An Optimization-based Algorithm for Non-stationary Kernel Bandits without Prior Knowledge

Kihyuk Hong
University of Michigan

Yuhang Li
University of Michigan

Ambuj Tewari
University of Michigan

Abstract

We propose an algorithm for non-stationary kernel bandits that does not require prior knowledge of the degree of non-stationarity. The algorithm follows randomized strategies obtained by solving optimization problems that balance exploration and exploitation. It adapts to non-stationarity by restarting when a change in the reward function is detected. Our algorithm enjoys a tighter dynamic regret bound than previous work on non-stationary kernel bandits. Moreover, when applied to non-stationary linear bandits by using a linear kernel, our algorithm is nearly minimax optimal, solving an open problem in the non-stationary linear bandit literature. We extend our algorithm to use a neural network for dynamically adapting the feature mapping to observed data. We prove a dynamic regret bound of the extension using the neural tangent kernel theory. We demonstrate empirically that our algorithm and the extension can adapt to varying degrees of non-stationarity.

1 INTRODUCTION

The linear bandit (LB) problem (Dani et al. 2008) and the kernel bandit (KB) problem (Srinivas et al. 2010) are important paradigms for sequential decision making under uncertainty. They extend the multi-armed bandit (MAB) problem (Robbins 1952) by modeling the reward function with the side information of each arm provided as a feature vector. LB assumes the reward function is linear. KB extends LB to model non-linearity by assuming the reward function lies in the RKHS induced by a kernel.

A recent line of work studies the non-stationary variants of LB and KB where the reward functions can vary over time subject to two main types of non-stationarity budgets: the number of changes and the total variation in the sequence of reward functions. A common algorithm design principle for adapting to non-stationarity is the principle of forgetting the past. It has been applied to the non-stationary MAB to design nearly minimax optimal algorithms (Garivier et al. 2011; Besbes et al. 2014). Similarly, the principle has been applied to the non-stationary LB (Cheung et al. 2019; Russac et al. 2019; Zhao et al. 2020; Kim et al. 2020) and the non-stationary KB (Zhou et al. 2021; Deng et al. 2022).

Recently, Zhao et al. (2021) found an error in a key technical lemma by Cheung et al. (2019) that affects the concentration bound of regression-based reward estimates under non-stationarity. Unfortunately, the error is inherited by Russac et al. (2019), Zhao et al. (2020) and Kim et al. (2020). The corrected regret bounds of the affected papers are worse than what were originally reported. Since the correction, finding a nearly minimax optimal algorithm for the non-stationary LB setting has been an open problem. The same error affected the work on non-stationary KB by Zhou et al. (2021) and they had to correct their initially reported regret bound to a worse one.

Algorithms using the principle of forgetting require the knowledge of the non-stationarity budgets. For example, sliding window algorithms (Garivier et al. 2011; Cheung et al. 2019; Zhou et al. 2021) that forget the past by discarding data older than certain time window require the knowledge of the non-stationarity budgets to optimally tune the size of the window. Since having a prior knowledge of the non-stationarity budgets may not be realistic in practical settings, researchers have developed change detection based algorithms that do not require the knowledge of non-stationarity budgets. A seminal paper by Auer et al. (2018) demonstrates a change detection based algorithm for the non-stationary two-armed bandit setting. Their design principle has been applied to MAB (Auer et al. 2019) and the contextual bandit setting (Chen et al. 2019). More recently, Wei et al. (2021) proposed a reduction called MASTER that equips an algorithm designed for a stationary environment with change detection subroutines to adapt to non-stationarity without the knowledge of non-stationarity budgets. They provided a reduction of the OFUL algorithm (Abbasi-yadkori et al. 2011)

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

and claimed near-minimax optimality for non-stationary linear bandits. However, due to the aforementioned error, they had to correct their regret bound to a suboptimal one.

In this paper, we design an algorithm that sidesteps the error and recover the tighter dynamic regret bounds for non-stationary LB and KB that were once thought to be achieved. We make the following contributions.

- We design a novel optimization-based algorithm OPKB for stationary kernel bandits that uses inverse propensity score based reward estimates that sidestep the aforementioned error specific to regression based reward estimates.
- We design an algorithm ADA-OPKB that adapts OPKB to non-stationary settings using change detection. ADA-OPKB does not require the knowledge of the non-stationarity budgets and enjoys a dynamic regret bound tighter than previous work on non-stationary KB.
- We show ADA-OPKB is nearly minimax optimal in the non-stationary linear bandit setting, solving an open problem in the non-stationary linear bandit literature.
- We provide an extension of ADA-OPKB called ADA-OPNN that trains a neural network to dynamically adapt the feature mapping to observed data. We show a dynamic regret bound for ADA-OPNN when the width of the network is sufficiently large using the neural tangent kernel theory (Jacot et al. 2018).

1.1 Related Work

Non-stationary Linear/Kernel Bandits Common approaches for non-stationary bandits include restarting periodically, using recent data within fixed time window (sliding-window) and exponentially decaying past observations (discounting). These approaches require the knowledge of non-stationarity. Zhou et al. (2021) analyze restarting and sliding-window approaches for adapting a UCB-based algorithm for kernel bandits. Deng et al. (2022) analyze a discounting approach for kernel bandits. Russac et al. (2019), Cheung et al. (2019) and Zhao et al. (2020) propose discounting, sliding-window and restarting approaches for adapting a UCB-based algorithm for linear bandits respectively. Cheung et al. (2022) discuss restarting adversarial linear bandit algorithm. For the non-stationary setting where the learner does not have the knowledge of the non-stationarity, Cheung et al. (2019), Zhao et al. (2020) and Cheung et al. (2022) discuss bandit-over-bandit (BOB) reduction. Wei et al. (2021) propose a change detection based reduction (MASTER) and show a reduction of a UCB-based algorithm for linear bandits.

Optimization-based Algorithms First proposed for contextual bandits optimization-based algorithms solve optimization problems to find randomized strategies that balance exploration and exploitation (Dudik et al. 2011; Agarwal et al. 2014). The idea is adapted to linear bandits (Lattimore et al. 2017; Hao et al. 2020; Lee et al. 2021). Our paper is the first to apply the approach to kernel bandits.

2 PROBLEM STATEMENT

We consider a bandit problem where the learner and the nature interact sequentially for T time steps. At each time t, the learner plays an action x_t chosen from a finite set of actions $\mathcal{X} = \{a_1, \dots, a_N\} \subset \mathbb{R}^d$. Then the nature reveals a noisy reward $y_t = r_t(x_t) + \eta_t$ where $r_t : \mathcal{X} \to \mathbb{R}$ is an unknown reward function at time t and $\{\eta_t\}_{t=1}^T$ are independent zero-mean noises with a bound $|\eta_t| \le S$. [1]

Following the kernel bandit setting commonly used in the literature, we make the following regularity assumption on the reward functions.

Assumption A (Kernel bandit)**.** *The reward functions r_t live in the RKHS $\mathcal{H}$ induced by a continuous positive semi-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with $k(x, x) \le 1$ for all $x \in \mathcal{X}$. Their norms satisfy $\|r_t\|_{\mathcal{H}} \le B$ for all $t = 1, \dots, T$. The kernel k and the bounds S, B are known to the learner.*

Note that Assumption A implies $|r_t(x)| = \langle r_t, k(\cdot, x)\rangle_{\mathcal{H}} \le \|r_t\|_{\mathcal{H}}\|k(\cdot, x)\|_{\mathcal{H}} \le B$ for all $t = 1, \dots, T$ and $x \in \mathcal{X}$ by the reproducing property of RKHS and Cauchy-Schwarz. For the rest of the paper, when making Assumption A, we assume that the learner scales the problem (by $S + B$) so that $|r_t(x)| \le 1$ and $|y_t(x)| \le 1$ for simpler exposition.

Before the learner interacts with the nature, the nature chooses a sequence of reward functions $\{r_t\}_{t=1}^T$ subject to two types of non-stationarity budgets simultaneously. The first budget V_T limits the total variation of the sequence of reward functions: $\sum_{t=1}^{T-1} \|r_{t+1} - r_t\|_\infty \le V_T$. The second budget L_T limits the number of changes in the sequence of reward functions: $1 + \sum_{t=1}^{T-1} \mathbb{I}\{r_{t+1} \ne r_t\} \le L_T$.

The learner aims to minimize the *dynamic regret* $\textsc{Reg}_T := \sum_{t=1}^T (r_t(x_t^\star) - r_t(x_t))$ where $x_t^\star := \operatorname{argmax}_{x \in \mathcal{X}} r_t(x)$ is the best action at time t. Note that $\textsc{Reg}_T$ is the cumulative expected regret against the optimal strategy with full knowledge of the sequence of reward functions.

2.1 Preliminaries and Notations

Feature Mapping By Mercer's theorem, given a continuous positive semi-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a feature mapping $\psi : \mathcal{X} \to \ell^2$ with $k(x, x') = \langle \psi(x), \psi(x')\rangle$ for all $x, x' \in \mathcal{X}$. We say feature mappings φ_1 and φ_2 are *equivalent* if $\langle \varphi_1(x), \varphi_1(x')\rangle =$

[1] The boundedness noise assumption is for making use of the Freedman-style inequality (Lemma D.2). We can relax this assumption to a subgaussian noise assumption by modifying the Freedman-style inequality using a truncation argument. See Appendix D.7 for detail.

Table 1: Regret Bound Comparison of Algorithms for Non-stationary Kernel/Linear Bandits

Setting	Algorithm	Regret bound in $\widetilde{\mathcal{O}}(\cdot)$	Required knowledge
Kernel Bandit	R/SW-GPUCB (Zhou et al. 2021)	$\gamma_T^{\frac{7}{8}} T^{\frac{3}{4}}(1+V_T)^{\frac{1}{4}}$	V_T
	WGPUCB (Deng et al. 2022)	$\dot{\gamma}_T^{\frac{7}{8}} T^{\frac{3}{4}}(1+V_T)^{\frac{1}{4}}$ [2]	V_T
	GPUCB+MASTER (Appendix E)	$\min\{\gamma_T\sqrt{TL_T}, \gamma_T T^{\frac{2}{3}} V_T^{\frac{1}{3}} + \gamma_T\sqrt{T}\}$	
	ADA-OPKB (Ours)	$\min\{\sqrt{d\gamma_T TL_T}, d^{\frac{1}{3}}\gamma_T^{\frac{1}{3}} T^{\frac{2}{3}} V_T^{\frac{1}{3}} + \sqrt{d\gamma_T T}\}$	
Linear Bandit	D-LinUCB (Russac et al. 2019)	$d^{\frac{7}{8}} T^{\frac{3}{4}} V_T^{\frac{1}{4}} + d\sqrt{T}$	V_T
	SW-UCB+BOB (Cheung et al. 2019)	$d^{\frac{7}{8}} T^{\frac{3}{4}} V_T^{\frac{1}{4}} + d\sqrt{T}$	
	RestartUCB+BOB (Zhao et al. 2020)	$d^{\frac{7}{8}} T^{\frac{3}{4}} V_T^{\frac{1}{4}} + d\sqrt{T}$	
	Restart-Adv (Cheung et al. 2022)	$d^{\frac{2}{3}} T^{\frac{2}{3}} V_T^{\frac{1}{3}} + d\sqrt{T}$	V_T
	Restart-Adv+BOB (Cheung et al. 2022)	$d^{\frac{2}{3}} T^{\frac{2}{3}} V_T^{\frac{1}{3}} + d^{\frac{1}{2}} T^{\frac{3}{4}}$	
	LinUCB+MASTER (Wei et al. 2021)	$\min\{d\sqrt{TL_T}, dT^{\frac{2}{3}} V_T^{\frac{1}{3}} + d\sqrt{T}\}$	
	ADA-OPKB (Ours)	$\min\{d\sqrt{TL_T}, d^{\frac{2}{3}} T^{\frac{2}{3}} V_T^{\frac{1}{3}} + d\sqrt{T}\}$	

$\langle \varphi_2(x), \varphi_2(x') \rangle$ for all $x, x' \in \mathcal{X}$. Given a feature mapping ψ, we can always find an equivalent N-dimensional feature mapping $\varphi : \mathcal{X} \to \mathbb{R}^N$. For example, we can decompose the kernel matrix $K = \{\langle \psi(a_i), \psi(a_j) \rangle\}_{i,j \in [N]}$ into $K = \Phi\Phi^T$ using the Cholesky decomposition where $\Phi \in \mathbb{R}^{N \times N}$, then take $\varphi(a_i) = \Phi^T e_i$ for all $i = 1, \ldots, N$.

Maximum Information Gain The maximum information gain (Srinivas et al. 2010) of the RKHS induced by a kernel k is defined as the maximum mutual information between observations $\{f(x_t) + \epsilon_t\}_{t=1}^T$ with $\epsilon_t \sim N(0,1)$ and $f(\cdot)$ sampled from a Gaussian process $GP(0, \sigma^{-1}k(\cdot,\cdot))$. It is a widely used dimensionality measure of RKHS. As done by Camilleri et al. (2021), we generalize the original definition to support T *fractional* observations, and define $\gamma_{\varphi,T} = \max_{P \in \mathcal{P}_\mathcal{X}} \log\det S_\varphi(TP/\sigma, 1)$ where $\mathcal{P}_\mathcal{A}$ is the set of probability distributions on $\mathcal{A}$, $S_\varphi(Q, \lambda) := \sum_{x \in \mathcal{X}} Q(x)\varphi(x)\varphi(x)^T + \lambda I$ and φ is an N-dimensional feature mapping of k. It can be shown that for equivalent feature mappings φ_1 and φ_2 of k, we have $\gamma_{\varphi_1,T} = \gamma_{\varphi_2,T}$ (see Appendix I). Hence, $\gamma_{\varphi,T}$ is fully determined by the underlying kernel k and does not depend on the particular choice of the feature mapping φ induced by the kernel. We suppress the subscript φ and write $S(\cdot,\cdot)$ and γ_T when clear from the context. For the connection between the original definition of the maximum information gain and our definition, see Appendix C.

Other Notations We use $[n]$ to denote $\{1, \ldots, n\}$. For a semi-positive definite matrix M and a vector x, we write $\|x\|_M^2 = x^T M x$. We denote by $\mathbb{E}_t[\cdot]$ and $\mathrm{Var}_t[\cdot]$ the conditional expectation and variance respectively given history up to time $t-1$. For an interval $\mathcal{I} = [s,t]$, we define $V_\mathcal{I} = \sum_{\tau=s}^{t-1} \|r_{\tau+1} - r_\tau\|_\infty$ and $L_\mathcal{I} = 1 + \sum_{\tau=s}^{t-1} \mathbb{I}\{r_{\tau+1} \neq r_\tau\}$.

3 MAIN RESULT

The main result of this paper provides a worst-case bound on the dynamic regret of our novel algorithm called ADA-OPKB for the non-stationary kernel bandit setting.

Theorem 3.1. *Under Assumption A, without the knowledge of non-stationarity budgets V_T and L_T, the dynamic regret of ADA-OPKB is bounded, with high probability, by*

$$\widetilde{\mathcal{O}}\Big(\min\Big\{\sqrt{\gamma_T L_T T \log N}, \\ (\gamma_T V_T \log N)^{1/3} T^{2/3} + \sqrt{\gamma_T T \log N}\Big\}\Big).$$

When the action set $\mathcal{X} \subset \mathbb{R}^d$ is an infinite bounded set, we can take a hypercube of side length R that contains $\mathcal{X}$ and discretize it into $\mathcal{O}((Rd/\epsilon)^d)$ hypercubes as done by Chowdhury et al. (2017) where ϵ is the maximum error of expected reward from discretization. Discretizing the action set with $\epsilon = 1/T$ and running ADA-OPKB on the discretized action set lead to a dynamic regret bound of $\widetilde{\mathcal{O}}(\min\{\sqrt{d\gamma_T L_T T}, (d\gamma_T V_T)^{1/3} T^{2/3} + \sqrt{d\gamma_T T}\})$. We use this bound to compare with previous work on the setting with an infinite action set.

We can reduce the kernel bandit setting to the linear bandit setting by using the linear kernel $k(x,x') = \langle x, x' \rangle$. As shown in Lemma C.3, the maximum information gain of the linear space is $\gamma_T = \mathcal{O}(d \log T)$ and the dynamic regret bound of ADA-OPKB that uses the linear kernel becomes $\widetilde{\mathcal{O}}(\min\{\sqrt{dL_T T \log N}, (dV_T \log N)^{1/3} T^{2/3} + \sqrt{dL_T T \log N}\})$ for the finite action set. For the infinite action set, we get $\widetilde{\mathcal{O}}(\min\{d\sqrt{L_T T}, d^{2/3} V_T^{1/3} T^{2/3} + d\sqrt{T}\})$ using the discretization technique.

[2] The dimensionality measure $\dot{\gamma}_T$ used in Deng et al. (2022) is

Relation to Previous Work Table 1 compares the regret bound of our work to the *corrected* regret bounds of previous works. The regret bound of ADA-OPKB for non-stationary kernel bandits is tighter than previous work. Applying to non-stationary linear bandits by using the linear kernel, ADA-OPKB nearly achieves the lower bound $\Omega(d^{\frac{2}{3}}V_T^{\frac{1}{3}}T^{\frac{2}{3}})$ (Cheung et al. 2019), solving an open problem of finding a nearly minimax optimal algorithm for non-stationary linear bandits. The best regret bound before our work is by Cheung et al. (2022) who discuss that an algorithm for adversarial linear bandits, e.g. Exp3 algorithm (Lattimore et al. 2020), equipped with periodic restarts (Restart-Adv) achieves $\widetilde{\mathcal{O}}(d^{\frac{2}{3}}T^{\frac{2}{3}}V_T^{\frac{1}{3}})$. However, it requires the knowledge of V_T to tune the frequency of restarts. They also discuss a bandit-over-bandit reduction of Restart-Adv (Restart-Adv+BOB) that does not require the knowledge of V_T. However, the reduction suffers an additional regret term of $d^{\frac{1}{2}}T^{\frac{3}{4}}$.

The dependence of γ_T in the regret bound for kernel bandits is crucial since γ_T can grow with T. For example, γ_T for the Matérn kernel with smoothness parameter ν scales as $\widetilde{\Theta}(T^{\frac{d}{2\nu+d}})$ (Vakili et al. 2021b). Previous works on non-stationary kernel bandits (Zhou et al. 2021; Deng et al. 2022) show a regret bound of order $\gamma_T^{7/8}T^{3/4}$, which may not be sublinear in T. For example, it is not sublinear in T for Matérn kernel when $\nu/d \leq 5/4$. Our improved regret bound for ADA-OPKB is of order $\min\{\gamma_T^{1/3}T^{2/3}, \sqrt{\gamma_T T}\}$, which is sublinear in T as long as γ_T is sublinear in T. As shown by Vakili et al. (2021b), γ_T is sublinear for a class of kernels of which eigenvalues decay polynomially or exponentially, which includes the Matérn kernel and the squared exponential kernel.

4 ALGORITHMS AND ANALYSES

We first study stationary kernel bandits where the reward functions do not vary over time.

4.1 OPKB: Optimization-based Algorithm for Stationary Kernel Bandits

Central to the OPKB algorithm is the optimization problem (OP) designed to return a randomized strategy that balances exploration and exploitation. OP uses an empirical suboptimality gap of each action computed based on the inverse propensity score (IPS) estimator (Camilleri et al. 2021).

Definition 4.1. *The inverse propensity score (IPS) estimator for the expected reward $r_t(x)$ with respect to φ using the observed reward y_t is defined as*

$$\widehat{\mathcal{R}}_{\varphi,t}(x) \coloneqq \varphi(x)^T S_\varphi(P_t, \sigma/T)^{-1}\varphi(x_t)y_t$$

for all $x \in \mathcal{X}$ where P_t is the randomized strategy used at time t. Averaging over an interval $\mathcal{I}$, we define $\widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x) \coloneqq \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\widehat{\mathcal{R}}_{\varphi,t}(x)$. The empirical suboptimality gap of action x from observations in $\mathcal{I}$ is defined as $\widehat{\Delta}_{\varphi,\mathcal{I}}(x) \coloneqq \max_{x'\in\mathcal{X}}\widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x') - \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x)$.

OP minimizes over $P \in \mathcal{P}_\mathcal{X}$ the objective function

$$\sum_{x\in\mathcal{X}} P(x)\widehat{\Delta}(x) - \frac{2}{\beta}\log\det S_\varphi(P, \sigma/T) \tag{1}$$

where the first term is the weighted average of the empirical suboptimality gaps that encourages exploitation and the second term is a regularizer that encourages exploration. That the second term encourages exploration can be seen by the property of the optimal design defined as follows.

Definition 4.2. *Given a set of actions $\mathcal{A} \subseteq \mathcal{X}$ and a feature mapping $\varphi : \mathcal{X} \to \mathbb{R}^p$, we define $\pi_\varphi(\mathcal{A}) \coloneqq \operatorname{argmax}_{P\in\mathcal{P}_\mathcal{A}}\log\det S_\varphi(P, \sigma/T)$ and call it the optimal design on $\mathcal{A}$ with respect to φ.*

The optimal design is a generalization of the Bayesian D-optimal design for linear models that maximizes $\log\det(\sum_{x\in\mathcal{A}}P(x)xx^T + R)$, where R is some regularizer. The Bayesian D-optimal design is one of the exploration strategies used in the Bayesian experimental design literature (Chaloner et al. 1995). As shown in the following lemma, by playing our definition of the optimal design $\pi_\varphi(\mathcal{A})$, we can uniformly bound the variance of the IPS estimators over all actions in $\mathcal{A}$. See Appendix D.2 for proof.

Lemma 4.3. *Consider an optimal design $\pi_\varphi(\mathcal{A})$ with respect to a feature mapping φ on a set of actions $\mathcal{A} \subseteq \mathcal{X}$. If we play an action sampled from $\pi_\varphi(\mathcal{A})$ at time t and observe y_t, then for all $x \in \mathcal{X}$, we have*

$$\operatorname{Var}(\widehat{\mathcal{R}}_{\varphi,t}(x)) \leq \|\varphi(x)\|^2_{S_\varphi(\pi_\varphi(\mathcal{A}),\sigma/T)^{-1}} \leq \gamma_{\varphi,T}.$$

The full OP algorithm is presented below. Note that due to the concavity of $\log\det(\cdot)$, the optimization problem used by OP and the optimal design can be solved efficiently, for example, by using the interior-point method in Vandenberghe et al. (1998).

Algorithm 1: OP: Optimization Problem

Input: φ, $\widehat{\Delta} = \{\widehat{\Delta}(x)\}_{x\in\mathcal{X}}$, α, β, T

1 Find a minimizer $P^\star \in \mathcal{P}_\mathcal{X}$ of (1).

2 Find $\mathcal{A} \leftarrow \{x \in \mathcal{X} : \widehat{\Delta}(x) \leq 2\alpha\gamma_{\varphi,T}/\beta\}$.

Return: The mixed strategy $Q = \frac{1}{2}P^\star + \frac{1}{2}\pi_\varphi(\mathcal{A})$

The parameter β controls the balance between exploration and exploitation. As stated in Lemma 4.4, the greater the β, the smaller the expected empirical regret $\sum_{x\in\mathcal{X}}Q(x)\widehat{\Delta}(x)$ and the greater the variance bound. See Appendix D.3 for

related to γ_T but they use a discounted kernel matrix computed with an approximate feature mapping for computing $\dot{\gamma}$.

the proof. Note that OP mixes the minimizer $P^\star$ with the optimal design on the set $\mathcal{A}$ computed in Line 2. This step is required to get the bound (4), which is the key to bound the bias of the reward estimator for the regret analysis.

Lemma 4.4. *The distribution Q returned by the algorithm* $\mathrm{OP}(\varphi, \widehat{\Delta}, \alpha, \beta, T)$ *satisfies*

$$\sum_{x\in\mathcal{X}} Q(x)\widehat{\Delta}(x) \le \frac{(1+\alpha)\gamma_{\varphi,T}}{\beta}, \tag{2}$$

$$\|\varphi(x)\|^2_{S_\varphi(Q,\sigma/T)^{-1}} \le \beta\widehat{\Delta}(x) + 2\gamma_{\varphi,T},\ \forall x\in\mathcal{X}, \tag{3}$$

$$\|\varphi(x)\|^2_{S_\varphi(Q,\sigma/T)^{-1}} \le \frac{\beta^2\widehat{\Delta}(x)^2}{2\alpha\gamma_{\varphi,T}} + 2\gamma_{\varphi,T},\ \forall x\in\mathcal{X}. \tag{4}$$

Now, we present the OPKB algorithm (Algorithm 2). OPKB takes a feature mapping φ as an input. Assuming the knowledge of the kernel k corresponding to the RKHS in which the reward function lies, we use any feature mapping $\varphi : \mathcal{X} \to \mathbb{R}^N$ equivalent to the feature mapping $\psi : \mathcal{X} \to \ell^2$ corresponding the kernel. The choice of φ among the feature mappings equivalent to ψ does not affect the algorithm and the analysis. See Appendix I for details. OPKB runs in blocks of doubling sizes. In the first block, it follows the optimal design for E time steps. Before starting a new block j, it computes the empirical suboptimality gaps using all past history, then runs OP to find the strategy $Q^{(j)}$ and mixes it with the optimal design. The mixed strategy $P^{(j)}$ is run in block j. Every block, OPKB increases the parameter β by a factor of $\sqrt{2}$ when calling OP to increase the degree of exploitation.

Algorithm 2: OPKB

Input: feature map φ, horizon T, confidence $\delta \in (0,1)$.
Definition: $\mu_j = c_1 2^{-j/2}$, $E = \lceil c_3\gamma_{\varphi,T}\log(C_0N/\delta)\rceil$,
$\beta_j = c_2\gamma_{\varphi,T}2^{j/2}$, $\alpha = c_4\sigma/\log(C_0N/\delta)$
Initialize: $t \leftarrow 1$, $P^{(0)} \leftarrow \pi_\varphi(\mathcal{X})$
1 **for** $j = 0, 1, \dots$ **do**
2 Set block $\mathcal{B}(j) \leftarrow [t, t+2^jE-1]$ and cumulative block $\mathcal{C}(j) \leftarrow \cup_{k=0}^{j}\mathcal{B}(k)$.
3 **if** $j \ge 1$ **then**
4 Compute empirical gap $\widehat{\Delta} \leftarrow \{\widehat{\Delta}_{\varphi,\mathcal{C}(j-1)}(x)\}_{x\in\mathcal{X}}$.
5 Find strategy $Q^{(j)} \leftarrow \mathrm{OP}(\varphi, \widehat{\Delta}, \alpha, \beta_j, T)$.
6 Set $P^{(j)} \leftarrow (1-\mu_j)Q^{(j)} + \mu_j\pi_\varphi(\mathcal{X})$.
7 **while** $t \in \mathcal{B}(j)$ **do**
8 Play $x_t \sim P^{(j)}$; receive y_t; increment $t \leftarrow t+1$.

4.2 Analysis of OPKB

For the analysis of OPKB, we use the following concentration bounds for the reward estimate $\widehat{\mathcal{R}}_{\varphi,\mathcal{C}(m)}(x)$ and the gap estimate $\widehat{\Delta}_{\varphi,\mathcal{C}(m)}$ shown under a more general setting of non-stationary kernel bandits. The proof is based on a Freedman-style inequality on the martingale difference sequence $\{\widehat{\mathcal{R}}_{\varphi,t}(x) - \mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi,t}(x)]\}_{t\in\mathcal{C}(j)}$. See Appendix D.5 for the full proof.

