Supplementary Material of the Paper: Learning in Distributed Contextual Linear Bandits Without Sharing the Context

A Proofs and Remarks for Section 3: Contextual Linear Bandits with Known Context Distribution

Remark. We note that, to reduce the multi-context problem to a single context problem in the case of known context distribution, the straightforward approach that replaces the actual context realizations $\{X_{t,a}\}_{a\in\mathcal{A}}$ with the fixed set $\{\mathbb{E}_{\mathcal{P}_a}[X_{t,a}]\}_{a\in\mathcal{A}}$, and uses this latter set as $\mathcal{X}$ in Λ, does not work and can lead to linear regret in some cases. For instance consider the case where $d=1, \mathcal{A}=\{1,2\}, X_{t,a}\in\{-1,1\}\forall a\in\mathcal{A}, \theta_\star=1$ and $X_{t,1}$ takes the value 1 with probability $3/4$ and -1 otherwise, while $X_{t,2}$ takes the values $1,-1$ with probability $1/2$. Then clearly $\langle\mathbb{E}_{\mathcal{P}_1}[X_{t,1}],\theta_\star\rangle > \langle\mathbb{E}_{\mathcal{P}_2}[X_{t,2}],\theta_\star\rangle$, however, choosing $a_t=1\forall t\in[T]$ leads to $\mathbb{E}[R_T]\geq T/8$ since it holds that $X_{t,1}=-1, X_{t,2}=1$ with probability $1/8$.

Downlink Communication. Note that in our setup we assume that the central learner does not have any communication constraints when communicating with the distributed agents. Yet our algorithm makes frugal use of this ability: the central learner only sends the updated parameter vector $\hat{\theta}_t$. We can quantize $\hat{\theta}_t$ without performance loss if the downlink were also communication constrained using $\approx 5d$ bits and an approach similar to the one in Algorithm 2 - yet we do not expand on this in this paper, as our focus is in minimizing uplink communication costs.

A.1 Proof of Theorem 1

Theorem 1. *Algorithm 1 uses 1 bit per reward and 0 bits per context. Under Assumption 1, it achieves a regret $R_T = \bar{R}_T(\Lambda) + O(\sqrt{T\log T})$ with probability at least $1-\frac{1}{T}$.*

Proof. It is obvious that the agent only sends 1 bit to the central learner to represent r_t using SQ_1, hence, the algorithm uses 0 bits per context and 1 bit per reward. We next bound the regret of our algorithm as following. The regret can be expressed as

$$\begin{aligned}
R_T &= \sum_{t=1}^{T} \max_{a\in\mathcal{A}} \langle X_{t,a},\theta_\star\rangle - \langle X_{t,a_t},\theta_\star\rangle \\
&= \sum_{t=1}^{T} \langle \arg\max_{X_{t,a}} \langle X_{t,a},\theta_\star\rangle,\theta_\star\rangle - \langle \arg\max_{X_{t,a}} \langle X_{t,a},\hat{\theta}_t\rangle,\theta_\star\rangle \\
&= \sum_{t=1}^{T} \left(\langle \arg\max_{X_{t,a}} \langle X_{t,a},\theta_\star\rangle,\theta_\star\rangle - \langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a},\theta_\star\rangle],\theta_\star\rangle \right) \\
&\quad - \left(\langle \arg\max_{X_{t,a}} \langle X_{t,a},\hat{\theta}_t\rangle,\theta_\star\rangle - \langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a},\hat{\theta}_t\rangle|\hat{\theta}_t],\theta_\star\rangle \right) \\
&\quad + \left(\langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a},\theta_\star\rangle],\theta_\star\rangle - \langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a},\hat{\theta}_t\rangle|\hat{\theta}_t],\theta_\star\rangle \right). \qquad (15)
\end{aligned}$$

