

---

# No-regret Exploration in Contextual Reinforcement Learning

---

**Aditya Modi**

Computer Science and Engineering  
University of Michigan

**Ambuj Tewari**

Department of Statistics  
University of Michigan

## Abstract

We consider the recently proposed reinforcement learning (RL) framework of Contextual Markov Decision Processes (CMDP), where the agent interacts with a (potentially adversarial) sequence of episodic tabular MDPs. In addition, a context vector determining the MDP parameters is available to the agent at the start of each episode, thereby allowing it to learn a context-dependent near-optimal policy. In this paper, we propose a no-regret online RL algorithm in the setting where the MDP parameters are obtained from the context using generalized linear mappings (GLMs). We propose and analyze optimistic and randomized exploration methods which make (time and space) efficient online updates. The GLM based model subsumes previous work in this area and also improves previous known bounds in the special case where the contextual mapping is linear. In addition, we demonstrate a generic template to derive confidence sets using an online learning oracle and give a lower bound for the setting.

## 1 INTRODUCTION

Recent advances in reinforcement learning (RL) methods have led to increased focus on finding practical RL applications. RL algorithms provide a set of tools for tackling sequential decision making problems with potential applications ranging from web advertising and portfolio optimization, to healthcare applications like adaptive drug treatment. However, despite the empirical success of RL in simulated domains such as boardgames and video games, it has seen limited use in real world applications because of the inherent trial-and-error nature of the paradigm. In addition to these concerns, for the applications listed above, we have to essentially design

adaptive methods for a *population* of users instead of a single system. For instance, optimizing adaptive drug treatment plans for an influx of patients has two key requirements: (1) ensure quickly learning good policies for each user and (2) share the observed outcome data efficiently across patients. Intuitively, we expect that frequently seen patient types (with some notion of similarity) can be adequately dealt with by using adaptive learning methods whereas difficult and rare cases could be carefully referred to experts to safely generate more data.

An efficient and plausible way to incorporate this heterogeneity is to include any distinguishing exogenous factors in form of a contextual information vector in the learning process. This information can include demographic, genomic features or individual measurements taken from lab tests. We model this setting using the framework of Contextual Markov Decision Processes (CMDPs) (Modi et al., 2018) where the learner has access to some *contextual features* at the start of every patient interaction. Similar settings have been studied with slight variations by Abbasi-Yadkori and Neu (2014); Hallak et al. (2015) and Dann et al. (2019). While the framework proposed in these works is innovative, there are a number of deficiencies in the available set of results. First, theoretical guarantees (PAC-style mistake bounds or regret bounds) sometimes hold only under a linearity assumption on the mapping between contexts and MDPs. This assumption is quite restrictive as it enforces additional constraints on the context features which are harder to satisfy in practice. Second, if non-linear mappings are introduced (Abbasi-Yadkori and Neu, 2014), the next state distributions are left un-normalized and therefore do not correctly model the context dependence of MDP dynamics.

We address these deficiencies by considering generalized linear models (GLMs) for mapping context features to MDP parameters (succinctly referred to as GLM-CMDP). We build upon the existing work on generalized linear bandits (Zhang et al., 2016) and propose UCRL2 (optimistic) and RLSVI (randomized) like algorithms with

regret analyses. Overall, our contributions are as follows:

- We provide optimistic and randomized regret minimizing algorithms for GLM-CMDPs. Our model subsumes/corrects previous CMDP frameworks and our analysis improves on the existing regret bounds by a factor of  $\mathcal{O}(\sqrt{S})$  in the linear case.
- The proposed algorithms use *efficient online updates*, both in terms of memory and time complexity, improving over typical OFU approaches whose running time scales linearly with number of rounds.
- We prove a regret lower bound for GLM-CMDP when a logistic or quadratic link function is used.
- We provide a generic way to convert any online no-regret algorithm for estimating GLM parameters to confidence sets. This allows an improvement in the regret incurred by our methods when the GLM parameters have additional structure (e.g., sparsity).

## 2 SETTING AND NOTATION

We consider episodic Markov decision processes, denoted by tuple  $(\mathcal{S}, \mathcal{A}, P, R, H)$  where  $\mathcal{S}$  and  $\mathcal{A}$  are finite state and action spaces,  $P(\cdot|s, a)$  the transition distribution,  $R(s, a)$  the reward function with mean  $r(s, a)$  and  $H$  is the horizon. Without loss of generality, we will consider a fixed start state for each episode. In the contextual MDP setting (Hallak et al., 2015; Modi et al., 2018), the agent interacts with a sequence of MDPs  $M_k$  (indexed by  $k$ ) whose dynamics and reward functions (denoted by  $P_k$  and  $R_k$ ) are determined by an observed context vector  $x_k \in \mathcal{X}$ . For notation, we use  $(s_{k,h}, a_{k,h}, r_{k,h}, s_{k,h+1})$  to denote the transition at step  $h$  in episode  $k$ . We denote the size of MDP parameters by the usual notation:  $|\mathcal{S}| = S$  and  $|\mathcal{A}| = A$ .

The value of a policy in an episode  $k$  is defined as the expected total return for  $H$  steps in MDP  $M_k$ :

$$v_k^\pi = \mathbb{E}_{M_k, \pi} \left[ \sum_{h=1}^H r_{k,h} \right]$$

The optimal policy for episode  $k$  is denoted by  $\pi_k^* := \arg \max_{\pi} v_k^\pi$  and its value as  $v_k^*$ . The agent’s goal in the CMDP setting is to learn a context dependent policy  $\pi : \mathcal{X} \times \mathcal{S} \rightarrow \mathcal{A}$  such that cumulative expected return over  $K$  episodes is maximized. We quantify the agent’s performance by the total regret incurred over a (potentially adversarial) sequence of  $K$  contexts:

$$R(K) := \sum_{k=1}^K v_k^* - v_k^{\pi_k} \quad (1)$$

Note that the regret here is defined with respect to the sequence of context dependent optimal policies.

**Additional notation.** For two matrices  $X$  and  $Y$ , the inner product is defined as  $\langle X, Y \rangle := \text{Tr}(X^\top Y)$ . For a vector  $x \in \mathbb{R}^d$  and a matrix  $A \in \mathbb{R}^{d \times d}$ , we define  $\|x\|_A^2 := x^\top A x$ . For matrices  $W \in \mathbb{R}^{m \times n}$  and  $X \in \mathbb{R}^{n \times n}$ , we define  $\|W\|_X^2 := \sum_{i=1}^m \|W^{(i)}\|_X^2$  where  $W^{(i)}$  is the  $i^{\text{th}}$  row of the matrix. Further, we reserve the notation  $\|W\|_F$  to denote the Frobenius norm of a matrix  $W$ . Any norm which appears without a subscript will denote the  $\ell_2$  norm for a vector and the Frobenius norm for a matrix.

