

Advantage Amplification in Slowly Evolving Latent-State Environments

Martin Mladenov¹, Ofer Meshi¹, Jayden Ooi¹, Dale Schuurmans^{1,2}, Craig Boutilier¹

¹Google AI, Mountain View, CA, USA

²Department of Computer Science, University of Alberta, Edmonton, AB, Canada

{mmladanov,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; Zhao *et al.*, 2017; Archak *et al.*, 2012; Mladenov *et al.*, 2017]. Typically, actions are the set of recommendable items¹; 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 latent state (e.g., uncover topic interests via exploration) and (b) influence the latent state (e.g., create new interests or improve awareness and 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, amplification seeks to overcome the error induced by state estimation by introducing (explicit or implicit) *temporal abstraction* across policy space. We require that policies take 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 or macros designed for the purpose of allowing 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 (Sec. 2) and formalize this as a POMDP (Sec. 3), defining several novel concepts,

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

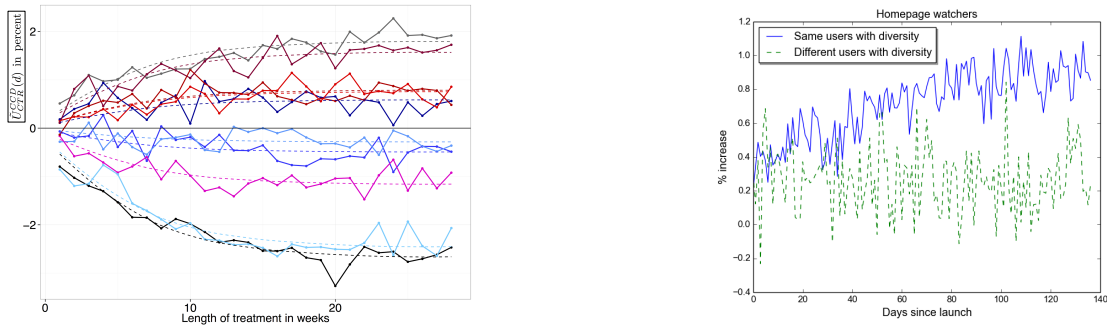


Figure 1: Gradual user response: (a) ad load/quality [Hohnhold *et al.*, 2015]; (b) YouTube recommendation diversity [Wilhelm *et al.*, 2018].

and show how low SNR interacts poorly with belief-state approximation (Sec. 4.1). We develop advantage amplification as a principle and 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. Experiments with stylized models show the effectiveness of these techniques.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 (hence ignores much of the true complexity of user satisfaction). 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 total positive-less-negative recommendations, with a discount $0 \leq \beta < 1$ applied to ensure that p is bounded: $p \in \left[\frac{-1}{1-\beta}, \frac{1}{1-\beta} \right]$. 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 her *degree of engagement* $g(s, a)$, and depends stochastically on both the quality of the recommendation, and her latent state s . 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 $a \in \{\text{Choc}, \text{Kale}\}$. We use $g(s, a)$ also to denote expected engagement. We require that *Choc* results in greater immediate (expected) engagement than *Kale*, $g(s, \text{Kale}) < g(s, \text{Choc})$, 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 *Kale* (and -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

We compute the Q-values of *Choc* and *Kale* for each satisfaction level s and plot them in Fig. 2a. 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. With additional noise, the Q-values become practically indistinguishable for a large range of satisfaction levels (Fig. 2b), which 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.

²Proofs, auxiliary lemmas and additional experiments are available in an extended version of the paper.

³Our 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.

⁴Such 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 serves a valuable role even when optimizing for individual users.

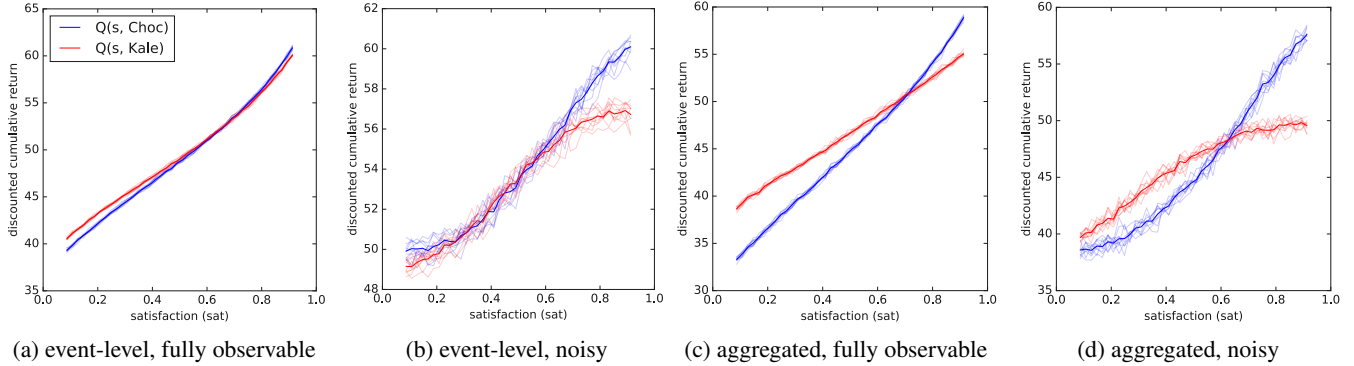


Figure 2: Q-values per satisfaction level in the Choc-Kale model.

