A. Theorem 1: Single parameter setting

A.1. Posterior distribution computation

$$
\begin{aligned}
& \Pr(\tilde{\mu}|r_i(t)) \\
\propto\ & \Pr(r_i(t)|\tilde{\mu})\Pr(\tilde{\mu}) \\
\propto\ & \exp\{-\frac{1}{2v^2}((r_i(t)-\tilde{\mu}^T b_i(t))^2 \\
& +(\tilde{\mu}-\hat{\mu}(t))^T B(t)(\tilde{\mu}-\hat{\mu}(t))\} \\
\propto\ & \exp\{-\frac{1}{2v^2}(r_i(t)^2+\tilde{\mu}^T b_i(t)b_i(t)^T\tilde{\mu} \\
& +\tilde{\mu}^T B(t)\tilde{\mu}-2\tilde{\mu}^T b_i(t)r_i(t)-2\tilde{\mu}^T B(t)\hat{\mu}(t))\} \\
\propto\ & \exp\{-\frac{1}{2v^2}(\tilde{\mu}^T B(t+1)\tilde{\mu}-2\tilde{\mu}^T B(t+1)\hat{\mu}(t+1))\} \\
\propto\ & \exp\{-\frac{1}{2v^2}(\tilde{\mu}-\hat{\mu}(t+1))^T B(t+1)(\tilde{\mu}-\hat{\mu}(t+1))\} \\
\propto\ & \mathcal{N}(\hat{\mu}(t+1), v^2 B(t+1)^{-1}).
\end{aligned}
$$

Therefore, the posterior distribution of μ at time $t+1$ is $\mathcal{N}(\hat{\mu}(t+1), v^2 B(t+1)^{-1})$.

A.2. Some concentration inequalities

Formula 7.1.13 from Abramowitz & Stegun (1964) can be used to derive the following concentration and anti-concentration inequalities for Gaussian distributed random variables.

Lemma 5. *(Abramowitz & Stegun, 1964) For a Gaussian distributed random variable Z with mean m and variance σ^2, for any $z \geq 1$,*

$$\frac{1}{2\sqrt{\pi}z}e^{-z^2/2} \leq \Pr(|Z-m| > z\sigma) \leq \frac{1}{\sqrt{\pi}z}e^{-z^2/2}.$$

Definition 9 (Super-martingale)**.** *A sequence of random variables $(Y_t; t \geq 0)$ is called a super-martingale corresponding to filtration $\mathcal{F}_t$, if for all t, Y_t is $\mathcal{F}_t$-measurable, and for $t \geq 1$,*

$$\mathbb{E}\left[Y_t - Y_{t-1} \mid \mathcal{F}_{t-1}\right] \leq 0.$$

Lemma 6 (Azuma-Hoeffding inequality)**.** *If a super-martingale $(Y_t; t \geq 0)$, corresponding to filtration $\mathcal{F}_t$, satisfies $|Y_t - Y_{t-1}| \leq c_t$ for some constant c_t, for all $t = 1, \ldots, T$, then for any $a \geq 0$,*

$$\Pr(Y_T - Y_0 \geq a) \leq e^{-\frac{a^2}{2\sum_{t=1}^T c_t^2}}.$$

The following lemma is implied by Theorem 1 in Abbasi-Yadkori et al. (2011):

Lemma 7. *(Abbasi-Yadkori et al., 2011) Let $(\mathcal{F}'_t; t \geq 0)$ be a filtration, $(m_t; t \geq 1)$ be an $\mathbb{R}^d$-valued stochastic process such that m_t is $(\mathcal{F}'_{t-1})$-measurable, $(\eta_t; t \geq 1)$ be a real-valued martingale difference process such that η_t is $(\mathcal{F}'_t)$-measurable. For $t \geq 0$, define $\xi_t = \sum_{\tau=1}^t m_\tau \eta_\tau$ and $M_t = I_d + \sum_{\tau=1}^t m_\tau m_\tau^T$, where I_d is the d-dimensional identity matrix. Assume η_t is conditionally R-sub-Gaussian.*

Then, for any $\delta' > 0$, $t \geq 0$, with probability at least $1-\delta'$,

$$||\xi_t||_{M_t^{-1}} \leq R\sqrt{d\ln\left(\frac{t+1}{\delta'}\right)},$$

where $||\xi_t||_{M_t^{-1}} = \sqrt{\xi_t^T M_t^{-1} \xi_t}$.

