Delayed Feedback in Generalised Linear Bandits Revisited

Benjamin Howson
Imperial College London

Ciara Pike-Burke
Imperial College London

Sarah Filippi
Imperial College London

Abstract

The stochastic generalised linear bandit is a well-understood model for sequential decision-making problems, with many algorithms achieving near-optimal regret guarantees under immediate feedback. However, the stringent requirement for immediate rewards is unmet in many real-world applications where the reward is almost always delayed. We study the phenomenon of delayed rewards in generalised linear bandits in a theoretical manner. We show that a natural adaptation of an optimistic algorithm to the delayed feedback setting can achieve regret of $\widetilde{\mathcal{O}}(d\sqrt{T}+d^{3/2}\mathbb{E}[\tau]\,)$, where $\mathbb{E}[\tau]$ denotes the expected delay, d is the dimension and T is the time horizon. This significantly improves upon existing approaches for this setting where the best known regret bound was $\widetilde{\mathcal{O}}(\sqrt{dT}\sqrt{d+\mathbb{E}[\tau]}\,)$. We verify our theoretical results through experiments on simulated data.

1 INTRODUCTION

Recently, bandit algorithms have found application in areas from dynamic pricing and healthcare to finance and recommender systems with great success (Misra et al., 2019; Durand et al., 2018; Shen et al., 2015; McInerney et al., 2018). There are many formulations of bandit problems. One of these is the stochastic generalised linear bandit, which captures a wide class of problems, such as when the rewards are counts, binary values or can take any real-valued number. The generalised linear bandit problem proceeds in rounds, where in each round, a learner must choose from a set of possible actions. After selecting an action, the learner receives feedback from the environment in the form of a reward which stochastically depends on the inner product of the action and some unknown parameter vector. The goal of the learner is to maximise their expected cumulative reward.

There are many provably efficient algorithms for the generalised linear bandit (Filippi et al., 2010; Abbasi-Yadkori et al., 2011; Li et al., 2017; Faury et al., 2020). Unfortunately, these existing algorithms require immediate feedback from the environment. This strict requirement for immediate rewards often goes unmet in practice. For example, in many recommender systems, the user must provide feedback to the learner while operating on a very different time scales; e.g. the learner can make thousands of recommendations per second, whereas a user may take several minutes to a couple of days to respond to the recommendation, if at all (Chapelle, 2014). Alternatively, practitioners might want to optimise for a longer-term measure of success (Han and Arndt, 2021), in which case the reward is not observable or even defined immediately. Delayed feedback also arises in clinical trials due to the time-consuming task of obtaining medical feedback and because patients do not respond to their prescribed treatment immediately.

In all the above settings, the reward for any given action returns at an unknown time in the future. Meanwhile, the learner must continue operating in the environment without feedback from many of their past choices. A natural model for this phenomenon is to introduce a random delay between taking action and receiving the reward. However, the delays pose significant theoretical challenges because standard tools for analysing bandit algorithms rely on utilising immediate feedback to reduce the uncertainty in the learner's estimation. Under delayed feedback, it is unclear how long the learner will have to wait before they gain information about the quality of an action, which hinders their future decision-making abilities.

These challenges have led to the development of algorithms specifically for delayed feedback in generalised linear bandits. However, to the best of our knowledge, these existing algorithms require *a-priori* knowledge of the expected delay (along with other quantities), strong assumptions on the delay distribution, restrictive assumptions on the action sets, or any combination thereof. Moreover, the best regret bound achievable by these algorithms is $\widetilde{\mathcal{O}}(\sqrt{dT}\sqrt{d+\mathbb{E}[\tau]})$ where T is the number of rounds, d is the dimension of the unknown parameter, and $\mathbb{E}[\tau]$ is

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

the expected delay. This result suggests that as the horizon increases, the impact of the delayed feedback will increase. This result is counter intuitive and starkly differs from the results in the K-armed bandit setting where the impact of the delay is independent of the horizon (Joulani et al., 2013). In this paper, we prove that a simple algorithm based on optimism can achieve a regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T} + d^{3/2}\mathbb{E}[\tau])$. This improves the penalty for delayed feedback from $\sqrt{dT\mathbb{E}[\tau]}$ in prior work to $d^{3/2}\mathbb{E}[\tau]$, separating the delay penalty from the horizon.

1.1 Related Work

The multi-armed bandit literature covers stochastically delayed feedback extensively. Both Joulani et al. (2013) and Mandel et al. (2015) propose queue-based approaches to adapt existing K-armed bandit algorithms to delayed feedback, each proving that the regret bound of the chosen algorithm only increases by an additive factor of $\mathbb{E}[\tau]$, where $\mathbb{E}[\tau]$ denotes the expected delay between playing an action and observing the corresponding reward. Pike-Burke et al. (2018) study another version of delayed feedback, where the rewards from various rounds are not only delayed but also aggregated. Vernade et al. (2017) consider the setting of delayed conversions, where actions associated with long delays can have censored feedback.

Comparatively, fewer theoretical results quantify the impact of delays beyond K-armed bandits. Vernade et al. (2020) consider a Bernoulli bandit with censored rewards whose expected value is linear in some unknown parameter vector. Combining Bernoulli rewards with delays makes it impossible to distinguish between a reward of zero and a delayed reward. Thus, the challenges they face are different to ours. Nevertheless, they deal with the delays by inflating the exploration bonus and handle the censoring by introducing a windowing parameter that sets rewards taking too long to return equal to zero. Dudik et al. (2011) develop a policy elimination algorithm capable of handling contextual information and prove a regret bound of the form $\widetilde{\mathcal{O}}(\sqrt{KT} + \sqrt{K}\tau_{\text{const}})$, where K is the number of actions and τ_{const} is a constant delay between playing an action and observing the corresponding reward. However, they remark that their algorithm is challenging to implement, requires perfect knowledge of the distribution of the contextual information, and needs a-priori knowledge of the constant delay.

Zhou et al. (2019) and follow-up work by Blanchet et al. (2020), that analyses that same algorithm, consider learning in the same setting as us. They propose an optimistic algorithm that inflates the exploration bonus by the square root of the number of missing feedbacks. They do this to account for the uncertainty arising from the missing rewards. Combining this bonus with an elegant argument allows them to use standard theoretical tools to handle the leading-order terms, namely the elliptical potential lemma. This lemma has found applications in the analysis of many linear bandit algorithms and is provably tight (Carpentier et al., 2020). However, due to the delay-dependent bonuses, their arguments lead to a multiplicative increase in the regret of the form $\widetilde{\mathcal{O}}(d\sqrt{T} + \sqrt{dT(\mathbb{E}[\tau] + M_\tau)})$, where M_τ is a known non-negative delay-dependent constant beyond which the delays have tails that are as heavy as (or lighter than) the exponential distribution. However, this algorithm requires prior knowledge of the expected delay and M_τ (along with other quantities). This theoretical result suggests that the impact of the delayed feedback increases with the horizon, which does not align with the intuition that the delays become irrelevant once the learner has observed enough feedback to obtain a "good" estimate of the underlying expected reward function.

1.2 Contributions

In this paper, we present a natural approach based on optimism that does not require any prior knowledge of the delays and achieves regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T} + d^{3/2}\mathbb{E}[\tau]\,)$, up to problem-specific constants. This result significantly improves upon the best-known theoretical results for generalized linear bandits with delayed feedback, whose regret bound is $\widetilde{\mathcal{O}}(d\sqrt{T} + \sqrt{dT\mathbb{E}[\tau]}\,)$. Further, our results align with what is seen in the K-armed bandit setting, where the delays only impact the worst-case performance by an additive penalty involving the expected delay (Joulani et al., 2013), and not the horizon T.

In contrast to prior work, we forgo the period of forced exploration which is present in many generalised linear bandit algorithms (Filippi et al., 2010; Li et al., 2017). Our algorithm is optimistic and constructs optimistic estimates using only *observations that have returned*. To do this, we develop delay-adapted confidence sets and prove that these are valid. Although this algorithm is natural, proving regret bounds for it is somewhat involved. In particular, the presence of delayed feedback obscures how selecting a suboptimal action in round t will improve the estimation in future rounds. To overcome these issues we provide a novel analysis centered around an elliptical potential lemma for delayed feedback, which may be of independent interest for bandit algorithms with complex feedback structures. We show that this technique leads to the stated regret bound. We also validate our theoretical findings experimentally in some simulated environments.

2 PROBLEM FORMULATION

The stochastic generalised linear bandit problem considers learning in an environment where the expected reward is a known function of the dot product between the action and the unknown parameter vector. Letting $X_t \in \mathcal{A}_t \subset \mathbb{R}^d$ and $Y_t \in \mathbb{R}$ be the action and reward associated with the t-th

round, we assume that the conditional distribution of the reward given the action belongs to the exponential family:

$$f\left(Y_t \mid X_t\right) \propto \exp\left(\frac{Y_t X_t^T \theta^* - b\left(X_t^T \theta^*\right)}{a\left(\phi\right)}\right) \quad (1)$$

where $\theta^* \in \mathbb{R}^d$ is an unknown parameter vector; a and b are known distribution specific functions; and ϕ is a known constant that is often referred to as the *dispersion parameter*. For distributions belonging to the exponential family, one can verify that:

$$\mathbb{E}\left[Y_t \mid X_t\right] = \mu\left(X_t^T \theta^*\right) = \dot{b}\left(X_t^T \theta^*\right).$$

Here, $\mu(\cdot)$ is a strictly increasing *link function* that relates the inner product of the action vector and the unknown parameter to the expected reward. For example, if the rewards are normally distributed, $\mu(z) = z$ and we recover the standard linear model. If the rewards are Bernoulli, then $\mu(z) = 1/(1+\exp(-z))$ and we have a logistic regression model.

In the stochastic setting, the learner selects an action $X_t \in \mathcal{A}_t \subset \mathbb{R}^d$ and receives noisy observations of the unknown expected reward function of the form $Y_t \sim f\left(Y_t \mid X_t\right)$ where

$$\eta_t \coloneqq Y_t - \mu\left(X_t^T \theta^*\right)$$

is the noise and is zero-mean conditional on past decisions and rewards. Section 2.2 formally states the assumptions we make on the link function and the noise.

The ultimate goal of the learner in the generalised linear bandit setting is to minimise the regret. Intuitively, this compares the expected reward of the action selected by the learner to the action with the highest expected reward. Mathematically, we define the regret of an algorithm in the generalised linear bandit setting as follows:

$$\hat{R}_T = \sum_{t=1}^{T} \mu\left(\langle X_t^*, \theta^* \rangle\right) - \mu\left(\langle X_t, \theta^* \rangle\right) \coloneqq \sum_{t=1}^{T} \hat{r}_t \quad (2)$$

where $X_t^* = \arg\max_{x \in \mathcal{A}_t} \{\mu\left(\langle x, \theta^* \rangle\right)\}$ is the action in the decision set $\mathcal{A}_t$ maximising the expected reward in the t-th round.

2.1 Delayed Feedback Learning Setting

Let $\tau_t \in [0, \infty)$ denote the random delay associated with the decision made in the t-th round. Then, the sequential decision-making procedure for generalized linear bandits under stochastically delayed feedback is as follows. For $t \in \{1, \cdots, T\}$:

1. The learner receives a decision set containing the context vectors: $\mathcal{A}_t \subset \mathbb{R}^d$ where $\|x\|_2 \leq 1$ for all $x \in \mathcal{A}_t$.
2. The learner selects a d-dimensional feature vector from the decision set: $X_t \in \mathcal{A}_t$.
3. Unbeknownst to the learner, the environment generates a random delay, a random reward and then schedules an observation time:
   - 3a. The random reward has the form: $Y_t = \mu\left(X_t^T \theta^*\right) + \eta_t$.
   - 3b. The environment schedules the observation time of the reward: $\lceil t + \tau_t \rceil$ where $\tau_t \sim f_\tau\left(\cdot\right)$.
5. The learner receives delayed rewards from its previous actions: $\{(s, Y_s) : t - 1 < s + \tau_s \leq t\}$.

From the above decision-making procedure, it is clear that the learner only has access to the rewards of the actions whose observation times are less than or equal to $t-1$ when making decisions in round t. Therefore, Y_s is observable to the learner in the rounds where $s+\tau_s \leq t-1$. Otherwise, it is missing. To that end, we define the σ-algebra generated by the set of observable information at the start of the t-th round as:

$$\mathcal{F}_{t-1} = \sigma\left(\left\{\left(X_s, C_s^{t-1}, Y_s C_s^{t-1}\right) : s \leq t-1\right\} \cup \mathcal{A}_t\right)$$

where

$$C_s^t = \mathbb{1}\left\{s + \tau_s \leq t\right\}$$

indicates whether the reward associated with the s-th round is observable at the end of the t-th round. Naturally, C_s^t is observable at the end of each round, as the learner can easily check which actions have and have not received feedback; this is standard in most works on delays in the bandit literature (Dudik et al., 2011; Joulani et al., 2013; Mandel et al., 2015; Zhou et al., 2019; Blanchet et al., 2020). Notably, C_s^t is $\mathcal{F}_t$-measurable, meaning the learner only has access to the indicators and the rewards observed at the end of rounds $1, \cdots, t-1$ when making decisions in round t.

2.2 Assumptions

We make the following assumptions on the noise and the link function. These are standard in the literature on linear and generalised linear bandits (Filippi et al., 2010; Abbasi-Yadkori et al., 2011; Li et al., 2017).

Assumption 1 (Subgaussian Noise)**.** *Let $R \geq 0$ and $|\eta_t| \leq R$ almost surely. Then, the moment generating function of the noise distribution conditional on the observed information must satisfy the following inequality:*

$$\mathbb{E}\left[\exp\left(\gamma\, \eta_t\right) | \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{1}{2}\gamma^2 R^2\right)$$

for all $\gamma \in \mathbb{R}$.