To bound R_T, we bound each of the three lines in the last expression. For the second term denoted by $\Sigma_t = \sum_{i=1}^{t} \left(\langle \arg\max_{X_{i,a}} \langle X_{i,a},\hat{\theta}_i\rangle,\theta_\star\rangle - \langle \mathbb{E}[\arg\max_{X_{i,a}} \langle X_{i,a},\hat{\theta}_i\rangle|\hat{\theta}_i],\theta_\star\rangle \right)$, we have that

$$\begin{aligned}
\mathbb{E}[\Sigma_{t+1}|\Sigma_t] &= \Sigma_t + \mathbb{E}\left[\langle \arg\max_{X_{t,a}} \langle X_{t,a},\hat{\theta}_t\rangle,\theta_\star\rangle - \langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a},\hat{\theta}_t\rangle|\hat{\theta}_t],\theta_\star\rangle|\Sigma_t \right] \\
&= \Sigma_t + \mathbb{E}\left[\mathbb{E}\left[\langle \arg\max_{X_{t,a}} \langle X_{t,a},\hat{\theta}_t\rangle,\theta_\star\rangle - \langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a},\hat{\theta}_t\rangle|\hat{\theta}_t],\theta_\star\rangle|\Sigma_t,\hat{\theta}_t \right] |\Sigma_t \right] \\
&= \Sigma_t + \mathbb{E}\left[\mathbb{E}\left[\langle \arg\max_{X_{t,a}} \langle X_{t,a},\hat{\theta}_t\rangle,\theta_\star\rangle - \langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a},\hat{\theta}_t\rangle|\hat{\theta}_t],\theta_\star\rangle|\hat{\theta}_t \right] |\Sigma_t \right] \\
&= \Sigma_t + \mathbb{E}\left[\langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a},\hat{\theta}_t\rangle|\hat{\theta}_t],\theta_\star\rangle - \langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a},\hat{\theta}_t\rangle|\hat{\theta}_t],\theta_\star\rangle|\Sigma_t \right]
\end{aligned}$$

$$= \Sigma_t. \tag{16}$$

We also have that

$$\begin{aligned}
|\Sigma_t - \Sigma_{t-1}| &\leq \|\langle \arg\max_{X_{t,a}} \langle X_{t,a}, \hat{\theta}_t\rangle, \theta_\star\rangle - \langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a}, \hat{\theta}_t\rangle], \theta_\star\rangle\| \|\theta_\star\| \\
&\leq \|\langle \arg\max_{X_{t,a}} \langle X_{t,a}, \hat{\theta}_t\rangle, \theta_\star\rangle\| + \|\langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a}, \hat{\theta}_t\rangle], \theta_\star\rangle\| \\
&\leq \|\arg\max_{X_{t,a}} \langle X_{t,a}, \hat{\theta}_t\rangle\| \|\theta_\star\| + \|\mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a}, \hat{\theta}_t\rangle]\| \|\theta_\star\| \leq 2.
\end{aligned} \tag{17}$$

Hence, Σ_t is a martingale with bounded difference. By Azuma–Hoeffding inequality [45], we have that $\|\Sigma_T\| \leq C\sqrt{T \log T}$ with probability at least $1 - \frac{1}{2T}$. Similarly, the first line in (15) is a martingale with bounded difference, hence, the following holds with probability at least $1 - \frac{1}{2T}$

$$\left| \sum_{t=1}^{T} \langle \arg\max_{X_{t,a}} \langle X_{t,a}, \theta_\star\rangle, \theta_\star\rangle - \langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a}, \theta_\star\rangle], \theta_\star\rangle \right| \leq C\sqrt{T \log T}. \tag{18}$$

By substituting in (15) and using the union bound we get that the following holds with probability at least $1 - \frac{1}{T}$

$$\begin{aligned}
R_T &\leq C\sqrt{T \log T} + \sum_{t=1}^{T} \left(\langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a}, \theta_\star\rangle], \theta_\star\rangle - \langle \mathbb{E}[\arg\max_{X_{t,a}} \langle X_{t,a}, \hat{\theta}_t\rangle | \hat{\theta}_t], \theta_\star\rangle \right) \\
&= C\sqrt{T \log T} + \sum_{t=1}^{T} \langle X^*(\theta_\star), \theta_\star\rangle - \langle X^*(\hat{\theta}_t), \theta_\star\rangle \\
&= C\sqrt{T \log T} + \sum_{t=1}^{T} \langle X^*(\theta_\star), \theta_\star\rangle - \langle X_t, \theta_\star\rangle.
\end{aligned} \tag{19}$$