A.3. Proof of Lemma 1

Bounding the probability of event $E^\mu(t)$: We use Lemma 7 with $m_t = b_{a(t)}(t)$, $\eta_t = r_{a(t)}(t) - b_{a(t)}(t)^T\mu$, $\mathcal{F}'_t = (a(\tau+1), m_{\tau+1}, \eta_\tau : \tau \leq t)$. (Note that effectively, $\mathcal{F}'_t$ has all the information, including the arms played, until time $t+1$, except for the reward of the arm played at time $t+1$). By the definition of $\mathcal{F}'_t$, m_t is $\mathcal{F}'_{t-1}$-measurable, and η_t is $\mathcal{F}'_t$-measurable. Also, η_t is conditionally R-sub-Gaussian due to the assumption mentioned in the problem settings (refer to Section 2.1), and is a martingale difference process:

$$\mathbb{E}\left[\eta_t|\mathcal{F}'_{t-1}\right] = \mathbb{E}[r_{a(t)}(t)|b_{a(t)}(t), a(t)] - b_{a(t)}(t)^T\mu = 0.$$

Also, this makes

$$M_t = I_d + \sum_{\tau=1}^t m_\tau m_\tau^T = I_d + \sum_{\tau=1}^t b_{a(\tau)}(\tau) b_{a(\tau)}(\tau)^T,$$

$$\xi_t = \sum_{\tau=1}^t m_\tau \eta_\tau = \sum_{\tau=1}^t b_{a(\tau)}(\tau)(r_{a(\tau)} - b_{a(\tau)}(\tau)^T\mu).$$

Note that $B(t) = M_{t-1}$, and $\hat{\mu}(t) - \mu = M_{t-1}^{-1}(\xi_{t-1} - \mu)$. Let for any vector $y \in \mathbb{R}$ and matrix $A \in \mathbb{R}^{d\times d}$, $||y||_A$ denote $\sqrt{y^T A y}$. Then, for all i,

$$
\begin{aligned}
|b_i(t)^T\hat{\mu}(t) - b_i(t)^T\mu| = |b_i(t)^T M_{t-1}^{-1}(\xi_{t-1}-\mu)| \leq \\
||b_i(t)||_{M_{t-1}^{-1}} ||\xi_{t-1}-\mu||_{M_{t-1}^{-1}} = \\
||b_i(t)||_{B(t)^{-1}} ||\xi_{t-1}-\mu||_{M_{t-1}^{-1}}.
\end{aligned}
$$

The inequality holds because M_{t-1}^{-1} is a positive definite matrix. Using Lemma 7, for any $\delta' > 0$, $t \geq 1$, with probability at least $1-\delta'$,

$$||\xi_{t-1}||_{M_{t-1}^{-1}} \leq R\sqrt{d\ln\left(\frac{t}{\delta'}\right)}.$$

Therefore, $||\xi_{t-1}-\mu||_{M_{t-1}^{-1}} \leq R\sqrt{d\ln\left(\frac{t}{\delta'}\right)} + ||\mu||_{M_{t-1}^{-1}} \leq R\sqrt{d\ln\left(\frac{T}{\delta'}\right)} + 1$. Substituting $\delta' = \frac{\delta}{T^2}$, we get that

with probability $1-\frac{\delta}{T^2}$, for all i,

$$\begin{aligned} & |b_i(t)^T\hat{\mu}(t) - b_i(t)^T\mu| \\ \leq\ & ||b_i(t)||_{B(t)^{-1}} \cdot \left(R\sqrt{d\ln\left(\frac{T}{\delta'}\right)}+1\right) \\ \leq\ & ||b_i(t)||_{B(t)^{-1}} \cdot \left(R\sqrt{d\ln(T^3)\ln(\frac{1}{\delta})}+1\right) \\ =\ & \ell(T)\ s_{t,i}. \end{aligned}$$

This proves the bound on the probability of $E^{\mu}(t)$.

Bounding the probability of event $E^{\theta}(t)$: For all i,

$$\begin{aligned} & |\theta_i(t) - b_i(t)^T\hat{\mu}(t)| \\ =\ & |b_i(t)^T\tilde{\mu}(t) - b_i(t)^T\hat{\mu}(t)| \\ =\ & |b_i(t)^T B(t)^{-1/2}B(t)^{1/2}(\tilde{\mu}(t)-\hat{\mu}(t))| \\ \leq\ & v\sqrt{b_i(t)^T B(t)^{-1}b_i(t)} \cdot ||\frac{1}{v}B(t)^{1/2}(\tilde{\mu}(t)-\hat{\mu}(t))||_2 \\ =\ & vs_{t,i} \cdot ||\frac{1}{v}B(t)^{1/2}(\tilde{\mu}(t)-\hat{\mu}(t))||_2. \end{aligned}$$

Therefore, we can prove the statement of the lemma by proving that $||\frac{1}{v}B(t)^{1/2}(\tilde{\mu}(t)-\hat{\mu}(t))||_2 \leq \sqrt{4d\ln(Td)}$ with probability at least $1-\frac{1}{T^2}$, given any filtration $\mathcal{F}_{t-1}$. Now, given any filtration, by definition $\frac{1}{v}\sqrt{B(t)}(\tilde{\mu}(t)-\hat{\mu}(t))$ is the d-dimensional standard normal variable, therefore using concentration of Gaussian random variables (Lemma 5),

$$\begin{aligned} & \Pr\left(||\frac{1}{v}B(t)^{1/2}(\tilde{\mu}(t)-\hat{\mu}(t))||_2 > \sqrt{4d\ln(Td)}\ \right) \\ \leq\ & d\frac{1}{\sqrt{\pi}\sqrt{4\ln(Td)}}e^{-(2\ln Td)} \\ \leq\ & \frac{1}{T^2}. \end{aligned}$$