Lemma 4.5. *With probability at least $1-\delta$, when running the OPKB algorithm, we have for all block indices $j = 0, 1, \dots$ and actions $x \in \mathcal{X}$ that*

$$|\widehat{\mathcal{R}}_{\varphi,\mathcal{C}(j)}(x) - \mathcal{R}_{\mathcal{C}(j)}(x)| \le \frac{1}{2}\Delta_{\mathcal{C}(j)}(x) + V_{\mathcal{C}(j)} + \frac{c_0\mu_j}{4} \tag{5}$$

$$\Delta_{\mathcal{C}(j)}(x) \le 2\widehat{\Delta}_{\varphi,\mathcal{C}(j)}(x) + 4V_{\mathcal{C}(j)} + c_0\mu_j \tag{6}$$

$$\widehat{\Delta}_{\varphi,\mathcal{C}(j)}(x) \le 2\Delta_{\mathcal{C}(j)}(x) + 4V_{\mathcal{C}(j)} + c_0\mu_j \tag{7}$$

where $\mathcal{C}(j)$ is the interval from time 1 to the end of block j, c_0 is a universal constant, $\mathcal{R}_\mathcal{I}(x) \coloneqq \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} r_t(x)$ is the average reward in $\mathcal{I}$ and $\Delta_\mathcal{I}(x) \coloneqq \max_{x'\in\mathcal{X}}\mathcal{R}_\mathcal{I}(x') - \mathcal{R}_\mathcal{I}(x)$.

Remark 1. Concentration bounds for regression-based reward estimates for the non-stationary LB and KB given by Lemma 2 in Zhao et al. (2021) and Lemma 1 in Zhou et al. (2021) are analogous to (5). However, their bounds have an additional factor of $\sqrt{d}$ and $\sqrt{\gamma_{\varphi,T}}$ respectively for the term $V_{\mathcal{C}(j)}$, leading to suboptimal regret bounds. Their concentration bounds were believed to have a constant factor for the term $V_{\mathcal{C}(j)}$, but they had to be corrected due to an error found by Zhao et al. (2021). The error is specific to regression-based reward estimates. See Zhao et al. (2021) for details. Our algorithm sidesteps the error by using IPS reward estimates instead of regression-based reward estimates. The main motivation for using randomized strategies in our algorithm is to use IPS reward estimates, which can only be constructed when randomized strategies are used.

Remark 2. Consider the stationary setting where $V_{\mathcal{C}(j)} = 0$. The expected one step regret when following $P^{(j)}$ is

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} P^{(j)}(x)\Delta_{\mathcal{C}(j-1)}(x) \\
&\quad\le \sum_{x\in\mathcal{X}} Q^{(j)}(x)\Delta_{\mathcal{C}(j-1)}(x) + 2\mu_j\sum_{x\in\mathcal{X}}\pi_\varphi(x) \\
&\quad\le 2\sum_{x\in\mathcal{X}} Q^{(j)}(x)\widehat{\Delta}_{\varphi,\mathcal{C}(j-1)}(x) + \mathcal{O}(\mu_j) \le \mathcal{O}(\mu_j)
\end{aligned}$$

where π_φ is the optimal design on $\mathcal{X}$, the first inequality uses $\Delta_{\mathcal{C}(j-1)} \le 2$ and the last inequality uses Lemma 4.4.

By the remark above, we can show the following theorem.

Theorem 4.6. *Under Assumption A with stationary reward functions $r_t(\cdot) = r(\cdot)$ for all $t \in [T]$, the dynamic regret bound of OPKB using a feature mapping induced by the kernel k is bounded with high probability by*

$$\mathrm{REG}_T \le \widetilde{\mathcal{O}}\left(\sqrt{\gamma_T T\log N}\right).$$

Proof sketch. By Remark 2, the expected regret of the block $\mathcal{B}(j)$ is $\mathcal{O}(|\mathcal{B}(j)|\sqrt{2^{-j}}) = \mathcal{O}(E\sqrt{2^j})$. Summing over all

blocks gives the bound $\mathcal{O}(\sqrt{\gamma_T T \log N})$ on the expected total regret. See Appendix D for a full proof. ■

Our regret bound for OPKB is order-optimal (Salgia et al. 2021) and matches work by Salgia et al. (2021), Camilleri et al. (2021), Li et al. (2022), and Valko et al. (2013). It is an improvement over Srinivas et al. (2010) and Chowdhury et al. (2017).

4.3 ADA-OPKB: Adapting OPKB to Non-Stationarity

In this section, we propose an algorithm called ADA-OPKB for the non-stationary kernel bandit setting that does not require the knowledge of the non-stationarity budgets.

Remark 3. Before our paper, the most natural attempt for designing an algorithm for non-stationary KB is to use the MASTER reduction (Wei et al. 2021) on GPUCB (Chowdhury et al. 2017), a UCB-based algorithm for stationary kernel bandits. This is because the MASTER reduction most naturally works for a UCB-based base algorithm. Also, the required analysis of GPUCB under non-stationary environment is available in the literature (Zhou et al. 2021). However, as shown in Appendix E, the reduction of GPUCB gives worse dynamic regret bound compared to ADA-OPKB due to the suboptimal concentration bound of regression based reward estimates.

ADA-OPKB adapts OPKB to non-stationarity by restarting upon detecting a significant change in reward functions. The key is to use past strategies as change detectors. Lemma 4.5 suggests that the strategy $P^{(j)}$ can detect changes in suboptimality gaps greater than $\sim \sqrt{2^{-j}}$ after running for $\sim 2^j$ time steps. ADA-OPKB replays older strategies with small indices to detect large changes fast and more recent strategies to detect small changes after running for longer time intervals. Algorithm 3 shows the full algorithm. Highlighted lines indicate the difference from OPKB.

Before starting a new block j, ADA-OPKB calls SCHEDULE (Algorithm 4), similar to the scheduler in Wei et al. (2021)), for determining when to use which of the strategies $P^{(0)}, \dots, P^{(j)}$. The procedure generates a set of replay intervals denoted by $(m, \mathcal{I})$ where m indicates the strategy index and $\mathcal{I}$ indicates the time interval scheduled for playing the strategy $P^{(m)}$. A replay schedule of index m has length $2^m E$ and there are 2^{j-m} slots in block j available to be scheduled. For each slot, the algorithm randomly schedule a replay of index m with probability $\sqrt{2^{m-j}}$. When multiple replay intervals are scheduled at a given time t, the algorithm selects the one with the smallest index. The strategy used at time t is denoted by m_t. Upon completion of a replay interval $\mathcal{I}$, the change detection test (8) is run. A restart is triggered if the test detects a significant change in reward functions. The test is based on the comparison of the empirical gap $\widehat{\Delta}_{\varphi,\mathcal{I}}$ and $\widehat{\Delta}_{\varphi,\mathcal{C}(k)}$ where $\mathcal{C}(k)$ is any cumulative block prior to $\mathcal{I}$.

Algorithm 3: ADA-OPKB: ADAptive Optimization Problem based Kernel Bandit Algorithm

Input: feature map φ, horizon T, confidence $\delta \in (0,1)$.
Definition: $\mu_j = c_1 2^{-j/2}$, $\beta_j = c_2 \gamma_{\varphi,T} 2^{j/2}$, $E = \lceil c_3 \gamma_{\varphi,T} \log(N/\delta) \rceil$, $\alpha = c_4 \sigma / \log(N/\delta)$
Initialize: $t \leftarrow 1$, epoch index $i \leftarrow 1$, $Q^{(0)} \leftarrow \pi_\varphi(\mathcal{X})$
1 **for** $j = 0, 1, \dots$ **do**
2 Set $\mathcal{B}(j) \leftarrow [t, t + 2^j E - 1]$ and $\mathcal{C}(j) \leftarrow \cup_{k=0}^{j} \mathcal{B}(k)$.
3 **if** $j \geq 1$ **then**
4 Compute $\widehat{\Delta} \leftarrow \{\widehat{\Delta}_{\varphi,\mathcal{C}(j-1)}(x)\}_{x \in \mathcal{X}}$.
5 Find strategy $Q^{(j)} \leftarrow \mathrm{OP}(\varphi, \widehat{\Delta}, \alpha, \beta_j, T)$.
6 Set $P^{(j)} \leftarrow (1 - \mu_j) Q^{(m_t)} + \mu_j \pi_\varphi(\mathcal{X})$.
7 Generate replay schedule $\mathcal{S} \leftarrow$ SCHEDULE(t, j).
8 **while** $t \in \mathcal{B}(j)$ **do**
9 $m_t \leftarrow \min\{m : (m, \mathcal{I}) \in \mathcal{S} \text{ with } t \in \mathcal{I}\}$; `// smallest index of scheduled intervals`
10 Play $x_t \sim P^{(m_t)}$; receive y_t; increment $t \leftarrow t + 1$.
11 If **Test** triggers a restart, increment i; go to Line 1.

Test: Trigger a restart if for any $(m, \mathcal{I}) \in \mathcal{S}$ with $\mathcal{I}$ ending at t and $k < j$, the following holds

$$\widehat{\Delta}_{\varphi,\mathcal{I}}(x) - 4\widehat{\Delta}_{\varphi,\mathcal{C}(k)}(x) > 4c_0\mu_{m \wedge k} \quad \text{or} \quad \widehat{\Delta}_{\varphi,\mathcal{C}(k)}(x) - 4\widehat{\Delta}_{\varphi,\mathcal{I}}(x) > 4c_0\mu_{m \wedge k}. \tag{8}$$

4.4 Analysis of ADA-OPKB

With the key lemmas proved for analyzing OPKB, we use ideas from Chen et al. (2019) and Wei et al. (2021) to analyze ADA-OPKB. We provide a sketch of the proof below. We suppress the dependency of the regret bound on γ_T and $\log N$ for simplicity. See Appendix F for the full proof.

Step 1: Interval Regret Using a martingale concentration, we can bound the regret of an interval $\mathcal{J}$ inside a block j as $\mathrm{REG}_\mathcal{J} \leq \mathcal{O}(\sum_{t \in \mathcal{J}} \mu_{m_t} + |\mathcal{J}| V_\mathcal{J} + |\mathcal{J}| \zeta_\mathcal{J})$ where $\zeta_\mathcal{J} := \max_{x \in \mathcal{X}}(\Delta_\mathcal{J}(x) - 8\widehat{\Delta}_{\mathcal{C}(j-1)}(x))$ measures the change in average reward in $\mathcal{J}$ compared to the previous block $j-1$.

Algorithm 4: SCHEDULE

Input: starting time t, block index j, base block size E
Initialize: $\mathcal{S} \leftarrow \{(j, [t, t + 2^j E - 1])\}$
1 **for** $\tau = 0, \dots, 2^j E - 1$ **do**
2 **for** $m = 0, \dots, j - 1$ **do**
3 **if** *τ is a multiple of $2^m E$* **then**
4 With probability $\frac{\sqrt{2^m}}{\sqrt{2^j}}$, add $(m, [t + \tau, t + \tau + 2^m E - 1])$ to $\mathcal{S}$.
Return: $\mathcal{S}$

See Appendix F.3 for the proof. Note that the interval regret is a sum of the expected one step regret assuming stationarity (Remark 2), the degree of non-stationarity within $\mathcal{J}$, and the magnitude of the change in reward function compared to the last block.

Step 2: Block Regret To bound the regret of a block j, we partition the block into nearly stationary intervals $\mathcal{J}_1, \dots, \mathcal{J}_\ell$ so that $V_{\mathcal{J}_i} \le \mu_{\mathcal{J}_i}$ where $\mu_{\mathcal{I}} := c|\mathcal{I}|^{-1/2}$. Summing over the interval regret of $\mathcal{J}_k$ in Step 1 and applying Cauchy-Schwarz, we get $\textsc{Reg}_{\mathcal{B}(j)} \le \mathcal{O}(\sum_{t\in\mathcal{B}(j)} \mu_{m_t} + \sqrt{\ell}|\mathcal{B}(j)|\mu_j + \sum_{k=1}^{\ell} |\mathcal{J}_k| \zeta_{\mathcal{J}_k} \mathbb{I}\{\zeta_{\mathcal{J}_k} > c'\mu_{\mathcal{J}_k}\})$. The first term $\sum_{t\in\mathcal{B}(j)} \mu_{m_t}$ can be shown to be $\widetilde{\mathcal{O}}(|\mathcal{B}(j)|\mu_j)$, which suggests the replays of past strategies are not overdone (Lemma F.6). To bound the third term, we use the property of change detection test that when $\zeta_{\mathcal{J}_k}$ is above $c'\mu_{\mathcal{J}_k}$ then replaying a suitable strategy within $\mathcal{J}_k$ triggers a restart (Lemma F.5). We can show that the replays of past strategies are done frequently enough to terminate the block before the third term gets too large, leading to a bound $\widetilde{\mathcal{O}}(\sqrt{\ell}|\mathcal{B}(j)|\mu_j)$ (proof of Lemma F.11). Finally, we can greedily construct a partition with $\ell = \widetilde{\mathcal{O}}(\min\{L_{\mathcal{B}(j)}, V_{\mathcal{B}(j)}^{2/3}|\mathcal{B}(j)|^{1/3}\})$ (Lemma F.10), which gives a block regret bound of $\widetilde{\mathcal{O}}(\min\{\sqrt{2^j L_{\mathcal{B}(j)}}, V_{\mathcal{B}(j)}^{1/3}(2^j)^{2/3}\})$ (Lemma F.11).

Step 3: Epoch Regret Since the block size is doubling, there can be at most $\mathcal{O}(\log_2 T)$ blocks in an epoch. Summing up regret bounds of the blocks and applying Cauchy-Schwarz and Hölder's inequalities, we can bound the epoch regret by $\widetilde{\mathcal{O}}(\min\{\sqrt{L_{\mathcal{E}_i}|\mathcal{E}_i|}, V_{\mathcal{E}_i}^{1/3}|\mathcal{E}_i|^{2/3}\})$ (Lemma F.13).

Step 4: Total Regret By the property of the change detection test, restarts can be triggered only when the degree of non-stationarity is large enough (Lemma F.3). Using this property, we can bound the number of epochs by $\widetilde{\mathcal{O}}(\min\{L_T, V_T^{2/3}T^{1/3}\}$ (Lemma F.12). The epoch regret bound in Step 3 gives total regret bound of $\widetilde{\mathcal{O}}(\min\{\sqrt{L_T T}, V_T^{1/3}T^{2/3}\})$ (Theorem 3.1).

5 DYNAMIC FEATURE MAPPING USING A NEURAL NETWORK

Recall that OPKB and ADA-OPKB use a fixed feature mapping induced by a kernel. In this section, we present extensions of OPKB and ADA-OPKB called OPNN and ADA-OPNN respectively that use *dynamic* feature mappings induced by a neural network trained using past history.

5.1 Preliminaries and Notations

Neural Network Following Zhou et al. (2020), we use a fully connected neural network with width m and depth L: $f(x; \boldsymbol{W}) = \sqrt{m}\boldsymbol{W}_L\sigma(\boldsymbol{W}_{L-1}\sigma(\cdots\sigma(\boldsymbol{W}_1 x)\cdots))$ where $\sigma(x) = \max\{x, 0\}$ is the ReLU activation function, $\boldsymbol{W}_1 \in \mathbb{R}^{m\times d}$, $\boldsymbol{W}_i \in \mathbb{R}^{m\times m}$ for $i = 2, \dots, L-1$, $\boldsymbol{W}_L \in \mathbb{R}^{m\times 1}$ and $\boldsymbol{W} = [\mathrm{vec}(\boldsymbol{W}_1)^T, \dots, \mathrm{vec}(\boldsymbol{W}_L)^T]^T \in \mathbb{R}^p$ with $p = m + md + m^2(L-1)$. We denote by $g(x; \boldsymbol{W}) = \nabla_{\boldsymbol{W}} f(x; \boldsymbol{W}) \in \mathbb{R}^p$ the gradient of the neural network function. We call $g(\cdot; \boldsymbol{W})$ the feature mapping induced by the neural network f with parameter $\boldsymbol{W}$. Each entry of the initial weights $\boldsymbol{W}^{(0)}$ of the network is sampled independently from $\mathcal{N}(0, 2/m)$.

Neural Tangent Kernel By Jacot et al. (2018), $\langle g(x; \boldsymbol{W}^{(0)}), g(x'; \boldsymbol{W}^{(0)})\rangle$ converges in probability to $H(x, x')$ for all $x, x' \in \mathcal{X}$ where the deterministic kernel $H(\cdot,\cdot)$ is called the *neural tangent kernel*. We denote by $\boldsymbol{H} = \{H(x, x')\}_{x,x'\in\mathcal{X}}$ the neural tangent kernel matrix.

Algorithm 5: OPNN: Optimization Problem based algorithm using Neural Network

Input: network width m, network depth L, time horizon T, confidence level $\delta \in (0, 1)$.

Initialize: $t \leftarrow 1$, initialize network weights $\boldsymbol{W}^{(0)}$, compute feature mapping $\varphi^{(0)}$ equivalent to $g(\cdot; \boldsymbol{W}^{(0)})$, find optimal design $P^{(0)} \leftarrow \pi_{\varphi^{(0)}}(\mathcal{X})$

Definition: $\mu_j = c_1 2^{-j/2}$, $\beta_j = c_2\gamma_{\varphi^{(0)},T}2^{j/2}$, $E = c_3\gamma_{\varphi^{(0)},T}\log(C_0 N/\delta)$, $\alpha = c_4\sigma/\log(C_0 N/\delta)$

for $j = 0, 1, \dots$ **do**
- Set $\mathcal{B}(j) \leftarrow [t, t + 2^j E - 1]$ and $\mathcal{C}(j) \leftarrow \cup_{k=0}^{j}\mathcal{B}(k)$.
- **if** $j \ge 1$ **then**
  - $\boldsymbol{W}^{(j)} \leftarrow \textsc{TrainNN}(\{(x_\tau, y_\tau)\}_{\tau\in\mathcal{C}(j-1)}, \boldsymbol{W}^{(0)})$.
  - Find a mapping $\varphi^{(j)}$ equivalent to $g(\cdot; \boldsymbol{W}^{(j)})/\sqrt{m}$.
  - Compute $\widehat{\Delta} \leftarrow \{\widehat{\Delta}_{\varphi^{(j)},\mathcal{C}(j-1)}(x)\}_{x\in\mathcal{X}}$.
  - Find strategy $Q^{(j)} \leftarrow \mathrm{OP}(\varphi^{(j)}, \widehat{\Delta}, \alpha, \beta_j, T)$.
  - Set $P^{(j)} \leftarrow (1-\mu_j)Q^{(j)} + \mu_j\pi_{\varphi^{(j)}}(\mathcal{X})$.
- **while** $t \in \mathcal{B}(j)$ **do**
  - Play $x_t \sim P^{(j)}$. Receive y_t. Increment $t \leftarrow t + 1$.

For the analysis of OPNN and ADA-OPNN, we make the following assumptions. The first assumption is on the invertibility of the neural tangent kernel matrix $\boldsymbol{H}$.

Assumption B. *For some $\lambda_0 > 0$, we have $\boldsymbol{H} \succcurlyeq \lambda_0\boldsymbol{I}$.*

This is a mild assumption commonly made when analyzing neural networks (Du et al. 2019a; Arora et al. 2019) and for analyzing neural bandit algorithms (Salgia et al. 2022; Zhou et al. 2020; Zhang et al. 2020; Gu et al. 2021; Kassraie et al. 2021). It is satisfied, for example, as long as no two actions in $\mathcal{X}$ are parallel (see Theorem 3.1 in Du et al. (2019b)). The second assumption is on the regularity of the reward functions commonly made in the neural bandits literature (Zhou et al. 2020; Zhang et al. 2020; Gu et al. 2021).

Assumption C. *We have $\sqrt{\boldsymbol{r}_t^T\boldsymbol{H}^{-1}\boldsymbol{r}_t} \le B$ for all $t = 1, \dots, T$ where $\boldsymbol{r}_t = (r_t(a_1), \dots, r_t(a_N))$.*

5.2 OPNN and ADA-OPNN

Unlike OPKB that uses a fixed feature mapping determined by a prespecified kernel, OPNN (Algorithm 5) uses the feature mapping induced by a neural network trained using past history. For the initial block, OPNN uses the feature mapping induced by the initial weight $\boldsymbol{W}^{(0)}$. Before starting a new block, OPNN trains the neural network with all past history using the procedure TRAINNN (Algorithm 6) and recomputes the feature mapping using the newly trained weight. The TRAINNN algorithm takes in training history and perform J steps of gradient descent on the squared error loss regularized by L2 distance of the weight $\boldsymbol{W}$ from the initial weight $\boldsymbol{W}^{(0)}$. Rest of the algorithm is the same as OPKB.

To adapt to non-stationarity, ADA-OPNN equips OPNN with change detection just as ADA-OPKB does with OPKB. See Appendix B for the full algorithm of ADA-OPNN.

Algorithm 6: TRAINNN: train neural network

Input: training history $\{(x_t, y_t)\}_{t\in\mathcal{I}}$, regularization parameter λ, step size η, number of gradient descent steps J, network width m, initial parameter $\boldsymbol{W}^{(0)}$

Define $\mathcal{L}(\boldsymbol{W}) = \sum_{t\in\mathcal{I}}(f(x_t;\boldsymbol{W}) - y_t)^2/2 + m\lambda\|\boldsymbol{W} - \boldsymbol{W}^{(0)}\|_2^2/2$.

for $j = 0, \dots, J-1$ **do**

  $\boldsymbol{W}^{(j+1)} \leftarrow \boldsymbol{W}^{(j)} - \eta\nabla\mathcal{L}(\boldsymbol{W}^{(j)})$.

Return: $\boldsymbol{W}^{(J)}$.

5.3 Analysis of OPNN and ADA-OPNN

Jacot et al. (2018) show that the neural tangent kernel stays constant during training in the infinite network width limit. Hence, in the infinite width limit, OPNN and ADA-OPNN are equivalent to OPKB and ADA-OPKB respectively that use the feature mapping corresponding to the kernel H. We can expect that in the finite width regime, the regret bound for OPNN and ADA-OPNN are the same as that for OPKB and ADA-OPKB respectively as long as the network width is large enough. Theorem 5.1 and Theorem G.1 confirm this. See Appendix G for the full proof.

Remark 4. The current NTK theory limits us to work in the infinite width regime where the feature mapping remains fixed. However, we empirically show in Appendix J that using the dynamic feature mapping induced by a finite width neural network is beneficial. This finding is consistent with numerous empirical results demonstrated by Fort et al. (2020) and Lee et al. (2020) in the supervised learning setting. We leave the analysis beyond the infinite width regime as future work.

Theorem 5.1 (Informal)**.** *Under Assumption B and Assumption C, the ADA-OPNN algorithm using a neural network of sufficiently large width achieves a dynamic regret bound of*

$$\widetilde{\mathcal{O}}(\min\{\sqrt{\gamma_T L_T T\log N}, (\gamma_T V_T \log N)^{1/3}T^{2/3} + \sqrt{\gamma_T T\log N}\})$$

with high probability, where γ_T is the maximum information gain corresponding to the neural tangent kernel H of the neural network used in the algorithm.

Relation to Previous Work Our regret bound of ADA-OPNN becomes $\widetilde{\mathcal{O}}(\sqrt{\gamma_T T\log N})$ when adapted to the stationary setting, which is an improvement over previous work (Zhou et al. 2020; Gu et al. 2021; Jia et al. 2022) by a factor of $\sqrt{\gamma_T}$ and is comparable to work by Kassraie et al. (2021).

6 EXPERIMENTS

The most notable feature of our algorithms is that they can adapt to non-stationarity without prior knowledge of the degree of non-stationarity. In this section, we illustrate this feature by comparing to previous work SW-GPUCB (Zhou et al. 2021) and WGPUCB (Deng et al. 2022), both of which require the knowledge of the degree of non-stationarity to tune parameters. For the parameter tuning and the experiments, we used an internal cluster of nodes with 20-core 2.40 GHz CPU and Tesla V100 GPU. The total amount of computing time was around 300 hours.

Experiment Design We run all algorithms in two environments: an environment with a single switch and the other with two switches. We first tune the algorithms for the first environment. Then, we run the tuned algorithms on the second environment to see how the algorithms adapt to the new non-stationarity.

Environments We run all simulations for $T = 10000$ rounds. For each simulation, we randomly sample an action set of size $N = 100$ from the unit sphere in $\mathbb{R}^d$. We follow Chowdhury et al. (2017) and sample the reward vector $\{r(x)\}_{x\in\mathcal{X}}$ from the multivariate normal distribution $\mathcal{N}(0, \boldsymbol{K})$ where $\boldsymbol{K} = \{k(x,x')\}_{x,x'\in\mathcal{X}}$ and k is the radial basis function kernel with length scale 0.2. We scale the reward vector so that the maximum absolute reward is 0.8, We sample the noises η_t from $\mathcal{N}(0, 0.1^2)$. We run experiments on two environments: the first environment has a single switch at time 3000 and the second environment has switches at time 1500 and 5000.

Algorithm Tuning We tune SW-GPUCB, WGPUCB, ADA-OPKB, ADA-OPNN on the first environment with a single switch. For SW-GPUCB, we do a grid search for λ over the range $\{0.01, 0.02, 0.05, 0.1, \dots, 100\}$, the UCB scale parameter v over $[0.001, 1]$, and the window size over $\{100, 200, 500, 1000, \dots, 10000\}$. See

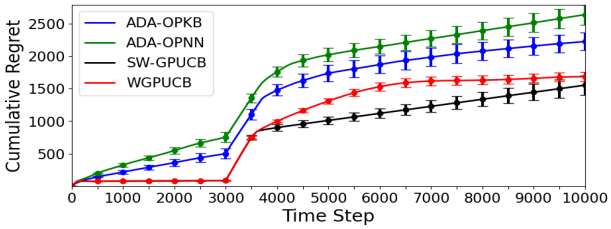
Algorithm 8 for the definition of λ. For WG-PUCB, we do a grid search for λ over the range $\{0.01, 0.02, 0.05, 0.1, \ldots, 100\}$, the UCB scale parameter over $\{0.001, 0.002, 0.005, 0.01, \ldots, 1\}$, and the discounting factor over $\{0.99, 0.995, 0.999, 0.9995, 0.9999\}$. See Algorithm 8 for the definition of λ. For ADA-OPKB and ADA-OPNN, we do a grid search for σ over $\{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}$ and c_0, c_1, c_2, c_3, c_4 over $\{0.001, 0.002, 0.005, 0.01, \ldots, 100\}$. For ADA-OPNN, we do a grid search for the learning rate η over $\{10^{-9}, 10^{-8}, 10^{-7}\}$, training steps J over $\{100, 1000, 10000\}$ and regularization parameter λ over $\{1, 10, 100, 1000\}$. We use a neural network of depth $L = 3$ and width $m = 2048$.

Remark 5. Compared to SW-GPUCB and WGPUCB, ADA-OPKB and ADA-OPNN have many parameters to tune. We leave designing a simpler algorithm with less parameters that does not require the knowledge of non-stationarity as future work.

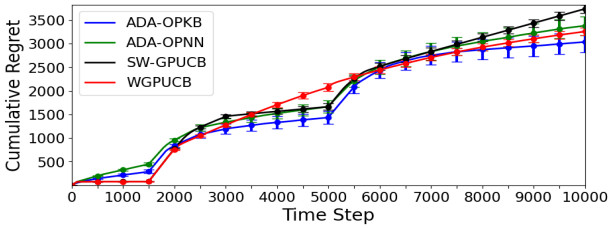
Results The cumulative regrets of SW-GPUCB, WG-PUCB, ADA-OPKB and ADA-OPNN averaged over 25 random seeds are shown in Figure 1. Error bars indicate standard errors of the means. Plot (a) shows the performances of the algorithms tuned under the first environment (a single switch). We remark that SW-GPUCB outperforms ADA-OPKB and ADA-OPNN in the initial stationary interval because ADA-OPKB and ADA-OPNN have overhead of running change detections. We conjecture that ADA-OPNN performs worse than ADA-OPKB due to kernel mismatch: ADA-OPKB uses the kernel used by the nature for drawing reward functions while ADA-OPNN does not.

