Advantage Amplification in Slowly Evolving Latent-State Environments

Martin Mladenov[1] , **Ofer Meshi**[1] , **Jayden Ooi**[1] , **Dale Schuurmans**[1,2] , **Craig Boutilier**[1]
[1]Google AI, Mountain View, CA, USA
[2]Department of Computer Science, University of Alberta, Edmonton, AB, Canada
{mmldanenov,meshi,jayden,schuurmans,cboutilier}@google.com

Abstract

Latent-state environments with long horizons, such as those faced by recommender systems, pose significant challenges for reinforcement learning (RL). In this work, we identify and analyze several key hurdles for RL in such environments, including belief state error and small action advantage. We develop a general principle called *advantage amplification* that can overcome these hurdles through the use of temporal abstraction. We propose several aggregation methods and prove they induce amplification in certain settings. We also bound the loss in optimality incurred by our methods in environments where latent state evolves slowly and demonstrate their performance empirically in a stylized user-modeling task.

1 Introduction

Long-term value (LTV) estimation and optimization is of increasing importance in the design of *recommender systems (RSs)*, and other user-facing systems. Often the problem is framed as a *Markov decision process (MDP)* and solved using MDP algorithms or *reinforcement learning (RL)* [Shani *et al.*, 2005; Taghipour *et al.*, 2007; Choi *et al.*, 2018; Archak *et al.*, 2012; Mladenov *et al.*, 2017; Zhao *et al.*, 2018; Gauci *et al.*, 2018; Ie *et al.*, 2019]. Typically, the action set is the set of recommendable items;[1] states reflect information about the user (e.g., static attributes, past interactions, context/query); and rewards measure some form of user engagement (e.g., clicks, views, time spent, purchase). Such *event-level models* have seen some success, but current state-of-the-art is limited to very short horizons.

When dealing with long-term user behavior, it is vital to consider the impact of recommendations on user *latent state* (e.g., satisfaction, latent interests, or item awareness) which often governs both immediate and long-term behavior. Indeed, the main promise of using RL/MDP models for RSs is to: (a) identify user latent state (e.g., uncover user interests in new topics via exploration, or estimate user satisfaction); and (b) influence the latent state (e.g., create new interests, improve awareness or increase satisfaction). That said, evidence is emerging that at least some aspects of user latent state *evolve very slowly*. For example, Hohnhold et al. [2015] show that varying ad quality and ad load induces slow, but inexorable (positive or negative) changes in user click propensity over a period of months, while Wilhelm et al. [2018] show that explicitly diversifying recommendations in YouTube induces similarly slow, persistent changes in user engagement (see such slow "user learning" curves in Fig. 1).

Event-level RL in such settings is challenging for several reasons. First, the effective horizon over which an RS policy influences the latent state can extend up to $O(10^4$–$10^5)$ state transitions. Indeed, the cumulative effect of recommendations is vital for LTV optimization, but the long-term impact of any *single* recommendation is often dwarfed by immediate reward differences. Second, the MDP is partially observable, requiring some form of belief state estimation. Third, the impact of latent state on immediate observable behavior is often small and very noisy—the problems have a low *signal-to-noise ratio (SNR)*. We detail below how these factors interact.

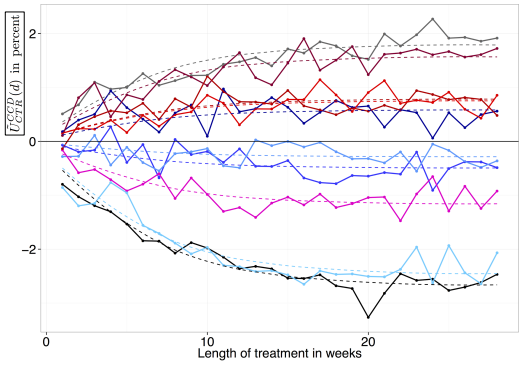
Given the importance of LTV optimization in RSs, we propose a new technique called *advantage amplification* to overcome these challenges. Intuitively, advantage amplification seeks to overcome the error induced by state estimation by introducing (explicit or implicit) *temporal abstraction* across policy space. We require that policies "commit" to taking a series of actions, thus allowing more accurate value estimation by mitigating the cumulative effects of state-estimation error.

We first consider *temporal aggregation*, where an action is held fixed for a short horizon. We show that this can lead to significant amplification of the advantage differences between abstract actions (relative to event-level actions). This is a form of MDP/RL temporal abstraction as used in *hierarchical RL* [Sutton *et al.*, 1999; Barto and Mahadevan, 2003] and can be viewed as options designed to allow distinction of good and bad behaviors in latent-state domains with low SNR (rather than, say, for subgoal achievement). We generalize this by analyzing policies with (artificial) action *switching costs*, which induces similar amplification with more flexibility.

Limiting policies to temporally abstract actions induces potential sub-optimality [Parr, 1998; Hauskrecht *et al.*, 1998]. However, since the underlying latent state often evolves slowly w.r.t. the event horizon in RS settings, we identify a "smoothness" property that is used to bound the induced error of advantage amplification.

Our contributions are as follows. We introduce a stylized model of slow user learning in RSs in Sec. 2. We formalize this

[1]Item *slates* are often recommended, but we ignore this here.



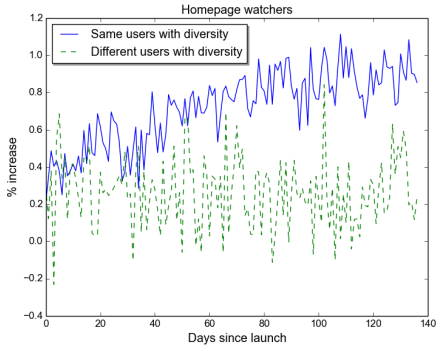


Figure 1: Gradual user response: (a) ad load/quality [Hohnhold *et al.*, 2015]; (b) YouTube recommendation diversity [Wilhelm *et al.*, 2018].

model as a POMDP in Sec. 3 and define several properties that we exploit in our analysis. In Sec. 4.1, we show that low SNR in the POMDP interacts poorly with belief-state approximation, and develop advantage amplification as a principle. We prove that *action aggregation* (Sec. 4.2) and *switching cost regularization* (Sec. 4.3) provide strong amplification guarantees with minimal policy loss under suitable conditions and suggest potential extensions in Sec. 4.4. Experiments with stylized models in Sec. 4.5 show the effectiveness of our methods.[2]

2 User Satisfaction: An Illustrative Example

Before formalizing our problem, we describe a stylized model reflecting the dynamics of *user satisfaction* as a user interacts with an RS. The model is intentionally stylized to help illustrate key concepts underlying the formal model and analysis developed in the sequel. While it ignores much of the true complexity of user satisfaction and RS interaction, its core elements permeate many recommendation domains. Finally, though focused on user-RS engagement, the principles apply more broadly to any latent-state system with low SNR and slowly evolving latent state.