A.4. Proof of Lemma 2

Given event $E^{\mu}(t)$, $|b_{a^*(t)}(t)^T\hat{\mu}(t) - b_{a^*(t)}(t)^T\mu| \leq \ell(T)s_{t,a^*(t)}$. And, since Gaussian random variable $\theta_{a^*(t)}(t)$ has mean $b_{a^*(t)}(t)^T\hat{\mu}(t)$ and standard deviation $vs_{t,a^*(t)}$, using anti-concentration inequality in Lemma 5,

$$\begin{aligned} & \Pr\left(\theta_{a^*(t)}(t) \geq b_{a^*(t)}(t)^T\mu + \ell(T)s_{t,a^*(t)} \,\middle|\, \mathcal{F}_{t-1}\right) \\ =\ & \Pr\left(\frac{\theta_{a^*(t)}(t)-b_{a^*(t)}(t)^T\hat{\mu}(t)}{vs_{t,a^*(t)}} \right. \\ & \qquad \left. \geq \frac{b_{a^*(t)}(t)^T\mu+\ell(T)s_{t,a^*(t)}-b_{a^*(t)}(t)^T\hat{\mu}(t)}{vs_{t,a^*(t)}} \,\middle|\, \mathcal{F}_{t-1}\right) \\ \geq\ & \frac{1}{4\sqrt{\pi}}e^{-Z_t^2}. \end{aligned}$$

Where

$$\begin{aligned} |Z_t| &= \left|\frac{b_{a^*(t)}(t)^T\mu - b_{a^*(t)}(t)^T\hat{\mu}(t) - \ell(T)s_{t,a^*(t)}}{vs_{t,a^*(t)}}\right| \\ &\leq \frac{2\left(\ell(T)\right)s_{t,a^*(t)}}{vs_{t,a^*(t)}} \\ &= \frac{2\left(R\sqrt{d\ln(T^3)\ln(\frac{1}{\delta})}+1\right)}{R\sqrt{\frac{24}{\epsilon}d\ln(\frac{1}{\delta})}} \\ &\leq \sqrt{\frac{\epsilon}{2}}(\ln T+1). \end{aligned}$$

So

$$\Pr\left(\theta_{a^*(t)}(t) \geq b_{a^*(t)}(t)^T\mu \,\middle|\, \mathcal{F}_{t-1}\right) \geq \frac{1}{4\sqrt{\pi}}e^{-\frac{\epsilon}{2}(\ln T+1)} = \frac{1}{4e\sqrt{\pi}T^{\frac{\epsilon}{2}}}.$$

A.5. Missing details from Section 3.2

To derive the inequality $\sum_{t=1}^{T} s_{t,a(t)} \leq 5\sqrt{dT\ln T}$, we use the following result, implied by the referred lemma in Auer (2002).

Lemma 8. *(Auer, 2002, Lemma 11). Let $A' = A + xx^T$, where $x \in \mathbb{R}^d, A, A' \in \mathbb{R}^{d\times d}$, and all the eigenvalues $\lambda_j, j = 1,\ldots,d$ of A are greater than or equal to 1. Then, the eigenvalues $\lambda'_j, j = 1,\ldots,d$ of A' can be arranged so that $\lambda_j \leq \lambda'_j$ for all j, and*

$$x^T A^{-1}x \leq 10\sum_{j=1}^{d}\frac{\lambda'_j - \lambda_j}{\lambda_j}.$$

Let $\lambda_{j,t}$ denote the eigenvalues of $B(t)$. Note that $B(t+1) = B(t) + b_{a(t)}(t)b_{a(t)}(t)^T$, and $\lambda_{j,t} \geq 1, \forall j$. Therefore, above implies

$$s_{t,a(t)}^2 \leq 10\sum_{j=1}^{d}\frac{\lambda_{j,t+1}-\lambda_{j,t}}{\lambda_{j,t}}.$$

This allows us to derive the given inequality after some algebraic computations following along the lines of Lemma 3 of Chu et al. (2011).