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 $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T, \mathcal{Z}, O, R, \mathbf{b}_0, \gamma \rangle$ [Smallwood and Sondik, 1973]. States \mathcal{S} reflect user latent state and other observable aspects of the domain: in the CK model, this is simply a user’s current satisfaction s . Actions \mathcal{A} are recommendable items: in CK, we distinguish only *Choc* from *Kale*. The transition kernel $T(s, a, s')$ in the CK model is $T(s', a, s) = 1$ 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 $r_{\max} = \max_{s,a} R(s, a)$), \mathbf{b}_0 the initial state distribution, and $\gamma \in [0, 1)$ the 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* $\mathbf{b} \in \Delta(\mathcal{S})$. 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. Let \mathcal{B} denote the set of (realizable) belief states. We also require a mapping $U : \mathcal{B} \times \mathcal{A} \times \mathcal{Z} \rightarrow \mathcal{B}$ that describes the update $U(\mathbf{b}, a, z)$ of any $\mathbf{b} \in \mathcal{B}$ given $a \in \mathcal{A}, z \in \mathcal{Z}$. The pair (\mathcal{B}, U) defines our *representation*.

A (*stochastic*) *policy* is a mapping $\pi : \mathcal{B} \rightarrow \Delta(\mathcal{A})$ that selects an action distribution $\pi(\mathbf{b})$ for execution given belief \mathbf{b} ; we write $\pi(a|\mathbf{b})$ 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

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

We define $Q^\pi(\mathbf{b}, a)$ by fixing a in Eq. 1 (rather than taking an expectation). An *optimal policy* $\pi^* = \sup_{\pi} V^\pi$ over \mathcal{B} has value (resp., Q) function V^* (resp., Q^*). Optimal

⁵This is easily randomized if desired.

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

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 $\pi(\mathbf{b}) = \arg \max_a Q(\mathbf{b}, a)$ and its *induced value function* is $V(\mathbf{b}) = \max_a Q(\mathbf{b}, a) = Q(\mathbf{b}, a^*(\mathbf{b}))$. The *advantage function* $A(a, \mathbf{b}) = V^*(\mathbf{b}) - Q^*(\mathbf{b}, a)$ reflects the difference in the expected value of doing a at \mathbf{b} (and then acting optimally) vs. acting optimally at \mathbf{b} [Baird III, 1999]. If a_2 is the second-best action at \mathbf{b} , the *advantage* of that belief state is $A(\mathbf{b}) = V^*(\mathbf{b}) - Q^*(\mathbf{b}, a_2)$.

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 \mathbf{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 reflected in the data). In this case, any RL method that treats \mathbf{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]:

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

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:

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

(the denominator is treated as 0 if no \mathbf{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 $\mathbf{b}, \mathbf{b}' \in \mathcal{B}$, and $a \in A$ s.t. $T(\mathbf{b}', a, \mathbf{b}) > 0$, we have

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

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 now 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 describe two realizations, temporal aggregation (Sec. 4.2) and switching cost (Sec. 4.3).

4.1 The Impact of Low SNR on RL

The bound Eq. (2) can help assess the impact of low SNR on RL. Assume that policies, values 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 in the domain.

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}^π that differ from its *true* values Q^π ; and (b) the estimates \hat{Q}^π (hence, the choice of π itself) will depend on ρ . We bound the loss of π w.r.t. the optimal π^* (with exact filtering) as follows. First, for any (belief) state-action pair (\mathbf{b}, a) , suppose the maximum difference between its inferred and optimal Q-values is bounded for any ρ : $|Q^*(\mathbf{b}, a) - Q^\pi(\mathbf{b}, a)| \leq \delta$. By Eq. (2) we set

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

If \mathbf{b} has small advantage $A(\mathbf{b}) \leq 2\delta$, under behavior policy ρ , the estimate $\hat{Q}(\mathbf{b}, a_2)$ (the second-best action) can exceed that of $\hat{Q}(\mathbf{b}, a^*(\mathbf{b}))$; hence π executes a_2 . If π visits \mathbf{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 $A(\mathbf{b})$. To illustrate, consider two extreme regimes. First, if $A(\mathbf{b}) \geq 2\delta$ at all $\mathbf{b} \in \mathcal{B}$, 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 $A(\mathbf{b})$ 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, the stylized CK model) often have such small (but non-trivial) advantages across a wide range of state space.

The latter situation is illustrated on Fig. 2. In Fig. 2a, the Q-values of the CK model are plotted against the level of satisfaction (as if satisfaction were fully observable). The small advantages are notable. Fig. 2b shows the Q-value estimates for 10 independent tabular Q-learning reruns (the thin lines show the individual runs, the thick lines show the average) where noise is added to s . 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 \mathbf{b} is either small, $A(\mathbf{b}) \leq \sigma$, or large, $A(\mathbf{b}) > \Sigma$ for some constants $\sigma \ll 2\delta \leq \Sigma$. The induced policy incurs a loss of σ at small-advantage 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 $\frac{\epsilon r_{\max}}{(1-\gamma)^2}$ error (Eq. 3) depending on σ .

