently the case that

[9] $x_t = 0$ occurs with a probability of zero, and so we ignore such a realization.

no update occurs and so $w_{t+1} = w_t$. This is the case if the recommender's error $|x_t(\theta - m_t)|$ is small, and therefore, the user follows the recommendation blindly, given any $\theta \in [l_t, u_t]$.

4.3 Failure of Binary Recommendations

We present our first main theorem, which evaluates the order of total regret under the binary recommendation policy.

Theorem 2 (Regret Bound of Binary Recommendation)**.** There exists a $\tilde{\Theta}(1)$ (polylogarithmic) function[10] $f(T)$ of T such that

$$\mathbb{E}[\mathrm{Reg}(T)] \geq T/f(T).$$

Theorem 2 shows that the total regret is $\tilde{\Omega}(T)$, which implies that users suffer from a large per-round regret even in the long run.

While each user precisely knows his static payoff $x_t\theta$, he has access to the dynamic payoff z_t only via recommendation. To help the user make the best decision, the recommender must identify which arm is better as a whole. The recommender must therefore learn about the state θ in order to figure out the value of $r_t(1) = x_t\theta + z_t$ through the users' feedback b_t. The more the recommender learns about θ, the less informative is the users' feedback: rational users hardly deviate from (moderately) accurate recommendations because the recommender's information advantage (in terms of information about the dynamic payoff term) tends to dominate the estimation error. Consequently, when recommendations are accurate, deviations are rarely observed, and the recommender has few opportunities to improve her estimator $\hat{\theta}$.

In the following, we provide two lemmas that characterize the problem and then discuss how we derive Theorem 2 from these lemmas.

Lemma 3 (Lower Bound on Regret per Round)**.** The following inequality holds:

$$\mathbb{E}[\mathrm{reg}(t)] \geq C_{\mathrm{reg}}|\theta - m_t|^2,$$

where $C_{\mathrm{reg}} > 0$ is a universal constant.[11]

Since the recommender does not know θ, she substitutes m_t for θ to determine her recommendation. The probability that the recommender fails to recommend the superior

[10] $\tilde{O}, \tilde{\Omega}$, and $\tilde{\Theta}$ are Landau notations that ignore polylogarithmic factors. We often treat these factors as if they were constant because polylogarithmic factors grow very slowly ($o(N^\epsilon)$ for any exponent $\epsilon > 0$).

[11] A universal constant is a value that does not depend on any model parameters.

arm is proportional to $|\theta - m_t|$, and the loss from such an event is also proportional to $|\theta - m_t|$. Accordingly, the per-round expected regret is at the rate of $O(|\theta - m_t|^2)$.

Lemma 4 (Upper Bound on Probability of Update)**.** There exists a universal constant $C_{\text{update}} > 0$ such that, for all $w_t \leq C_{\text{update}}$,

$$\mathbb{P}[(a_{t+1}, b_{t+1}) \neq (a_t, b_t)] \leq \exp\left(-\frac{C_{\text{update}}}{w_t}\right).$$

User t compares two factors for making his decision: (i) the recommender's estimation error of the static payoff term $|x_t(\theta - m_t)|$ and (ii) the recommender's information advantage about the dynamic payoff term z_t. When the former factor is small, the user blindly obeys the recommendation, and the user's decision does not provide additional information. Since $w_t > |\theta - m_t|$, the former factor is bounded by $x_t w_t$. For a user's decision to be informative, x_t must be $\Omega(1/w_t)$ (so that $|x_t(\theta - m_t)|$ becomes larger than a threshold value). Since x_t follows a normal distribution, the probability that such a context arrives decreases exponentially in $1/w_t$.[12]

Lemma 4 states that the recommender's learning stalls when w_t is moderately small. In particular, if $w_t = 2C_{\text{update}}/(\log T) = \Theta(1/(\log T))$, then the probability of her belief update is $1/T^2$. This implies that no update occurs in the next T rounds with a probability at least $1 - 1/T$.

We use these lemmas to obtain the total regret bound presented in Theorem 2. First, Lemma 4 implies the update of θ is likely to stall when it reaches $w_t = |\theta - m_t| = \Theta(1/(\log T))$. Given $|\theta - m_t| = \Theta(1/(\log T))$, Lemma 3 implies the per-round (expected) regret is $\Theta(1/(\log T)^2)$. Consequently, the order of total regret is $\Omega(T/(\log T)^2) = \tilde{\Omega}(T)$, implying that users suffer from large per-round regrets even in the long run. This is how we obtained the regret bound presented as Theorem 2.

5 Ternary Recommendations

5.1 Policy

Section 4 considered a binary message space, $A = \{-1, 1\} = B$. We have shown that the recommender fails to learn θ with the binary message space. In this section, we consider an alternative recommender system with a ternary message space, $A = \{-1, 0, 1\} = B \cup \{0\}$.

[12] A similar result holds whenever x_t follows a sub-Gaussian distribution, where the probability of observing x_t decays at an exponential rate with respect to $|x_t|$. Conversely, when the distribution of x_t is heavy-tailed, the conclusion of Lemma 4 may not hold.

This ternary message space allows the recommender to inform users that she is "on the fence." When the recommender is confident in her recommendation, she sends either $a_t = -1$ or $a_t = 1$. If the recommender predicts that the user should be approximately indifferent between two arms, then she sends $a_t = 0$ instead.

Specifically, we focus on the following recommendation policy. We introduce a sequence of parameters $(\epsilon_t)_{t=1}^T$, where $\epsilon_t > 0$ for all $t \in [T]$, that determines whether the recommender is confident in her prediction. If $\hat{r}_t(1) > \epsilon_t$, then the recommender is confident about the superiority of arm 1, and therefore recommends arm 1: $a_t = 1$. Conversely, if $\hat{r}_t(1) < -\epsilon_t$, then the recommender is confident about the superiority of arm -1, and therefore recommends arm -1: $a_t = -1$. In the third case, i.e., $-\epsilon_t < \hat{r}_t(1) < \epsilon_t$, the recommender honestly states that she is on the fence; she sends the message $a_t = 0$, implying that she predicts that the payoffs associated with arms 1 and -1 will be similar. This recommendation policy is summarized as follows:

$$a_t = \begin{cases} 1 & \text{if } x_t m_t + z_t > \epsilon_t; \\ 0 & \text{if } \epsilon_t > x_t m_t + z_t > -\epsilon_t; \\ -1 & \text{if } x_t m_t + z_t < -\epsilon_t. \end{cases}$$

User t's posterior belief about z_t is given by (i) $z_t \sim \mathcal{N}^{\text{tr}}(-\infty, -x_t m_t - \epsilon_t)$ given $a_t = -1$; (ii) $z_t \sim \mathcal{N}^{\text{tr}}(-x_t m_t - \epsilon_t, -x_t m_t + \epsilon_t)$ given $a_t = 0$; and (iii) $z_t \sim \mathcal{N}^{\text{tr}}(-x_t m_t + \epsilon_t, \infty)$ given $a_t = 1$. Accordingly, the conditional expectation of z_t with respect to the posterior distribution is as follows.

$$Z_t := \mathbb{E}[z_t | m_t, a_t] = \begin{cases} \mathbb{E}_{z \sim \mathcal{N}^{\text{tr}}(-\infty, -x_t m_t - \epsilon_t)}[z] & \text{if } a_t = -1; \\ \mathbb{E}_{z \sim \mathcal{N}^{\text{tr}}(-x_t m_t - \epsilon_t, -x_t m_t + \epsilon_t)}[z] & \text{if } a_t = 0; \\ \mathbb{E}_{z \sim \mathcal{N}^{\text{tr}}(-x_t m_t + \epsilon_t, \infty)}[z] & \text{if } a_t = 1, \end{cases}$$

which is illustrated in Figure 2. Given the new specifications of a_t and Z_t, the user's decision rule for choosing b_t (given in Eq. (1)) and the belief update rule for deciding (l_{t+1}, u_{t+1}) (given in Eq. (2) and (3)) are unchanged.

5.2 Success of Ternary Recommendations

The following theorem characterizes the total regret achieved by the ternary recommendation policy.

Theorem 5 (Regret Bound of Ternary Recommendation)**.** Let $\epsilon_t = (3/4) w_t$. Then, the regret is bounded as:

$$\mathbb{E}[\text{Reg}(T)] \leq C_{\text{ter}},$$

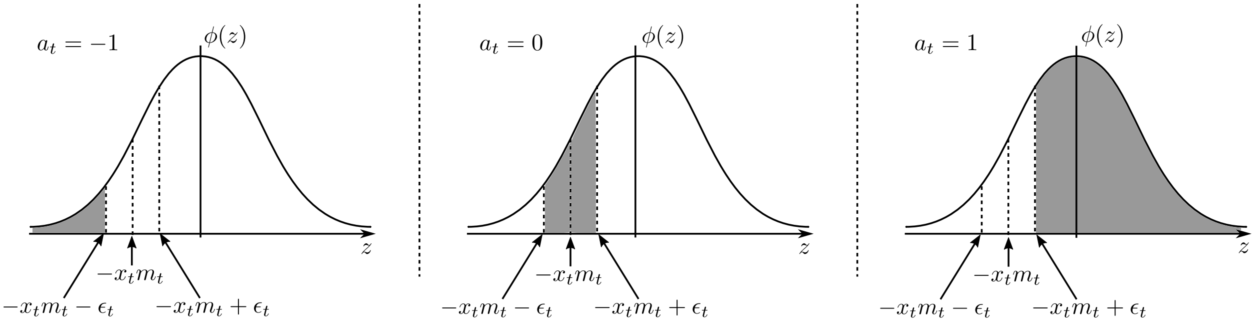

Figure 2: Region of possible z_t for $a_t = -1$ (left), $a_t = 0$ (middle), and $a_t = 1$ (right) when $x_t m_t > 0$. The quantity Z_t is the expected value of the gray region.

where $C_{\text{ter}} > 0$ is a universal constant.