Our model captures the relationship between a user and an RS over an extended period (e.g., a content recommender of news, video, or music) through *overall user satisfaction*, which is not known to the RS. We hypothesize that satisfaction is one (of several) key latent factors that impacts user engagement; and since new treatments often induce slow-moving or delayed effects on user behavior, we assume this latent variable evolves slowly as a function of the quality of the content consumed [Hohnhold *et al.*, 2015] (and see Fig. 1 (left)). Finally, the model captures the tension between (often low-quality) content that encourages short-term engagement (e.g., manipulative, provocative or distracting content) at the expense of long-term engagement; and high-quality content that promotes long-term usage but can sacrifice near-term engagement.

Our model includes two classes of recommendable items. Some items induce high immediate engagement, but degrade user engagement over the long run. We dub these "Chocolate" (*Choc*)—immediately appealing but not very "nutritious." Other items—dubbed "Kale," less attractive, but more "nutritious"—induce lower immediate engagement but tend to improve long-term engagement.[3] We call this the *Choc-Kale model* (CK). A stationary, stochastic policy can be represented by a single scalar $0 \leq \pi \leq 1$ representing the probability of taking action *Choc*. We sometimes refer to *Choc* as a "negative" and *Kale* as a "positive" recommendation.

We use a single latent variable $s \in [0, 1]$ to capture a user's overall satisfaction with the RS. Satisfaction is driven by *net positive exposure* p, which measures a user's total (discounted) accrued positive and negative recommendations, with a discount $0 \leq \beta < 1$ applied to ensure that p is bounded: $p \in \left[\frac{-1}{1-\beta}, \frac{1}{1-\beta}\right]$. (The dynamics of p is detailed below.) We view p as a user's learned perception of the RS and s as how this influences gradual changes in engagement.

A user response to a recommendation a is given by her *degree of engagement* $g(s, a)$, and depends stochastically on both the quality of the recommendation, and her latent state s. Engagement g is a random function; e.g., responses might be normally distributed: $g(s, a) \sim N(s \cdot \mu_a, \sigma_a^2)$ for $a \in \{Choc, Kale\}$. We abuse notation and sometimes let $g(s, a)$ denote *expected* engagement. We require that *Choc* results in greater immediate (expected) engagement than *Kale*, i.e., $g(s, Kale) < g(s, Choc)$, for any fixed s.

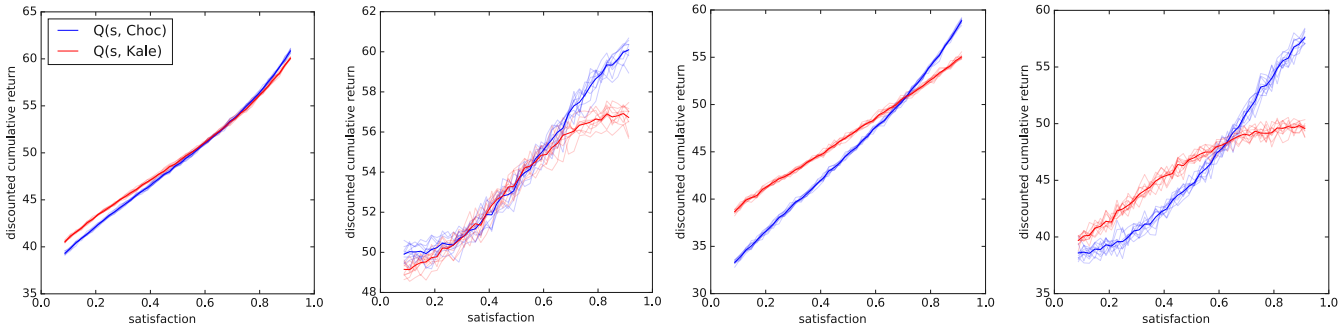
The dynamics of p is straightforward. A *Kale* exposure increases p by 1 and *Choc* decreases it by 1, with discounting: $p_{t+1} \leftarrow \beta p_t + 1$ with a *Kale* recommendation and $p_{t+1} \leftarrow \beta p_t - 1$ with *Choc*. Satisfaction s is a user-learned function of p and follows a sigmoidal learning curve: $s(p) = 1/(1 + e^{-\tau p})$, where τ is a temperature/learning rate parameter. Other learning curves are possible, but the sigmoidal model captures both positive and negative exponential learning as hypothesized in psychological-learning literature [Thurstone, 1919; Jaber, 2006] and as observed in the empirical curves in Fig. 1.[4]

To illustrate the workings of this model, we compute the Q-values of *Choc* and *Kale* for each satisfaction level s and

[2] Proofs, auxiliary lemmas and additional experiments are available in an extended version of the paper [Mladenov *et al.*, 2019].

[3] Our model allows a real-valued continuum of items (e.g., degree of Choc between $[0, 1]$ as in our experiments) like measures of ad quality. We use the binary form to streamline our initial exposition.

[4] Such learning curves are often reflective of aggregate behavior, obscuring individual differences that are much less "smooth." However, unless cues are available that allow us to model such individual differences, the aggregate model is our best resource, even when optimizing for individual users. Indeed, prediction of user responses in RSs usually relies on fairly coarse user features—and not user identity—and already relies on similar aggregate behavior.



(a) event-level, fully observable (b) event-level, noisy (c) aggregated, fully observable (d) aggregated, noisy

Figure 2: Q-values as a function of satisfaction level in the Choc-Kale model. Model parameters: $\beta = 0.9, \tau = 0.25, \mu_{\text{Choc}} = 8, \mu_{\text{Kale}} = 2$, $\sigma_{\text{Choc}} = \sigma_{\text{Kale}} = 0.5$. Observations (satisfaction s) are corrupted with additive Gaussian noise (mean 0, stddev. 0.3) truncated on $[-1, 1]$.

plot them in Fig. 2a (see the figure caption for model parameters). We observe that when satisfaction is low, *Kale* is a better recommendation, and above some level *Choc* becomes preferable, as expected. We also see that for any s the difference in Q-values is rather small. In complex domains, any required function approximation will limit our ability to accurately model the Q-function. With additional noise, the Q-values become practically indistinguishable for a large range of satisfaction levels, as shown in Fig. 2b). This illustrates the hardness of RL in this setting.