Plot (b) shows the performances of the algorithms on the second environment (switches at time 1500 and 5000). SW-GPUCB optimally tuned for the single switch environment (window size 3000), performs worse than ADA-OPKB and ADA-OPNN in the new environment. WGPUCB optimally tuned for the single switch environment (discounting factor of 0.9995) performs similarly to ADA-OPNN but is outperformed by ADA-OPKB. This experiment highlights the fact that ADA-OPKB and ADA-OPNN can adapt to new non-stationarity better than SW-GPUCB and WGPUCB.

For an experiment that demonstrates the benefit of dynamically updating feature mapping for OPNN, and an experiment under a slowly varying environment, see Appendix J.



(a) Environment 1: single switch



(b) Environment 2: two switches

Figure 1: Cumulative regret comparison of algorithms in non-stationary environments

7 CONCLUSION

In this paper, we propose an algorithm for non-stationary kernel bandits that does not require the knowledge of non-stationary budgets, and show a simultaneous dynamic regret bound in terms of the budgets on the total variation and the number of changes in reward functions. The dynamic regret bound is tighter than previous work on the non-stationary kernel bandit setting. Also, our algorithm is nearly minimax optimal in the non-stationary linear bandit setting when run with a linear kernel. We provide an extension of our algorithm using a neural network. An interesting future work would be to adapt to a new non-stationary measure that tracks the number of times the identity of the best arm changes, which is a smaller measure than the number of changes in the reward functions. We believe the reward estimate based change detection algorithm and its analysis in this paper is suitable for this extension.

8 Acknowledgements

We acknowledge the support of NSF via grant IIS-2007055.

References

Abbasi-yadkori, Yasin, Dávid Pál, and Csaba Szepesvári (2011). "Improved Algorithms for Linear Stochastic Bandits". In: *Advances in Neural Information Processing Systems*. Vol. 24.

Agarwal, Alekh, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire (2014). "Taming the monster: a fast and simple algorithm for contextual bandits". In: *International Conference on Machine Learning*, pp. 1638–1646.

Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song (2019). "A convergence theory for deep learning via over-parameterization". In: *International Conference on Machine Learning*, pp. 242–252.

Arora, Sanjeev, Simon S. Du, Wei Hu, Zhiyuan Li, Russ R. Salakhutdinov, and Ruosong Wang (2019). "On ex-

act computation with an infinitely wide neural net". In: *Advances in Neural Information Processing Systems* 32.

Auer, P. and R. Ortner (2018). "Adaptively tracking the best arm with an unknown number of distribution changes". In: *European Workshop on Reinforcement Learning*.

Auer, Peter, Pratik Gajane, and Ronald Ortner (2019). "Adaptively tracking the best bandit arm with an unknown number of distribution changes". In: *Conference on Learning Theory*, pp. 138–158.

Besbes, Omar, Yonatan Gur, and Assaf Zeevi (2014). "Stochastic multi-armed-bandit problem with non-stationary rewards". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, pp. 199–207.

Beygelzimer, Alina, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire (2011). "Contextual bandit algorithms with supervised learning guarantees". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26.

Camilleri, Romain, Kevin Jamieson, and Julian Katz-Samuels (2021). "High-dimensional experimental design and kernel bandits". In: *International Conference on Machine Learning*, pp. 1227–1237.

Chaloner, Kathryn and Isabella Verdinelli (1995). "Bayesian experimental design: a review". In: *Statistical Science* 10.3, pp. 273–304.

Chen, Yifang, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei (2019). "A new algorithm for non-stationary contextual bandits: efficient, optimal and parameter-free". In: *Conference on Learning Theory*, pp. 696–726.

Cheung, Wang Chi, David Simchi-Levi, and Ruihao Zhu (2019). "Learning to optimize under non-stationarity". In: *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1079–1087.

Cheung, Wang Chi, David Simchi-Levi, and Ruihao Zhu (2022). "Hedging the Drift: Learning to Optimize Under Nonstationarity". In: *Management Science* 68.3, pp. 1696–1713.

Chowdhury, Sayak Ray and Aditya Gopalan (2017). "On kernelized multi-armed bandits". In: *International Conference on Machine Learning*, pp. 844–853.

Dani, Varsha, Thomas Hayes, and Sham Kakade (2008). "Stochastic linear optimization under bandit feedback". In: *21st Annual Conference on Learning Theory*, pp. 355–366.

Deng, Yuntian, Xingyu Zhou, Baekjin Kim, Ambuj Tewari, Abhishek Gupta, and Ness Shroff (2022). "Weighted gaussian process bandits for non-stationary environments". In: *The 25nd International Conference on Artificial Intelligence and Statistics*.

Du, Simon, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai (2019a). "Gradient descent finds global minima of deep neural networks". In: *Proceedings of the 36th International Conference on Machine Learning*, pp. 1675–1685.

Du, Simon, Xiyu Zhai, Barnabas Poczos, and Aarti Singh (2019b). "Gradient Descent Provably Optimizes Over-parameterized Neural Networks". In.

Dudik, Miroslav, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang (2011). "Efficient optimal learning for contextual bandits". In: *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 169–178.

Fort, Stanislav, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli (2020). "Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel". In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 5850–5861.

Garivier, Aurélien and Eric Moulines (2011). "On upper-confidence bound policies for switching bandit problems". In: *Proceedings of the 22nd international conference on Algorithmic learning theory*, pp. 174–188.

Gu, Quanquan, Amin Karbasi, Khashayar Khosravi, Vahab Mirrokni, and Dongruo Zhou (2021). "Batched neural bandits". In: *arXiv:2102.13028 [cs, stat]*.

Hao, Botao, Tor Lattimore, and Csaba Szepesvari (2020). "Adaptive exploration in linear contextual bandit". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 3536–3545.

Hazan, Elad, Amit Agarwal, and Satyen Kale (2007). "Logarithmic regret algorithms for online convex optimization". In: *Machine Learning* 69.2-3, pp. 169–192.

Jacot, Arthur, Franck Gabriel, and Clement Hongler (2018). "Neural tangent kernel: convergence and generalization in neural networks". In: *Advances in Neural Information Processing Systems*. Vol. 31.

Jia, Yiling, Weitong Zhang, Dongruo Zhou, Quanquan Gu, and Hongning Wang (2022). "Learning Neural Contextual Bandits through Perturbed Rewards". In.

Kassraie, Parnian and Andreas Krause (2021). "Neural contextual bandits without regret". In: *arXiv:2107.03144 [cs, stat]*.

Kim, Baekjin and Ambuj Tewari (2020). "Randomized exploration for non-stationary stochastic linear bandits". In: *Conference on Uncertainty in Artificial Intelligence*, pp. 71–80.

Lattimore, Tor and Csaba Szepesvari (2017). "The end of optimism? An asymptotic analysis of finite-armed linear bandits". In: *Proceedings of the 20th International Con-*

ference on Artificial Intelligence and Statistics, pp. 728–737.

Lattimore, Tor and Csaba Szepesvári (2020). *Bandit Algorithms*.

Lee, Chung-Wei, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang (2021). "Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously". In: *International Conference on Machine Learning*, pp. 6142–6151.

Lee, Jaehoon, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein (2020). "Finite versus infinite neural networks: an empirical study". In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 15156–15172.

Li, Zihan and Jonathan Scarlett (2022). "Gaussian Process Bandit Optimization with Few Batches". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 92–107.

Luo, Haipeng, Chen-Yu Wei, Alekh Agarwal, and John Langford (2018). "Efficient contextual bandits in non-stationary worlds". In: *Conference On Learning Theory*, pp. 1739–1776.

Robbins, Herbert (1952). "Some aspects of the sequential design of experiments". In: *Bulletin of the American Mathematical Society* 58.5, pp. 527–535.

Russac, Yoan, Claire Vernade, and Olivier Cappé (2019). "Weighted linear bandits for non-stationary environments". In: *Advances in Neural Information Processing Systems*. Vol. 32.

Salgia, Sudeep, Sattar Vakili, and Qing Zhao (2021). "A Domain-Shrinking based Bayesian Optimization Algorithm with Order-Optimal Regret Performance". In: *Advances in Neural Information Processing Systems*. Vol. 34, pp. 28836–28847.

Salgia, Sudeep, Sattar Vakili, and Qing Zhao (2022). *Provably and Practically Efficient Neural Contextual Bandits*.

Srinivas, Niranjan, Andreas Krause, Sham Kakade, and Matthias Seeger (2010). "Gaussian process optimization in the bandit setting: no regret and experimental design". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 1015–1022.

Vakili, Sattar, Michael Bromberg, Jezabel Garcia, Da-shan Shiu, and Alberto Bernacchia (2021a). "Uniform generalization bounds for overparameterized neural networks". In: *arXiv:2109.06099 [cs, stat]*.

Vakili, Sattar, Kia Khezeli, and Victor Picheny (2021b). "On information gain and regret bounds in gaussian process bandits". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 82–90.

Valko, Michal, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini (2013). "Finite-time analysis of kernelised contextual bandits". In: *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 654–663.

Vandenberghe, Lieven, Stephen Boyd, and Shao-Po Wu (1998). "Determinant maximization with linear matrix inequality constraints". In: *SIAM Journal on Matrix Analysis and Applications* 19.2, pp. 499–533.

Wei, Chen-Yu and Haipeng Luo (2021). "Non-stationary reinforcement learning without prior knowledge: an optimal black-box approach". In: *Conference on Learning Theory*.

Zhang, Weitong, Dongruo Zhou, Lihong Li, and Quanquan Gu (2020). "Neural thompson sampling". In.

Zhao, Peng and Lijun Zhang (2021). "Non-stationary linear bandits revisited". In: *arXiv:2103.05324 [cs]*.

Zhao, Peng, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou (2020). "A simple approach for non-stationary linear bandits". In: *International Conference on Artificial Intelligence and Statistics*, pp. 746–755.

Zhou, Dongruo, Lihong Li, and Quanquan Gu (2020). "Neural contextual bandits with ucb-based exploration". In: *International Conference on Machine Learning*, pp. 11492–11502.

Zhou, Xingyu and Ness Shroff (2021). "No-regret algorithms for time-varying bayesian optimization". In: *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6.

Supplementary Materials

A Notation Table

Notation	Definition	Explanation
$S_\varphi(Q,\lambda)$	$\sum_{x\in\mathcal{X}} Q(x)\varphi(x)\varphi(x)^T + \lambda I$	
$\gamma_{\varphi,T}$	$\max_{P\in\mathcal{P}_\mathcal{X}} \log\det S_\varphi(TP/\sigma, 1)$	Information gain with respect to φ
$V_{[s,t]}$	$\sum_{\tau=s}^{t-1} \|r_{\tau+1} - r_\tau\|_\infty$	Total variation in interval $[s,t]$
$L_{[s,t]}$	$\sum_{\tau=s}^{t-1} \mathbb{I}\{r_{\tau+1} \neq r_\tau\}$	Number of arm switches in $[s,t]$
$\pi_\varphi(\mathcal{A})$	$\operatorname{argmax}_{P\in\mathcal{P}_\mathcal{A}} \log\det S_\varphi(P,\sigma/T)$	Optimal design on $\mathcal{A}$ with respect to φ
$\mathcal{R}_\mathcal{I}(x)$	$\frac{1}{\lvert\mathcal{I}\rvert}\sum_{t\in\mathcal{I}} r_t(x)$	Average reward of arm x over interval $\mathcal{I}$
$\Delta_t(x)$	$\max_{x'\in\mathcal{X}} r_t(x') - r_t(x)$	Optimality gap of x at time t
$\Delta_\mathcal{I}(x)$	$\max_{x'\in\mathcal{X}} \mathcal{R}_\mathcal{I}(x') - \mathcal{R}_\mathcal{I}(x)$	Average optimality gap over the interval $\mathcal{I}$
$\widehat{\mathcal{R}}_{\varphi,t}(x)$	$\varphi(x)^T S_\varphi(P_t,\sigma/T)^{-1}\varphi(x_t)y_t$	IPS estimator for $r_t(x)$ with respect to φ
$\widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x)$	$\frac{1}{\lvert\mathcal{I}\rvert}\sum_{t\in\mathcal{I}} \widehat{\mathcal{R}}_{\varphi,t}(x)$	IPS estimator for average reward of x over the interval $\mathcal{I}$
$\widehat{\Delta}_{\varphi,\mathcal{I}}(x)$	$\max_{x'\in\mathcal{X}} \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x') - \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x)$	Estimated optimality gap of arm x over the interval $\mathcal{I}$

B Omitted algorithms

The ADA-OPNN algorithm adapts the OPNN algorithm to the non-stationary environment by equipping change detection.

Algorithm 7: ADA-OPNN: ADAptive Optimization Problem based algorithm using Neural Network

Input: network width m, network depth L, time horizon T, confidence level $\delta \in (0,1)$.
Definition: $\mu_j = c_1 2^{-j/2}$, $\beta_j = c_2\gamma_{\varphi,T}2^{j/2}$, $E = \lceil c_3\gamma_{\varphi,T}\log(C_1N/\delta)\rceil$, $\alpha = c_4\sigma/\log(C_1N/\delta)$
Initialize: time step $t \leftarrow 1$, epoch index $i \leftarrow 1$, initial strategy $Q^{(0)} \leftarrow \pi_\varphi(\mathcal{X})$

1 **for** $j = 0, 1, \ldots$ **do**
2 Set block $\mathcal{B}(j) \leftarrow [t, t + 2^j E - 1]$ and cumulative block $\mathcal{C}(j) \leftarrow \cup_{k=0}^{j}\mathcal{B}(k)$.
3 **if** $j \geq 1$ **then**
4 $\boldsymbol{W}^{(j)} \leftarrow \textsc{TrainNN}(\{(x_\tau, y_\tau)\}_{\tau\in\mathcal{C}(j-1)}, \boldsymbol{W}^{(0)})$
5 Find a feature mapping $\varphi^{(j)}$ equivalent to $g(\cdot;\boldsymbol{W}^{(j)})/\sqrt{m}$
6 Compute the empirical gap $\widehat{\Delta} \leftarrow \{\widehat{\Delta}_{\varphi^{(j)},\mathcal{C}(j-1)}(x)\}_{x\in\mathcal{X}}$ using all past history in epoch i.
7 Find strategy $Q^{(j)} \leftarrow \text{OP}(\varphi^{(j)}, \widehat{\Delta}, \alpha, \beta_j, T)$; Set $P^{(j)} \leftarrow (1-\mu_j)Q^{(m_t)} + \mu_j\pi_{\varphi^{(j)}}(\mathcal{X})$.
8 Generate replay schedule $\mathcal{S} \leftarrow \textsc{Schedule}(t, j)$.
9 **while** $t \in \mathcal{B}(j)$ **do**
10 $m_t \leftarrow \min\{m : (m,\mathcal{I}) \in \mathcal{S} \text{ with } t \in \mathcal{I}\}$; // smallest index of scheduled intervals
11 Record $P_t \leftarrow P^{(m_t)}$. Play $x_t \sim P_t$ and receive reward y_t; Increment $t \leftarrow t+1$.
12 If **Test** with $\varphi = \varphi^{(j)}$ triggers a restart then increment i and go to Line 1.

Test: Trigger a restart if for any $(m,\mathcal{I}) \in \mathcal{S}$ with $\mathcal{I}$ ending at t and $k < j$, the following holds

$$\widehat{\Delta}_{\varphi,\mathcal{I}}(x) - 4\widehat{\Delta}_{\varphi,\mathcal{C}(k)}(x) > 4c_0\mu_{m\wedge k} \quad \text{or} \quad \widehat{\Delta}_{\varphi,\mathcal{C}(k)}(x) - 4\widehat{\Delta}_{\varphi,\mathcal{I}}(x) > 4c_0\mu_{m\wedge k}.$$

C Maximum information gain

In this section, we summarize the properties of the maximum information gain used in this paper. The original definition of the maximum information gain by Srinivas et al. (2010) is

$$\bar{\gamma}_T = \max_{x_1,\dots,x_T \in \mathcal{X}} \frac{1}{2} \log\det(\sigma^{-1} K_T + I_T)$$

where $K_T = [k(x_i, x_j)]_{i,j\in[T]}$ and $I_T \in \mathbb{R}^{T\times T}$ is the identity matrix. For ease of exposition, we drop the factor $\frac{1}{2}$ that appears in the original definition of $\bar{\gamma}_T$. In this paper, we define the *continuous* version of the maximum information gain γ_T as follows

$$\gamma_{\varphi,T} = \max_{P\in\mathcal{P}_\mathcal{X}} \log\det S_\varphi(\sigma^{-1} T P, I_N)$$

where $\varphi : \mathcal{X} \to \mathbb{R}^N$ is a feature mapping corresponding to the kernel k such that $k(x, x') = \langle\varphi(x), \varphi(x')\rangle$. To see the connection of $\gamma_{\varphi,T}$ to the original definition $\bar{\gamma}_T$, note that $K_T = HKH^T$ where $K = [k(a_i, a_j)]_{i,j\in[N]}$ and $H \in \{0, 1\}^{T\times N}$ with $H_{ti} = \mathbb{I}\{x_t = a_i\}$ is the history matrix that indicates whether the action a_i is played at time t for $t \in [T]$ and $i \in [N]$. Using the notation $\Phi = [\varphi(a_1)\cdots\varphi(a_N)]^T \in \mathbb{R}^{N\times N}$ such that $K = \Phi\Phi^T$, we have by the Sylvester's determinant identity $\det(I + AB) = \det(I + BA)$ that

$$\begin{aligned}
\log\det(\sigma^{-1} K_T + I_T) &= \log\det(\sigma^{-1} H\Phi\Phi^T H^T + I_T) \\
&= \log\det(\sigma^{-1}\Phi^T H^T H\Phi + I_N) \\
&= \log\det(\sigma^{-1}\Phi^T D_N\Phi + I_N) \\
&= \log\det S_\varphi(\sigma^{-1} T P_N, I_N)
\end{aligned}$$

where $D_N = H^T H = \mathrm{diag}(n_1, \dots, n_N)$ with n_i denoting how many times a_i appears in the sequence $x_1, \dots, x_T$ and $P_N = D_N/T$ is the relative frequency of the actions. Hence,

$$\bar{\gamma}_T = \max_{P\in\mathcal{P}_{T,\mathcal{X}}} \frac{1}{2} \log\det S_\varphi(\sigma^{-1} T P, I_N)$$

where the maximization is over $\mathcal{P}_{T,\mathcal{X}} \coloneqq \{P \in \mathcal{P}_\mathcal{X} : P(a_i) = n_i/T \text{ for all } i \in [N] \text{ with } n_i \in \mathbb{Z}\}$. It follows that our definition $\gamma_{\varphi,T}$ is a continuous version of the maximum information gain in the sense that it maximizes over $\mathcal{P}_\mathcal{X}$ instead of the discretized probability space $\mathcal{P}_{T,\mathcal{X}}$.

A direct consequence is that $\gamma_{\varphi,T} \geq \bar{\gamma}_T$. To get an upper bound on $\gamma_{\varphi,T}$ we can use Theorem 3 in Vakili et al. (2021b) that shows an upper bound of $\bar{\gamma}_T$ in terms of the eigendecay of the kernel $k(\cdot,\cdot)$. It can be seen that their proof can be easily adapted to the continuous version, which leads to upper bounds for common kernels in the following lemma.

Lemma C.1 (Theorem 3 in Vakili et al. (2021b))**.** *For the Matérn-ν kernel and the SE kernel, the maximum information gain is upper bounded by*

$$\begin{aligned}
\gamma_{\varphi,T} &= \mathcal{O}\left(T^{\frac{d}{2\nu+d}} \log^{\frac{2\nu}{2\nu+d}}(T)\right), \quad \textit{for Matérn-}\nu\textit{ kernel} \\
\gamma_{\varphi,T} &= \mathcal{O}\left(\log^{d+1}(T)\right), \quad \textit{for SE kernel.}
\end{aligned}$$

Similarly, adapting the proof of Theorem 2 in Vakili et al. (2021a), we get an upper bound on the maximum information gain for the neural tangent kernel of a ReLU network as follows.

Lemma C.2. *For the neural tangent kernel of a ReLU network, the maximum information gain is upper bounded by*

$$\gamma_{\varphi,T} = \mathcal{O}\left(T^{\frac{d-1}{d}} \log^{\frac{1}{d}}(T)\right).$$

For the linear kernel, we get the following upper bound on the maximum information gain.

Lemma C.3. *For the identity feature mapping $\varphi(x) = x$ for all $x \in \mathcal{X} \subset \mathbb{R}^d$ corresponding to the linear kernel $k(x, x') = \langle x, x'\rangle$, we have $\gamma_{\varphi,T} \leq \mathcal{O}(d \log T)$.*

Proof. Using the identity $\det(A) \leq (\mathrm{Tr}(A)/d)^d$ for a positive semi-definite matrix $A \in \mathbb{R}^{d\times d}$, which can be seen by the AM-GM inequality on the eigenvalues of A, we have

$$\begin{aligned}
\log\det S_\varphi(\sigma^{-1}TP, 1) &\leq d\log(\mathrm{Tr}(S_\varphi(\sigma^{-1}TP,1))/d) \\
&= d\log\left(\frac{1}{d}\,\mathrm{Tr}\left(\frac{T}{\sigma}\sum_{x\in\mathcal{X}} P(x)xx^T + I_d\right)\right) \\
&= d\log\left(\frac{1}{d}\left(\frac{T}{\sigma}\sum_{x\in\mathcal{X}} P(x)\|x\|_2^2 + d\right)\right) \\
&\leq d\log\left(\frac{T}{\sigma d} + 1\right) = \mathcal{O}(d\log T)
\end{aligned}$$

where the second inequality follows by the assumption that $\|x\|_2 \leq 1$. Taking the maximum over $P \in \mathcal{P}_\mathcal{X}$ completes the proof. ■

Lemma C.4. *For any feature mapping φ and any $P \in \mathcal{P}_\mathcal{X}$, we have*

$$\sum_{x\in\mathcal{X}} P(x)\|\varphi(x)\|^2_{S_\varphi(P,\sigma/T)^{-1}} \leq \gamma_{\varphi,T}.$$

Proof. We can rewrite the left hand side as

$$\begin{aligned}
\sum_{x\in\mathcal{X}} P(x)\|\varphi(x)\|^2_{S_\varphi(P,\sigma/T)^{-1}} &= \mathrm{Tr}\left(\sum_{x\in\mathcal{X}} S_\varphi(P,\sigma/T)^{-1}P(x)\varphi(x)\varphi(x)^T\right) \\
&= \mathrm{Tr}\left(S_\varphi(P,\sigma/T)^{-1}(S_\varphi(P,\sigma/T) - (\sigma/T)I_N)\right) \\
&\leq \log\det S_\varphi(P,\sigma/T) - \log\det(\sigma/T)I_N \\
&= \log\det S_\varphi((T/\sigma)P, 1) \leq \gamma_{\varphi,T}
\end{aligned}$$

where the first inequality uses the identity $\mathrm{Tr}(A^{-1}(A-B)) \leq \log\det A - \log\det B$ for $A \succcurlyeq B \succcurlyeq 0$ (Lemma 12 in Hazan et al. (2007)). ■

D Analysis of OPKB

In this section, we prove the high probability dynamic regret bound of the OPKB algorithm under the stationary kernel bandit setting stated below.

D.1 Constants and notations

We use the following parameters in this section (and in Section F) for ease of exposition of the proof: $c_0 = 40 + 16\sqrt{\alpha}$, $c_1 = \frac{1}{2}$, $c_2 = \frac{1}{10+4\sqrt{\alpha}}$, $c_3 = 4$, $c_4 = \frac{1}{4}$ so that $\mu_j = 2^{-(j+2)/2}$, $\beta_j = \frac{\gamma_{\varphi,T}}{10+4\sqrt{\alpha}}2^{j/2}$, $C_0 = 8T\log_2 T$, $E = \lceil 4\gamma_{\varphi,T}\log(C_0N/\delta)\rceil$, $\alpha = \sigma/(4\log(C_0N/\delta))$. We define $\xi_j = \frac{\mu_j}{4\gamma_{\varphi,T}}$. We frequently use the identities

$$c_0\beta_j\mu_j = 2\gamma_{\varphi,T}, \quad \xi_j\gamma_{\varphi,T} = \frac{\mu_j}{4}, \quad \mu_j\beta_j = \frac{2\gamma_{\varphi,T}}{c_0} \leq \frac{\gamma_{\varphi,T}}{20}, \quad \xi_j\beta_j = \frac{1}{2c_0} \leq \frac{1}{80}.$$

We denote by $\mathcal{R}_\mathcal{I}(x) = \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} r_t(x)$ the average reward of action x in interval $\mathcal{I}$. We define $\Delta_\mathcal{I}(x) = \max_{x'\in\mathcal{X}} \mathcal{R}_\mathcal{I}(x') - \mathcal{R}_\mathcal{I}(x)$.