This result contrasts with Theorem 2, which shows that the binary recommendation policy suffers from $\tilde{\Omega}(T)$ regret. Theorem 5 implies that a slight expansion of the message space drastically improves the efficiency of the recommendation policy.

When the message space is binary, the recommender can inform user t only of whether or not $x_t m_t + z_t > 0$; thus, user t cannot figure out how large $|x_t m_t + z_t|$ is. However, the value of $|x_t m_t + z_t|$ is indeed crucial to know for whether the user should deviate from the recommendation. If the user were to know that $|x_t m_t + z_t|$ is close to zero, then he would have a stronger motivation to defy the recommendation because the estimation error $|x_t(\theta - m_t)|$ matters. Accordingly, by informing users of the magnitude of $|x_t m_t + z_t|$, the recommender could "encourage" users to deviate from her recommendation so as to exploit the users' knowledge more efficiently.

Our ternary recommendation policy effectively achieves the scenario described above. When $|x_t m_t + z_t|$ is smaller than the threshold value ϵ_t, the recommender informs the agent this fact by submitting message $a_t = 0$. The user's choice after observing $a_t = 0$ is useful for making a better estimate of θ. The more accurate estimate on θ the recommender makes, the smaller the value of ϵ_t she chooses; this in turn leads to her sending $a_t = 0$ less often. In the end, she is able to recommend a truly superior arm for every user.

The following lemmas characterize the key properties that we use in Theorem 5. The first lemma, Lemma 6, computes the probability that $a_t = 0$ is sent.

Lemma 6 (Probability of $a_t = 0$)**.** The following equality holds:

$$\mathbb{P}[a_t = 0] = C_{\text{OtF}} \epsilon_t,$$

where $C_{\text{OtF}} > 0$ is a universal constant.

Lemma 6 states that the probability that $a_t = 0$ is recommended is linear in ϵ_t. This result immediately follows from the fact that (i) z_t follows a standard normal distribution, and (ii) $a_t = 0$ is sent when $z_t \in (-x_t m_t - \epsilon_t, -x_t m_t + \epsilon_t)$.

The second lemma, Lemma 7, bounds the per-round regret, $\mathrm{reg}(t)$, using a quadratic function of the policy parameter, ϵ_t, and the width of the confidence region, w_t.

Lemma 7 (Upper Bound on Regret per Round)**.** The following inequality holds:

$$\mathbb{E}[\mathrm{reg}(t)] \le C_{\mathrm{regt}}(\max(\epsilon_t, w_t))^2,$$

where $C_{\mathrm{regt}} > 0$ is a universal constant.

When an arm is recommended (i.e., when $a_t \neq 0$), then we can apply the analysis for binary recommendations to derive the per-round expected regret of $O(w_t^2)$. The message $a_t = 0$ is sent with probability $\Theta(\epsilon_t)$ (by Lemma 6). Since $a_t = 0$ is sent only when the utility is (approximately) indifferent between two arms, the per-round regret is bounded by $\epsilon_t + w_t$ in this case. Hence, the per-round expected regret for this case is $O(\max(\epsilon_t, w_t)^2)$.

Finally, the third lemma, Lemma 8, guarantees that when ϵ_t is selected appropriately, then for every time that $a_t = 0$ is selected, the confidence region shrinks geometrically.

Lemma 8 (Geometric Update)**.** Let $\epsilon_t = (3/4)w_t$. Then, the following inequality holds:

$$\mathbb{P}[w_{t+1} \le C_{w,1} w_t \,|\, a_t = 0] \ge C_{w,2},$$

where

$$(C_{w,1}, C_{w,2}) = \left(\frac{7}{8}, \frac{3}{4}\frac{e^{-2}}{\sqrt{2\pi}}\right).$$

The intuition for the geometric gain after $a_t = 0$ is as follows. When $\epsilon_t \approx 0$, then, the user, upon observing $a_t = 0$ (i.e., $x_t m_t + z_t$ is very close to zero), can accurately figure out the realization of the dynamic payoff term: $z_t \approx -x_t m_t$. Given this, the user chooses $b_t = 1$ if $x_t\theta + z_t \approx x_t(\theta - m_t) > 0$ and $b_t = -1$ otherwise; in other words, by observing b_t, the recommender can identify whether or not $\theta > m_t$. Since m_t is the median of the confidence region $[l_t, u_t]$, this observation halves the width of the confidence region. Accordingly, when the recommender chooses a sufficiently small ϵ_t, a geometric update is achieved.

While a smaller ϵ_t results in a larger update after $a_t = 0$ is sent, we cannot set $\epsilon_t = 0$ because in that case the probability of sending $a_t = 0$ becomes zero. The policy parameter ϵ_t must be chosen to balance this trade-off. Lemma 8 shows that $\epsilon_t = (3/4)w_t$ is an appropriate choice in the sense that it achieves a constant per-round probability of geometric updates.

The proof outline of Theorem 5 is as follows. By Lemma 6, the probability of $a_t = 0$ is $\Theta(\epsilon_t) = \Theta(w_t)$. Together with Lemma 8, it follows that, in round t, with probability $\Theta(w_t)$, the width of confidence region w_t shrinks geometrically to $w_{t+1} = C_{w,1} w_t$ or smaller. This leads an exponential convergence of $\hat{\theta}_t$ to the total number of users to which the recommender has sent $a_t = 0$. Finally, when $\epsilon_t = (3/4)w_t$, Lemma 7 ensures that the per-round regret is $O(w_t^2)$. Let us refer to an interval between two geometric intervals as an *epoch*. Since a geometric update occurs with probability $\Theta(w_t)$, the expected number of rounds contained in one epoch is $\Theta(1/w_t)$. The regret incurred per round is $\Theta(w_t^2)$; thus, the total regret incurred in one epoch is $\Theta(w_t^2 \times 1/w_t) = \Theta(w_t)$. Accordingly, the expected regret associated with each epoch is bounded by a geometric sequence whose common ratio is $C_{w,1} = 7/8 < 1$. The total regret is the sum of the regret from all the epochs. Consequently, the total regret is bounded by the sum of a geometric series, which converges to a constant.

6 Simulations

This section provides the simulation results. For each path, we draw θ from $\mathrm{Unif}[-1, 1]$, and x_t, z_t from the standard normal distribution i.i.d. for $T = 10,000$ rounds. We analyze how regret $\mathrm{Reg}(t)$ and the width of the confidence region w_t evolve over time under the binary and ternary recommendation policies.

For all graphs in this section, the lines are averages over 500 sample paths, and the shaded areas cover between 25 and 75 percentiles over runs. The whiskers drawn at the final round ($T = 10,000$) are the two-sigma confidence intervals of the average values. [13]

6.1 Regret

We plot the cumulative regret, $\mathrm{Reg}(t)$, in Figures 3 and 4. As proved in Theorem 2, the simulation result implies almost-linear growth of the cumulative regret under the binary recommendation: $\mathrm{Reg}(T) \coloneqq \tilde{\Omega}(T)$. In contrast, under the ternary recommendation, the cumulative regret is bounded by a constant. This observation is also consistent with Theorem 5, which shows that $\mathrm{Reg}(T) = O(1)$ under the ternary recommendation. The simulation result additionally demonstrates that there is a large quantitative difference between the magnitude of total regret incurred by these two policies. For $T = 10,000$, while the binary recommendation policy produces $\mathrm{Reg}(T) \approx 10$ on average, that under the ternary recommendation policy produces $\mathrm{Reg}(T) \approx 0.2$. Moreover, we can infer from Figures 3 and 4 that the difference

[13] The implementation of the simulation is available at https://github.com/jkomiyama/deviationbasedlearning.

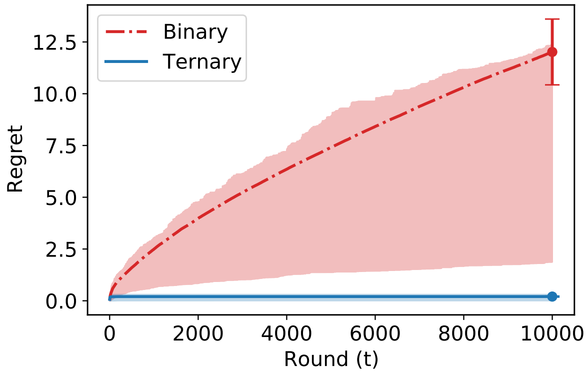

Figure 3: The evolution of cumulative regret $\mathrm{Reg}(t)$ under binary and ternary recommendation.