3 Problem Statement

We outline a basic latent-state control problem as a *partially observable MDP (POMDP)* that encompasses the notions above. We highlight several properties that play a key role in the analysis of latent-state RL we develop in the next section.

We consider environments that can be modeled as a POMDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T, \mathcal{Z}, O, R, \mathfrak{b}_0, \gamma \rangle$ [Smallwood and Sondik, 1973]. States $\mathcal{S}$ reflect user latent state and other observable aspects of the domain. In the stylized CK model, the state is simply a user's current satisfaction s, but more faithful RS models will use much richer state representations. Actions $\mathcal{A}$ are recommendable items: in CK, we distinguish only *Choc* from *Kale*. The transition kernel $T(s, a, s')$ captures state dynamics: in the CK model we use a simple deterministic model, $T(s', a, s) = 1$ if $s' = (1 + \exp(\beta \log(1 - 1/s) - \beta\tau a))^{-1}$, where a is 1 (resp., -1) for action *Kale* (resp., *Choc*).[5] Observations $\mathcal{Z}$ reflect observable user behavior and $O(s, a, z)$ the probability of $z \in \mathcal{Z}$ when a is taken at state s. In CK, $\mathcal{Z}$ is the observed engagement with a recommendation while O reflects the random realization of $g(s, a)$. The immediate reward $R(s, a)$ is (expected) user engagement (we let $r_{\max} = \max_{s,a} R(s, a)$), $\mathfrak{b}_0$ the initial state distribution, and $\gamma \in [0, 1)$ is a discount factor.

In this POMDP, an RS does not have access to the true state s, but must generate policies that depend only on the sequence of past action-observation pairs—let $\mathcal{H}^*$ be the set of all finite such sequences $(a_t, z_t)_{t \in \mathbb{N}}$. Any such *history* can be summarized, via optimal Bayes filtering, as a distribution or *belief state* $\mathfrak{b} \in \Delta(\mathcal{S})$. More generally, this "belief state" can be *any summarization* of $\mathcal{H}^*$ used to make decisions. It may be, say, a collection of sufficient statistics, or a deep recurrent embedding of history. We assume some *belief state representation* $(\mathcal{B}, U)$, where $\mathcal{B}$ is the set of (realizable) belief states, and the mapping $U : \mathcal{B} \times \mathcal{A} \times \mathcal{Z} \to \mathcal{B}$ describes the update $U(\mathfrak{b}, a, z)$ of any $\mathfrak{b} \in \mathcal{B}$ given $a \in \mathcal{A}, z \in \mathcal{Z}$.

A *(stochastic) policy* is a mapping $\pi : \mathcal{B} \to \Delta(\mathcal{A})$ that selects an action distribution $\pi(\mathfrak{b})$ for execution given belief $\mathfrak{b}$; we write $\pi(a|\mathfrak{b})$ to indicate the probability of action a. Deterministic policies are defined in the usual way. The *value* of a policy π is given by the standard recurrence:[6]

$$V^\pi(\mathfrak{b}) = \mathop{\mathbb{E}}_{a \sim \pi(\mathfrak{b})} \left[R(\mathfrak{b}, a) + \gamma \sum_{z \in \mathcal{Z}} \Pr(z|\mathfrak{b}, a) V^\pi(U(\mathfrak{b}, a, z)) \right] \quad (1)$$

We define $Q^\pi(\mathfrak{b}, a)$ by fixing a in Eq. (1) (rather than taking the action distribution $\pi(\mathfrak{b})$). An *optimal policy* $\pi^* = \sup V^\pi$ over $\mathcal{B}$ has value (resp., Q) function V^* (resp., Q^*). Optimal policies and values can be computed using dynamic programming or learned using (partially observable) RL methods. When we learn a Q-function Q, whether exactly or approximately, the *policy induced by* Q is the greedy policy $\pi(\mathfrak{b}) = \arg\max_a Q(\mathfrak{b}, a)$ and its *induced value function* is $V(\mathfrak{b}) = \max_a Q(\mathfrak{b}, a) = Q(\mathfrak{b}, a^*(\mathfrak{b}))$. The *advantage function* $A(a, \mathfrak{b}) = V^*(\mathfrak{b}) - Q^*(\mathfrak{b}, a)$ reflects the difference in the expected value of taking action a at $\mathfrak{b}$ (and then acting optimally) vs. acting optimally at $\mathfrak{b}$ [Baird III, 1999]. If a_2 is the second-best action at $\mathfrak{b}$, the *advantage* at that belief state is $A(\mathfrak{b}) = V^*(\mathfrak{b}) - Q^*(\mathfrak{b}, a_2)$.

Eq. (1) assumes optimal Bayesian filtering, i.e., the representation $(\mathcal{B}, U)$ must be such that the (implicit) expectations over R and O are exact for any history that maps to $\mathfrak{b}$. Unfortunately, exact recursive state estimation is intractable, except for special cases (e.g., linear-Gaussian control). As a consequence, *approximation schemes* are used in practice (e.g., variational projections [Boyen and Koller, 1998]; fixed-length histories, incl. treating observations as state [Singh *et al.*, 1994]; learned PSRs [Littman and Sutton, 2002]; recursive policy/Q-function representations [Downey *et al.*, 2017]). Approximate histories render the process non-Markovian; as such, a counterfactually estimated Q-value of a policy (e.g., using offline data) differs from its *true* value due to modified latent-state dynamics (not

[5] This is easily randomized if desired.