Assumption 2 (Link Function)**.** *The link function $\mu : \mathbb{R} \to \mathbb{R}$ is known a-priori and is twice differentiable with first and second derivatives bounded by L_μ and M_μ, respectively. Further,*

$$\kappa \coloneqq \inf\left\{\dot{\mu}\left(\langle x, \theta \rangle\right) : (x, \theta) \in \mathcal{A} \times \Theta\right\} > 0.$$

where Θ is the set of all possible parameter vectors.

Assumption 1 implies that the noise distribution has light tails. Assumption 2 implies that the link function is L_μ-Lipschitz. One can interpret the condition on κ as guaranteeing that it is possible to distinguish between two actions whose expected rewards are arbitrarily close to one another. Indeed, R, L_μ and κ all feature in the theoretical analysis and regret bounds.

It will also be necessary for the delays to satisfy some assumptions (see Section 3.3). In particular, we assume the following holds.

Assumption 3 (Subexponential Delays)**.** *The delays are non-negative, independent and identically distributed (v, b)-subexponential random variables. That is, their moment generating function satisfies the following inequality:*

$$\mathbb{E}\left[\exp\left(\gamma\left(\tau_t - \mathbb{E}\left[\tau_t\right]\right)\right)\right] \leq \exp\left(\frac{1}{2}v^2\gamma^2\right)$$

for some non-negative v and b, and all $|\gamma| \leq 1/b$.

The class of distributions with subexponential tail behaviour is broad enough to include many heavy-tailed distributions, such as the χ^2 and Exponential distributions. Importantly, Assumption 3 aligns with the empirical evidence suggesting that delays have exponential-like tails in practice (Chapelle, 2014). However, other tail bounds on the delays can be used if they exist. Furthermore, it is possible to relax this assumption to only requiring that the delays have a finite (unknown) expected value by considering the expected regret, a weaker theoretical guarantee.

2.3 Notation

Throughout, $\|x\|_p$ denotes the p-norm of an arbitrary vector $x \in \mathbb{R}^d$. For $A, B \in \mathbb{R}^{d\times d}$, we denote $\|x\|_A = \sqrt{x^T A x}$ and adopt the following notation for positive (semi)-definite matrices:

- $A \succeq 0$ (positive semi-definite) $\iff$ $\|x\|_A^2 \geq 0$ for all $x \in \mathbb{R}^d$.
- $A \succeq B \iff \|x\|_A^2 \geq \|x\|_B^2$ for all $x \in \mathbb{R}^d$.

Additionally, $\lambda_i(A)$ and $\sigma_i(A)$ denote the i-th largest eigenvalue and the i-th largest singular value of matrix A, respectively. Finally, we denote the first and second derivatives of a real-valued function f by $\dot{f}$ and $\ddot{f}$, respectively.

3 DELAYED OFU-GLM

In this section, we describe a provably efficient algorithm for generalised linear bandits with stochastic delays. We base our approach on the optimistic principle and show that delays only cause an additive increase in the regret bound. This is in contrast to the multiplicative effect seen in existing work (Blanchet et al., 2020).

Due to the delays, it is necessary to introduce some additional notation that discriminates between rounds whose feedback has or has not been observed. Denote the number of missing rewards at the end of the t-th round by:

$$G_t = \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s > t\right\} \ .$$

Further, we define the total, observed and missing design matrices as

$$\bar{V}_t = \lambda I + \sum_{s=1}^{t} X_s X_s^T \tag{3}$$

$$\bar{W}_t = \lambda I + \sum_{s=1}^{t} \mathbb{1}\{s + \tau_s \leq t\} X_s X_s^T \tag{4}$$

$$Z_t = \sum_{s=1}^{t} \mathbb{1}\{s + \tau_s > t\} X_s X_s^T \,, \tag{5}$$

respectively. Here, $\lambda > 0$ is a regularisation parameter. Briefly, $\bar{V}_t$ is the total design matrix and contains information relating to all past choices. Whereas $\bar{W}_t$ and Z_t include information about actions with and without observed rewards, respectively. It is easy to see that the total, observed and missing design matrices must satisfy the following relationship:

$$\bar{V}_t = \bar{W}_t + Z_t \ . \tag{6}$$

Thus, when there are no delays, the total and observed design matrices are equivalent to each other, and the missing design matrix is full of zeros.

3.1 Estimation Procedure

As is standard when fitting generalised linear models, we use maximum likelihood estimation to estimate the unknown parameter of the environment. However, we make several adjustments to the estimator to account for delayed feedback.

First note that not all actions played will have received feedback. To mitigate this issue, we ignore the actions with missing feedback in our estimation procedure. Secondly, many existing algorithms for generalised linear bandits use a phase of pure exploration (Filippi et al., 2010; Li et al., 2017). This exploration phase lasts until the observed design matrix is of full rank, which ensures a unique maximiser of the likelihood function exists. Since we choose to ignore actions with missing feedback, the length of this exploration phase will increase depending on delay distribution. To avoid waiting for an exploration phase to pass, we introduce a penalisation term into the objective function, an idea that we borrow from the linear bandits where one can derive a closed-form penalised maximum likelihood estimator (Abbasi-Yadkori et al., 2011; Chu et al., 2011). In

the generalised linear setting, this trick equates to penalising the log-likelihood function and has found use for logistic bandits under immediate feedback (Jun et al., 2017; Faury et al., 2020). From Equation (1) and the conditional independence of the rewards given past actions, one can write the penalised log-likelihood as follows:

$$\mathcal{L}_t(\theta, \alpha) = \sum_{s=1}^{t} C_s^t \log\left(f\left(Y_t \,|X_t\right)\right) - \frac{\alpha}{2}\|\theta\|_2^2 \quad (7)$$

Equation (7) always has a unique maximiser due to the introduction of $\alpha > 0$, which means we can leverage new information from the very first round. One can easily verify that the maximiser is the solution of the following equation:

$$\left[\sum_{s=1}^{t} C_s^t \left(Y_s - \mu\left(X_t^T\theta\right)\right) X_s\right] - \alpha\, a(\phi)\, \theta = 0 \quad (8)$$

where $a(\phi)$ is a known function of the dispersion parameter of the reward distribution. We denote the solution of Equation (8) by $\hat{\theta}_t$. To implement the optimistic principle, we construct confidence sets around our estimators and prove that this set contains θ^* with high probability.

Lemma 1. *Let $\lambda = \alpha\, a(\phi)/\kappa$ and assume that $\|\theta^*\|_2 \leq m_1$. Then, with probability at least $1-\delta$, for all rounds $t \geq 0$:*

$$\|\hat{\theta}_t - \theta^*\|_{\bar{W}_t} \leq \sqrt{\lambda} m_1 + \frac{R}{\kappa}\sqrt{2\log\left(\frac{\det\left(\bar{W}_t\right)^{1/2}}{\delta\,\lambda^{d/2}}\right)}$$

Proof Sketch. Firstly, we account for regularising the log-likelihood function, which we do in Lemmas 6 and 7 of Appendix A. These lemmas allow us to separate noise-related terms from those introduced by biasing our estimator with the regularisation term. Subsequently, we show that the noise-related terms satisfy the martingale property under the information structure created by the delays. This result allows us to apply existing results for self-normalising processes (de la Peña et al., 2004; Abbasi-Yadkori et al., 2011). See Appendix A for a full proof. □

By Lemma 1, defining the confidence sets as:

$$\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \|\hat{\theta}_t - \theta\|_{\bar{W}_t} \leq \sqrt{\beta_t}\right\} \quad (9)$$

with

$$\sqrt{\beta_t} = \sqrt{\lambda} m_1 + \frac{R}{\kappa}\sqrt{2\log\left(\frac{\det\left(\bar{W}_t\right)^{1/2}}{\delta\,\lambda^{d/2}}\right)} \quad (10)$$

guarantees that $\mathbb{P}(\exists\, t \geq 0 : \theta^* \notin \mathcal{C}_t) \leq 1-\delta$.

Algorithm 1 Delayed OFU-GLM

Input: model parameters d, $a(\phi)$, m_1, κ, and tuning parameters $\alpha > 0$ and $\delta \in (0,1)$.
Initialise: $\hat{\theta}_0 = \vec{0}$, $\lambda = \frac{\alpha\, a(\phi)}{\kappa}$ and $\bar{W}_0 = \lambda I$
for $t = 1$ **to** T **do**
  Play X_t where:

$$(X_t, \tilde{\theta}_t) = \underset{(x,\theta)\in\mathcal{A}_t\times\mathcal{C}_{t-1}}{\arg\max}\; x^T\theta$$

  Receive the (possible empty) set of delayed rewards.
  Update $\bar{W}_t$, $\hat{\theta}_t$ and β_t via Equations (4), (8) and (10).
end for

3.2 Delayed OFU for Generalised Linear Bandits

Algorithm 1 presents the pseudo-code for our algorithm, Delayed OFU-GLM. It requires several input parameters that we briefly discuss below.

Firstly, algorithm requires knowledge of $a(\phi)$, a known function of the dispersion parameter of the reward distribution. For Bernoulli and Poisson rewards, one can show that $a(\phi) = 1$. In the Gaussian case, this parameter is the variance of the reward distribution $a(\phi) = R^2$, which all optimistic algorithms require to define the confidence sets.

Secondly, $m_1 \geq \|\theta^*\|_2$ is an upper bound on the ℓ_2-norm of the unknown parameter vector that features in many existing algorithms for the immediate feedback setting (Abbasi-Yadkori et al., 2011; Jun et al., 2017; Faury et al., 2020). Note that since (Zhou et al., 2019) uses a period of explicit exploration, they do not need this hyperparameter. Instead, they require knowledge of the delay distribution to define the length of the exploration phase.

Finally, κ quantifies the smallest possible rate of change in the expected reward function. For Linear bandits with Gaussian rewards, $\kappa = 1$. For other distributions, one can replace this quantity with a lower bound and our theoretical results will still hold. For Logistic bandits, one can utilise the fact that the first derivative of the link function is symmetric about zero and decreasing to show that: $\kappa \geq m_2 \coloneqq \dot{\mu}(m_1)$. For Poisson bandits, by the definition of the inner product, $\kappa \geq m_2 \coloneqq \exp(-m_1)$. Indeed, many optimistic algorithms for generalised linear bandits require this hyperparameter, as it features in the definition of the confidence sets. Recent work removes the need to specify this hyperparameter for the logistic bandit (Faury et al., 2020).

3.3 Regret Bounds for Delayed OFU-GLM

In this subsection, we state and prove a worst-case regret bound for our algorithm. Specifically, Algorithm 1 only suffers an additive penalty caused by the delays under the assumptions outlined in Section 2.2.

Theorem 1. *Suppose $\|x\|_2 \leq 1$ for all $x \in \cup_{t=1}^{\infty} \mathcal{A}_t$, and Assumptions 1, 2 and 3 hold. Then, with probability greater than $1 - 3\delta$, Delayed OFU-GLM with any regularisation parameter $\lambda = \alpha\, a(\phi)/\kappa \geq 1$ has pseudo-regret that satisfies:*

$$\hat{R}_T \leq \tilde{\mathcal{O}}\left(\frac{dRL_\mu}{\kappa}\sqrt{T} + \frac{d^{3/2}RL_\mu\left(\mathbb{E}\left[\tau\right] + \min\left\{v, b\right\}\right)}{\kappa}\right)$$

where v and b are the subexponential parameters of the delay distribution.

Proof. First, we bound the per-round pseudo-regret. In Algorithm 1,the action selected by the algorithm is optimistic with probability $1 - \delta$. Therefore,

$$\begin{aligned}
\hat{r}_t &= \mu\left(\langle \theta^*,\, X_t^* \rangle\right) - \mu\left(\langle \theta^*,\, X_t \rangle\right) \\
&\leq L_\mu\left(\langle \theta^*,\, X_t^* \rangle - \langle \theta^*,\, X_t \rangle\right) && \text{(Assumption 2)} \\
&\leq L_\mu\left(\langle \tilde{\theta}_t - \theta^*,\, X_t \rangle\right) && \text{(Lemma 1)}
\end{aligned}$$

where the final inequality holds with probability at least $1 - \delta$ across all rounds due to the definition of the confidence sets and the action-selection procedure in Algorithm 1. Adding and subtracting the maximum likelihood estimator gives:

$$\begin{aligned}
\hat{r}_t &\leq L_\mu\left(\langle \tilde{\theta}_t - \hat{\theta}_{t-1},\, X_t \rangle + \langle \hat{\theta}_{t-1} - \theta^*,\, X_t \rangle\right) \\
&\leq L_\mu \|\tilde{\theta}_t - \hat{\theta}_{t-1}\|_{\bar{W}_{t-1}} \|X_t\|_{\bar{W}_{t-1}^{-1}} + \\
&\quad L_\mu \|\hat{\theta}_{t-1} - \theta^*\|_{\bar{W}_{t-1}} \|X_t\|_{\bar{W}_{t-1}^{-1}} && \text{(Hölder's)} \\
&\leq 2L_\mu \sqrt{\beta_{t-1}} \|X_t\|_{\bar{W}_{t-1}^{-1}} && \text{(Definition of } \mathcal{C}_{t-1}) \\
&\leq 2L_\mu \sqrt{\beta_T} \|X_t\|_{\bar{W}_{t-1}^{-1}} && (\beta_1 \leq \beta_2 \leq \cdots \leq \beta_T)
\end{aligned}$$

Therefore, we have that the pseudo-regret has the following upper bound:

$$\hat{R}_T \leq 2L_\mu \sqrt{\beta_T} \sum_{t=1}^{T} \|X_t\|_{\bar{W}_{t-1}^{-1}} \tag{11}$$

with probability at least $1 - \delta$. Usually, an application of Cauchy-Schwarz and the elliptical potential lemma handles the remaining summation. This algebraic argument completes the proof in the immediate feedback setting and provides a tight upper bound on the term in question (Carpentier et al., 2020). However, the elliptical potential lemma requires that the learner updates the design matrix at the end of every round with the most recent action.