D.2 Proof of Lemma 4.3

Proof of Lemma 4.3. For ease of exposition, we write $\pi^\star = \pi_\varphi(\mathcal{A})$. Recall that the optimal design $\pi^\star$ is a maximizer of $\log\det S_\varphi(P,\sigma/T)$ subject to $P(x) \geq 0$ for all $x \in \mathcal{A}$ and $\sum_{x\in\mathcal{A}} P(x) = 1$. Introducing Lagrange multipliers λ_x for the

conditions $P(x) \geq 0$ for all $x \in \mathcal{X}$ and λ for $\sum_{x\in\mathcal{A}} P(x) = 1$, the KKT optimality conditions give

$$\begin{aligned} \|\varphi(x)\|^2_{S_\varphi(\pi^\star,\sigma/T)^{-1}} + \lambda_x - \lambda = 0, \quad \text{for all } x \in \mathcal{A} && \text{(Stationarity)} \\ \lambda_x \geq 0, \quad \text{for all } x \in \mathcal{A} && \text{(Dual feasibility)} \\ \pi^\star(x)\lambda_x = 0, \quad \text{for all } x \in \mathcal{A} && \text{(Complementary slackness)} \end{aligned}$$

where we use the fact that $\frac{\partial}{\partial P(x)} \log\det S_\varphi(P, \sigma/T) = \|\varphi(x)\|^2_{S_\varphi(P,\sigma/T)^{-1}}$. Multiplying $\pi^\star(x)$ to the stationarity condition and summing over $x \in \mathcal{A}$, we get

$$\begin{aligned} 0 &= \sum_{x\in\mathcal{A}} \pi^\star(x)\|\varphi(x)\|^2_{S_\varphi(\pi^\star,\sigma/T)^{-1}} + \sum_{x\in\mathcal{A}} \pi^\star(x)\lambda_x - \lambda \sum_{x\in\mathcal{A}} \pi^\star(x) \\ &= \sum_{x\in\mathcal{A}} \pi^\star(x)\|\varphi(x)\|^2_{S_\varphi(\pi^\star,\sigma/T)^{-1}} - \lambda \end{aligned}$$

where the second equality uses the complementary slackness conditions. Hence,

$$\lambda = \sum_{x\in\mathcal{A}} \pi^\star(x)\|\varphi(x)\|^2_{S_\varphi(\pi^\star,\sigma/T)^{-1}} \leq \max_{P\in\mathcal{P}_\mathcal{X}} \sum_{x\in\mathcal{X}} P(x)\|\varphi(x)\|^2_{S_\varphi(P,\sigma/T)^{-1}} = \gamma_{\varphi,T}.$$

Using this result $\lambda \leq \gamma_{\varphi,T}$ to the stationarity conditions and using the dual feasibility conditions $\lambda_x \geq 0$, we get $\|\varphi(x)\|^2_{S_\varphi(\pi^\star,\sigma/T)^{-1}} = \lambda - \lambda_x \leq \lambda \leq \gamma_{\varphi,T}$ for all $x \in \mathcal{X}$ as desired. For the proof of $\mathrm{Var}(\widehat{\mathcal{R}}_{\varphi,t}(x)) \leq \|\varphi(x)\|^2_{S_\varphi(\pi_\varphi(\mathcal{A}),\sigma/T)^{-1}}$, refer to the proof of Lemma D.3. ■

D.3 Proof of Lemma 4.4

Proof of Lemma 4.4. Recall that the strategy returned by the algorithm $OP(\varphi, \widehat{\Delta}, \alpha, \beta, T)$ is $Q = \frac{1}{2}P^\star + \frac{1}{2}\pi^\star$ where $P^\star$ is the minimizer of $J(P) = \sum_{x\in\mathcal{X}} P(x)\widehat{\Delta}(x) - \frac{2}{\beta}\log\det S_\varphi(P, \sigma/T)$ among $\mathcal{P}_\mathcal{X}$ and we write $\pi^\star = \pi_\varphi(\mathcal{A})$ where $\mathcal{A} = \{x \in \mathcal{X} : \widehat{\Delta}(x) \leq 2\alpha\gamma_{\varphi,T}/\beta\}$. Since the empirical gap estimates satisfy $\widehat{\Delta}(x) \geq 0$ for all $x \in \mathcal{X}$ and there exists $\hat{x} \in \mathcal{X}$ with $\widehat{\Delta}(\hat{x}) = 0$, we can check that $P^\star$ is also a minimizer among the set of sub-distributions $\widetilde{\mathcal{P}}_\mathcal{X} = \{P \in \mathbb{R}^\mathcal{X} : P(x) \geq 0 \text{ for all } x \in \mathcal{X}, \sum_{x\in\mathcal{X}} P(x) \leq 1\}$. This can be seen by noting that for any sub-distribution $\widetilde{P}$, the proper distribution P obtained by increasing the weight of the empirically best action $\hat{x}$ satisfies $J(\widetilde{P}) \geq J(P)$. Introducing Lagrange multipliers λ_x for the conditions $P(x) \geq 0$ for all $x \in \mathcal{X}$ and λ for $\sum_{x\in\mathcal{X}} P(x) \leq 1$, the KKT optimality conditions give

$$\begin{aligned} \widehat{\Delta}(x) - \frac{2}{\beta}\|\varphi(x)\|^2_{S_\varphi(P^\star,\sigma/T)^{-1}} - \lambda_x + \lambda = 0, \quad \text{for all } x \in \mathcal{X} && \text{(Stationarity)} \\ \lambda_x \geq 0, \quad \text{for all } x \in \mathcal{X} && \text{(Dual feasibility)} \\ \lambda \geq 0 && \text{(Dual feasibility)} \\ P^\star(x)\lambda_x = 0, \quad \text{for all } x \in \mathcal{X}. && \text{(Complementary slackness)} \end{aligned}$$

Multiplying $P^\star(x)$ to the stationarity conditions and summing over $x \in \mathcal{X}$, we get

$$\begin{aligned} 0 &= \sum_{x\in\mathcal{X}} P^\star(x)\widehat{\Delta}(x) - \frac{2}{\beta}\sum_{x\in\mathcal{X}} P^\star(x)\|\varphi(x)\|^2_{S_\varphi(P^\star,\sigma/T)^{-1}} - \sum_{x\in\mathcal{X}} P^\star(x)\lambda_x + \lambda\sum_{x\in\mathcal{X}} P^\star(x) \\ &= \sum_{x\in\mathcal{X}} P^\star(x)\widehat{\Delta}(x) - \frac{2}{\beta}\sum_{x\in\mathcal{X}} P^\star(x)\|\varphi(x)\|^2_{S_\varphi(P^\star,\sigma/T)^{-1}} + \lambda \end{aligned} \tag{9}$$

where the second equality uses the complementary slackness conditions. Rearranging and using the dual feasibility condition $\lambda \geq 0$, we get

$$\sum_{x\in\mathcal{X}} P^\star(x)\widehat{\Delta}(x) = \frac{2}{\beta}\sum_{x\in\mathcal{X}} P^\star(x)\|\varphi(x)\|^2_{S_\varphi(P^\star,\sigma/T)^{-1}} - \lambda \leq \frac{2\gamma_{\varphi,T}}{\beta}.$$

It follows that $Q = \frac{1}{2}P^\star + \frac{1}{2}\pi^\star$ satisfies

$$\begin{aligned} \sum_{x\in\mathcal{X}} Q(x)\widehat{\Delta}(x) &= \frac{1}{2}\sum_{x\in\mathcal{X}} P^\star(x)\widehat{\Delta}(x) + \frac{1}{2}\sum_{x\in\mathcal{A}} \pi^\star(x)\widehat{\Delta}(x) \\ &\leq \frac{1}{2}\frac{2\gamma_{\varphi,T}}{\beta} + \frac{1}{2}\frac{2\alpha\gamma_{\varphi,T}}{\beta} = \frac{(1+\alpha)\gamma_{\varphi,T}}{\beta} \end{aligned}$$

where the inequality uses the fact that $\widehat{\Delta}(x) \le 2\alpha\gamma_{\varphi,T}/\beta$ for $x \in \mathcal{A}$ by the definition of $\mathcal{A}$. This proves the first inequality of the lemma. Also, since the empirical gaps satisfy $\widehat{\Delta}(x) \ge 0$ for all $x \in \mathcal{X}$, rearranging (9) gives

$$\lambda = \frac{2}{\beta}\sum_{x\in\mathcal{X}} P^\star(x)\|\varphi(x)\|^2_{S_\varphi(P^\star,\sigma/T)^{-1}} - \sum_{x\in\mathcal{X}} P^\star(x)\widehat{\Delta}(x) \le \frac{2\gamma_{\varphi,T}}{\beta}.$$

Hence, by the stationarity condition, we have for each $x \in \mathcal{X}$ that

$$\|\varphi(x)\|^2_{S_\varphi(P^\star,\sigma/T)^{-1}} = \frac{\beta\widehat{\Delta}(x)}{2} - \frac{\beta\lambda_x}{2} + \frac{\beta\lambda}{2} \le \frac{\beta\widehat{\Delta}(x)}{2} + \gamma_{\varphi,T}$$

where we use the dual feasibility condition $\lambda_x \ge 0$. Using the fact that $S_\varphi(Q,\sigma/T) \succcurlyeq \frac{1}{2}S_\varphi(P^\star,\sigma/T)$ gives the second inequality of the lemma. Finally, for the third inequality of the lemma, we argue for the cases $x \in \mathcal{A}$ and $x \notin \mathcal{A}$ separately. If $x \in \mathcal{A}$, then using $S_\varphi(Q,\sigma/T) \succcurlyeq \frac{1}{2}S_\varphi(\pi^\star,\sigma/T)$, we get $\|\varphi(x)\|^2_{S_\varphi(Q,\sigma/T)^{-1}} \le 2\|\varphi(x)\|^2_{S_\varphi(\pi^\star,\sigma/T)^{-1}} \le 2\gamma_{\varphi,T} \le \frac{\beta^2\widehat{\Delta}^2(x)}{2\alpha\gamma_{\varphi,T}} + 2\gamma_{\varphi,T}$. If $x \notin \mathcal{A}$, then we have $\widehat{\Delta}(x) > 2\alpha\gamma_{\varphi,T}/\beta$ by the definition of $\mathcal{A}$. Hence, $1 < \frac{\beta\widehat{\Delta}(x)}{2\alpha\gamma_{\varphi,T}}$ and the second inequality of the lemma gives $\|\varphi(x)\|^2_{S_\varphi(Q,\sigma/T)^{-1}} \le \beta\widehat{\Delta}(x) + 2\gamma_{\varphi,T} \le \frac{\beta^2\widehat{\Delta}^2(x)}{2\alpha\gamma_{\varphi,T}} + 2\gamma_{\varphi,T}$, as desired. ■

D.4 Concentration bound for reward estimates

In this subsection, we prove the following concentration bound for reward estimates.

Lemma D.1. *Let $\mathcal{I} \subseteq [1,T]$ be a time interval. Let m_t be the strategy index used by OPKB at time t. Let j be the maximum strategy index used in $\mathcal{I}$ such that $m_t \le j$ for all $t \in \mathcal{I}$. Then, with probability at least $1 - \frac{2\delta}{C}$, we have*

$$|\widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x) - \mathcal{R}_{\mathcal{I}}(x)| \le \frac{\xi_j}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\|\varphi(x)\|^2_{S_\varphi(P_t,\sigma/T)^{-1}} + \frac{\log(CN/\delta)}{\xi_j|\mathcal{I}|} + \frac{\sqrt{\sigma/T}}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\|\varphi(x)\|_{S_\varphi(P_t,\sigma/T)^{-1}}$$

for all $x \in \mathcal{X}$ where $\xi_j = \mu_j/(4\gamma_{\varphi,T})$.

The proof relies on the following Freedman-style martingale inequality. See Theorem 1 in Beygelzimer et al. 2011 for the proof of this inequality.

Lemma D.2 (Freedman)**.** *Let $X_1,\dots,X_n \in \mathbb{R}$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_0,\mathcal{F}_1,\dots$. Assume $X_i \le R$ a.s. for all i. Then for any $\delta \in (0,1)$ and $\xi \in [0,1/R]$, we have with probability at least $1-\delta$ that*

$$\sum_{i=1}^n X_i \le \xi V + \frac{\log(1/\delta)}{\xi},$$

where $V = \sum_{i=1}^n \mathbb{E}[X_i^2 \mid \mathcal{F}_{i-1}]$.

To apply the Freedman inequality, we analyze the distribution of the IPS estimator $\widehat{\mathcal{R}}_{\varphi,t}(x)$ in the following lemma.

Lemma D.3. *Suppose the reward function $r_t(\cdot)$ lies in a RKHS with a feature mapping $\psi : \mathcal{X} \to \ell^2$. Let $\varphi : \mathcal{X} \to \mathbb{R}^N$ be a feature mapping equivalent to ψ. Let m_t be the strategy index used at time t and $P_t = P^{(m_t)}$ be the strategy used at time t. Then, the IPS estimator $\widehat{\mathcal{R}}_{\varphi,t}(x) = \varphi(x)^T S_\varphi(P_t,\sigma/T)^{-1}\varphi(x_t)^T y_t$ satisfies*

$$\begin{aligned}
|\widehat{\mathcal{R}}_{\varphi,t}(x)| &\le \frac{\gamma_{\varphi,T}}{\mu_{m_t}} \\
|\mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi,t}(x)] - r_t(x)| &\le \sqrt{\sigma/T}\|\varphi(x)\|_{S_\varphi(P_t,\sigma/T)^{-1}} \\
\mathrm{Var}_t[\widehat{\mathcal{R}}_{\varphi,t}(x)] &\le \|\varphi(x)\|^2_{S_\varphi(P_t,\sigma/T)^{-1}}
\end{aligned}$$

where $\mathbb{E}_t$ and Var_t are the conditional expectation and the conditional variance given the history before time t respectively.

Proof. The first claim follows by

$$|\widehat{\mathcal{R}}_{\varphi,t}(x)| = \left|\varphi(x)^T S(P_t,\sigma/T)^{-1}\varphi(x_t)y_t\right| \le \|\varphi(x)\|_{S(P_t,\sigma/T)^{-1}}\|\varphi(x_t)\|_{S(P_t,\sigma/T)^{-1}} \le \frac{\gamma_{\varphi,T}}{\mu_{m_t}}$$

where the first inequality uses the assumption $|y_t| \le 1$ and the Cauchy-Schwarz inequality, and the second inequality uses $S_\varphi(P_t, \sigma/T) = (1-\mu_{m_t})S(Q^{(m_t)}, \sigma/T) + \mu_{m_t} S_\varphi(\pi_\varphi(\mathcal{X}), \sigma/T) \succcurlyeq \mu_{m_t} S(\pi_\varphi(\mathcal{X}), \sigma/T)$ and Lemma 4.3.

To show the second claim, let $\theta_t \in \ell^2$ be the parameter such that $r_t(x) = \langle \psi(x), \theta_t \rangle$ for all $x \in \mathcal{X}$. Since P_t is completely determined given history up to $t-1$, we have

$$\begin{aligned}
\mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi,t}(x)] &= \mathbb{E}_t[\psi(x)^T S_\psi(P_t, \sigma/T)^{-1} \psi(x_t)(\psi(x_t)^T \theta_t + \eta_t)] \\
&= \psi(x)^T S_\psi(P_t, \sigma/T)^{-1} \mathbb{E}_t[\psi(x_t)\psi(x_t)^T]\theta_t \\
&= \psi(x)^T S_\psi(P_t, \sigma/T)^{-1} (S_\psi(P_t, \sigma/T) - (\sigma/T) I)\theta_t \\
&= r_t(x) - (\sigma/T)\psi(x)^T S_\psi(P_t, \sigma/T)^{-1}\theta_t
\end{aligned}$$

where the first equality is by Lemma I.1 and the third equality uses the fact that the strategy P_t is deterministic given the history up to time t. The second claim follows by the bound

$$\begin{aligned}
(\sigma/T)\left|\psi(x)^T S_\psi(P_t, \sigma/T)^{-1}\theta_t\right| &\le (\sigma/T)\,\|\psi(x)\|_{S_\psi(P_t,\sigma/T)^{-1}}\,\|\theta_t\|_{S_\psi(P_t,\sigma/T)^{-1}} \\
&\le \sqrt{\sigma/T}\,\|\varphi(x)\|_{S_\varphi(P_t,\sigma/T)^{-1}}
\end{aligned}$$

where the first inequality is by the Cauchy-Schwarz inequality and the last inequality uses $S_\psi(P_t, \sigma/T)^{-1} \preccurlyeq (T/\sigma) I$, the assumption that $\|\theta_t\|_2 \le 1$ and Lemma I.1.

Finally, the third claim follows by

$$\begin{aligned}
\mathrm{Var}_t[\widehat{\mathcal{R}}_{\varphi,t}(x)] &\le \mathbb{E}_t[\{\varphi(x)^T S_\varphi(P_t, \sigma/T)^{-1}\varphi(x_t)\}^2 y_t^2] \\
&\le \varphi(x)^T S_\varphi(P_t, \sigma/T)^{-1} \mathbb{E}_t[\varphi(x_t)\varphi(x_t)^T] S_\varphi(P_t, \sigma/T)^{-1}\varphi(x) \\
&= \varphi(x)^T S_\varphi(P_t, \sigma/T)^{-1} S_\varphi(P_t, 0) S_\varphi(P_t, \sigma/T)^{-1}\varphi(x) \\
&\le \|\varphi(x)\|^2_{S_\varphi(P_t,\sigma/T)^{-1}}.
\end{aligned}$$

where the second inequality uses the assumption $|y_t| \le 1$ and the last inequality uses $S_\varphi(P_t, 0) \preccurlyeq S_\varphi(P_t, \sigma/T)$. ■

We are now ready to prove Lemma D.1.

Proof of Lemma D.1. Fix an action $x \in \mathcal{X}$ and consider a martingale difference sequence $\{z_{t,x}\}_{t\in\mathcal{I}}$ where $z_{t,x} = \widehat{\mathcal{R}}_{\varphi,t}(x) - \mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi,t}(x)]$. We can bound $z_{t,x}$ for all $t \in \mathcal{I}$ by

$$z_{t,x} \le |\widehat{\mathcal{R}}_{\varphi,t}(x)| + |\mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi,t}(x)]| \le |\widehat{\mathcal{R}}_{\varphi,t}(x)| + \mathbb{E}_t[|\widehat{\mathcal{R}}_{\varphi,t}(x)|] \le \frac{2\gamma_{\varphi,T}}{\mu_j}$$

where the last inequality uses Lemma D.3 and $m_t \le j$. Also, by Lemma D.3, we have

$$\mathrm{Var}_t[z_{t,x}] = \mathrm{Var}_t[\widehat{\mathcal{R}}_{\varphi,t}(x)] \le \|\varphi(x)\|^2_{S_\varphi(P_t,\sigma/T)^{-1}}.$$

Using the Freedman inequality (Lemma D.2) on $\{z_{t,x}\}_{t\in\mathcal{I}}$ with $\xi = \frac{\mu_j}{4\gamma_{\varphi,T}} = \xi_j$, we get with probability at least $1 - \frac{\delta}{CN}$ that

$$\begin{aligned}
\widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x) - \mathcal{R}_\mathcal{I}(x) &= \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}(z_{t,x} + \mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi,t}(x)] - \mathcal{R}_\mathcal{I}(x)) \\
&\le \frac{\xi_j}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\|\varphi(x)\|^2_{S_\varphi(P_t,\sigma/T)^{-1}} + \frac{\log(CN/\delta)}{\xi_j|\mathcal{I}|} + \frac{\sqrt{\sigma/T}}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\|\varphi(x)\|_{S_\varphi(P_t,\sigma/T)^{-1}}
\end{aligned}$$

where we use Lemma D.3 to bound the bias term $\mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi,t}(x)] - \mathcal{R}_\mathcal{I}(x)$. A union bound over all $x \in \mathcal{X}$ and the reverse case $\mathcal{R}_\mathcal{I}(x) - \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x)$ completes the proof. ■

Choosing $C = C_0 = 8T\log_2 T$, we get by a union bound that for all intervals of sizes $E, 2E, 2^2E, \ldots$ and $(2^2-1)E, (2^3-1)E, \ldots$, the concentration bound in Lemma D.1 holds with probability at least $1-\delta$. For ease of exposition, we define the following event.

Definition D.4 (EVENT_1)**.** *Denote by* EVENT_1 *the event that*

$$|\widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x) - \mathcal{R}_{\mathcal{I}}(x)| \le \frac{\xi_j}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} \|\varphi(x)\|^2_{S_\varphi(P_t,\sigma/T)^{-1}} + \frac{\log(C_0N/\delta)}{\xi_j|\mathcal{I}|} + \frac{\sqrt{\sigma/T}}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} \|\varphi(x)\|_{S_\varphi(P_t,\sigma/T)^{-1}}$$

holds for all intervals $\mathcal{I}\subset[T]$ *of sizes* 2^jE *for all* $j=0,1,\dots$ *and* $(2^j-1)E$ *for all* $j=1,2,\dots$.

By the previous argument, EVENT_1 holds with probability at least $1-\delta$.

D.5 Proof of Lemma 4.5

The following lemma bounds the optimality gaps of an action in two intervals by the total variation of the reward function throughout an interval that spans the two intervals. The proof is adapted from Lemma 13 by Luo et al. (2018) and Lemma 8 by Chen et al. (2019).

Lemma D.5. *For any interval* $\mathcal{I}$*, any of its sub-intervals* $\mathcal{I}_1,\mathcal{I}_2\subseteq\mathcal{I}$ *and any* $x\in\mathcal{X}$*, we have*

$$|\Delta_{\mathcal{I}_1}(x) - \Delta_{\mathcal{I}_2}(x)| \le 2V_{\mathcal{I}}.$$

Proof. For all $x\in\mathcal{X}$, we have

$$\begin{aligned}
|\mathcal{R}_{\mathcal{I}_1}(x) - \mathcal{R}_{\mathcal{I}_2}(x)| &= \left|\frac{1}{|\mathcal{I}_1|}\sum_{s\in\mathcal{I}_1} r_s(x) - \frac{1}{|\mathcal{I}_2|}\sum_{t\in\mathcal{I}_2} r_t(x)\right| \\
&= \frac{1}{|\mathcal{I}_1||\mathcal{I}_2|}\left|\sum_{s\in\mathcal{I}_1}\sum_{t\in\mathcal{I}_2}(r_s(x)-r_t(x))\right| \\
&\le \frac{1}{|\mathcal{I}_1||\mathcal{I}_2|}\sum_{s\in\mathcal{I}_1}\sum_{t\in\mathcal{I}_2}|r_s(x)-r_t(x)| \le V_{\mathcal{I}}
\end{aligned}$$

where the last inequality follows since $|r_s(x)-r_t(x)|\le\sum_{\tau=s}^{t-1}|r_{\tau+1}(x)-r_\tau(x)|\le V_{\mathcal{I}}$. Hence,

$$-V_{\mathcal{I}} \le \mathcal{R}_{\mathcal{I}_1}(x) - \mathcal{R}_{\mathcal{I}_2}(x), \mathcal{R}_{\mathcal{I}_1}(x^\star_{\mathcal{I}_1}) - \mathcal{R}_{\mathcal{I}_2}(x^\star_{\mathcal{I}_1}), \mathcal{R}_{\mathcal{I}_1}(x^\star_{\mathcal{I}_2}) - \mathcal{R}_{\mathcal{I}_2}(x^\star_{\mathcal{I}_2}) \le V_{\mathcal{I}}$$

where we use the notation $x^\star_{\mathcal{I}} = \text{argmax}_{x'\in\mathcal{X}}\mathcal{R}_{\mathcal{I}}(x')$. It follows that

$$-V_{\mathcal{I}} \le \mathcal{R}_{\mathcal{I}_1}(x^\star_{\mathcal{I}_2}) - \mathcal{R}_{\mathcal{I}_2}(x^\star_{\mathcal{I}_2}) \le \mathcal{R}_{\mathcal{I}_1}(x^\star_{\mathcal{I}_1}) - \mathcal{R}_{\mathcal{I}_2}(x^\star_{\mathcal{I}_2}) \le \mathcal{R}_{\mathcal{I}_1}(x^\star_{\mathcal{I}_1}) - \mathcal{R}_{\mathcal{I}_2}(x^\star_{\mathcal{I}_1}) \le V_{\mathcal{I}}$$

where we use the optimality of $x^\star_{\mathcal{I}_1}$ and $x^\star_{\mathcal{I}_2}$. Hence, for all $x\in\mathcal{X}$,

$$\begin{aligned}
|\Delta_{\mathcal{I}_1}(x) - \Delta_{\mathcal{I}_2}(x)| &= |\mathcal{R}_{\mathcal{I}_1}(x^\star_{\mathcal{I}_1}) - \mathcal{R}_{\mathcal{I}_1}(x) - \mathcal{R}_{\mathcal{I}_2}(x^\star_{\mathcal{I}_2}) + \mathcal{R}_{\mathcal{I}_2}(x)| \\
&\le |\mathcal{R}_{\mathcal{I}_1}(x^\star_{\mathcal{I}_1}) - \mathcal{R}_{\mathcal{I}_2}(x^\star_{\mathcal{I}_2})| + |\mathcal{R}_{\mathcal{I}_1}(x) - \mathcal{R}_{\mathcal{I}_2}(x)| \le 2V_{\mathcal{I}}.
\end{aligned}$$

■

Now, we are ready to prove Lemma 4.5.

Proof of Lemma 4.5. Assume that the event EVENT_1 holds. We prove by induction on the block index j. For the base case $j=0$, note that the strategy used in block $\mathcal{B}(0)$ is $\pi_\varphi(\mathcal{X})$. Under the event EVENT_1, using the result $\|\varphi(x)\|^2_{S_\varphi(\pi_\varphi(\mathcal{X}),\sigma/T)^{-1}} \le \gamma_{\varphi,T}$ from Lemma 4.3 gives

$$|\widehat{\mathcal{R}}_{\varphi,\mathcal{B}(0)}(x) - \mathcal{R}_{\mathcal{B}(0)}(x)| \le \xi_0\gamma_{\varphi,T} + \frac{\log(C_0N/\delta)}{\xi_0|\mathcal{B}(0)|} + \sqrt{\frac{\sigma\gamma_{\varphi,T}}{T}} \le \frac{c_0}{4}\mu_0$$

where the last inequality follows by $\xi_0 = \frac{1}{8\gamma_{\varphi,T}}$, $|\mathcal{B}(0)| = E \ge 4\gamma_{\varphi,T}\log(C_0N/\delta)$ and $\sqrt{\sigma\gamma_{\varphi,T}/T} \le 2\sqrt{\alpha}\mu_0$. This proves the base case for the bound (5).

Now, suppose the bound (5) holds for the block indices 0, 1, ..., j. Then, for any $m = 0, \dots, j$, using the notations $x^\star = \operatorname{argmax}_{x\in\mathcal{X}} \mathcal{R}_{\mathcal{C}(m)}(x)$ and $\hat{x} = \operatorname{argmax}_{x\in\mathcal{X}} \widehat{\mathcal{R}}_{\varphi,\mathcal{C}(m)}(x)$, we have

$$\begin{aligned}\Delta_{\mathcal{C}(m)}(x) - \widehat{\Delta}_{\varphi,\mathcal{C}(m)}(x) &= \mathcal{R}_{\mathcal{C}(m)}(x^\star) - \mathcal{R}_{\mathcal{C}(m)}(x) - \widehat{\mathcal{R}}_{\varphi,\mathcal{C}(m)}(\hat{x}) + \widehat{\mathcal{R}}_{\varphi,\mathcal{C}(m)}(x) \\ &\le \mathcal{R}_{\mathcal{C}(m)}(x^\star) - \mathcal{R}_{\mathcal{C}(m)}(x) - \widehat{\mathcal{R}}_{\varphi,\mathcal{C}(m)}(x^\star) + \widehat{\mathcal{R}}_{\varphi,\mathcal{C}(m)}(x) \\ &\le \frac{1}{2}\Delta_{\mathcal{C}(m)}(x) + 2V_{\mathcal{C}(m)} + \frac{c_0}{2}\mu_m\end{aligned}$$

where the first inequality uses the optimality of $\hat{x}$, and the second inequality uses the induction hypothesis and the fact that $\Delta_{\mathcal{C}(m)}(x^\star) = 0$. Rearranging gives the bound (6) for the blocks $0, \dots, j$. Similarly, for $m = 0, \dots, j$, we have

$$\begin{aligned}\widehat{\Delta}_{\varphi,\mathcal{C}(m)}(x) - \Delta_{\mathcal{C}(m)}(x) &\le \widehat{\mathcal{R}}_{\varphi,\mathcal{C}(m)}(\hat{x}) - \widehat{\mathcal{R}}_{\varphi,\mathcal{C}(m)}(x) - \mathcal{R}_{\mathcal{C}(m)}(\hat{x}) + \mathcal{R}_{\mathcal{C}(m)}(x) \\ &\le \frac{1}{2}\Delta_{\mathcal{C}(m)}(\hat{x}) + \frac{1}{2}\Delta_{\mathcal{C}(m)}(x) + 2V_{\mathcal{C}(m)} + \frac{c_0}{2}\mu_m \\ &\le \frac{1}{2}\Delta_{\mathcal{C}(m)}(x) + 4V_{\mathcal{C}(m)} + c_0\mu_m\end{aligned}$$

where the first inequality uses the optimality of $x^\star$, the second inequality uses the induction hypothesis and the last inequality uses the bound (6) we showed and the optimality of $\hat{x}$ to bound $\Delta_{\mathcal{C}(m)}(\hat{x}) \le 2\widehat{\Delta}_{\varphi,\mathcal{C}(j)}(x) + 4V_{\mathcal{C}(j)} + c_0\mu_j = 4V_{\mathcal{C}(j)} + c_0\mu_j$. Rearranging gives the bound (7) for the blocks $0, \dots, j$.