We also have by definition of $X^*(\theta_\star)$ that for any given $\theta, \theta_\star$

$$\begin{aligned}
\langle X^*(\theta_\star), \theta_\star\rangle &= \mathbb{E}[\max_{X_{t,a}} \langle X_{t,a}, \theta_\star\rangle] \\
&\geq \mathbb{E}[\langle \arg\max_{X_{t,a}} \langle X_{t,a}, \theta\rangle, \theta_\star\rangle] = \langle X^*(\theta), \theta_\star\rangle.
\end{aligned} \tag{20}$$

Hence, we have that $\max_{X \in \mathcal{X}} \langle X, \theta_\star\rangle = \langle X^*(\theta_\star), \theta_\star\rangle$. By substituting in (19), we get that

$$R_T \leq C\sqrt{T \log T} + \sum_{t=1}^{T} \max_{X \in \mathcal{X}} \langle X, \theta_\star\rangle - \langle X_t, \theta_\star\rangle = C\sqrt{T \log T} + R_T(\Lambda), \tag{21}$$

where $R_T(\Lambda)$ is the regret of the subroutine Λ. □

A.2 Proof of Corollary 1

Corollary 1. Suppose we are given $\tilde{X}^\star$ that satisfies (12). Then, there exists an algorithm Λ for which Algorithm 1 achieves $R_T = \tilde{O}(d\sqrt{T} + \epsilon T\sqrt{d})$ with probability at least $1 - \frac{1}{T}$.

Proof. Λ assumes that the reward r_t is generated according to $\langle \tilde{X}^\star(\hat{\theta}_t), \theta_\star\rangle + \eta_t$, while it is actually generated according to

$$r_t = \langle X^\star(\hat{\theta}_t), \theta_\star\rangle + \eta_t = \langle \tilde{X}^\star(\hat{\theta}_t), \theta_\star\rangle + \eta_t + f(\hat{\theta}_t), \tag{22}$$

where $f(\hat{\theta}_t) = \langle X^\star(\hat{\theta}_t) - \tilde{X}^\star(\hat{\theta}_t), \theta_\star\rangle$. We have that

$$|f(\hat{\theta}_t)| \leq \|X^\star(\hat{\theta}_t) - \tilde{X}^\star(\hat{\theta}_t)\| \|\theta_\star\| \leq \epsilon. \tag{23}$$

Hence, the rewards follow a misspecified linear bandit model [24]. It was shown in [24] that for the single context case, there is an algorithm Λ that achieves $R_T(\Lambda) = \tilde{O}(d\sqrt{T} + \epsilon T)$ with probability at least $1 - \frac{1}{T}$. The corollary follows from Theorem 1 by noting that $R_T(\Lambda)$ is defined based on the true $X^\star$ as in (5). □

B Proofs of Section 4: Contextual Linear Bandits with Unknown Context Distribution

B.1 Proof of Theorem 2

Theorem 2. *Algorithm 2 satisfies that for all t: $X_t \in \mathcal{Q}$; and $B_t \leq 1 + \log_2(2d+1) + 5.03d$ bits. Under assumptions 1, 2, it achieves a regret $R_T = O(d\sqrt{T \log T})$ with probability at least $1 - \frac{1}{T}$.*

Proof. We start by proving some properties about the quantized values $\hat{X}_t, \hat{r}_t, \hat{X}_t^2$. We first note that be definition of SQ, we have that

$$m|(\hat{X}_t - X_{t,a_t})_j| \leq 1. \tag{24}$$

Hence,

$$\begin{aligned}|(\hat{X}_t^2 - X_{t,a_t}^2)_j| &= |(\hat{X}_t^2 - (\hat{X}_t + X_{t,a_t} - \hat{X}_t)^2)_j| = |(2(X_{t,a_t} - \hat{X}_t)\hat{X}_t + (X_{t,a_t} - \hat{X}_t)^2)_i| \\ &\leq 2|(X_{t,a_t} - \hat{X}_t)_i||(\hat{X}_t)_i| + |((X_{t,a_t} - \hat{X}_t)^2)_i| \leq \frac{2}{m} + \frac{1}{m^2} \leq \frac{3}{m}. \end{aligned} \tag{25}$$