If the environment is smooth, *action aggregation* can be used to restructure a problem falling in the second regime to one in this 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 \mathbf{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 \mathbf{b} , all actions are 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 $Q^*(\mathbf{b}, a^{\times k})$ is the same as that of its event-level counterpart $Q^*(\mathbf{b}, a)$, 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 \mathbf{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 aggregating over all states induces some loss due to the inability to switch actions quickly. We bound that cost in computing σ and Σ . 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 \mathbf{b} , $|V^*(\mathbf{b}) - V^{\times k*}(\mathbf{b})| \leq \frac{2kL}{1-\gamma}$, where $V^{\times k*}(\mathbf{b})$ is the value of state \mathbf{b} under an optimal k -aggregate policy.⁷*

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

This theorem is proved by constructing a policy which switches actions every k events and showing that it has bounded regret. 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 $A(\mathbf{b}) \geq 2kL$, the k -aggregate-horizon advantage is bounded below:*

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

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 an 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 Sec. 4.4, 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 $L \rightarrow 0$ and amplification at every state is guaranteed.*

4.3 Temporal Regularization: Switching Cost

As discussed above, temporal aggregation is guaranteed to improve learning in slow environments. It has, however, certain practical drawbacks due to its inflexibility. One such drawback is that, in the non-Markovian setting induced by belief-state approximation, training data should ideally be collected using a k -aggregated behavior policy.⁸ 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, but are harmless from an SNR perspective if they induce large advantage gaps. 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:

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

⁸This is unnecessary if the system is Markovian, since (s, a, r, s') tuples may be reordered to emulate any behavioral policy.

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 $R(\mathbf{b}, a_{t-1}, a_t) = R(\mathbf{b}, a_t) - T \cdot \mathbb{1}[a_t \neq a_{t-1}]$.

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 in the case with action aggregation.

In the case of problems with 2 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 \mathbf{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 2-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].

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

Theorem 4. *Let κ be as in Thm. 3. 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 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, 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 faulty state estimation, we corrupt the satisfaction level s with noise drawn from a Gaussian (mean 0, stdev. σ_N), truncated on $[-1, 1]$. As we increase σ_N , state estimation becomes worse. To mitigate this effect, we apply aggregations of 3, 5 actions at discounts of $\gamma = 0.95$ and 0.99 (Fig. 3a,b) and switching costs of 1, 2, 3 (Fig. 3c). For each parameter setting, we train 10 tabular policies with Q -learning, discretizing state space into 50 buckets. For each training run, we roll-out 30000 event-level transitions, exploring using actions taken uniformly at random—aggregated actions in the aggregation setting—then evaluate the discounted return of

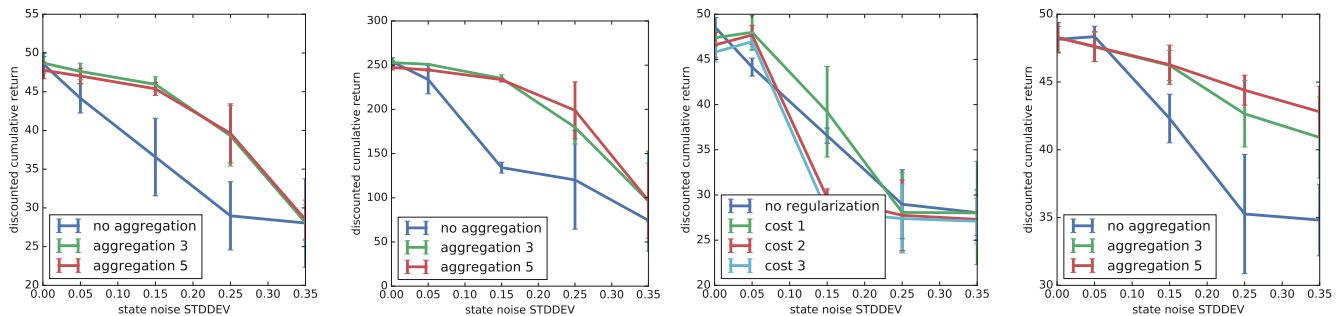


Figure 3: Experimental results.

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 the 95% confidence interval) as a function of the σ_N . We see that action aggregation has a profound effect on solution quality, improving the performance of the policy up to a factor of almost 2 (for $\gamma = 0.99$). Switching cost regularization has a more subtle effect, providing more modest improvements in performance. We observe that over-regularized policies actually perform worse than the unregularized policy. We conjecture that this stark 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 a more “options-oriented” perspective on the problem. Here, recommendable items have a continuous “kaleness” score between 0 and 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 $\lambda = 0.2$. The results are shown in Fig. 3d and exhibit 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 seems to be 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 theory employs temporal aggregation to organize the policy space into a hierarchy, where a pre-specified 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], who also informally suggest that reduced control frequency can improve SNR; however the task of characterizing this phenomenon formally has not been addressed to the best of our

knowledge. The *learning to repeat* framework (see [Sharma *et al.*, 2017] and references therein) provide 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 reinforcement learning algorithms was first studied by Farahmand [?]. Bellemare *et al.* [?] observed that the action gap can be manipulated to improve the quality of learned policies by introducing local policy consistency constraints to the Bellman operator. Their considerations are, however, not bound to specific environment properties.

Finally, 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 this work, however, regularization is simply a price to pay to achieve a side goal (e.g., better optimization/statistical efficiency). While stochastic policies often perform better than deterministic ones when state estimation is deficient [Singh *et al.*, 1994], and 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.