### 2.1 GENERALIZED LINEAR MODEL FOR CMDPs

Using a linear mapping of the predictors is a simple and ubiquitous approach for modeling contextual/dynamical dependence in sequential decision making problems. Linear models are also well known for being interpretable and explainable, properties which are very valuable in our motivating settings. Similarly, we also utilize this structural simplicity of linearity and model the categorical output space  $(p(\cdot|s, a))$  in a contextual MDP using generalized linear mappings. Specifically, for each pair  $s, a \in \mathcal{S} \times \mathcal{A}$ , there exists a weight matrix  $W_{sa} \in \mathcal{W} \subseteq \mathbb{R}^{S \times d}$  where  $\mathcal{W}$  is a convex set. For any context  $x_k \in \mathbb{R}^d$ , the next state distribution for the pair is specified by a GLM:

$$P_k(\cdot|s, a) = \nabla \Phi(W_{sa} x_k) \quad (2)$$

where  $\Phi(\cdot) : \mathbb{R}^S \rightarrow \mathbb{R}$  is the link function of the GLM<sup>1</sup>. We will assume that this link function is convex which is always the case for a canonical exponential family (Lauritzen, 1996). For rewards, we assume that each mean reward is given by a linear function<sup>2</sup> of the context:  $r_k(s, a) := \theta_{sa}^\top x_k$  where  $\theta \in \Theta \subseteq \mathbb{R}^d$ . In addition, we will make the following assumptions about the link function.

**Assumption 2.1.** *The function  $\Phi(\cdot)$  is  $\alpha$ -strongly convex and  $\beta$ -strongly smooth, that is:*

$$\Phi(v) \geq \Phi(u) + \langle \nabla \Phi(u), v - u \rangle + \frac{\alpha}{2} \|u - v\|_2^2 \quad (3)$$

$$\Phi(v) \leq \Phi(u) + \langle \nabla \Phi(u), v - u \rangle + \frac{\beta}{2} \|u - v\|_2^2 \quad (4)$$

We will see that this assumption is critical for constructing the confidence sets used in our algorithm. We make another assumption about the size of the weight matrices  $W_{sa}^*$  and contexts  $x_k$ :

<sup>1</sup>We abuse the term GLM here as we don’t necessarily consider a complementary exponential family model in eq. (2)

<sup>2</sup> Similar results can be derived for GLM reward functions.

**Assumption 2.2.** For all episodes  $k$ , we have  $\|x_k\|_2 \leq R$  and for all state-action pairs  $(s, a)$ ,  $\|W_{sa}^{(i)}\|_2 \leq B_p$  and  $\|\theta_{sa}\|_2 \leq B_r$ . So, we have  $\|Wx_k\|_\infty \leq B_p R$  for all  $W \in \mathcal{W}$ .

The following two contextual MDP models are special cases of our setting:

**Example 2.3** (Multinomial logit model, Agarwal (2013)). Each next state is sampled from a categorical distribution with probabilities<sup>3</sup>:

$$P_x(s_i|s, a) = \frac{\exp(W_{sa}^{(i)} x)}{\sum_{j=1}^S \exp(W_{sa}^{(j)} x)}$$

The link function for this case can be given as  $\Phi(y) = \log(\sum_{i=1}^S \exp(y_i))$  which can be shown to be strongly convex with  $\alpha = \frac{1}{\exp(BR)S^2}$  and smooth with  $\beta = 1$ .

**Example 2.4** (Linear combination of MDPs, Modi et al. (2018)). Each MDP is obtained by a linear combination of  $d$  base MDPs  $\{(\mathcal{S}, \mathcal{A}, P^i, R^i, H)\}_{i=1}^d$ . Here,  $x_k \in \Delta_{d-1}$ <sup>4</sup>, and  $P_k(\cdot|s, a) := \sum_{i=1}^d x_{ki} P^i(\cdot|s, a)$ . The link function for this can be shown to be:

$$\Phi(y) = \frac{1}{2} \|y\|_2^2$$

which is strongly convex and smooth with parameters  $\alpha = \beta = 1$ . Moreover,  $W_{sa}$  here is the  $S \times d$  matrix containing each next state distribution in a column. We have,  $B_p \leq \sqrt{d}$ ,  $\|W_{sa}\|_F \leq \sqrt{d}$  and  $\|W_{sa} x_k\|_2 \leq 1$ .

### 3 ONLINE ESTIMATES AND CONFIDENCE SET CONSTRUCTION

In order to obtain a no-regret algorithm for our setting, we will follow the popular *optimism in the face of uncertainty* (OFU) approach which relies on the construction of confidence sets for MDP parameters at the beginning of each episode. We focus on deriving these confidence sets for the next state distributions for all state action pairs. We assume that the link function  $\Phi$  and values  $\alpha$ ,  $B$  and  $R$  are known a priori. The confidence sets are constructed and used in the following manner in the OFU template for MDPs: at the beginning of each episode  $k = 1, 2, \dots, K$ :

- For each  $(s, a)$ , compute an estimate of transition distribution  $\hat{P}_k(\cdot|s, a)$  and mean reward  $\hat{r}_k(s, a)$  along with confidence sets  $\mathcal{P}$  and  $\mathcal{R}$  such that  $P_k(\cdot|s, a) \in \mathcal{P}$  and  $r_k(s, a) \in \mathcal{R}$  with high probability.

<sup>3</sup>Without loss of generality, we can set the last row  $W_{sa}^{(S)}$  of the weight matrix to be 0 to avoid an overparameterized system.

<sup>4</sup> $\Delta_{d-1}$  denotes the simplex  $\{x \in \mathbb{R}^d : \|x\|_1 = 1, x \geq 0\}$ .

- Compute an optimistic policy  $\pi_k$  using the confidence sets and unroll a trajectory in  $M_k$  with  $\pi_k$ . Using observed transitions, update the estimates and confidence sets.

Therefore, in the GLM-CMDP setup, estimating transition distributions and reward functions is the same as estimating the underlying parameters  $W_{sa}$  and  $\theta_{sa}$  for each pair  $(s, a)$ . Likewise, any confidence set  $\mathcal{W}_{sa}$  for  $W_{sa}$  ( $\Theta_{sa}$  for  $\theta_{sa}$ ) can be translated into a confidence set of transition distributions.