Now, for the block index $j+1$, $\textsc{Event}_1$ gives

$$\begin{aligned}|\widehat{\mathcal{R}}_{\varphi,\mathcal{C}(j+1)}(x) - \mathcal{R}_{\mathcal{C}(j+1)}(x)| \le& \frac{\xi_{j+1}}{|\mathcal{C}(j+1)|}\sum_{t\in\mathcal{C}(j+1)} \|\varphi(x)\|^2_{S_\varphi(P_t,\sigma/T)^{-1}} + \frac{\log(CN/\delta)}{\xi_{j+1}|\mathcal{C}(j+1)|} \\ &+ \frac{\sqrt{\sigma/T}}{|\mathcal{C}(j+1)|}\sum_{t\in\mathcal{C}(j+1)} \|\varphi(x)\|_{S_\varphi(P_t,\sigma/T)^{-1}}. \qquad (10)\end{aligned}$$

To bound the first term, we use Lemma 4.4 and the bound (7) we showed for blocks $0, \dots, j$ to get

$$\begin{aligned}\xi_{j+1}\|\varphi(x)\|^2_{S_\varphi(P_t,\sigma/T)^{-1}} &\le 2\xi_{j+1}\|\varphi(x)\|^2_{S_\varphi(Q^{(m_t)},\sigma/T)^{-1}} \\ &\le 2\xi_{j+1}(\beta_{m_t}\widehat{\Delta}_{\varphi,\mathcal{C}(m_t-1)}(x) + 2\gamma_{\varphi,T}) \\ &\le 2\xi_{j+1}(\beta_{m_t}(2\Delta_{\mathcal{C}(m_t-1)}(x) + 4V_{\mathcal{C}(m_t-1)} + c_0\mu_{m_t-1}) + 2\gamma_{\varphi,T}) \\ &\le \frac{1}{20}\Delta_{\mathcal{C}(m_t-1)}(x) + \frac{1}{10}V_{\mathcal{C}(m_t-1)} + \frac{3}{2}\mu_{j+1} \qquad (11) \\ &\le \frac{1}{20}\Delta_{\mathcal{C}(j+1)}(x) + \frac{1}{5}V_{\mathcal{C}(j+1)} + 2\mu_{j+1}\end{aligned}$$

where the second to last inequality follows by a simple calculation using identities in Section D.1 and the fact that $m_t \le j+1$ for $t \in \mathcal{C}(j+1)$ and the last inequality follows by Lemma D.5.

The second term can be bounded by

$$\frac{\log(CN/\delta)}{\xi_{j+1}|\mathcal{C}(j+1)|} = \frac{4\gamma_{\varphi,T}\log(CN/\delta)}{\mu_{j+1}E\cdot 2^{j+1}} \le \frac{1}{\mu_{j+1}2^{j+1}} = 4\mu_{j+1}. \qquad (12)$$

The third term can be bounded using Lemma 4.4 and the bound (7):

$$\begin{aligned}\sqrt{\sigma/T}\|\varphi(x)\|_{S_\varphi(P_t,\sigma/T)^{-1}} &\le \sqrt{2\sigma/T}\|\varphi(x)\|_{S_\varphi(Q^{(m_t)},\sigma/T)^{-1}} \\ &\le \frac{\sqrt{\gamma}}{\sqrt{\alpha\gamma_{\varphi,T}T}}\beta_{m_t}\widehat{\Delta}_{\varphi,\mathcal{C}(m_t-1)}(x) + \frac{2\sqrt{\sigma\gamma_{\varphi,T}}}{\sqrt{T}} \\ &\le \frac{2\mu_{j+1}}{\gamma_{\varphi,T}}\beta_{m_t}(2\Delta_{\mathcal{C}(m_t-1)}(x) + 4V_{\mathcal{C}(m_t-1)} + c_0\mu_{m_t-1}) + 4\sqrt{\alpha}\mu_{j+1} \\ &\le \frac{1}{5}\Delta_{\mathcal{C}(m_t-1)}(x) + \frac{2}{5}V_{\mathcal{C}(m_t-1)} + (4+4\sqrt{\alpha})\mu_{j+1} \qquad (13) \\ &\le \frac{1}{5}\Delta_{\mathcal{C}(j+1)}(x) + \frac{4}{5}V_{\mathcal{C}(j+1)} + (4+4\sqrt{\alpha})\mu_{j+1}\end{aligned}$$

where the second inequality uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and the third inequality uses $\sqrt{\sigma\gamma_{\varphi,T}/T} \le 2\sqrt{\alpha}\mu_j$ for any block index j and the second to last inequality follows by a simple calculation and the last inequality follows by Lemma D.5.

Using these three bounds, we can further bound (10) by $|\widehat{\mathcal{R}}_{\varphi,\mathcal{C}(j+1)}(x) - \mathcal{R}_{\mathcal{C}}(j+1)(x)| \le \frac{1}{2}\Delta_{\mathcal{C}(j+1)}(x) + V_{\mathcal{C}(j+1)} + \frac{c_0}{4}\mu_{j+1}$, which proves the bound (5) for the block $j+1$. By induction, the proof is complete.

■

D.6 Proof of Theorem 4.6

Proof of Theorem 4.6. We bound the regret of each block $\mathcal{B}(j)$ separately. Using the Azuma-Hoeffding inequality on a martingale difference sequence $\{\mathbb{E}_t[r(x_t)] - r(x_t)\}_{t\in\mathbb{B}(j)}$, we get

$$\textsc{Reg}_{\mathcal{B}(j)} = \sum_{t\in\mathcal{B}(j)} (r(x^\star) - r(x_t)) \le \sum_{t\in\mathcal{B}(j)} (r(x^\star) - \mathbb{E}_t[r(x_t)]) + \widetilde{\mathcal{O}}(\sqrt{2^j E})$$

where we use $r(\cdot)$ to denote the stationary reward function and $x^\star = \operatorname{argmax}_{x\in\mathcal{X}} r(x)$. Since $P_t = (1-\mu_j)Q^{(j)} + \mu_j \pi_\varphi(\mathcal{X})$ for $t \in \mathcal{B}(j)$, using Lemma 4.5 with $V_{\mathcal{C}(j)} = 0$, we get with high probability that

$$r(x^\star) - \mathbb{E}_t[r(x_t)] = \sum_{x\in\mathcal{X}} P_t(x)\Delta_{\mathcal{C}(j-1)}(x) \le 2\sum_{x\in\mathcal{X}} Q^{(j)}(x)\widehat{\Delta}_{\varphi,\mathcal{C}(j-1)}(x) + \mathcal{O}(\mu_j) \le \mathcal{O}(\mu_j)$$

where the last inequality uses Lemma 4.4 and $1/\beta_j = \mathcal{O}(\mu_j)$. Summing over $t \in \mathcal{B}(j)$, we get $\textsc{Reg}_{\mathcal{B}(j)} \le \widetilde{\mathcal{O}}(E\sqrt{2^j})$. Summing over j and applying Cauchy-Schwarz, we get $\textsc{Reg}_T = \widetilde{\mathcal{O}}(E\sqrt{T/E}) = \widetilde{\mathcal{O}}(\sqrt{\gamma_T T \log N})$. ■

D.7 Subgaussian case

For the analysis with subgaussian noises, we can use the following modified Freedman-style inequality.

Lemma D.6. *Let $X_1, \dots, X_n \in \mathbb{R}$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_0, \mathcal{F}_1, \dots$. Assume X_i are σ-subguassian. Then for any $\delta \in (0,1)$ and $\xi \in [0, 1/\sqrt{2\sigma^2\log(n/\delta)}]$, we have with probability at least $1-2\delta$ that*

$$\sum_{i=1}^n X_i \le \xi V + \frac{\log(1/\delta)}{\xi},$$

where $V = \sum_{i=1}^n \mathbb{E}[X_i^2 \mid \mathcal{F}_{i-1}]$.

Proof. The proof closely follows the proof of the original Freedman-style inequality by Beygelzimer et al. (2011). Since $X_1, \dots, X_n$ are σ-subguassian, we have $X_t \le B = \sqrt{2\sigma^2\log(n/\delta)}$ for all $t = 1, \dots, n$ with probability at least $1-\delta$. Define $\widetilde{X}_t = \min\{X_t, B\}$ for $i = 1, \dots, n$. Then,

$$\mathbb{E}_t[\exp(\xi\widetilde{X}_t)] \le \mathbb{E}_t[1 + \xi\widetilde{X}_t + \xi^2\widetilde{X}_t^2] \le 1 + \xi^2\mathbb{E}_t[\widetilde{X}_t^2] \le \exp(\xi^2\mathbb{E}_t[\widetilde{X}_t^2]) \le \exp(\xi^2\mathbb{E}_t[X_t^2]) \tag{14}$$

where the first inequality uses the fact that $\xi \le 1/B$ and the identity $e^z \le 1 + z + z^2$ for $z \le 1$. Define $Z_0 = 1$ and $Z_t = Z_{t-1}\exp(\xi\widetilde{X}_t - \xi^2\mathbb{E}_t[X_t^2])$. Then,

$$\mathbb{E}_t[Z_t] = Z_{t-1}\exp(-\xi^2\mathbb{E}_t[X_t^2])\mathbb{E}_t[\exp(\xi\widetilde{X}_t)] \le 1$$

where the last inequality holds by (14). Hence, we have $\mathbb{E}[Z_n] \le 1$ and by Markov inequality, $P(Z_n \ge 1/\delta) \le \delta$. Note that by recursive definition, we have $Z_n = \exp(\xi\sum_{t=1}^n \widetilde{X}_t - \xi^2\sum_{t=1}^n \mathbb{E}_t X_t^2)$. Hence, $\sum_{t=1}^n \widetilde{X}_t \le \xi\sum_{t=1}^n \mathbb{E}_t X_t^2 + \log(1/\delta)/\xi$ with probability at least $1-\delta$.

By the previous argument that $X_t \le B$ for all $t = 1, \dots, n$ with probability at least $1-\delta$, we have $\sum X_t^2 = \sum \widetilde{X}_t^2$ with probability at least $1-\delta$. By a union bound, we have $\sum X_t^2 = \sum \widetilde{X}_t^2 \le \xi V + \log(1/\delta)/\xi$ with probability at least $1-2\delta$ as desired. ■

E MASTER reduction of GPUCB

Wei et al. (2021) introduce the MASTER reduction that converts a base algorithm into an algorithm that adapts to non-stationarity. They prove that if a base algorithm satisfies Condition E.1 for a constant ω, then the converted algorithm satisfies the dynamic regret bound displayed in Theorem E.2 without prior knowledge of the non-stationarity budgets.

Condition E.1 (Adapted from Assumption 1' in Wei et al. (2021))**.** *For any $t = 1, \dots, T$, as long as $\omega V_{[1,t]} \le \rho(t)$, the base algorithm can produce $\tilde{f}_t$ using history up to $t-1$ that satisfies*

$$\tilde{f}_t \ge \min_{\tau \in [1,t]} \max_{x \in \mathcal{X}} r_t(x) - \omega V_{[1,t]} \quad \text{and} \quad \frac{1}{t}\sum_{\tau=1}^{t}(\tilde{f}_\tau - y_\tau) \le c\rho(t) + c\omega V_{[1,t]}$$

with probability at least $1 - \frac{\delta}{T}$ where $\rho(t) \ge \frac{1}{\sqrt{t}}$, $t\rho(t)$ is non-decreasing in t, ω is some function of the parameters, and c is a universal constant.

Theorem E.2 (Adapted from Theorem 2 in Wei et al. (2021))**.** *If a base algorithm satisfies Condition E.1 with $t\rho(t) = g_1\sqrt{t} + g_2$, then the algorithm obtained by the MASTER reduction guarantees with high probability that*

$$\textsc{Reg}_T = \widetilde{\mathcal{O}}\left(\min\left\{(g_1 + g_1^{-1}g_2)\sqrt{L_T T}, (g_1^{2/3} + g_2 g_1^{-4/3})\omega^{1/3} V_T^{1/3} T^{2/3} + (g_1 + g_1^{-1}g_2)\sqrt{T}\right\}\right).$$

Now, we show that the GPUCB algorithm (Chowdhury et al. 2017) satisfies Condition E.1, and provide the resulting dynamic regret bounds.

The GPUCB algorithm (Algorithm 8) is a UCB-based algorithm for stationary kernel bandits introduced by Chowdhury et al. (2017). They use a surrogate prior model $GP(0, k(\cdot,\cdot))$ on f and use the posterior distribution $GP(\mu_t(\cdot), k_t(\cdot,\cdot))$ given observed rewards up to time t for designing the upper confidence bounds of reward estimates. It can be shown that

$$\mu_t(x) = \varphi(x)^T \Phi^T (\Phi\Phi^T + \lambda I)^{-1} y_{1:t}, \quad k_t(x,x') = k(x,x') - \varphi(x)^T \Phi^T (\Phi\Phi^T + \lambda I)^{-1} \Phi \varphi(x')$$

where φ is a feature mapping induced by the kernel k, $\Phi = [\varphi(x_1) \cdots \varphi(x_t)]^T$ and $y_{1:t} = (y_1, \dots, y_t)$.

Algorithm 8: GPUCB Chowdhury et al. 2017

Input: kernel k, confidence level $\delta \in (0,1)$, regularization parameter λ

for $t = 1, \dots, T$ **do**

  Set $\beta_t \leftarrow 1 + \sqrt{2(\gamma_{t-1} + 1 + \log(1/\delta))}$ and $\sigma_t^2 \leftarrow k_t(x,x)$

  Play $x_t = \operatorname{argmax}_{x \in \mathcal{X}} \mu_{t-1}(x) + \beta_t \sigma_{t-1}(x)$ and receive reward y_t.

The following lemma shows that GPUCB satisfies Condition E.1.

Lemma E.3. *The GPUCB algorithm satisfies Condition E.1 with $\tilde{f}_t = \max_{x \in \mathcal{X}}(\mu_{t-1}(x) + \beta_t \sigma_{t-1}(x))$, $\rho(t) = \beta_t \sqrt{\gamma_{t-1}\log(T/\delta)/t}$ and $\omega = \gamma_T \sqrt{\log(T/\delta)}$.*

Proof. Let $W_t \coloneqq \sum_{s=1}^{t} \varphi(x_s)\varphi(x_s)^T + \lambda I$. It can be shown that $\sigma_t(x) = \sqrt{k_t(x,x)} = \sqrt{\lambda}\|\varphi(x)\|_{W_t^{-1}}$. Following the proof of Lemma 1 in Zhou et al. (2021), we get

$$|r_t(x) - \mu_{t-1}(x)| \le \left|\varphi(x)^T W_{t-1}^{-1} \sum_{s=1}^{t-1} \varphi(x_s)\varphi(x_s)^T(\theta_t - \theta_s)\right| + \beta_t \|\varphi(x)\|_{W_{t-1}^{-1}}$$

Following the corrected version of the analysis for the reduction of OFUL in Wei et al. (2021), we get

$$\begin{aligned}
\left|\varphi(x)^T W_{t-1}^{-1} \sum_{s=1}^{t-1} \varphi(x_s)\varphi(x_s)^T(\theta_t - \theta_s)\right| &\le \sum_{s=1}^{t-1} |\varphi(x)^T W_{t-1}^{-1} \varphi(x_s)||\varphi(x_s)^T(\theta_t - \theta_s)| \\
&\le V_{[1,t]} \|\varphi(x)\|_{W_{t-1}^{-1}} \sum_{s=1}^{t-1} \|\varphi(x_s)\|_{W_{t-1}^{-1}} \\
&\le V_{[1,t]} \|\varphi(x)\|_{W_{t-1}^{-1}} \sqrt{(t-1)\sum_{s=1}^{t-1} \|\varphi(x_s)\|_{W_{t-1}^{-1}}^2} \\
&\le V_{[1,t]} \|\varphi(x)\|_{W_{t-1}^{-1}} \sqrt{t\gamma_{t-1}}
\end{aligned}$$

where the second inequality is by Cauchy-Schwarz, $|\theta_t - \theta_s| \le V_{[1,t]}$ and the assumption $\|\varphi(x_s)\| \le 1$. The third inequality is by Cauchy-Schwarz. The last inequality is by

$$\sum_{s=1}^{t-1} \|\varphi(x_s)\|^2_{W_{t-1}^{-1}} = \sum_{s=1}^{t-1} U(x_s)\|\varphi(x_s)\|^2_{S_\varphi(U,\sigma/(t-1))^{-1}} \le \gamma_{t-1} \tag{15}$$

where U is the uniform distribution on $\{x_1, \dots, x_{t-1}\}$ and the inequality is by Lemma C.4. Hence,

$$|r_t(x) - \mu_{t-1}(x)| \le (V_{[1,t]}\sqrt{t\gamma_{t-1}} + \beta_t)\|\varphi(x)\|_{W_{t-1}^{-1}} \le 2\beta_t\|\varphi(x)\|_{W_{t-1}^{-1}} = 2\beta_t\sigma_{t-1}(x)/\sqrt{\lambda}$$

where the last inequality uses $V_{[1,t]} \le \rho(t)/\omega \le \beta_t/\sqrt{t\gamma_T}$. Thus,

$$\begin{aligned}
\sum_{\tau=1}^{t}(\tilde{f}_\tau - y_\tau) &= \sum_{\tau=1}^{t}(\tilde{f}_\tau - r_\tau(x_\tau)) + \sum_{\tau=1}^{t}(r_\tau(x_\tau) - y_\tau) \\
&= \sum_{\tau=1}^{t}(\mu_{\tau-1}(x_\tau) - r_\tau(x_\tau)) + \sum_{\tau=1}^{t}\beta_\tau\sigma_{\tau-1}(x_\tau) + \mathcal{O}(\sqrt{t\log(T/\delta)}) \\
&= \mathcal{O}(\sum_{\tau=1}^{t}\beta_\tau\sigma_{\tau-1}(x_\tau) + \sqrt{t\log(T/\delta)}) \\
&= \mathcal{O}(\beta_t\sqrt{t\gamma_T\log(T/\delta)})
\end{aligned}$$

where the second equality uses the fact that $\tilde{f}_\tau = \mu_{\tau-1}(x_\tau) + \beta_\tau\sigma_{\tau-1}(x_\tau)$ due to the optimism principle of the algorithm. The last equality uses

$$\sum_{\tau=1}^{t}\beta_\tau\sigma_{\tau-1}(x_\tau) \le \beta_t\sum_{\tau=1}^{t}\sigma_{\tau-1}(x_\tau) \le \mathcal{O}(\beta_t\sqrt{t\gamma_t})$$

where the last inequality uses Lemma 4 in Chowdhury et al. (2017). This verifies the second condition in Condition E.1. Also,

$$\tilde{f}_t = \max_{x\in\mathcal{X}}(\mu_{t-1}(x) + \beta_t\sigma_{t-1}(x)) \ge \max_{x\in\mathcal{X}} r_t(x) \ge \min_{\tau\in[1,t]}\max_{x\in\mathcal{X}} r_\tau(x)$$

where the first inequality uses Theorem 2 in Chowdhury et al. (2017). This shows the first condition, completing the proof. ■

The previous lemma allows invoking the MASTER reduction for GPUCB, which gives a dynamic regret bound of

$$\textsc{Reg}_T \le \widetilde{\mathcal{O}}(\min\{\gamma_T\sqrt{L_T T}, \gamma_T V^{1/3}T^{2/3} + \gamma_T\sqrt{T}\}).$$

F Analysis of ADA-OPKB

For ease of exposition, we use the same set of parameters listed in Section D.1.

F.1 Change detection

In this subsection, we prove properties of the change detection rules used in ADA-OPKB.

Lemma F.1. *Assume the event* $\textsc{Event}_1$ *holds. Then, we have for any* $x \in \mathcal{X}$ *and replay interval* $(m, \mathcal{I})$ *that*

$$\begin{aligned}
\Delta_{\mathcal{I}}(x) &\le 2\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + c_0\mu_m + 4V_{[\tau_i,t]} \\
\widehat{\Delta}_{\varphi,\mathcal{I}}(x) &\le 2\Delta_{\mathcal{I}}(x) + c_0\mu_m + 4V_{[\tau_i,t]}
\end{aligned}$$

where τ_i *is the starting time of the epoch* i *in which* $\mathcal{I}$ *is scheduled and* t *is the end of the interval* $\mathcal{I}$.

Proof. Consider a replay interval $(m, \mathcal{I})$ scheduled in a block $\mathcal{B}(j)$ in epoch i and let τ_i be the starting time of the epoch i and t be the end time of $\mathcal{I}$. Following the calculation in (11) in the proof of Lemma 4.5, we get

$$\begin{aligned}\xi_m \|\varphi(x)\|^2_{S_\varphi(P_t,\sigma/T)^{-1}} &\le \frac{1}{20}\Delta_{\mathcal{C}(m_t-1)}(x) + \frac{1}{10}V_{\mathcal{C}(m_t-1)} + \frac{3}{2}\mu_m \\ &\le \frac{1}{20}\Delta_{\mathcal{I}}(x) + \frac{1}{5}V_{[\tau_i,t]} + 2\mu_m\end{aligned}$$

where the second inequality uses Lemma D.5 and the fact that both $\mathcal{C}(m_t - 1)$ and $\mathcal{I}$ lie in $[\tau_i, t]$. Likewise, following the calculation in (13) in the proof of Lemma 4.5 and using Lemma D.5, we get

$$\begin{aligned}\sqrt{\sigma/T}\|\varphi(x)\|_{S_\varphi(P_t,\sigma/T)^{-1}} &\le \frac{1}{5}\Delta_{\mathcal{C}(m_t-1)}(x) + \frac{2}{5}V_{\mathcal{C}(m_t-1)} + (4+4\sqrt{\alpha})\mu_m \\ &\le \frac{1}{5}\Delta_{\mathcal{I}}(x) + \frac{4}{5}V_{[\tau_i,t]} + (4+4\sqrt{\alpha})\mu_m.\end{aligned}$$

Note that m is the maximum strategy index used in $\mathcal{I}$ due to the index selection logic in Line 9 in Algorithm 3. Hence, under the event $\textsc{Event}_1$, the two bounds above and the bound (12) give

$$\begin{aligned}&|\widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x) - \mathcal{R}_{\mathcal{I}}(x)| \\ &\quad\le \frac{\xi_m}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\|\varphi(x)\|^2_{S_\varphi(P_t,\sigma/T)^{-1}} + \frac{\log(CN/\delta)}{\xi_m|\mathcal{I}|} + \frac{\sqrt{\sigma/T}}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\|\varphi(x)\|_{S_\varphi(P_t,\sigma/T)^{-1}} \\ &\quad\le \frac{1}{2}\Delta_{\mathcal{I}}(x) + V_{[\tau_i,t]} + \frac{c_0}{4}\mu_m. \qquad (16)\end{aligned}$$

Denoting $\hat{x} = \operatorname{argmax}_{x'\in\mathcal{X}} \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x')$ and $x^\star = \operatorname{argmax}_{x'\in\mathcal{X}} \mathcal{R}_{\mathcal{I}}(x)$, we have

$$\begin{aligned}\Delta_{\mathcal{I}}(x) - \widehat{\Delta}_{\varphi,\mathcal{I}}(x) &= \mathcal{R}_{\mathcal{I}}(x^\star) - \mathcal{R}_{\mathcal{I}}(x) - \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(\hat{x}) + \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x) \\ &\le \mathcal{R}_{\mathcal{I}}(x^\star) - \mathcal{R}_{\mathcal{I}}(x) - \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x^\star) + \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x) \\ &\le \frac{1}{2}\Delta_{\mathcal{I}}(x) + 2V_{[\tau_i,t]} + \frac{c_0}{2}\mu_m\end{aligned}$$

where the first inequality uses the optimality of $\hat{x}$ and the second inequality uses the bound (16) and $\Delta_{\mathcal{I}}(x^\star) = 0$. Rearranging proves the first inequality of the lemma. The second inequality can be shown by

$$\begin{aligned}\widehat{\Delta}_{\varphi,\mathcal{I}}(x) - \Delta_{\mathcal{I}}(x) &\le \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(\hat{x}) - \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x) - \mathcal{R}_{\mathcal{I}}(\hat{x}) + \mathcal{R}_{\mathcal{I}}(x) \\ &\le \frac{1}{2}\Delta_{\mathcal{I}}(\hat{x}) + \frac{1}{2}\Delta_{\mathcal{I}}(x) + 2V_{[\tau_i,t]} + \frac{c_0}{2}\mu_m \\ &\le \frac{1}{2}\Delta_{\mathcal{I}}(x) + 4V_{[\tau_i,t]} + c_0\mu_m\end{aligned}$$

where the first inequality uses the optimality of $x^\star$, the second inequality uses the bound (16) and the last inequality uses the first inequality of the lemma. Rearranging proves the second inequality of the lemma.

■

Lemma F.2. *Let $(m, \mathcal{I})$ be a replay interval scheduled in $\mathcal{S}$ for block j in some epoch i. If no restart is triggered by this replay interval when performing the change detection test at the end of $\mathcal{I}$, we have with probability at least $1 - \delta$ for all $x \in \mathcal{X}$ that*

$$\begin{aligned}\widehat{\Delta}_{\varphi,\mathcal{I}}(x) &\le 2\Delta_{\mathcal{I}}(x) + 4c_0\mu_m \\ \Delta_{\mathcal{I}}(x) &\le 2\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 4c_0\mu_m\end{aligned}$$

Proof. Suppose no restart is triggered by the test (8) for $(m, \mathcal{I})$. Then, $\widehat{\Delta}_{\varphi,\mathcal{I}}(x) - 4\widehat{\Delta}_{\varphi,\mathcal{C}(k)}(x) \le 4c_0\mu_{m\wedge k}$ and

$\widehat{\Delta}_{\varphi,\mathcal{C}(k)}(x) - 4\widehat{\Delta}_{\varphi,\mathcal{I}}(x) \le 4c_0\mu_{m\wedge k}$ for all $k = 0, \dots, j-1$. Hence, for $t \in \mathcal{I}$, we have

$$\begin{aligned}
\xi_m \|\varphi(x)\|^2_{S_\varphi(P_t,\sigma/T)^{-1}} &\le 2\xi_m \|\varphi(x)\|^2_{S_\varphi(Q^{(m_t)},\sigma/T)^{-1}} \\
&\le 2\xi_m(\beta_{m_t}\widehat{\Delta}_{\varphi,\mathcal{C}(m_t-1)}(x) + 2\gamma_{\varphi,T}) \\
&\le 2\xi_m(\beta_{m_t}(4\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 4c_0\mu_{(m_t-1)\wedge m}) + 2\gamma_{\varphi,T}) \\
&\le 8\xi_m\beta_m\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 8\sqrt{2}c_0\xi_m\beta_{m_t}\mu_{m_t} + 4\xi_m\gamma_{\varphi,T} \\
&\le \frac{1}{10}\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 10\mu_m
\end{aligned}$$

where the first inequality uses $\frac{1}{2}S_\varphi(Q^{(m_t)}, \sigma/T) \preccurlyeq S_\varphi(P_t, \sigma/T)$, the second inequality uses Lemma 4.4, the fourth inequality uses $m_t \le m$ and $\mu_{m_t-1} = \sqrt{2}\mu_{m_t}$. The last inequality holds by simple calculation. Similarly,

$$\begin{aligned}
\sqrt{\sigma/T}\|\varphi(x)\|_{S_\varphi(P_t,\sigma/T)^{-1}} &\le \sqrt{2\sigma/T}\|\varphi(x)\|_{S_\varphi(Q^{(m_t)},\sigma/T)^{-1}} \\
&\le \frac{\sqrt{\sigma}}{\sqrt{\alpha\gamma_{\varphi,T}T}}\beta_{m_t}\widehat{\Delta}_{\varphi,\mathcal{C}(m_t-1)}(x) + \frac{2\sqrt{\sigma\gamma_{\varphi,T}}}{\sqrt{T}} \\
&\le \frac{2\mu_m}{\gamma_{\varphi,T}}\beta_{m_t}(4\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 4c_0\mu_{(m_t-1)\wedge m}) + 4\sqrt{\alpha}\mu_m \\
&\le \frac{8\mu_m\beta_m}{\gamma_{\varphi,T}}\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + \frac{8\sqrt{2}c_0\mu_m\beta_{m_t}\mu_{m_t}}{\gamma_{\varphi,T}} + 4\sqrt{\alpha}\mu_m \\
&\le \frac{2}{5}\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 24\mu_m + 4\sqrt{\alpha}\mu_m.
\end{aligned}$$

where the second inequality uses Lemma 4.4 and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, the third inequality uses $\sqrt{\sigma\gamma_{\varphi,T}/T} \le 2\sqrt{\alpha}\mu_j$

Under the event EVENT_1, the two bounds above and the bound (12) give

$$\begin{aligned}
&|\widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x) - \mathcal{R}_{\mathcal{I}}(x)| \\
&\quad\le \frac{\xi_m}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\|\varphi(x)\|^2_{S_\varphi(P_t,\sigma/T)^{-1}} + \frac{\log(CN/\delta)}{\xi_m|\mathcal{I}|} + \frac{\sqrt{\sigma/T}}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\|\varphi(x)\|_{S_\varphi(P_t,\sigma/T)^{-1}} \\
&\quad\le \frac{1}{2}\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 38\mu_m + 4\sqrt{\alpha}\mu_m \le \frac{1}{2}\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + c_0\mu_m
\end{aligned} \tag{17}$$

Denoting $\hat{x} = \operatorname{argmax}_{x'\in\mathcal{X}}\widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x')$ and $x^\star = \operatorname{argmax}_{x'\in\mathcal{X}}\mathcal{R}_{\mathcal{I}}(x)$, we have

$$\widehat{\Delta}_{\varphi,\mathcal{I}}(x) - \Delta_{\mathcal{I}}(x) \le \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(\hat{x}) - \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x) - \mathcal{R}_{\mathcal{I}}(\hat{x}) + \mathcal{R}_{\mathcal{I}}(x) \le \frac{1}{2}\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 2c_0\mu_m$$

where the first inequality uses the optimality of $x^\star$ and the second inequality uses $\widehat{\Delta}_{\varphi,\mathcal{I}}(\hat{x}) = 0$. Rearranging gives the first inequality of the lemma. Using this result, we get

$$\begin{aligned}
\Delta_{\mathcal{I}}(x) - \widehat{\Delta}_{\varphi,\mathcal{I}}(x) &= \mathcal{R}_{\mathcal{I}}(x^\star) - \mathcal{R}_{\mathcal{I}}(x) - \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(\hat{x}) + \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x) \\
&\le \mathcal{R}_{\mathcal{I}}(x^\star) - \mathcal{R}_{\mathcal{I}}(x) - \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x^\star) + \widehat{\mathcal{R}}_{\varphi,\mathcal{I}}(x) \\
&\le \frac{1}{2}\widehat{\Delta}_{\varphi,\mathcal{I}}(x^\star) + \frac{1}{2}\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 2c_0\mu_m \\
&\le \frac{1}{2}(2\Delta_{\mathcal{I}}(x^\star) + 4c_0\mu_m) + \frac{1}{2}\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 2c_0\mu_m \\
&= \frac{1}{2}\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 4c_0\mu_m
\end{aligned}$$

where the first inequality uses the optimality of $\hat{x}$ and the second inequality uses the bound (17) and the last equality uses $\Delta_{\mathcal{I}}(x^\star) = 0$. Rearranging gives the second inequality of the lemma.