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: (i) the study of soft-policy regularization for amplification (preliminary results are presented in the long version of this paper); (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.

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.
- [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, and 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.
- [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*.
- [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.
- [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, and T. Lu. Logistic Markov decision processes. *IJCAI-17*, pp.2486–2493, Melbourne, 2017.
- [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. I. 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 and D. Veredas. Temporal aggregation of univariate and multivariate time series models: A survey. *J. Econ. Surveys*, 22(3):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. D. Smallwood and E. J. 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, Minneapolis, 2007.
- [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.*, 2017] X. Zhao, L. Zhang, Z. Ding, D. Yin, Y. Zhao, and J. Tang. Deep reinforcement learning for list-wise recommendations. *arXiv:1801.00209*, 2017.

A Technical Results and Additional Material

A.1 Analysis of Action Aggregation

We start with an observation that frames the subsequent analysis.

Lemma 1 (Counterfactual Q-values). *Let π and ρ be two deterministic policies and let the Q-function of π , Q^π be known. Then, for any state $\mathbf{b} \in \mathcal{B}$:*

$$\begin{aligned} V^\pi(\mathbf{b}) - V^\rho(\mathbf{b}) &= Q^\pi(\mathbf{b}, \pi(\mathbf{b})) - Q^\rho(\mathbf{b}, \rho(\mathbf{b})) \\ &= \mathbb{E}_\rho \left[\sum_{i=0}^{\infty} \gamma^i (Q^\pi(\mathbf{b}_i, \pi(\mathbf{b}_i)) - Q^\pi(\mathbf{b}_i, \rho(\mathbf{b}_i))) \right], \end{aligned}$$

where \mathbb{E}_ρ the expectation is over trajectories $(\mathbf{b}_i)_{i \in \mathbb{N}}$ generated by ρ starting at \mathbf{b} ($\mathbf{b}_0 = \mathbf{b}$).

Proof. Let $\rho\pi$ be a non-stationary policy that starting at \mathbf{b} executes $\rho(\mathbf{b})$ exactly once, then follows π forever. Then,

$$\begin{aligned} Q^\pi(\mathbf{b}, \pi(\mathbf{b})) - Q^\rho(\mathbf{b}, \rho(\mathbf{b})) &= Q^\pi(\mathbf{b}, \pi(\mathbf{b})) - Q^{\rho\pi}(\mathbf{b}, \rho\pi(\mathbf{b})) \\ &\quad + Q^{\rho\pi}(\mathbf{b}, \rho\pi(\mathbf{b})) - Q^\rho(\mathbf{b}, \rho(\mathbf{b})) \\ &= Q^\pi(\mathbf{b}, \pi(\mathbf{b})) - Q^\pi(\mathbf{b}, \rho(\mathbf{b})) \\ &\quad + Q^\pi(\mathbf{b}, \rho(\mathbf{b})) - Q^\rho(\mathbf{b}, \rho(\mathbf{b})) \\ &= Q^\pi(\mathbf{b}, \pi(\mathbf{b})) - Q^\pi(\mathbf{b}, \rho(\mathbf{b})) \\ &\quad + \gamma \mathbb{E}_{\mathbf{b}' \sim \rho} [Q^\pi(\mathbf{b}', \pi(\mathbf{b}')) - Q^\rho(\mathbf{b}', \rho(\mathbf{b}'))]. \end{aligned}$$

Linearity of expectation and the boundedness of both Q functions ensures that recursive application to $Q^\pi(\mathbf{b}', \pi(\mathbf{b}')) - Q^\rho(\mathbf{b}', \rho(\mathbf{b}'))$ converges to the desired quantity. \square

This allows us to measure the behavior of k -aggregate policies (in terms of advantages) relative to atomic ones.

Lemma 2. *Let \mathbf{b} be a state for which the optimal action a does not change within k steps. In other words, for all $\mathbf{b}' \in \mathcal{B}^{\times k}(\mathbf{b})$, $\pi^*(\mathbf{b}') = a$, where $\mathcal{B}^{\times k}(\mathbf{b})$ denotes all states reachable from \mathbf{b} in k steps under some action sequence. Then for any other action a' ,*

$$\begin{aligned} Q^*(\mathbf{b}, a^{\times k}) - Q^*(\mathbf{b}, a'^{\times k}) &= \mathbb{E} \left[\sum_{i=0}^{k-1} \gamma^i (Q^*(\mathbf{b}_i^{a'}, a) - Q^*(\mathbf{b}_i^{a'}, a')) \right] \end{aligned}$$

where the expectation is over the trajectory $\mathbf{b}_1^{a'}, \dots, \mathbf{b}_k^{a'}$ of states that follow after \mathbf{b} when taking action a' and $\mathbf{b}_0^{a'} = \mathbf{b}$ for notational convenience.