In our final algorithm for GLM-CMDP, we will use the method from this section for estimating the next state distribution for each state-action pair. The reward parameter  $\theta_{sa}$  and confidence set  $\Theta_{sa}$  is estimated using the linear bandit estimator (Lattimore and Szepesvri (2020), Chap. 20). Here, we solely focus on the following online estimation problem without any reference to the CMDP setup. Specifically, given a link function  $\Phi$ , the learner observes a sequence of contexts  $x_t \in \mathcal{X}$  ( $t = 1, 2, \dots$ ) and a sample  $y_t$  drawn from the distribution  $P_t \equiv \nabla \Phi(W^* x_t)$  over a finite domain of size  $S$ . Here, we use  $W^*$  to denote the true parameter for the given GLM model. The learner's task is to compute an estimate  $W_t$  for  $W^*$  and a confidence set  $\mathcal{W}_t$  after any such  $t$  samples. We frame this as an online optimization problem with the following loss sequence (based on the negative log-likelihood):

$$l_t(W; x_t, y_t) = \Phi(Wx_t) - y_t^\top Wx_t \quad (5)$$

where  $y_t$  is the one-hot representation of the observed sample in round  $t$ . This loss function preserves the strong convexity of  $\Phi$  with respect to  $Wx_t$  and is a proper loss function (Agarwal, 2013):

$$\arg \min_W \mathbb{E}[l_t(W; x_t, y_t)|x_t] = W^* \quad (6)$$

Since our aim is computational and memory efficiency, we carefully follow the Online Newton Step (Hazan et al., 2007) based method proposed for 0/1 rewards with logistic link function in Zhang et al. (2016). While deriving the confidence set in this extension to GLMs, we use properties of categorical vectors in various places in the analysis which eventually saves a factor of  $S$ . The online update scheme is shown in Algorithm 1. Interestingly, note that for tabular MDPs, where  $d = \alpha = 1$  and  $\Phi(y) = \frac{1}{2} \|y\|_2^2$ , with  $\eta = 1$ , we would recover the empirical average distribution as the online estimate. Along with the estimate  $W_{t+1}$ , we can also construct a high probability confidence set as follows:

**Theorem 3.1** (Confidence set for  $W^*$ ). *In Algorithm 1, for all timesteps  $t = 1, 2, \dots$ , with probability at least  $1 - \delta$ , we have:*

$$\|W_{t+1} - W^*\|_{Z_{t+1}} \leq \sqrt{\gamma_{t+1}} \quad (8)$$

---

**Algorithm 1** Online parameter estimation for GLMs
 

---

- 1: **Input:**  $\Phi, \alpha, \eta$
- 2: Set  $W_1 \leftarrow \mathbf{0}, Z_1 \leftarrow \lambda \mathbb{I}_d$
- 3: **for**  $t = 1, 2, \dots$  **do**
- 4:   Observe  $x_t$  and sample  $y_t \sim P_t(\cdot)$
- 5:   Compute new estimate  $W_{t+1}$ :

$$\arg \min_{W \in \mathcal{W}} \frac{\|W - W_t\|_{Z_{t+1}}^2}{2} + \eta \langle \nabla l_t(W_t x_t) x_t^\top, W - W_t \rangle \quad (7)$$

$$\text{where } Z_{t+1} = Z_t + \frac{\eta \alpha}{2} x_t x_t^\top.$$


---

where

$$\gamma_{t+1} = \lambda B^2 + 8\eta B_p R + 2\eta \left[ \left( \frac{4}{\alpha} + \frac{8}{3} B_p R \right) \tau_t + \frac{4}{\alpha} \log \frac{\det(Z_{t+1})}{\det(Z_1)} \right] \quad (9)$$

with  $\tau_t = \log(2[2 \log St]t^2/\delta)$  and  $B = \max_{W \in \mathcal{W}} \|W\|_F$ .

Any upper bound for  $\|W^*\|_F^2$  can be substituted for  $B$  the confidence width in eq (9). The term  $\gamma_t$  depends on the size of the true weight matrix, strong convexity parameter  $\frac{1}{\alpha}$  and the log determinant of the covariance matrix. We will later show that the last term grows at a  $\mathcal{O}(d \log t)$  rate. Therefore, overall  $\gamma_t$  scales as  $\mathcal{O}(S + \frac{d}{\alpha} \log^2 t)$ . The complete proof can be found in Appendix A.

Algorithm 1 only stores the empirical covariance matrix and solves the optimization problem (7) for the current context. Since  $\mathcal{W}$  is convex, this is a tractable problem and can be solved via any off-the-shelf optimizer up to desired accuracy. The total computation time for each context and all  $(s, a)$  pairs is  $\mathcal{O}(\text{poly}(S, A, d))$  with no dependence on  $t$ . Furthermore, we only store  $SA$ -many matrices of size  $S \times d$  and covariance matrices of sizes  $d \times d$ . Thus, both time and memory complexity of the method are bounded by  $\mathcal{O}(\text{poly}(S, A, H, d))$  per episode.

## 4 NO-REGRET ALGORITHMS FOR GLM-CMDP

### 4.1 OPTIMISTIC REINFORCEMENT LEARNING FOR GLM CMDP

In this section, we describe the OFU based online learning algorithm which leverages the confidence sets as described in the previous section. Not surprisingly, our algorithm is similar to the algorithm of Dann et al. (2019) and Abbasi-Yadkori and Neu (2014) and follows the standard format for no-regret bounds in MDPs. In all discussions about CMDPs, we will again use  $x_k \in \mathcal{X}$  to denote the context for episode  $k$  and use Algorithm 1 from the previous section to estimate the corresponding MDP  $M_k$ .

Specifically, for each state-action pair  $(s, a)$ , we use all observed transitions to estimate  $W_{sa}$  and  $\theta_{sa}$ . We compute and store the quantities used in Algorithm 1 for each  $(s, a)$ : we use  $\widehat{W}_{k,sa}$  to denote the parameter estimate for  $W_{sa}$  at the beginning of the  $k^{\text{th}}$  episode. Similarly, we use the notation  $\gamma_{k,sa}$  and  $Z_{k,sa}$  for the other terms. Using the estimate  $\widehat{W}_{k,sa}$  and the confidence set, we compute the confidence interval for  $P_k(\cdot|s, a)$ :

$$\begin{aligned} \xi_{k,sa}^{(p)} &:= \|P_k(\cdot|s, a) - \widehat{P}_k(\cdot|s, a)\|_1 \\ &\leq \beta \sqrt{S} \|W_{sa} - \widehat{W}_{k,sa}\|_{Z_{k,sa}} \|x_k\|_{Z_{k,sa}^{-1}} \\ &\leq \beta \sqrt{S} \sqrt{\gamma_{k,sa}} \|x_k\|_{Z_{k,sa}^{-1}} \end{aligned} \quad (10)$$

where in the definition of  $\gamma_{k,sa}$  we use  $\delta = \delta_p$ . It is again easy to see that for tabular MDPs with  $d = 1$ , we recover a similar confidence interval as used in Jaksch et al. (2010). For rewards, using the results from linear contextual bandit literature (Lattimore and Szepesvri (2020), Theorem 20.5), we use the following confidence interval:

$$\begin{aligned} \xi_{k,sa}^{(r)} &:= |r_k(s, a) - \hat{r}_k(s, a)| \\ &= \underbrace{\left( \sqrt{\lambda d} + \sqrt{\frac{1}{4} \log \frac{\det Z_{k,sa}}{\delta_r^2 \det \lambda I}} \right)}_{:= \zeta_{k,sa}} \|x_k\|_{Z_{k,sa}^{-1}} \end{aligned} \quad (11)$$

In GLM-ORL, we use these confidence intervals to compute an optimistic policy (Lines 9-15). The computed value function is optimistic as we add the total uncertainty as a bonus (Line 11) during each Bellman backup. For any step  $h$ , we clip the optimistic estimate between  $[0, H - h]$  during Bellman backups (Line 13<sup>5</sup>). After unrolling an episode using  $\pi_k$ , we update the parameter estimates and confidence sets for every observed  $(s, a)$  pair.