To obtain bounds for the other definition of regret in Remark 1, we observe that because $\mathbb{E}[r_i(t)|\mathcal{F}_{t-1}] = b_i(t)^T\mu$ for all i, the expected value of regret$'(t)$ given $\mathcal{F}_{t-1}$ for this definition of regret(t) is same as before. More precisely, for $\mathcal{F}_{t-1}$ such that $E^{\mu}(t)$ holds,

$$\begin{aligned} & \mathbb{E}\left[\text{regret}'(t) \,|\, \mathcal{F}_{t-1}\right] \\ =\ & \mathbb{E}\left[\text{regret}(t) \,|\, \mathcal{F}_{t-1}\right] \\ =\ & \mathbb{E}\left[r_{a^*(t)}(t) - r_{a(t)}(t) \,\middle|\, \mathcal{F}_{t-1}\right] \\ =\ & \mathbb{E}\left[b_{a^*(t)}(t)^T\mu - b_{a(t)}(t)^T\mu \,\middle|\, \mathcal{F}_{t-1}\right]. \end{aligned}$$

And, $\mathbb{E}\left[\text{regret}'(t) \mid \mathcal{F}_{t-1}\right] = 0$ for other $\mathcal{F}_{t-1}$. Therefore, Lemma 4 holds as it is, and Y_t defined in Definition 8 is a super-martingale with respect to this new definition of regret(t) as well. Now, if $|r_i(t) - b_i(t)^T\mu| \leq R$, for all i, then $|\text{regret}'(t)| \leq 2R$ and $|Y_t - Y_{t-1}| \leq \frac{6}{p}\frac{g(T)^2}{\ell(T)} + 2R$, and we can apply Azuma-Hoeffding inequality exactly as in the proof of Theorem 1 to obtain regret bounds of the same order as Theorem 1 for the new definition. The results extend to the more general R-sub-Gaussian condition on $r_i(t)$, using a simple extension of Azuma-Hoeffding inequality; we omit the proof of that extension.

B. Theorem 3: Modified algorithm

Below is a description of the algorithm for single parameter setting which, instead of generating a single $\tilde{\mu}(t)$ and setting $\theta_i(t)$ as $b_i(t)^T\tilde{\mu}(t)$, as was done in Algorithm 1, generates $\theta_i(t), i = 1, \ldots, N$ as N independent samples with the same marginal distributions as $b_i(t)^T\tilde{\mu}(t)$.

Algorithm 2 Modified Thompson Sampling

Set $B = I_d, \hat{\mu} = 0_d$, $f = 0_d$.
for all $t = 1, 2, \ldots,$ **do**
  For each arm $i = 1, \ldots, N$, sample $\theta_i(t)$ independently from distribution $\mathcal{N}(b_i(t)^T\hat{\mu}, v^2 b_i(t)^T B^{-1} b_i(t))$.
  Play arm $a(t) := \arg\max_i \theta_i(t)$ and observe reward r_t.
  Update $B = B + b_{a(t)}(t)b_{a(t)}(t)^T$, $f = f + b_{a(t)}(t)r_t$, $\hat{\mu} = B^{-1}f$.
end for

In the regret analysis of this algorithm, we will be able to utilize the independence of the $\theta_i(t)$'s to bound the probability of playing saturated arms in terms of the probability of playing *optimal arm* (see Lemma 11). In comparison, in the proof of Theorem 1, we bounded the probability of playing saturated arms in terms of the probability of playing *unsaturated arms* which includes the optimal arm. This difference in the analysis is the key to our improved regret bound for this algorithm.

In the proof below, except when explicitly redefined, notations are as before (refer to notations table in Appendix A).

Definition 10. *Define $\ell(T)$ and v as before, but redfine $g(T) = \sqrt{4\ln(NT)}\, v + \ell(T)$.*

Definition 11. *An arm i is called* saturated *at time t if $\Delta_i(t) > g(T)s_{t,i}$, and* unsaturated *otherwise. Observe that by definition, $a^*(t)$ is an unsaturated arm at time t.*

Definition 12. *Define event $E^\mu(t)$ as before, but redefine $E^\theta(t)$ as the event that*

$$\forall i, |\theta_i(t) - b_i(t)^T\hat{\mu}(t)| \leq \sqrt{4\ln(NT)}\, v s_{t,i}.$$

Lemma 9. *For all t, $0 < \delta < 1$, $\Pr(E^\mu(t)) \geq 1 - \frac{\delta}{T^2}$. And, for all possible filtrations $\mathcal{F}_{t-1}$, $\Pr(E^\theta(t)|\mathcal{F}_{t-1}) \geq 1 - \frac{1}{T^2}$.*

Proof. The probability bound for $E^\mu(t)$ can be proven using a concentration inequality given by Abbasi-Yadkori et al. (2011), as before in the proof of Theorem

1. The probability bound for $E^\theta(t)$ can be proven using the concentration inequality for Gaussian random variables from Abramowitz & Stegun (1964) stated as Lemma 5 in Appendix A.2 . □

The next lemma lower bounds the probability that the sample $\theta_{a^*(t)}(t)$ for the optimal arm at time t will exceed $b_{a^*(t)}(t)^T\mu$.