Proof. Let π be a policy that executes a for k steps and then reverts to the optimal policy and ρ be a policy that executes b for k steps and then reverts to the optimal policy. We apply Lemma 1 noting that π coincides with the optimal policy, hence $Q^\pi = Q^*$. \square

Lemma 2 establishes that advantage of k -aggregate actions is the compound discounted advantage of the atomic ones. This, combined with the smoothness of the optimal Q-function

allows us to analyze the effects of action aggregation. In particular, suppose that we have found a state \mathbf{b} such that $A(\mathbf{b}) = Q^*(\mathbf{b}, a) - \max_{a' \neq a} Q^*(\mathbf{b}, a') \geq 2kL$. The smoothness of the Q-function allows us to infer that for any action taken, the advantage at the next state will be at least $2kL - 2L$, $2kL - 4L$ after 2 steps and so on (this guarantees the advantage gap can't close in less than k steps). Hence if we were to replace atomic actions with k -repeated actions at states with advantage of more than $2kL$, the following can be observed:

Lemma 3. *Consider an MDP with an L -smooth optimal Q-function Q^* , and the reparametrization by holding actions fixed for k steps whenever the advantage gap is greater than $2kL$. Then, at each state, either:*

$$\begin{aligned} A^{\times k}(\mathbf{b}) &:= Q^*(\mathbf{b}, a^{\times k}) - Q^*(\mathbf{b}, a'^{\times k}) \\ &\geq A(\mathbf{b}) \frac{1 - \gamma^k}{1 - \gamma} \\ &\quad - 2\gamma L \frac{1 - (1 + k - \gamma k)\gamma^k}{(1 - \gamma)^2}, \end{aligned}$$

or $A(\mathbf{b}) \leq 2kL$.

Proof. As established, $A^{\times k}(\mathbf{b}) = E[\sum_{i=0}^{k-1} \gamma^i A(\mathbf{b}_i^{a'})]$. Smoothness of the Q-function guarantees that $A(\mathbf{b}_{i+1}^{a'}) \geq A(\mathbf{b}_i^{a'}) - 2L$, resp. $A(\mathbf{b}_k^{a'}) \geq A(\mathbf{b}) - 2kL$ as discussed above. Hence

$$A^{\times k}(\mathbf{b}) = \mathbb{E} \left[\sum_{i=0}^{k-1} \gamma^i A(\mathbf{b}_i^{a'}) \right] \quad (5)$$

$$\geq \mathbb{E} \left[\sum_{i=0}^{k-1} \gamma^i (A(\mathbf{b}) - 2iL) \right] \quad (6)$$

$$= A(\mathbf{b}) \frac{1 - \gamma^k}{1 - \gamma} - 2L \frac{\gamma - (1 + k - \gamma k)\gamma^{k+1}}{(1 - \gamma)^2}. \quad (7)$$

Here, the first expression comes from the finite sum geometric series formula, and the second term reflects the fact that

$$\sum_{i=0}^{k-1} i\gamma^i = \frac{\gamma - (1 + k - \gamma k)\gamma^{k+1}}{(1 - \gamma)^2}. \quad \square$$

We have thus far considered the situation where aggregation is state-dependent, based on the knowledge of the advantage amplitude. A more realistic implementation of aggregation would be to fix some global static k a priori. While the reasoning is similar to the above, there is an added complexity that the aggregation itself comes at a price – the values of certain states will be reduced (relative to the best atomic policy) due to the inability to rapidly switch actions. This additional cost must then be factored into the bounds.

We first bound the cost of the best slow policy under smooth Q assumptions.

Theorem 1. *Let k be a fixed horizon and Q^* , the event-level optimal Q-function be L -smooth. Then for all \mathbf{b} , $|V^*(\mathbf{b}) -$*

$V^{\times k^*}(\mathbf{b})| \leq \frac{2kL}{1-\gamma}$, where $V^{\times k^*}(\mathbf{b})$ is the value of state \mathbf{b} under an optimal k -aggregate policy⁹.

Proof. We lower bound the return of the optimal k -aggregate policy by exhibiting a (not-necessarily optimal) k -aggregate policy with sufficient returns. In particular, let π be a deterministic optimal atomic policy, and ρ be the policy that at state \mathbf{b}_t executes $\pi(\mathbf{b}_t)$ for the next k steps, then executes $\pi(\mathbf{b}_{t+k})$, etc. Following Lemma 1, we have that

$$\begin{aligned} & V^\pi(\mathbf{b}) - V^\rho(\mathbf{b}) \\ &= \mathbb{E}_\rho \left[\sum_{i=0}^{k-1} \gamma^i (Q^\pi(\mathbf{b}_i, \pi(\mathbf{b}_i)) - Q^\rho(\mathbf{b}_i, \rho(\mathbf{b}_i))) \right]. \end{aligned}$$

Let $(\mathbf{b}_{ik})_{i \in \mathbb{N}}$ be the states at which ρ may switch actions. At any \mathbf{b}_{ik} , if $A(\mathbf{b}_{ik}) \geq 2kL$, due to smoothness of Q^* , we know that a remains optimal for the next k steps thus $\rho(\mathbf{b}_{ik+j}) = \pi(\mathbf{b}_{ik+j})$ and $Q^*(\mathbf{b}_{ik+j}, \pi(\mathbf{b}_{ik+j})) - Q^*(\mathbf{b}_{ik+j}, \rho(\mathbf{b}_{ik+j})) = 0$ until the next decision point for ρ (i.e. for $j \leq k-1$). Conversely, if $A(\mathbf{b}_{ik}) < 2kL$, then in the worst case, assuming that the Q-value of the optimal action $Q^*(\mathbf{b}_{ik+j}, a)$ decreases by L with every step j and the Q-value of a suboptimal action $Q^*(\mathbf{b}_{ik+j}, a')$ increases respectively by L , then for any of the next k steps, $Q^*(\mathbf{b}_{ik+j}, a) - Q^*(\mathbf{b}_{ik+j}, a') \geq -2kL$. Hence, for all i , $|Q^*(\mathbf{b}_i, \pi(a)) - Q^*(\mathbf{b}_i, \rho(a))|$ is either 0 or less than $2kL$, leading to

$$V^\pi(\mathbf{b}) - V^\rho(\mathbf{b}) \leq \sum_{i=0}^{k-1} \gamma^i 2kL = \frac{2kL}{1-\gamma}.$$

□

We now factor in the aggregation loss from Thm. 1 into Lemma 3, which together characterize the amplification properties of action aggregation.