For any sequence of  $K$  contexts, we can guarantee the following regret bound:

**Theorem 4.1** (Regret of GLM-ORL). *For any  $\delta \in (0, 1)$ , if Algorithm 2 is run with the estimation method 1, then for all  $K \in \mathbb{N}$  and with probability at least  $1 - \delta$ , the regret  $R(K)$  is:*

$$\tilde{\mathcal{O}} \left( \left( \frac{\sqrt{d} \max_{s,a} \|W_{sa}\|_F + d}{\sqrt{\alpha}} \right) \beta S H^2 \sqrt{AK} \log \frac{KHd}{\lambda \delta} \right)$$

If  $\|W^{(i)}\|$  is bounded by  $B_p$ , we get  $\|W_{sa}\|_F^2 \leq S B_p^2$ , whereas, for the linear case (Ex. 2.4),  $\|W_{sa}\|_F^2 \leq \sqrt{d}$ . Substituting the bounds on  $\|W_{sa}\|_F^2$ , we get:

**Corollary 4.2** (Multinomial logit model). *For example 2.3, we have  $\|W\|_F \leq B\sqrt{S}$ ,  $\alpha = \frac{1}{\exp(BR)S^2}$  and  $\beta = 1$ . Therefore, the regret bound of Algorithm 2 is  $\tilde{\mathcal{O}}(dS^3 H^2 \sqrt{AK})$ .*

<sup>5</sup>We use the notation  $a \wedge b$  to denote  $\min(a, b)$  and  $a \vee b$  for  $\max(a, b)$ .

**Corollary 4.3** (Regret bound for linear combination case). *For example 2.4, with  $\|W\|_F \leq \sqrt{d}$ , the regret bound of Algorithm 2 is  $\tilde{O}(dSH^2\sqrt{AK})$ .*

**Algorithm 2** GLM-ORL (GLM Optimistic Reinforcement Learning)

- 
- 1: **Input:**  $\mathcal{S}, \mathcal{A}, H, \Phi, d, \mathcal{W}, \lambda, \delta$
  - 2:  $\delta' = \frac{\delta}{2SA+SH}, \tilde{V}_{k,H+1}(s) = 0 \forall s \in \mathcal{S}, k \in \mathbb{N}$
  - 3: **for**  $k \leftarrow 1, 2, 3, \dots$  **do**
  - 4:   Observe current context  $x_k$
  - 5:   **for**  $s \in \mathcal{S}, a \in \mathcal{A}$  **do**
  - 6:      $\hat{P}_k(\cdot|s, a) \leftarrow \nabla \Phi(\hat{W}_{k,sa} x_k)$
  - 7:      $\hat{r}_k(s, a) \leftarrow \langle \hat{\theta}_{k,sa}, x_k \rangle$
  - 8:     Compute conf. intervals using eqns. (10), (11)
  - 9:   **for**  $h \leftarrow H, H-1, \dots, 1$ , and  $s \in \mathcal{S}$  **do**
  - 10:     **for**  $a \in \mathcal{A}$  **do**
  - 11:        $\varphi = \|\tilde{V}_{k,h+1}\|_\infty \xi_{k,sa}^{(p)} + \xi_{k,sa}^{(r)}$
  - 12:        $\tilde{Q}_{k,h}(s, a) = \hat{P}_{k,sa}^\top \tilde{V}_{k,h+1} + \hat{r}_k(s, a) + \varphi$
  - 13:        $\tilde{Q}_{k,h}(s, a) = 0 \vee (\tilde{Q}_{k,h}(s, a) \wedge V_h^{\max})$
  - 14:        $\pi_{k,h}(s) = \arg \max_a \tilde{Q}_{k,h}(s, a)$
  - 15:        $\tilde{V}_{k,h}(s) = \tilde{Q}_{k,h}(s, \pi_{k,h}(s))$
  - 16:     Unroll a trajectory in  $M_k$  using  $\pi_k$
  - 17:     Update  $\hat{W}_{sa}$  and  $\hat{\theta}_{sa}$  for observed samples.
- 

In Corollary 4.3, the bound is worse by a factor of  $\sqrt{H}$  when compared to the  $\tilde{O}(HS\sqrt{AKH})$  bound of UCRL2 for tabular MDPs ( $d = 1$ ). This factor is incurred while bounding the sum of confidence widths in eq. (28) (in UCRL2 it is  $O(\sqrt{SAKH})$ ).

#### 4.1.1 Proof of Theorem 4.1

We provide the key lemmas used in the analysis with the complete proof in Appendix B.1. Here, we assume that transition probability estimates are valid with probability at least  $1 - \delta_p$  and reward estimates with  $1 - \delta_r$  for all  $(s, a)$  for all episodes. We first begin by showing that the computed policy's value is optimistic.

**Lemma 4.4** (Optimism). *If all the confidence intervals as computed in Algorithm 2 are valid for all episodes  $k$ , then for all  $k$  and  $h \in [H]$  and  $s, a \in \mathcal{S} \times \mathcal{A}$ , we have:*

$$\tilde{Q}_{k,h}(s, a) \geq Q_{k,h}^*(s, a)$$

*Proof.* We show this via an inductive argument. For every episode, the lemma is true trivially for  $H + 1$ . Assume that it is true for  $h + 1$ . For  $h$ , we have:

$$\begin{aligned} & \tilde{Q}_{k,h}(s, a) - Q_{k,h}^*(s, a) \\ &= (\hat{P}_k(s, a)^\top \tilde{V}_{k,h+1} + \hat{r}_k(s, a) + \varphi_{k,h}(s, a)) \wedge V_h^{\max} \\ & \quad - P_k(s, a)^\top V_{k,h+1}^* - r_k(s, a) \end{aligned}$$

We use the fact that when  $\tilde{Q}_{k,h}(s, a) = V_h^{\max}$ , the lemma is trivially satisfied. When  $\tilde{Q}_{k,h}(s, a) < V_h^{\max}$ , we have:

$$\begin{aligned} & \tilde{Q}_{k,h}(s, a) - Q_{k,h}^*(s, a) \\ &= \hat{r}_k(s, a) - r_k(s, a) + \hat{P}_k(s, a)^\top (\tilde{V}_{k,h+1} - V_{k,h+1}^*) \\ & \quad + \varphi_{k,h}(s, a) - (P_k(s, a) - \hat{P}_k(s, a))^\top V_{k,h+1}^* \\ & \geq -|\hat{r}_k(s, a) - r_k(s, a)| + \varphi_{k,h}(s, a) \\ & \quad - \|P_k(s, a) - \hat{P}_k(s, a)\|_1 \|\tilde{V}_{k,h+1}\|_\infty \geq 0 \end{aligned}$$