We also have that

$$\begin{aligned}\mathbb{E}[\hat{X}_t^{(D)}|X_{t,a_t}^2] &= \mathbb{E}[\hat{X}_t^2 + e_t^2|X_{t,a_t}^2] = \mathbb{E}[\mathbb{E}[\hat{X}_t^2 + e_t^2|X_{t,a_t}^2, \hat{X}_t]|X_{t,a_t}^2] \\ &= \mathbb{E}[\hat{X}_t^2 + \mathbb{E}[e_t^2|X_{t,a_t}^2 - \hat{X}_t^2]|X_{t,a_t}^2] = \mathbb{E}[\hat{X}_t^2 + X_{t,a_t}^2 - \hat{X}_t^2|X_{t,a_t}^2] = X_{t,a_t}^2. \end{aligned} \tag{26}$$

In summary, from this and the definition of SQ, we get that

$$\begin{aligned}&m|(\hat{X}_t - X_{t,a_t})_j| \leq 1, \mathbb{E}[\hat{X}_t|X_{t,a_t}] = X_{t,a_t} \\ &m|\hat{X}_t^{(D)} - X_{t,a_t}^2| \leq 3, \mathbb{E}[\hat{X}_t^{(D)}|X_{t,a_t}^2] = X_{t,a_t}^2 \\ &|\hat{r}_t - r_t| \leq 1, \mathbb{E}[\hat{r}_t|r_t] = r_t \end{aligned} \tag{27}$$

We next show that $X_t \in \text{dom}(h)$. By definition of SQ, we have that $X_t \in \mathbb{N}^d$. We also have that

$$\begin{aligned}\|X_t\|_1 &= \sum_{i=1}^d m|(X_{t,a_t})_i| \leq 1 + \lfloor\sqrt{d}|(X_{t,a_t})_i|\rfloor \leq d + \sum_{i=1}^d \lfloor\sqrt{d}|(X_{t,a})_i|\rfloor^2 \\ &\leq d + d\|X_{t,a_t}\|^2 \leq 2d. \end{aligned} \tag{28}$$

Therefore, we have that $X_t \in \mathcal{Q} = \text{dom}(h)$.

We next show the upper bound on the number of bits B_t. We have that $\hat{r}_t$ uses 1 bit, the sign vector s_t uses d bits, e_t^2 uses d bits and X_t uses $\log(|\mathcal{Q}|)$ bits. We bound $|\mathcal{Q}|$ as follows. The number of non-negative solutions for the equation $\|a\|_1 = x$ for $a \in \mathbb{N}^d, x \in \mathbb{N}$ is $\binom{d+x-1}{x} \leq \binom{d+x}{x} = \binom{d+x}{d}$, hence,

$$|\mathcal{Q}| \leq (2d+1)\binom{3d}{d} \leq (2d+1)(e\frac{3d}{d})^d. \tag{29}$$

Hence, we have that

$$B_t \leq 1 + \log(2d+1) + (2 + \log(3e))d. \tag{30}$$

We next show the regret bound. We start by bounding the regret in iteration t by the distance between $\theta_\star, \hat{\theta}_{t-1}$. From step 7 of Algorithm 2, we have that $\langle X_{t,a_t}, \hat{\theta}_{t-1}\rangle \geq \langle X_{t,a}, \hat{\theta}_{t-1}\rangle \forall a \in \mathcal{A}$, hence, we have that

$$\begin{aligned}\max_{a\in\mathcal{A}}\langle X_{t,a}, \theta_\star\rangle - \langle X_{t,a_t}, \theta_\star\rangle &\leq \max_{a\in\mathcal{A}}\langle X_{t,a} - X_{t,a_t}, \theta_\star\rangle - \max_{a\in\mathcal{A}}\langle X_{t,a} - X_{t,a_t}, \hat{\theta}_{t-1}\rangle \\ &\leq \max_{a\in\mathcal{A}}\|X_{t,a} - X_{t,a_t}\|\|\theta_\star - \hat{\theta}_{t-1}\| \leq 2\|\theta_\star - \hat{\theta}_{t-1}\|. \end{aligned} \tag{31}$$