Theorem 2. *In an L smooth MDP, let k be a fixed repetition horizon. For any state where the advantage gap $A(\mathbf{b})$ greater than $2kL$, the fixed-horizon advantage is lower-bounded as follows:*

$$\begin{aligned} & A(\mathbf{b})^{\times k} \\ & \geq A(\mathbf{b}) \frac{1-\gamma^k}{1-\gamma} - 2L \frac{\gamma - (1+k-\gamma k)\gamma^{k+1}}{(1-\gamma)^2} \\ & \quad - \frac{2kL}{1-\gamma}. \end{aligned}$$

Proof. Let $\bar{Q}^{\times k^*}$ be the losslessly amplified Q-function as in Lemma 3. Since $A(\mathbf{b}) \geq 2kL$, action a is optimal under an atomic optimal policy for the next k periods and due to Thm. 1,

$$Q^{\times k^*}(\mathbf{b}, a^{\times k}) \geq Q^*(\mathbf{b}, a) - \frac{2kL}{1-\gamma} = \bar{Q}^{\times k^*}(\mathbf{b}, a^{\times k}) - \frac{2kL}{1-\gamma}.$$

Moreover, for any $a' \neq a$

$$Q^{\times k^*}(\mathbf{b}, a'^{\times k}) \leq \bar{Q}^{\times k^*}(\mathbf{b}, a'^{\times k}),$$

⁹Note that the reparametrized problem in which a decision can be made every k atomic steps is also an MDP, so the notion of an optimal value function (one that provides the largest possible value at every state) resp. optimal deterministic policy is well-defined.

as the expected return following the aggregation period of k can only be lower than the one of the losslessly amplified problem. Thus:

$$\begin{aligned} & Q^{\times k^*}(\mathbf{b}, a^{\times k}) - Q^{\times k^*}(\mathbf{b}, a'^{\times k}) \\ & \geq \mathbb{E} \left[\sum_i^k \gamma^i (Q^*(\mathbf{b}_i, a) - Q^*(\mathbf{b}_i, a')) \right] - \frac{2kL}{1-\gamma}. \end{aligned}$$

The bound from the expectation term comes from Lemma 3. □

A.2 Analysis of Switching Cost

Let Q^* be an L -smooth optimal Q-function of a POMDP with an optimal deterministic policy π^* and T be a switching cost. Consider the following policy ρ : at \mathbf{b}_0 , ρ adopts the optimal atomic action $a_0 = \rho_0(\mathbf{b}_0) = \pi^*(\mathbf{b}_0)$. Also at $t = 0$, ρ calculates the time until its regret for repeating a_0 relative to following the optimal policy exceeds, T , i.e.

$$\begin{aligned} \omega_0 = & \arg \min_t \left\{ t : \sum_{i=0}^t \gamma^i \mathbb{E} [Q^*(\mathbf{b}_i, \pi^*(\mathbf{b}_i)) \right. \\ & \left. - Q^*(\mathbf{b}_i, a_0) | \mathbf{b}_0] \geq T \right\}, \end{aligned}$$

(where the expectation is, as before, over trajectories of belief states generated by executing a_0 , conditioned on the realization of \mathbf{b}_0) and then repeats a_0 until $t = \omega_0$. At $t = \omega_0$, ρ queries the optimal policy for $a_{\omega_0} = \pi^*(\mathbf{b}_{\omega_0})$, executes it until time ω_1 and so on. In other words, ρ repeats an action adopted from the optimal policy until its regret for not having followed the optimal policy exceeds T and then switches.

The return of ρ is equivalent to the return of a policy that would pay T upfront to switch to and follow the optimal policy for ω steps. We first bound the loss of this policy and then argue that it also upper-bounds the loss of the optimal switching cost policy.

Lemma 4. *The total regret of ρ relative to π^* is no more than*

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

Proof. Starting from Lemma 1, the total loss of ρ relative to π^* is

$$\sum_{i=0}^{\infty} \gamma^i \mathbb{E}_{\mathbf{b}_i \sim \rho} [Q^*(\mathbf{b}_i, \pi^*(\mathbf{b}_i)) - Q^*(\mathbf{b}_i, \rho_i)].$$

We now argue that $\mathbb{E}_{\mathbf{b}_i \sim \rho} [Q^*(\mathbf{b}_i, \pi^*(\mathbf{b}_i)) - Q^*(\mathbf{b}_i, \rho_i)]$ is bounded at any step i . Observe that the action taken by the policy at time i depends only on the belief state realization $\mathbf{b}_{\omega_{-1}}, \omega_{-1}$ being the most recent time point at which the policy was allowed to switch actions. Thus, by the law of total probability,

$$\begin{aligned} & \mathbb{E}_{\mathbf{b}_i \sim \rho} [Q^*(\mathbf{b}_i, \pi^*(\mathbf{b}_i)) - Q^*(\mathbf{b}_i, \rho_i)] \\ &= \mathbb{E}_{\mathbf{b}_{\omega_{-1}} \sim \rho} \mathbb{E}_{\mathbf{b}_i \sim \rho} [Q^*(\mathbf{b}_i, \pi^*(\mathbf{b}_i)) - Q^*(\mathbf{b}_i, \rho_i) | \mathbf{b}_{\omega_{-1}}]. \end{aligned}$$