The last step uses the guarantee on confidence intervals and the inductive assumption for  $h + 1$ . Therefore, the estimated  $Q$ -values are optimistic by induction.  $\square$

Using this optimism guarantee, we can bound the instantaneous regret  $\Delta_k$  in episode  $k$  as:  $V_{k,1}^*(s) - V_{k,1}^{\pi_k}(s) \leq \tilde{V}_{k,1}(s) - V_{k,1}^{\pi_k}(s)$ . With  $\tilde{V}$  as the upper bound, we can bound the total regret with the following Lemma:

**Lemma 4.5.** *In the event that the confidence sets are valid for all episodes, then with probability at least  $1 - SH\delta_1$ , the total regret  $R(K)$  can be bounded by*

$$\begin{aligned} R(K) &\leq SH \sqrt{K \log \frac{6 \log 2K}{\delta_1}} \\ & \quad + \sum_{k=1}^K \sum_{h=1}^H (2\varphi_{k,h}(s_{k,h}, a_{k,h}) \wedge V_h^{\max}) \end{aligned} \quad (12)$$

The proof is given in the appendix. The second term in ineq. (12) can now be bounded as follows:

$$\begin{aligned} & \sum_{k=1}^K \sum_{h=1}^H (2\varphi(s_{k,h}, a_{k,h}) \wedge V_h^{\max}) \\ & \leq \sum_{k=1}^K \sum_{h=1}^H (2\xi_{k,s_{k,h},a_{k,h}}^{(r)} \wedge V_h^{\max}) \\ & \quad + \sum_{k=1}^K \sum_{h=1}^H (2V_{h+1}^{\max} \xi_{k,s_{k,h},a_{k,h}}^{(p)} \wedge V_h^{\max}) \end{aligned} \quad (13)$$

We ignore the reward estimation error in eq. (13) as it leads to lower order terms. The second expression can be again bounded as follows:

$$\begin{aligned} & \sum_{k=1}^K \sum_{h=1}^H (2V_{h+1}^{\max} \xi_{k,s_{k,h},a_{k,h}}^{(p)} \wedge V_h^{\max}) \\ & \leq 2 \sum_{k,h} V_h^{\max} \left( 1 \wedge \beta \sqrt{S\gamma_k(s_{k,h}, a_{k,h})} \|x_k\|_{Z_{k,sa,h}^{-1}} \right) \end{aligned} \quad (14)$$

Using Lemma B.4, we see that

$$\begin{aligned}\gamma_k(s, a) &:= f_\Phi(k, \delta_p) + \frac{8\eta}{\alpha} \log \frac{\det(Z_{k,sa})}{\det(Z_{1,sa})} \\ &\leq \frac{\eta\alpha}{2S} + f_\Phi(KH, \delta_p) + \frac{8\eta}{\alpha} \log \frac{\det(Z_{K+1,sa})}{\det(Z_{1,sa})} \\ &\leq \frac{\eta\alpha}{2S} + f_\Phi(KH, \delta_p) + \frac{8\eta d}{\alpha} \log \left( 1 + \frac{KHR^2}{\lambda d} \right)\end{aligned}$$

We use  $f_\Phi(k, \delta_p)$  to refer to the  $Z_k$  independent terms in eq. (9). Setting  $\bar{\gamma}_K$  to the last expression guarantees that  $\frac{2S\bar{\gamma}_K}{\eta\alpha} \geq 1$ . We can now bound the term in eq. (14) as:

$$\begin{aligned}2\beta V_1^{\max} &\sqrt{\frac{2S\bar{\gamma}_K}{\eta\alpha}} \sum_{k,h} \left( 1 \wedge \sqrt{\frac{\eta\alpha}{2}} \|x_k\|_{Z_{k,sa,h}^{-1}} \right) \\ &\leq 2\beta V_1^{\max} \sqrt{\frac{2S\bar{\gamma}_K KH}{\eta\alpha}} \sqrt{\sum_{k,h} \left( 1 \wedge \frac{\eta\alpha}{2} \|x_k\|_{Z_{k,sa,h}^{-1}}^2 \right)}\end{aligned}\quad (15)$$

Ineq. (15) follows by using Cauchy-Schwarz inequality. Finally, by using Lemma B.4 in Appendix B.1, we can bound the term as

$$\begin{aligned}&\sum_{k=1}^K \sum_{h=1}^H (2V_{h+1}^{\max} \zeta_{k,s_k,h,a_k,h}^{(p)} \wedge V_h^{\max}) \\ &= 4\beta V_1^{\max} \sqrt{\frac{2S\bar{\gamma}_K KH}{\eta\alpha}} \sqrt{2HSAd \log \left( 1 + \frac{KHR^2}{\lambda d} \right)}\end{aligned}$$

Now, after setting the failure probabilities  $\delta_1 = \delta_p = \delta_r = \delta/(2SA + SH)$  and taking a union bound over all events, we get the total failure probability as  $\delta$ . Therefore, with probability at least  $1 - \delta$ , we can bound the regret of GLM-ORL as

$$R(K) = \tilde{\mathcal{O}} \left( \left( \frac{\sqrt{d} \max_{s,a} \|W_{sa}^*\|_F}{\sqrt{\alpha}} + \frac{d}{\alpha} \right) \beta SH^2 \sqrt{AK} \right)$$

where  $\max_{s,a} \|W_{sa}^*\|_F$  is replaced by the problem dependent upper bound assumed to be known a priori.<sup>6</sup>

#### 4.1.2 Mistake bound for GLM-ORL

The regret analysis shows that the total value loss suffered by the agent is sublinear in  $K$ , and therefore, goes to 0 on average. However, this can still lead to infinitely many episodes where the sub-optimality gap is larger than a desired threshold  $\epsilon$ , given that it occurs relatively infrequently. It is still desirable, for practical purposes, to

<sup>6</sup>An improved dependence on  $\sum_{s,a} \|W_{sa}^*\|_F$  can be obtained instead of  $S \max_{s,a} \|W_{sa}^*\|_F$  in the regret bound.

analyze how frequently can the agent incur such mistakes. Here, a mistake is defined as an episode in which the value of the learner's policy  $\pi_k$  is not  $\epsilon$ -optimal, i.e.,  $V_k^* - V_k^{\pi_k} \geq \epsilon$ . In our setting, we can show the following result.