Lemma 10. *For any filtration $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ is true,*

$$\Pr\left(\theta_{a^*(t)}(t) > b_{a^*(t)}(t)^T\mu \,\middle|\, \mathcal{F}_{t-1}\right) \geq \frac{1}{4e\sqrt{\pi}T^\epsilon}.$$

Proof. Given event $E^\mu(t)$, $|b_{a^*(t)}(t)^T\hat{\mu}(t) - b_{a^*(t)}(t)^T\mu| \leq (\ell(T))\, s_{t,a^*(t)}$. And, since Gaussian random variable $\theta_{a^*(t)}(t)$ has mean $b_{a^*(t)}(t)^T\hat{\mu}(t)$ and standard deviation $vs_{t,a^*(t)}$, using anti-concentration inequality in Lemma 5,

$$\begin{aligned}
& \Pr\left(\theta_{a^*(t)}(t) \geq b_{a^*(t)}(t)^T\mu \,\middle|\, \mathcal{F}_{t-1}\right) \\
= & \Pr\left(\frac{\theta_{a^*(t)}(t) - b_{a^*(t)}(t)^T\hat{\mu}(t)}{vs_{t,a^*(t)}} \right. \\
& \qquad \left. \geq \frac{b_{a^*(t)}(t)^T\mu - b_{a^*(t)}(t)^T\hat{\mu}(t)}{vs_{t,a^*(t)}} \,\middle|\, \mathcal{F}_{t-1}\right) \\
\geq & \frac{1}{4\sqrt{\pi}} e^{-Z_t^2}.
\end{aligned}$$

Where

$$\begin{aligned}
|Z_t| &= \left|\frac{b_{a^*(t)}(t)^T\mu - b_{a^*(t)}(t)^T\hat{\mu}(t)}{vs_{t,a^*(t)}}\right| \\
&\leq \frac{\ell(T)s_{t,a^*(t)}}{vs_{t,a^*(t)}} \\
&= \frac{\left(R\sqrt{d\ln(T^3)\ln(\frac{1}{\delta})}+1\right)}{R\sqrt{\frac{24}{\epsilon}d\ln(\frac{1}{\delta})}} \\
&\leq \sqrt{\frac{\epsilon}{2}}(\ln T + 1).
\end{aligned}$$

So

$$\Pr\left(\theta_{a^*(t)}(t) \geq b_{a^*(t)}(t)^T\mu \,\middle|\, \mathcal{F}_{t-1}\right) \geq \frac{1}{4\sqrt{\pi}} e^{-\frac{\epsilon}{2}(\ln T+1)} = \frac{1}{4e\sqrt{\pi}T^{\frac{\epsilon}{2}}}.$$

□

The following lemma bounds the probability of playing a saturated arm in terms of the probability of playing the optimal arm.

Lemma 11. *Given any filtration $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ is true,*

$$\Pr\left(a(t) \in C(t) \mid \mathcal{F}_{t-1}\right) \leq \frac{1}{p}\Pr\left(a(t) = a^*(t) \mid \mathcal{F}_{t-1}\right) + \frac{1}{pT^2},$$

where $p = \frac{1}{4e\sqrt{\pi}T^\epsilon}$.

Proof.

$$\begin{aligned}
& \Pr\left(a(t) = a^*(t) \mid \mathcal{F}_{t-1}\right) \\
= & \Pr\left(\theta_{a^*(t)}(t) \geq \theta_j(t), \forall j \neq a^*(t) \,\middle|\, \mathcal{F}_{t-1}\right) \\
\geq & \Pr\left(\exists i \in C(t), \theta_{a^*(t)}(t) \geq \theta_i(t), \right. \\
& \qquad \left. \theta_i(t) \geq \theta_j(t), \forall j \neq a^*(t) \mid \mathcal{F}_{t-1}\right) \\
\geq & \Pr\left(\theta_{a^*(t)}(t) \geq b_{a^*(t)}(t)^T\mu, \right. \\
& \qquad \left. \exists i \in C(t), \theta_i(t) \geq \theta_j(t), \forall j \neq a^*(t) \mid \mathcal{F}_{t-1}\right) \\
& - \Pr\left(\theta_i(t) > b_{a^*(t)}(t)^T\mu, \exists i \in C(t) \,\middle|\, \mathcal{F}_{t-1}\right) \\
= & \Pr\left(\theta_{a^*(t)}(t) \geq b_{a^*(t)}(t)^T\mu\right) \\
& \cdot \Pr\left(\exists i \in C(t), \theta_i(t) \geq \theta_j(t), \forall j \neq a^*(t) \mid \mathcal{F}_{t-1}\right) \\
& - \Pr\left(\theta_i(t) > b_{a^*(t)}(t)^T\mu, \exists i \in C(t) \,\middle|\, \mathcal{F}_{t-1}\right) \\
\geq & p \cdot \Pr\left(a(t) \in C(t) \mid \mathcal{F}_{t-1}\right) - \Pr\left(\overline{E^\theta(t)} \,\middle|\, \mathcal{F}_{t-1}\right) \\
\geq & p \cdot \Pr\left(a(t) \in C(t) \mid \mathcal{F}_{t-1}\right) - \frac{1}{T^2}.
\end{aligned}$$

