Supplementary Material: Simple Regret Minimization for Contextual Bandits

Aniket Anand Deshmukh [*1] **Srinagesh Sharma** [*2] **James W. Cutler** [3] **Mark Moldwin** [4] **Clayton Scott** [5]

1. Remarks

- In Section 4, we formalize the notion of History H_t in the main paper.
- Detailed experiments and results about different α are in Section 6.

2. Estimating Expected Rewards and Confidence Estimates

A key ingredient of our extension is an estimate of f_a, for each a, based on the current history. We use kernel methods to estimate f_a. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric positive definite kernel function on $\mathcal{X}$, $\mathcal{H}$ be the corresponding RKHS and $\phi(x) = k(\cdot, x)$ be the associated canonical feature map. Let $\phi(X_{a,t}) := [\phi(x_j)]_{j \in D_{a,t}}$. We define the kernel matrix associated with $X_{a,t}$ as $K_{a,t} := \phi(X_{a,t})^T \phi(X_{a,t}) \in \mathbb{R}^{N_{a,t} \times N_{a,t}}$ and the kernel vector of context x as $k_{a,t}(x) := \phi(X_{a,t})^T \phi(x)$. Let $I_{a,t}$ be the identity matrix of size $N_{a,t}$. We estimate f_a at time t, via kernel ridge regression, i.e.,

$$\hat{f}_{a,t}(x) = \arg\min_{f_a \in \mathcal{H}} \sum_{j \in D_{a,t}} (f_a(x_j) - r_j)^2 + \lambda \|f_a\|^2.$$

The solution to this optimization problem is $\hat{f}_{a,t}(x) = k_{a,t}(x)^T (K_{a,t} + \lambda I_{a,t})^{-1} Y_{a,t}$. Furthermore, Durand et al. (2018) establish a confidence interval for $f_a(x)$ in terms of $\hat{f}_{a,t}(x)$ and the "variance" $\hat{\sigma}^2_{a,t}(x) := k(x,x) - k_{a,t}(x)^T (K_{a,t} + \lambda I_{a,t})^{-1} k_{a,t}(x)$.

Theorem 2.1 (Restatement of Theorem 2.1 in (Durand et al., 2018))**.** *Consider the contextual bandit scenario described in the main paper. For any $\beta > 0$, with probability at least $1 - e^{-\beta^2}$, it holds simultaneously over all $x \in \mathcal{X}$ and all $t \leq T$,*

$$|f_a(x) - \hat{f}_{a,t}(x)| \leq (C_1 \beta + C_2) \frac{\hat{\sigma}_{a,t}(x)}{\sqrt{\lambda}}, \qquad (1)$$

where $C_1 = \rho\sqrt{2}$ and
$C_2 = \rho\sqrt{\sum_{\tau=2}^{T} \ln(1 + \frac{1}{\lambda}\hat{\sigma}_{a,\tau-1}(x_\tau))} + \sqrt{\lambda}\|f_a\|_{\mathcal{H}}$.

In the contextual bandit setting in Durand et al. (2018), for any $\delta \in (0, 1]$, Theorem 2.1 in Durand et al. (2018) establishes that with probability at least $1 - \delta$, it holds simultaneously over all $x \in \mathcal{X}$ and $t \geq 0$,

$$|f_a(x) - \hat{f}_{a,t}(x)| \leq \frac{\hat{\sigma}_{a,t}(x)}{\sqrt{\lambda}} \left[\sqrt{\lambda}\|f_a\|_{\mathcal{H}} + \rho\sqrt{2\ln(1/\delta) + 2\gamma_t(\lambda)} \right],$$

where $\gamma_t(\lambda) = \frac{1}{2}\sum_{\tau=1}^{t} \ln(1 + \frac{1}{\lambda}\hat{\sigma}_{a,\tau-1}(x_\tau))$

For $T \geq t$, one can replace t in the log terms with T. Then $\forall x, \forall t \geq 1$, we have

$$1 - \delta \leq \mathbb{P}\left(|f_a(x) - \hat{f}_{a,t}(x)| \leq \frac{\hat{\sigma}_{a,t}(x)}{\sqrt{\lambda}} \left[\sqrt{\lambda}\|f_a\|_{\mathcal{H}} + \rho\sqrt{2\ln(1/\delta) + 2\gamma_T(\lambda)} \right] \right).$$

Let $\delta = e^{-\beta^2}$. In that case,

$$1 - e^{-\beta^2} \leq \mathbb{P}\left(|f_a(x) - \hat{f}_{a,t}(x)| \leq \frac{\hat{\sigma}_{a,t}(x)}{\sqrt{\lambda}} \left[\sqrt{\lambda}\|f_a\|_{\mathcal{H}} + \rho\sqrt{2\beta^2 + 2\gamma_T(\lambda)} \right] \right).$$

Using triangle inequality $\sqrt{p+q} \leq \sqrt{p} + \sqrt{q}$ for any $p, q \geq 0$,

$$1 - e^{-\beta^2} \leq \mathbb{P}\left(|f_a(x) - \hat{f}_{a,t}(x)| \leq \frac{\hat{\sigma}_{a,t}(x)}{\sqrt{\lambda}} \left[\sqrt{\lambda}\|f_a\|_{\mathcal{H}} + \rho\sqrt{2\beta^2} + \rho\sqrt{2\gamma_T(\lambda)} \right] \right).$$

[*]Equal contribution [1]Bing Ads, Microsoft AI & Research, USA [2]Microsoft Search, Assistant & Intelligence, Microsoft, USA [3]Department of Aerospace Engineering, University of Michigan Ann Arbor, USA [4]Climate and Space Engineering, University of Michigan Ann Arbor, USA [5]Department of EECS, University of Michigan Ann Arbor, USA. Correspondence to: Aniket Anand Deshmukh <aniketde@umich.edu>, Srinagesh Sharma <srinag@umich.edu>.

Proceedings of the 2^{nd} Exploration in Reinforcement Learning Workshop at the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Let $C_1 = \rho\sqrt{2}$ and $C_2 = \sqrt{\lambda}\|f_a\|_{\mathcal{H}} + \rho\sqrt{2\gamma_T(\lambda)}$. Hence, we have

$$1 - e^{-\beta^2} \leq \mathbb{P}\left(|f_a(x) - \hat{f}_{a,t}(x)| \leq \frac{\hat{\sigma}_{a,t}(x)}{\sqrt{\lambda}}[C_1\beta + C_2] \right).$$

In the later Section 5.5 we show that $C_2 = O(\rho\sqrt{\ln T})$. For convenience, we denote the width of the confidence interval $s_{a,t}(x) := 2(C_1\beta + C_2)\frac{\hat{\sigma}_{a,t}(x)}{\sqrt{\lambda}}$. Thus, the upper and lower confidence bounds of $f_a(x)$ are $U_{a,t}(x) := \hat{f}_{a,t}(x) + \frac{s_{a,t}(x)}{2}$ and $L_{a,t}(x) := \hat{f}_{a,t}(x) - \frac{s_{a,t}(x)}{2}$. The upper confidence bound is the most optimistic estimate of the reward and the lower confidence bound is the most pessimistic estimate of the reward.

3. Comparison of Contextual-Gap and Kernel-UCB

In this section, we illustrate the difference between the policies of Kernel-UCB (which minimizes cumulative regret) and exploration phase of Contextual-Gap (which aims to minimize simple regret). At each time step, Contextual-Gap selects one of two arms: $J_t(x)$, the arm with highest pessimistic reward estimate, or $j_t(x)$, the arm excluding $J_t(x)$ with highest optimistic reward estimate. Kernel-UCB, in contrast, selects the arm with the highest optimistic reward estimate (i.e., with the maximum upper confidence bound).

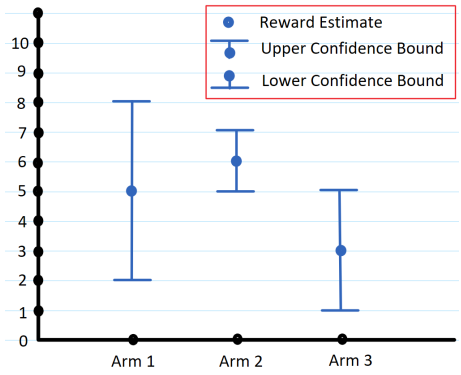


Figure 1: case 1

Consider a three arm scenario at some time τ with context x_τ. Suppose that the estimated rewards and confidence intervals are as in Figures 1 and 2, reflecting two different cases.

- Case 1 (Figure 1): In this case, Kernel-UCB would pick arm 1, because it has the maximum upper confidence bound. Kernel-UCB's policy is designed to be optimistic in the case of uncertainty. In the Contextual-Gap, we first calculate $J_\tau(x_\tau)$ which minimizes $B_{a,\tau}(x_\tau)$. Note that $B_{1,\tau}(x_\tau) = U_{2,\tau}(x_\tau) - L_{1,\tau}(x_\tau) = 7 - 2 = 5$, $B_{2,\tau}(x_\tau) = 3$ and $B_{3,\tau}(x_\tau) = 7$. In this case, $J_\tau(x_\tau) = 2$ and hence $j_\tau(x_\tau) = 1$. Finally, Contextual-Gap would choose among arm 1 and arm 2, and would finally choose arm 1 because it has the largest confidence interval. Hence, in case 1, Contextual-Gap chooses the same arm as that of Kernel-UCB.

- Case 2 (Figure 2): In this case, Kernel-UCB would pick arm 1. Note that $B_{1,\tau}(x_\tau) = U_{2,\tau}(x_\tau) - L_{1,\tau}(x_\tau) = 7 - 4 = 3$, $B_{2,\tau}(x_\tau) = 7$ and $B_{3,\tau}(x_\tau) = 4$. Then $J_\tau(x_\tau) = 1$ and hence $j_\tau(x_\tau) = 2$. Finally, Contextual-Gap chooses arm 2, because it has the widest confidence interval. Hence, in case 2, Contextual-Gap chooses a different arm compared to that of Kernel-UCB.

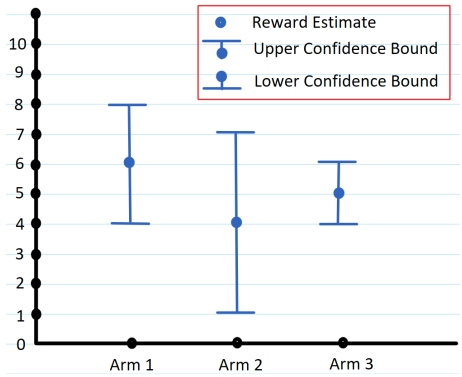


Figure 2: case 2

Clearly, the use of the lower confidence bound along with upper confidence bound allows Contextual-Gap to explore more than kernel-UCB. However, Contextual-Gap doesn't explore just any arm, but rather it explores only among arms with some likelihood of being optimal. The following section details high probability bounds on the simple regret of the Contextual-Gap algorithm.