[6] Here $R(\mathfrak{b}, a)$ and $\Pr(z|\mathfrak{b}, a)$ are given by expectations of R and O, respectively, w.r.t. $s \sim \mathfrak{b}$ if $\mathfrak{b} \in \Delta(\mathcal{S})$. The interpretation for other representations is discussed below.

reflected in the data). In this case, any RL method that treats $\mathfrak{b}$ as (Markovian) state induces a suboptimal policy. We can bound the induced suboptimality using *ε-sufficient statistics* [Francois-Lavet *et al.*, 2017]. A function $\phi : \mathcal{H}^* \rightarrow \mathcal{B}$ is an ε-sufficient statistic if, for all $H_t \in \mathcal{H}^*$,

$$|p(s_{t+1}|H_t) - p(s_{t+1}|\phi(H_t))|_{\text{TV}} < \varepsilon \ ,$$

where $|\cdot|_{\text{TV}}$ is the total variation distance. If ϕ is ε-sufficient, then any MDP/RL algorithm that constructs an "optimal" value function $\hat{V}$ over $\mathcal{B}$ incurs a bounded loss w.r.t. V^* [Francois-Lavet *et al.*, 2017]:

$$\left|V^*(\phi(H)) - \hat{V}(\phi(H))\right| \leq \frac{2\varepsilon r_{\max}}{(1-\gamma)^3}. \tag{2}$$

The errors in Q-value estimation induced by limitations of $\mathcal{B}$ are irresolvable (i.e., they are a form of model *bias*), in contrast to error induced by limited data. Moreover, any RL method relying only on offline data is subject to the above bound, regardless of whether the Q-values are estimated directly or not. The impact of this error on model performance can be related to certain properties of the underlying domain as we outline below. A useful quantity for this purpose is the *signal-to-noise ratio (SNR)* of a POMDP, defined as:

$$\mathfrak{S} \triangleq \frac{\sup_{\mathfrak{b}} A(\mathfrak{b})}{\sup\{A(\mathfrak{b}) : A(\mathfrak{b}) \leq 2\varepsilon r_{\max}/(1-\gamma)^2\}} - 1,$$

(the denominator is treated as 0 if no $\mathfrak{b}$ meets the condition).

As discussed above, many aspects of user latent state, such as satisfaction, evolve slowly. We say a POMDP is *L-smooth* if, for all $\mathfrak{b}, \mathfrak{b}' \in \mathcal{B}$, and $a \in A$ s.t. $T(\mathfrak{b}', a, \mathfrak{b}) > 0$, we have

$$|Q^*(\mathfrak{b}, a) - Q^*(\mathfrak{b}', a)| \leq L.$$

Smoothness ensures that for any state reachable under an action a, the optimal Q-value of a does not change much.

4 Advantage Amplification

We detail how low SNR causes difficulty for RL in POMDPs, especially with long horizons (Sec. 4.1). We introduce the principle of *advantage amplification* to address it (Sec. 4.2) and analyze two realizations of this principle, *temporal aggregation* in Sec. 4.2 and *switching cost* in Sec. 4.3. We suggest several extensions in Sec. 4.4 and conclude with an empirical illustration of our mechanisms in Sec. 4.5.

4.1 The Impact of Low SNR on RL

The bound given by Eq. (2) can help assess the impact of low SNR on RL. Assume that policies, values and/or Q-functions are learned using an approximate belief representation $(\mathcal{B}, U)$ that is ε-sufficient. We first show that the error induced by $(\mathcal{B}, U)$ is tightly coupled to optimal action advantages.

Consider an RL agent that learns Q-values using a behavior (data-generating) policy ρ. The non-Markovian nature of $(\mathcal{B}, U)$ means that: (a) the resulting estimated-optimal policy π will have *estimated* values $\hat{Q}^\pi$ that differ from its *true* values Q^π; and (b) the estimates $\hat{Q}^\pi$ (hence, the choice of π itself) will depend on ρ. We bound the loss of π w.r.t. the optimal π^*—here π^* assumes exact filtering—as follows. First, for any (belief) state-action pair $(\mathfrak{b}, a)$, let the maximum difference between its inferred and optimal Q-values is bounded for any ρ: $|Q^*(\mathfrak{b}, a) - Q^\pi(\mathfrak{b}, a)| \leq \delta$. Using Eq. (2), we set

$$\delta = \frac{\varepsilon Q_{\max}}{1-\gamma} \leq \frac{\varepsilon r_{\max}}{(1-\gamma)^2}. \tag{3}$$

If $A(\mathfrak{b}) \leq 2\delta$ (i.e., $\mathfrak{b}$ has small advantage), under behavior policy ρ, the estimate $\hat{Q}(\mathfrak{b}, a_2)$ (the second-best action) can exceed that of $\hat{Q}(\mathfrak{b}, a^*(\mathfrak{b}))$, in which case π executes a_2. If π visits $\mathfrak{b}$ (or states with similarly small advantages) at a constant rate, the loss w.r.t. π^* compounds, inducing $O(\frac{2\delta}{1-\gamma})$ error.

The tightness of the second part of the argument depends on the structure of the advantage function $A(\mathfrak{b})$. To illustrate, consider two extreme regimes. First, if $A(\mathfrak{b}) \geq 2\delta$ at all $\mathfrak{b} \in \mathcal{B}$, i.e., if SNR $\mathfrak{S} = \infty$, state estimation error has no impact on the recovered policy and incurs no loss. In the second regime, if all $A(\mathfrak{b})$ are less than (but on the order of) 2δ, i.e., if $\mathfrak{S} = 0$, then the inequality is tight provided ρ saturates the state-action error bound. We will see below that low-SNR environments with long horizons (e.g., practical RSs, or our stylized CK model) often have such small (but non-trivial) advantages across a wide range of state space.

The latter situation is illustrated in Fig. 2. In Fig. 2a, the Q-values of the CK model are plotted against the level of satisfaction (treating it as fully observable). The small advantages are notable. Fig. 2b shows the Q-value estimates for 10 independent tabular Q-learning runs when noise is added to the estimated belief state s (the thin lines show individual runs, the thick lines show the average) . The corrupted Q-values at all but the highest satisfaction levels are essentially indistinguishable, leading to extremely poor policies.

4.2 Temporal Abstraction: Action Aggregation

There is a third regime in which state error is relatively benign. Suppose the advantage at each state $\mathfrak{b}$ is either small, $A(\mathfrak{b}) \leq \sigma$, or large, $A(\mathfrak{b}) > \Sigma$ for some constants $\sigma \ll 2\delta \leq \Sigma$. The induced policy incurs a loss of σ at small-advantage belief states, and no loss on states with large advantages. This leads to a compounded loss of at most $\frac{\sigma}{1-\gamma}$, which may be much smaller than the $\frac{\varepsilon r_{\max}}{(1-\gamma)^2}$ error in Eq. (3), depending on σ.