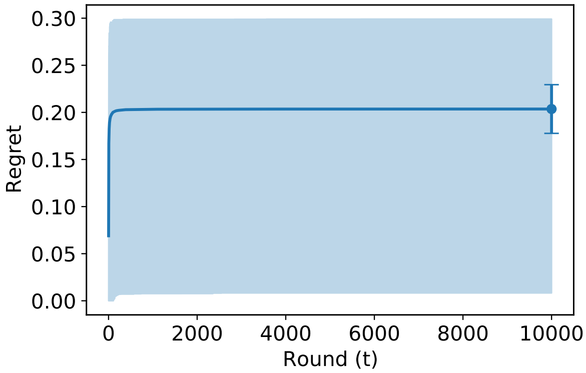

Figure 4: The evolution of cumulative regret $\mathrm{Reg}(t)$ under ternary recommendation. The blue line in this figure is identical to the blue line in Figure 3.

would grow almost linearly as we further increase the number of rounds T.

6.2 Learning

We now investigate how the width of the confidence region, w_t, is updated. First, we focus on when and how frequently updates occur. We count the occurrence of belief updates, i.e., the number of rounds such that $w_{t+1} < w_t$. Among all rounds in which updates occur, (i) the set of rounds in which the user followed the recommendation is denoted by $\mathcal{T}_{\mathrm{Obey}}$ (obedience), (ii) the set of rounds in which the user deviated from the recommendation is denoted by $\mathcal{T}_{\mathrm{Deviate}}$ (deviation), and (iii) the set of rounds in which the recommender did not recommend a particular action is denoted by $\mathcal{T}_{\mathrm{OtF}}$ (on the fence). More formally,

$$\begin{aligned}
\mathcal{T}_{\mathrm{Obey}}(t) &\coloneqq \{s \in [t] : w_{s+1} < w_s \text{ and } a_s = b_s\};\\
\mathcal{T}_{\mathrm{Deviate}}(t) &\coloneqq \{s \in [t] : w_{s+1} < w_s, a_s \neq 0 \text{ and } a_s \neq b_s\};\\
\mathcal{T}_{\mathrm{OtF}}(t) &\coloneqq \{s \in [t] : w_{s+1} < w_s \text{ and } a_s = 0\}.
\end{aligned}$$

Note that $|\mathcal{T}_{\mathrm{OtF}}| = 0$ for the case of the binary recommendations since $a_t = 0$ is never sent.

Figures 5 and 6 plot the number of updates, $|\mathcal{T}_{\mathrm{Obey}}(t)|$, $|\mathcal{T}_{\mathrm{Deviate}}(t)|$, $|\mathcal{T}_{\mathrm{OtF}}(t)|$, and their total, $|\mathcal{T}_{\mathrm{Obey}}(t)|+|\mathcal{T}_{\mathrm{Deviate}}(t)|+|\mathcal{T}_{\mathrm{OtF}}(t)|$. Figure 5 exhibits the case of binary recommendations. The updates by obedience occurs more often than the updates by deviation.

Figure 6 represents the case of ternary recommendations. Since the recommender recommends an arm only if she is confident about it, users follow the recommendation blindly whenever an arm is recommended; therefore, an update occurs only if the recommender con-

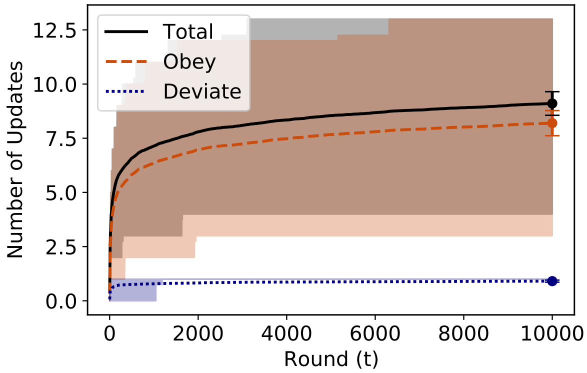

Figure 5: Number of updates under binary recommendations.

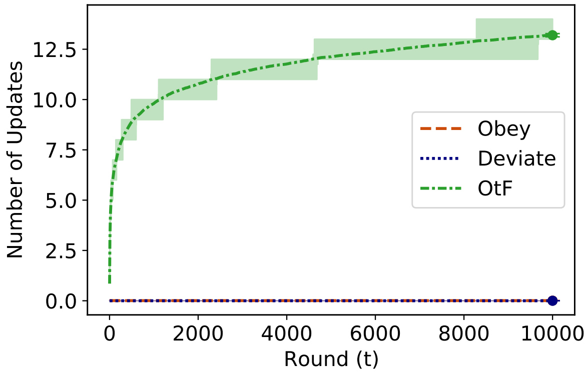

Figure 6: Number of updates under ternary recommendations. We observe $|\mathcal{T}_{\text{Obey}}(T)| = |\mathcal{T}_{\text{Deviate}}(T)| \approx 0$; thus, the total number of updates until round t is approximately equal to $|\mathcal{T}_{\text{OtF}}(t)|$.

fesses that she is on the fence. The opportunities for her learning are mostly concentrated at the beginning of the game, but updates occur occasionally even in the later stages of the game. On average, updates occur more frequently than in the case of binary recommendations.

Next, we evaluate the total amount of information acquired from each recommendation. We measure the *accuracy* of the estimation in round t by

$$\text{ACC}(t) \coloneqq -\log\left(w_{t+1}/2\right).$$

The value w_{t+1} is the width of the confidence region after the round-t update. Note that $w_1 = u_1 - l_1 = 1 - (-1) = 2$, and therefore, ACC(0) is normalized to zero.

We define the *accuracy gain* from each recommendation as follows:

$$\text{ACC}_{\text{obey}}(t) \coloneqq -\sum_{s \in \mathcal{T}_{\text{Obey}}} \log(w_{s+1}/w_s);$$

$$\text{ACC}_{\text{deviate}}(t) \coloneqq -\sum_{s \in \mathcal{T}_{\text{Deviate}}} \log(w_{s+1}/w_s);$$

$$\text{ACC}_{\text{otf}}(t) \coloneqq -\sum_{s \in \mathcal{T}_{\text{OtF}}} \log(w_{s+1}/w_s).$$

Note that it is always the case that $\text{ACC}(t) = \text{ACC}_{\text{obey}}(t) + \text{ACC}_{\text{deviate}}(t) + \text{ACC}_{\text{otf}}(t)$.

Figures 7 and 8 depict the accuracy gain from each recommendation under the cases of binary recommendations and ternary recommendations. As illustrated in Figure 5 under the binary recommendation policy, learning from obedience occurs more frequently than learning

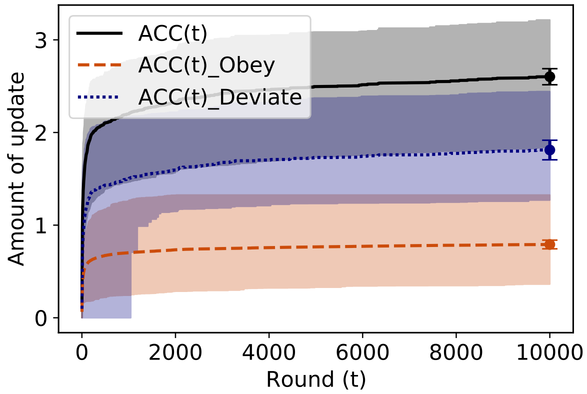

Figure 7: The breakdown of accuracy gains under binary recommendations.

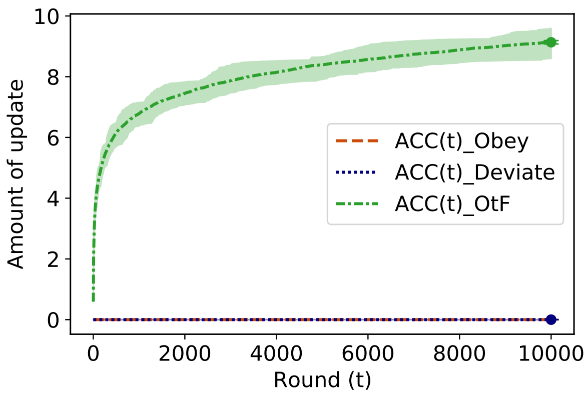

Figure 8: The breakdown of accuracy gains under ternary recommendations. Almost all the accuracy gains are from $a_t = 0$; thus, $\text{ACC}(t) \approx \text{ACC}_{\text{otf}}(t)$.

from deviations. Nevertheless, Figure 7 shows that the recommender acquires information more from deviations than obedience. This is because once a deviation occurs, it is much more informative than obedience (as implied by Theorem 1).

Figure 8 reveals that, under the ternary recommendation, almost all the accuracy gains are obtained when the recommender signals being on the fence. For any stage of the game, the learning rate is higher than that under binary recommendations, and the difference is quantitatively large. In round 10,000, the average accuracy under the ternary policy becomes larger than that under the binary policy by (roughly) six points, which implies that w_T under the ternary policy is $e^6 \approx 403$ times smaller than w_T under the binary policy.