4. Probabilistic Setting and Martingale Lemma

For the theoretical results, the following general probabilistic framework is adopted, following Abbasi-Yadkori et al. (2011) and Durand et al. (2018). We formalize the notion of history H_t defined in the Section 3 of the main paper using filtration. A filtration is a sequence of σ-algebras $\{\mathcal{F}_t\}_{t=1}^{\infty}$ such that $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \cdots$. Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtration such that x_t is $\mathcal{F}_{t-1}$ measurable,

and ζ_t is $\mathcal{F}_t$ measurable. For example, one may take $\mathcal{F}_t := \sigma(x_1, x_2, \cdots, x_{t+1}, \zeta_1, \zeta_2, \cdots, \zeta_t)$, i.e., $\mathcal{F}_t$ is the $\sigma-$algebra generated by $x_1, x_2, \cdots, x_{t+1}, \zeta_1, \zeta_2, \cdots, \zeta_t$.

We assume that ζ_t is a zero mean, ρ-conditionally sub-Gaussian random variable, i.e., ζ_t is such that for some $\rho > 0$ and $\forall \gamma \in \mathbb{R}$,

$$\mathbb{E}[e^{\gamma \zeta_t} | \mathcal{F}_{t-1}] \leq \exp\left(\frac{\gamma^2 \rho^2}{2}\right). \tag{2}$$

Definition 4.1 (Definition 4.11 in Motwani & Raghavan (1995))**.** *Let $(\Sigma, \mathcal{F}, Pr)$ be a probability space with filtration $\mathcal{F}_0, \mathcal{F}_1, \ldots$. Suppose that $Z_0, Z_1, \ldots$ are random variables such that for all $i > 0$, Z_i is $\mathcal{F}_i$ measurable. The sequence $Z_0, Z_1, \ldots$ is a martingale provided for all $i \geq 0$,*

$$\mathbb{E}[Z_{i+1} | \mathcal{F}_i] = Z_i.$$

Lemma 4.2 (Theorem 4.12 in Motwani & Raghavan (1995))**.** *Any subsequence of a martingale is also a martingale (relative to the corresponding subsequence of the underlying filter).*

The above Lemma is important because we construct confidence intervals for each arm separately. Note that we define a subset of time indices ($D_{a,t}$ of each arm a), when the arm a was selected. Based on these indices we can form subsequences of the main context $\{x_t\}_{t=1}^{\infty}$ and noise sequence $\{\zeta_t\}_{t=1}^{\infty}$ such that the assumptions on the main sequence hold for subsequences.

5. Learning Theoretic Analysis

We now analyze high probability simple regret bounds which depend on the gap quantity $\Delta_a(x) := |\max_{i \neq a} f_i(x) - f_a(x)|$. The bounds are presented in the non-i.i.d setting described in Section 3 of the paper. For the confidence interval to be useful, it needs to shrink to zero with high probability over the feature space as each arm is pulled more and more. This requires the smallest non-zero eigenvalue of the sample covariance matrix of the data for each arm to be lower bounded by a certain value. We make an assumption that allows for such a lower bound, and use it to prove that the confidence intervals shrink with high probability under certain assumptions. Finally, we bound the simple regret using the result of shrinking confidence interval, the gap quantity, and the special exploration strategy described in Algorithm in the main paper. We now make additional assumptions to the problem setting.

A I $\{\mathcal{X}_t\}_{t \geq 1} \subset \mathbb{R}^d$, is a random process on compact space endowed with a finite positive Borel measure.

A II Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is bounded by a constant L, the canonical feature map $\phi : \mathcal{X} \to \mathcal{H}$ of k is a continuous function, and $\mathcal{H}$ is separable.

We denote $\mathbb{E}_{t-1}[\cdot] := \mathbb{E}[\cdot | x_1, x_2, \ldots, x_{t-1}]$ and by $\lambda_r(A)$ the r^{th} largest eigenvalue of a compact self adjoint operator A. For a context x, the operator $\phi(x)\phi(x)^T : \mathcal{H} \to \mathcal{H}$ is a compact self-adjoint operator. Based on this notation, we make the following assumption:

A III There exists a subspace of dimension d^* with projection P, and a constant $\lambda_x > 0$, such that $\forall t$, $\lambda_r(P^T \mathbb{E}_{t-1}[\phi(x_t)\phi(x_t)^T]P) > \lambda_x$ for $r \leq d^*$ and $\lambda_r((I-P)^T \mathbb{E}_{t-1}[\phi(x_t)\phi(x_t)^T](I-P)) = 0, \forall r > d^*$.

Assumption **A III** facilitates the generalization of Bayes gap (Hoffman et al., 2014) to the kernel setting with non-i.i.d, time varying contexts. It allows us to lower bound, with high probability, the r^{th} eigenvalue of the cumulative second moment operator $S_t := \sum_{s=1}^{t} \phi(x_s)\phi(x_s)^T$ so that it is possible to learn the reward behavior in the low energy directions of the context at the same rate as the high energy ones with high probability.

5.1. Lower Bound on r^{th} Eigenvalue

We now provide a lower bound on the r^{th} eigenvalue of a compact self-adjoint operator. There are similar results in the setting where reward is a linear function of context, including Lemma 2 in Gentile et al. (2014) and Lemma 7 in Li & Zhang (2018) which provides lowest eigenvalue bounds with the assumption of linear reward and full rank covariance, and Theorem 2.2 in Tu & Recht (2017) which assumes more structure to the contexts generated. There are results similar to Lemma 2 in Gentile et al. (2014), Gentile et al. (2017) and Korda et al. (2016). We extend these results to the setting of a compact self-adjoint operator scenario with data occupying a finite dimensional subspace.

Let $W_t := \sum_{s=1}^{t} \mathbb{E}_{s-1}[(\phi(x_s)\phi(x_s)^T)^2] - (\mathbb{E}_{s-1}[\phi(x_s)\phi(x_s)^T])^2$. By construction and Assumption **A III** we can show that W_t has d^* non-zero eigenvalues (See Section 4.1 in the supplementary material).

Lemma 5.1 (Lower bound on r^{th} Eigen-value of compact self-adjoint operators)**.** *Let $x_t \in \mathcal{X}$, $t \geq 1$ be generated sequentially from a random process. Assume that conditions **A I**-**A III** hold. Let $p(t) = \min(-t, 1)$ and $\forall b \geq 0, a > \frac{1}{6}(L^2 + \sqrt{L^4 + 36b})$ let $\tilde{d} := 50 \sum_{r=1}^{d^*} p(-\frac{a\lambda_r(\mathbb{E}W_t)}{L^2 b}) \leq 50 d^*$. Let*

$$A(t, \delta) = \log \frac{(tL^4 + 1)(tL^4 + 3)\tilde{d}}{\delta},$$

and

$$h(t, \delta) = \left(t\lambda_x - \frac{L^2}{3}\sqrt{18tA(t, \delta) + A(t, \delta)^2}) - \frac{L^2}{3}A(t, \delta)\right).$$

Then for any $\delta > 0$,

$$\lambda_r(S_t) \geq h(t, \delta)_+$$

holds for all $t > 0$ with probability at least $1 - \delta$. Furthermore, if L =1, $r \leq d^*$ and $0 < \delta \leq \frac{1}{8}$, then the event

$$\lambda_r(S_t) \geq \frac{t\lambda_x}{2}, \forall t \geq \frac{256}{\lambda_x^2}\log(\frac{128\tilde{d}}{\lambda_x^2\delta}),$$

holds with probability at least $1 - \delta$.

First we state the Lemmas that we use to prove Lemma 5.1.

Lemma 5.2 (Lemma 9 in Li & Zhang (2018))**.** *If $a > 0, b > 0, ab \geq e$, then for all $t \geq 2a\log(ab)$,*

$$t \geq a\log(bt). \tag{3}$$

Lemma 5.3 (Lemma 1.1 in Zi-Zong (2009))**.** *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix partitioned according to*

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{12}^T & A_{22} \end{array}\right],$$

where $A_{11} \in \mathbb{R}^{(n-1)\times(n-1)}, A_{12} \in \mathbb{R}^{(n-1)}$ and $A_{22} \in \mathbb{R}^1$. Then $\det(A) = \det(A_{11})(A_{22} - A_{12}^T A_{11}^{-1} A_{12})$.

Lemma 5.4 (Special case of extended Horn's inequality (Theorem 4.5 of Bercovici et al. (2009)))**.** *Let A, B be compact self-adjoint operators. Then for any $p \geq 1$,*

$$\lambda_p(A + B) \leq \lambda_1(A) + \lambda_p(B). \tag{4}$$

Theorem 5.5 (Freedman's inequality for self adjoint operators, Thm 3.2 & section 3.2 in Minsker (2017))**.** *Let $\{\Phi_t\}_{t=1,\ldots}$ be a sequence of self-adjoint Hilbert Schmidt operators $\Phi_t : \mathcal{H} \to \mathcal{H}$ acting on a seperable Hilbert space ($\mathbb{E}\Phi$ is a operator such that $\langle(\mathbb{E}\Phi)z_1, z_2\rangle_{\mathcal{H}} = \mathbb{E}\langle\Phi z_1, z_2\rangle_{\mathcal{H}}$ for any $z_1, z_2 \in \mathcal{H}$). Additionally, assume that $\{\Phi_t\}_{t=1,\ldots}$ is a martingale difference sequence of self adjoint operators such that $\|\Phi_t\| \leq L^2$ almost surely for all $1 \leq t \leq T$ and some positive $L \in \mathbb{R}$. Denote by $W_t = \sum_{s=1}^t \mathbb{E}_{s-1}[\Phi_s^2]$ and $p(t) = \min(-t, 1)$. Then for any $a \geq \frac{1}{6}(L^2 + \sqrt{L^4 + 36b}), b \geq 0$,*

$$\mathbb{P}\left(\|\sum_{j=1}^t \Phi_j\| > a \text{ and } \lambda_1(W_t) \leq b\right) \leq \tilde{d}\cdot\exp\left(-\frac{a^2/2}{b + aL^2/3}\right),$$

where $\|\cdot\|$ is the operator norm and $\tilde{d} := 50\sum_{r=1}^{\infty}(p(-\frac{a\lambda_r(\mathbb{E}W_t)}{L^2 b}))$.