If the environment is smooth, *action aggregation* can be used to tramsform a problem falling in the second regime into one that falls in the third regime, with σ depending on the level of smoothness. This can significantly reduce the impact of estimation error on policy quality by turning the problem into one that is essentially Markovian. More specifically, if at state $\mathfrak{b}$, we know that the optimal (stationary) policy takes action a for the next k decision periods, we consider a reparameterization $\mathcal{M}^{\times k}$ of the belief-state MDP where, at $\mathfrak{b}$, any action taken must be executed k times in a row, no matter what the subsequent k states are. In this new problem, the Q-value of the optimal repeated action $Q^*(\mathfrak{b}, a^{\times k})$ is the same as that of its event-level counterpart $Q^*(\mathfrak{b}, a)$, since the same sequence of expected rewards will be generated. Conversely, all suboptimal actions incur a cumulative *reduction* in Q-value in $\mathcal{M}^{\times k}$ since their suboptimality *compounds* over k periods. Thus, in $\mathcal{M}^{\times k}$, the optimal policy $\pi^{\times k *}$ generates the same cumulative discounted return as the event-level optimal policy, while the

advantage of $a^{\times k}$ over any other repeated action $a'^{\times k}$ at $\mathfrak{b}$ is larger than that of a over a' in the event-level problem.

To derive bounds, note that, for an L-smooth POMDP, at any state where the advantage is at least $2kL$, the optimal action persists for the next k periods (its Q-value can decrease by at most L while that of the second-best can at most increase by L). If we apply aggregation only at such states, the advantage increases to some value Σ, putting us in regime 3 (i.e., the advantage is either less than $\sigma = 2kL$ or more than Σ). Of course, we cannot "cherry-pick" only states with high advantage for aggregation, but instead apply this action aggregation uniformly over $\mathcal{B}$. Naturally, aggregating over all states induces some loss due to the inability to switch actions quickly. We bound that cost when determining σ and Σ. to complete the analysis. This allows us to first lower bound the regret of the best k-aggregate policy:

Theorem 1. *Let k be a fixed horizon, and let Q^*—the event-level optimal Q function—be L-smooth. Then for all $\mathfrak{b}$, $|V^*(\mathfrak{b}) - V^{\times k*}(\mathfrak{b})| \leq \frac{2kL}{1-\gamma}$, where $V^{\times k*}(\mathfrak{b})$ is the value of state $\mathfrak{b}$ under an optimal k-aggregate policy.*[7]

This theorem is proved by constructing a policy which switches actions every k events and showing that it has bounded regret.[8] This policy, at the start of any k-event period, adopts the optimal action from the unaggregated MDP at the initiating state. Due to smoothness, Q-values cannot drift by more than kL during the period, after which the policy corrects itself. This, together with the reasoning above, offers an amplification guarantee:

Theorem 2. *In an L-smooth MDP, let k be a fixed repetition horizon. For any belief state where $A(\mathfrak{b}) \geq 2kL$, the k-aggregate-horizon advantage is bounded below as follows:*

$$Q^{\times k*}(\mathfrak{b}, a^{\times k}) - Q^{\times k*}(\mathfrak{b}, a'^{\times k}) \geq A(\mathfrak{b})\frac{1-\gamma^k}{1-\gamma} - 2L\frac{\gamma - (1+k-\gamma k)\gamma^{k+1}}{(1-\gamma)^2} - \frac{2kL}{1-\gamma}.$$

This result is especially useful when the event-level advantage is more than $\sigma = \frac{2kL}{1-\gamma}$. In this case, an aggregation horizon of k can mitigate the adverse effects of approximating belief state with an ε-sufficient statistic for any ε up to

$$\varepsilon_{\max} \leq L\frac{k(\gamma - \gamma^k) - \gamma(1 - (1+k-\gamma k)\gamma^k)}{r_{\max}},$$

at the cost of the aggregation loss of $\frac{2kL}{1-\gamma}$.

Figs. 2c and 2d illustrate the benefit of action aggregation: they show the Q-values of the k-aggregated CK model with $k = 5$ with both perfect and imperfect state estimation, respectively (the amount of noise is the same as in Fig. 2b). As we show in Sec. 4.5, the recovered policies incur very little loss due to state estimation error.

We conclude with the following observation.

Corollary 1. *Optimal repeating policies are near-optimal for the event-level problem as $L \to 0$ and amplification at every state is guaranteed.*

[7]The reparameterized problem is also an MDP, so the optimal value function and deterministic policy are well-defined.

[8]See the full paper [Mladenov *et al.*, 2019] for proofs of all results.

4.3 Temporal Regularization: Switching Cost

While temporal aggregation is guaranteed to improve learning in slow environments, it has certain practical drawbacks due to its inflexibility. One is that, in the non-Markovian setting induced by belief-state approximation, training data should ideally be collected using a k-aggregated behavior policy.[9] Another drawback arises if the L-smoothness assumption is partially violated. For example, if certain rare events cause large changes in state or reward for short periods, the changes in Q-values may be abrupt. Notice that such changes are harmless from an SNR perspective if they induce large advantage gaps; but an agent "committed" to a constant action during an aggregation period is unable to react to such events. We thus propose a more flexible advantage amplification mechanism, namely, a *switching-cost regularizer*. Intuitively, instead of fixing an aggregation horizon, we impose a fictitious cost (or penalty) T on the agent whenever it changes its action.

More formally, the goal in the *switching-cost (belief-state) MDP* is to find an optimal policy defined as:

$$\pi^* = \arg\max_{\pi} \sum_t \gamma^t \mathbb{E}_\pi \left(R_t - T \cdot \mathbb{1}[a_t \neq a_{t-1}]\right). \quad (4)$$

This problem is Markovian in the extended state space $\mathcal{B} \times A$ representing the *current* (belief) state and the *previously* executed action. This state space allows the switching penalty to be incorporated into the reward function as $R(\mathfrak{b}, a_{t-1}, a_t) = R(\mathfrak{b}, a_t) - T \cdot \mathbb{1}[a_t \neq a_{t-1}]$.

The switching cost induces an implicit *adaptive action aggregation*—after executing action a, the agent will keep repeating a until the cumulative advantage of switching to a different action exceeds the switching cost T. We can use this insight to bound the maximum regret of such a policy (relative to the optimal event-level policy) and also provide an amplification guarantee, as is the case with action aggregation.