7 Concluding Remarks

In this paper, we propose deviation-based learning, a novel approach for training recommender systems. Our approach is built upon a simple idea. When a user deviates from a recommendation, he is aware of the merit of the recommended option, but has concluded that another option provides him with a better payoff. This event indicates that the recommender has misestimated the user's preference, and so the recommender can update her estimate based on this information. Conversely, if a user follows a recommendation even though the recommender is not perfectly confident, then the recommender can increase her confidence. The deviation-based learning is effective when (i) payoffs are unobservable, (ii) there are many knowledgeable experts, and (iii) the recommender can easily identify the set of experts.

Our analysis reveals that the size of the message space is crucial for the efficiency of learning. In a stylized model with two arms, we demonstrated that a binary message space results in a large welfare loss. After the recommender is trained to some extent, users start to follow her recommendations blindly, and users' decisions are uninformative for advancing the recommender's learning. This effect significantly slows down her learning, and the total regret grows almost linearly in the number of users. In contrast, when the message space is ternary, the recommender can sometimes disclose the fact that she predicts that two arms will produce similar payoffs. With the ternary message space, the total regret is bounded by a constant (which does not depend on the number of users).

Our analysis of the binary recommendation policy also provides a simple but useful caveat: the recommender should not consider the rate at which users follow recommendations to be a key performance indicator. When the recommender has an information advantage, the user may follow a recommendation even when it does not fully respect his own information and preference. Accordingly, if such a performance indicator is used, then the recommendation system may incur a large welfare loss, and the recommender may not be able to realize this fact.

Future studies could investigate deviation-based learning in more complex environments. In practice, observable contexts (x_t) are often multi-dimensional. Furthermore, users' payoffs are rarely linear in the parameter (θ), and their functional form may be unknown ex ante; thus, the recommender may have to adopt a nonparametric approach. While we believe that the insight obtained from our stylized model will be informative in general environments, more comprehensive and exhaustive analyses are necessary for practical applications.

References

Adomavicius, G. and A. Tuzhilin (2005): "Toward the Next Generation of Recommender Systems: a Survey of the State-of-the-Art and Possible Extensions," *IEEE Transactions on Knowledge and Data Engineering*, 17, 734–749.

Ahuja, R. K. and J. B. Orlin (2001): "Inverse Optimization," *Operations Research*, 49, 771–783.

Bergemann, D. and S. Morris (2016a): "Bayes Correlated Equilibrium and the Comparison of Information Structures in Games," *Theoretical Economics*, 11, 487–522.

——— (2016b): "Information Design, Bayesian Persuasion, and Bayes Correlated Equilibrium," *American Economic Review*, 106, 586–91.

Besbes, O., Y. Fonseca, and I. Lobel (2021): "Contextual Inverse Optimization: Offline and Online Learning," *CoRR*, abs/2106.14015.

Che, Y.-K. and J. Hörner (2017): "Recommender Systems as Mechanisms for Social Learning," *The Quarterly Journal of Economics*, 133, 871–925.

Chen, J., H. Dong, X. Wang, F. Feng, M. Wang, and X. He (2020): "Bias and Debias in Recommender System: A Survey and Future Directions," .

Feller, W. (1968): *An Introduction to Probability Theory and Its Applications.*, vol. 1 of *Third edition*, New York: John Wiley & Sons Inc.

Kamenica, E. and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590–2615.

Khoury, R. E. (2019): "Google Maps Hits 5 Billion Downloads on the Play Store, Does It after YouTube but Before the Google App," *Android Police*, https://www.androidpolice.com/2019/03/09/google-maps-hits-5-billion-downloads-on-the-play-store-does-it-after-youtube-but-before-the-google-app/.

Kremer, I., Y. Mansour, and M. Perry (2014): "Implementing the 'Wisdom of the Crowd'," *Journal of Political Economy*, 122, 988–1012.

Lai, T. and H. Robbins (1985): "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics*, 6, 4–22.

Luca, M. and G. Zervas (2016): "Fake It till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 62, 3412–3427.

Marlin, B. M. and R. S. Zemel (2009): "Collaborative Prediction and Ranking with Non-random Missing Data," in *Proceedings of the Third ACM Conference on Recommender Systems*, 5–12.

Mayzlin, D., Y. Dover, and J. Chevalier (2014): "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104, 2421–55.

Muchnik, L., S. Aral, and S. J. Taylor (2013): "Social Influence Bias: A Randomized Experiment," *Science*, 341, 647–651.

Myerson, R. B. (1982): "Optimal Coordination Mechanisms in Generalized Principal–Agent Problems," *Journal of Mathematical Economics*, 10, 67–81.

Ng, A. Y. and S. J. Russell (2000): "Algorithms for Inverse Reinforcement Learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, 663–670.

Salganik, M. J., P. S. Dodds, and D. J. Watts (2006): "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science*, 311, 854–856.

Samuelson, P. A. (1938): "A Note on the Pure Theory of Consumer's Behaviour," *Economica*, 5, 61–71.

Sauré, D. and J. P. Vielma (2019): "Ellipsoidal Methods for Adaptive Choice-Based Conjoint Analysis," *Operations Research*, 67, 315–338.

Sinha, A., D. F. Gleich, and K. Ramani (2016): "Deconvolving Feedback Loops in Recommender Systems," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 29.

Suro, M. A. (2018): "How to Avoid Sketchy Neighborhoods When Driving," https://www.richmiser.com/avoid-sketchy-neighborhoods/. Accessed on 07/18/2021.

Sutton, R. S. and A. G. Barto (2018): *Reinforcement Learning: An Introduction*, The MIT Press, second ed.

Thompson, W. R. (1933): "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples," *Biometrika*, 25, 285–294.

Toubia, O., J. Hauser, and R. Garcia (2007): "Probabilistic Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis: Theory and Application," *Marketing Science*, 26, 596–610.

Appendix

A Proofs

A.1 Proof of Theorem 1

Proof of Theorem 1. For ease of discussion, we assume $x_t > 0$. The case of $x_t < 0$ can be proved in a similar manner.

Case 1. $a_t = b_t = 1$.

We have

$$\begin{aligned} w_{t+1} &= u_{t+1} - l_{t+1} \\ &= u_t - \max(l_t, (-Z_t/x_t)) \quad \text{(by Eq. (2) and (3))} \\ &> u_t - m_t \quad \text{(by } m_t > l_t \text{ and } a_t = 1 \text{ if and only if } m_t > -z_t/x_t\text{)} \\ &= (1/2)w_t. \end{aligned} \tag{6}$$

Case 2. $a_t = b_t = -1$

We have

$$\begin{aligned} w_{t+1} &= u_{t+1} - l_{t+1} \\ &= \min(u_t, (-Z_t/x_t)) - l_t \quad \text{(by Eq. (2) and (3))} \\ &< m_t - l_t \qquad \text{(by } m_t < u_t \text{ and } a_t = -1 \text{ if and only if } m_t < -z_t/x_t\text{)} \\ &= (1/2)w_t. \end{aligned} \tag{7}$$

Eq. (4) follows from Eq. (6) and (7).

Case 3. $a_t = -1$, $b_t = 1$

We have

$$\begin{aligned} w_{t+1} &= u_{t+1} - l_{t+1} \\ &\leq u_t - (-Z_t/x_t) \quad \text{(by Eq. (2) and (3))} \\ &< u_t - m_t \quad \text{(by } a_t = -1\text{)} \\ &= (1/2)w_t. \end{aligned} \tag{8}$$

Case 4. $a_t = 1$, $b_t = -1$

We have

$$\begin{aligned} w_{t+1} &= u_{t+1} - l_{t+1} \\ &\leq (-Z_t/x_t) - l_t \quad \text{(by Eq. (2) and (3))} \\ &< m_t - l_t \quad \text{(by } a_t = 1\text{)} \\ &= (1/2)w_t. \end{aligned} \tag{9}$$

Eq. (5) follows from Eq. (8) and (9).

□

A.2 Proof of Theorem 2

Proof of Theorem 2. Let $C_1 = (1/2)C_{\text{update}}$ and

$$\mathcal{Z}(t) := \left\{ w_t \leq \frac{C_1}{\log T}, |\theta - m_t| \geq \frac{C_1}{2\log T} \right\}.$$

In the following, we first show the following inequality.

Claim 2.a.

$$\mathbb{P}\left[\mathcal{Z}(3)\right] \geq \frac{C_2}{\text{polylog}(T)}$$

for some constant $C_2 > 0$.

Proof. Let

$$\begin{aligned} \mathcal{A} &:= \left\{ u_2 \leq \theta + \frac{C_1}{6\log T} \right\}, \\ \mathcal{B} &:= \left\{ \theta - \frac{5C_1}{6\log T} \leq l_3 \leq \theta - \frac{2C_1}{3\log T} \right\}. \end{aligned}$$

Note that $\mathcal{A} \cap \mathcal{B} \subseteq \mathcal{Z}(3)$.

In order to evaluate the probability that $\mathcal{Z}(3)$ occurs, in the following, we evaluate the probability that $\mathcal{A}$ and $\mathcal{B}$ occur, assuming $\theta > 2C_1/(\log T)$ (which occurs with probability $\Theta(1)$ for sufficiently large T).

Claim 2.b. $\mathbb{P}[\mathcal{A}] = \Theta\left(1/(\log T)\right)$.