Note that $\tilde{d}$ is a function of t but it's upper bounded by d^* which is the rank of $\mathbb{E}_{s-1}[\phi(x)\phi(x)^T]$.

5.1.1. Proof of Lemma 5.1

Lemma 7 in Li & Zhang (2018) gives the lower bound on minimum eigenvalue (finite dimensional case) when reward depends linearly on context. We extend it to r^{th} largest eigenvalue (infinite dimensional case) and the case when reward depends non-linearly on context.

Proof. $\mathcal{X} \subset \mathbb{R}^d$ is a compact space endowed with a finite positive Borel measure. For a continuous kernel k the canonical feature map ϕ is a continuous function $\phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is a separable Hilbert space (See section 2 of Micchelli et al. (2006) for a construction such that $\mathcal{H}$ is separable). In such a setting $\phi(\mathcal{X})$ is also compact space with a finite positive Borel measure (Micchelli et al., 2006). We now define a few terms on $\phi(\mathcal{X})$.

Define the random variable $\Phi_t := \mathbb{E}_{t-1}[\phi(x_t)\phi(x_t)^T] - \phi(x_t)\phi(x_t)^T$. Let $Z_t := \sum_{s=1}^t \Phi_s = \sum_{s=1}^t \mathbb{E}_{s-1}[\phi(x_t)\phi(x_t)^T] - S_t = V_t - S_t$.

By construction, $\{Z_t\}_{t=1,2,\ldots}$ is a martingale and $\{\Phi_s\}_{s=1,2,\ldots}$ is the martingale difference sequence. Notice that $\lambda_1(\Phi_t) \leq L^2$. To use the Freedman's inequality, we lower bound the operator norm of Z_t, $\|Z_t\|$ and upper bound the largest eigenvalue of W_t, $\lambda_1(W_t)$. Let $\nu(A) = \max_i|\lambda_i(A)|$ be the spectral radius of operator A. We work with the spectral radius because it is not necessary that Z_t is a positive definite operator. It is well known that

$$\nu(A) \leq \|A\|. \tag{5}$$

By assumption **A III**, $\mathbb{E}_{s-1}[\phi(x)\phi(x)^T]$ lies in a fixed d^* dimensional subspace with its eigenvalues $\lambda_r(\mathbb{E}_{s-1}[\phi(x)\phi(x)^T]) > \lambda_x$ for $r \leq d^*$. Thus, for $V_t = \sum_{s=1}^t \mathbb{E}_{s-1}[\phi(x)\phi(x)^T]$, $\lambda_r(V_t) \geq t\lambda_x$.

Bound on $\|Z_t\|$: By definition, $V_t = Z_t + S_t$. Hence, $\lambda_r(V_t) \leq \lambda_1(Z_t) + \lambda_r(S_t)$ by using Horn's inequality (Lemma 5.4).

$$\begin{aligned} \lambda_1(Z_t) &\geq \lambda_r(V_t) - \lambda_r(S_t) \\ \lambda_1(Z_t) &\geq t\lambda_x - \lambda_r(S_t) \\ \nu(Z_t) &\geq t\lambda_x - \lambda_r(S_t), \end{aligned}$$

where the second step is due to **A III** and the third step is by definition of spectral radius. By Eqn. (5), we have

$$\|Z_t\| \geq t\lambda_x - \lambda_r(S_t). \tag{6}$$

Bound on $\lambda_1(W_t)$: To bound the term $\lambda_1(W_t)$, write

$$\begin{aligned} W_t &= \sum_{s=1}^t \mathbb{E}_{s-1}[\Phi_s^2] \\ &= \sum_{s=1}^t \mathbb{E}_{s-1}\left[(\mathbb{E}_{s-1}[\phi(x_s)\phi(x_s)^T] - \phi(x_s)\phi(x_s)^T)^2\right]. \end{aligned}$$

By using square expansion,

$$\begin{aligned} W_t &= \sum_{s=1}^{t} \mathbb{E}_{s-1}\big[(\mathbb{E}_{s-1}[\phi(x_s)\phi(x_s)^T]^2 + (\phi(x_s)\phi(x_s)^T)^2 \\ &\quad -\mathbb{E}_{s-1}[\phi(x_s)\phi(x_s)^T](\phi(x_s)\phi(x_s)^T) \\ &\quad -(\phi(x_s)\phi(x_s)^T)\mathbb{E}_{s-1}[\phi(x_s)\phi(x_s)^T]\big] \\ &= \sum_{s=1}^{t} \mathbb{E}_{s-1}[(\phi(x_s)\phi(x_s)^T)^2] - \mathbb{E}_{s-1}[\phi(x_s)\phi(x_s)^T]^2. \end{aligned}$$

Taking norm on both sides,

$$\|W_t\| = \|\sum_{s=1}^{t} \mathbb{E}_{s-1}[(\phi(x_s)\phi(x_s)^T)^2] - \mathbb{E}_{s-1}[\phi(x_s)\phi(x_s)^T]^2\|.$$

As both terms on the right hand side are positive semi-definite matrices,

$$\|W_t\| \leq \|\sum_{s=1}^{t} \mathbb{E}_{s-1}[(\phi(x_s)\phi(x_s)^T)^2]\|.$$

Next, we use convexity properties of norms to get the upper bound.

$$\begin{aligned} \|W_t\| &\leq \sum_{s=1}^{t} \|\mathbb{E}_{s-1}[(\phi(x_s)\phi(x_s)^T)^2]\| \\ &= \sum_{s=1}^{t} \|\mathbb{E}_{s-1}[(\phi(x_s)\big(\phi(x_s)^T\phi(x_s)\big)\phi(x_s)^T)]\| \\ &\leq L^2 \sum_{s=1}^{t} \|\mathbb{E}_{s-1}[(\phi(x_s)\phi(x_s)^T)]\| \\ &\leq L^2 \sum_{s=1}^{t} \mathbb{E}_{s-1}[\|(\phi(x_s)\phi(x_s)^T)\|] \end{aligned}$$

where the first step is due to the triangle inequality and the third step is due to the upper bound $\|\phi(x)\| \leq L$, the fourth step is due to the convexity of the operator norm and Jensen's inequality. Using the properties of Hilbert Schmidt operators, we can write

$$\begin{aligned} \mathbb{E}_{s-1}[\|(\phi(x_s)\phi(x_s)^T)\|] &\leq \mathbb{E}_{s-1}[\|(\phi(x_s)\phi(x_s)^T)\|_{HS}] \\ &= \mathbb{E}_{s-1}[\|\phi(x_s)\|^2] \leq L^2 \end{aligned}$$

Therefore, we can bound the norm $\|W_t\|$ as

$$\begin{aligned} \|W_t\| &\leq L^2 \sum_{s=1}^{t} L^2 \\ &= tL^4, \end{aligned}$$

Again, by using Eqn. (5), we have

$$\lambda_1(W_t) \leq tL^4. \tag{7}$$

Now, we shall construct a parameter A such that

$$\frac{a^2/2}{b + aL^2/3} \geq A. \tag{8}$$

For this inequality to hold, one can see, by its quadratic solution, $a \geq f(A,b) := \frac{1}{3}AL^2 + \sqrt{\frac{1}{9}A^2L^4 + 2Ab}$. Note that for $A > 1$, the condition of $a \geq f(A,b)$ also satisfies the conditions of Friedman's inequality in Theorem 5.5.

Let $A(m,\delta) = \log \frac{(m+1)(m+3)}{\delta}$ and P be the probability of event $\Big[\exists t : \lambda_r(\boldsymbol{S}_t) \leq t\lambda_x - f(A(tL^4,\delta), tL^4)\Big]$.

$$\begin{aligned} &P \\ &= \mathbb{P}\Big[\exists t : \lambda_r(\boldsymbol{S}_t) \leq t\lambda_x - f(A(tL^4,\delta), tL^4)\Big] && (9) \\ &\leq \mathbb{P}\Big[\exists t : \lambda_r(\boldsymbol{S}_t) \leq t\lambda_x \\ &\quad -f(A(\lambda_1(W_t),\delta), \lambda_1(W_t))\Big] && (10) \\ &\leq \sum_{m=0}^{\infty} \mathbb{P}\Big[\exists t : \lambda_r(\boldsymbol{S}_t) \leq t\lambda_x \\ &\quad -f(A(m,\delta),m), \lambda_1(W_t) \leq m\Big] && (11) \\ &\leq \sum_{m=0}^{\infty} \mathbb{P}\Big[\exists t : \|Z_t\| \geq f(A(m,\delta),m), \\ &\quad \lambda_1(W_t) \leq m\Big] && (12) \\ &\leq \tilde{d} \sum_{m=0}^{\infty} \exp\left(-A(m,\delta)\right) && (13) \\ &= \tilde{d} \sum_{m=0}^{\infty} \frac{\delta}{(m+1)(m+3)} \\ &\leq \tilde{d} \cdot \delta, && (14) \end{aligned}$$

where (10) is because A is increasing in m, f is increasing in A, b, and Eqn. (7). Eqn. (11) is by application of the union bound over all the events for which $\lambda_1(W_t) \leq m$. Also, Eqn. (12) is due to Eqn. (6) and Eqn. (13) is due to Theorem 5.5.

The result is obtained by replacing δ by $\frac{\delta}{\tilde{d}}$.