This is not the case for the summation in (11), as the feedback associated with the most recent action is not necessarily observable immediately and is, therefore, not used to increment the observed design matrix. Moreover, there will likely be rounds where no feedback arrives at all and rounds where multiple feedbacks return to the learner, meaning that the matrix determinant lemma does not hold; a key argument in the proof. Consequently, we introduce the following technical lemmas that aid in bounding the summation.

Lemma 2. *Let $\lambda = \alpha\, a(\phi)/\kappa > 0$. Then, $\bar{W}_t$ and $\bar{V}_t$ are invertible and have inverses that satisfy the following relationship:*

$$\bar{W}_t^{-1} = \bar{V}_t^{-1} + \bar{V}_t^{-1} Z_t\, \bar{W}_t^{-1} = \bar{V}_t^{-1} + M_t$$

where $M_t \coloneqq \bar{V}_t^{-1} Z_t\, \bar{W}_t^{-1}$.

Proof. See Appendix B. □

Lemma 3. *Let $\{\tau_t\}_{t=1}^{\infty}$ be an arbitrary sequence of non-negative random variables. Then, for $\lambda = \alpha\, a(\phi)/\kappa \geq 1$:*

$$\sum_{t=1}^{T} \|X_t\|_{M_{t-1}} \leq \sum_{t=1}^{T} \frac{1 + G_* + \tau_t}{2} \|X_t\|_{\bar{V}_{t-1}^{-1}}^2$$

where $G_* = \max\{G_t : t \leq T\}$.

Proof. See Appendix B. □

Lemma 2 relates the inverse of the observed design matrix to the inverse of the total design matrix and a product of three matrices. This allows us to separate the usual elliptical potential from terms involving the delays by application of the triangle inequality. Then, Lemma 3 shows that we can relate the remaining summation to a lower-order term. More concretely,

$$\begin{aligned}
&\sum_{t=1}^{T} \|X_t\|_{\bar{W}_{t-1}^{-1}} = \sum_{t=1}^{T} \|X_t\|_{\bar{V}_{t-1}^{-1} + M_t} && \text{(Lemma 2)} \\
&\leq \sum_{t=1}^{T} (\|X_t\|_{\bar{V}_{t-1}^{-1}} + \|X_t\|_{M_{t-1}}) && \text{(Triangle Inequality)} \\
&\leq \sum_{t=1}^{T} \|X_t\|_{\bar{V}_{t-1}^{-1}} + \sum_{t=1}^{T} \frac{1 + G_* + \tau_t}{2} \|X_t\|_{\bar{V}_{t-1}^{-1}}^2 && (12)
\end{aligned}$$

where the final inequality follows from Lemma 3. The above reveals that we must bound the number of missing rewards at the end of the t-th round and the delay, which we do in the following lemmas.

Lemma 4. *Define $G_t = \sum_{s=1}^{t} \mathbb{1}\{s + \tau_s > t\}$ and let $\{\tau_t\}_{t=1}^{\infty}$ be a sequence of independent and identically distributed random variables with a finite expectation and define:*

$$\psi_\tau^t \coloneqq \frac{4}{3}\log\left(\frac{3t}{2\delta}\right) + 2\sqrt{2\mathbb{E}\left[\tau\right]\log\left(\frac{3t}{2\delta}\right)}.$$

Then,

$$\mathbb{P}\left(\exists\, t \geq 1 : G_t \leq \mathbb{E}[\tau] + \psi_\tau^t\right) \leq 1 - \delta.$$

Proof. The proof follows the same arguments used in multi-armed bandits (Joulani et al., 2013). However, we include a simple extension to accommodate for continuous delay distributions. See Appendix B. □

Lemma 5. *Let $\{\tau_t\}_{t=1}^{\infty}$ satisfy Assumption 3 and define:*

$$D_\tau^t = \min\left\{\sqrt{2v^2\log\left(\frac{3t}{2\delta}\right)}, 2b\log\left(\frac{3t}{2\delta}\right)\right\}$$

Then,

$$\mathbb{P}\left(\exists\, t \geq 1 : \tau_t \leq \mathbb{E}\left[\tau\right] + D_\tau^t\right) \leq 1-\delta$$

Proof. The above follows from a standard tail bound for subexponential random variables (Wainwright, 2019) and a union bound. □

Applying Lemmas 4 and 5, combined with the observation that $\psi_\tau \coloneqq \psi_\tau^T \geq \psi_\tau^t$ and $D_\tau \coloneqq D_\tau^T \geq D_\tau^t$ for all $t \leq T$ allows us to bound the delays and the maximum number of missing rewards in Equation (12) with high probability. By setting $D_\tau^+ = 1 + 2\mathbb{E}[\tau] + D_\tau + \psi_\tau$, we have that:

$$\begin{aligned}(12) &\leq \sum_{t=1}^{T}\|X_t\|_{\bar{V}_{t-1}^{-1}} + \frac{D_\tau^+}{2}\sum_{t=1}^{T}\|X_t\|^2_{\bar{V}_{t-1}^{-1}} \\ &\leq \sqrt{T\sum_{t=1}^{T}\|X_t\|^2_{\bar{V}_{t-1}^{-1}}} + \frac{D_\tau^+}{2}\sum_{t=1}^{T}\|X_t\|^2_{\bar{V}_{t-1}^{-1}}\end{aligned}$$

with probability $1-2\delta$, where the final inequality follows from an application of Cauchy-Schwarz. Now, the total design matrix is incremented by the most recent action at the end of every round. Therefore, we can apply the elliptical potential lemma, which bounds the remaining summation terms as follows:

$$\sum_{t=1}^{T}\|X_t\|^2_{\bar{V}_{t-1}^{-1}} \leq 2d\log\left(\frac{d\lambda+T}{d\lambda}\right) = 2dL$$

where $L \coloneqq \log((d\lambda+T)/d\lambda)$. For completeness, we provide a statement and proof of this well-known result in Appendix D. Therefore,

$$\sum_{t=1}^{T}\|X_t\|_{\bar{W}_{t-1}^{-1}} \leq \sqrt{2dTL} + dLD_\tau^+.$$

From Equation (11), it is clear that all that remains is to upper bound the width of the confidence set at the end of the final round. Recall $\bar{V}_t \succeq \bar{W}_t$, because the observed design matrix is a partial sum of positive semi-definite matrices that make up the total design matrix. Therefore,

$$\begin{aligned}\sqrt{\beta_T} &\leq \sqrt{\lambda}m_1 + \frac{R}{\kappa}\sqrt{2\log\left(\frac{|\bar{V}_T|^{1/2}}{\lambda^{d/2}}\right) + 2\log\left(\frac{1}{\delta}\right)} \\ &\leq \sqrt{\lambda}m_1 + \frac{R}{\kappa}\sqrt{2dL + 2\log\left(\frac{1}{\delta}\right)}\end{aligned}$$

where the inequality follows from Lemma 15 of Appendix D. Bringing everything together,

$$\begin{aligned}\hat{R}_T &\leq 2L_\mu\sqrt{\beta_T}\sum_{t=1}^{T}\|X_t\|_{\bar{W}_{t-1}^{-1}} \\ &\leq 2L_\mu\sqrt{\beta_T}\left(\sqrt{2dTL} + dLD_\tau^+\right)\end{aligned}$$

Substituting $D_\tau^+ = 1 + 2\mathbb{E}[\tau] + D_\tau + \psi_\tau$ and our upper bound on $\sqrt{\beta_T}$ into the above, and omitting poly-logarithmic factors gives:

$$\widetilde{\mathcal{O}}\left(\frac{dRL_\mu}{\kappa}\sqrt{T} + \frac{d^{3/2}RL_\mu\left(\mathbb{E}\left[\tau\right] + \min\left\{v,b\right\}\right)}{\kappa}\right)$$

completing the proof. □

Remark 1. *Under Assumptions 1 and 2, one can relax the assumption on the delays from subexponential to only requiring a finite expected value if we only consider a weaker notion of regret, namely the expected regret. Formally, for a fixed θ^* and any delay distribution with a finite expected value:*

$$\mathbb{E}\left[\hat{R}_T\right] \leq \widetilde{\mathcal{O}}\left(\frac{dRL_\mu}{\kappa}\sqrt{T} + \frac{d^{3/2}RL_\mu\mathbb{E}[\tau]}{\kappa}\right)$$

where we take the expectation over the randomness of the rewards and the delays. This result follows from standard arguments; e.g. by setting $\delta = 1/T$ and using the definition of the confidence sets. Eventually, we end up taking the expectation of Equation (12) *with respect to the rewards and delays.*

Remark 2. *In the proof, we focused on the confidence sets given in Lemma 1. At the heart of this confidence set is a high probability bound on:*

$$\left\|\sum_{s=1}^{t}\mathbb{1}\{s+\tau_s \leq t\}X_s\eta_s\right\|_{\bar{W}_t^{-1}}$$

which we prove is a non-negative supermartingale under the information structure imposed on the learner by the delays. Many other algorithms utilise slightly different techniques to bound an identical term (Filippi et al., 2010; Li et al., 2017) or one that differs by the choice of weight in the norm (Faury et al., 2020) to define confidence sets. By Lemma 8, Algorithm 1 ensures that these confidence sets are valid in the delayed feedback setting too. Thus, combining our theoretical results within their analyses will yield a similar additive delay-dependent quantity in the regret bounds under delayed feedback.

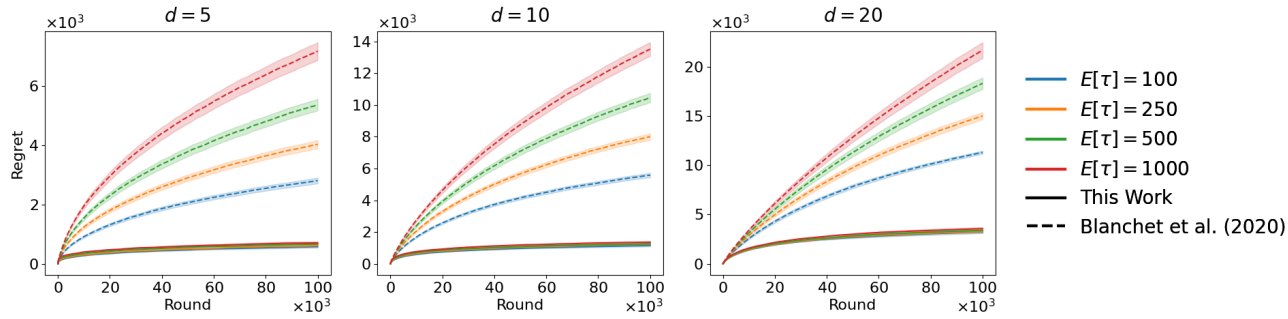

Figure 1: Linear Bandit & Exponentially Distributed Delays.

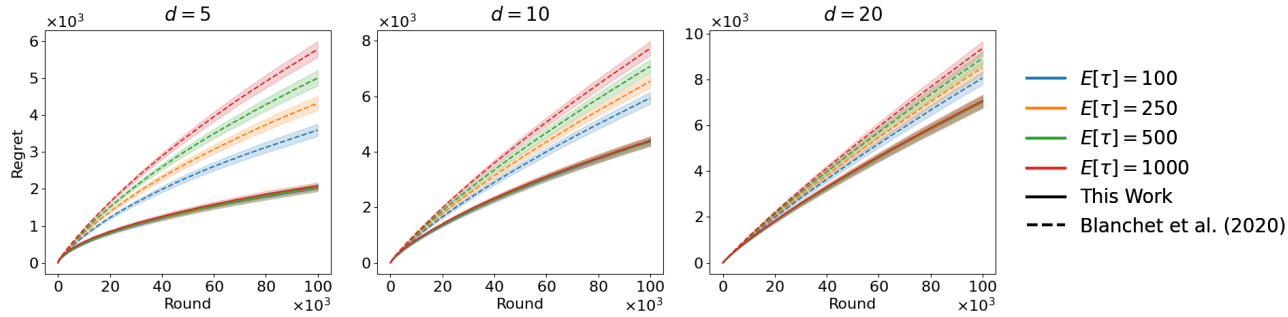

Figure 2: Logistic Bandit & Exponentially Distributed Delays.

4 EXPERIMENTAL RESULTS

We conduct simulated experiments to empirically investigate the impact of delayed feedback in Linear and Logistic bandits. We compare our algorithmic ideas to other approaches for the setting of delayed feedback in generalised linear bandits, which inflate the exploration bonus by the number of missing rewards (Blanchet et al., 2020).

In our experiments, we consider $d \in \{5, 10, 20\}$ and fix $T = 100,000$. At the start of the simulations, we randomly sample θ^* from the unit ball for the Linear and Logistic bandit environments so that it remains fixed across each independent run of our experiments. The decision set in each round is a random sample of $K = 100$ actions from the unit ball. We choose the confidence parameter for each algorithm so that the theoretical guarantees hold with probability 0.95 by setting $\delta = 0.05/3$. All results are averaged over 30 independent runs and the shaded region in all figures represent the standard errors of the estimates.

We consider several delay distributions to investigate the impact of the delays on the performance of each algorithm, namely:

- Exponential(λ) with $\lambda = 1/\mathbb{E}[\tau]$,
- Uniform(a, b) with $a = 0$ and $b = 2\mathbb{E}[\tau]$,
- Pareto($a, x_m = 1$) with $a = (1 + \mathbb{E}[\tau])/\mathbb{E}[\tau]$.

For each delay distribution, we consider expected values of $\mathbb{E}[\tau] = \{100, 250, 500, 1000\}$. Notably, Assumption 3 holds for the uniform and exponential distributions. However, it does not hold for the Pareto distribution. Blanchet et al. (2020) make a similar subexponential assumption on the delays, meaning that their theoretical guarantees do not hold for Pareto delays either.

Figures 1 and 2 illustrate the results of our experiments for exponentially distributed delays. Appendix C shows similar results for the other delay distributions and expected delays considered. The empirical results show that our approach out-performs existing algorithms designed for the same problem setting. These results are consistent with the theoretical guarantees, where the delayed feedback causes an additive penalty for our algorithm and a larger multiplicative penalty for the approach of Blanchet et al. (2020).