In the case of problems with two actions, we can analyze the action of the switching-cost regularizer in a relatively intuitive way. As with Thm. 1, we bound the regret induced by the switching cost by constructing a policy that behaves as if it had to pay T with every action switch. In particular, the optimal policy under this penalty adopts the action of the event-level optimal policy at some state $\mathfrak{b}_t$, then holds it fixed until its expected regret for not switching to a *different* action dictated by the event-level optimal policy exceeds T. Suppose the time at which this switch occurs is $(t + \omega)$. The regret of this agent is no more than the regret of an agent with the option of paying T upfront in order to follow the event-level optimal policy for ω steps. We can show that the same regret bound holds if the agent were paying to switch to the best fixed action for ω steps instead of following the event-level optimal policy. This allows derivation of the following bound:

Theorem 3. *The regret of the optimal switching cost policy for a two-action MDP is less than $\frac{2\kappa L}{1-\gamma}$, where*

$$\kappa = \frac{\log\gamma + (\gamma-1)W\left(\frac{\gamma^{1/(1-\gamma)}}{\gamma-1}\left(\frac{(1-\gamma)^2}{2\gamma L}T - 1\right)\log\gamma\right)}{(\gamma-1)\log\gamma},$$

and where W is the Lambert W-function [Corless et al.*, 1996].*

[9]This is unnecessary if the system is Markovian, since (s, a, r, s') tuples may be reordered to emulate any behavioral policy.

This leads to an amplification result, analogous to Thm. 2:

Theorem 4. *Let κ be as in Thm. 3. In a two-action MDP, any state whose advantage in the event-level optimal policy is at least $(1 + \frac{1}{1-\gamma})2\kappa L$ has an advantage of at least $2T$ in the switching-cost regularized optimal policy.*

4.4 Discussion

Before empirically testing the two mechanisms above, we suggest several more general extensions of these proposals which we believe offer opportunities for additional research. We have demonstrated that the L-smoothness of the environment dynamics can be incorporated into an RL algorithm like Q-learning to improve the quality of the learned policy. In essence, we do this in one of two ways: we restrict the policy space to those policies that repeat the same action multiple times; or we add an explicit penalty for switching from one action to another. Both have the effect of injecting a specific bias into the learning process which makes the resulting policy less susceptible to the effects of an inaccurate (or incomplete) belief state representation.

Both the temporal (action) aggregation and switching cost mechanisms are framed in the context of deterministic policies. It is well-known that under conditions of approximate or error-prone state representation, stochastic policies may perform better than deterministic ones [Singh *et al.*, 1994]. One important generalization of our approaches is to develop regularization methods for stochastic policies. We can extend the basic form of policy regularization to stochastic policies rather easily by introducing a penalty that is a function of the difference (e.g., KL-divergence) between consecutive action distributions $\pi(s_t)$ and $\pi(s_{t+1})$. On the one hand, this is a natural generalization of temporal (action) aggregation, exploiting the continuous nature of the "softened" decision space in a way that is not generally possible in the case of deterministic policies[10]. On the other hand, this also generalizes the switching cost mechanism in a natural way. Such a mechanism is likely to result in softmax-like policies and provides an interesting avenue for future research.

A second line of generalization is broader—action aggregation can be seen as a hierarchical approach to RL, using a particular class of *macros* or *options*. Unlike most work in options, the options are constructed with the explicit aim of reducing loss due to an inaccurate belief state representation. Extending our existing mechanisms to develop general principles for constructing arbitrary macros for state-noise mitigation could be of significant value for RL for complex, partially observable environments.

4.5 Empirical Illustration

We experiment with synthetic models to demonstrate the theoretical results above. In a first experiment, we apply both action aggregation and switching cost regularization to the simple *Choc-Kale* POMDP from Sec. 2 with parameters $\beta = 0.9, \tau = 0.25, \mu_{\text{Choc}} = 8, \mu_{\text{Kale}} = 2$, and $\sigma_{\text{Choc}} = \sigma_{\text{Kale}} = 0.5$. To illustrate the effects of approximate belief-state estimation, we corrupt the satisfaction level s with zero-mean Gaussian noise truncated on $[-1, 1]$. As we increase the variance σ_N of the noise distribution (we vary it in this experiment), state-estimation accuracy decreases.

To mitigate the effect of state-estimation error, we apply temporal aggregation of 3 and 5 actions using discounts of $\gamma = 0.95$ and 0.99 (Fig. 3a,b); and enforce switching costs of 1, 2 and 3 (Fig. 3c). For each parameter setting, we train 10 policies using tabular Q-learning, discretizing state space into 50 buckets. For each training run, we train using 30000 event-level transitions, exploring using actions taken uniformly at random—aggregated actions in the aggregation setting. Once trained, we evaluate the discounted return of each policy using 100 Monte Carlo rollouts of length 1000.

Figs. 3a, b and c show the average performance across the 10 training runs (with 95% confidence intervals) as a function of σ_N. We see that action aggregation has a profound effect on solution quality, improving policy performance by nearly a factor of 2 with $\gamma = 0.99$. Switching cost regularization has a more subtle effect, providing more modest improvements in performance. We conjecture that this difference in performance is due to action aggregation having a double effect on the value estimates—apart from amplification, it also provides a more favorable behavioral policy.

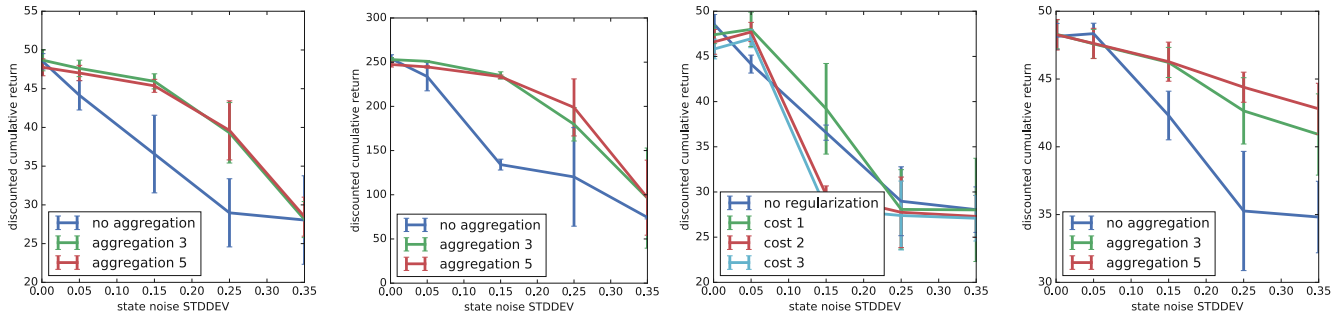
A second experiment takes an "options-oriented" perspective. Recommendable items now have a continuous *"kaleness" score* in $[0, 1]$, with item i's score denoted $v(i)$. At each time step, a set of 7 items is drawn from a $[0, 1]$-truncated Gaussian with mean equal to the kaleness score of the previously consumed item. The RL agent sets a *target kaleness score* $\theta \in \{0, 0.25, 0.5, 0.75, 1\}$ (its action space). This translates to a specific "presentation" of the 7 items to the user such that the user is nudged to consume an item whose score is closer to the target. Specifically, the user chooses an item i using a softmax distribution: $P(i) \propto \exp(-|v(i) - \theta|/\lambda)$, with temperature $\lambda = 0.2$. We compare the performance of action aggregation with 3 and 5 actions with that of unregularized policy in Fig. 3d: aggregation exhibits a comparable level of improvement as in the binary-action case.