For the second part. Let $\tilde{\lambda}_x := \frac{\lambda_x}{L}$. By definition of L, $\tilde{\lambda}_x \leq 1$. Let $t \geq \frac{256}{\tilde{\lambda}_x^2} \log \frac{128\tilde{d}}{\tilde{\lambda}_x^2\delta}$. Then by using the Lemma 5.2,

$$t \geq \frac{128}{\tilde{\lambda}_x^2} \log \frac{t\tilde{d}}{\delta}. \tag{15}$$

Rearranging the terms, we get

$$\frac{t\tilde{\lambda}_x^2}{4} \geq 32 \log \frac{t\tilde{d}}{\delta}$$

Taking square root and then multiplying by $\sqrt{t}$ on both sides

$$\begin{aligned}\frac{t\tilde{\lambda}_x}{2} &\geq \sqrt{32t\log\frac{t\tilde{d}}{\delta}} \\ &= \frac{2}{3}\sqrt{72t\log\frac{t\tilde{d}}{\delta}} \\ &= \frac{2}{3}\sqrt{36t\log\frac{t\tilde{d}}{\delta}+36t\log\frac{t\tilde{d}}{\delta}}.\end{aligned}$$

Using equation (15),

$$\begin{aligned}\frac{t\tilde{\lambda}_x}{2} &\geq \frac{2}{3}\sqrt{36t\log\frac{t\tilde{d}}{\delta}+\frac{36\cdot 128}{\tilde{\lambda}_x^2}\Big(\log\frac{t\tilde{d}}{\delta}\Big)^2} \\ &= \frac{2}{3}\sqrt{36t\log\frac{t\tilde{d}}{\delta}+\frac{36\cdot 32}{\tilde{\lambda}_x^2}4\Big(\log\frac{t\tilde{d}}{\delta}\Big)^2}.\end{aligned}$$

Since $\tilde{\lambda}_x^2 \leq 1$ we have

$$\begin{aligned}\frac{t\tilde{\lambda}_x}{2} &\geq \frac{2}{3}\sqrt{36t\log\frac{t\tilde{d}}{\delta}+(36\cdot 32)\Big(2\log\frac{t\tilde{d}}{\delta}\Big)^2} \\ &> \frac{2}{3}\sqrt{18t\cdot 2\log\frac{t\tilde{d}}{\delta}+\Big(2\log\frac{t\tilde{d}}{\delta}\Big)^2}. \qquad (16)\end{aligned}$$

Now we use the condition on δ as stated in the Theorem statement: $0 \leq \delta \leq \frac{1}{8}$. We can see that

$$\frac{1}{8} \leq \frac{t^2\tilde{d}^2}{(t+1)(t+3)}, \qquad (17)$$

because $\frac{t^2\tilde{d}^2}{(t+1)(t+3)}$ is a monotonically increasing function for both $t, \tilde{d}$ for $t, \tilde{d} \geq 1$. Simplifying Eqn. (17), we get

$$\frac{t^2\tilde{d}^2}{\delta^2} \geq \frac{(t+1)(t+3)}{\delta}. \qquad (18)$$

Taking log of both sides,

$$2\log\frac{t\tilde{d}}{\delta} \geq \log(\frac{(t+1)(t+3)}{\delta}) = A(t,\delta).$$

Without loss of generality, we will assume that $L = 1$. From Eqn. (18) and Eqn. (16), we have

$$\begin{aligned}\frac{t\lambda_x}{2} &\geq \frac{2}{3}\sqrt{18t\cdot A(t,\delta)+A(t,\delta)^2} \\ &= \frac{1}{3}\sqrt{18t\cdot A(t,\delta)+A(t,\delta)^2} \\ &\quad +\frac{1}{3}\sqrt{18t\cdot A(t,\delta)+A(t,\delta)^2} \\ &\geq \frac{1}{3}\sqrt{18t\cdot A(t,\delta)+A(t,\delta)^2}+\frac{1}{3}A(t,\delta)\end{aligned}$$

Therefore,

$$\frac{t\lambda_x}{2} \geq f(A(t,\delta),t). \qquad (19)$$

Equations (9) and (19) complete the proof. □

5.2. Monotonic Upper bound of $s_{a,t}(x)$

Lemma 5.1 provides high probability lower bounds on the minimum nonzero eigenvalue of the cumulative second moment operator S_t. Using the preceding lemma and the confidence interval defined in Theorem 2.1, it is possible to provide high probability monotonic bounds on the confidence interval widths $s_{a,t}(x)$.

Lemma 5.6 (Monotonic upper bound of $s_{a,t}(x_t)$)**.** *Consider a contextual bandit simple regret minimization problem with assumptions **A I-A III** and fix T. Assume $\|\phi(x)\| \leq 1$, $\lambda > 0$ and $\forall a \in [A]$, $N_{a,t} > N_\lambda := \max\Big(\frac{2(1-\lambda)}{\lambda_x}, d^*, \frac{256}{\lambda_x^2}\log(\frac{128\tilde{d}}{\lambda_x^2\delta})\Big)$. Then, for any $0 < \delta \leq \frac{1}{8}$,*

$$s_{a,t}(x_t)^2 \leq g_{a,t}(N_{a,t})$$

with probability at least $1-\delta$, for the monotonically decreasing function $g_{a,t}$ defined as $g_{a,t}(N_{a,t}) := 8(C_1\beta + C_2)^2\Big(\frac{1}{\lambda+N_{a,t}\lambda_x/2}\Big)$.

The condition $N_{a,t} > N_\lambda$ results in a minimum number of tries that arm a has to be selected before any bound will hold. In $N_\lambda := \max\Big(\frac{2(1-\lambda)}{\lambda_x}, d^*, \frac{256}{\lambda_x^2}\log(\frac{128\tilde{d}}{\lambda_x^2\delta})\Big)$, the first and third term in the max are needed so that we can give concentration bounds on eigenvalues and prove that the confidence width shrinks. The second term is needed because one has to get at least d^* contexts for every arm so that at least some energy is added to the lowest eigenvalues.

Lemma 5.7. *[Arithmetic Mean-Geometric Mean Inequality (Steele, 2004)] For every sequence of nonnegative real numbers $a_1, a_2, ...a_n$ one has*

$$(\prod_{i=1}^{n} a_i)^{1/n} \leq \frac{\sum_{i=1} a_i}{n}$$

with equality if and only if $a_1 = a_2 = ... = a_n$.

Lemma 5.8. *If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$, and $\mu_1 \geq 0, \mu_2 \geq 0 \cdots \mu_d \geq 0$ such that $\sum_j \mu_j = L$ and $\lambda_d \geq L$ then*

$$\prod_{i=1}^{d}\Big(1+\frac{\mu_i}{\lambda_i}\Big) - 1 \leq \frac{2L}{\lambda_d}.$$

Proof. By replacing each λ_i with the smallest element λ_d

we get,

$$\begin{aligned}\prod_{i=1}^{d}\left(1+\frac{\mu_i}{\lambda_i}\right)-1 &\leq \prod_{i=1}^{d}\left(1+\frac{\mu_i}{\lambda_d}\right)-1\\ &= \prod_{i=1}^{d}\left(\frac{\lambda_d+\mu_i}{\lambda_d}\right)-1\\ &= \left(\frac{\prod_{i=1}^{d}(\lambda_d+\mu_i)}{\lambda_d^d}\right)-1\\ &\leq \left(\frac{\sum_{i=1}^{d}(\lambda_d+\mu_i)}{d\lambda_d}\right)^d-1\\ &= \left(\frac{d\lambda_d+L}{d\lambda_d}\right)^d-1\\ &= \left(1+\frac{L}{d\lambda_d}\right)^d-1\\ &\leq e^{L/\lambda_d}-1,\end{aligned}$$

where the fourth inequality is by Lemma 5.7 and last inequality holds because $(1+\frac{a}{x})^x$ approaches e^a as $x\to\infty$ and $(1+\frac{a}{x})^x$ is a monotonically increasing function of x.

By $e^x \leq 1+2x$ for $x\in[0,1]$ and the assumption that $\lambda_d \geq L$,

$$\begin{aligned}\prod_{i=1}^{d}\left(1+\frac{\mu_i}{\lambda_i}\right)-1 &\leq e^{L/\lambda_d}-1\\ &\leq 1+\frac{2L}{\lambda_d}-1\\ &= \frac{2L}{\lambda_d}.\end{aligned}$$

□

5.2.1. Proof of Lemma 5.6

Proof. We will assume that $L=1$. We write

$$K_{a,t+1}+\lambda_{I_{a,t+1}} = \left[\begin{array}{c|c} K_{a,t}+\lambda_{I_{a,t}} & k_{a,t}(x) \\ \hline k_{a,t}(x)^T & k(x,x)+\lambda \end{array}\right].$$

Let $\mu_i = \lambda_i(K_{a,t+1}+\lambda_{I_{a,t+1}}) - \lambda_i(K_{a,t}+\lambda_{I_{a,t}})$.

Using Lemma 5.3,

$$\begin{aligned}&\det(K_{a,t+1}+\lambda_{I_{a,t+1}})\\ =\ &\det(K_{a,t}+\lambda_{I_{a,t}})\Big(k(x,x)+\lambda\\ &-k_{a,t}(x)^T(K_{a,t}+\lambda_{I_{a,t}})^{-1}k_{a,t}(x)\Big).\end{aligned}$$

Rearranging,

$$\begin{aligned}&k(x,x)-k_{a,t}(x)^T(K_{a,t}+\lambda_{I_{a,t}})^{-1}k_{a,t}(x)\\ =\ &\frac{\det(K_{a,t+1}+\lambda_{I_{a,t+1}})}{\det(K_{a,t}+\lambda_{I_{a,t}})}-\lambda.\end{aligned}$$

Dividing both sides by λ,

$$\begin{aligned}&\frac{k(x,x)-k_{a,t}(x)^T(K_{a,t}+\lambda_{I_{a,t}})^{-1}k_{a,t}(x)}{\lambda}\\ =\ &\frac{\det(K_{a,t+1}+\lambda_{I_{a,t+1}})}{\lambda\det(K_{a,t}+\lambda_{I_{a,t}})}-1. \qquad (20)\end{aligned}$$

Notice that the left hand side is equal to $\frac{\hat{\sigma}_{a,t}(x)}{\lambda}$. Using the definitions of $s_{a,t}(x)$ and $\hat{\sigma}_{a,t}(x)$, we can write,

$$\begin{aligned}s_{a,t}(x)^2 &= 4(C_1\beta+C_2)^2\frac{\hat{\sigma}_{a,t}(x)^2}{\lambda}\\ &= 4(C_1\beta+C_2)^2\Big(\frac{\det(K_{a,t+1}+\lambda I_{a,t+1})}{\lambda\det(K_{a,t}+\lambda I_{a,t})}-1\Big)\\ &= 4(C_1\beta+C_2)^2\Big(\frac{\prod_{i=1}^{N_{a,t}+1}\lambda_{i,a,t+1}}{\lambda\prod_{i=1}^{N_{a,t}}\lambda_{i,a,t}}-1\Big)\end{aligned}$$

By assumption in the statement of the Lemma, $N_{a,t}\geq d^*$. Hence, all eigenvalues above d^* are λ.