Figures 3 and 4 show the regret at the end of the final round as a function of the expected delay for our algorithm. Although Assumption 3 does not hold for delays drawn from the Pareto distribution, our algorithm still provides good performance for the various values of $\mathbb{E}[\tau]$ considered by our experiments. Notably, these empirical results are consistent with the expected regret guarantee stated in Remark 1, which only requires that the delays have a finite expected value. Our experiments also suggest that the penalty for Pareto delays is lesser than the other distributions under investigation. This observation may be due to our particular parameterisation of the Pareto distribution producing many

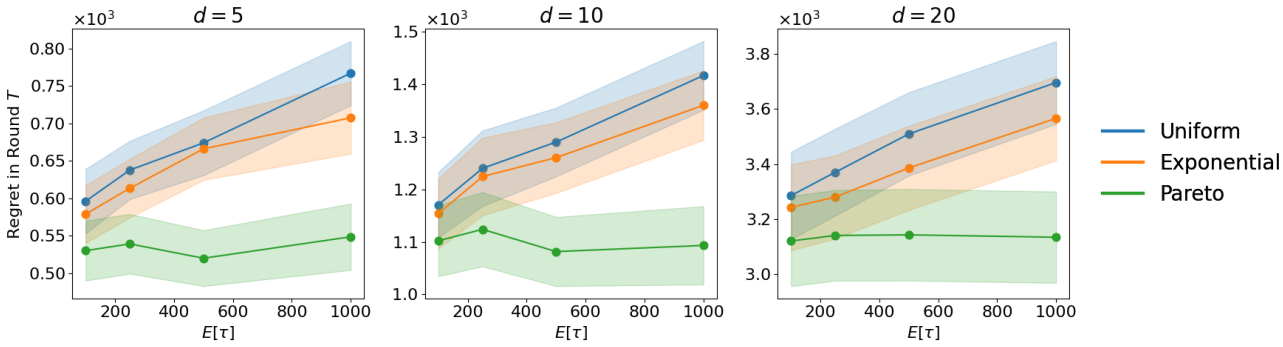

Figure 3: Final Round Regret vs. Expected Delay in Linear Bandits.

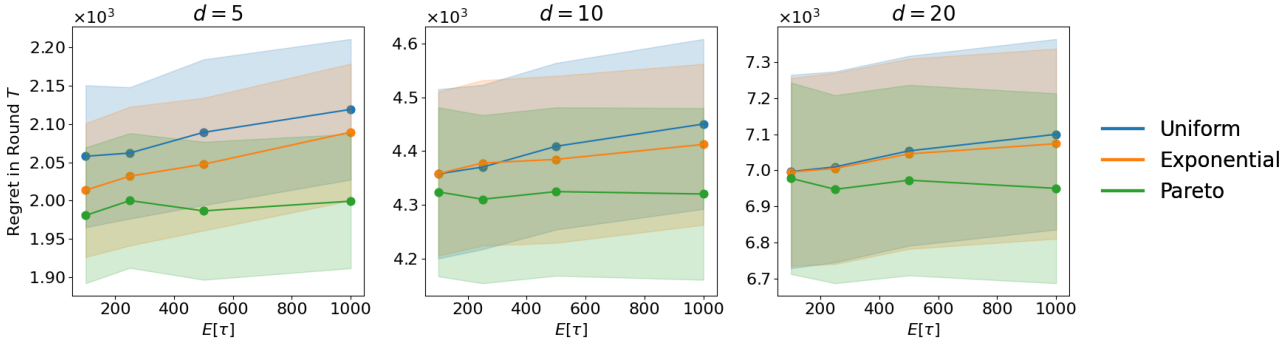

Figure 4: Final Round Regret vs. Expected Delay in Logistic Bandits.

small delays; indeed, $\mathbb{P}(\tau_t \leq 20) \geq 0.95$ for all the expected delays considered by our experiments. The same is not true for the other distributions.

5 CONCLUSION

In this work, we studied the impact of delayed feedback on algorithms for generalised linear bandits. Under Assumption 3, that the delays are subexponential random variables, we designed an optimistic algorithm whose worst-case regret bound increases by an additive term involving the expected delay. We obtain a similar result for the expected regret, which only requires that the delays have a finite expected value.

These theoretical results significantly improve on prior work, where existing algorithms suffer a multiplicative penalty and require a-priori knowledge of the delay distribution as input. Reducing the delay dependence from multiplicative to additive was possible by introducing a novel technique to carefully separate the delays from the difficulty of the learning problem. Doing so allowed us to define tighter confidence sets than existing algorithms, leading to better theoretical guarantees and superior empirical performance. Indeed, the theoretical techniques introduced in this paper might be useful in other bandit problems with complex feedback structures.

Our result nearly recovers the additive delay penalty from multi-armed bandits, despite the additional difficulties of our setting. Whether or not it is possible to remove the d-dependence entirely remains an interesting open question. Another open question relates to relaxing our assumptions on the delays. Namely, can we get high probability bounds that only require that the delays have a finite expected value? We anticipate that addressing these open questions may require adjustments to our theoretical techniques or different algorithmic approaches.

Finally, we expect that similar results hold for a Thompson Sampling version of our algorithm. Combining techniques found in Russo and Van Roy (2014) with those in this paper will likely give similar guarantees for the Bayesian regret.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful feedback that helped greatly improved the clarity and quality of the manuscript.

BH is funded by EPSRC through the Modern Statistics and Statistical Machine Learning CDT. Grant number EP/S023151/1.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Jose Blanchet, Renyuan Xu, and Zhengyuan Zhou. Delay-Adaptive Learning in Generalized Linear Contextual Bandits, 2020. URL https://arxiv.org/abs/2003.05174.

Alexandra Carpentier, Claire Vernade, and Yasin Abbasi-Yadkori. The Elliptical Potential Lemma Revisited. *ArXiv*, abs/2010.10182, 2020.

Olivier Chapelle. Modeling Delayed Feedback in Display Advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 1097–1105, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual Bandits with Linear Payoff Functions. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/chu11a.html.

Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3):1902 – 1933, 2004.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient Optimal Learning for Contextual Bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 169–178, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D. Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 67–82. PMLR, 17–18 Aug 2018. URL https://proceedings.mlr.press/v85/durand18a.html.

Louis Faury, Marc Abeille, Clement Calauzenes, and Olivier Fercoq. Improved Optimistic Algorithms for Logistic Bandits. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3052–3060. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/faury20a.html.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf.

Benjamin Han and Carl Arndt. Budget Allocation as a Multi-Agent System of Contextual & Continuous Bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 2937–2945, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467124. URL https://doi.org/10.1145/3447548.3467124.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online Learning under Delayed Feedback. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1453–1461, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable Generalized Linear Bandits: Online Computation and Hashing. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/28dd2c7955ce926456240b2ff0100bde-Paper.pdf.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2071–2080. PMLR, 06–11 Aug 2017.

Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. The Queue Method: Handling Delay, Heuristics, Prior Data, and Evaluation in Bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015.

James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. Explore, exploit, and explain: Personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 31–39, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359016. doi: 10.1145/3240323.3240354. URL https://doi.org/10.1145/3240323.3240354.

Kanishka Misra, Eric Schwartz, and Jacob Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38, 03 2019. doi: 10.1287/mksc.2018.1129.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4105–4113. PMLR, 10–15 Jul 2018.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014. doi: 10.1287/moor.2014.0650. URL https://doi.org/10.1287/moor.2014.0650.

Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. Portfolio choices with orthogonal bandit learning. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 974–980. AAAI Press, 2015. ISBN 9781577357384.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic Bandit Models for Delayed Conversions. In *Conference on Uncertainty in Artificial Intelligence*, Sydney, Australia, August 2017. URL https://hal.archives-ouvertes.fr/hal-01545667.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brückner. Linear Bandits with Stochastic Delayed Feedback. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9712–9721. PMLR, 13–18 Jul 2020.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in Generalized Linear Contextual Bandits with Stochastic Delays. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

A CONFIDENCE SETS

Here, we show that the confidence sets are valid under delayed feedback. To that end, we define the following σ-algebra:

$$\mathcal{F}_{t-1} = \sigma\left(\left\{\left(X_s, C_s^{t-1}, Y_s C_s^{t-1}\right) : s \leq t-1\right\} \cup \mathcal{A}_t\right) \tag{13}$$

Consequently, Y_t is $\mathcal{F}_t$-measurable. Further, X_t is $\mathcal{F}_{t-1}$-measurable. For notational purposes, we find it useful to define:

$$g_t(\theta) = \alpha\, a(\phi)\, \theta + \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} \mu\left(\langle X_t, \theta \rangle\right) X_s$$

as well as the second derivative of the negative log likelihood:

$$H_t(\theta) = \alpha\, a(\phi) + \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} \dot{\mu}\left(\langle X_s, \theta \rangle\right) X_s X_s^T$$

Now, $\hat{\theta}_t$ is the vector satisfying the following equality:

$$\begin{aligned}
\frac{\partial \mathcal{L}_t(\theta, \alpha)}{\partial \theta} &= \left[\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} \left(Y_s - \mu\left(X_t^T \theta\right)\right) X_s\right] - \alpha\, a(\phi)\, \theta \\
&= \left[\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} Y_s X_s\right] - g_t(\theta) \\
&= 0
\end{aligned} \tag{14}$$

Lemma 1. *Let $\lambda = \alpha\, a(\phi)/\kappa$ and assume that $\|\theta^*\|_2 \leq m_1$. Then, with probability at least $1-\delta$, for all rounds $t \geq 0$:*

$$\|\hat{\theta}_t - \theta^*\|_{\bar{W}_t} \leq \sqrt{\lambda} m_1 + \frac{R}{\kappa} \sqrt{2 \log\left(\frac{\det\left(\bar{W}_t\right)^{1/2}}{\delta\, \lambda^{d/2}}\right)}$$

Proof. By Lemmas 6 and 7 of Appendix A.1, we have that:

$$\|\hat{\theta}_t - \theta^*\|_{\bar{W}_t} \leq \sqrt{\lambda} \|\theta^*\|_2 + \frac{1}{\kappa} \|S_t\|_{\bar{W}_t^{-1}}$$

where

$$S_t = \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} X_s \eta_s$$

Further Lemma 8 of A.1 reveals that the last term in the above is a non-negative supermartingale under the delayed feedback information structure. Therefore, we are able to use known methods for bounding self-normalised vector-valued martingales (de la Peña et al., 2004; Abbasi-Yadkori et al., 2011). Let ω be a stopping time with respect to the filtration. Applying Lemma 9 of Abbasi-Yadkori et al. (2011) to the stopped martingale gives:

$$\mathbb{P}\left(\|S_\omega\|_{\bar{W}_\omega^{-1}} \geq R \sqrt{\log\left(\frac{\det\left(\bar{W}_\omega\right)}{\lambda^d}\right) + 2 \log\left(\frac{1}{\delta}\right)}\right) \leq \delta$$

Since Lemma 8 guarantees that the stopped supermartingale is well-defined, regardless of whether the stopping time is finite, the above inequality holds across all rounds without the need for a union bound. That is:

$$\mathbb{P}\left(\exists\, t \geq 0 : \|S_t\|_{\bar{W}_t^{-1}} \geq R \sqrt{\log\left(\frac{\det\left(\bar{W}_t\right)}{\lambda^d}\right) + 2 \log\left(\frac{1}{\delta}\right)}\right) \leq \delta \tag{15}$$

Therefore, with probability at least $1-\delta$:

$$\begin{aligned}\|\hat{\theta}_t - \theta^*\|_{\bar{W}_t} &\leq \sqrt{\lambda}\|\theta^*\|_2 + \frac{1}{\kappa}\|S_t\|_{\bar{W}_t^{-1}} && \text{(Lemmas 6 \& 7)}\\ &\leq \sqrt{\lambda}\|\theta^*\|_2 + \frac{R}{\kappa}\sqrt{\log\left(\frac{\det\left(\bar{W}_t\right)}{\lambda^d}\right) + 2\log\left(\frac{1}{\delta}\right)}\end{aligned}$$

as required. □

A.1 Supporting Lemmas

Proving Lemma 1 requires several supporting lemmas. Firstly, the confidence sets are in terms of $\hat{\theta}_t$ and θ^*. Conversely, Equation (14) reveals that our estimation procedure involves $g_t(\hat{\theta})$ and $g_t(\theta^*)$. The following lemma allowed us to relate these two quantities to one another.