Proof. Recall that $(l_1, u_1, m_1) = (-1, 1, 0)$. Let $\sigma_1 = \int_0^\infty 2\phi(x)dx$, which is equal to $-Z_1$

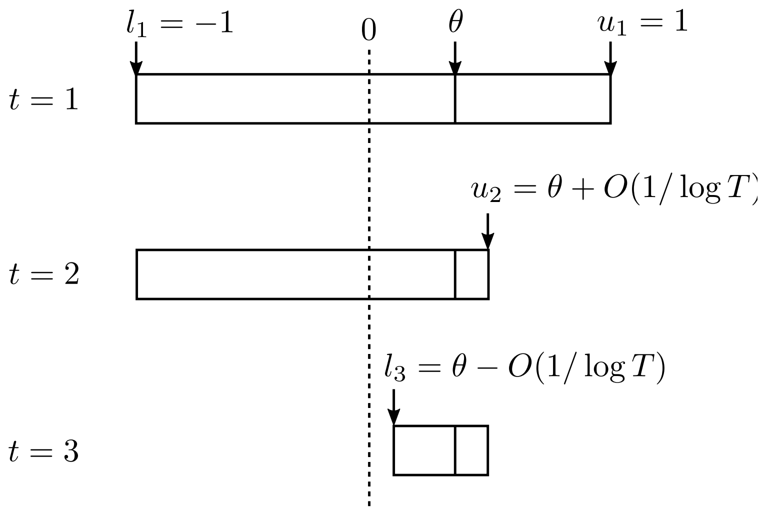

Figure 9: Illustration of $(l_t, u_t)_{t=1,2,3}$ in the instance of Theorem 2. Here, $w_3 = O(1/\log T)$ holds, which implies that (l_t, u_t) has very small chance of being updated again.

given $b_1 = -1$. Under $a_1 = -1$, $Z_1 = -\sigma_1$. Let

$$\mathcal{A}' := \{z_1 < 0\} \cap \left\{ \frac{\sigma_1}{\theta + \frac{C_1}{6 \log T}} \leq x_1 \leq \frac{\sigma_1}{\theta} \right\}.$$

In the following, we show that $\mathcal{A}'$ implies $\mathcal{A}$ and $\mathbb{P}[\mathcal{A}'] = \Omega(1/\log T)$.

$$\begin{aligned} \mathcal{A}' &= \mathcal{A}' \cap \{a_1 = -1\} \quad (\text{by } x_1 m_1 + z_1 = z_1 < 0) \\ &= \mathcal{A}' \cap \{a_1 = -1, b_1 = -1\} \quad (\text{by } x_1\theta + Z_1 = x_1\theta - \sigma_1 < 0) \\ &= \mathcal{A}' \cap \left\{ a_1 = -1, b_1 = -1, u_2 \leq \theta + \frac{C_1}{6 \log T} \right\} \quad (\text{by Eq. (3)}) \\ &\subseteq \mathcal{A}. \end{aligned} \tag{10}$$

Therefore,

$$\begin{aligned} \mathbb{P}[\mathcal{A}] &\geq \mathbb{P}[\mathcal{A}'] \quad (\text{by Eq. (10)}) \\ &= \mathbb{P}[z_1 < 0] \times \mathbb{P}\left[\frac{\sigma_1}{\theta + \frac{C_1}{6 \log T}} \leq x_1 \leq \frac{\sigma_1}{\theta} \right] \\ &= \frac{1}{2} \mathbb{P}\left[\frac{\sigma_1}{\theta + \frac{C_1}{6 \log T}} \leq x_1 \leq \frac{\sigma_1}{\theta} \right] \\ &= \frac{1}{2} \int_{\frac{\sigma_1}{\theta + \frac{C_1}{6 \log T}}}^{\frac{\sigma_1}{\theta}} \phi(x) dx \end{aligned}$$

$$
\begin{aligned}
&\geq \frac{\sigma_1}{2\theta^2}\frac{C_1}{6\log T} \times \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\sigma_1^2}{2\theta^2}\right) \quad (\text{for } \theta \geq C_1/(6\log T)) \\
&= \Theta\left(\frac{1}{\log T}\right) \quad (\text{since } \sigma_1, C_1, \theta = \Theta(1)).
\end{aligned}
$$

□

Claim 2.c. The probability $\mathbb{P}[\mathcal{B}|\mathcal{A}] = \Theta\left(1/(\log T)\right)$.

Proof. Let

$$
\mathcal{B}' = \{x_2 > 0\} \cap \{x_2 m_2 + z_2 < 0\} \cap \left\{\frac{-Z_2}{\theta - \frac{2C_1}{3\log T}} \leq x_2 \leq \frac{-Z_2}{\theta - \frac{5C_1}{6\log T}}\right\}.
$$

We have

$$
\begin{aligned}
\mathcal{B}' &= \mathcal{B}' \cap \{a_2 = -1\} \quad (\text{by } x_2 m_2 + z_2 < 0) \\
&= \mathcal{B}' \cap \{a_2 = -1, b_2 = 1\} \quad (\text{by } x_2\theta + Z_2 > 2C_1/(3\log T) > 0) \\
&= \mathcal{B}' \cap \left\{a_2 = -1, b_2 = 1, \theta - \frac{2C_1}{3\log T} \leq l_3 \leq \theta - \frac{5C_1}{6\log T}\right\} \quad (\text{by Eq. (2)}) \\
&\subseteq \mathcal{B}.
\end{aligned}
$$

Furthermore, by using the fact that $-Z_2 \in (0, \sigma_1) = \Theta(1)$ and $\sigma_1 = \Theta(1)$, we have the following under $\{l_2 < 0, x_2 m_2 < 0\}$:

$$
\begin{aligned}
\mathbb{P}[\mathcal{B}|\mathcal{A}] &\geq \mathbb{P}[\mathcal{B}'|\mathcal{A}] \\
&= \mathbb{P}\left[\frac{-Z_2}{\theta - \frac{2C_1}{3\log T}} \leq x_2 \leq \frac{-Z_2}{\theta - \frac{5C_1}{6\log T}}, x_2 m_2 + z_2 < 0\right] \\
&\geq \mathbb{P}\left[\frac{-Z_2}{\theta - \frac{2C_1}{3\log T}} \leq x_2 \leq \frac{-Z_2}{\theta - \frac{5C_1}{6\log T}}\right] \times \frac{1}{2} \\
&\qquad (\text{by } x_2 m_2 \leq 0) \\
&\geq \frac{C_1}{6\log T}\frac{-Z_2}{2\theta^2}\phi\left(\frac{-2Z_2}{\theta}\right) \times \frac{1}{2} \\
&\qquad (\text{for } \theta \geq 2 \times \frac{5C_1}{6\log T}) \\
&= \Theta\left(\frac{1}{\log T}\right). \quad (\text{since } Z_2, C_1, \theta = \Theta(1))
\end{aligned}
$$

□

Combining these claims, we have

$$\mathbb{P}[\mathcal{Z}(3)] \geq \mathbb{P}[\mathcal{A} \cap \mathcal{B}] = \mathbb{P}[\mathcal{A}] \times \mathbb{P}[\mathcal{B}|\mathcal{A}] = \Omega\left(\frac{1}{\log T} \times \frac{1}{\log T}\right) = \Omega\left(\frac{1}{(\log T)^2}\right), \tag{11}$$

as desired. □

Note that, by Lemma 4, $\mathcal{Z}(3)$ implies $\mathcal{Z}(4) \cap \mathcal{Z}(5) \cap \cdots \cap \mathcal{Z}(T)$ with probability at least $1 - 1/T$. It follows that $\mathbb{E}[\mathrm{Reg}(T)]$ is bounded as

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(T)] &\geq \mathbb{E}\left[\sum_{t=3}^{T} \mathrm{reg}(t) \middle| \mathcal{Z}(3)\right] \Omega\left(\frac{1}{(\log T)^2}\right) \quad \text{(by Eq. (11))} \\
&\geq \left(1 - \frac{1}{T}\right) \mathbb{E}\left[\sum_{t=3}^{T} \mathrm{reg}(t) \middle| \bigcap_{t=3}^{T} \mathcal{Z}(t)\right] \Omega\left(\frac{1}{(\log T)^2}\right) \\
&\qquad \text{(by Lemma 4 and construction of } \mathcal{Z}(3)) \\
&= \Theta(1) \times \Omega\left(\frac{T}{(\log T)^2}\right) \times \Omega\left(\frac{1}{(\log T)^2}\right) \\
&\qquad \text{(by Lemma 3, } \mathcal{Z}(t) \text{ implies } \mathbb{E}[\mathrm{reg}(t)] = \Omega(w_t^2) = \Omega(1/(\log T)^2)) \\
&= \Omega\left(\frac{T}{\mathrm{polylog}(T)}\right).
\end{aligned}$$

□

A.3 Lemma 9

We prove a lemma that is useful to prove Lemmas 3 and 4.

Lemma 9 (Gap between Z_t and $-x_t m_t$: Binary Case)**.** There exist universal constants $C_{\mathrm{l}}, C_{\mathrm{u}} > 0$ such that the following inequalities hold.