**Theorem 4.6** (Bound on the number of mistakes). *For any number of episodes  $K$ ,  $\delta \in (0, 1)$  and  $\epsilon \in (0, H)$ , with probability at least  $1 - \delta$ , the number of episodes where GLM-ORL's policy  $\pi_k$  is not  $\epsilon$ -optimal is bounded by*

$$\mathcal{O} \left( \frac{dS^2 AH^5 \log(KH)}{\epsilon^2} \left( \frac{d \log^2(KH)}{\alpha} + S \right) \right)$$

ignoring  $\mathcal{O}(\text{poly}(\log \log KH))$  terms.

We defer the proof to Appendix C. Note that this term depends poly-logarithmically on  $K$  and therefore increases with time. The algorithm doesn't need to know the value of  $\epsilon$  and result holds for all  $\epsilon$ . This differs from the standard mistake bound style PAC guarantees where a finite upper bound is given. Dann et al. (2019) argued that this is due to the non-shrinking nature of the constructed confidence sets. As such, showing such a result for CMDPs requires a non-trivial construction of confidence sets and falls beyond the scope of this paper.

## 4.2 RANDOMIZED EXPLORATION FOR GLM-CMDP

Empirical investigations in bandit and MDP literature has shown that optimism based exploration methods typically over-explore, often resulting in sub-optimal empirical performance. In contrast, Thompson sampling based methods which use randomization during exploration have been shown to have an empirical advantage with slightly worse regret guarantees. Recently, Russo (2019) showed that even with such randomized exploration methods, one can achieve a worst-case regret bound instead of the typical Bayesian regret guarantees. In this section, we show that the same is true for GLM-CMDP where a randomized reward bonus can be used for exploration. We build upon their work to propose an RLSVI style method (Algorithm 3) and analyze its expected regret. The main difference between Algorithm 2 and Algorithm 3 is that instead of the fixed bonus  $\varphi$  (Line 11) in the former, GLM-RLSVI samples a random reward bonus in Line 12 for each  $(s, a)$  from the distribution  $N(0, HS\varphi^2)$ . The variance term  $\varphi$  is set to a sufficiently high value, such that, the resulting policy is optimistic with constant probability. We use a slightly modified version of the confidence sets as follows:

$$\begin{aligned}\bar{\xi}_{k,sa}^{(p)} &:= 2 \wedge \left( \beta \sqrt{S} \sqrt{\gamma_{k,sa}} \|x_k\|_{Z_{k,sa}^{-1}} \right) \\ \bar{\xi}_{k,sa}^{(r)} &:= B_r R \wedge \left( \tau_{k,sa} \|x_k\|_{Z_{k,sa}^{-1}} \right)\end{aligned}$$

---

**Algorithm 3** GLM-RLSVI

---

- 1: **Input:**  $\mathcal{S}, \mathcal{A}, H, \Phi, d, \mathcal{W}, \lambda$
- 2:  $\bar{V}_{k,H+1}(s) = 0 \forall s \in \mathcal{S}, k \in \mathbb{N}$
- 3: **for**  $k \leftarrow 1, 2, 3, \dots$  **do**
- 4:   Observe current context  $x_k$
- 5:   **for**  $s \in \mathcal{S}, a \in \mathcal{A}$  **do**
- 6:      $\hat{P}_k(\cdot|s, a) \leftarrow \nabla \Phi(\widehat{W}_{k,sa} x_k)$
- 7:      $\hat{r}_k(s, a) \leftarrow \langle \hat{\theta}_{k,sa}, x_k \rangle$
- 8:     Compute conf. intervals using eqns. (10), (11)
- 9:   **for**  $h \leftarrow H, H-1, \dots, 1$ , and  $s \in \mathcal{S}$  **do**
- 10:    **for**  $a \in \mathcal{A}$  **do**
- 11:      $\varphi = (H-h)\bar{\xi}_{k,sa}^{(p)} + \bar{\xi}_{k,sa}^{(r)}$
- 12:     Draw sample  $b_{k,h}(s, a) \sim N(0, SH\varphi)$
- 13:      $\bar{Q}_{k,h}(s, a) = \hat{P}_{k,sa}^\top \bar{V}_{k,h+1} + \hat{r}_k(s, a) + b_{k,h}(s, a)$
- 14:      $\pi_{k,h}(s) = \arg \max_a \bar{Q}_{k,h}(s, a)$
- 15:      $\bar{V}_{k,h}(s) = \bar{Q}_{k,h}(s, \pi_{k,h}(s))$
- 16:    Unroll a trajectory in  $M_k$  using  $\pi_k$ .
- 17:    Update  $\widehat{W}_{sa}$  and  $\hat{\theta}_{sa}$  for observed samples.

---

The algorithm, thus, generates exploration policies by using perturbed rewards for planning. Similarly to Russo (2019), we can show the following bound for the expected regret incurred by GLM-RLSVI:

**Theorem 4.7.** *For any contextual MDP with given link function  $\Phi$ , in Algorithm 3, if the MDP parameters for  $M_k$  are estimated using Algorithm 1, with reward bonuses  $b_{k,h}(s, a) \sim N(0, SH\varphi_{k,h}(s, a))$  where  $\varphi_{k,h}(s, a)$  is defined in Line. 11, the algorithm satisfies:*

$$\begin{aligned} \bar{R}(K) &= \mathbb{E} \left[ \sum_{k=1}^K V_k^* - V_k^{\pi_k} \right] \\ &= \tilde{\mathcal{O}} \left( \left( \frac{\sqrt{d} \max_{s,a} \|W_{sa}^*\|_F}{\sqrt{\alpha}} + \frac{d}{\alpha} \right) \beta \sqrt{H^7 S^3 AK} \right) \end{aligned}$$

The proof of the regret bound is given in Appendix B.2. Our regret bound is again worse by a factor of  $\sqrt{H}$  when compared to the  $\tilde{\mathcal{O}}(H^3 S^{3/2} \sqrt{AK})$  bound from Russo (2019) for the tabular case. Therefore, such randomized bonus based exploration algorithms can also be used in the CMDP framework with similar regret guarantees as the tabular case.

## 5 LOWER BOUND FOR GLM CMDP

In this section, we show a regret lower bound by constructing a family of hard instances for the GLM-CMDP problem. We build upon the construction of Osband and Van Roy (2016) and Jaksch et al. (2010) for the analysis<sup>7</sup>:

<sup>7</sup>The proof is deferred to the appendix due to space constraints.

**Theorem 5.1.** *For any algorithm  $\mathbf{A}$ , there exists CMDP's with  $S$  states,  $A$  actions, horizon  $H$  and  $K \geq dSA$  for logit and linear combination case, such that the expected regret of  $\mathbf{A}$  (for any sequence of initial states  $\in \mathcal{S}^K$ ) after  $K$  episodes is:*

$$\mathbb{E}[R(K; \mathbf{A}, M_{1:K}, s_{1:K})] = \Omega(H\sqrt{dSAK})$$

The lower bound has the usual dependence on MDP parameters in the tabular MDP case, with an additional  $\mathcal{O}(\sqrt{d})$  dependence on the context dimension. Thus, our upper bounds have a gap of  $\mathcal{O}(H\sqrt{dS})$  with the lower bound even in the arguably simpler case of Example 2.4.