Lemma 6. *Let $\theta_1 \in \mathbb{R}^d$ and $\theta_2 \in \mathbb{R}^d$ be arbitrary vectors, and $\lambda = \alpha a(\phi)/\kappa$. Then, the following inequality holds:*

$$\kappa \left\|\theta_1 - \theta_2\right\|_{\bar{W}_t} \leq \left\|g\left(\theta_1\right) - g\left(\theta_2\right)\right\|_{\bar{W}_t^{-1}}$$

Proof. Similarly to Filippi et al. (2010), we apply the mean value theorem to the terms inside the norm on the right-hand side of the above, which allows us to related them to the original vectors. Expanding $g\left(\theta_1\right)$ and $g\left(\theta_2\right)$ reveals that:

$$\begin{aligned}g\left(\theta_1\right) - g\left(\theta_2\right) &= \alpha\, a(\phi)\, \theta_1 - \alpha\, a(\phi)\, \theta_2 + \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\}\left[\mu\left(\langle X_t, \theta_1 \rangle\right) - \mu\left(\langle X_t, \theta_2 \rangle\right)\right] X_s\\ &= \alpha\, a(\phi)\, \theta_1 - \alpha\, a(\phi)\, \theta_2 + \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\}\, \dot{\mu}\left(\langle X_t, \bar{\theta} \rangle\right) X_s X_s^T \left(\theta_1 - \theta_2\right)\\ &= \left[\alpha\, a(\phi) + \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\}\, \dot{\mu}\left(\langle X_t, \bar{\theta} \rangle\right) X_s X_s^T\right]\left(\theta_1 - \theta_2\right)\\ &= H_t\left(\bar{\theta}\right)\left(\theta_1 - \theta_2\right) \qquad (16)\end{aligned}$$

where the second equality follows from the mean value theorem for some $\bar{\theta} \in (\theta_2, \theta_1)$. Rewriting the Hessian for some $\theta \in \mathbb{R}^d$ and recalling that $\kappa \leq \dot{\mu}(z)$ reveals that:

$$H_t\left(\theta\right) = \alpha\, a(\phi) + \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} \dot{\mu}\left(\langle X_s, \theta \rangle\right) X_s X_s^T \succeq \kappa\left[\frac{\alpha\, a(\phi)}{\kappa} + \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} X_s X_s^T\right] = \kappa \bar{W}_t \qquad (17)$$

From Equation (17), we immediately have that $H_t^{-1}(\theta) \preceq \bar{W}_t^{-1}/\kappa$. Combining Equation (16) with the partial ordering of Equation (17) gives:

$$\begin{aligned}\left\|\theta_1 - \theta_2\right\|_{\kappa\bar{W}_t} &\leq \left\|\theta_1 - \theta_2\right\|_{H_t\left(\bar{\theta}\right)} && \left(\kappa\bar{W}_t \preceq H_t(\theta)\right)\\ &= \left\|H_t^{1/2}\left(\bar{\theta}\right)\left(\theta_1 - \theta_2\right)\right\|_2 && \left(\|x\|_A = \|A^{1/2}x\|_2\right)\\ &= \left\|H_t^{-1/2}\left(\bar{\theta}\right)\left(g_t\left(\theta_1\right) - g_t\left(\theta_2\right)\right)\right\|_2 && \text{(Equation (16))}\\ &= \left\|g_t\left(\theta_1\right) - g_t\left(\theta_2\right)\right\|_{H_t^{-1}\left(\bar{\theta}\right)} && \left(\|A^{1/2}x\|_2 = \|x\|_A\right)\\ &\leq \left\|g_t\left(\theta_1\right) - g_t\left(\theta_2\right)\right\|_{\frac{1}{\kappa}\bar{W}_t^{-1}} && \left(H_t^{-1}(\theta) \preceq \bar{W}_t^{-1}/\kappa\right)\end{aligned}$$

Therefore, using homogeneity property of norms on the first and last terms of the above reveals that:

$$\sqrt{\kappa}\|\theta_1 - \theta_2\|_{\bar{W}_t} = \left\|\theta_1 - \theta_2\right\|_{\kappa\bar{W}_t} \leq \left\|g_t\left(\theta_1\right) - g_t\left(\theta_2\right)\right\|_{\frac{1}{\kappa}\bar{W}_t^{-1}} = \frac{1}{\sqrt{\kappa}}\left\|g_t\left(\theta_1\right) - g_t\left(\theta_2\right)\right\|_{\bar{W}_t^{-1}}$$

Bringing all κ's to the left hand side side gives the stated result. □

Lemma 7. *Let θ^* be the unknown parameter of the environment and $\hat{\theta}_t$ be the solution to* (14)*. Further, define $\lambda = \alpha\, a(\phi)/\kappa$ and*

$$S_t = \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} X_s \eta_s$$

Then,

$$\|\hat{\theta}_t - \theta^*\|_{\bar{W}_t} \leq \sqrt{\lambda}\|\theta^*\|_2 + \frac{1}{\kappa}\|S_t\|_{\bar{W}_t^{-1}}$$

Proof. By Lemma 6, we have that:

$$\|\hat{\theta}_t - \theta^*\|_{\bar{W}_t} \leq \frac{1}{\kappa}\left\| g\left(\hat{\theta}_t\right) - g\left(\theta^*\right)\right\|_{\bar{W}_t^{-1}} \tag{18}$$

Since $\hat{\theta}_t$ is the solution to (14), it follows that:

$$\left[\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} Y_s X_s\right] - g_t\left(\hat{\theta}_t\right) = 0 \iff g_t\left(\hat{\theta}_t\right) = \left[\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} Y_s X_s\right]$$

Substituting the above into (18) gives:

$$\begin{aligned}
\left\|\hat{\theta}_t - \theta^*\right\|_{\bar{W}_t} &\leq \frac{1}{\kappa}\left\| g\left(\hat{\theta}_t\right) - g\left(\theta^*\right)\right\|_{\bar{W}_t^{-1}} \\
&= \frac{1}{\kappa}\left\|\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} Y_s X_s - g\left(\theta^*\right)\right\|_{\bar{W}_t^{-1}} \\
&= \frac{1}{\kappa}\left\|\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} Y_s X_s - \left[\alpha\, a(\phi)\, \theta^* + \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} \mu\left(\langle X_t, \theta^* \rangle\right) X_s\right]\right\|_{\bar{W}_t^{-1}} \\
&= \frac{1}{\kappa}\left\|-\alpha\, a(\phi)\, \theta^* + \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} \left[Y_s - \mu\left(\langle X_s, \theta^* \rangle\right)\right] X_s\right\|_{\bar{W}_t^{-1}} \\
&= \frac{1}{\kappa}\left\|-\alpha\, a(\phi)\, \theta^* + \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} \left[\mu\left(\langle X_s, \theta^* \rangle\right) + \eta_s - \mu\left(\langle X_s, \theta^* \rangle\right)\right] X_s\right\|_{\bar{W}_t^{-1}} \\
&= \frac{1}{\kappa}\left\|-\alpha\, a(\phi)\, \theta^* + \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} X_s \eta_s\right\|_{\bar{W}_t^{-1}} = \frac{1}{\kappa}\left\|S_t - \alpha\phi\theta^*\right\|_{\bar{W}_t^{-1}} \\
&\leq \frac{1}{\kappa}\left\|\alpha\, a(\phi)\, \theta^*\right\|_{\bar{W}_t^{-1}} + \frac{1}{\kappa}\|S_t\|_{\bar{W}_t^{-1}} \\
&\leq \frac{1}{\kappa}\sqrt{\frac{\alpha^2\, a(\phi)^2}{\lambda}}\|\theta^*\|_2 + \frac{1}{\kappa}\|S_t\|_{\bar{W}_t^{-1}} \\
&= \sqrt{\lambda}\|\theta^*\|_2 + \frac{1}{\kappa}\|S_t\|_{\bar{W}_t^{-1}}
\end{aligned}$$

where the final inequality follows from the fact that $\bar{W}_t^{-1} \preceq \lambda^{-1} I$, and the final equality follows from the fact that $\lambda = \alpha\, a(\phi)/\kappa$. □

All that remains is bounding the norm involving the noise terms with high probability, which is the second term in the result stated in Lemma 7. By Fenchel Duality, we have that (Abbasi-Yadkori et al., 2011):

$$\begin{aligned}
\frac{1}{2}\|S_t\|_{\bar{W}_t^{-1}}^2 &= \max_{x \in \mathbb{R}^d}\left\{\langle x, S_t\rangle - \frac{1}{2}\|x\|_{\bar{W}_t}^2\right\} \\
&= \max_{x \in \mathbb{R}^d}\left\{\log\left(\exp\left(\langle x, S_t\rangle - \frac{1}{2}\|x\|_{\bar{W}_t}^2\right)\right)\right\}
\end{aligned}$$

$$= \log\left(\max_{x\in\mathbb{R}^d}\left\{\exp\left(\langle x, S_t\rangle - \frac{1}{2}\|x\|^2_{\bar{W}_t}\right)\right\}\right) \tag{19}$$

Equation (19) suggests that it would be useful to obtain a high probability bound on the following random variable:

$$M_t(x) = \exp\left(\frac{1}{R}\langle x, S_t\rangle - \frac{1}{2}\|x\|^2_{W_t}\right)$$

where

$$W_t = \bar{W}_t - \lambda I = \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\} X_s X_s^T$$

for an arbitrary vector $x \in \mathbb{R}^d$. To do so, we first establish the following supermartingale argument, which is essential in showing the validity of the confidence sets. Due to the delayed feedback, we cannot directly use results from the immediate feedback setting. Therefore, we make the necessary adjustments to account for the delays.

Lemma 8. *Let $x \in \mathbb{R}^d$ be an arbitrary vector and define:*

$$M_t(x) = \exp\left(\frac{1}{R}\langle x, S_t\rangle - \frac{1}{2}\|x\|^2_{W_t}\right) = \exp\left(\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\}\left(\frac{\langle x, X_s\rangle \eta_s}{R} - \frac{1}{2}\langle x, X_s\rangle^2\right)\right)$$

Let ω be a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Then, $M_\omega(x)$ is almost surely well-defined and $\mathbb{E}\left[M_\omega(x)\right] \leq 1$.

Proof. Recall that $C_s^t = \mathbb{1}\left\{s + \tau_s \leq t\right\}$. We start by re-writing $M_t(x)$ in terms of $M_{t-1}(x)$:

$$\begin{aligned}
M_t(x) &= \exp\left(\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s \leq t\right\}\left(\frac{\langle x, X_s\rangle \eta_s}{R} - \frac{1}{2}\langle x, X_s\rangle^2\right)\right) \\
&= \exp\left(\sum_{s=1}^{t} C_s^{t-1}\left(\frac{\langle x, X_s\rangle \eta_s}{R} - \frac{1}{2}\langle x, X_s\rangle^2\right) + \sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s > t-1\right\}\left(\frac{\langle x, X_s\rangle \eta_s}{R} - \frac{1}{2}\langle x, X_s\rangle^2\right)\right) \\
&= \exp\left(\sum_{s=1}^{t} C_s^{t-1}\left(\frac{\langle x, X_s\rangle \eta_s}{R} - \frac{1}{2}\langle x, X_s\rangle^2\right)\right)\exp\left(\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s > t-1\right\}\left(\frac{\langle x, X_s\rangle \eta_s}{R} - \frac{1}{2}\langle x, X_s\rangle^2\right)\right) \\
&= \exp\left(\sum_{s=1}^{t-1} C_s^{t-1}\left(\frac{\langle x, X_s\rangle \eta_s}{R} - \frac{1}{2}\langle x, X_s\rangle^2\right)\right)\exp\left(\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s > t-1\right\}\left(\frac{\langle x, X_s\rangle \eta_s}{R} - \frac{1}{2}\langle x, X_s\rangle^2\right)\right) \\
&= M_{t-1}(x)\exp\left(\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s > t-1\right\}\left(\frac{\langle x, X_s\rangle \eta_s}{R} - \frac{1}{2}\langle x, X_s\rangle^2\right)\right)
\end{aligned}$$

where the penultimate equality follows from the fact that $\mathbb{1}\{t + \tau_t \leq t-1\} = 0$ as the delays are non-negative random variables, allowing us to stop the first summation at round $t-1$ by pulling the corresponding $\exp(0)$ out of the the summation and utilising that $\exp(a+b) = \exp(a)\exp(b)$.

Recall that $C_s^{t-1} = \mathbb{1}\left\{s + \tau_s \leq t-1\right\}$ is $\mathcal{F}_{t-1}$-measurable. Since $\mathcal{F}_{t-1}$ is a σ-algebra, $\mathbb{1}\left\{s + \tau_s > t-1\right\}$ must also be measurable. Utilising this fact, it is clear that everything except the noise terms are $\mathcal{F}_{t-1}$-measurable. Assumption 1 guarantees the noise is subgaussian, therefore:

$$\begin{aligned}
&\mathbb{E}\left[M_t(x) \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[M_{t-1}(x)\exp\left(\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s > t-1\right\}\left(\frac{\langle x, X_s\rangle \eta_s}{R} - \frac{1}{2}\langle x, X_s\rangle^2\right)\right) \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= M_{t-1}(x)\,\mathbb{E}\left[\exp\left(\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s > t-1\right\}\left(\frac{\langle x, X_s\rangle \eta_s}{R} - \frac{1}{2}\langle x, X_s\rangle^2\right)\right) \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= M_{t-1}(x)\exp\left(-\frac{1}{2}\sum_{s=1}^{t} \mathbb{1}\left\{s + \tau_s > t-1\right\}\langle x, X_s\rangle^2\right)\,\mathbb{E}\left[\exp\left(\sum_{s=1}^{t} \frac{\mathbb{1}\left\{s + \tau_s > t-1\right\}\langle x, X_s\rangle \eta_s}{R}\right) \,\middle|\, \mathcal{F}_{t-1}\right]
\end{aligned}$$

$$\leq M_{t-1}(x)\exp\left(-\frac{1}{2}\sum_{s=1}^{t}\mathbb{1}\left\{s+\tau_s>t-1\right\}\langle x,X_s\rangle^2\right)\exp\left(\frac{1}{2}\sum_{s=1}^{t}\mathbb{1}\left\{s+\tau_s>t-1\right\}\langle x,X_s\rangle^2\right)$$
$$= M_{t-1}(x)$$

showing that $\{M_t(x)\}_{t=0}^{\infty}$ is indeed a non-negative supermartingale.[1] For conciseness, denote:

$$P_t = \exp\left(\sum_{s=1}^{t}\mathbb{1}\left\{s+\tau_s>t-1\right\}\left(\frac{\langle x,X_s\rangle\,\eta_s}{R}-\frac{1}{2}\langle x,X_s\rangle^2\right)\right)$$