The second equality follows from the independence of $\theta_1(t), \ldots, \theta_N(t)$ given $\mathcal{F}_{t-1}$. This independence holds because given the current distributions, which are fixed on fixing the filtration $\mathcal{F}_{t-1}$, the algorithms samples $\theta_1(t), \ldots, \theta_N(t)$ are independently from their respective distributions. In the second last inequality, for the first term we used the lower bound provided by Lemma 10. For the second term, we used the observation that if $E^\theta(t)$ and $E^\mu(t)$ are true, then by the definition of these events and the definition of saturated arms, it holds that

$$\forall i \in C(t), \theta_i(t) \leq b_i(t)^T\mu + g(T)s_{t,i} \leq b_i(t)^T\mu + \Delta_i(t) = b_{a^*(t)}(t)^T\mu.$$

Therefore, given an $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ is true, $\theta_i(t)$ for some saturated arm i can be larger than $b_{a^*(t)}(t)^T\mu$ only if $E^\theta(t)$ is false. □

Next, we establish a super-martingale process that will form the basis of our proof of the high-probability regret bound.

Definition 13. *Let*

$$X_t := \frac{regret'(t)}{g(T)} - \frac{1}{p}I(a(t) = a^*(t))s_{t,a^*(t)} - s_{t,a(t)} - \frac{2}{pT^2},$$

$$Y_t := \sum_{w=1}^{t} X_w,$$

where $p = \frac{1}{4e\sqrt{\pi}T^\epsilon}$.

Lemma 12. *$(Y_t; t = 0, \ldots, T)$ is a super-martingale process with respect to filtration $\mathcal{F}_t$.*

Proof. We need to prove that for all $t \in [1, T]$, and any $\mathcal{F}_{t-1}$, $\mathbb{E}[Y_t - Y_{t-1}|\mathcal{F}_{t-1}] \leq 0$, i.e.

$$\frac{1}{g(T)}\mathbb{E}[\text{regret}'(t)|\mathcal{F}_{t-1}] \leq \frac{\Pr(a(t)=a^*(t) \mid \mathcal{F}_{t-1})}{p} s_{t,a^*(t)} + \mathbb{E}\left[s_{t,a(t)} \,\middle|\, \mathcal{F}_{t-1}\right] + \frac{2}{pT^2}.$$

If $\mathcal{F}_{t-1}$ is such that $E^\mu(t)$ is not true, then $\text{regret}'(t) = \text{regret}(t) \cdot I(E^\mu(t)) = 0$, and the above inequality holds trivially. So, we consider $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ holds.

We observe that if the events $E^\mu(t), E^\theta(t)$ are true, then $\Delta_{a(t)}(t) \leq g(T)(s_{t,a(t)} + s_{t,a^*(t)})$. This is because if an arm i is played at time t, then it must be true that $\theta_i(t) \geq \theta_{a^*(t)}(t)$. And, if $E^\theta(t)$ and $E^\mu(t)$ are true, then,

$$\begin{aligned} b_i(t)^T\mu &\geq \theta_i(t) - g(T)s_{t,i} \\ &\geq \theta_{a^*(t)}(t) - g(T)s_{t,i} \\ &\geq b_{a^*(t)}(t)^T\mu - g(T)s_{t,a^*(t)} - g(T)s_{t,i}. \end{aligned}$$

Therefore, given a filtration $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ is true, either $\Delta_{a(t)}(t) \leq g(T)(s_{t,a(t)} + s_{t,a^*(t)})$ or $E^\theta(t)$ is false. Also, by the definition of unsaturated arms, for every unsaturated arm i, $\Delta_i(t) \leq g(T)s_{t,i}$. Using these observations,

$$\begin{aligned} &\mathbb{E}\left[\text{regret}'(t) \mid \mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[\Delta_{a(t)}(t) I(a(t) \in C(t)) \,\middle|\, \mathcal{F}_{t-1}\right] \\ &\quad + \mathbb{E}\left[\Delta_{a(t)}(t) I(a(t) \notin C(t)) \,\middle|\, \mathcal{F}_{t-1}\right] \\ &\leq g(T)\mathbb{E}\left[(s_{t,a^*(t)} + s_{t,a(t)}) I(a(t) \in C(t)) \,\middle|\, \mathcal{F}_{t-1}\right] \\ &\quad + \Pr\left(\overline{E^\theta(t)} \,\middle|\, \mathcal{F}_{t-1}\right) \\ &\quad + g(T)\mathbb{E}\left[s_{t,a(t)} I(a(t) \notin C(t)) \,\middle|\, \mathcal{F}_{t-1}\right] \\ &\leq g(T)s_{t,a^*(t)} \Pr\left(a(t) \in C(t) \mid \mathcal{F}_{t-1}\right) \\ &\quad + \tfrac{1}{T^2} + g(T)\mathbb{E}\left[s_{t,a(t)} \,\middle|\, \mathcal{F}_{t-1}\right] \\ &\leq g(T)s_{t,a^*(t)} \cdot \tfrac{1}{p} \Pr\left(a(t) = a^*(t) \mid \mathcal{F}_{t-1}\right) + g(T)\tfrac{2}{pT^2} \\ &\quad + g(T)\mathbb{E}\left[s_{t,a(t)} \,\middle|\, \mathcal{F}_{t-1}\right]. \end{aligned}$$