5 Related Work

The study of time series at different scales of granularity has a long-standing history in econometrics, where the main object of interest is the behavior of various autoregressive models under aggregation [Silvestrini and Veredas, 2008]. However, the behavior of aggregated systems under control does not seem to have been investigated in that field.

In RL, time granularity arises in several contexts. Classical semi-MDP/options methods employ temporal aggregation to organize the policy space into a hierarchy, where a prespecified sub-policy, or *option*, is executed for some period of time (termination is generally part of the option specification) [Sutton *et al.*, 1999]. That options might help with partial observability ("state noise") has been suggested—e.g., Daniel et al. [2016] informally suggest that reduced control frequency can improve SNR. However, a formal characterization of this phenomenon has not been addressed to the best of our knowl-

[10]For actions in some metric space, e.g., as in many continuous control tasks, "smoother" aggregation methods are sensible even in deterministic policy space.



(a) Temporal aggr., $\gamma = 0.95$. (b) Temporal aggr., $\gamma = 0.99$. (c) Switching cost, $\gamma = 0.95$. (d) Options-oriented, $\gamma = 0.95$.

Figure 3: Experimental results. Cumulative discounted returns of event-level, aggregated and regularized policies on the Choc-Kale POMDP.

edge. The *learning to repeat* framework (see [Sharma *et al.*, 2017] and references therein) provides a modeling perspective that allows an agent to choose an action-repetition granularity as part of the action space itself, but does not study these models theoretically. SNR has played a role in RL, but in different ways than studied here, e.g., as applied to policy gradient (rather than as a property of the domain) [Roberts and Tedrake, 2009].

The effect of the advantage magnitude (also called *action gap*) on the quality and convergence of RL algorithms was first studied by Farahmand [2011]. Bellemare et al. [2016] observed that the action gap can be manipulated to improve the quality of learned polices by introducing local policy consistency constraints to the Bellman operator. Their considerations are, however, not bound to specific environment properties.

Our framework is closely related with the study of regularization in RL and its benefits when dealing with POMDPs. Typically, an entropy-based penalty (or KL-divergence w.r.t. to a behavioral policy) is added to the reward to induce a stochastic policy. This is usually justified in one of several ways: inducing exploration [Nachum *et al.*, 2017]; accelerating optimization by making improvements monotone [Schulman *et al.*, 2015]; and smoothing the Bellman equation and improving sample efficiency [Fox *et al.*, 2016]. Of special relevance is the work of Thodoroff et al. [2018], who, akin to this work, exploit the sequential dependence of Q-values for better Q-value estimation. In all of this work, regularization is typically treated as a price to be paid to achieve an auxiliary goal (e.g., better optimization or improved statistical efficiency). While stochastic policies often perform better than deterministic ones when state estimation is approximate or error-prone [Singh *et al.*, 1994]—indeed, methods that exploit this have been proposed in restricting settings (e.g., corrupted rewards [Everitt *et al.*, 2017])—the connection to regularization has not been made explicitly to the best of our knowledge. An alternative way of dealing with limited state representations is to directly optimize the policy return via policy gradient.

Finally, LTV optimization for user satisfaction is often discussed in the *safe RL* literature (see, e.g., [Theocharous *et al.*, 2014; Theocharous *et al.*, 2015]). Typically, the aim is to guarantee a certain level of performance, or improvement relative to a behavioral policy, before deploying a policy learned in an off-line fashion. Our work is somewhat orthogonal (and compatible) with these approaches. Work on the general application of RL to RSs (see Sec. 1) is relevant as well.

6 Concluding Remarks

We have developed a framework for studying the impact of belief-state approximation in latent-state RL problems, especially suited to slowly evolving, highly noisy (low SNR) domains like recommender systems. We introduced *advantage amplification* and proposed and analyzed two conceptually simple realizations of it. Empirical study on a stylized domain demonstrated the tradeoffs and gains they might offer.

There are a variety of interesting avenues suggested by this work (see Sec. 4.4 for further discussion): (i) the study of soft-policy regularization for amplification; (ii) developing techniques for constructing more general "options" (beyond aggregation) for amplification; (iii) developing amplification methods for arbitrary sources of modeling error; (iv) conducting more extensive empirical analysis on real-world domains.

Acknowledgments. Thanks to the reviewers for helpful feedback.

References

[Archak *et al.*, 2012] N. Archak, V. Mirrokni, and S. Muthukrishnan. Budget optimization for online campaigns with positive carryover effects. *WINE-12*, pp.86–99, Liverpool, 2012.

[Baird III, 1999] L. C. Baird III. *Reinforcement Learning Through Gradient Descent*. PhD thesis, US Air Force Academy, 1999.

[Barto and Mahadevan, 2003] A. G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13(1-2):41–77, 2003.

[Bellemare *et al.*, 2016] M. G. Bellemare, G. Ostrovski, A. Guez, P. S. Thomas, and R. Munos. Increasing the action gap: New operators for reinforcement learning. *AAAI-16*, pp.1476–1483, 2016.

[Boyen and Koller, 1998] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. *UAI-98*, pp.33–42, 1998.

[Choi *et al.*, 2018] S. Choi, H. Ha, U. Hwang, C. Kim, J. Ha, S. Yoon. Reinforcement learning-based recommender system using biclustering technique. *arXiv:1801.05532*, 2018.

[Corless *et al.*, 1996] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the Lambert W function. *Adv. in Comp. Math.*, 5(1):329–359, 1996.

[Daniel *et al.*, 2016] C. Daniel, G. Neumann, O. Kroemer, and J. Peters. Hierarchical relative entropy policy search. *JMLR*, 17(93):1–50, 2016.