By replacing all eigenvalues $\lambda_{i,a,\tau}$ by λ for $\tau=\{t,t+1\}$ and $i>d^*$, we get

$$s_{a,t}(x)^2 = 4(C_1\beta+C_2)^2\Big(\prod_{i=1}^{d^*}\frac{\lambda_{i,a,t+1}}{\lambda_{i,a,t}}-1\Big).$$

Note that $\lambda_{i,a,t+1}=\lambda_{i,a,t}+\mu_i$. By replacing $\lambda_{i,a,t+1}$, we get

$$\begin{aligned}s_{a,t}(x)^2 &= 4(C_1\beta+C_2)^2\Big(\prod_{i=1}^{d^*}\frac{\lambda_{i,a,t}+\mu_i}{\lambda_{i,a,t}}-1\Big)\\ &= 4(C_1\beta+C_2)^2\left(\prod_{i=1}^{d^*}\left(1+\frac{\mu_i}{\lambda_{i,a,t}}\right)-1\right)\\ &\leq 4(C_1\beta+C_2)^2\left(1+\frac{2L}{\lambda_{d^*,a,t}}-1\right),\end{aligned}$$

where the third inequality is due to Lemma 5.8.

For $L=1$,

$$\begin{aligned}s_{a,t}(x)^2 &\leq 4(C_1\beta+C_2)^2\left(1+\frac{2}{\lambda_{d^*,a,t}}-1\right)\\ &= 4(C_1\beta+C_2)^2\left(\frac{2}{\lambda_{d^*,a,t}}\right).\end{aligned}$$

Note that $\lambda_{d^*,a,t} = \lambda_{d^*}(K_{a,t+1}+\lambda_{I_{a,t+1}}) = \lambda_{d^*}(K_{a,t+1})+\lambda$. By Lemma 5.1 in main paper $\lambda_{d^*}(K_{a,t+1})\geq N_{a,t}\lambda_x$. We can apply Lemma 5.8 only when

$$\frac{1}{\lambda+N_{a,t}\lambda_x/2} < 1$$

or

$$N_{a,t} \quad > \quad \frac{2(1-\lambda)}{\lambda_x}.$$

The assumption in the statement of the lemma satisfies the above equation. Hence, we have

$$\begin{aligned} s_{a,t}(x)^2 &\leq 4(C_1\beta + C_2)^2\left(\frac{2}{\lambda + N_{a,t}\lambda_x/2}\right) \\ &= 8(C_1\beta + C_2)^2\left(\frac{1}{\lambda + N_{a,t}\lambda_x/2}\right) \\ &= g_{a,t}(N_{a,t}). \end{aligned}$$

This concludes the proof. □

5.3. Closed form of $g_{a,t}^{-1}(s)$

Now we calculate a closed form expression of $N_{a,t}$. Setting the upper bound on confidence in the Theorem 4.1 in main paper to s, we calculate the inverse in terms of $N_{a,t}$,

$$8(C_1\beta + C_2)^2\left(\frac{1}{\lambda + N_{a,t}\lambda_x/2}\right) \quad = \quad s^2.$$

Rearranging all the terms, we get

$$\begin{aligned} 8(C_1\beta + C_2)^2 &= s^2(\lambda + N_{a,t}\lambda_x/2) \\ (\lambda + N_{a,t}\lambda_x/2) &= \frac{8(C_1\beta + C_2)^2}{s^2} \\ N_{a,t} &= \frac{16(C_1\beta + C_2)^2}{s^2\lambda_x} - \frac{2\lambda}{\lambda_x}. \end{aligned}$$

Define

$$g_{a,t}^{-1}(s) = \frac{16(C_1\beta + C_2)^2}{s^2\lambda_x} - \frac{2\lambda}{\lambda_x}. \tag{21}$$

5.4. Simple Regret Analysis

These high probability monotonic upper bounds on the confidence estimate can be used to upper bound the simple regret. The upper bound depends on a context-based hardness quantity defined for each arm a (similar to Hoffman et al. (2014)) as

$$H_{a,\epsilon}(x) = \max(\frac{1}{2}(\Delta_a(x) + \epsilon), \epsilon). \tag{22}$$

Denote its lowest value as $H_{a,\epsilon} := \inf_{x\in\mathcal{X}} H_{a,\epsilon}(x)$. Let total hardness be defined as $H_\epsilon := \sum_{a\in[A]} H_{a,\epsilon}^{-2}$ (Note that $H_\epsilon \leq \frac{A}{\epsilon^2}$). The recommended arm after time $t \geq T$ is defined as

$$\Omega(x) = J_{\arg\min_{AN_\lambda+1\leq\tau\leq T} B_{J_\tau(x_t),t}(x_t)}(x_t)$$

from Algorithm in the main paper. We now upper bound the simple regret as follows:

Theorem 5.9. *Consider a contextual bandit problem as defined in the main paper with assumptions **A I-A III**. For $0 < \delta \leq \frac{1}{8}$, $\epsilon > 0$ and $N_\lambda := \max\left(\frac{2(1-\lambda)}{\lambda_x}, d^*, \frac{256}{\lambda_x^2}\log(\frac{128\tilde{d}}{\lambda_x^2\delta})\right)$, let*

$$\beta = \sqrt{\frac{\lambda_x(T - N_\lambda(A-1)) + 2A\lambda}{16C_1^2H_\epsilon}} - \frac{C_2}{C_1}. \tag{23}$$

For all $t > T$ and $\epsilon > 0$,

$$\mathbb{P}(R_{\Omega(x_t)}(x_t) < \epsilon | x_t) \geq 1 - A(T - AN_\lambda)e^{-\beta^2} - A\delta. \tag{24}$$

Note that the term C_2 in (23) grows logarithmically in T (see supplementary material). For β to be positive, T should be greater than $\frac{16H_\epsilon C_2^2 - 2A\lambda}{\lambda_x} + N_\lambda(A-1)$. We compare the term $e^{-\beta^2}$ in our bound with the uniform sampling technique in Guan & Jiang (2018) which leads to a bound that decay like $Ce^{-cT^{\frac{2}{(d_1+d)}}} \geq Ce^{-cT^{\frac{2}{(2+d)}}}$, where $d_1 \geq 2$, d is the context dimension, and C and c are constants. In our case, the decay rate has the form $C'Te^{-c'T}$ for constants C', c'. Clearly, our bound is superior for $\forall d \geq 1$. We can also compare Theorem 5.9 with Bayes Gap (Hoffman et al., 2014) and UGapEb (Gabillon et al., 2012) which provide simple regret guarantees in the multi-armed bandit setting. Bayes Gap and UGapEb have regret bounds of order $O(ATe^{-\frac{T-A}{H_\epsilon}})$ and we provide bounds of order $O(A(T - AN_\lambda)e^{-\frac{T-AN_\lambda}{H_\epsilon}})$. Ignoring other constants, our method has the additional term N_λ which is required because algorithm needs to see enough number of contexts to get information about context space and to become confident in the reward estimates in that context space. The simple regret bound is also dependent on the gap between the arms. A larger gap quantity Δ_a implies a larger $H_{a,\epsilon}$ which implies that quantity $e^{-\frac{1}{H_\epsilon}}$ is small. This means that a larger gap quantity leads to a faster rate.

Note that there are two choices for simple regret analysis: 1) bounding the simple regret uniformly (Theorem 5.9) and 2) average simple regret $\left(\sum_{t>T} R_{\Omega(x_t)}(x_t)\right)$. We bound the simple regret uniformly and it may require stronger distributional assumptions (e.g. Assumption **A III**) compared to average simple regret. Furthermore, we provide uniform bounds and not average simple regret bounds since our problem setting of simple regret minimization and the motivating application require performance guarantees for every time step during exploitation, as opposed to average simple regret guarantees.

Lemma 5.10 (Value of β)**.** *Assume the conditions in Theorem 4.1 and Lemma 4.2 in main paper. If $\sum_{a\in[A]} g_{a,t}^{-1}(H_{a\epsilon}) = T - N_\lambda(A-1)$, then*

$$\beta = \sqrt{\frac{\lambda_x(T - N_\lambda(A-1)) + 2A\lambda}{16C_1^2H_\epsilon}} - \frac{C_2}{C_1}. \tag{25}$$

Proof. We have

$$\sum_{a\in[A]} g_{a,t}^{-1}(H_{a\epsilon}) \quad = \quad T - N_\lambda(A-1).$$

By using Eqn. (21),

$$\sum_{a\in[A]} \frac{16(C_1\beta+C_2)^2}{H_{a\epsilon}^2\lambda_x} - \frac{2\lambda}{\lambda_x} \quad = \quad T - N_\lambda(A-1)$$

$$\frac{16(C_1\beta+C_2)^2}{\lambda_x}\sum_{a\in[A]} \frac{1}{H_{a\epsilon}^2} - \frac{2A\lambda}{\lambda_x} \quad = \quad T - N_\lambda(A-1).$$

By using definition of H_ϵ,

$$\frac{16(C_1\beta+C_2)^2H_\epsilon}{\lambda_x} - \frac{2A\lambda}{\lambda_x} \quad = \quad T - N_\lambda(A-1)$$

Rearranging the terms,

$$\begin{aligned} 16(C_1\beta+C_2)^2H_\epsilon &= \lambda_x(T-N_\lambda(A-1)) + 2A\lambda \\ (C_1\beta+C_2)^2 &= \frac{\lambda_x(T-N_\lambda(A-1)) + 2A\lambda}{16H_\epsilon} \\ \beta &= \sqrt{\frac{\lambda_x(T-N_\lambda(A-1)) + 2A\lambda}{16C_1^2H_\epsilon}} - \frac{C_2}{C_1} \end{aligned}$$

□

5.5. Proof of Theorem 5.9

Let $[A] = \{1, ..., A\}$. We define a feasible set $A'(x) \subseteq [A]$ such that elements of $A'(x)$ contain possible set of arms that may be pulled if context x was observed at all times $AN_\lambda < t \leq T$. The set $A'(x)$ is used to discount the arms that will never be pulled with context x.

Proof. The proof broadly follows the same structure presented in Theorem 2 of Hoffman et al. (2014). We will provide the simple regret bound at the recommendation of time $T+1$, since the algorithm operates in a pure exploitation setting, the recommended arm $\Omega(x_{T+2})$ will follow the same properties.

Fix $x \in \mathcal{X}$ such that x can be generated from the filtration. We define the event $\mathcal{E}_{a,t}(x)$ to be the event in which for arm $a \leq A$, $f_a(x)$ lies between the upper and lower confidence bounds given $x_1, x_2, ..., x_{t-1}$ More precisely,

$$\mathcal{E}_{a,t}(x) = \{L_{a,t}(x_t) \leq f_a(x) \leq U_{a,t}(x) | x_1, x_2, \cdots, x_{t-1}\}.$$

For events $\mathcal{E}_{a,t}$, from Theorem 4.1 of the main paper,

$$\mathbb{P}(\mathcal{E}_{a,t}(x)) \geq 1 - e^{-\beta^2}.$$

Let $N_{a,T}$ denote the number of times each arm has been tried upto time T. Clearly $\sum_{a=1}^{A} N_{a,T} = T$. Also, note that we try each arm at least N_λ number of times before we run our algorithm. We define event $\mathcal{E}$ as $\mathcal{E} := \bigcup_{a\leq A, AN_\lambda < t \leq T} \mathcal{E}_{a,t}(x)$. By the union bound we can show that

$$\mathbb{P}(\mathcal{E}) \geq 1 - A(T - AN_\lambda)e^{-\beta^2}.$$

The next part of the proof works by contradiction.