1. If $\mathrm{sgn}(x_t m_t) a_t < 0$, then

$$C_{\mathrm{l}} \min(1, 1/|x_t|) < a_t(Z_t + x_t m_t) < C_{\mathrm{u}}. \tag{12}$$

2. If $\mathrm{sgn}(x_t m_t) a_t > 0$, then

$$C_{\mathrm{l}} < a_t(Z_t + x_t m_t). \tag{13}$$

The term $\min(1, 1/|x|)$ in Eq. (12) is derived from the fact that $e^{-x^2/2}$ decays faster for

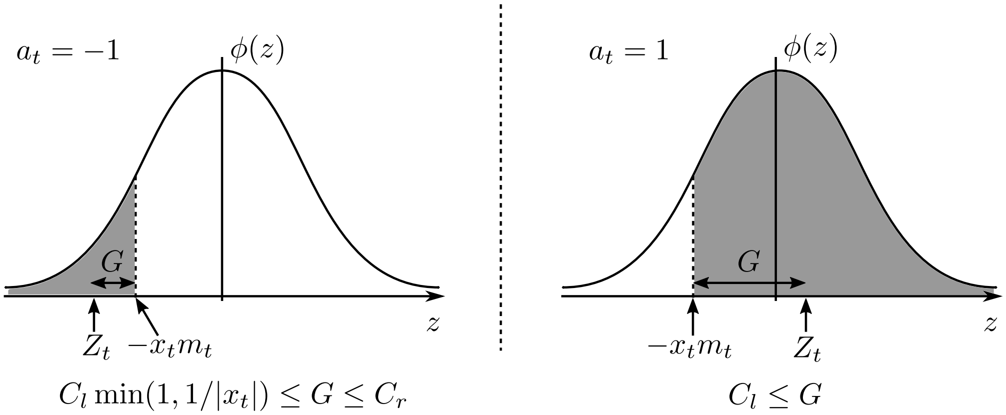

Figure 10: Illustration of Lemma 9 when $x_t m_t > 0$. The lemma bounds $G := a_t(Z_t + x_t m_t)$. The left figure corresponds to Eq. (12), whereas the right figure corresponds to Eq. (13).

a large $|x|$. It is analogous to the equation

$$\frac{1}{(x+1)^2 - x^2} = \frac{1}{2x+1} \geq \frac{1}{3} \min\left(1, \frac{1}{x}\right)$$

for $x > 0$.

Proof of Lemma 9. For ease of discussion, we assume $x_t m_t \geq 0$ (which aligns with Figure 1). The case of $x_t m_t < 0$ can be dealt with the same discussion.

Let $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ be the pdf of the standard normal distribution and $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ be the error function. Let $a_t = -1$ and $C_2 = \min(1, 1/|x_t m_t|)$. Then,

$$\begin{aligned} Z_t + x_t m_t &= \frac{\int_{-\infty}^{-x_t m_t} (z + x_t m_t)\phi(z) dz}{\int_{-\infty}^{-x_t m_t} \phi(z) dz} \\ &\leq \frac{1}{\phi(-x_t m_t)} \int_{-\infty}^{-x_t m_t} (z + x_t m_t)\phi(z) dz \\ &\leq \frac{1}{\phi(-x_t m_t)} \int_{-x_t m_t - 2C_2}^{-x_t m_t - C_2} (z + x_t m_t)\phi(z) dz \\ &\leq \frac{-C_2}{\phi(-x_t m_t)} \int_{-x_t m_t - 2C_2}^{-x_t m_t - C_2} \phi(z) dz \\ &\leq \frac{-C_2}{\phi(-x_t m_t)} \min_{z \in [-x_t m_t - 2C_2, -x_t m_t - C_2]} \phi(z) \end{aligned}$$

$$
\begin{aligned}
&\le -C_2 \min_{z\in[-x_t m_t - 2C_2, -x_t m_t - C_2]} e^{-(3/2)} = -C_2 e^{-(3/2)} \\
&\qquad \text{(by } e^{-(x+a)^2/2}/e^{-x^2/2} = e^{-xa-a^2/2} \text{ and } |x_t m_t| C_2 \le 1) \\
&= -e^{-(3/2)} \min(1, 1/|x_t m_t|) \le -e^{-(3/2)} \min(1, 1/|x_t|),
\end{aligned}
$$

which implies the first inequality[14] of Eq. (12).

Moreover,

$$
\begin{aligned}
Z_t + x_t m_t &= \mathbb{E}_{z\sim\mathcal{N}^{\mathrm{tr}}_{-\infty,-x_t m_t}}[z] + x_t m_t \\
&= \frac{\int_{-\infty}^{-x_t m_t} (z + x_t m_t)\phi(z)dz}{\int_{-\infty}^{-x_t m_t} \phi(z)dz} \\
&\ge \frac{\int_{-\infty}^{0} z\phi(z)dz}{\int_{-\infty}^{0} \phi(z)dz} \\
&\qquad \text{(by } \phi(x+c)/\phi(x) \le \phi(c) \text{ for any } x, c \le 0) \\
&= -\frac{\int_{0}^{\infty} z\phi(z)dz}{\int_{0}^{\infty} \phi(z)dz} \\
&= -\sqrt{\frac{2}{\pi}},
\end{aligned}
$$

which is a constant and implies the second inequality[15] of Eq. (12).

If $a_t = 1$, then

$$
\begin{aligned}
Z_t + x_t m_t &= \mathbb{E}_{z\in\mathcal{N}^{\mathrm{tr}}(-x_t m_t, \infty)}[z] + x_t m_t \\
&\ge \frac{1}{2}\mathbb{E}_{z\in\mathcal{N}^{\mathrm{tr}}(0,\infty)}[z] \\
&= \sqrt{\frac{1}{2\pi}},
\end{aligned}
$$

which implies Eq. (13).

$\square$

A.4 Proof of Lemma 3

Proof of Lemma 3. Without loss of generality, we assume $m_t \ge 0$. (Otherwise, by using the symmetry of the model, we may flip the sign of variables as $(l_t, u_t, \theta) = (-u_t, -l_t, -\theta)$ and apply the same analysis to obtain the same result.) For the ease of discussion, we assume

[14] Note that $a_t = -1$ and the inequality here is flipped.

[15] Again, $a_t = -1$ and the inequality here is flipped.

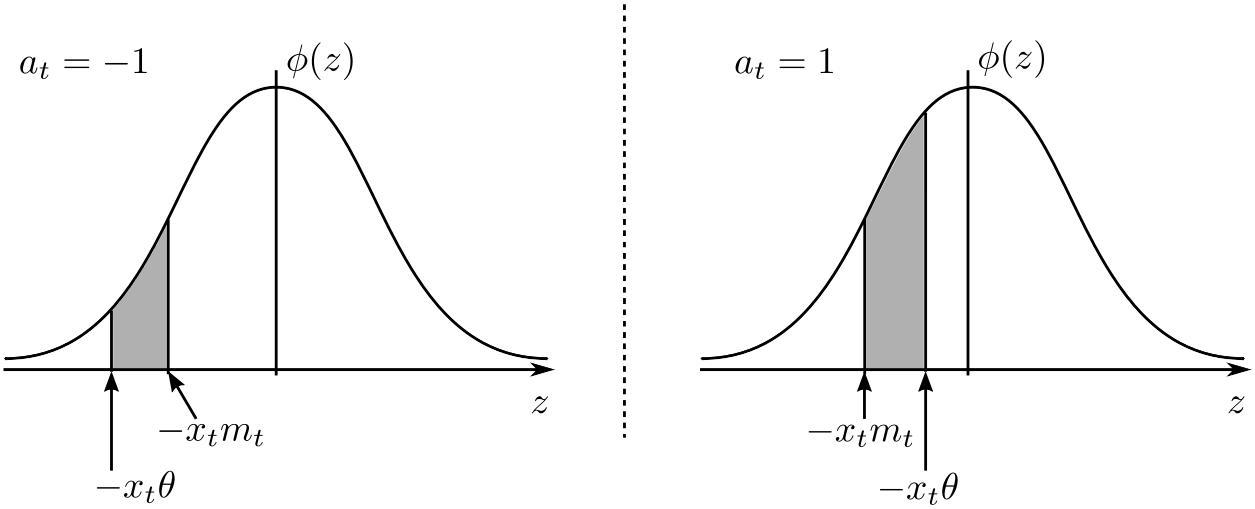

Figure 11: Recommendation a_t is determined by the sign of $x_t m_t + z_t$, whereas the true superior arm is determined by $x_t\theta + z_t$. When $z_t \in [-x_t m_t, -x_t\theta]$, the recommender fails to recommend the superior arm.

$m_t - \theta > 0$. (In the case of $m_t - \theta < 0$, we can follow essentially the same discussion as the case of $m_t - \theta > 0$.)

For $t \in [T]$ and $l > 0$, let

$$\mathcal{C}(t,l) := \{-x_t m_t < z_t < -x_t\theta - l\} \cap \{x_t(m_t - \theta) < C_1\}.$$

Claim 4.a. $\mathcal{C}(t,l) \subseteq \{\mathrm{reg}(t) \geq l\}$.