## 6 IMPROVED CONFIDENCE SETS FOR STRUCTURED SPACES

In Section 3, we derived confidence sets for  $W^*$  for the case when it lies in a bounded set. However, in many cases, we have additional prior knowledge about the problem in terms of possible constraints over the set  $\mathcal{W}$ . For example, consider a healthcare scenario where the context vector contains the genomic encoding of the patient. For treating any ailment, it is fair to assume that the patient's response to the treatment and the progression in general depends on a few genes rather than the entire genome which suggests a sparse dependence of the transition model on the context vector  $x$ . In terms of the parameter  $W^*$ , this translates as complete columns of the matrix being zeroed out for the irrelevant indices. Thus, it is desirable to construct confidence sets which take this specific structure into account and give more problem dependent bounds.

In this section, we show that it is possible to convert a generic regret guarantee of an online learner to a confidence set. If the online learner adapts to the structure of  $\mathcal{W}$ , we would get the aforementioned improvement. The conversion proof presented here is reminiscent of the techniques used in Abbasi-Yadkori et al. (2012) and Jun et al. (2017) with close resemblance to the latter. For this section, we use  $X_t$  to denote the  $t \times d$  shaped matrix with each row as  $x_i$  and  $C_t$  as  $t \times S$  shaped matrix with each row  $i$  being  $(W_i x_i)^\top$ <sup>8</sup>. Also, set  $\bar{W}_t := Z_{t+1}^{-1} X_t^\top C_t$ . Using a similar notation as before, we can give the following guarantee.

**Theorem 6.1** (Multinomial GLM Online-to-confidence set conversion). *Assume that loss function  $l_i$  defined in eq. (5) is  $\alpha$ -strongly convex with respect to  $Wx$ . If an online learning oracle takes in the sequence  $\{x_i, y_i\}_{i=1}^t$ , and produces outputs  $\{W_i\}_{i=1}^t$  for an input sequence*

<sup>8</sup>We again solely consider the estimation problem for a single  $(s, a)$  pair and study a  $t$ -indexed online estimation problem.

$\{x_i, y_i\}_{i=1}^t$ , such that:

$$\sum_{i=1}^t l_i(W_i) - l_i(W) \leq B_t \quad \forall W \in \mathcal{W}, t > 0,$$

then with  $\bar{W}_t$  as defined above, with probability at least  $1 - \delta$ , for all  $t \geq 1$ , we have

$$\|W^* - \bar{W}_t\|_{Z_{t+1}}^2 \leq \gamma_t$$

where  $\gamma_t := \gamma'_t(B_t) + \lambda B^2 S - (\|C_t\|_F^2 - \langle \bar{W}_t, X_t^\top C_t \rangle)$ ,

$$\gamma'_t(B_t) := 1 + \frac{4}{\alpha} B_t + \frac{8}{\alpha^2} \log \left( \frac{1}{\delta} \sqrt{4 + \frac{8B_t}{\alpha} + \frac{16}{\alpha^4 \delta^2}} \right).$$

The complete proof can be found in Appendix E. Note that, all quantities required in the expression  $\gamma_t$  can be incrementally computed. The required quantities are  $Z_t$  and  $Z_t^{-1}$  along with  $X_t^\top C_t$  which are incrementally updated with  $O(\text{poly}(S, d))$  computation. Also, we note that this confidence set is meaningful when  $B_t$  is poly-logarithmic in  $t$  which is possible for strongly convex losses as shown in Jun et al. (2017). The dependence on  $S$  and  $d$  is the same as the previous construction, but the dependence on the strong convexity parameter is worse.

**Column sparsity of  $W^*$**  Similar to sparse stochastic linear bandit, as discussed in Abbasi-Yadkori et al. (2012), one can use an online learning method with the group norm regularizer ( $\|W\|_{2,1}$ ). Therefore, if an efficient online no-regret algorithm has an improved dependence on the sparsity coefficient  $p$ , we can get an  $O(\sqrt{p} \log d)$  size confidence set. This will improve the final regret bound to  $\tilde{O}(\sqrt{pdT})$  as observed in the linear bandit case. To our knowledge, even in the sparse adversarial linear regression setting, obtaining an efficient and sparsity aware regret bound is an open problem.

## 7 DISCUSSION

Here, we discuss the obtained regret guarantees for our methods along with the related work. Further, we outline the algorithmic/analysis components which are different from the tabular MDP case and lead to interesting open questions for future work.

### 7.1 RELATED WORK

**Contextual MDP** To our knowledge, Hallak et al. (2015) first used the term contextual MDPs and studied the case when the context space is finite and the context is not observed during interaction. They propose CECE, a clustering based learning method and analyze its regret. Modi et al. (2018) generalized the CMDP framework and

proved the PAC exploration bounds under smoothness and linearity assumptions over the contextual mapping. Their PAC bound is incomparable to our regret bound as a no-regret algorithm can make arbitrarily many mistakes  $\Delta_k \geq \epsilon$  as long as it does so sufficiently less frequently.

Our work can be best compared with Abbasi-Yadkori and Neu (2014) and Dann et al. (2019) who propose regret minimizing methods for CMDPs. Abbasi-Yadkori and Neu (2014) consider an online learning scenario where the values  $p_k(s'|s, a)$  are parameterized by a GLM. The authors give a no-regret algorithm which uses confidence sets based on Abbasi-Yadkori et al. (2012). However, their next state distributions are not normalized which leads to invalid next state distributions. Due to these modelling errors, their results cannot be directly compared with our analysis. Even if we ignore their modelling error, in the linear combination case, we get an  $\tilde{O}(S\sqrt{A})$  improvement. Similarly, Dann et al. (2019) proposed an OFU based method `ORLC-SI` for the linear combination case. Their regret bound is  $\tilde{O}(\sqrt{S})$  worse than our bound for `GLM-ORL`. In addition, the work also showed that obtaining a finite mistake bound guarantees for such CMDPs requires a non-trivial and novel confidence set construction. In this paper, we show that a  $\text{polylog}(K)$  mistake bound can still be obtained. For a quick comparison, Table 1 shows the results from the two papers.

**(Generalized) linear bandit** Our reward model is based on the (stochastic) linear bandit problem first studied by Abe et al. (2003). Our work borrows key results from Abbasi-Yadkori et al. (2011) for both the reward estimator and during analysis for the GLM case. Extending the linear bandit problem, Filippi et al. (2010) first proposed the generalized linear contextual bandit setting and showed a  $\mathcal{O}(d\sqrt{T})$  regret bound. We, however, leverage the approach from Zhang et al. (2016) and Jun et al. (2017) who also studied the logistic bandit and GLM Bernoulli bandit case. We extend their proposed algorithm and analysis to a generic categorical GLM setting. Consequently, our bounds also incur a dependence on the strong convexity parameter  $\frac{1}{\alpha}$  of the GLM which was recently shown to be unavoidable by Foster et al. (2018) for proper learning in the closely related online logistic regression problem.