Then, by the law of total expectation:

$$\begin{aligned}\mathbb{E}\left[M_t(x)\right] &= \mathbb{E}\left[M_{t-1}(x)\,P_t\right] = \mathbb{E}\left[\mathbb{E}\left[M_{t-1}(x)\,P_t\,|\,\mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[M_{t-1}(x)\,\mathbb{E}\left[P_t\,|\,\mathcal{F}_{t-1}\right]\right]\\ &\leq \mathbb{E}\left[M_{t-1}(x)\right] = \mathbb{E}\left[M_{t-2}(x)\,P_{t-1}\right] = \mathbb{E}\left[\mathbb{E}\left[M_{t-2}(x)\,P_{t-1}\,|\,\mathcal{F}_{t-2}\right]\right] = \mathbb{E}\left[M_{t-2}(x)\,\mathbb{E}\left[P_{t-1}\,|\,\mathcal{F}_{t-2}\right]\right]\\ &\vdots\\ &\leq \mathbb{E}\left[M_1(x)\right] = \mathbb{E}\left[\mathbb{E}\left[M_1(x)\,|\,\mathcal{F}_0\right]\right]\\ &\leq 1\end{aligned}$$

where we define $M_0(x)=1$. By the convergence theorem for non-negative supermartingales:

$$M_{\infty}^{x} = \lim_{t\to\infty} M_t^x$$

is almost surely well-defined. Hence, M_{ω}^{x} is almost surely well-defined, regardless of whether the stopping time is finite or not. Now, Fatou's Lemma tells us that:

$$\mathbb{E}\left[M_{\omega}(x)\right] = \mathbb{E}\left[\liminf_{t\to\infty} M_{\min\{t,\omega\}}(x)\right] \leq \liminf_{t\to\infty}\mathbb{E}\left[M_{\min\{t,\omega\}}(x)\right]$$

Combining the right hand side of the above with the law of total expectation reveals that for any $t\geq 0$:

$$\begin{aligned}\mathbb{E}\left[M_{\min\{t,\omega\}}(x)\right] = \mathbb{E}\left[\mathbb{E}\left[M_{\min\{t,\omega\}}(x)\,|\,\mathcal{F}_{t-1}\right]\right] &\leq \mathbb{E}\left[M_{\min\{t-1,\omega\}}(x)\right] = \mathbb{E}\left[\mathbb{E}\left[M_{\min\{t-1,\omega\}}(x)\,|\,\mathcal{F}_{t-1}\right]\right]\\ &\vdots\\ &\leq \mathbb{E}\left[M_{\min\{0,\omega\}}(x)\right] = \mathbb{E}\left[M_0(x)\right] = 1\end{aligned}$$

Therefore, $\mathbb{E}[M_{\omega}^{x}]\leq 1$, as required. □

B MISSING PROOFS OF THEOREM 1

Proving Theorem 1 required the introduction of four technical lemmas. These lemmas are crucial in showing that our algorithms only suffer from an additive penalty due to the delays.

Lemma 2. *Let $\lambda = \alpha\, a(\phi)/\kappa > 0$. Then, $\bar{W}_t$ and $\bar{V}_t$ are invertible and have inverses that satisfy the following relationship:*

$$\bar{W}_t^{-1} = \bar{V}_t^{-1} + \bar{V}_t^{-1} Z_t\,\bar{W}_t^{-1} = \bar{V}_t^{-1} + M_t$$

where $M_t \coloneqq \bar{V}_t^{-1} Z_t\,\bar{W}_t^{-1}$.

Proof. From Equations (3), (4) and (5), and $\lambda I \succ 0$, we have that the total and observed gram matrices are symmetric and positive-definite. They are symmetric because they are the sum of symmetric matrices. And they are positive-definite because they are the sum of a positive-definite matrix and a positive semi-definite matrix. Thus, the first part of the lemma follows from the fact that all symmetric positive-definite matrices are invertible.

Next, we move on to the second claim of the lemma. From Equations (3), (4) and (5), we have that the total, observed, and missing design matrices satisfy the following relationship:

$$\bar{V}_t = \lambda I + \sum_{s=1}^{t} X_s X_s^T\left(\mathbb{1}\{s+\tau_s\leq t\} + \mathbb{1}\{s+\tau_s>t\}\right) = \bar{W}_t + Z_t \tag{20}$$

[1] By definition, the exponential function is always positive. Hence, $M_t(x)$ is always non-negative.

We prove the second statement in the lemma as follows:

$$\begin{aligned}\bar{W}_t^{-1} &= \bar{V}_t^{-1} + \bar{W}_t^{-1} - \bar{V}_t^{-1} \\ &= \bar{V}_t^{-1} + \bar{V}_t^{-1}\bar{V}_t\bar{W}_t^{-1} - \bar{V}_t^{-1}\bar{W}_t\bar{W}_t^{-1} \\ &= \bar{V}_t^{-1} + \bar{V}_t^{-1}\left(\bar{V}_t - \bar{W}_t\right)\bar{W}_t^{-1} \\ &= \bar{V}_t^{-1} + \bar{V}_t^{-1} Z_t\, \bar{W}_t^{-1},\end{aligned}$$

where the final equality follows from rearranging Equation (20). ■

Lemma 3. *Let $\{\tau_t\}_{t=1}^{\infty}$ be an arbitrary sequence of non-negative random variables. Then, for $\lambda = \alpha\, a(\phi)/\kappa \geq 1$:*

$$\sum_{t=1}^{T} \|X_t\|_{M_{t-1}} \leq \sum_{t=1}^{T} \frac{1 + G_* + \tau_t}{2} \|X_t\|^2_{\bar{V}_{t-1}^{-1}}$$

where $G_* = \max\{G_t : t \leq T\}$.

Proof. Firstly, we rewrite the norm as follows:

$$\begin{aligned}\|X_t\|_{M_{t-1}} &= \|X_t\|_{\bar{V}_{t-1}^{-1} Z_{t-1} \bar{W}_{t-1}^{-1}} \\ &= \sqrt{X_t^T \bar{V}_{t-1}^{-1} Z_{t-1} \bar{W}_{t-1}^{-1} X_t} \\ &= \sqrt{\operatorname{Tr}\left(X_t^T \bar{V}_{t-1}^{-1} Z_{t-1} \bar{W}_{t-1}^{-1} X_t\right)} \\ &= \sqrt{\operatorname{Tr}\left(\bar{V}_{t-1}^{-1} Z_{t-1} \bar{W}_{t-1}^{-1} X_t X_t^T\right)}\end{aligned}$$

Let $A = \bar{V}_{t-1}^{-1} Z_{t-1}$ and $B = \bar{W}_{t-1} X_t X_t^T$. Then Lemma 9 guarantees that A and B have non-negative eigenvalues, meaning:

$$\operatorname{Tr}(AB) = \operatorname{Tr}(AB^{1/2}B^{1/2}) = \operatorname{Tr}(B^{1/2}AB^{1/2}) \leq \operatorname{Tr}(B^{1/2}(\operatorname{Tr}(A))IB^{1/2}) = \operatorname{Tr}(A)\operatorname{Tr}(B)$$

Therefore

$$\begin{aligned}\|X_t\|_{M_{t-1}} &\leq \sqrt{\operatorname{Tr}\left(\bar{W}_{t-1}^{-1} X_t X_t^T\right)\operatorname{Tr}\left(\bar{V}_{t-1}^{-1} Z_{t-1}\right)} \\ &\leq \frac{1}{2}\operatorname{Tr}\left(\bar{W}_{t-1}^{-1} X_t X_t^T\right) + \frac{1}{2}\operatorname{Tr}\left(\bar{V}_{t-1}^{-1} Z_{t-1}\right) && \text{(AM-GM Inequality)} \\ &= \frac{1}{2}\|X_t\|^2_{\bar{W}_{t-1}^{-1}} + \frac{1}{2}\sum_{s-1}^{t-1} \mathbb{1}\left\{s + \tau_s > t-1\right\} \|X_s\|^2_{\bar{V}_{t-1}^{-1}} && \text{(Equation (5))} \\ &\leq \frac{1 + G_{t-1}}{2}\|X_t\|^2_{\bar{V}_{t-1}^{-1}} + \frac{1}{2}\sum_{s-1}^{t-1} \mathbb{1}\left\{s + \tau_s > t-1\right\} \|X_s\|^2_{\bar{V}_{t-1}^{-1}} && \text{(Lemma 2 \& 11 and } \lambda \geq 1) \\ &\leq \frac{1 + G_*}{2}\|X_t\|^2_{V_{t-1}^{-1}} + \frac{1}{2}\sum_{s=1}^{t-1} \mathbb{1}\left\{s + \tau_s > t-1\right\} \|X_s\|^2_{V_{t-1}^{-1}} && (G_t \leq G_*)\end{aligned}$$

Now, we are ready to reintroduce the outer summation. Doing so gives:

$$\begin{aligned}\sum_{t=1}^{T}\|X_t\|_{M_{t-1}} &\leq \frac{1 + G_*}{2}\sum_{t=1}^{T}\|X_t\|^2_{V_{t-1}^{-1}} + \frac{1}{2}\sum_{t=1}^{T}\sum_{s=1}^{t-1} \mathbb{1}\left\{s + \tau_s > t-1\right\} \|X_s\|^2_{V_{t-1}^{-1}} \\ &\leq \frac{1 + G_*}{2}\sum_{t=1}^{T}\|X_t\|^2_{V_{t-1}^{-1}} + \frac{1}{2}\sum_{t=1}^{T}\sum_{s=1}^{t-1} \mathbb{1}\left\{s + \tau_s > t-1\right\} \|X_s\|^2_{V_{s-1}^{-1}} && (\bar{V}_t \succeq V_s \text{ for } t \geq s) \\ &\leq \frac{1 + G_*}{2}\sum_{t=1}^{T}\|X_t\|^2_{V_{t-1}^{-1}} + \frac{1}{2}\sum_{t=1}^{T} \tau_t \|X_t\|^2_{V_{t-1}^{-1}}\end{aligned}$$

The final equality follows from expanding the two summations and realising that the indicator ensures each term contributes to the summation τ_t times. Simply rearranging the above terms gives the final result. ■

Lemma 4. *Define $G_t = \sum_{s=1}^{t} \mathbb{1}\{s + \tau_s > t\}$ and let $\{\tau_t\}_{t=1}^{\infty}$ be a sequence of independent and identically distributed random variables with a finite expectation and define:*

$$\psi_\tau^t \coloneqq \frac{4}{3}\log\left(\frac{3t}{2\delta}\right) + 2\sqrt{2\mathbb{E}\left[\tau\right]\log\left(\frac{3t}{2\delta}\right)}.$$

Then,

$$\mathbb{P}\left(\exists\, t \geq 1 : G_t \leq \mathbb{E}[\tau] + \psi_\tau^t\right) \leq 1 - \delta.$$

Proof. The proof of this claim is similar to that found in work done on multi-armed bandits (Joulani et al., 2013). However, we extend the result so it holds for continuous delay distributions. Bernstein's inequality gives the following tail bound on sums of subgaussian random variables:

$$\mathbb{P}\left(G_t - \mathbb{E}\left[G_t\right] \geq \frac{2}{3}\log\left(\frac{1}{\delta'}\right) + 2\sqrt{\mathbb{V}\left[G_t\right]\log\left(\frac{1}{\delta'}\right)}\right) \leq \delta'$$

Setting $\delta' = 6\delta/\pi^2 t^2$ and taking a union bound over all possible rounds reveals that:

$$\mathbb{P}\left(\exists\, t \in \mathbb{N}_1 : G_t - \mathbb{E}\left[G_t\right] \geq \frac{2}{3}\log\left(\frac{1}{\delta'}\right) + 2\sqrt{\mathbb{V}\left[G_t\right]\log\left(\frac{1}{\delta'}\right)}\right) \leq \sum_{t=1}^{\infty}\delta' = \frac{6\delta}{\pi^2}\sum_{t=1}^{\infty}\frac{1}{t^2} = \delta$$

Therefore, with probability $1-\delta$:

$$G_t \leq \mathbb{E}\left[G_t\right] + \frac{4}{3}\log\left(\frac{2t}{3\delta}\right) + 2\sqrt{2\mathbb{V}\left[G_t\right]\log\left(\frac{2t}{3\delta}\right)}$$

for any $t \in \mathbb{N}_1$. All that remains is to show that expectation and variance of the number of missing feedbacks is smaller than the expected delay. By assumption, the delays are independent. Therefore, each of the indicator variables involved in the definition of G_t are independent. Considering its expectation reveals that:

$$\begin{aligned}
\mathbb{E}\left[G_t\right] &= \sum_{s=1}^{t}\mathbb{E}\left[\mathbb{1}\left\{s + \tau_s > t\right\}\right] = \sum_{s=1}^{t}\mathbb{P}\left[s + \tau_s > t\right] = \sum_{i=0}^{t-1}\mathbb{P}\left[\,\tau_{t-i} > i\,\right] \\
&\leq \sum_{i=0}^{\infty}\mathbb{P}\left[\,\tau > i\,\right] = \sum_{i=0}^{\infty}\sum_{j=i+1}^{\infty}\mathbb{P}\left[\,\tau = j\,\right] = \sum_{j=1}^{\infty}\sum_{i=0}^{j-1}\mathbb{P}\left[\,\tau = j\,\right] \\
&= \sum_{j=1}^{\infty} j\,\mathbb{P}\left[\,\tau = j\,\right] = \mathbb{E}\left[\tau\right]
\end{aligned}$$

for discrete delay distributions. For continuous delay distributions, we can obtain a similar result by utilising the fact that the complement of the cumulative distribution function is non-increasing:

$$\begin{aligned}
\mathbb{E}\left[G_t\right] &= \sum_{s=1}^{t}\mathbb{E}\left[\mathbb{1}\left\{s + \tau_s > t\right\}\right] = \sum_{s=1}^{t}\mathbb{P}\left[\tau_s > t - s\right] = \sum_{x=0}^{t-1}\mathbb{P}\left[\tau > x\right] \\
&\leq 1 + \int_0^t \mathbb{P}\left[\,\tau > x\,\right]dx = 1 + \int_0^t\int_x^\infty f_\tau(y)\,dy\,dx && \text{(Setting } x = t - s\text{)} \\
&\leq 1 + \int_0^\infty\int_x^\infty f_\tau(x)\,dy\,dx = 1 + \int_0^\infty\int_0^y f_\tau(y)\,dx\,dy && \text{(Tonelli's Theorem)} \\
&= 1 + \int_0^\infty \left[x f_\tau(y)\right]_0^y dy = 1 + \int_0^\infty y f_\tau(y) dy \\
&= 1 + \mathbb{E}\left[\tau\right]
\end{aligned}$$

Similarly, looking at the variance reveals that:

$$\mathbb{V}\left[G_t\right] = \sum_{s=1}^{t}\mathbb{V}\left[\mathbb{1}\left\{s + \tau_s \geq t\right\}\right] \leq \sum_{s=1}^{t}\mathbb{E}\left[\mathbb{1}\left\{s + \tau_s \geq t\right\}^2\right] = \mathbb{E}\left[G_t\right],$$

which is smaller than the expected delay. Therefore,

$$G_t \leq 1 + \mathbb{E}\left[\tau\right] + \frac{4}{3}\log\left(\frac{2t}{3\delta}\right) + 2\sqrt{2\mathbb{E}\left[G_t\right]\log\left(\frac{2t}{3\delta}\right)}$$

as required. ■

B.1 Supporting Lemmas

Proving Lemma 3 requires Lemma 11, which itself requires two additional results. We state and prove all three of these results in this subsection.