Let $\epsilon > 0$. The recommended arm at the end of time T for context x is defined as follows: let $t^* := \arg\min_{AN_\lambda < t \leq T} B_{J_t(x),t}(x)$ then the recommended arm is $\Omega(x) := J_{t^*}(x)$.

Conditioned on event $\mathcal{E}$, we will assume that the event $R_{\Omega(x)}(x) > \epsilon$ is true and arrive at a contradiction with high probability. Note that if $R_{\Omega(x)}(x) > \epsilon$, the recommended arm $\Omega(x)$ is necessarily sub-optimal (regret is zero for the optimal arm).

Define $M_{a,T}(x)$ as number of times arm $a \in [A]$ would be selected in $AN_\lambda < t \leq T$, if we had seen context x at all those times. Hence, $\sum_{a\in A'(x)} M_{a,T}(x) = T - AN_\lambda$. Also, note that $N_{a,T}(x) = M_{a,T}(x) + N_\lambda$ for $a \in A'(x)$ and $N_{a,T}(x) = N_\lambda$ otherwise. Let $t_a = t_a(x)$ be the last time instant for which arm $a \in A'(x)$ may have been selected using the Contextual-Gap algorithm if context x was observed throughout.

The following holds for the recommended arm $\Omega(x)$ with context x:

$$\begin{aligned} \min(0, s_{a,t_a}(x) - \Delta_a(x)) + s_{a,t_a}(x) &\geq B_{J_{t_a}(x),t_a}(x) \\ &\geq B_{\Omega(x),T+1}(x) \\ &\geq R_{\Omega(x)}(x) \\ &> \epsilon. \end{aligned}$$

Where the first inequality holds due to Lemma 5.14, the second inequality holds by definition of $B_{\Omega(x),T+1}$, the third inequality holds due to Lemma 5.11 and the last inequality holds due to the event $R_{\Omega(x)} > \epsilon$. The preceding inequality can also be written as

$$\begin{aligned} &s_{a,t_a}(x) > 2s_{a,t_a}(x) - \Delta_a(x) > \epsilon, && \text{if } \Delta_a(x) > s_{a,t_a}(x). \\ &2s_{a,t_a}(x) - \Delta_a(x) > s_{a,t_a}(x) > \epsilon, && \text{if } \Delta_a(x) < s_{a,t_a}(x). \end{aligned}$$

This leads to the following bound on the confidence diameter of $a \in [A]$,

$$s_{a,t_a}(x) > \max(\frac{1}{2}(\Delta_a(x) + \epsilon), \epsilon) =: H_{a\epsilon}(x).$$

For any arm a, we consider the final number of arm pulls $M_{a,T}(x) + N_\lambda$. From Lemma 5.2 of the main paper we can write, using the strict monotonicity and there by invertibility of $g_{a,T}$, with probability at least $1 - \delta$ as

$$\begin{aligned} M_{a,T}(x) + N_\lambda &\leq g_{a,T}^{-1}(s_{a,t_a}(x)) \\ &< g_{a,T}^{-1}(H_{a\epsilon}(x)) \\ &\leq g_{a,T}^{-1}(H_{a\epsilon}), \end{aligned}$$

where $H_{a\epsilon} = \inf_x H_{a\epsilon}(x)$. Last two equations hold as $g_{a,T}$ is a monotonically decreasing function. By summing both sides with respect to $a \in A'(x)$ we can write

$$T - AN_\lambda + |A'(x)|N_\lambda \quad < \quad \sum_{a \in A'(x)} g_{a,T}^{-1}(H_{a\epsilon}),$$

We can make RHS even bigger by adding terms $a \in [A] \backslash A'(x)$. Hence, we get

$$T - (A - |A'(x)|)N_\lambda \quad < \quad \sum_{a \in [A]} g_{a,T}^{-1}(H_{a\epsilon}).$$

We can make LHS even smaller by noting that minimum value of $|A'(x)|$ is one.

$$T - AN_\lambda + N_\lambda \quad < \quad \sum_{a \in [A]} g_{a,T}^{-1}(H_{a\epsilon}).$$

Rearranging the terms, we get

$$\begin{aligned} T - AN_\lambda + N_\lambda \quad &< \quad \sum_{a \in [A]} g_{a,T}^{-1}(H_{a\epsilon}) \\ T - N_\lambda(A-1) \quad &< \quad \sum_{a \in [A]} g_{a,T}^{-1}(H_{a\epsilon}). \end{aligned}$$

which contradicts our definition of $g_{a,T}$ in the theorem statement. Therefore $R_{\Omega(x)}(x) \leq \epsilon$.

From the preceding argument we have that if $\sum_{a \in [A]} g_{a,T}^{-1}(H_{a\epsilon}) \leq T - N_\lambda(A-1)$, then for any $x \in \mathcal{X}$ generated from the filtration,

$$\mathbb{P}(R_{\Omega(x)} < \epsilon | x) \geq 1 - A(T - AN_\lambda)e^{-\beta^2} - A\delta.$$

In the above equation, $1 - A(T - AN_\lambda)e^{-\beta^2}$ is from the event $\mathcal{E}$ and $1 - A\delta$ is due to the fact that the monotonic upper bounds holds only with probability $1-\delta$ for each of the arms. Setting β such that $\sum_{a \in [A]} g_{a,T}^{-1}(H_{a\epsilon}) = T - N_\lambda(A-1)$ (See Lemma 5.10), we have for

$$\beta = \sqrt{\frac{\lambda_x(T - N_\lambda(A-1)) + 2A\lambda}{16 C_1^2 H_\epsilon}} - \frac{C_2}{C_1},$$

that

$$\mathbb{P}(R_{\Omega(x)} < \epsilon | x) \geq 1 - A(T - AN_\lambda)e^{-\beta^2} - A\delta,$$

for $C_1 = \rho\sqrt{2}$ and $C_2 = \rho\sqrt{\sum_{\tau=2}^{T} \ln(1 + \frac{1}{\lambda}\hat{\sigma}_{a,\tau-1}(x_\tau))} + \sqrt{\lambda}\|f_a\|_{\mathcal{H}}$.

Since C_2 depends on T, to complete the proof and validity of the bound, we will show that C_2 grows logarithmically in T. When assumption **A III** holds and $\|\phi(x)\| \leq 1$, similar to the analysis in Abbasi-Yadkori et al. (2011); Durand et al. (2018), we have

$$\begin{aligned} C_2 &= \rho\sqrt{\sum_{\tau=2}^{T} \ln(1 + \frac{1}{\lambda}\hat{\sigma}_{a,\tau-1}(x_\tau))} + \sqrt{\lambda}\|f_a\|_{\mathcal{H}} \\ &= \rho\sqrt{\sum_{\tau=2}^{T} \ln(1 + \frac{1}{\lambda}\phi(x_\tau)^T(I + \frac{1}{\lambda}K_{a,\tau-1})^{-1}\phi(x_\tau))} \\ &+ \sqrt{\lambda}\|f_a\|_{\mathcal{H}} \\ &= \rho\sqrt{\ln(\det(I + \frac{1}{\lambda}K_{a,T}))} + \sqrt{\lambda}\|f_a\|_{\mathcal{H}} \\ &\leq \rho\sqrt{d^* \ln\left(\frac{1}{d^*}\left(1 + \frac{T}{\lambda}\right)\right)} + \sqrt{\lambda}\|f_a\|_{\mathcal{H}}. \end{aligned}$$

Since C_2 depends on $\sqrt{\ln(T)}$, we fix $C_2 = O(\rho\sqrt{\ln(T)})$. As $T \to \infty$ the RHS of the probability bound goes to unity and we have the resulting theorem.

□

5.6. Lemmas over event $\mathcal{E}$

For arm a at time t, we define event $\mathcal{E}_{a,t}$ as

$$\mathcal{E}_{a,t}(x) = \{L_{a,t}(x_t) \leq f_a(x) \leq U_{a,t}(x) | x_1, x_2, \cdots, x_{t-1}\}.$$

We define event $\mathcal{E}$ as $\mathcal{E} := \bigcup_{a \leq A, AN_\lambda < t \leq T} \mathcal{E}_{a,t}(x)$

The following theorems operate under the assumption the event $\mathcal{E}$ holds. We provide two properties of the terms in the algorithm that will be of help in the proofs:

- $B_{J_t(x)} = U_{j_t(x),t}(x) - L_{J_t(x),t}(x)$
- $U_{a,t}(x) = L_{a,t}(x) + s_{a,t}(x)$

Lemma 5.11. *Over event $\mathcal{E}$, for any sub-optimal arm $a(x) \neq a^*(x)$ at any time $t \leq T$, the simple regret of pulling that arm is upper bounded by the $B_{a,t}(x)$,*

Proof.

$$\begin{aligned} B_{a,t}(x) &= \max_{i \neq a} U_{i,t}(x) - L_{a,t}(x) \\ &\geq \max_{i \neq a} f_i(x) - f_{a,t}(x) = f^*(x) - f_a(x) = R_a(x). \end{aligned}$$

The first inequality holds due to the definition of event $\mathcal{E}$ and the equality holds since we are only considering sub-optimal arms. □

Note that the preceding lemma need not hold for the optimal arm, for which $R_a(x) = 0$ and it is not necessary that $B_{a,t}(x) \geq 0$.

Lemma 5.12. *Consider the contextual bandit setting proposed in the main paper. Over event $\mathcal{E}$, for any time t and context $x \in \mathcal{X}$, the following statements hold for the arm $a = a_t$ to be selected:*

$$\text{if } a = j_t(x), \text{ then } L_{j_t(x),t}(x) \leq L_{J_t(x),t}(x),$$
$$\text{if } a = J_t(x), \text{ then } U_{j_t(x),t}(x) \leq U_{J_t(x),t}(x).$$

Proof. We consider two cases based on which of the two candidate arms $j_t(x), J_t(x)$ is selected.