We next bound the distance $\|\theta_\star - \hat{\theta}_{t-1}\|$. Let us denote $e_t = \hat{X}_t - X_{t,a_t}, \hat{\eta}_t = \eta_t + (\hat{r}_t - r_t), E_t = X_{t,a_t}X_{t,a_t}^T - (V_t - V_{t-1})$. We have that

$$\|\theta_\star - \hat{\theta}_t\| = \|\theta_\star - V_t^{-1}\sum_{i=1}^t \hat{r}_i\hat{X}_i\| = \|\theta_\star - V_t^{-1}\sum_{i=1}^t (X_{i,a_i}X_{i,a_i}^T\theta_\star + r_ie_i + \hat{\eta}_iX_{i,a_i} + \hat{\eta}_ie_i)\|$$

$$
\begin{aligned}
&= \|V_t^{-1} \sum_{i=1}^{t} (E_i\theta_\star + r_i e_i + \hat{\eta}_i X_{i,a_i} + \hat{\eta}_i e_i)\| \\
&\leq \|V_t^{-1}\|(\|\sum_{i=1}^{t} E_i\| + (|r_i| + |\eta_i|)\|\sum_{i=1}^{t} e_i\| + \|\sum_{i=1}^{t} \hat{\eta}_i X_{i,a_i}\| \\
&\leq \|V_t^{-1}\|(\|\sum_{i=1}^{t} E_i\| + (1 + |\eta_i|)\|\sum_{i=1}^{t} e_i\| + \|\sum_{i=1}^{t} \hat{\eta}_i X_{i,a_i}\|. \qquad (32)
\end{aligned}
$$

We next bound each of the values in the last expression. As η_i is subgaussian we have that with probability at least $1 - \frac{1}{5T^2}$, we have that $|\eta_i| \leq C \log T \forall i \in [T]$. We also have that, using (27), $S_t^e = \sum_{i=1}^{t} e_i$ is a martingale with bounded difference, hence, by Azuma–Hoeffding inequality, we get that with probability at least $1 - \frac{1}{5dT^2}$ we have that $|(S_t^e)_j| \leq \frac{C}{\sqrt{d}}\sqrt{t\log(dT)}$; note that $|(e_t)_i| \leq \frac{1}{\sqrt{d}}$. Hence, by the union bound we get that with probability at least $1 - \frac{1}{5T^2}$ we have that $\|\sum_{i=1}^{t} e_i\| \leq C\sqrt{t\log(dT)}$. Similarly, conditioned on $X_{1,a_1}, ..., X_{t,a_t}$, $\sum_{i=1}^{t} \hat{\eta}_i X_{i,a_i}$ is a martingale with bounded difference, hence, with probability at least $1 - \frac{1}{5dT^2}$ we have that $|(\sum_{i=1}^{t} \hat{\eta}_i X_{i,a_i})_j| \leq C\sqrt{\sum_{i=1}^{t}(X_{i,a_i})_j^2 \log(dT)}$. Hence, with probability at least $1 - \frac{1}{5T^2}$ we have that $\|\sum_{i=1}^{t} \hat{\eta}_i X_{i,a_i}\| \leq C\sqrt{\sum_{i=1}^{t}\|X_{i,a_i}\|^2 \log(dT)} \leq C\sqrt{t\log(dT)}$. Summing up, we get that with probability at least $1 - \frac{3}{5T^2}$

$$
\|\theta_\star - \hat{\theta}_t\| \leq \|V_t^{-1}\|(\|\sum_{i=1}^{t} E_i\| + C\sqrt{t\log(dT)}). \qquad (33)
$$

It remains to bound $\|V_t^{-1}\|, \|\sum_{i=1}^{t} E_i\|$ which we do in the following by starting with $\|\sum_{i=1}^{t} E_i\|$. We have that

$$
\begin{aligned}
E_i &= X_{i,a_i}X_{i,a_i}^T - \hat{X}_i\hat{X}_i^T + \text{diag}(\hat{X}_i\hat{X}_i^T) - \text{diag}(\hat{X}_t^{(D)}) \\
&= \text{diag}(\hat{X}_i\hat{X}_i^T) - 2X_{i,a_i}e_i^T - e_ie_i^T - \text{diag}(\hat{X}_t^{(D)}) \\
&= 2\text{diag}(X_{i,a_i}e_i^T) - 2X_{i,a_i}e_i^T - (e_ie_i^T - \text{diag}(e_ie_i^T)) - \text{diag}(\hat{X}_t^{(D)} - X_{i,a_i}^2). \qquad (34)
\end{aligned}
$$