The last inequality uses Lemma 11. In the first and the last inequality, we also used that for all i, $\Delta_i(t) \leq 1$, and $0 \leq s_{t,a^*(t)} \leq ||b_{a^*(t)}(t)|| \leq 1$. □

Now, we are ready to prove Theorem 3.

B.1. Proof of Theorem 3

We observe that the absolute value of the first three terms in the definition of X_t bounded by $1/p$, and the last term is bounded by $2/p$, therefore the super-martingale Y_t has bounded difference $|Y_t - Y_{t-1}| \leq \frac{5}{p}$, for all $t \geq 1$, and thus apply Azuma-Hoeffding inequality, to obtain that with probability $1 - \frac{\delta}{2}$,

$$\begin{aligned} &\textstyle\sum_{t=1}^T \frac{1}{g(T)}\text{regret}'(t) \\ &\leq \textstyle\sum_{t=1}^T \left(\frac{1}{p} I(a(t) = a^*(t)) s_{t,a^*(t)} + s_{t,a(t)} + \frac{2}{pT^2}\right) \\ &\quad + \tfrac{5}{p}\sqrt{2T\ln(2/\delta)} \\ &= \tfrac{1}{p}\textstyle\sum_{t=1}^T s_{t,a(t)} + \sum_{t=1}^T s_{t,a(t)} + \frac{2}{pT} + \frac{5}{p}\sqrt{2T\ln(\frac{2}{\delta})} \\ &= O(\sqrt{T^{1+\epsilon} d \ln T} + \sqrt{T^{1+\epsilon}\ln\tfrac{1}{\delta}}). \end{aligned}$$

Here, we used that $\sum_{t=1}^T s_{t,a(t)} \leq 5\sqrt{dT\ln T}$, which can be derived along the lines of Lemma 3 of Chu et al. (2011) using Lemma 11 of Auer (2002). Also, because $E^\mu(t)$ holds for all t with probability at least $1 - \frac{\delta}{2}$ (Lemma 9), $\text{regret}'(t) = \text{regret}(t)$ for all t with probability at least $1 - \frac{\delta}{2}$. Hence, with probability $1 - \delta$,

$$\mathcal{R}(T) = \textstyle\sum_{t=1}^T \text{regret}(t) = \sum_{t=1}^T \text{regret}'(t) = \left(d\sqrt{\frac{T^{1+\epsilon}\ln(N)}{\epsilon}}\ln(T)\ln(\tfrac{1}{\delta}).\right)$$

C. Theorem 2: N different parameters

Theorem 2 considers the setting where each arm i is associated with a parameter $\mu_i \in \mathbb{R}^d$, where possibly $\mu_i \neq \mu_{i'}$ for two different arms i and i'. In this case, Thompson Sampling would maintain a separate estimate of mean $\hat{\mu}_i(t)$, and $B_i(t)$ for each arm i which would be updated only at the time instances when i is played. We appropriately modify some of the previous definitions:

$$B_i(t) = I_d + \sum_{\tau=1:a(\tau)=i}^{t-1} b_i(\tau) b_i(\tau)^T,$$

$$\hat{\mu}_i(t) = B_i(t)^{-1} \left(\sum_{\tau=1:a(\tau)=i}^{t-1} b_i(\tau) r_i(\tau) \right),$$

$$s_{t,i} = \sqrt{b_i(t)^T B_i(t)^{-1} b_i(t)}.$$

The posterior distribution for each arm i at time t would be $\mathcal{N}(b_i(t)^T \hat{\mu}_i(t), v^2\ b_i(t)^T B_i(t)^{-1} b_i(t))$. And, the TS algorithm is now stated as follows.

Algorithm 3 Thompson Sampling for Contextual bandits with N parameters

Set $B_i = I_d, \hat{\mu}_i = 0_d$, $i = 1, \ldots, N$, $f_i = 0_d$.
for $t = 1, 2, \ldots,$ **do**
  For each arm $i = 1, \ldots, N$, sample $\theta_i(t)$ independently from distribution $\mathcal{N}(b_i(t)^T \hat{\mu}_i, v^2\ b_i(t)^T B_i^{-1} b_i(t))$.
  Play arm $a(t) := \arg\max_i \theta_i(t)$ and observe reward r_t.
  Update $B_{a(t)} = B_{a(t)} + b_{a(t)}(t) b_{a(t)}(t)^T$, $f_{a(t)} = f_{a(t)} + b_{a(t)}(t) r_t$, $\hat{\mu}_{a(t)} = B_{a(t)}^{-1} f_{a(t)}$.
end for

The optimal arm $a^*(t)$ is now the arm that maximizes $b_i(t)^T \mu_i$, and the regret at time t is defined as

$$\text{regret}(t) = b_{a^*(t)}(t)^T \mu_{a^*(t)} - b_{a(t)}(t)^T \mu_{a(t)}.$$

The regret analysis closely follows the proof of Theorem 3, described in the previous section. Below, we describe only the changes required.