Case 1: $a = j_t(x)$ is selected. The proof works by contradiction. Assume that $L_{j_t(x),t}(x) > L_{J_t(x),t}(x)$. From the arm selection rule we have $s_{j_t(x),t}(x) \geq s_{J_t(x),t}(x)$. Based on this we can deduce that $U_{j_t(x),t}(x) \geq U_{J_t(x),t}(x)$. As a result,

$$\begin{aligned} B_{j_t(x),t}(x) &= \max_{i \neq j_t(x)} U_{i,t}(x) - L_{j_t(x),t}(x) \\ &< \max_{i \neq J_t(x)} U_{i,t}(x) - L_{J_t(x),t}(x) = B_{J_t(x),t}(x). \end{aligned}$$

The above inequality holds because the arm $j_t(x)$ must necessarily have the highest upper bound over all the arms. However, this contradicts the definition of $B_{J_t(x),t}(x)$ and as a result it must hold that $L_{j_t(x),t}(x) \leq L_{J_t(x),t}(x)$.

Case 2: $a = J_t(x)$ is selected. The proof works by contradiction. Assume that $U_{j_t(x),t}(x) > U_{J_t(x),t}(x)$. From the arm selection rule we have $s_{J_t(x),t}(x) \geq s_{j_t(x),t}(x)$. Based on this we can deduce that $L_{J_t(x),t}(x) \leq L_{j_t(x),t}(x)$. As a result, similar to Case 1,

$$\begin{aligned} B_{j_t(x),t}(x) &= \max_{j \neq j_t(x)} U_{j,t}(x) - L_{j_t(x),t}(x) \\ &< \max_{j \neq J_t(x)} U_{j,t}(x) - L_{J_t(x),t}(x) = B_{J_t(x),t}(x). \end{aligned}$$

The above inequality holds because the arm $j_t(x)$ must necessarily be have the highest upper bound over all the arms. However, this contradicts the definition of $B_{J_t(x),t}(x)$ and as a result it must hold that $U_{j_t(x),t}(x) \leq U_{J_t(x),t}(x)$. □

Corollary 5.13. *For context x, if arm $a = a_t(x)$ is pulled at time t, then $B_{J_t(x),t}(x)$ is bounded above by the uncertainty of arm a, i.e.,*

$$B_{J_t(x),t}(x) \leq s_{a,t}(x).$$

Proof. By construction of the algorithm $a \in \{j_t(x), J_t(x)\}$. If $a = j_t(x)$, then using the definition of $B_{J_t(x),t}(x)$ and Lemma 5.12, we can write

$$\begin{aligned} B_{J_t(x),t}(x) &= U_{j_t(x),t}(x) - L_{J_t(x),t}(x) \\ &\leq U_{j_t(x),t}(x) - L_{j_t(x),t}(x) = s_{a,t}(x). \end{aligned}$$

Similarly, for $a = J_t(x)$,

$$\begin{aligned} B_{J_t(x),t}(x) &= U_{j_t(x),t}(x) - L_{J_t(x),t}(x) \\ &\leq U_{J(x),t}(x) - L_{J_t(x),t}(x) = s_{a,t}(x). \end{aligned}$$

□

Lemma 5.14. *On event $\mathcal{E}$, for any time $t \leq T$ and for arm $a = a_t(x)$ the following bounds hold for the minimal gap*

$$B_{J_t(x),t}(x) \leq \min(0, s_{a,t}(x) - \Delta_a(x)) + s_{a,t}(x).$$

Proof. The arm to be pulled is restricted to $a \in \{j_t(x), J_t(x)\}$. The optimal arm for the context x at time t can either belong to $\{j_t(x), J_t(x)\}$ or be equal to some other arm. This results in 6 cases:

1. $a = j_t(x), a^* = j_t(x)$
2. $a = j_t(x), a^* = J_t(x)$
3. $a = j_t(x), a^* \notin \{j_t(x), J_t(x)\}$
4. $a = J_t(x), a^* = j_t(x)$
5. $a = J_t(x), a^* = J_t(x)$
6. $a = J_t(x), a^* \notin \{j_t(x), J_t(x)\}$

We define $f^*(x) := f_{a^*}(x)$ as the expected reward associated with the best arm and $f_{(a)}(x)$ as the expected reward of the a^{th} best arm.

Case 1: The following sequence of inequalities holds:

$$\begin{aligned} f_{(2)}(x) &\geq f_{J_t(x)}(x) \\ &\geq L_{J_t(x),t}(x) \\ &\geq L_{j_t(x),t}(x) \\ &\geq f_a(x) - s_{a,t}(x). \end{aligned}$$

The first inequality follows from the assumption that $a = a^* = j_t(x)$, the chosen and optimal arm has the highest upper confidence bound, and therefore, the expected reward of arm $J_t(x)$ can be at most that of the second best arm. The second inequality follows from event $\mathcal{E}$, the third inequality follows from 5.12. The last inequality follows from event $\mathcal{E}$. Using the above string of inequalities and the definition of $\Delta_a(x)$, we can write

$$s_{a,t} - (f_a(x) - f_{(2)}(x)) = s_{a,t} - \Delta_a(x) \geq 0.$$

The result holds for case 1 with the application of Corollary 5.13.

Case 2: $a = j_t(x), a^* = J_t(x)$. We can write

$$\begin{aligned} B_{J_t(x),t}(x) &= U_{j_t(x),t}(x) - L_{J_t(x),t}(x) \\ &\leq f_{j_t(x)}(x) + s_{j_t(x),t}(x) \\ &\quad - f_{J_t(x)}(x) + s_{J_t(x),t}(x) \\ &\leq f_a(x) - f^*(x) + 2s_{a,t}(x). \end{aligned}$$

The first inequality follows from event $\mathcal{E}$ and the second inequality holds because the selected arm has a larger uncertainty. From the definition of $\Delta_a(x)$,

$$\begin{aligned} B_{J_t(x),t}(x) &\leq 2s_{a,t}(x) - \Delta_a(x) \\ &\leq s_{a,t}(x) + \min(0, s_{a,t} - \Delta_a(x)). \end{aligned}$$

Where the inequality follows from Corollary 5.13.

Case 3: $a = j_t(x), a^* \notin \{j_t(x), J_t(x)\}$. We can write the following sequence of inequalities

$$f_{j_t(x)}(x) + s_{j_t(x),t}(x) \geq U_{j_t(x),t}(x) \geq U_{a^*} \geq f^*.$$

The first and third inequalities hold due to event $\mathcal{E}$, the second inequality holds by definition as $j_t(x)$ has the highest upper bound on any arm other than $J_t(x)$ neither of which is the optimal arm in this case. From the first and last inequalities, we obtain

$$s_{a,t}(x) - (f^* - f_{a,t}(x)) \geq 0,$$

or $s_{a,t}(x) - \Delta_a(x) \geq 0$. The result follows from Corollary 5.13.

Case 4: $a = J_t(x), a^* = j_t(x)$. We can write

$$\begin{aligned} B_{J_t(x),t}(x) &= U_{j_t(x),t}(x) - L_{J_t(x),t}(x) \\ &\leq f_{j_t(x)}(x) + s_{j_t(x),t}(x) \\ &\quad - f_{J_t(x)}(x) + s_{J_t(x),t}(x) \\ &\leq f_a(x) - f^*(x) + 2s_{a,t}(x). \end{aligned}$$

The first inequality follows from event $\mathcal{E}$ and the second inequality holds because the selected arm has a larger uncertainty. From the definition of $\Delta_a(x)$,

$$\begin{aligned} B_{J_t(x),t}(x) &\leq 2s_{a,t}(x) - \Delta_a(x) \\ &\leq s_{a,t}(x) + \min(0, s_{a,t} - \Delta_a(x)). \end{aligned}$$

Where the inequality follows from Corollary 5.13.

Case 5: $a = J_t(x), a^* = J_t(x)$. The following sequence of inequalities holds:

$$\begin{aligned} f_a(x) + s_{a,t}(x) &\geq U_{J_t(x),t}(x) \\ &\geq U_{j_t(x),t}(x) \\ &\geq f_{j_t(x)}(x) \\ &\geq f_{(2)}(x). \end{aligned}$$

The first and third inequalities follow from event $\mathcal{E}$, the second inequality is a consequence of Lemma 5.12, the fourth inequality follows from the fact that since $J_t(x)$ is the optimal arm, the upper bound and the arm selected should be as good as the second arm. Using the above chain of inequalities, we can write

$$s_{a,t}(x) - (f_{(2)}(x) - f_a(x)) = s_{a,t}(x) - \Delta_a(x) \geq 0.$$

Case 6: $a = J_t(x), a^* \notin \{j_t(x), J_t(x)\}$. We can write the following sequence of inequalities

$$f_{J_t(x)}(x) + s_{J_t(x),t}(x) \geq U_{J_t(x),t}(x) \geq U_{a^*,t}(x) \geq f^*.$$

The first and third inequalities hold due to event $\mathcal{E}$, the second inequality holds by definition as $J_t(x)$ has the highest upper bound on any arm when $a = J_t(x)$ due to Lemma 5.12 and $J_t(x)$ is not optimal in this case. From the first and last inequalities, we obtain

$$s_{a,t}(x) - (f^* - f_{a,t}(x)) \geq 0,$$

or $s_{a,t}(x) - \Delta_a(x) \geq 0$. The result follows from Corollary 5.13.

□

6. Experimental Details, Discussion and Additional Experimental Results

The algorithm was implemented with the best arm chosen with a history of one i.e., $\Omega(x_t) = J_T(x_t)$. For speed and scalability in implementation, the kernel inverse for arm a, $(K_{a,t} + \lambda I_{a,t})^{-1}$ and the kernel vector $k_{a,t}(x)$ updates were implemented as rank one updates. To tune kernel bandwidth and regularization parameters, we use following procedure: The dataset was split into two parts for hold-out (HO) and evaluation (EV). Each part was further split into two phases: exploration and exploitation. The value of T selected in both the hold-out and evaluation datasets were of similar magnitude. On the hold-out dataset, a grid search was used to set the tuning parameters by optimizing the average simple regret of the exploitation phase. The tuned parameters were used with the evaluation datasets to generate the plots. The code is available online to reproduce all results [1]. As our implementation performs rank one updates of the kernel matrix and its inverse, our algorithm has $O(T^2)$ as both computational and memory complexity in the worst case scenario, where T is the length of the exploration phase.