Hence, we have that

$$
\begin{aligned}
\|\sum_{i=1}^{t} E_i\| \leq& 2\|\sum_{i=1}^{t} \text{diag}(X_{i,a_i}e_i^T)\| + 2\|\sum_{i=1}^{t} X_{i,a_i}e_i^T\| \\
&+ \|\sum_{i=1}^{t} e_ie_i^T - \text{diag}(e_ie_i^T)\| + \|\sum_{i=1}^{t} \text{diag}(\hat{X}_t^{(D)} - X_{i,a_i}^2)\|. \qquad (35)
\end{aligned}
$$

We have that, using (27), conditioned on $X_{1,a_1}, ..., X_{t,a_t}$, $\sum_{i=1}^{t} \text{diag}(X_{i,a_i}e_i^T)$ is a martingale with bounded difference, hence, similar to what we did before using Azuma–Hoeffding inequality and the union bound we get that with probability at least $1 - \frac{1}{20T^2}$, we have that $\|\sum_{i=1}^{t} \text{diag}(X_{i,a_i}e_i^T)\| \leq \frac{C}{\sqrt{d}}\sqrt{t\log(dT)}$. Similarly, with probability at least $1 - \frac{1}{20T^2}$, we have that$\|\sum_{i=1}^{t} \text{diag}(\hat{X}_t^{(D)} - X_{i,a_i}^2)\| \leq \frac{C}{\sqrt{d}}\sqrt{t\log(dT)}$. We next turn to bounding $\|\sum_{i=1}^{t} X_{i,a_i}e_i^T\|$. Conditioned on $X_{1,a_1}, ..., X_{t,a_t}$, we have that by Azuma–Hoeffding, with probability at least $1 - \frac{1}{d^2T^2}$, we have

$$
|(\sum_{i=1}^{t} X_{i,a_i}e_i^T)_{jk}| \leq \frac{C}{\sqrt{d}}\sqrt{\sum_{i=1}^{t}(X_{i,a_i})_j^2 \log(dT)}. \qquad (36)
$$

We notice that taking the absolute value of all elements of a matrix does not decrease its maximum eigenvalue, hence, by the union bound we have that with probability at least $1 - \frac{1}{20T^2}$ we have that

$$
\|\sum_{i=1}^{t} X_{i,a_i}e_i^T\| \leq \frac{C\sqrt{\log(dT)}}{\sqrt{d}}\|\mathbf{1}[\sqrt{\sum_{i=1}^{t}(X_{i,a_i})_1^2}, ..., \sqrt{\sum_{i=1}^{t}(X_{i,a_i})_d^2}]\|
$$

$$\leq \frac{C\sqrt{\log(dT)}}{\sqrt{d}}\sqrt{d\sum_{i=1}^{t}\|X_{i,a_i}\|^2} \leq C\sqrt{t\log(dT)}. \tag{37}$$

To bound $\|\sum_{i=1}^{t} e_i e_i^T - \text{diag}(e_i e_i^T)\|$, we notice that for all elements except the diagonal we have that $\mathbb{E}[(e_i)_j(e_i)_k] = \mathbb{E}[(e_i)_j(e_i)_k|X_{t,a_t}] = \mathbb{E}[(e_i)_k|X_{t,a_t}]\mathbb{E}[(e_i)_j|X_{t,a_t}] = 0, j \neq k$. Hence, it can be shown that $\sum_{i=1}^{t}(e_i)_j(e_i)_k$ is a martingale with bounded difference for $j \neq k$, hence, with probability at least $1 - \frac{1}{20d^2T^2}$, we have that $|\sum_{i=1}^{t}(e_i)_j(e_i)_k| \leq \frac{C}{d}\sqrt{t\log(dT)}$. Hence, by the union bound we get that with probability at least $1 - \frac{1}{20T^2}$

$$\|\sum_{i=1}^{t} e_i e_i^T - \text{diag}(e_i e_i^T)\| \leq \frac{C\sqrt{t\log(dT)}}{d}\|\mathbf{1}\mathbf{1}^T\| \leq C\log(dT). \tag{38}$$