The events $E^\mu(t)$ will now be defined with respect to concentration of all $\hat{\mu}_i(t)$ around their respective means. That is,

$$E^\mu(t) : \forall i, b_i(t)^T \hat{\mu}_i(t) \in [b_i(t)^T \mu_i - (\ell(T))\, s_{t,i},\quad b_i(t)^T \mu_i + (\ell(T))\, s_{t,i}].$$

Similarly, $E^\theta(t)$ will be the event that

$$\forall i, \theta_i(t) \in [b_i(t)^T \hat{\mu}_i(t) - \sqrt{4 \ln(NT)}\ vs_{t,i},\quad b_i(t)^T \hat{\mu}_i(t) + \sqrt{4 \ln(NT)}\ vs_{t,i}].$$

It is easy to observe that the statements of Lemma 9 and the super-martingale property established by Lemma 12 will hold as it is for these new definitions. The only difference will appear in the bound for $\sum_t s_{t,a(t)}$ used in the proof of Theorem 3. For the case of N different parameters, we will get a bound of $O(\sqrt{NTd \ln T})$ on this quantity.

Let $n_i(T)$ be the number of times arm i is played by time T. Then using Lemma 8, for any two consecutive time steps t, t' at which arm i is played,

$$s_{t,a(t)}^2 \leq 10 \sum_{j=1}^{d} \frac{\lambda_{j,t'} - \lambda_{j,t}}{\lambda_{j,t}}.$$

This allows us to derive the following lemma along the lines of Lemma 3 of Chu et al. (2011).

Lemma 13. *(Chu et al., 2011, Lemma 3) For $i = 1, \ldots, N$,*

$$\sum_{t=1:a(t)=i}^{T} s_{t,a(t)} \leq 5 \sqrt{d n_i(T) \ln(n_i(T))}.$$

Using the above lemma,

$$\begin{aligned} \sum_{t=1}^{T} s_{t,a(t)} &= \sum_{i=1}^{N} \sum_{t=1:a(t)=i}^{T} s_{t,a(t)} \\ &\leq \sum_{i=1}^{N} 5 \sqrt{n_i(T) d \ln T} \\ &\leq 5 \sqrt{N} \sqrt{\sum_i n_i(T)} \sqrt{d \ln T} \\ &= 5 \sqrt{NTd \ln T}. \end{aligned}$$

Therefore, following the same lines as proof of Theorem 3, we will get a regret bound of $\tilde{O}(d \sqrt{\frac{NT^{1+\epsilon}}{\epsilon}})$.

D. Conclusions

We provided a theoretical analysis of Thompson Sampling for the stochastic contextual bandits problem with linear payoffs. Our results resolve some open questions regarding the theoretical guarantees for Thompson Sampling, and establish that even for the contextual version of the stochastic MAB problem, TS achieves regret bounds close to the state-of-the-art methods. We used a novel martingale-based analysis technique which is arguably simpler than the techniques in the past work on TS (Agrawal & Goyal, 2012; Kaufmann et al., 2012), and is amenable to extensions.

In the algorithm in this paper, Gaussian priors were used, so that $\tilde{\mu}(t)$ was generated from a Gaussian distribution. However, the analysis techniques in this paper are extendable to an algorithm that uses a prior distribution other than the Gaussian distribution. The only distribution specific properties we have used in the analysis are the concentration and anti-concentration inequalities for Gaussian distributed random variables (Lemma 5), which were used to prove Lemma 1 and Lemma 2 respectively. If any other distribution provides similar tail inequalities, to allow us proving these lemmas, these can be used as a black box in the analysis, and the regret bounds can be reproduced for that distribution.

Several questions remain open. A tighter analysis that can remove the dependence on ϵ is desirable. We believe that our techniques would adapt to provide such bounds for the *expected regret*. Other avenues to explore are contextual bandits with *generalized* linear models considered in Filippi et al. (2010), the setting with delayed and batched feedback, and the *agnostic* case of contextual bandits with linear payoffs. The agnostic case refers to the setting which does not make the realizability assumption that there exists a vector μ_i for each i for which $\mathbb{E}[r_i(t)|b_i(t)] = b_i(t)^T\mu_i$. To our knowledge, no existing algorithm has been shown to have non-trivial regret bounds for the agnostic case.