■

For the rest of the analysis, we define $\mu_{\mathcal{I}} \coloneqq c_1(|\mathcal{I}|/E)^{-1/2}$ so that $\mu_j = \mu_{\mathcal{B}(j)}$.

Lemma F.3. *Assume the event* $\textsc{Event}_1$ *holds. Consider an epoch* i *that starts at time* τ_i*. If* $V_{[\tau_i,t]} \le \mu_{[\tau_i,t]}$ *holds for some time* $t \ge \tau_i$*, then no restart is triggered in* $[\tau_i, t]$*.*

Proof. It is enough to show that none of the end of replay intervals that lie within $[\tau_i, t]$ trigger a restart when running the change detection test (8). Suppose $\mathcal{S}$ is the replay schedule in a block j. Suppose s is the end of a replay interval $(m, \mathcal{I}) \in \mathcal{S}$ with $\mathcal{I} \subseteq [\tau_i, t]$. Then, by Lemma 4.5 and Lemma F.1 (which hold under $\textsc{Event}_1$), we have for any $k < j$ that

$$\begin{aligned}
\widehat{\Delta}_{\varphi,\mathcal{I}}(x) &\le 2\Delta_{\mathcal{I}}(x) + c_0\mu_m + 4V_{[\tau_i,s]} \\
&\le 2\Delta_{\mathcal{C}(k)}(x) + c_0\mu_m + 8V_{[\tau_i,s]} \\
&\le 4\widehat{\Delta}_{\varphi,\mathcal{C}(k)}(x) + 8V_{\mathcal{C}(k)} + 2c_0\mu_k + c_0\mu_m + 8V_{[\tau_i,s]} \\
&\le 4\widehat{\Delta}_{\varphi,\mathcal{C}(k)}(x) + 3c_0\mu_{m\wedge k} + 16V_{[\tau_i,t]} \\
&\le 4\widehat{\Delta}_{\varphi,\mathcal{C}(k)}(x) + 4c_0\mu_{m\wedge k}
\end{aligned}$$

where the second inequality uses Lemma D.5 and the last inequality uses $V_{[\tau_i,t]} \le \mu_{[\tau_i,t]} \le \mu_m \le \mu_{m\wedge k}$. Similarly, we have

$$\begin{aligned}
\widehat{\Delta}_{\varphi,\mathcal{C}(k)}(x) &\le 2\Delta_{\mathcal{C}(k)}(x) + c_0\mu_k + 4V_{\mathcal{C}(k)} \\
&\le 2\Delta_{\mathcal{I}}(x) + c_0\mu_k + 8V_{[\tau_i,t]} \\
&\le 4\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 8V_{[\tau_i,s]} + 2c_0\mu_m + c_0\mu_k + 8V_{[\tau_i,t]} \\
&\le 4\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 3c_0\mu_{m\wedge k} + 16V_{[\tau_i,t]} \\
&\le 4\widehat{\Delta}_{\varphi,\mathcal{I}}(x) + 4c_0\mu_{m\wedge k}.
\end{aligned}$$

Hence, no restart is triggered by the replay interval $(m, \mathcal{I})$. Since this holds for any $(m, \mathcal{I}) \in \mathcal{S}$, proof is complete. ■

Definition F.4 (Excess regret)**.** *Let* $\mathcal{J}$ *be an interval, not necessarily a replay interval, that lies in a block* $\mathcal{B}(j)$ *with* $j \ge 1$ *in an epoch* i*. We define the excess regret of* $\mathcal{J}$ *with respect to a feature mapping* φ *as*

$$\zeta_{\varphi,\mathcal{J}} = \max_{x\in\mathcal{X}} \left(\Delta_{\mathcal{J}}(x) - 8\widehat{\Delta}_{\varphi,\mathcal{C}(j-1)}(x)\right).$$

Lemma F.5. *Assume* $\textsc{Event}_1$ *holds. Let* $\mathcal{J}$ *be an interval that lies within a block* $\mathcal{B}(j)$ *with* $V_{\mathcal{J}} \le \mu_{\mathcal{J}}$ *and* $\zeta_{\varphi,\mathcal{J}} > D_1\mu_{\mathcal{J}}$ *where* $D_1 = 25c_0$*. Then, there exists an index* $m^\star \in \{0, \dots, j\}$ *such that* $D_1\mu_{m^\star+1} < \zeta_{\varphi,\mathcal{J}} \le D_1\mu_{m^\star}$ *and* $2^{m^\star}E < |\mathcal{J}|$*. Moreover, any replay interval* $\mathcal{I}$ *of index* $m^\star$ *with* $\mathcal{I} \subseteq \mathcal{J}$ *triggers a restart.*

Proof. We show that there exists $m^\star$ such that $D_1\mu_{m^\star+1} < \zeta_{\varphi,\mathcal{J}} \le D_1\mu_{m^\star}$. By the definition of the excess regret, we have $\zeta_{\varphi,\mathcal{J}} \le \max_{x\in\mathcal{X}} \Delta_{\mathcal{J}}(x) \le 2 \le D_1\mu_0$. Also, by the assumption that $\zeta_{\varphi,\mathcal{J}} > D_1\mu_{\mathcal{J}} \ge D_1\mu_j$ where the last inequality follows since $\mathcal{J} \subseteq \mathcal{B}(j)$, we have $D_1\mu_j < \zeta_{\varphi,\mathcal{J}} \le D_1\mu_0$. It follows that there exists $m^\star \in \{0, \dots, j\}$ such that $D_1\mu_{m^\star+1} < \zeta_{\varphi,\mathcal{J}} \le D_1\mu_{m^\star}$. Also, such $m^\star$ satisfies $D_1\mu_{\mathcal{J}} < \zeta_{\varphi,\mathcal{J}} \le D_1\mu_{m^\star}$ and it follows that $|\mathcal{J}| > 2^{m^\star}E$ as desired.

Now, we show that any replay interval $\mathcal{I} \subseteq \mathcal{J}$ of index $m^\star$ determined above triggers a restart. We argue by contradiction. Suppose that no restart is triggered after running a replay interval $\mathcal{I} \subseteq \mathcal{J}$ of index $m^\star$. By the definition of the excess regret, there exists $x' \in \mathcal{X}$ such that $\zeta_{\varphi,\mathcal{J}} = \Delta_{\mathcal{J}}(x') - 8\widehat{\Delta}_{\varphi,\mathcal{C}(j-1)}(x')$. Hence, by Lemma D.5, we have

$$\begin{aligned}
\Delta_{\mathcal{I}}(x') &\ge \Delta_{\mathcal{J}}(x') - 2V_{\mathcal{J}} \\
&\ge 8\widehat{\Delta}_{\varphi,\mathcal{C}(j-1)}(x') + \zeta_{\varphi,\mathcal{J}} - 2\mu_{\mathcal{J}} \\
&> 8\widehat{\Delta}_{\varphi,\mathcal{C}(j-1)}(x') + D_1\mu_{m^\star+1} - 2\mu_{\mathcal{I}}.
\end{aligned}$$

Moreover, by Lemma F.2, we have $\Delta_{\mathcal{I}}(x') \le 2\widehat{\Delta}_{\varphi,\mathcal{I}}(x') + 4c_0\mu_{m^\star}$ under $\textsc{Event}_1$. Rearranging the lower bound and the upper bound of $\Delta_{\mathcal{I}}(x')$ we just found, we get

$$\widehat{\Delta}_{\varphi,\mathcal{I}}(x') > 4\widehat{\Delta}_{\varphi,\mathcal{C}(j-1)}(x') + \frac{D_1}{2}\mu_{m^\star+1} - 2c_0\mu_{m^\star} - \mu_{\mathcal{I}} \ge 4\widehat{\Delta}_{\varphi,\mathcal{C}(j-1)}(x') + 4c_0\mu_{m^\star}$$

which must have triggered a restart by the test (8). This contradicts the assumption that no restart is triggered, completing the proof. ■

F.2 Replay schedule

In this subsection, we analyze the behavior of the replay schedule. Consider a replay schedule $\mathcal{S}$ for a block $\mathcal{B}(j)$ in an epoch i. The following lemma shows that the sum of the errors μ_{m_t} over the block $\mathcal{B}(j)$ when following the schedule $\mathcal{S}$ is similar to the sum of the errors when using the latest strategy over the entire block.

Lemma F.6. *With probability at least $1-\delta$, for any block $\mathcal{B}(j)$ in any epoch i defined by ADA-OPKB, we have*

$$\sum_{t\in\mathcal{B}(j)} \mu_{m_t} = \widetilde{\mathcal{O}}(|\mathcal{B}(j)|\mu_j) = \widetilde{\mathcal{O}}(\sqrt{2^j}\gamma_T \log N).$$

Proof. Consider a block $\mathcal{B}(j)$ in an epoch i and its replay schedule $\mathcal{S}$. Then,

$$\sum_{t\in\mathcal{B}(j)} \mu_{m_t} = \mathcal{O}\left(\sum_{t\in\mathcal{B}(j)} 2^{-m_t/2}\right) = \mathcal{O}\left(\sum_{m=0}^{j} 2^{-m/2} \sum_{t\in\mathcal{B}(j)} \mathbb{I}\{m_t = m\}\right). \tag{18}$$

Note that the sum $\sum_{t\in\mathcal{B}(j)} \mathbb{I}\{m_t = m\}$ counts the number of times the replay index m is chosen when following the schedule $\mathcal{S}$. This sum is bounded by the sum of lengths of all replay intervals of index m in $\mathcal{S}$. Since a replay interval of index m has length $2^m E$, the maximum possible number of replay intervals of index m is $|\mathcal{B}(j)|/(2^m E) = 2^{j-m}$. Denote by $Z_k^{(m)}$, $k = 1, \dots, 2^{j-m}$ a Bernoulli random variable that indicates whether the k-th candidate replay interval of index m is scheduled in $\mathcal{S}$. By the replay scheduling algorithm (Algorithm 4) used by ADA-OPKB, $Z_k^{(m)}$ are independent with success probability $p = \sqrt{2^{m-j}}$. Hence, with probability at least $1 - \frac{\delta}{4T(\log_2 T)^2}$, we have

$$\sum_{t\in\mathcal{B}(j)} \mathbb{I}\{m_t = m\} \le (2^m E) \sum_{k=1}^{2^{j-m}} Z_k^{(m)} \le \widetilde{\mathcal{O}}(E\sqrt{2^{j+m}})$$

where we use the Hoeffding's inequality to bound

$$\sum_{k=1}^{2^{j-m}} Z_k^{(m)} \le 2^{j-m}p + \sqrt{\frac{2^{j-m}\log(T(\log_2 T)^2/\delta)}{2}} = \widetilde{\mathcal{O}}(\sqrt{2^{j-m}})$$

with probability at least $1 - \frac{\delta}{T(\log_2 T)^2}$. Applying a union bound over the possible choices of replay index m, we can further bound (18) with probability at least $1 - \frac{\delta}{T\log_2 T}$ by

$$\sum_{t\in\mathcal{B}(j)} \mu_{m_t} = \mathcal{O}\left(\sum_{m=0}^{j} 2^{-m/2} \sum_{t\in\mathcal{B}(j)} \mathbb{I}\{m_t = m\}\right) \le \widetilde{\mathcal{O}}\left(jE\sqrt{2^j}\right) \le \widetilde{\mathcal{O}}\left(\sqrt{2^j\gamma_T \log N}\right)$$

where we use the fact that the block index j is bounded by $\log_2(T/E)$. Applying a union bound over all possible choices of the starting time of $\mathcal{B}(j)$ and the block index j completes the proof. ∎

F.3 Regret of an interval

Lemma F.7. *With probability at least $1-\delta$, for all intervals $\mathcal{J} \subseteq [T]$, we have*

$$\sum_{t\in\mathcal{J}} (r_t(x_t^\star) - r_t(x_t)) \le \sum_{t\in\mathcal{J}} (r_t(x_t^\star) - \mathbb{E}_t[r_t(x_t)]) + \sqrt{8|\mathcal{J}|\log(T^2/\delta)} \tag{19}$$

where $x_t^\star = \operatorname{argmax}_{x\in\mathcal{X}} r_t(x)$.

Proof. The result follows by applying the Azuma-Hoeffding inequality on the martingale difference sequence $\{\mathbb{E}_t[r_t(x_t)] - r_t(x_t)\}_{t\in\mathcal{J}}$, using the fact that $|\mathbb{E}_t[r_t(x_t)] - r_t(x_t)| \le 2$. ∎

Definition F.8 (EVENT$_2$)**.** *Define* EVENT$_2$ *as the event that the bound (19) holds for all intervals $\mathcal{J} \subseteq [T]$.*

Lemma F.9. *With probability at least $1-\delta$, for any interval $\mathcal{J}$ that lies in any block $\mathcal{B}(j)$ with $j \geq 1$ in any epoch i, the regret in any sub-interval $\mathcal{J}' \subseteq \mathcal{J}$ is bounded by*

$$\textsc{Reg}_{\mathcal{J}'} \leq \mathcal{O}\left(\sum_{t\in\mathcal{J}'} \mu_{m_t} + |\mathcal{J}'|\mu_{\mathcal{J}'} + |\mathcal{J}'|V_{\mathcal{J}} + |\mathcal{J}'|\zeta_{\varphi,\mathcal{J}}\mathbb{I}\{\zeta_{\varphi,\mathcal{J}} > D_1\mu_{\mathcal{J}}\}\right)$$

where $D_1 = 25c_0$.

Proof. Fix epoch i and consider an interval $\mathcal{J}$ that lies within a block j. Under $\textsc{Event}_2$, we have

$$\begin{aligned}
\sum_{t\in\mathcal{J}'} (r_t(x_t^\star) - r_t(x_t)) &\leq \sum_{t\in\mathcal{J}'} (r_t(x_t^\star) - \mathbb{E}_t[r_t(x_t)]) + \sqrt{8|\mathcal{J}'|\log(4T^2/\delta)} \\
&= \sum_{t\in\mathcal{J}'}\sum_{x\in\mathcal{X}} P_t(x)\Delta_t(x) + \sqrt{8|\mathcal{J}'|\log(4T^2/\delta)} \\
&\leq \mathcal{O}\left(\sum_{t\in\mathcal{J}'}\sum_{x\in\mathcal{X}} Q^{(m_t)}(x)\Delta_t(x) + \sum_{t\in\mathcal{J}'}\mu_{m_t} + |\mathcal{J}'|\mu_{\mathcal{J}'}\right)
\end{aligned}$$

where the last inequality uses $\mu_{\mathcal{J}'} = \mathcal{O}(1/\sqrt{|\mathcal{J}'|/E}) = \mathcal{O}(1/\sqrt{|\mathcal{J}'|/\log(T/\delta)})$, $P_t = (1-\mu_{m_t})Q^{(m_t)} + \mu_{m_t}\pi_{\mathcal{X}}$ and $|\Delta_t(x)| \leq 2$. The first term in the bound above can be bounded by

$$\begin{aligned}
\sum_{x\in\mathcal{X}} Q^{(m_t)}(x)\Delta_t(x) &\leq \sum_{x\in\mathcal{X}} Q^{(m_t)}(x)\Delta_{\mathcal{J}}(x) + 2V_{\mathcal{J}} \\
&\leq \sum_{x\in\mathcal{X}} Q^{(m_t)}(x)\left(8\widehat{\Delta}_{\varphi,\mathcal{C}(j-1)}(x) + \zeta_{\varphi,\mathcal{J}}\right) + 2V_{\mathcal{J}} \\
&\leq 8\sum_{x\in\mathcal{X}} Q^{(m_t)}(x)\left(4\widehat{\Delta}_{\varphi,\mathcal{C}(m_t-1)}(x) + 4c\mu_{m_t-1}\right) + \zeta_{\varphi,\mathcal{J}} + 2V_{\mathcal{J}} \\
&\leq \mathcal{O}\left(\frac{(1+\alpha)\gamma_{\varphi,T}}{\beta_{m_t-1}} + \mu_{m_t-1} + \zeta_{\varphi,\mathcal{J}} + V_{\mathcal{J}}\right) \leq \mathcal{O}\left(\mu_{m_t} + \zeta_{\varphi,\mathcal{J}} + V_{\mathcal{J}}\right)
\end{aligned}$$

where the first inequality uses Lemma D.5, the second inequality uses the definition of $\zeta_{\varphi,\mathcal{J}}$ and the third inequality uses the fact that no restart is triggered by the block $\mathcal{B}(j-1)$. The second to last inequality uses Lemma 4.4. We can further bound the regret as

$$\sum_{t\in\mathcal{J}'} (r_t(x_t^\star) - r_t(x_t)) \leq \widetilde{\mathcal{O}}\left(\sum_{t\in\mathcal{J}'}\mu_{m_t} + |\mathcal{J}'|\zeta_{\varphi,\mathcal{J}} + |\mathcal{J}'|V_{\mathcal{J}} + |\mathcal{J}'|\mu_{\mathcal{J}'}\right).$$

Noting that $|\mathcal{J}'|\zeta_{\varphi,\mathcal{J}} \leq |\mathcal{J}'|\zeta_{\varphi,\mathcal{J}}\mathbb{I}\{\zeta_{\varphi,\mathcal{J}} > D_1\mu_{\mathcal{J}}\} + D_1\mu_{\mathcal{J}}|\mathcal{J}'|$ completes the proof. ■

F.4 Regret of a block

In this section, we fix a block $\mathcal{J}$ in an epoch i and bound its regret. The strategy is to partition the block into nearly-stationary intervals to use the interval regret bound we found in Lemma F.9, and argue that the change detection test does not allow the non-stationarity to accumulate without being detected. First, we show that given an arbitrary interval $\mathcal{J}$, we can partition it into nearly-stationary intervals $\mathcal{J}_1, \ldots, \mathcal{J}_\ell$ while controlling the size of the partition ℓ. For ease of exposition, we write $\gamma_T = \gamma_{\varphi,T}$.

Lemma F.10. *Given an interval $\mathcal{J}$, we can partition it into a set of intervals $\{\mathcal{J}_1, \ldots, \mathcal{J}_\ell\}$ such that $V_{\mathcal{J}_k} \leq \mu_{\mathcal{J}_k}$ for all $k = 1, \ldots, \ell$ and*

$$\ell \leq \min\left\{L_{\mathcal{J}}, \left(\frac{1}{2}\gamma_T\log(C_1N/\delta)\right)^{-1/3} V_{\mathcal{J}}^{2/3}|\mathcal{J}|^{1/3} + 1\right\}.$$

Proof. Following the same procedures described in the proof of Lemma 5 by Chen et al. (2019) and the proof of Lemma 19 by Wei et al. (2021), we partition $\mathcal{J}$ by taking intervals consecutively from the beginning of $\mathcal{J}$ in a greedy manner. Specifically, given that the first $k-1$ intervals we took are $\mathcal{J}_1 = [s_1, e_1], \ldots, \mathcal{J}_{k-1} = [s_{k-1}, e_{k-1}]$, we take the next interval $\mathcal{J}_k = [s_k, e_k]$ with $s_k = e_{k-1} + 1$ (or set s_k to the beginning of $\mathcal{J}$ if $k = 1$) that satisfies $V_{[s_k,e_k]} \leq \mu_{[s_k,e_k]}$

and $V_{[s_k,e_k]} > \mu_{[s_k,e_k+1]}$. In other words, $\mathcal{J}_k$ is the maximal interval that immediately follows $\mathcal{J}_{k-1}$ and satisfies $V_{[s_k,e_k]} \le \mu_{[s_k,e_k]}$. We repeat this greedy procedure until the end of $\mathcal{J}$ is reached.

We first show that the number of intervals ℓ in the partition obtained by the procedure must satisfy $\ell \le L_{\mathcal{J}}$. To see this, consider the partition $\{\mathcal{I}_1, \dots, \mathcal{I}_{L_{\mathcal{J}}}\}$ of $\mathcal{J}$ where each $\mathcal{I}_k$, $k = 1, \dots, L_{\mathcal{J}}$ are stationary, that is $V_{\mathcal{I}_k} = 0$. Then, each interval $\mathcal{J}_k$ must contain at least one end point of a stationary interval. Otherwise, $\mathcal{J}_k$ must end within some stationary interval $\mathcal{I}_{k'}$ and does not contain the end point of the stationary interval. This contradicts with the greedy procedure because when the procedure constructs $\mathcal{J}_k$, it must have taken time steps at least until the end point of the stationary interval $\mathcal{I}_{k'}$ since doing so does not affect $V_{\mathcal{J}_k}$. Also, each end point of the stationary interval is contained in exactly one of $\mathcal{J}_1, \dots, \mathcal{J}_\ell$. Hence, there is a surjection from $\{\mathcal{I}_1, \dots, \mathcal{I}_{L_{\mathcal{J}}}\}$ to $\{\mathcal{J}_1, \dots, \mathcal{J}_\ell\}$ and it follows that $\ell \le L_{\mathcal{J}}$.

Now, we show that $\ell \le (\frac{1}{2}\gamma_T \log(C_1 N/\delta))^{-1/3} V_{\mathcal{J}}^{2/3} |\mathcal{J}|^{1/3} + 1$. Recall that for any interval $\mathcal{I}$, $\mu_{\mathcal{I}}$ is defined as $\mu_{\mathcal{I}} = \frac{1}{2}\sqrt{E}|\mathcal{I}|^{-1/2}$ where $E = \lceil 4\gamma_T \log(C_1 N/\delta) \rceil$. Hence,

$$V_{\mathcal{J}} \ge \sum_{k=1}^{\ell-1} V_{[s_k,e_k]} > \sum_{k=1}^{\ell-1} \mu_{[s_k,e_k+1]} = \frac{\sqrt{E}}{2} \sum_{k=1}^{\ell-1} (|\mathcal{J}_k| + 1)^{-1/2} \ge \sqrt{\frac{1}{2}\gamma_T \log(C_1 N/\delta)} \sum_{k=1}^{\ell-1} |\mathcal{J}_k|^{-1/2}$$

where the second inequality follows by the greedy procedure and the last inequality follows since $(x+1)^{-1/2} \ge (2x)^{-1/2} = \frac{1}{\sqrt{2}} x^{-1/2}$ for all $x \ge 1$. By the Hölder's inequality, we have

$$\ell - 1 \le \left(\sum_{k=1}^{\ell-1} |\mathcal{J}_k|^{-1/2} \right)^{2/3} \left(\sum_{k=1}^{\ell-1} |\mathcal{J}_k| \right)^{1/3} \le \left(\frac{1}{2}\gamma_T \log(C_1 N/\delta) \right)^{-1/3} V_{\mathcal{J}}^{2/3} |\mathcal{J}|^{1/3}$$

and the desired bound for ℓ follows. This completes the proof. ■

Note that the block $\mathcal{B}(j)$ defined by ADA-OPKB spans exactly $2^j \cdot E$ time steps whether the block runs past the time horizon or a restart is triggered before the block ends. Denote by $\mathcal{B}'(j)$ the actual block run as part of epoch i before a restart is triggered or the time horizon is reached.

Lemma F.11. *Consider a block $\mathcal{B}(j)$ in an epoch i defined by ADA-OPKB. Let $\mathcal{B}'(j)$ be the actual block run as part of epoch i before a restart is triggered or the time horizon is reached. With probability at least $1 - 2\delta$, we have*

$$\textsc{Reg}_{\mathcal{B}'(j)} = \widetilde{\mathcal{O}}\left(\min\left\{ (\gamma_T \log N)\sqrt{2^j L_{\mathcal{B}'(j)}}, (\gamma_T \log N) V_{\mathcal{B}'(j)}^{1/3} (2^j)^{2/3} + (\gamma_T \log N)\sqrt{2^j} \right\} \right).$$

Proof. For ease of exposition, we suppress the subscript φ and write γ_T and $\zeta_{\mathcal{J}}$ instead of $\gamma_{\varphi,T}$ and $\zeta_{\varphi,\mathcal{J}}$. Assume $\textsc{Event}_1$ holds. Using the procedure described in Lemma F.10, we partition $\mathcal{B}(j)$ into $\mathcal{J}_1, \dots, \mathcal{J}_\ell$ such that $V_{\mathcal{J}_k} \le \mu_{\mathcal{J}_k}$ for all $k = 1, \dots, \ell$. Let $\mathcal{J}'_1, \dots, \mathcal{J}'_{\ell'}$ be the non-empty intervals $\mathcal{J}'_k = \mathcal{J}_k \cap \mathcal{B}'(j)$ that partition $\mathcal{B}'(j)$. Using the interval regret bound in Lemma F.9 with the fact that $\mathcal{J}'_k \subseteq \mathcal{J}_k$ and using $V_{\mathcal{J}_k} \le \mu_{\mathcal{J}_k} \le \mu_{\mathcal{J}'_k}$, we get

$$\textsc{Reg}_{\mathcal{B}'(j)} = \sum_{k=1}^{\ell'} \textsc{Reg}_{\mathcal{J}'_k} \le \mathcal{O}\left(\sum_{t \in \mathcal{B}'(j)} \mu_{m_t} + \sum_{k=1}^{\ell'} |\mathcal{J}'_k| \mu_{\mathcal{J}'_k} + \sum_{k=1}^{\ell'} |\mathcal{J}'_k| \zeta_{\mathcal{J}_k} \mathbb{I}\{\zeta_{\mathcal{J}_k} > D_1 \mu_{\mathcal{J}_k}\} \right). \tag{20}$$

The first term can be bounded using Lemma F.6 by $\sum_{t \in \mathcal{B}'(j)} \mu_{m_t} \le \widetilde{\mathcal{O}}(\sqrt{2^j}\gamma_T \log N)$ with probability at least $1 - \delta$.