Proof.

$$\begin{aligned}
&\{-x_t m_t < z_t < -x_t\theta - l\} \cap \{x_t(m_t - \theta) < C_1\}\\
&= \{0 < x_t m_t + z_t\} \cap \{x_t\theta + z_t < -l\} \cap \{x_t(m_t - \theta) < C_1\}\\
&\subseteq \{0 < x_t m_t + z_t\} \cap \{x_t\theta + z_t < -l\} \cap \{x_t\theta + Z_t > 0\}\\
&\qquad \text{(by Eq. (13))}\\
&= \{x_t\theta + z_t < -l\} \cap \{b_t^* = -1\} \cap \{a_t = 1\} \cap \{b_t = 1\}.
\end{aligned}$$

It follows from $b_t \neq b_t^*$ and $x_t\theta + z_t < -l$ that $\mathrm{reg}(t) > l$. □

Claim 4.b. $\mathbb{P}[\mathcal{C}(t,l)] \geq \Theta(m_t - \theta - l)$.

Proof. By using the fact that $C_{\rm l}$ is a universal constant and $1 \geq m_t > \theta \geq -1$, we have

$$\begin{aligned}
\mathbb{P}[x_t(m_t - \theta) < C_{\rm l}] &\geq \mathbb{P}[2x_t < C_{\rm l}, x_t > 0] \\
&\geq \mathbb{P}[C_{\rm l}/2 < 2x_t < C_{\rm l}, x_t > 0] \\
&\geq \int_{C_{\rm l}/4}^{C_{\rm l}/2} \phi(x)dx \\
&= \Theta(1).
\end{aligned}$$

Moreover, for any $C_{\rm l}/4 \leq x_t \leq C_{\rm l}/2$,

$$\begin{aligned}
\mathbb{P}[-x_t m_t < z_t < -x_t\theta - l] &= \int_{-x_t m_t}^{-x_t\theta - l} \phi(z)dz \\
&= x_t(m_t - \theta - l) \min_{z \in \{-x_t m_t, -x_t\theta - l\}} \phi(z) \\
&= \Theta(m_t - \theta - l).
\end{aligned}$$

Therefore, $\mathbb{P}[\mathcal{C}(t,l)] \geq \Theta(1) \times \Theta(m_t - \theta - l) = \Theta(m_t - \theta - l)$. □

Combining Claims 4.a and 4.b, the regret is bounded as follows:

$$\begin{aligned}
\mathbb{E}[\text{reg}(t)] &\geq \int_{l=0}^{\infty} \mathbb{P}[\text{reg}(t) \geq l]dl \\
&\geq \int_{l=0}^{\infty} \mathbb{P}[\mathcal{C}(t,l)]dl \quad \text{(by Claim 4.a)} \\
&\geq \int_{l=0}^{m_t - \theta} \Theta(m_t - \theta - l)dl \quad \text{(by Claim 4.b)} \\
&= \Omega((m_t - \theta)^2).
\end{aligned}$$

□

A.5 Proof of Lemma 4

Proof of Lemma 4. Let

$$\begin{aligned}
\mathcal{U}_1(t) &= \{b_t \text{sgn}(x_t) < 0\} \cap \{b_t > -Z_t/x_t\}, \\
\mathcal{U}_2(t) &= \{b_t \text{sgn}(x_t) > 0\} \cap \{a_t < -Z_t/x_t\}, \\
\mathcal{U}(t) &= \mathcal{U}_1(t) \cup \mathcal{U}_2(t).
\end{aligned}$$

By the update rule (Eq. (2) and (3)), event $\mathcal{U}(t)$ is equivalent to $(a_{t+1}, b_{t+1}) \neq (a_t, b_t)$.

Eq. (12) and (13) in Lemma 9 imply

$$|Z_t - x_t m_t| \geq C_l \min(1, 1/|x_t|). \tag{14}$$

Accordingly,

$$\begin{aligned}
\mathcal{U}(t) &= \mathcal{U}_1(t) \cup \mathcal{U}_2(t) \\
&\subseteq \{b_t > -Z_t/x_t\} \cup \{a_t < -Z_t/x_t\} \\
&\subseteq \{w_t/2 > C_l \min(1/|x_t|, 1/|x_t|^2)\} \\
&\quad \text{(by } u_t - m_t = m_t - l_t = w_t/2 \text{ and Eq. (14)).}
\end{aligned}$$

For a sufficiently small w_t,[16]

$$\begin{aligned}
\mathbb{P}[\mathcal{U}(t)] &\leq \mathbb{P}\left[w_t > \frac{2C_l}{x_t^2}\right] \\
&= \mathbb{P}\left[x_t^2 > \frac{2C_l}{w_t}\right] \\
&= 2\Phi^c\left(\sqrt{\frac{2C_l}{w_t}}\right) \\
&\leq \exp\left(-\frac{2C_l}{w_t}\right) \times 2\Phi^c(0) \\
&= \exp\left(-\frac{2C_l}{w_t}\right),
\end{aligned}$$

which completes the proof. □

A.6 Proof of Theorem 5

Proof of Theorem 5. Let

$$\mathcal{E}(t) = \{w_{t+1} \leq C_{w,1} w_t\}.$$

Lemmas 6 and 8 imply that there exists a universal constant $C_{\text{shrink}} > 0$ such that

$$\mathbb{P}[\mathcal{E}(t)] \geq C_{\text{shrink}} w_t. \tag{15}$$

Lemma 7 states that

$$\mathbb{E}[\text{reg}(t)] \leq C_{\text{regt}} w_t^2. \tag{16}$$

[16] $w_t \leq 2C$ is enough to assure $\{w_t/2 > C_l/|x_t|\} \subseteq \{w_t/2 > C_l/|x_t|^2\}$.

For $s \in 1, 2, \dots,$ let

$$\mathcal{P}_s(t) = \{C_{w,1}^{s} \le w_t \le C_{w,1}^{s-1}\},$$
$$\mathrm{Reg}_s(T) = \sum_{t=1}^{T} \mathrm{reg}(t)\mathbf{1}[\mathcal{P}(t)].$$

Let t_s be the first round in which $\mathcal{P}_s(t)$ holds. Then, for each round $t = t_s+1, t_s+2, \dots,$ we have the following:

1. Eq. (15) implies that, with probability at least $C_{\mathrm{shrink}} C_{w,1}^{s}$, $\mathcal{E}(t)$ occurs. Furthermore, once $\mathcal{E}(t)$ occurs, $\mathcal{P}_s(t')$ never occurs again for round $t' > t$.
2. Eq. (16) implies that the expected regret per round is at most $C_{\mathrm{regt}}(C_{w,1}^{s-1})^2$.

Accordingly, it follows that

$$\begin{aligned}
&\mathbb{E}[\mathrm{Reg}_s(T)] \\
&\le C_{\mathrm{regt}}(C_{w,1}^{s-1})^2 \left(1 + (1 - C_{\mathrm{shrink}} C_{w,1}^{s}) + (1 - C_{\mathrm{shrink}} C_{w,1}^{s})^2 + (1 - C_{\mathrm{shrink}} C_{w,1}^{s})^3 + \dots\right) \\
&= \frac{C_{\mathrm{regt}}(C_{w,1}^{s-1})^2}{C_{\mathrm{shrink}} C_{w,1}^{s}} \\
&= \frac{C_{\mathrm{regt}}}{C_{\mathrm{shrink}} C_{w,1}^{2}} C_{w,1}^{s}. \qquad (17)
\end{aligned}$$

The regret is bounded as

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(T)] &= \sum_{s=1}^{\infty} \mathbb{E}[\mathrm{Reg}_s(T)] \\
&= \frac{C_{\mathrm{regt}}}{C_{\mathrm{shrink}} C_{w,1}^{2}} \sum_{s=1}^{\infty} C_{w,1}^{s} \quad \text{(by Eq. (17))} \\
&= \frac{C_{\mathrm{regt}}}{C_{\mathrm{shrink}} C_{w,1}^{2}} \frac{1}{1 - C_{w,1}},
\end{aligned}$$

which is a constant. □

A.7 Proof of Lemma 6

Proof of Lemma 6. We have

$$\mathbb{P}[a_t = 0] = \mathbb{P}\left[|x_t m_t + z_t| \le \epsilon_t\right]$$

$$
\begin{aligned}
&= \mathbb{P}\left[\int_{-\epsilon_t/(1+m_t^2)}^{\epsilon_t/(1+m_t^2)} \phi(x)dx\right] \quad (\text{by } x_t m_t + z_t \sim \mathcal{N}(0, 1+m_t^2) \text{ given } m_t) \\
&= \Theta(\epsilon_t) \quad (\text{by } 1 \leq (1+m_t^2) \leq 2 \text{ and } \phi(x) \leq 1),
\end{aligned}
$$

which completes the proof. □

A.8 Proof of Lemma 7

We first introduce the following lemmas.

Lemma 10 (Gap between Z_t and $-x_t m_t$: Ternary Case)**.** There exist universal constants $C_l > 0$ such that the following inequalities hold.

1. If $\operatorname{sgn}(x_t m_t) a_t < 0$, then

$$
C_l \min(1, 1/|x_t|) < a_t(Z_t + x_t m_t). \tag{18}
$$

2. If $\operatorname{sgn}(x_t m_t) a_t > 0$, then

$$
C_l < a_t(Z_t + x_t m_t). \tag{19}
$$

Lemma 10 is a version of Lemma 9 for the ternary recommendation. We omit the proof of Lemma 10 because it follows the same steps as Lemma 9.