**Regret analysis in tabular MDPs** Auer and Ortner (2007) first proposed a no-regret online learning algorithm for average reward infinite horizon MDPs, and the problem has been extensively studied afterwards. More recently, there has been an increased focus on fixed horizon problems where the gap between the upper and lower bounds has been effectively closed. Azar et al. (2017) and Dann et al. (2019), both provide optimal regret guarantees

Algorithm	$R^{\text{Linear}}(K)$	$R^{\text{Logit}}(K)$	$P_x(\cdot s, a)$ <b>normalized</b>
Algorithm 1 (Abbasi-Yadkori and Neu, 2014)	$\tilde{\mathcal{O}}(dH^3S^2A\sqrt{K})$	$\times$	$\times$
ORLC-SI (Dann et al., 2019)	$\tilde{\mathcal{O}}(dH^2S^{3/2}\sqrt{AK})$	$\times$	$\times$
GLM-ORL (this work)	$\tilde{\mathcal{O}}(dH^2S\sqrt{AK})$	$\tilde{\mathcal{O}}(dH^2S^3\sqrt{AK})$	$\checkmark$

Table 1: Comparison of regret guarantees for CMDPs. Last column denotes whether the transition dynamics  $P_x(\cdot|s, a)$  are normalized in the model or not.

( $\tilde{\mathcal{O}}(H\sqrt{SAK})$ ) for tabular MDPs. Another series of papers (Osband et al., 2013, 2016; Russo et al., 2018) study Thompson sampling based randomized exploration methods and prove Bayesian regret bounds. Russo (2019) recently proved a worst case regret bound for RLSVI-style methods (Osband et al., 2016). The algorithm template and proof structure of GLM-RLSVI is borrowed from their work.

**Feature-based linear MDP** Yang and Wang (2019a) consider an RL setting where the MDP transition dynamics are low-rank. Specifically, given state-action features  $\phi(s, a)$ , they assume a setting where  $p(s'|s, a) := \sum_{i=1}^d \phi_i(s, a)\nu_i(s')$  where  $\nu_i$  are  $d$  base distributions over the state space. This structural assumption guarantees that the  $Q^\pi(s, a)$  value functions are linear in the state-action features for every policy. Yang and Wang (2019b); Jin et al. (2019) have recently proposed regret minimizing algorithms for the linear MDP setting. Although, their algorithmic structure is similar to ours (linear bandit based bonuses), the linear MDP setting is only superficially related to CMDP. In our case, the value functions are not linear in the contextual features for every policy and/or context. Thus, the two MDP frameworks and their regret analyses are incomparable.

## 7.2 CLOSING THE REGRET GAP

From the lower bound in Section 5, it is clear that the regret bound of GLM-ORL is sub-optimal by a factor of  $\tilde{\mathcal{O}}(H\sqrt{dS})$ . As mentioned previously, for episodic MDPs, Azar et al. (2017) and Dann et al. (2019) propose minimax-optimal algorithms. The key technique in these analyzes is to directly build a confidence interval for the value functions and use a refined analysis using empirical Bernstein bonuses based on state-action visit counts saves a factor of  $\mathcal{O}(\sqrt{HS})$ . In our case, we use a Hoeffding style bonus for learning the next state distributions to derive confidence sets for the value function. Further, the value functions in GLM-CMDP do not have a nice structure as a function of the context variable and therefore, these techniques do not trivially extend to CMDPs. Similarly, the dependence on context dimension  $d$  is typically

resolved by dividing the samples into phases which make them statistically independent (Auer, 2002; Chu et al., 2011; Li et al., 2017). However, for CMDPs, these filtering steps cannot be easily performed while ensuring long horizon optimistic planning. Thus, tightening the regret bounds for CMDPs is highly non-trivial and we leave this for future work.

## 8 CONCLUSION AND FUTURE WORK

In this paper, we have proposed optimistic and randomized no-regret algorithms for contextual MDPs which are parameterized by generalized linear models. We provide an efficient online Newton step (ONS) based update method for constructing confidence sets used in the algorithms. This work also outlines potential future directions: close the regret gap for tabular CMDPs, devise an efficient and sparsity aware regret bound and investigate whether a near-optimal mistake and regret bound can be obtained simultaneously. Lastly, extension of the framework to non-tabular MDPs is an interesting problem for future work.

## Acknowledgements

AM thanks Satinder Singh and Alekh Agarwal for helpful discussions. This work was supported in part by a grant from the Open Philanthropy Project to the Center for Human-Compatible AI, and in part by NSF grant CAREER IIS-1452099. AT would like to acknowledge the support of a Sloan Research Fellowship.

## References

- Abbasi-Yadkori, Y. and Neu, G. (2014). Online learning in MDPs with side information. *arXiv preprint arXiv:1406.6812*.
- Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.
- Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. (2012).

- Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9.
- Abe, N., Biermann, A. W., and Long, P. M. (2003). Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293.
- Agarwal, A. (2013). Selective sampling algorithms for cost-sensitive multiclass prediction. In *International Conference on Machine Learning*, pages 1220–1228.
- Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.
- Auer, P. and Ortner, R. (2007). Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 49–56.
- Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org.
- Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214.
- Dann, C., Li, L., Wei, W., and Brunskill, E. (2019). Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516.
- Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594.
- Foster, D. J., Kale, S., Luo, H., Mohri, M., and Sridharan, K. (2018). Logistic regression: The importance of being improper. In *Conference On Learning Theory*, pages 167–208.
- Hallak, A., Di Castro, D., and Mannor, S. (2015). Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*.
- Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192.
- Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600.
- Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2019). Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*.
- Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. (2017). Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, pages 99–109.
- Lattimore, T. and Szepesvri, C. (2020). *Bandit Algorithms*. Cambridge University Press.
- Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.
- Li, L., Lu, Y., and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080.
- Modi, A., Jiang, N., Singh, S., and Tewari, A. (2018). Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pages 597–618.
- Osband, I., Russo, D., and Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011.
- Osband, I. and Van Roy, B. (2016). On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*.
- Osband, I., Van Roy, B., and Wen, Z. (2016). Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386.
- Russo, D. (2019). Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, pages 14410–14420.
- Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. (2018). A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.
- Yang, L. and Wang, M. (2019a). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004.
- Yang, L. F. and Wang, M. (2019b). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*.
- Zhang, L., Yang, T., Jin, R., Xiao, Y., and Zhou, Z.-h. (2016). Online stochastic linear optimization under one-bit feedback. In *International Conference on Machine Learning*, pages 392–401.