In the above, we have just rewritten the loss relative to the last switching point ω_{-1} . We now show that the conditional expectation $\mathbb{E}_{\mathbf{b}_i \sim \rho} [Q^*(\mathbf{b}_i, \pi^*(\mathbf{b}_i)) - Q^*(\mathbf{b}_i, \rho_i) | \mathbf{b}_{\omega_{-1}}]$ is bounded for any realization $\mathbf{b}_{\omega_{-1}}$. Recall that at time ω_{-1} , ρ picks the action $a_{\omega_{-1}} = \pi^*(\omega_{-1})$ and executes it until time ω , the time when its expected regret exceeds T , and then switches. Thus, the highest achievable per-event expected regret conditioned on $\mathbf{b}_{\omega_{-1}}$ occurs when at time ω_{-1} , there exists an alternative action $a' \neq a_{\omega_{-1}}$ having the same Q-value, i.e. $Q^*(\mathbf{b}_{\omega_{-1}}, a_{\omega_{-1}}) = Q^*(\mathbf{b}_{\omega_{-1}}, a')$, and then, in subsequent time steps, the expected Q-value of a' , $\mathbb{E}_{\mathbf{b}_i \sim \rho} [Q^*(\mathbf{b}_i, a') | \mathbf{b}_{\omega_{-1}}]$, starts increasing at the maximum rate of L while the Q-value of $a_{\omega_{-1}}$ starts decreasing at the same rate. The maximum rate of increase of the expectation is justified since if the Q-value of a' is L -smooth over individual sample paths, i.e. $|Q^*(\mathbf{b}_i, a) - Q^*(\mathbf{b}_{i+1}, a)| \leq L$, the sequence of expectations is also L -smooth, $|\mathbb{E}_{\mathbf{b}_i \sim \rho} [Q^*(\mathbf{b}_i, a) | \mathbf{b}_{\omega_{-1}}] - \mathbb{E}_{\mathbf{b}_{i+1} | \mathbf{b}_{\omega_{-1}} \sim \rho} [Q^*(\mathbf{b}_{i+1}, a) | \mathbf{b}_{\omega_{-1}}]| \leq L$. By the same reasoning as in Lemma 3 (but this time for the sequence of expectations, rather than the sequence of realizations), the regret for not following π^* after k steps is bounded by $2\gamma L \frac{1+k\gamma^k - (1+k)\gamma^k}{(1-\gamma)^2}$. Let us calculate the minimum k (under assumptions of monotone regret increase at the maximum rate) for which this regret exceeds T as $\lceil \kappa \rceil$, where κ is the solution of $2\gamma L \frac{1+k\gamma^k - (1+k)\gamma^k}{(1-\gamma)^2} = T$.

$$\begin{aligned} 2\gamma L \frac{1+k\gamma^{k+1} - (1+k)\gamma^k}{(1-\gamma)^2} &= T \\ \rightarrow k\gamma^{k+1} - (1+k)\gamma^k &= \frac{(1-\gamma)^2}{2\gamma L} T - 1 \\ \rightarrow (\gamma-1)k\gamma^k - \gamma^k &= \frac{(1-\gamma)^2}{2\gamma L} T - 1 \\ \rightarrow k &= \frac{\log \gamma + (\gamma-1)W\left(\frac{\gamma^{1/(1-\gamma)}}{\gamma-1} \left(\frac{(1-\gamma)^2}{2\gamma L} T - 1\right) \log \gamma\right)}{(\gamma-1) \log \gamma}, \end{aligned}$$

where W is the Lambert W-function. Knowing κ , we can deduce that the maximum per-event expected regret, $\mathbb{E}_{s_i \sim \rho} [Q^*(s_i, \pi^*(s_i)) - Q^*(s_i, \rho_i) | \mathbf{b}_{\omega_{-1}}]$ is no more than

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

under worst-case assumptions. This leads to an overall regret bound of $\epsilon/(1-\gamma)$. \square

So far, we have produced an agent that pays a cost T to switch to the optimal policy for some time-specific horizon ω_t and bounded its regret. This is not quite what we need, since ρ gets more than one action switch for free within the horizon over which it is allowed to follow the optimal policy. We now argue that for a slow environment, this bound also holds for an agent which pays on every switch. To do so, we need to produce a reasonable policy that pays for every action switch.

Let $A(\mathbf{b}, \pi^*, a^{\times k})$ be the advantage of following the optimal policy for k turns over repeating a for the same time and

then switching to the optimal policy (resp. the regret for repeating a over following the optimal policy). That is, following Lemma 1,

$$\begin{aligned} A(\mathbf{b}, \pi^*, a^{\times k}) &= \\ \mathbb{E}_{\mathbf{b}_0, \dots, \mathbf{b}_k} \left[\sum_{i=0}^k \gamma^i (Q^*(\mathbf{b}_i, \pi^*(\mathbf{b}_i)) - Q^*(\mathbf{b}_i, a)) \right], \end{aligned}$$

where the expectation is over trajectories generated by taking action a and $\mathbf{b}_0 = \mathbf{b}$.