Lemma 11 (Expected Regret from Choosing the Inferior Arm)**.** The following inequality holds:

$$
\mathbb{E}[\text{reg}(t)\mathbf{1}[b_t^* \neq b_t, a_t \neq 0]] = O(w_t^2).
$$

Proof of Lemma 11. We have

$$
\{b_t^* \neq b_t, a_t \neq 0\} \subseteq \{b_t^* \neq a_t, a_t \neq 0\} \cup \{a_t \neq b_t, a_t \neq 0\},
$$

and we bound each of the terms in the right hand side.

Claim 8.a. $\mathbb{E}[\text{reg}(t)\mathbf{1}[b_t^* \neq a_t, a_t \neq 0]] = O(w_t^2)$.

Proof.

$$
\begin{aligned}
\{b_t^* \neq a_t, a_t \neq 0\} \subseteq \{b_t^* \neq a_t\} & \\
&= \{\operatorname{sgn}(x_t\theta + z_t) \neq \operatorname{sgn}(x_t m_t + z_t)\} \\
&= \{z_t \in [\min(-x_t\theta, -x_t m_t), \max(-x_t\theta, -x_t m_t)]\},
\end{aligned}
$$

and thus, conditioning on x_t, we have

$$\begin{aligned}
&\mathbb{P}[z_t \in [\min(-x_t\theta, -x_t m_t), \max(-x_t\theta, -x_t m_t)]\,|x_t] \\
&\leq \int_{z=\min(-x_t\theta,-x_t m_t)}^{\max(-x_t\theta,-x_t m_t)} \phi(z)dz \\
&\leq \int_{z=\min(-x_t\theta,-x_t m_t)}^{\max(-x_t\theta,-x_t m_t)} dz = |x_t(\theta - m_t)|, \qquad (20)
\end{aligned}$$

where we have used the fact that $\phi(z) \leq 1$. The event $b_t^* \neq a_t$ implies $\mathrm{reg}(t) \leq x_t w_t$, and marginalizing Eq. (20) over x_t, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{reg}(t)\mathbf{1}[b_t^* \neq a_t, a_t \neq 0]] &\leq \int_{x=-\infty}^{\infty} \phi(x)|x^2 w_t(\theta - m_t)|dx \\
&\leq \int_{x=-\infty}^{\infty} \phi(x)x^2 w_t^2 dx \\
&= w_t^2 \int_{x=-\infty}^{\infty} \phi(x)x^2 dx = O\left(w_t^2\right),
\end{aligned}$$

as desired. □

Claim 8.b. $\mathbb{P}[a_t \neq b_t, a_t \neq 0] = O(w_t^2)$.

Proof.

$$\begin{aligned}
&\{a_t \neq b_t, a_t \neq 0\} \\
&= \{\mathrm{sgn}(x_t m_t + z_t) \neq \mathrm{sgn}(x_t\theta + Z_t), a_t \neq 0\} \\
&\subseteq \{x_t m_t + z_t > 0, x_t\theta + Z_t < 0\} \cup \{x_t m_t + z_t < 0, x_t\theta + Z_t > 0\} \\
&\subseteq \{x_t\theta - x_t m_t + C_1 \min(1, 1/|x_t|) < 0\} \cup \{x_t\theta - x_t m_t - C_1 \min(1, 1/|x_t|) > 0\} \\
&\qquad \text{(by Eq. (18) and Eq. (19))} \\
&= \{|x_t\theta - x_t m_t| \geq C_1 \min(1, 1/|x_t|)\},
\end{aligned}$$

and thus

$$\begin{aligned}
\mathbb{P}[a_t \neq b_t, a_t \neq 0] &\leq \mathbb{P}[|x_t\theta - x_t m_t| \geq C_1 \min(1, 1/|x_t|)] \\
&= \mathbb{P}\left[|x_t|^2 \geq \frac{C_1}{|\theta - m_t|}\right] \\
&\leq \mathbb{P}\left[|x_t|^2 \geq \frac{C_1}{w_t}\right]
\end{aligned}$$

$$
\begin{aligned}
&= 2\Phi^c\left(\sqrt{\frac{C_1}{w_t}}\right) \\
&\leq e^{-\frac{w_t}{2C_1}} \\
&= O(w_t^2). \quad \text{(An exponential decays faster than any polynomial)}
\end{aligned}
$$

□

(Proof of Lemma 11, continued.) Combining Claims 8.a and 8.b, we have

$$
\begin{aligned}
\mathbb{E}[\mathrm{reg}(t)\mathbf{1}[b_t^* \neq b_t, a_t \neq 0]] &\leq \mathbb{E}[\mathrm{reg}(t)\mathbf{1}[b_t^* \neq a_t, a_t \neq 0]] + \mathbb{E}[\mathrm{reg}(t)\mathbf{1}[a_t \neq b_t, a_t \neq 0]] \\
&= O(w_t^2).
\end{aligned}
$$

□

Proof of Lemma 7.

$$
\begin{aligned}
\mathbb{E}[\mathrm{reg}(t)] &\leq \mathbb{E}[\mathbf{1}[a_t = 0]\mathrm{reg}(t)] + \mathbb{E}[\mathbf{1}[b_t \neq b_t^*, a_t \neq 0]\mathrm{reg}(t)] \\
&\leq \mathbb{E}[\mathbf{1}[a_t = 0]|x_t\theta + z_t|] + \mathbb{E}[\mathbf{1}[b_t \neq b_t^*, a_t \neq 0]\mathrm{reg}(t)] \\
&\leq \mathbb{P}[a_t = 0](\epsilon_t + w_t) + \mathbb{E}[\mathbf{1}[b_t \neq b_t^*, a_t \neq 0]\mathrm{reg}(t)] \\
&\quad \text{(by } a_t = 0 \text{ implies } |x_t m_t + z_t| \leq \epsilon_t \text{ and } |x_t\theta + z_t| - |x_t m_t + z_t| \leq |x_t w_t|) \\
&\leq O((\epsilon_t + w_t)^2) + \mathbb{E}[\mathbf{1}[b_t \neq b_t^*, a_t \neq 0]\mathrm{reg}(t)] \quad \text{(by Lemma 6)} \\
&\leq O((\epsilon_t + w_t)^2) + O(w_t^2) \quad \text{(by Lemma 11)} \\
&= O((\max(\epsilon_t, w_t))^2).
\end{aligned}
$$

□

A.9 Proof of Lemma 8

Proof of Lemma 8. Let

$$\mathcal{X}(t) := \{x_t \geq 2\}.$$

Claim 9.a. $\mathcal{X}(t)$ and $a_t = 0$ implies $\{w_{t+1} \leq (7/8)w_t\}$.

Proof. Eq. (2) and (3) imply that $l_{t+1} = \max(l_t, -Z_t/x_t)$ or $u_{t+1} = \min(u_t, -Z_t/x_t)$ always holds. By using this, we have

$$
\begin{aligned}
&\{\mathcal{X}(t), a_t = 0\} \\
&:= \{x_t \geq 2, a_t = 0\}
\end{aligned}
$$

$$
\begin{aligned}
&= \{x_t \geq 2, |x_t m_t + Z_t| \leq \epsilon_t\} \\
&\subseteq \{|m_t + Z_t/x_t| \leq \epsilon_t/2\} \\
&= \{|m_t + Z_t/x_t| \leq \epsilon_t/2\} \cap \{l_{t+1} = \max(l_t, -Z_t/x_t) \cup u_{t+1} = \min(u_t, -Z_t/x_t)\} \\
&\qquad \text{(by Eq. (2) and (3))} \\
&\subseteq \{l_{t+1} \geq m_t - \epsilon_t/2 \cup u_{t+1} \leq m_t + \epsilon_t/2\} \\
&= \{l_{t+1} \geq m_t - (3/8)w_t \cup u_{t+1} \leq m_t + (3/8)w_t\}.
\end{aligned}
$$

Moreover, by $w_t/2 = u_t - m_t = m_t - l_t$, we have

$$\{l_{t+1} \geq m_t - (3/8)w_t \cup u_{t+1} \leq m_t + (3/8)w_t\} \subseteq \{w_{t+1} \leq (7/8)w_t\}.$$

□

Claim 9.b. $\Pr[\mathcal{X}(t)] = \Theta(1)$.

Proof.

$$\Pr[\mathcal{X}(t)] = \Phi^c(2) \geq \frac{3}{4}\frac{e^{-2}}{\sqrt{2\pi}},$$

where the last transformation uses the results in Feller (1968). □

Combining Claims 9.a and 9.b, we have

$$
\begin{aligned}
\Pr[w_{t+1} \leq (7/8)w_t | a_t = 0] &= \Pr[w_{t+1} \leq (7/8)w_t, a_t = 0 | a_t = 0] \\
&\leq \Pr[\mathcal{X}(t) | a_t = 0] \quad \text{(by Claim 9.a)} \\
&= \Pr[\mathcal{X}(t)] \quad \text{(by $\mathcal{X}(t)$ and a_t are independent)} \\
&\geq \frac{3}{4}\frac{e^{-2}}{\sqrt{2\pi}}. \quad \text{(by Claim 9.b)}
\end{aligned}
$$

□