[Downey *et al.*, 2017] C. Downey, A. Hefny, B. Boots, G. J. Gordon, and B. Li. Predictive state recurrent neural networks. *NIPS-17*, pp.6053–6064. 2017.

[Everitt *et al.*, 2017] T. Everitt, V. Krakovna, L. Orseau, and S. Legg. Reinforcement learning with a corrupted reward channel. *IJCAI-17*, pp.4705–4713, Melbourne, 2017.

[Farahmand, 2011] A. Farahmand. Action-gap phenomenon in reinforcement learning. *NIPS-11*, pp.172–180.

[Fox *et al.*, 2016] R. Fox, A. Pakman, and N. Tishby. Taming the noise in reinforcement learning via soft updates. *UAI-16*, pp.1889–1897, New York, 2016.

[Francois-Lavet *et al.*, 2017] V. Francois-Lavet, G. Rabusseau, J. Pineau, D. Ernst, and R. Fonteneau. On overfitting and asymptotic bias in batch reinforcement learning with partial observability, 2017. To appear, *JAIR*.

[Gauci *et al.*, 2018] J. Gauci, E. Conti, Y. Liang, K. Virochsiri, Y. He, Z. Kaden, V. Narayanan, and X. Ye. Horizon: Facebook's open source applied reinforcement learning platform. arXiv:1312.5602, 2018.

[Hauskrecht *et al.*, 1998] M. Hauskrecht, N. Meuleau, L. P. Kaelbling, T. Dean, and C. Boutilier. Hierarchical solution of Markov decision processes using macro-actions. *UAI-98*, pp.220–229, Madison, WI, 1998.

[Hohnhold *et al.*, 2015] H. Hohnhold, D. O'Brien, and D. Tang. Focusing on the long-term: It's good for users and business. *KDD-15*, pp.1849–1858, Sydney, 2015.

[Ie *et al.*, 2019] E. Ie, V. Jain, J. Wang, S. Navrekar, R. Agarwal, R. Wu, H. Cheng, T. Chandra, and C. Boutilier. SlateQ: A tractable decomposition for reinforcement learning with recommendation sets. *IJCAI-19*, 2019. To appear.

[Jaber, 2006] M. Y Jaber. Learning and forgetting models and their applications. *Handbook of Industrial and Systems Engineering*, 30(1):30–127, 2006.

[Littman and Sutton, 2002] M. L. Littman and R. S. Sutton. Predictive representations of state. *NIPS-02*, pp.1555–1561, Vancouver, 2002.

[Mladenov *et al.*, 2017] M. Mladenov, C. Boutilier, D. Schuurmans, O. Meshi, G. Elidan, T. Lu. Logistic Markov decision processes. *IJCAI-17*, pp.2486–2493, 2017.

[Mladenov *et al.*, 2019] M. Mladenov, O. Meshi, J. Ooi, D. Schuurmans, and C. Boutilier. Advantage amplification in slowly evolving latent-state environments. arXiv preprint, 2019.

[Nachum *et al.*, 2017] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans. Bridging the gap between value and policy based reinforcement learning. *NIPS-17*, pp.1476–1483, Long Beach, CA, 2017.

[Parr, 1998] R. Parr. Flexible decomposition algorithms for weakly coupled Markov decision processes. *UAI-98*, pp.422–430, Madison, WI, 1998.

[Roberts and Tedrake, 2009] J. W. Roberts and R. Tedrake. Signal-to-noise ratio analysis of policy gradient algorithms. *NIPS-09*, pp.1361–1368, Vancouver, 2009.

[Schulman *et al.*, 2015] J. Schulman, S. L., P. Abbeel, M. I. Jordan, and P. Moritz. Trust region policy optimization. *ICML-15*, pp.1889–1897, Sydney, 2015.

[Shani *et al.*, 2005] G. Shani, D. Heckerman, and R. Brafman. An MDP-based recommender system. *JMLR*, 6:1265–1295, 2005.

[Sharma *et al.*, 2017] S. Sharma, A. S. Lakshminarayanan, and B. Ravindran. Learning to repeat: Fine-grained action repetition for deep reinforcement learning. *ICLR-17*, Toulon, France, 2017.

[Silvestrini and Veredas, 2008] A. Silvestrini, D. Veredas. Temporal aggregation of univariate and multivariate time series models: A survey. *J. Econ. Surv.*, 22:458–497, 2008.

[Singh *et al.*, 1994] S. P. Singh, T. Jaakkola, and M. I. Jordan. Learning without state-estimation in partially observable Markovian decision processes. *ICML-94*, pp.284–292, New Brunswick, NJ, 1994.

[Smallwood and Sondik, 1973] R. Smallwood, E. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Op. Res.*, 21:1071–1088, 1973.

[Sutton *et al.*, 1999] R. S. Sutton, D. Precup, and S. P. Singh. Between MDPs and Semi-MDPs: Learning, planning, and representing knowledge at multiple temporal scales. *Artif. Intel.*, 112:181–211, 1999.

[Taghipour *et al.*, 2007] N. Taghipour, A. Kardan, and S. S. Ghidary. Usage-based web recommendations: A reinforcement learning approach. *RecSys07*, pp.113–120, 2007.

[Theocharous *et al.*, 2014] G. Theocharous, P. S. Thomas, and M. Ghavamzadeh. Safe policy search. *ICML-2014 Workshop on Customer Life-Time Value Optimization in Digital Marketing*, pp.1806–1812, 2014.

[Theocharous *et al.*, 2015] G. Theocharous, P. S. Thomas, and M. Ghavamzadeh. Personalized ad recommendation systems for life-time value optimization with guarantees. *IJCAI-15*, pp.1806–1812, 2015.

[Thodoroff *et al.*, 2018] P. Thodoroff, A. Durand, J. Pineau, and D. Precup. Temporal regularization for Markov decision process. *NeurIPS-18*, pp.1784–1794, Montreal, 2018.

[Thurstone, 1919] L. L. Thurstone. The learning curve equation. *Psychological Monographs*, 26(3):i, 1919.

[Wilhelm *et al.*, 2018] M. Wilhelm, A. Ramanathan, A. Bonomo, S. Jain, E. H. Chi, and J. Gillenwater. Practical diversified recommendations on youtube with determinantal point processes. *CIKM18*, pp.2165–2173, Torino, 2018.

[Zhao *et al.*, 2018] X. Zhao, L. Xia, L. Zhang, Z. Ding, D. Yin, and J. Tang. Deep reinforcement learning for page-wise recommendations. *RecSys-18*, pp.95–103, 2018.