The second term can be bounded using $\mu_{\mathcal{I}} = \mathcal{O}(\sqrt{E/|\mathcal{I}|})$ as

$$\sum_{k=1}^{\ell'} |\mathcal{J}'_k| \mu_{\mathcal{J}'_k} \le \mathcal{O}\left(\sum_{k=1}^{\ell'} \sqrt{|\mathcal{J}'_k| E} \right) \le \mathcal{O}\left(\sqrt{\ell' |\mathcal{B}'(j)| E} \right) \le \mathcal{O}\left(E\sqrt{\ell' 2^j} \right)$$

where the second inequality uses Cauchy-Schwarz and the last inequality uses $|\mathcal{B}'(j)| \le |\mathcal{B}(j)|$.

The third term is bounded using Lemma F.5 which shows that there exists $m_k^\star \in \{0, \dots, j\}$ with

$$2^{m_k^\star} < |\mathcal{J}_k|/E \quad \text{and} \quad D_1 \mu_{m_k^\star + 1} < \zeta_{\mathcal{J}_k} \le D_1 \mu_{m_k^\star} \tag{21}$$

such that running a replay interval of index $m_k^\star$ inside $\mathcal{J}_k$ triggers a restart. Denote by $n_k^{(m)}$ the number of replay intervals of index m that can be scheduled completely inside $\mathcal{J}_k'$. Then,

$$n_k^{(m)} \geq (|\mathcal{J}_k'| - 3 \cdot 2^m E)_+ / (2^m E) \tag{22}$$

for all $m = 0, \dots, j$ where $(\cdot)_+ = \max\{0, \cdot\}$. Hence,

$$\begin{aligned}|\mathcal{J}_k'|\zeta_{\mathcal{J}_k}\mathbb{I}\{\zeta_{\mathcal{J}_k} > D_1\mu_{\mathcal{J}_k}\} &\leq 3 \cdot 2^{m_k^\star} E D_1 \mu_{m_k^\star} + (|\mathcal{J}_k'| - 3 \cdot 2^{m_k^\star} E)_+ D_1 \mu_{m_k^\star} \mathbb{I}\{\zeta_{\mathcal{J}_k} > D_1\mu_{\mathcal{J}_k}\} \\ &\leq \mathcal{O}\left(\sqrt{E|\mathcal{J}_k|} + E\sqrt{2^{m_k^\star}} n_k^{(m_k^\star)} \mathbb{I}\{\zeta_{\mathcal{J}_k} > D_1\mu_{\mathcal{J}_k}\}\right)\end{aligned}$$

where the first inequality uses $\zeta_{\mathcal{J}_k} \leq D_1 \mu_{m_k^\star}$ stated in (21) and the second inequality uses $2^{m_k^\star}\mu_{m_k^\star} = \mathcal{O}(\sqrt{2^{m_k^\star}}) \leq \mathcal{O}(\sqrt{|\mathcal{J}_k|/E})$ which follows by (21) and the lower bound of $n_k^{(m)}$ shown in (22). Summing over $k = 1, \dots, \ell'$ and writing the event $\{\zeta_{\mathcal{J}_k} > D_1\mu_{\mathcal{J}_k}\}$ as A_k for convenience, we get

$$\begin{aligned}\sum_{k=1}^{\ell'} |\mathcal{J}_k'|\zeta_{\mathcal{J}_k}\mathbb{I}\{A_k\} &\leq \mathcal{O}\left(\sum_{k=1}^{\ell'} \sqrt{E|\mathcal{J}_k|} + E\sum_{k=1}^{\ell'} \sqrt{2^{m_k^\star}} n_k^{(m_k^\star)} \mathbb{I}\{A_k\}\right) \\ &\leq \mathcal{O}\left(E\sqrt{\ell' 2^j} + E\sum_{m=0}^{j} \sqrt{2^m} \sum_{k=1}^{\ell'} n_k^{(m)} \mathbb{I}\{A_k, m_k^\star = m\}\right)\end{aligned} \tag{23}$$

where the second inequality uses Cauchy-Schwarz and $\sum_{k=1}^{\ell'} |\mathcal{J}_k| \leq |\mathcal{B}(j)| = 2^j E$. Denoting by $Z_{k,l}^{(m)}$, $l = 1, \dots, n_k^{(m)}$ the Bernoulli random variable that indicates whether the l-th replay interval among the $n_k^{(m)}$ candidate replay intervals within $\mathcal{J}_k'$ is scheduled, we have

$$\begin{aligned}\sum_{k=1}^{\ell'} n_k^{(m)} \mathbb{I}\{A_k, m_k^\star = m\} &= \sum_{k=1}^{\ell'} n_k^{(m)} \mathbb{I}\{A_k, m_k^\star = m, Z_{k,1}^{(m)} = 0, \dots, Z_{k,n_k^{(m)}}^{(m)} = 0\} \\ &\leq \sum_{k=1}^{\ell'} n_k^{(m)} \mathbb{I}\{Z_{k,1}^{(m)} = 0, \dots, Z_{k,n_k^{(m)}}^{(m)} = 0\} \leq \widetilde{\mathcal{O}}(\sqrt{2^{j-m}})\end{aligned}$$

where the first equality follows under $\textsc{Event}_1$ by Lemma F.5 since if any of $Z_{k,l}^{(m)} = 1$ for $m = m_k^\star$, a restart must have been triggered before reaching the end of $\mathcal{J}_k'$. The last inequality follows since the second to last term is a geometric random variable with trials $Z_{1,1}^{(m)}, \dots, Z_{1,n_1^{(m)}}^{(m)}, \dots, Z_{\ell',n_{\ell'}^{(m)}}^{(m)}$ with success probability $\sqrt{2^{m-j}}$, which is bounded with probability at least $1 - \delta$ by $\widetilde{\mathcal{O}}(\sqrt{2^{j-m}})$. We can further bound the third term (23) by $\sum_{k=1}^{\ell'} |\mathcal{J}_k'|\zeta_{\mathcal{J}_k}\mathbb{I}\{A_k\} \leq \widetilde{\mathcal{O}}\left(E\sqrt{\ell' 2^j}\right)$ where we use $j \leq \log_2(T/E)$. Summing the three bounds we found for the terms in (20), we get $\textsc{Reg}_{\mathcal{B}'(j)} = \widetilde{\mathcal{O}}(E\sqrt{\ell' 2^j})$. Bounding ℓ' using Lemma F.10 completes the proof. ■

F.5 Proof of Theorem 3.1

Lemma F.12. *Assume* $\textsc{Event}_1$ *holds. The number of epochs H when running ADA-OPKB is bounded by*

$$H \leq \min\left\{L_T, \left(\frac{1}{2}\gamma_T \log(C_1 N/\delta)\right)^{-1/3} V_T^{2/3} T^{1/3} + 1\right\}.$$

Proof. Let $\{\mathcal{J}_k\}_{k=1}^{\ell}$ with $\ell \leq \min\left\{L_T, (\frac{1}{2}\gamma_T \log(C_0 N/\delta))^{-1/3} V_T^{2/3} T^{1/3} + 1\right\}$ be a partition of $[T]$ where $V_{\mathcal{J}_k} \leq \mu_{\mathcal{J}_k}$ for all $k = 1, \dots, \ell$. Such a partition exists by Lemma F.10. Let $\mathcal{E}_1, \dots, \mathcal{E}_H$ be all the intervals spanned by the epochs in $[1, T]$. Note that if an epoch i starts inside an interval $\mathcal{J}_k$, then the epoch must continue at least until the end of $\mathcal{J}_k$ since the total variation in $\mathcal{E}_i \cap \mathcal{J}_k$ is upper bounded by $V_{\mathcal{E}_i \cap \mathcal{J}_k} \leq V_{\mathcal{J}_k} \leq \mu_{\mathcal{J}_k} \leq \mu_{\mathcal{E}_i \cap \mathcal{J}_k}$ and no restart is triggered under $\textsc{Event}_1$ in $\mathcal{E}_i \cap \mathcal{J}_k$ due to Lemma F.3. Hence, each $\mathcal{E}_i$ contains the end point of at least one interval $\mathcal{J}_k$. Also, trivially, the end point of each $\mathcal{J}_k$ is contained in exactly one epoch. Hence, there is a surjection from $\{\mathcal{J}_1, \dots, \mathcal{J}_\ell\}$ to $\{\mathcal{E}_1, \dots, \mathcal{E}_H\}$ and it follows that $H \leq \ell$. This completes the proof. ■

Lemma F.13. *Given an epoch i in ADA-OPKB, let $\mathcal{E}_i$ be the interval spanned by the epoch. Then, with high probability, we have*

$$\textsc{Reg}_{\mathcal{E}_i} = \widetilde{\mathcal{O}}\left(\min\left\{\sqrt{\gamma_T L_{\mathcal{E}_i}|\mathcal{E}_i|\log N}, (\gamma_T V_{\mathcal{E}_i}\log N)^{1/3}|\mathcal{E}_i|^{2/3} + \sqrt{\gamma_T \log N|\mathcal{E}_i|}\right\}\right).$$

Proof. Let J_i be the index of the last block in epoch i. Then $\mathcal{E}_i = \cup_{j=0}^{J_i}\mathcal{B}'(j)$ where $\mathcal{B}'(j) = \mathcal{B}(j)\cap\mathcal{E}_i$ and $\mathcal{B}(j)$ is the j-th block defined by ADA-OPKB in epoch i. Since $|\mathcal{E}_i| = \sum_{j=0}^{J_i}|\mathcal{B}'(j)| = E(1+2+\cdots+2^{J_i-1}) + |\mathcal{B}'(J_i)| \geq (2^{J_i}-1)E$, we have $2^{J_i}-1 \leq |\mathcal{E}_i|/E$. Hence, using the regret bound of $\mathcal{B}'(j)$ in terms of $L_{\mathcal{B}'(j)}$ provided by Lemma F.11 and $\textsc{Reg}_{\mathcal{E}_i} = \sum_{j=0}^{J_i}\textsc{Reg}_{\mathcal{B}'(j)}$, we have with high probability that

$$\textsc{Reg}_{\mathcal{E}_i} \leq \widetilde{\mathcal{O}}\left(E\sum_{j=0}^{J_i}\sqrt{2^j L_{\mathcal{B}'(j)}}\right) \leq \widetilde{\mathcal{O}}\left(E\sqrt{2^{J_i}-1}\sqrt{L_{\mathcal{E}_i}+J_i}\right) \leq \widetilde{\mathcal{O}}\left(\sqrt{E|\mathcal{E}_i|L_{\mathcal{E}_i}}\right)$$

where the second inequality uses Cauchy-Schwarz and the fact that $\sum_{j=0}^{J_i}L_{\mathcal{B}'(j)} \leq L_{\mathcal{E}_i}+J_i$, and the last inequality uses $J_i \leq \log_2(T/E)$. This shows the first bound of the lemma.

To show the second bound in terms of $V_{\mathcal{E}_i}$, we use the regret bound of $\mathcal{B}'(j)$ in terms of $V_{\mathcal{B}'(j)}$ provided by Lemma F.11 to get with high probability that

$$\begin{aligned}\textsc{Reg}_{\mathcal{E}_i} &\leq \widetilde{\mathcal{O}}\left(E\sum_{j=0}^{J_i}V_{\mathcal{B}'(j)}^{1/3}(2^j)^{2/3}\right) + \widetilde{\mathcal{O}}\left(E\sum_{j=0}^{J_i}\sqrt{2^j}\right)\\ &\leq \widetilde{\mathcal{O}}\left(EV_{\mathcal{E}_i}^{1/3}(2^{J_i}-1)^{2/3}\right) + \widetilde{\mathcal{O}}\left(E\sqrt{J_i(2_i^J-1)}\right) \leq \widetilde{\mathcal{O}}\left(E^{1/3}V_{\mathcal{E}_i}^{1/3}|\mathcal{E}_i|^{2/3} + \sqrt{E|\mathcal{E}_i|}\right)\end{aligned}$$

where the second inequality uses the Hölder's inequality and the Cauchy-Schwarz inequality, and the third inequality uses the bound $2^{J_i}-1 \leq |\mathcal{E}_i|/E$ and $J_i \leq \log_2(T/E)$. This completes the proof. ■

Now, we are ready to prove Theorem 3.1. To bound the total dynamic regret, we bound the sum of the epoch regret bounds and use the bound on the number of epochs as shown below.

Proof of Theorem 3.1. Using the epoch regret bound in Lemma F.13 and the bound on the number of epochs H in Lemma F.12, we can bound $\textsc{Reg}_T = \sum_{i=1}^{H}\textsc{Reg}_{\mathcal{E}_i}$ as follows. First, using the epoch regret bound in terms of $L_{\mathcal{E}_i}$, we get with high probability that

$$\textsc{Reg}_T \leq \widetilde{\mathcal{O}}\left(\sqrt{E}\sum_{i=1}^{H}\sqrt{L_{\mathcal{E}_i}|\mathcal{E}_i|}\right) \leq \widetilde{\mathcal{O}}\left(\sqrt{E}\sqrt{L_T+H}\sqrt{T}\right) \leq \widetilde{\mathcal{O}}\left(\sqrt{EL_TT}\right)$$

where the second inequality uses Cauchy-Schwarz and the last inequality uses the bound $H \leq L_T$.

Now, using the epoch regret bound in terms of $V_{\mathcal{E}_i}$, we get

$$\textsc{Reg}_T \leq \widetilde{\mathcal{O}}\left(E^{1/3}\sum_{i=1}^{H}V_{\mathcal{E}_i}^{1/3}|\mathcal{E}_i|^{2/3} + \sqrt{E}\sum_{i=1}^{H}\sqrt{|\mathcal{E}_i|}\right) \leq \widetilde{\mathcal{O}}\left(E^{1/3}V_T^{1/3}T^{2/3} + \sqrt{EHT}\right)$$

where the second inequality uses the Hölder's inequality and the Cauchy-Schwarz inequality. Further bounding by $H \leq \mathcal{O}(1+E^{-1/3}V_T^{2/3}T^{1/3})$ completes the proof. ■

G Analysis of OPNN

In this section, we prove the following theorem that states a regret bound for the OPNN algorithm under the general stationary bandit setting.

Theorem G.1 (c.f. Theorem 4.6)**.** *Consider the general stationary bandit setting described in Section 2. Assume Assumption D and Assumption E hold. If we run the OPNN algorithm using a neural network with width m and depth L, the dynamic regret is bounded by*

$$\textsc{Reg}_T \leq \widetilde{\mathcal{O}}\left(\sqrt{\gamma_T T\log N}\right)$$

with probability at least $1-\delta$ where γ_T is the maximum information gain with respect to the neural tangent kernel of the neural network as long as $m \geq \mathrm{poly}(T, L, N, \lambda^{-1}, \lambda_0^{-1}, \log(1/\delta))$.

The key insight for the analysis of OPNN is that in the infinite network width regime, OPNN is equivalent to OPKB with the neural tangent kernel $\boldsymbol{H}$ defined as follows.

Definition G.2 (Jacot et al. (2018), Arora et al. (2019)). *Consider a fully connected neural network of depth L with the ReLU activation function σ. For all $a_i, a_j \in \mathcal{X}$, define covariance matrices $\Sigma^{(l)}$ and derivative covariance matrices $\dot{\Sigma}^{(l)}$ for $l = 0, \dots, L$ recursively as follows:*

$$\Sigma_{ij}^{(0)} = \langle a_i, a_j \rangle, \quad \boldsymbol{A}_{ij}^{(l)} = \begin{pmatrix} \Sigma_{ii}^{(l-1)} & \Sigma_{ii}^{(l-1)} \\ \Sigma_{ji}^{(l-1)} & \Sigma_{jj}^{(l-1)} \end{pmatrix}$$
$$\Sigma_{ij}^{(l)} = 2\mathbb{E}_{(u,v)\sim N(0, \boldsymbol{A}_{ij}^{(l)})}[\sigma(u), \sigma(v)]$$
$$\dot{\Sigma}_{ij}^{(l)} = 2\mathbb{E}_{(u,v)\sim N(0, \boldsymbol{A}_{ij}^{(l)})}[\dot{\sigma}(u), \dot{\sigma}(v)]$$

where $\dot{\sigma}$ is the derivative of the activation function. The neural tangent kernel $\boldsymbol{H}$ for the network is defined as

$$\boldsymbol{H}_{ij} = \sum_{l=1}^{L} \left(\Sigma_{ij}^{(l-1)} \cdot \prod_{l'=l}^{L} \dot{\Sigma}_{ij}^{(l')} \right).$$

For the analysis, we make the following technical assumptions.

Assumption D. *The neural tangent kernel matrix is positive definite with $\boldsymbol{H} \succcurlyeq \lambda_0 \boldsymbol{I}$ for some $\lambda_0 > 0$.*

The assumption that the neural tangent kernel matrix is positive definite is a mild assumption commonly made when analyzing neural networks (Du et al. 2019a; Arora et al. 2019). The assumption is satisfied, for example, as long as the actions are normalized to $\|a_i\|_2 = 1$ for all $i \in [N]$ and no two actions in $\mathcal{X}$ are parallel (Du et al. 2019b).

We impose regularity assumption on the reward functions as follows.

Assumption E. *For all $t \in [T]$, we have $\sqrt{2\boldsymbol{r}_t \boldsymbol{H}^{-1} \boldsymbol{r}_t} \le B$ for some constant B where $\boldsymbol{r}_t = (r_t(a_1), \dots, r_t(a_N))$ is the vector of reward function values at time t. We assume that the learner knows the upper bound B and scales the problem so that $\sqrt{2\boldsymbol{r}_t \boldsymbol{H}^{-1} \boldsymbol{r}_t} \le 1$ for all $t \in [T]$.*

This assumption is common in the neural bandits literature (Zhou et al. 2020; Zhang et al. 2020; Gu et al. 2021). As discussed by Zhou et al. (2020), if $\boldsymbol{r}_t$ lies in the RKHS $\mathcal{H}$ induced by the neural tangent kernel, the quantity $\sqrt{\boldsymbol{r}_t \boldsymbol{H}^{-1} \boldsymbol{r}_t}$ is upper bounded by the RKHS norm $\|\boldsymbol{r}_t\|_{\mathcal{H}}$. In this sense, the upper bound on $\sqrt{2\boldsymbol{r}_t \boldsymbol{H}^{-1} \boldsymbol{r}_t}$ imposes regularity on the reward functions.

G.1 NTK theory from previous work

We first review results related to the neural tangent kernel in previous work. The lemmas provided in this subsection are adapted from Zhou et al. (2020) which uses results in Allen-Zhu et al. (2019) and Arora et al. (2019).

Lemma G.3 (Lemma B.5 by Zhou et al. (2020)). *With high probability, we have*

$$\|g(x; \boldsymbol{W}) - g(x; \boldsymbol{W}^{(0)})\|_2 \le \mathcal{O}\left(\sqrt{\log m} T^{1/6} m^{-1/6} \lambda^{-1/6} L^3 \|g(x; \boldsymbol{W}^{(0)})\|_2 \right)$$

for all $\|\boldsymbol{W} - \boldsymbol{W}^{(0)}\|_2 \le 2\sqrt{T/(m\lambda)}$ as long as $m \ge \text{poly}(T, L, \lambda^{-1})$.

Lemma G.4 (Lemma B.6 by Zhou et al. (2020)). *With high probability, we have*

$$\|g(x; \boldsymbol{W})\|_2 \le \mathcal{O}(\sqrt{mL})$$

for all $\|\boldsymbol{W} - \boldsymbol{W}^{(0)}\|_2 \le 2\sqrt{T/(m\lambda)}$ and $x \in \mathcal{X}$ as long as $m \ge \text{poly}(T, L, \lambda^{-1})$.

Lemma G.5 (Lemma 5.2 by Zhou et al. (2020)). *Let W be a parameter trained by* TRAINNN *(Algorithm 6). Then, with probability at least $1 - \delta$, we have*

$$\|\boldsymbol{W} - \boldsymbol{W}^{(0)}\|_2 \le \mathcal{O}(\sqrt{T/(m\lambda)})$$

as long as $m \ge \text{poly}(T, L, \lambda^{-1}, \log(1/\delta))$.

Lemma G.6 (Lemma 5.1 by Zhou et al. (2020)). *With probability at least $1-\delta$, there exists $\boldsymbol{W}_t^\star \in \mathbb{R}^p$ such that*

$$r_t(x) = \langle g(x;\boldsymbol{W}^{(0)}), \boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)} \rangle \quad \text{and} \quad \sqrt{m}\|\boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}\|_2 \le \sqrt{2\boldsymbol{r}_t^T \boldsymbol{H}^{-1}\boldsymbol{r}_t} \le \sqrt{2N/\lambda_0}$$

for all $t \in [T]$ and $x \in \mathcal{X}$ as long as $m \ge \text{poly}(T, L, N, \lambda_0^{-1}, \log(1/\delta))$ where $\boldsymbol{H}$ is the neural tangent kernel matrix, λ_0 is the minimum eigenvalue of $\boldsymbol{H}$ and $\boldsymbol{r}_t = [r_t(a_1)\cdots r_t(a_N)]^T$.

Lemma G.7 (Lemma B.1 by Zhou et al. (2020)). *Let $\boldsymbol{H}$ be the neural tangent kernel matrix and let $\boldsymbol{G}_0 = [g(a_1;\boldsymbol{W}^{(0)})\cdots g(a_N;\boldsymbol{W}^{(0)})]/\sqrt{m} \in \mathbb{R}^{p\times N}$. Then, with probability at least $1-\delta$, we have*

$$\|\boldsymbol{G}_0^T\boldsymbol{G}_0 - \boldsymbol{H}\|_F \le N\epsilon$$

as long as $m \ge \text{poly}(L, \epsilon^{-1}, \log(1/\delta))$.

Lemma G.8 (Lemma B.2 by Zhou et al. (2020)). *Let $\boldsymbol{W}$ be the parameter trained by the algorithm* TRAINNN *with learning rate $\eta \le \mathcal{O}((m\lambda + TmL)^{-1})$ and initial weight $\boldsymbol{W}^{(0)}$. Then, with probability at least $1-\delta$, we have*

$$\|\boldsymbol{W} - \boldsymbol{W}^{(0)}\|_2 \le 2\sqrt{T/(m\lambda)}$$

as long as the network width satisfies $m \ge \text{poly}(T, L, \lambda^{-1}, \log(1/\delta))$.

G.2 More NTK theory

Lemma G.9. *Let $\boldsymbol{W}$ be close to the initial weight $\boldsymbol{W}^{(0)}$ such that $\|\boldsymbol{W} - \boldsymbol{W}^{(0)}\|_2 \le 2\sqrt{T/(m\lambda)}$. Let $\boldsymbol{G} = [g(a_1;\boldsymbol{W})\cdots g(a_N;\boldsymbol{W})]/\sqrt{m}$ and $\boldsymbol{G}_0 = [g(a_1;\boldsymbol{W}^{(0)})\cdots g(a_N;\boldsymbol{W}^{(0)})]/\sqrt{m}$. Then, with probability at least $1-\delta$, as long as $m \ge \text{poly}(T, L, \lambda^{-1})$, we have*

$$\|\boldsymbol{G}_0^T\boldsymbol{G}_0 - \boldsymbol{G}^T\boldsymbol{G}\|_F \le \mathcal{O}\left(m^{-1/3}(\log m)T^{1/3}N^2\lambda^{-1/3}L^8\right).$$

Proof. For ease of exposition, we write $g(\cdot) = g(\cdot;\boldsymbol{W})$ and $g_0(\cdot) = g(\cdot;\boldsymbol{W}^{(0)})$. Note that

$$\begin{aligned}
\|\boldsymbol{G}_0^T\boldsymbol{G}_0 - \boldsymbol{G}^T\boldsymbol{G}\|_F^2 &= \frac{1}{m^2}\sum_{i,j\in[N]}\left(\langle g(a_i), g(a_j)\rangle - \langle g_0(a_i), g_0(a_j)\rangle\right)^2 \\
&= \frac{1}{m^2}\sum_{i,j\in[N]}\left(\langle g(a_i) - g_0(a_i), g(a_j)\rangle - \langle g_0(a_i), g_0(a_j) - g(a_j)\rangle\right)^2 \\
&\le \frac{2}{m^2}\sum_{i,j\in[N]}\left(\|g(a_i) - g_0(a_i)\|_2^2\|g(a_j)\|_2^2 + \|g_0(a_i)\|_2^2\|g_0(a_j) - g(a_j)\|_2^2\right) \\
&\le \frac{2}{m^2}N^2\mathcal{O}\left((\log m)T^{1/3}m^{5/3}\lambda^{-1/3}L^8\right) \\
&= \mathcal{O}\left((\log m)T^{1/3}N^2m^{-1/3}\lambda^{-1/3}L^8\right)
\end{aligned}$$

where the first inequality follows by Cauchy-Schwarz and $(a-b)^2 \le 2a^2 + 2b^2$, and the second inequality follows by Lemma G.3 and Lemma G.4. ∎

Lemma G.10. *Consider a weight $\boldsymbol{W}$ close to the initial weight $\boldsymbol{W}^{(0)}$ such that $\|\boldsymbol{W} - \boldsymbol{W}^{(0)}\|_2 \le 2\sqrt{T/(m\lambda)}$. Let φ and $\varphi^{(0)}$ be feature mappings equivalent to $g(\cdot;\boldsymbol{W})/\sqrt{m}$ and $g(\cdot;\boldsymbol{W}^{(0)})/\sqrt{m}$ respectively. Then, as long as $m \ge \text{poly}(T, L, \lambda^{-1})$, we have*

$$\gamma_{\varphi,T} \le \gamma_{\varphi^{(0)},T} + \mathcal{O}\left(m^{-1/3}(\log m)T^{4/3}N^{5/2}\lambda^{-1/3}L^8\right).$$

Proof. Let $\boldsymbol{G} = [g(a_1;\boldsymbol{W})\cdots g(a_N;\boldsymbol{W})]/\sqrt{m}$ and $\boldsymbol{G}_0 = [g(a_1;\boldsymbol{W}^{(0)})\cdots g(a_N;\boldsymbol{W}^{(0)})]/\sqrt{m}$. For ease of exposition, write $S = S_{g(\cdot;\boldsymbol{W})/\sqrt{m}}$ and $S_0 = S_{g(\cdot;\boldsymbol{W}^{(0)})/\sqrt{m}}$. Then, $S(TP/\sigma, 1) = \frac{T}{\sigma}\boldsymbol{G}D_P\boldsymbol{G}^T + I_p$ and $S_0(TP/\sigma, 1) =$

$\frac{T}{\sigma}\boldsymbol{G}_0 D_P \boldsymbol{G}_0^T + I_p$ where $D_P = \mathrm{diag}(P(a_1), \dots, P(a_N)) \in \mathbb{R}^{N\times N}$. Using the Sylvester's identity $\log\det(I+AB) = \log\det(I+BA)$, we have

$$\begin{aligned}
\log\det S(TP/\sigma, 1) &= \log\det((T/\sigma)GD_PG^T + I_p) = \log\det((T/\sigma)D_PG^TG + I_N) \\
&= \log\det((T/\sigma)D_PG_0^TG_0 + I_N + (T/\sigma)D_PG^TG - (T/\sigma)D_PG_0^TG_0) \\
&\le \log\det((T/\sigma)D_PG_0^TG_0 + I_N) \\
&\quad + \langle((T/\sigma)D_PG_0^TG_0 + I_N)^{-1}, (T/\sigma)D_PG^TG - (T/\sigma)D_PG_0^TG_0)\rangle \\
&\le \log\det((T/\sigma)D_PG_0^TG_0 + I_N) \\
&\quad + \|((T/\sigma)D_PG_0^TG_0 + I_N)^{-1}\|_F \|(T/\sigma)D_PG^TG - (T/\sigma)D_PG_0^TG_0)\|_F
\end{aligned}$$

where the first inequality follows by the concavity of $\log\det(\cdot)$ and the last inequality follows by Cauchy-Schwarz. To bound the second term on the right hand side, we can bound the first factor by

$$\|((T/\sigma)D_PG_0^TG_0 + I_N)^{-1}\|_F \le \sqrt{N}\|((T/\sigma)D_PG_0^TG_0 + I_N)^{-1}\|_2 \le \sqrt{N}$$

where the first inequality uses the identity $\|A\|_F \le \sqrt{N}\|A\|_2$ for $A \in \mathbb{R}^{N\times N}$ and the second inequality uses $(T/\sigma)D_PG_0^TG_0 + I_N \succcurlyeq I_N$. Also, we can bound the second factor by

$$(T/\sigma)\|D_P(G^TG - G_0^TG_0)\|_F \le (T/\sigma)\|D_P\|_2\|G^TG - G_0^TG_0\|_F \le (T/\sigma)\|G^TG - G_0^TG_0\|_F$$

where the first inequality uses the identity $\|AB\|_F \le \|A\|_2\|B\|_F$ and the second inequality uses $\|D_P\|_2 \le 1$. Using the bound of the two factors and using Lemma G.9 for bounding $\|G^TG - G_0^TG_0\|_F$ gives

$$\log\det S(TP/\sigma, 1) \le \log\det S_0(TP/\sigma, 1) + \mathcal{O}\left((\log m)T^{4/3}N^{5/2}m^{-1/3}\lambda^{-1/3}L^8\right).$$

Maximizing over $P \in \mathcal{P}_\mathcal{X}$ on the left hand side and denoting the maximizer by $P^\star$, we get

$$\begin{aligned}
\gamma_{\varphi,T} &\le \log\det S_0(TP^\star/\sigma, 1) + \mathcal{O}\left((\log m)T^{4/3}N^{5/2}m^{-1/3}\lambda^{-1/3}L^8\right) \\
&\le \gamma_{\varphi^{(0)},T} + \mathcal{O}\left((\log m)T^{4/3}N^{5/2}m^{-1/3}\lambda^{-1/3}L^8\right)
\end{aligned}$$

where the second inequality follows since $\gamma_{\varphi^{(0)},T}$ maximizes $\log\det S_0(TP/\sigma, 1)$ over $P \in \mathcal{P}_\mathcal{X}$. This completes the proof. ■

Lemma G.11. *Let $\boldsymbol{W}$ be a parameter returned by the* TRAINNN *algorithm. For each $t \in [T]$, let $\boldsymbol{W}_t^\star$ be a parameter that satisfies $r_t(x) = \langle g(x;\boldsymbol{W}^{(0)}), \boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}\rangle$ for all $x \in \mathcal{X}$ and $\|\boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}\|_2 \le \sqrt{2\boldsymbol{r}_t^T\boldsymbol{H}^{-1}\boldsymbol{r}_t/m}$. Such a parameter $\boldsymbol{W}_t^\star$ exists by Lemma G.6. Then, with probability at least $1-\delta$, we have*

$$|r_t(x) - \langle g(x;\boldsymbol{W}), \boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}\rangle| \le \epsilon$$

for all $t \in [T]$ and $x \in \mathcal{X}$ as long as $m \ge \mathrm{poly}(T, L, N, \lambda^{-1}, \lambda_0^{-1}, \log(1/\delta), \epsilon^{-1})$.