Lemma 9. *Let $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times d}$ be two symmetric positive semi-definite matrices. Then, $A^{1/2}BA^{1/2}$ and AB share the same set of eigenvalues. Further, these eigenvalues are all non-negative.*

Proof. Since A is positive semi-definite, we can utilise the spectral decomposition to show that: $AB = A^{1/2}A^{1/2}B$. Suppose AB has an eigenvalue equal to λ. Then, there exists a non-zero eigenvector such that:

$$AB\vec{v} = A^{1/2}A^{1/2}B\vec{v} = \lambda\vec{v}$$

Pre-multiplying both sides of the above equation by the same matrix gives:

$$A^{1/2}BAB\vec{v} = A^{1/2}BA^{1/2}\left(A^{1/2}B\right)\vec{v} = \lambda\left(A^{1/2}B\right)\vec{v}$$

Thus, AB and $A^{1/2}BA^{1/2}$ share the same set of eigenvalues, albeit with different eigenvectors, verifying the first statement of the lemma. Now, $A^{1/2}BA^{1/2}$ is symmetric, because:

$$\left(A^{1/2}BA^{1/2}\right)^T = (A^{1/2})^T B^T (A^{1/2})^T = A^{1/2}BA^{1/2}$$

Further, it is positive semi-definite, because:

$$\underbrace{x^T A^{1/2}}_{\tilde{x}^T} B \underbrace{A^{1/2}x}_{\tilde{x}} = \tilde{x}^T B\tilde{x} \geq 0$$

The final inequality follows from the fact that B is positive semi-definite. Therefore, $A^{1/2}BA^{1/2}$ must have non-negative eigenvalues, as it is symmetric and positive semi-definite. Recall AB and $A^{1/2}BA^{1/2}$ shares the same set of eigenvalues. Therefore, AB has non-negative eigenvalues too. □

Lemma 10. *Let Z_t and G_t be the missing design matrix and the number of missing feedbacks at the end of the t-th round, respectively. Then, $\lambda_1(Z_t) \leq G_t$.*

Proof. By Equations (3), (4) and (5), we have that:

$$Z_t = \sum_{s\leq t} \mathbb{1}\{s + \tau_s > t\} X_s X_s^T,$$

Clearly, $X_s X_s^T$ is a symmetric matrix, as it is the outer product of two vectors. The Courant–Fischer–Weyl min-max principle shows that:

$$\lambda_1\left(X_s X_s^T\right) \leq \|X_s\|_2^2 \leq 1,$$

where the final inequality follows from assuming that the vectors are appropriately normalised. Applying Weyl's inequality repeatedly to each symmetric matrix in the summation and utilising the above result gives:

$$\lambda_1\left(Z_t\right) \leq \sum_{s\leq t} \mathbb{1}\{s + \tau_s > t\}\lambda_1\left(I\right) = G_t$$

as required. □

Lemma 11. *Let $\bar{V}_t$, $\bar{W}_t$ and Z_t be the total, observed and missing gram matrices, respectively. Then,*

$$\frac{G_t}{\lambda}\,\bar{V}_t^{-1} \succeq \bar{V}_t^{-1} Z_t\,\bar{W}_t^{-1} = M_t$$

Proof. Firstly, $A \succeq B \iff A - B \succeq 0$. Therefore, we focus on proving that the difference between the two matrices is positive semi-definite. That is, we prove that:

$$D_t = \frac{G_t}{\lambda}\,\bar{V}_t^{-1} - \bar{V}_t^{-1} Z_t\,\bar{W}_t^{-1} \tag{21}$$

is positive semi-definite. To do so, we take a four-stepped approach. Below is an overview of these four steps.

1. Firstly, we show that the matrix of (21) is symmetric.
2. Next, we define a similar matrix and prove that it has the same set of eigenvalues as that of (21).
3. Then, we show that all the eigenvalues of the similar matrix are non-negative.
4. Finally, we chain the above three steps in reverse order and recall basic facts about symmetric matrices to prove the claim.

Step 1. Indeed, D_t is the difference of two symmetric matrices. From Equations (3), (4) and (5), the first matrix is symmetric, as it is just the total gram matrix scaled by a constant. Also, the second matrix is symmetric because it is the difference between the two symmetric matrices:

$$\bar{V}_t^{-1} Z_t \bar{W}_t^{-1} = \bar{V}_t^{-1}\left(\bar{V}_t - \bar{W}_t\right)\bar{W}_t^{-1} = \bar{W}_t^{-1} - \bar{V}_t^{-1}$$

Therefore, D_t is symmetric, as it is the difference between two symmetric matrices. Indeed, a symmetric matrix must have all non-negative eigenvalues for positive semi-definiteness to hold. Thus, it is sufficient to find a matrix with the same eigenvalues and show that its quadratic form is greater than or equal to zero, for which non-negative eigenvalues is a necessary condition.

Step 2. To that end, we define the following matrix:

$$\tilde{D}_t \coloneqq \bar{V}_t^{-1/2}\left(\frac{G_t}{\lambda} I - Z_t\,\bar{W}_t^{-1}\right)\bar{V}_t^{-1/2}$$

Applying Lemma 9 with $A = \bar{V}_t$ and $B = (G_t/\lambda)I - Z_t\bar{W}_t^{-1}$ reveals that D_t and $\tilde{D}_t$ share the same set of eigenvalues, albeit with different eigenvectors.

Step 3. Showing $\tilde{D}_t \succeq 0$ proves it must have non-negative eigenvalues, as this is a necessary condition for the positive semi-definiteness of an arbitrary (possibly non-symmetric) matrix. Utilising the definition of positive semi-definiteness, we can verify whether or not this holds by checking if: $x^T\tilde{D}_t x \geq 0$. To do so, we first decompose the matrix into the sum of symmetric and anti-symmetric matrices:

$$\tilde{D}_t = \frac{1}{2}\left(\tilde{D}_t + \tilde{D}_t^T\right) + \frac{1}{2}\left(\tilde{D}_t - \tilde{D}_t^T\right),$$

Then, we use the fact that:

$$y = x^T\left(\tilde{D}_t - \tilde{D}_t^T\right)x = \left(x^T\left(\tilde{D}_t - \tilde{D}_t^T\right)x\right)^T = x^T\left(\tilde{D}_t - \tilde{D}_t^T\right)^T x = -x^T\left(\tilde{D}_t - \tilde{D}_t^T\right)x = -y\,,$$

which holds if and only if $y = 0$. Doing so gives:

$$\begin{aligned} x^T\tilde{D}_t x &= \frac{1}{2}x^T\left(\tilde{D}_t + \tilde{D}_t^T\right)x + \frac{1}{2}x^T\left(\tilde{D}_t - \tilde{D}_t^T\right)x = \frac{1}{2}x^T\left(\tilde{D}_t + \tilde{D}_t^T\right)x \\ &= \frac{1}{2}x^T\left(\bar{V}_t^{-1/2}\left(\frac{2G_t}{\lambda} I - Z_t\,\bar{W}_t^{-1} - \bar{W}_t^{-1}Z_t\right)\bar{V}_t^{-1/2}\right)x \end{aligned}$$

$$= \frac{G_t}{\lambda} x^T \bar{V}_t^{-1} x - \frac{1}{2} x^T \bar{V}_t^{-1/2} \left(Z_t \bar{W}_t^{-1} + \bar{W}_t^{-1} Z_t \right) \bar{V}_t^{-1/2} x$$
$$= \frac{G_t}{\lambda} \|x\|^2_{\bar{V}_t^{-1}} - \frac{1}{2} x^T \bar{V}_t^{-1/2} \left(Z_t \bar{W}_t^{-1} + \bar{W}_t^{-1} Z_t \right) \bar{V}_t^{-1/2} x$$

Now, $A = Z_t \bar{W}_t^{-1} + \bar{W}_t^{-1} Z_t$ is a real-value symmetric matrix. Therefore, its eigendecomposition is given by $A = Q \Lambda Q^T$ where Λ is a diagonal matrix containing its eigenvalues and Q is an orthogonal matrix whose columns contain its unit eigenvectors.

$$\begin{aligned} x^T \tilde{D}_t x &= \frac{G_t}{\lambda} \|x\|^2_{\bar{V}_t^{-1}} - \frac{1}{2} x^T \bar{V}_t^{-1/2} Q \Lambda Q^T \bar{V}_t^{-1/2} x \\ &= \frac{G_t}{\lambda} \|x\|^2_{\bar{V}_t^{-1}} - \frac{1}{2} \tilde{x}^T \Lambda \tilde{x} && \text{(Setting } \tilde{x} = Q^T \bar{V}_t^{-1/2} x\text{)} \\ &= \frac{G_t}{\lambda} \|x\|^2_{\bar{V}_t^{-1}} - \frac{1}{2} \sum_{i=1}^{d} \lambda_i \tilde{x}_i^2 && (22) \end{aligned}$$

where λ_i is the i-th largest eigenvalue of matrix $A = Z_t \bar{W}_t^{-1} + \bar{W}_t^{-1} Z_t$. Indeed, A is a symmetric matrix positive definite matrix. It is symmetric because it is the sum of a matrix and its transpose and it is positive semi-definite because:

$$\|x\|^2_A = 2 \underbrace{x^T Z_t}_{\in \mathbb{R}^d} (\bar{W}_t^{-1}) \underbrace{Z_t x}_{\in \mathbb{R}^d}$$

and $\bar{W}_t^{-1} \succ 0$. Thus, it follows that the singular values A are the absolute values of its eigenvalues. Define σ_i as the i-th largest singular value of matrix $Z_t W_t \left(\lambda_*\right)^{-1} + W_t \left(\lambda_*\right)^{-1} Z_t$. Then, we have that:

$$\begin{aligned} (22) &\geq \frac{G_t}{\lambda} \|x\|^2_{V_t(\lambda_*)^{-1}} - \frac{1}{2} \sum_{i=1}^{d} \sigma_i \tilde{x}_i^2 && \text{(Since } x \leq |x|\text{)} \\ &\geq \frac{G_t}{\lambda} \|x\|^2_{\bar{V}_t^{-1}} - \frac{1}{2} \sigma_{\max} \|x\|^2_{\bar{V}_t^{-1}} && \text{(Definition of } \tilde{x}\text{)} \\ &= \frac{G_t}{\lambda} \|x\|^2_{\bar{V}_t^{-1}} - \frac{1}{2} \left\| Z_t \bar{W}_t^{-1} + \bar{W}_t^{-1} Z_t \right\|_2 \|x\|^2_{\bar{V}_t^{-1}} && \text{(For } A \in \mathbb{R}^{d \times d}\text{: } \sigma_{\max} = \|A\|_2\text{)} \\ &\geq \frac{G_t}{\lambda} \|x\|^2_{\bar{V}_t^{-1}} - \frac{1}{2} \left(\|Z_t \bar{W}_t^{-1}\|_2 + \|\bar{W}_t^{-1} Z_t\|_2 \right) \|x\|^2_{\bar{V}_t^{-1}} && \text{(Matrix Norm is Sub-additive)} \\ &\geq \frac{G_t}{\lambda} \|x\|^2_{\bar{V}_t^{-1}} - \|Z_t\|_2 \|\bar{W}_t^{-1}\|_2 \|x\|^2_{\bar{V}_t^{-1}} && \text{(Matrix Norm is Sub-multiplicative)} \end{aligned}$$

By definition, $\|\bar{W}_t^{-1}\|_2 \leq 1/\lambda$. Further, Lemma 10 tells us that $\|Z_t\|_2 \leq G_t$. Substituting this into the above gives:

$$\begin{aligned} x^T \tilde{D}_t x &\geq \frac{G_t}{\lambda} \|x\|^2_{\bar{V}_t^{-1}} - \|Z_t\|_2 \|\bar{W}_t^{-1}\|_2 \|x\|^2_{\bar{V}_t^{-1}} \\ &\geq \frac{G_t}{\lambda} \|x\|^2_{\bar{V}_t^{-1}} - \frac{G_t}{\lambda_*} \|x\|^2_{\bar{V}_t^{-1}} \\ &= 0 \end{aligned}$$

Step 4. Now, Step 3 shows $\tilde{D}_t$ is positive semi-definite, implying all of its eigenvalues are non-negative. Therefore, Step 2 shows D_t has non-negative eigenvalues. Finally, Step 1 shows D_t is a symmetric matrix. Since D_t is symmetric, non-negative eigenvalues implies positive semi-definiteness. Finally, $D_t \succeq 0$ implies that:

$$\frac{G_t}{\lambda} \bar{V}_t^{-1} \succeq \bar{V}_t^{-1} Z_t \bar{W}_t = M_t \,,$$

as required. ■

C ADDITIONAL EXPERIMENTAL RESULTS

Here, we present additional experimental results for the linear and logistic bandits under delayed feedback with delays drawn from the uniform and Pareto distributions.