Hence, from (35) and the union bound we have that with probability at least $1 - \frac{1}{5T^2}$

$$\|\sum_{i=1}^{t} E_i\| \leq C\sqrt{t\log(dT)}. \tag{39}$$

We next turn to bounding $\|V_t^{-1}\|$. We have from (39), and Assumption 2, and the union bound, the following holds with probability at least $1 - \frac{2}{5T^2}$

$$\|V_t\| = \|\sum_{i=1}^{t} X_{i,a_i}X_{i,a_i}^T - E_i\| \geq \|\sum_{i=1}^{t} X_{i,a_i}X_{i,a_i}^T\| - \|E_i\| \geq C(\frac{t}{d} - \sqrt{t\log(dT)}). \tag{40}$$

Hence for $t \geq 4\log(dT)$, we have that with probability at least $1 - \frac{2}{5T^2}$, it holds that $\|V_t\| \geq C\frac{t}{2d}$, and hence,

$$\|V_t^{-1}\| \leq C\frac{d}{t}. \tag{41}$$

Hence, from (32) and the union bound, the following holds with probability at least $1 - \frac{1}{T^2}$

$$\|\theta_\star - \hat{\theta}_t\| \leq Cd\frac{\sqrt{\log(dT)}}{\sqrt{t}}. \tag{42}$$

Therefore, from (31) and the union bound again we have that the following holds with probability at least $1 - \frac{1}{T}$

$$\begin{aligned} R_T &\leq \sum_{t=1}^{T} Cd\frac{\sqrt{\log(dT)}}{\sqrt{t}} \leq Cd\sqrt{\log(dT)}(1 + \int_{t=1}^{T}\frac{1}{\sqrt{t}}dt) \\ &\leq 2Cd\sqrt{T\log(dT)}. \end{aligned} \tag{43}$$

□

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[2] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[3] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30, 2017.

[4] J. Alman and V. V. Williams. A refined laser method and faster matrix multiplication. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 522–539. SIAM, 2021.

[5] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745, 2011.

[6] V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, 1987.

[7] M. H. Anisi, G. Abdul-Salaam, and A. H. Abdullah. A survey of wireless sensor network approaches and their energy consumption for monitoring farm fields in precision agriculture. *Precision Agriculture*, 16(2):216–238, 2015.

[8] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[9] B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53, 2004.

[10] D. Bouneffouf, I. Rish, and G. A. Cecchi. Bandit models of human behavior: Reward processing in mental disorders. In *International Conference on Artificial General Intelligence*, pages 237–248. Springer, 2017.

[11] Q. Ding, C.-J. Hsieh, and J. Sharpnack. An efficient algorithm for generalized linear bandit: Online stochastic gradient descent and thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 1585–1593. PMLR, 2021.

[12] A. Durand, C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis, and J. Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pages 67–82. PMLR, 2018.

[13] R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[14] P. Elias. Universal codeword sets and representations of the integers. *IEEE transactions on information theory*, 21(2):194–203, 1975.

[15] A. Gersho and R. M. Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

[16] R. M. Gray and T. G. Stockham. Dithered quantizers. *IEEE Transactions on Information Theory*, 39(3):805–812, 1993.

[17] Y. Han, Z. Zhou, Z. Zhou, J. Blanchet, P. W. Glynn, and Y. Ye. Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*, 2020.

[18] O. A. Hanna, Y. H. Ezzeldin, C. Fragouli, and S. Diggavi. Quantization of distributed data for learning. *IEEE Journal on Selected Areas in Information Theory*, 2(3):987–1001, 2021.

[19] O. A. Hanna, Y. H. Ezzeldin, T. Sadjadpour, C. Fragouli, and S. Diggavi. On distributed quantization for classification. *IEEE Journal on Selected Areas in Information Theory*, 1(1):237–249, 2020.

[20] O. A. Hanna, A. M. Girgis, C. Fragouli, and S. Diggavi. Differentially private stochastic linear bandits:(almost) for free. *arXiv preprint arXiv:2207.03445*, 2022.