6.1. Synthetic Dataset

We present results of contextual simple regret minimization for a synthetic dataset. At every time step, we observe a one dimensional feature vector $x_t \sim \mathcal{U}[0, 2\pi]$, where $\mathcal{U}$ is a uniform distribution. There are 20 arms and reward for each arm a is $r_{a,t} := \sin(a * x_t)$, where $a = [1, 2..., 20]$. The arm with the highest reward at time t is the best arm. At every time step, we only observe the reward for the arm that the algorithm selects.

Since the dataset is i.i.d in nature, multiple simple regret evaluations are performed by shuffling the evaluation

[1] The code to reproduce our results is available at https://github.com/aniketde/ContextualGap

dataset, and the average curves are reported. Note that the algorithms have been cross validated for simple regret minimization. The plots are generated by varying the length of the exploration phase and keeping the exploitation dataset constant for evaluation of simple regret. It can be seen that the simple regret of the Contextual-Gap converges faster than the simple regret of other baselines.

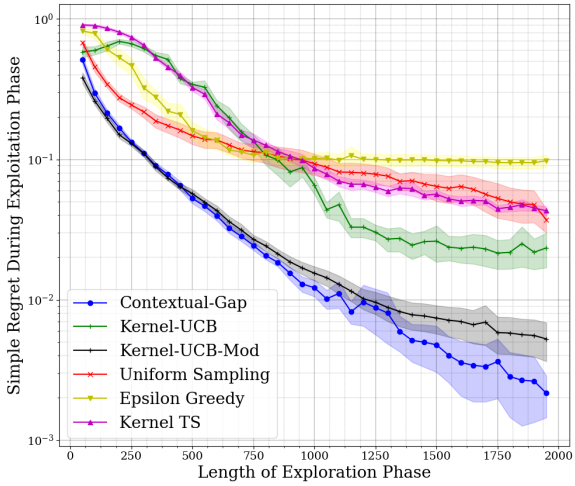
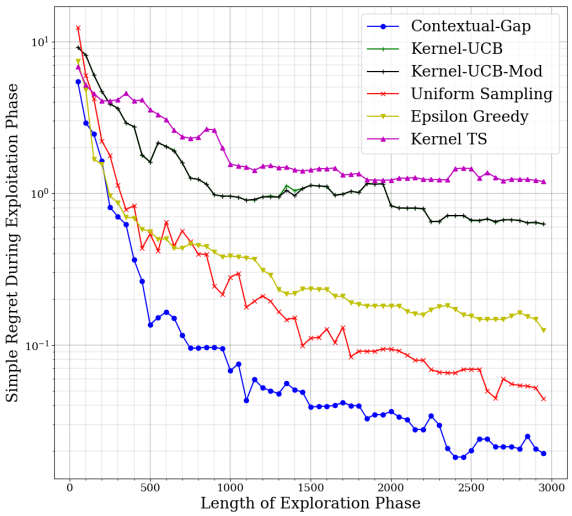
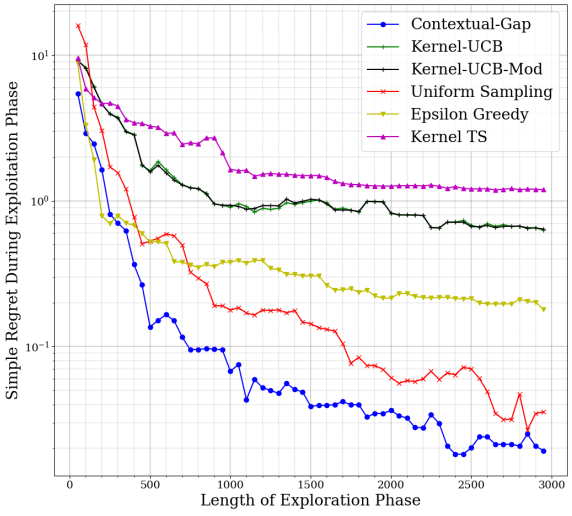


Figure 3: Average Simple Regret Evaluation on Synthetic Dataset



(a) Spacecraft dataset

Figure 4: Simple Regret evaluation with $\alpha = 0.1$



(a) Spacecraft dataset

Figure 5: Simple Regret evaluation with $\alpha = 0.5$

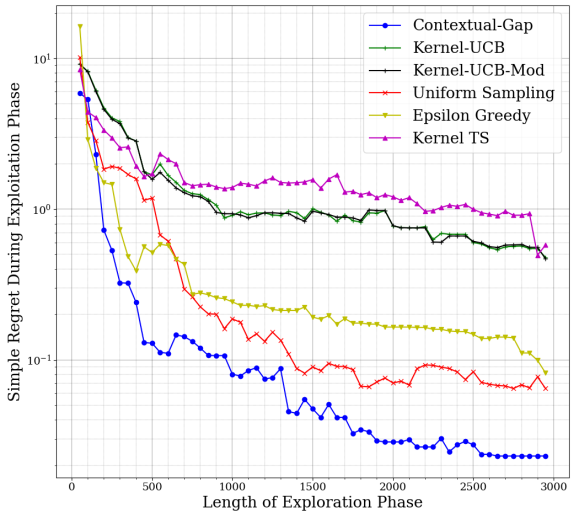
6.2. Experimental Spacecraft Magnetic Field Dataset

6.2.1. Results for different α

The following plots (Figures 4, 5 and 6) provide results for different α values. Note that, the hyper parameter computed with cross validation for $\alpha = 1$ were retained for the evaluation runs of different values of α. It can be seen that Contextual Gap performs consistently better for all the datasets under consideration.

6.2.2. History during exploitation

A comparison of the average simple regret variation of Contextual gap with a history of the past 25 data points (instead of 1) is shown in Figure 7. It can be seen that there exists only minor differences in contextual gap runs with history.



(a) Spacecraft dataset

Figure 6: Simple Regret evaluation with $\alpha = 2$

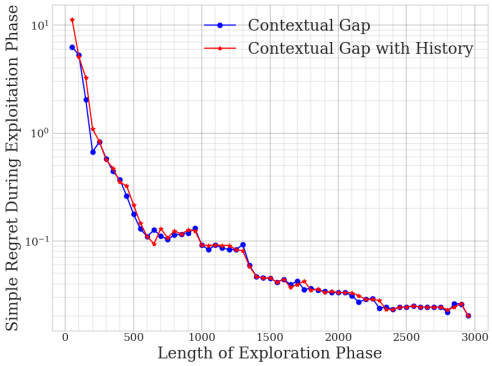


Figure 7: Comparison with Contextual Gap algorithm with recent history for Spacecraft dataset

6.2.3. Worst Case Simple Regret

The contextual gap algorithm presented is a solution to simple regret minimization, and not average simple regret minimization. Hence,we present the worst case simple regret among all the data present in the exploitation phase as additional empirical evidence of simple regret minimization (Figure 8).

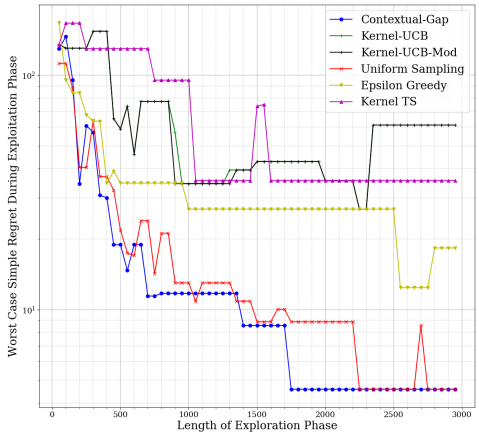


Figure 8: Worst Case Simple Regret Evaluation on Spacecraft Magnetic Field Dataset

6.2.4. Histogram of Arm Selection

Histogram of number times the best, second best and third best (worst) arms are selected during exploration is shown in Figure 9. As expected, algorithms designed to minimize cumulative regret focus on the best arm more and Contextual-Gap explores best and second best arms.

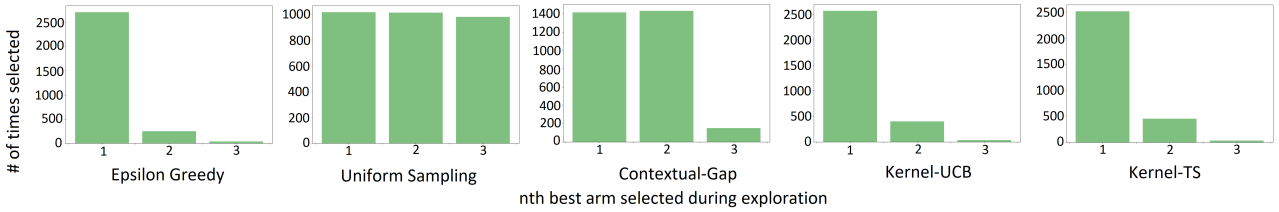


Figure 9: Histogram of arm selection during exploration

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Bercovici, H., Li, W., and Timotin, D. The horn conjecture for sums of compact self-adjoint operators. *American Journal of Mathematics*, 131(6):1543–1567, 2009.

Durand, A., Maillard, O.-A., and Pineau, J. Streaming kernel regression with provably adaptive mean, variance, and regularization. *Journal of Machine Learning Research*, 19(August), 2018.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pp. 3212–3220, 2012.

Gentile, C., Li, S., and Zappella, G. Online clustering of bandits. In *International Conference on Machine Learning*, pp. 757–765, 2014.

Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., and Etrue, E. On context-dependent clustering of bandits. In *International Conference on Machine Learning*, pp. 1253–1262, 2017.

Guan, M. Y. and Jiang, H. Nonparametric stochastic contextual bandits. In *The 32nd AAAI Conference on Artificial Intelligence*, 2018.

Hoffman, M., Shahriari, B., and Freitas, N. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*, pp. 365–374, 2014.

Korda, N., Szörényi, B., and Shuai, L. Distributed clustering of linear bandits in peer to peer networks. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, volume 48, pp. 1301–1309. International Machine Learning Societ, 2016.

Li, S. and Zhang, S. Online clustering of contextual cascading bandits. In *The 32nd AAAI Conference on Artificial Intelligence*, 2018.

Micchelli, C. A., Xu, Y., and Zhang, H. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

Minsker, S. On some extensions of Bernsteins inequality for self-adjoint operators. *Statistics & Probability Letters*, 127(C):111–119, 2017.

Motwani, R. and Raghavan, P. *Tail Inequalities*, pp. 67100. Cambridge University Press, 1995. doi: 10.1017/CBO9780511814075.005.

Steele, J. M. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

Tu, S. and Recht, B. Least-squares temporal difference learning for the linear quadratic regulator. *arXiv preprint arXiv:1712.08642*, 2017.

Zi-Zong, Y. Schur complements and determinant inequalities. 2009.