C.1 Uniform Delays

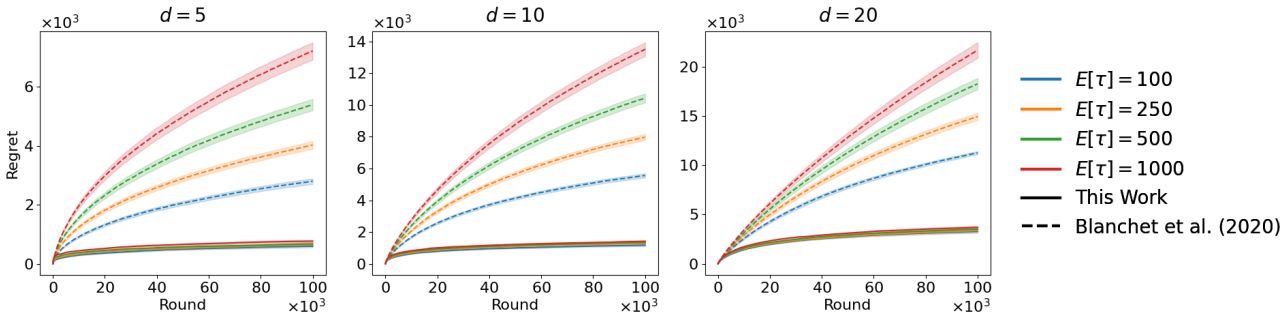

Figure 5: Linear Bandit & Uniform Delays.

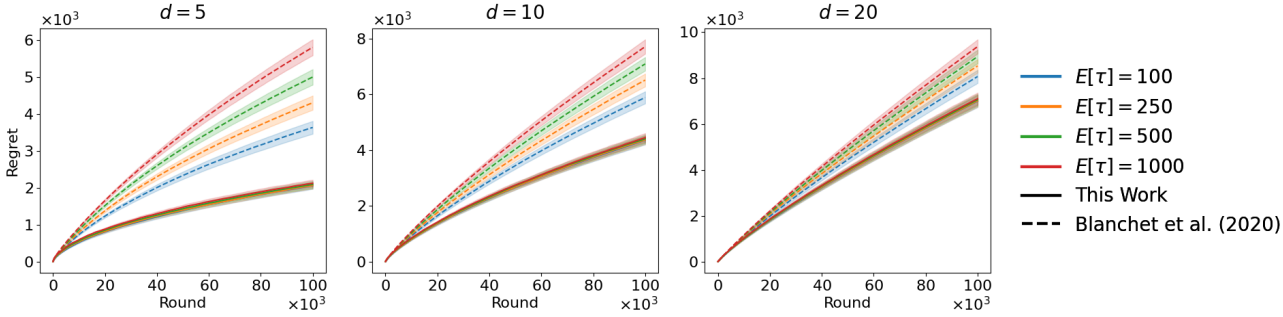

Figure 6: Logistic Bandit & Uniform Delays.

C.2 Pareto Delays

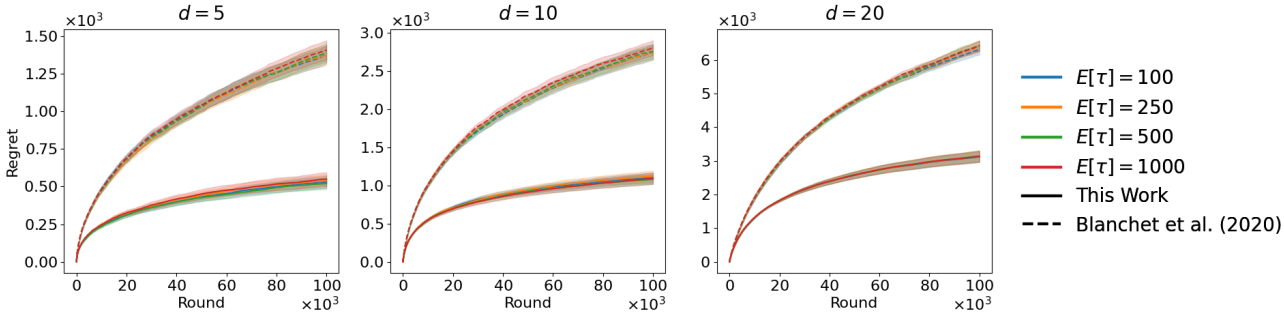

Figure 7: Linear Bandit & Pareto Delays.

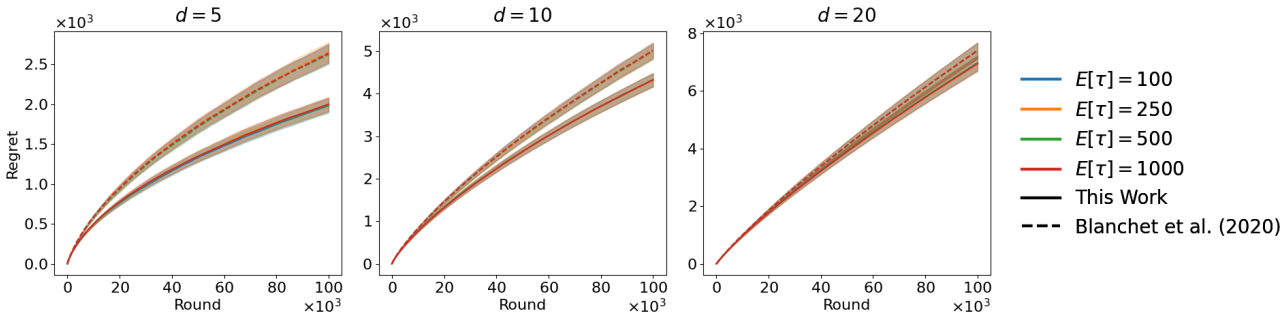

Figure 8: Logistic Bandit & Pareto Delays

D STANDARD RESULTS

Here, we present a selection of well-known tail bounds for subgaussian and subexponential random variables that find use in our paper. Additionally, we provide proof of the elliptical potential lemma.

Lemma 12 (Bernstein's Inequality). *Let $\{X_t\}_{t=1}^n$ be a sequence of independent and identically distributed σ_t-subgaussian random variables. Define $S_n = X_1 + X_2 + \cdots + X_n$. Then,*

$$\mathbb{P}\left(S_n - \mathbb{E}\left[S_n\right] \geq \frac{2}{3}\log\left(\frac{1}{\delta}\right) + 2\sqrt{\mathbb{V}\left[S_n\right]\log\left(\frac{1}{\delta}\right)}\right) \leq \delta$$

Lemma 13 (Subexponential Tail Bounds). *Suppose $\{\tau_t\}_{t=1}^\infty$ are (v, α) subexponential random variables. Then,*

$$\mathbb{P}\left(\tau_t - \mathbb{E}\left[\tau_t\right] \geq \epsilon\right) \leq \exp\left(-\frac{1}{2}\min\left\{\frac{\epsilon^2}{v^2}, \frac{\epsilon}{\alpha}\right\}\right)$$

Therefore, with probability $1-\delta$:

$$\tau_t \leq \mathbb{E}\left[\tau_t\right] + \min\left\{\sqrt{2v^2\log\left(2t/3\delta\right)}, 2\alpha\log\left(2t/3\delta'\right)\right\}$$

for any $t \in \mathbb{N}_1$.

D.1 Elliptical Potential Lemma

Lemma 14 (Elliptical Potential Lemma). *Let $\{X_t\}_{t=1}^\infty$ be an arbitrary sequence of d-dimensional vectors such that $\|X_t\|_2^2 \leq 1$. Define $V_0 = \lambda I$, $V_t = \sum_{s=1}^t X_s X_s^T$ and $\bar{V}_t = V_0 + V_t$. Then,*

$$\sum_{t=1}^T \|X_t\|^2_{\bar{V}_{t-1}^{-1}} \leq 2\log\left(\frac{\det\left(\bar{V}_T\right)}{\det\left(\bar{V}_0\right)}\right) \leq 2d\log\left(\frac{d\lambda + T}{d\lambda}\right)$$

for $\lambda \geq 1/2$.

Proof. For completeness, we provided a detailed proof of the elliptical potential lemma using the arguments of Abbasi-Yadkori et al. (2011). However, we note that this is not the only way to obtain the stated result. Carpentier et al. (2020) prove the lemma using insights from linear algebra.

Firstly, notice that:

$$\|x\|^2_{\bar{V}_t} = \lambda\|x\|_2^2 + \sum_{s=1}^t \left(x^T X_s\right)\left(X_s^T x\right) = \lambda\|x\|_2^2 + \sum_{s=1}^t \|X_s^T x\|_2^2 \geq \lambda\|x\|_2^2 > 0 \implies \|x\|^2_{\bar{V}_t^{-1}} > 0$$

Additionally, $\lambda \geq 1/2$ and $\|X_t\|_2^2 \leq 1$. Therefore,

$$\|x\|^2_{\bar{V}_t^{-1}} \leq \|x\|^2_{\bar{V}_0^{-1}} = \frac{1}{\lambda}\|x\|_2^2 \leq \frac{1}{\lambda}$$

Consequently,

$$0 < \|X_t\|^2_{\bar{V}_t^{-1}} \leq \frac{1}{\lambda} \leq 2$$

Since $x < 2\ln(1+x)$ for any $0 < x \leq 2$, we have that:

$$\sum_{t=1}^{T} \|X_t\|^2_{\bar{V}_t^{-1}} \leq 2\sum_{t=1}^{T} \log\left(1 + \|X_t\|^2_{\bar{V}_t^{-1}}\right) = 2\log\left(\prod_{t=1}^{T}\left(1 + \|X_t\|^2_{\bar{V}_t^{-1}}\right)\right) \tag{23}$$

Now, proving the first inequality amounts to relating the term inside the logarithm to the determinants of the matrices. By Definition, we have that:

$$\bar{V}_t = \bar{V}_{t-1} + X_t X_t^T = \bar{V}_{t-1}^{1/2}\left(I + \bar{V}_{t-1}^{-1/2} X_t X_t^T \bar{V}_{t-1}^{-1/2}\right)\bar{V}_{t-1}^{1/2}$$

and

$$\begin{aligned}
\det\left(\bar{V}_n\right) &= \det\left(\bar{V}_{n-1}^{1/2}\left(I + \bar{V}_{n-1}^{-1/2} X_n X_n^T \bar{V}_{n-1}^{-1/2}\right)\bar{V}_{n-1}^{1/2}\right) \\
&= \det\left(\bar{V}_{n-1}\right)\det\left(I + \bar{V}_{n-1}^{-1/2} X_n X_n^T \bar{V}_{n-1}^{-1/2}\right) && \text{(Properties of Determinants)} \\
&= \det\left(\bar{V}_{n-1}\right)\left(1 + X_n^T \bar{V}_{n-1}^{-1} X_n\right) && \text{(Matrix Determinant Lemma)} \\
&= \det\left(\bar{V}_{n-1}\right)\left(1 + \|X_n\|^2_{\bar{V}_{n-1}^{-1}}\right) && \text{(By Positive Definiteness)} \\
&= \det\left(V_0\right)\prod_{t=1}^{n}\left(1 + \|X_t\|^2_{\bar{V}_{t-1}^{-1}}\right)
\end{aligned}$$

Rearranging and plugging this into (23) gives:

$$\sum_{t=1}^{T} \|X_t\|^2_{\bar{V}_t^{-1}} \leq 2\log\left(\prod_{t=1}^{T}\left(1 + \|X_t\|^2_{\bar{V}_t^{-1}}\right)\right) = 2\log\left(\frac{\det\left(V_T\right)}{\det\left(V_0\right)}\right)$$

proving the first inequality. Lemma 15 proves the second inequality, completing the proof. □

Lemma 15. *Let $\{X_t\}_{t=1}^{\infty}$ be an arbitrary sequence of d-dimensional vectors such that $\|X_t\|_2^2 \leq 1$. Define $V_0 = \lambda I$, $V_t = \sum_{s=1}^{t} X_s X_s^T$ and $\bar{V}_t = V_0 + V_t$. Then,*

$$2\log\left(\frac{\det\left(\bar{V}_T\right)}{\det\left(\bar{V}_0\right)}\right) \leq 2d\log\left(\frac{d\lambda + T}{d\lambda}\right)$$

for $\lambda \geq 1/2$.

Proof. Let $\bar{V}_T$ have eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Then,

$$\begin{aligned}
2\log\left(\frac{\det\left(V_T\right)}{\det\left(V_0\right)}\right) &= 2\log\left(\frac{\prod_{i=1}^{d}\lambda_i}{\lambda^d}\right) \\
&\leq 2\log\left(\frac{\left(\frac{1}{d}\sum_{i=1}^{d}\lambda_i\right)^d}{\lambda^d}\right) = 2\log\left(\left(\frac{\sum_{i=1}^{d}\lambda_i}{d\lambda}\right)^d\right) && \text{(AM-GM Inequality)} \\
&= 2\log\left(\left(\frac{\mathrm{Tr}\left(\bar{V}_T\right)}{d\lambda}\right)^d\right) = 2\log\left(\left(\frac{\mathrm{Tr}\left(V_0 + \sum_{t=1}^{T} X_t X_t^T\right)}{d\lambda}\right)^d\right) \\
&= 2\log\left(\left(\frac{\mathrm{Tr}\left(V_0\right) + \sum_{t=1}^{T}\mathrm{Tr}\left(X_t X_t^T\right)}{d\lambda}\right)^d\right) = 2\log\left(\left(\frac{d\lambda + \sum_{t=1}^{T}\mathrm{Tr}\left(X_t^T X_t\right)}{d\lambda}\right)^d\right)
\end{aligned}$$

$$
\begin{aligned}
&= 2\log\left(\left(\frac{d\lambda + \sum_{t=1}^{T}\|X_t\|_2^2}{d\lambda}\right)^d\right) \\
&\leq 2\log\left(\left(\frac{d\lambda + T}{d\lambda}\right)^d\right) \\
&= 2d\log\left(\frac{d\lambda + T}{d\lambda}\right)
\end{aligned}
$$

which completes the proof. □