[21] O. A. Hanna, L. Yang, and C. Fragouli. Solving multi-arm bandit using a few bits of communication. In *International Conference on Artificial Intelligence and Statistics*, pages 11215–11236. PMLR, 2022.

[22] T. J. Jech. *The axiom of choice*. Courier Corporation, 2008.

[23] P. C. Landgren. *Distributed multi-agent multi-armed bandits*. PhD thesis, Princeton University, 2019.

[24] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[25] T. Le, C. Szepesvari, and R. Zheng. Sequential learning for multi-channel wireless network monitoring with channel switching costs. *IEEE Transactions on Signal Processing*, 62(22):5919–5929, 2014.

[26] F. Li, D. Yu, H. Yang, J. Yu, H. Karl, and X. Cheng. Multi-armed-bandit-based spectrum scheduling algorithms in wireless networks: A survey. *IEEE Wireless Communications*, 27(1):24–30, 2020.

[27] L. Li, Y. Lu, and D. Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

[28] Y. Li, Y. Wang, X. Chen, and Y. Zhou. Tight regret bounds for infinite-armed linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 370–378. PMLR, 2021.

[29] Y. Li, Y. Wang, and Y. Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, pages 2173–2174. PMLR, 2019.

[30] J. Mary, R. Gaudel, and P. Preux. Bandits and recommender systems. In *International Workshop on Machine Learning, Optimization and Big Data*, pages 325–336. Springer, 2015.

[31] P. Matikainen, P. M. Furlong, R. Sukthankar, and M. Hebert. Multi-armed recommendation bandits for selecting state machine policies for robotic systems. In *2013 IEEE International Conference on Robotics and Automation*, pages 4545–4551. IEEE, 2013.

[32] P. Mayekar and H. Tyagi. Ratq: A universal fixed-length quantizer for stochastic optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1399–1409. PMLR, 2020.

[33] T. D. Novlan, H. S. Dhillon, and J. G. Andrews. Analytical modeling of uplink cellular networks. *IEEE Transactions on Wireless Communications*, 12(6):2669–2679, 2013.

[34] A. N. Rafferty, H. Ying, and J. J. Williams. Bandit assignment for educational experiments: Benefits to students versus statistical power. In *International Conference on Artificial Intelligence in Education*, pages 286–290. Springer, 2018.

[35] W. Ren, X. Zhou, J. Liu, and N. B. Shroff. Multi-armed bandits with local differential privacy. *arXiv preprint arXiv:2007.03121*, 2020.

[36] Z. Ren and Z. Zhou. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *arXiv preprint arXiv:2008.11918*, 2020.

[37] P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[38] T. Sajed and O. Sheffet. An optimal private stochastic-mab algorithm based on optimal private stopping rule. In *International Conference on Machine Learning*, pages 5579–5588. PMLR, 2019.

[39] S. Sajeev, J. Huang, N. Karampatziakis, M. Hall, S. Kochman, and W. Chen. Contextual bandit applications in a customer support bot. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3522–3530, 2021.

[40] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*. Citeseer, 2014.

[41] S. Shahrampour, A. Rakhlin, and A. Jadbabaie. Multi-armed bandits in multi-agent networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2786–2790. IEEE, 2017.

[42] R. Shariff and O. Sheffet. Differentially private contextual linear bandits. *Advances in Neural Information Processing Systems*, 31, 2018.

[43] C. Shi and C. Shen. Federated multi-armed bandits. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[44] J. Tenenbaum, H. Kaplan, Y. Mansour, and U. Stemmer. Differentially private multi-armed bandits in the shuffle model. *Advances in Neural Information Processing Systems*, 34:24956–24967, 2021.

[45] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[46] Y. Wang, J. Hu, X. Chen, and L. Wang. Distributed bandit learning: Near-optimal regret with efficient communication. In *International Conference on Learning Representations*, 2019.

[47] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. *Advances in Neural Information Processing Systems*, 26, 2013.

[48] K. Zheng, T. Cai, W. Huang, Z. Li, and L. Wang. Locally differentially private (contextual) bandits learning. *Advances in Neural Information Processing Systems*, 33:12300–12310, 2020.