Furthermore, let $\omega(\mathbf{b}) = \arg \min_k \min_{a' \in A} k$ s.t. $A(\mathbf{b}, \pi^*, a_0^{\times k}) - A(\mathbf{b}, \pi^*, a'^{\times k}) \geq T$ if the former is feasible, else ∞ . In words, $\omega(\mathbf{b})$ is the shortest time until the regret of having repeated the current action a_0 exceeds the regret of having repeated some other action a' by T . Note that this does not need to be finite, since in some state, repeating any action for any amount of time might yield similar returns.

Based on this, we construct a policy σ that behaves as if it would pay T on every switch. At \mathbf{b}_0 , σ executes the atomic optimal action $\pi^*(\mathbf{b}_0)$. At the next state, \mathbf{b}_1 , if $\omega(\mathbf{b}_1)$ is infinite, σ repeats a_0 once. Otherwise, σ repeats a_0 $\omega(\mathbf{b}_1)$ times and then switches to a' , the action whose regret exceeds T after $\omega(\mathbf{b}_1)$ steps. This policy can do no better than a policy that pays T upfront to switch to a' and maintain it for $\omega(\mathbf{b}_1)$ steps. By extension, σ can generate no more return than the best switching cost policy. Now we can bound its loss.

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

Proof. Again we must scrutinize the sequence $(\mathbb{E}_{\mathbf{b}_i \sim \sigma} [Q^*(\mathbf{b}_i, \pi^*(\mathbf{b}_i)) - Q^*(\mathbf{b}_i, \sigma_i)])_{i \in \mathbb{N}}$ showing that it is bounded at any step i . We consider the contrapositive: suppose there existed i , such that

$$\mathbb{E}_{\mathbf{b}_i \sim \sigma} [Q^*(\mathbf{b}_i, \pi^*(\mathbf{b}_i)) - Q^*(\mathbf{b}_i, \sigma_i)] > 2\kappa L$$

with κ as in the previous theorem. By the law of total probability,

$$\begin{aligned} \mathbb{E}_{\mathbf{b}_i \sim \sigma} [Q^*(\mathbf{b}_i, \pi^*(\mathbf{b}_i)) - Q^*(\mathbf{b}_i, \sigma_i)] \\ = \mathbb{E}_{\mathbf{b}_{i-\kappa}} \mathbb{E}_{\mathbf{b}_i \sim \sigma} [Q^*(\mathbf{b}_i, \pi^*(\mathbf{b}_i)) - Q^*(\mathbf{b}_i, \sigma_i) | \mathbf{b}_{i-\kappa}] > 2\kappa L. \end{aligned}$$

The expectation over $\mathbf{b}_{i-\kappa}$ in the second line above implies that for at least one possible realization of $\mathbf{b}_{i-\kappa}$, the expected Q difference after κ periods exceeds $2\kappa L$, i.e.

$$\mathbb{E}_{\mathbf{b}_i \sim \sigma} [Q^*(\mathbf{b}_i, \pi^*(\mathbf{b}_i)) - Q^*(\mathbf{b}_i, \sigma_i) | \mathbf{b}_{i-\kappa}] \geq 2\kappa L. \text{ Let } \hat{\mathbf{b}} \text{ be such a realization of } \mathbf{b}_{i-\kappa} \text{ and let } a^* = \pi^*(\mathbf{b}_i) \text{ resp. } a = \sigma(\mathbf{b}_i). \text{ Due to the } L\text{-smoothness of the sequences of expected Q-values } \left(\mathbb{E}_{\mathbf{b}_j \sim \rho} [Q^*(\mathbf{b}_j, a^*) | \mathbf{b}_{i-\kappa} = \hat{\mathbf{b}}] \right)_{j \in [i-\kappa, \dots, i]},$$

$\left(\mathbb{E}_{\mathbf{b}_j \sim \rho} [Q^*(\mathbf{b}_j, a) | \mathbf{b}_{i-\kappa} = \hat{\mathbf{b}}] \right)_{j \in [i-\kappa, \dots, i]}$, we know that for all $j \in [i-\kappa, \dots, i]$, $\mathbb{E}_{\mathbf{b}_j \sim \rho} [Q^*(\mathbf{b}_j, a^*) | \mathbf{b}_{i-\kappa} = \hat{\mathbf{b}}] \geq \mathbb{E}_{\mathbf{b}_j \sim \rho} [Q^*(\mathbf{b}_j, a) | \mathbf{b}_{i-\kappa} = \hat{\mathbf{b}}]$. It is straightforward to verify that for any L -smooth sequences of real numbers $(A_i)_{i \in [1, \dots, k]} \geq (B_i)_{i \in [1, \dots, k]}$ of length κ , such that $A_\kappa - B_\kappa > 2\kappa L$, the discounted sum $\sum_{i=1}^\kappa \gamma^{i-1} (A_i - B_i)$ must

be greater than T . Hence, the cumulative expected regret of taking a^* vs. a following state $\hat{\mathbf{b}}$ must exceed T and σ must have switched by time i , which is a contradiction with the assumption that σ executes a .

□