Proof. By Lemma G.8, we have with high probability that $\|\boldsymbol{W} - \boldsymbol{W}^{(0)}\|_2 \le 2\sqrt{T/(m\lambda)}$ as long as $m \ge \mathrm{poly}(T, L, \lambda^{-1}, \log(1/\delta))$ which allows us to use Lemma G.3 and Lemm G.4 to get

$$\|g(x;\boldsymbol{W}^{(0)}) - g(x;\boldsymbol{W})\|_2 \le \mathcal{O}(\sqrt{\log m}T^{1/6}m^{1/3}\lambda^{-1/6}L^{7/2})$$

with high probability as long as $m \ge \mathrm{poly}(T, L, \lambda^{-1}, \log(1/\delta))$. Hence, we have

$$\begin{aligned}
|r_t(x) - \langle g(x;\boldsymbol{W}), \boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}\rangle| &= |\langle g(x;\boldsymbol{W}^{(0)}) - g(x;\boldsymbol{W}), \boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}\rangle| \\
&\le \|g(x;\boldsymbol{W}^{(0)}) - g(x;\boldsymbol{W})\|_2\|\boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}\|_2 \\
&\le \mathcal{O}(\sqrt{\log m}T^{1/6}m^{-1/6}\lambda^{-1/6}\lambda_0^{-1/2}N^{1/2}L^{7/2}) \le \epsilon
\end{aligned}$$

for all $t \in [T]$ and $x \in \mathcal{X}$ as long as $m \ge \mathrm{poly}(T, L, N, \lambda^{-1}, \lambda_0^{-1}, \epsilon^{-1})$. This completes the proof. ■

Lemma G.12. *Let $\boldsymbol{W}$ be close to the initial weight $\boldsymbol{W}^{(0)}$ such that $\|\boldsymbol{W} - \boldsymbol{W}^{(0)}\|_2 \leq 2\sqrt{T/(m\lambda)}$. Let $\varphi : \mathcal{X} \to \mathbb{R}^N$ be a feature mapping equivalent to $g(\cdot; \boldsymbol{W})/\sqrt{m}$. Then, we have*

$$\gamma_{\varphi,T} = \mathcal{O}(N \log(TL)).$$

Proof. Using the identity $\det A \leq (\frac{1}{N} \operatorname{Tr}(A))^N$ for positive semi-definite $A \in \mathbb{R}^{N \times N}$, we have

$$\begin{aligned}
\log \det S_\varphi(\sigma^{-1}TP, 1) &\leq N \log \left(\frac{1}{N} \operatorname{Tr}(S_\varphi(\sigma^{-1}TP, 1)) \right) \\
&= N \log \left(\frac{T}{\sigma N} \sum_{x \in \mathcal{X}} P(x) \|\varphi(x)\|_2^2 + 1 \right) \\
&\leq N \log \left(\frac{T}{\sigma N} \sum_{x \in \mathcal{X}} P(x) \mathcal{O}(L) + 1 \right) \\
&= \mathcal{O}(N \log(TL))
\end{aligned}$$

where the second inequality uses $\|\varphi(x)\|_2^2 = \|g(x; \boldsymbol{W})\|_2^2/m$ due to equivalence and Lemma G.4. This completes the proof. ∎

G.3 Concentration bound on reward estimates

In this subsection, we prove the following concentration bound for the reward estimate analogous to Lemma D.1.

Lemma G.13 (c.f. Lemma D.1)**.** *Let $\mathcal{I} \subseteq [T]$ be a time interval. Let m_t be the strategy index used at time t by OPNN and $\varphi^{(m)}$ the feature mapping computed by OPNN using data in the cumulative block $\mathcal{C}(m-1)$. Let $\boldsymbol{\varphi} = \{\varphi_t\}_{t \in \mathcal{I}}$ be the sequence of feature mappings used by OPNN where $\varphi_t = \varphi^{(m_t)}$. If j is such that $m_t \leq j$ for all $t \in \mathcal{I}$, then with probability at least $1 - \frac{2\delta}{C}$, we have for all $x \in \mathcal{X}$ that*

$$\begin{aligned}
|\widehat{\mathcal{R}}_{\boldsymbol{\varphi},\mathcal{I}}(x) - \mathcal{R}_{\mathcal{I}}(x)| \leq \frac{\xi_j}{|\mathcal{I}|} \sum_{t \in \mathcal{I}} \|\varphi_t(x)\|^2_{S_{\varphi_t}(P_t, \sigma/T)^{-1}} + \frac{\log(CN/\delta)}{\xi_j |\mathcal{I}|} \\
+ \sqrt{\frac{\sigma}{T}} \|\varphi_t(x)\|_{S_{\varphi_t}(P_t, \sigma/T)^{-1}} + \epsilon
\end{aligned}$$

as long as $m \geq$ poly$(T, L, N, \lambda^{-1}, \lambda_0^{-1}, \log(1/\delta), \epsilon^{-1})$ where $\xi_j = \mu_j/(4\gamma_{\varphi^{(0)},T})$ and $\widehat{\mathcal{R}}_{\boldsymbol{\varphi},\mathcal{I}} \coloneqq \frac{1}{|\mathcal{I}|} \sum_{t \in \mathcal{I}} \widehat{\mathcal{R}}_{\varphi_t,t}$.

First, we show the following distributional properties of the IPS estimator analogous to Lemma D.3.

Lemma G.14 (c.f. Lemma D.3)**.** *Let m_t be the strategy index used by OPNN at time t and let $\varphi_t : \mathcal{X} \to \mathbb{R}^N$ be the feature mapping equivalent to $g(\cdot; \boldsymbol{W}^{(m_t)})/\sqrt{m}$ used by OPNN at time t. Let $P_t = P^{(m_t)}$ be the strategy used at time t. Then, with probability at least $1 - \delta$, the IPS estimator $\widehat{\mathcal{R}}_{\varphi_t,t}(x)$ satisfies*

$$\begin{aligned}
|\widehat{\mathcal{R}}_{\varphi_t,t}(x)| &\leq \frac{\gamma_{\varphi_t,T}}{\mu_{m_t}} \\
|\mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi_t,t}(x)] - r_t(x)| &\leq \sqrt{\frac{\sigma}{T}} \|\varphi_t(x)\|_{S_{\varphi_t}(P_t, \sigma/T)^{-1}} + \epsilon \\
\operatorname{Var}_t[\widehat{\mathcal{R}}_{\varphi_t,t}(x)] &\leq \|\varphi_t(x)\|^2_{S_{\varphi_t}(P_t, \sigma/T)^{-1}}
\end{aligned}$$

for all $x \in \mathcal{X}$ and $t \in [T]$ as long as $m \geq$ poly$(T, L, N, \lambda^{-1}, \lambda_0^{-1}, \log(1/\delta), \epsilon^{-1})$.

Proof. The first and the third inequalities follow by the same proof as in Lemma D.3. We focus on the second inequality. By Lemma G.11, with probability at least $1 - \delta$, there exists $\xi_{t,x}$ with $|\xi_{t,x}| \leq \epsilon_0$ such that $r_t(x) = \langle g(x; \boldsymbol{W}^{(m_t)}), \boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)} \rangle + \xi_{t,x}$ for all $t \in [T]$ and $x \in \mathcal{X}$ as long as $m \geq \text{poly}(T, L, N, \lambda^{-1}, \lambda_0^{-1}, \log(1/\delta), \epsilon_0^{-1})$.

Writing $S_t = S_{g(\cdot;\boldsymbol{W}^{(m_t)})/\sqrt{m}}(P_t, \sigma/T)$ and $g_t(\cdot) = g(\cdot;\boldsymbol{W}^{(m_t)})/\sqrt{m}$ for convenience, we have

$$\begin{aligned}
\mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi_t,t}(x)] &= \mathbb{E}_t[\varphi_t(x)^T S_{\varphi_t}(P_t, \sigma/T)^{-1}\varphi_t(x_t) r_t(x_t)] \\
&= \mathbb{E}_t[g_t(x)^T S_t^{-1} g_t(x_t)(\sqrt{m} g_t(x_t)^T(\boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}) + \xi_{t,x})] \\
&= \sqrt{m} g_t(x)^T S_t^{-1}(S_t - (\sigma/T)I)(\boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}) + \mathbb{E}_t[g_t(x)^T S_t^{-1} g_t(x_t)\xi_{t,x}] \\
&= r_t(x) - \xi_{t,x} - \frac{\sigma\sqrt{m}}{T} g_t(x)^T S_t^{-1}(\boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}) + \mathbb{E}_t[g_t(x)^T S_t^{-1} g_t(x_t)\xi_{t,x}]
\end{aligned}$$

where the first equality uses the fact that the term with the noise η_t vanishes due to independence and the second equality uses Lemma I.1. Hence, writing $\widetilde{S}_t = S_{\varphi_t}(P_t, \sigma/T)$, we have

$$\begin{aligned}
|\mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi_t,t}(x)] - r_t(x)| &\le |\xi_{t,x}| + \frac{\sigma\sqrt{m}}{T}|g_t(x)^T S_t^{-1}(\boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)})| + \mathbb{E}_t[|g_t(x)^T S_t^{-1} g_t(x_t)\xi_{t,x}|] \\
&\le \epsilon_0 + \frac{\sigma\sqrt{m}}{T}\|g_t(x)\|_{S_t^{-1}}\|\boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}\|_{S_t^{-1}} + \epsilon_0 \mathbb{E}_t[\|g_t(x)\|_{S_t^{-1}}\|g_t(x_t)\|_{S_t^{-1}}] \\
&\le \epsilon_0 + \sqrt{\frac{\sigma m}{T}}\|\varphi_t(x)\|_{\widetilde{S}_t^{-1}}\|\boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}\|_2 + \epsilon_0 \mathbb{E}_t[\|\varphi_t(x)\|_{\widetilde{S}_t^{-1}}\|\varphi_t(x_t)\|_{\widetilde{S}_t^{-1}}]
\end{aligned}$$

where the second inequality uses Cauchy-Schwarz and the last inequality uses Lemma I.1 and $S_t^{-1} \preccurlyeq (T/\sigma)I$. Since $P_t = (1-\mu_{m_t})P^{(m_t)} + \mu_{m_t}\pi_{\varphi_t,\mathcal{X}} \succcurlyeq \mu_{m_t}\pi_{\varphi_t,\mathcal{X}}$, we have $\widetilde{S}_t \succcurlyeq \mu_{m_t} S_{\varphi_t}(\pi_{\varphi_t,\mathcal{X}}, \sigma/T)$ and it follows that

$$\|\varphi_t(x)\|_{\widetilde{S}_t^{-1}}^2 \le \frac{1}{\mu_{m_t}}\|\varphi_t(x)\|_{S_{\varphi_t}(\pi_{\varphi_t,\mathcal{X}},\sigma/T)^{-1}}^2 \le \frac{\gamma_{\varphi_t,T}}{\mu_{m_t}} \le CT^{1/2}N\log(TL)$$

for some constant C where the second inequality follows by Lemma 4.3 and the last inequality follows by $\mu_j \ge T^{-1/2}$ and Lemma G.12. Also, by Lemma G.6, we have $\sqrt{m}\|\boldsymbol{W}_t^\star - \boldsymbol{W}^{(0)}\|_2 \le 1$ with probability at least $1-\delta$. Hence, we can further bound the bias term by

$$\begin{aligned}
|\mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi_t,t}(x)] - r_t(x)| &\le \epsilon_0 + \sqrt{\frac{\sigma}{T}}\|\varphi_t(x)\|_{S_{\varphi_t}(P_t,\sigma/T)^{-1}} + \epsilon_0 CT^{1/2}N\log(TL) \\
&\le \sqrt{\frac{\sigma}{T}}\|\varphi_t(x)\|_{S_{\varphi_t}(P_t,\sigma/T)^{-1}} + \epsilon
\end{aligned}$$

where we set ϵ_0 sufficiently small such that $\epsilon_0 + \epsilon_0 CT^{1/2}N\log(TL) \le \epsilon$ and choose $m \ge \mathrm{poly}(T, L, N, \lambda^{-1}, \lambda_0^{-1}, \log(1/\delta), \epsilon^{-1})$ appropriately. This completes the proof. ■

Using the previous lemma, we are ready to prove Lemma G.13.

Proof of Lemma G.13. Fix an action $x \in \mathcal{X}$ and consider a martingale difference sequence $\{z_{t,x}\}_{t\in\mathcal{I}}$ where $z_{t,x} = \widehat{\mathcal{R}}_{\varphi_t,t}(x) - \mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi_t,t}(x)]$. We can bound $z_{t,x}$ for all $t \in \mathcal{I}$ by

$$z_{t,x} \le |\widehat{\mathcal{R}}_{\varphi_t,t}(x)| + \mathbb{E}_t[|\widehat{\mathcal{R}}_{\varphi_t,t}(x)|] \le \frac{2\gamma_{\varphi_t,T}}{\mu_{m_t}} \le \frac{4\gamma_{\varphi^{(0)},T}}{\mu_j}$$

where the second inequality uses Lemma G.14 to bound $|\widehat{\mathcal{R}}_{\varphi_t,t}(x)|$ and the last inequality uses Lemma G.10 to choose $m \ge \mathrm{poly}(T, N, \lambda, L)$ that satisfies $\gamma_{\varphi_t,T} \le 2\gamma_{\varphi^{(0)},T}$. Also, we have $\mathrm{Var}_t[z_{t,x}] = \mathrm{Var}_t[\widehat{\mathcal{R}}_{\varphi_t,t}(x)] \le \|\varphi_t(x)\|_{S_{\varphi_t}(P_t,\sigma/T)^{-1}}^2$ by Lemma G.14. Using the Freedman inequality (Lemma D.2) on $\{z_{t,x}\}_{t\in\mathcal{I}}$ we get with probability at least $1 - \frac{\delta}{CN}$ that

$$\begin{aligned}
\widehat{\mathcal{R}}_{\boldsymbol{\varphi},\mathcal{I}}(x) - \mathcal{R}_{\mathcal{I}}(x) &= \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}(z_{t,x} + \mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi_t,t}(x)] - \mathcal{R}_{\mathcal{I}}(x)) \\
&\le \frac{\xi_j}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\|\varphi_t(x)\|_{S_{\varphi_t}(P_t,\sigma/T)^{-1}}^2 + \frac{\log(CN/\delta)}{\xi_j|\mathcal{I}|} + \sqrt{\frac{\sigma}{T}}\|\varphi_t(x)\|_{S_{\varphi_t}(P_t,\sigma/T)^{-1}} + \epsilon
\end{aligned}$$

where $\xi_j = \mu_j/(4\gamma_{\varphi^{(0)},T})$ and we use Lemma G.14 to bound the bias term $\mathbb{E}_t[\widehat{\mathcal{R}}_{\varphi^{(m_t)},t}(x)] - \mathcal{R}_{\mathcal{I}}(x)$. A union bound over all $x \in \mathcal{X}$ and the reverse case $\mathcal{R}_{\mathcal{I}}(x) - \widehat{\mathcal{R}}_{\boldsymbol{\varphi},\mathcal{I}}(x)$ completes the proof. ■

Lemma G.15 (c.f. Lemma 4.5)**.** *Let m_t be the strategy index used at time t by OPNN and $\varphi^{(m)}$ the feature mapping computed by OPNN using data in the cumulative block $\mathcal{C}(m-1)$. Let $\boldsymbol{\varphi} = \{\varphi_t\}_{t\in\mathcal{I}}$ be the sequence of feature mappings used by OPNN where $\varphi_t = \varphi^{(m_t)}$. With high probability, when running the OPNN algorithm, we have for all block indices $j = 0, 1, \ldots$ and actions $x \in \mathcal{X}$ that*

$$|\widehat{\mathcal{R}}_{\boldsymbol{\varphi},\mathcal{C}(j)}(x) - \mathcal{R}_{\mathcal{C}(j)}(x)| \le \frac{1}{2}\Delta_{\mathcal{C}(j)}(x) + V_{\mathcal{C}(j)} + \frac{c_0}{4}\mu_j \tag{24}$$

$$\Delta_{\mathcal{C}(j)}(x) \le 2\widehat{\Delta}_{\boldsymbol{\varphi},\mathcal{C}(j)}(x) + 4V_{\mathcal{C}(j)} + c_0\mu_j \tag{25}$$

$$\widehat{\Delta}_{\boldsymbol{\varphi},\mathcal{C}(j)}(x) \le 2\Delta_{\mathcal{C}(j)}(x) + 4V_{\mathcal{C}(j)} + c_0\mu_j \tag{26}$$

where $c_0 = (40 + 16\sqrt{\alpha})$.

Proof. Apart from dealing with the error ϵ when applying Lemma G.14 due to the finiteness of the width of the network and bounding $\gamma_{\varphi_t,T} \le \gamma_{\varphi^{(0)},T} + \epsilon$ using Lemma G.10, the proof is exactly the same as that for Lemma 4.5. As for dealing with ϵ, we set $\epsilon \le 1$ by choosing $m \ge \text{poly}(T, L, N, \lambda^{-1}, \lambda_0^{-1}, \log(1/\delta), \epsilon^{-1})$ appropriately when applying Lemma G.14 and Lemma G.10. ∎

Now, we are ready to prove Theorem G.1.

Proof of Theorem G.1. The proof is exactly the same as that of Theorem 4.6. Instead of using Lemma 4.5 as in the proof of Theorem 4.6, we use Lemma G.15 for the reward estimate concentration bound and the suboptimality gap estimate concentration bound. ∎

H Analysis of ADA-OPNN

The analysis of ADA-OPNN is exactly the same as the analysis of ADA-OPKB presented in Section F with the following adjustments. In place of Lemma D.1 and Lemma 4.5 use Lemma G.13 and Lemma G.15.

I Equivalence of feature mappings

Recall that OPKB and ADA-OPKB use a feature mapping equivalent to a feature mapping corresponding to a given kernel. Also, OPNN and ADA-OPNN use a feature mapping equivalent to the feature mapping induced by the neural network. In this section, we show that the choice of feature mapping does not affect the algorithm and the analysis. Note that the algorithm and the analysis depend on the feature mapping φ only through the quantities $\|\varphi(x)\|^2_{S_\varphi(P,\lambda)^{-1}}$ and $\log\det S_\varphi(P,\lambda)$. The following lemmas show that these quantities are not affected by the choice of the equivalent feature mapping.

Lemma I.1. *Let $\psi : \mathcal{X} \to \ell^2$ (or $\psi : \mathcal{X} \to \mathbb{R}^p$) be a feature mapping. Let $\varphi : \mathcal{X} \to \mathbb{R}^N$ be an equivalent feature mapping. Then, for all $x, x' \in \mathcal{X}$, we have*

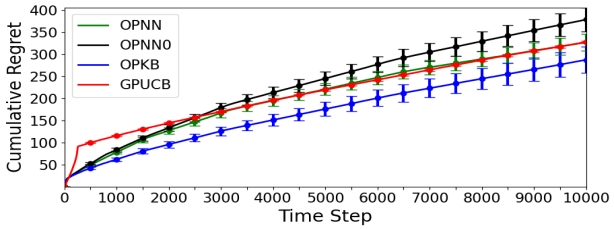
$$\varphi(x)^T S_\varphi(P,\lambda)^{-1}\varphi(x') = \psi(x)^T S_\psi(P,\lambda)^{-1}\psi(x').$$

Proof. We prove the more general case $\psi : \mathcal{X} \to \ell^2$. Let $\Phi = [\varphi(a_1)\cdots\varphi(a_N)]^T \in \mathbb{R}^{N\times N}$ and $\Psi = [\psi(a_1)\cdots\psi(a_N)]^T \in \mathbb{R}^{N\times\infty}$. The infinite matrix Ψ can be thought of a linear operator $\Psi : \ell^2 \to \mathbb{R}^N$ with $\Psi(\cdot) = (\langle\psi(a_1),\cdot\rangle, \ldots, \langle\psi(a_N),\cdot\rangle)$. We denote by $\Psi^T : \mathbb{R}^N \to \ell^2$ the linear operator with $\Psi^T(w) = \sum_{i=1}^N w_i\varphi(a_i)$. By the definition of equivalence of feature mappings, we have $\Phi\Phi^T = \Psi\Psi^T = K$ where $K = [\langle\psi(x),\psi(x')\rangle]_{x,x'\in\mathcal{X}}$ is the kernel matrix. Defining $D_P = \text{diag}(P(a_1),\ldots,P(a_N))$, we can write $S_\varphi(P,\lambda) = \Phi^T D_P\Phi + \lambda I_N$ and $S_\psi(P,\lambda) = \Psi^T D_P\Psi + \lambda I$. Note that

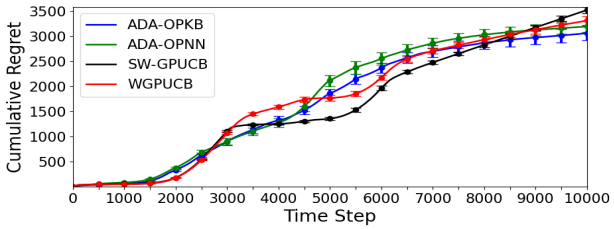
$$S_\psi(P,\lambda)\Psi^T = (\Psi^T D_P\Psi + \lambda I)\Psi^T = \Psi^T(D_P\Psi\Psi^T + \lambda I_N) = \Psi^T(D_P K + \lambda I_N).$$

Applying the inverses of $S_\psi(P,\lambda)$ and $(D_P K + \lambda I_N)$ on both sides, we get $\Psi^T(D_P K + \lambda I_N)^{-1} = S_\psi(P,\lambda)^{-1}\Psi^T$ It follows that

$$\begin{aligned}\psi(a_i)^T S_\psi(P,\lambda)^{-1}\psi(a_j) &= \langle\psi(a_i), S_\psi(P,\lambda)^{-1}\Psi^T e_j\rangle \\ &= \langle\psi(a_i), \Psi^T(D_P K + \lambda I_N)^{-1}e_j\rangle \\ &= \langle\psi(a_i), \Psi^T w\rangle\end{aligned}$$



(a) Stationary cosine bandit



(b) Slowly-varying cosine bandit

Figure 2: Cumulative regret comparison of algorithms in cosine bandit environments

where $e_j \in \mathbb{R}^N$ is the unit vector with jth entry 1 and $w = (D_P K + \lambda I_N)^{-1} e_j$. Since $\langle \psi(a_i), \Psi^T w \rangle = \langle \psi(a_i), \sum_{j=1}^N w_j \psi(a_j) \rangle = \sum_{j=1}^N w_j k(a_i, a_j) = e_i^T K w$, it follows by standard matrix algebra that

$$\begin{aligned}
\psi(a_i)^T S_\psi(P,\lambda)^{-1} \psi(a_j) &= e_i^T K w \\
&= e_i^T \Phi\Phi^T (D_P \Phi\Phi^T + \lambda I_N)^{-1} e_j \\
&= e_i^T \Phi (\Phi^T D_P \Phi + \lambda I_N)^{-1} \Phi^T e_j \\
&= \varphi(a_i)^T S_\varphi(P,\lambda)^{-1} \varphi(a_j)
\end{aligned}$$

for all $1 \le i, j \le N$ where the second to last equality uses the fact that $\Phi^T(D_P\Phi\Phi^T + \lambda I_N) = (\Phi^T D_P \Phi + \lambda I_N)\Phi^T$, which implies $\Phi^T(D_P\Phi\Phi^T + \lambda I_N)^{-1} = (\Phi^T D_P\Phi + \lambda I_N)^{-1}\Phi^T$. This completes the proof. ■

Lemma I.2. *Let $\varphi_1 : \mathcal{X} \to \mathbb{R}^{p_1}$ and $\varphi_2 : \mathcal{X} \to \mathbb{R}^{p_2}$ be equivalent feature mappings. Then, we have*

$$\log\det S_{\varphi_1}(P,\lambda) = \log\det S_{\varphi_2}(P,\lambda).$$

Proof. Let $\Phi_1 = [\varphi_1(a_1) \cdots \varphi_1(a_N)]^T \in \mathbb{R}^{N\times p_1}$ and $\Phi_2 = [\varphi_2(a_1) \cdots \varphi_2(a_N)]^T \in \mathbb{R}^{N\times p_2}$. By the definition of equivalence of feature mappings, we have $\Phi_1\Phi_1^T = \Phi_2\Phi_2^T = K$ for some kernel matrix $K \in \mathbb{R}^{N\times N}$. Defining $D_P = \mathrm{diag}(P(a_1), \ldots, P(a_N))$, we can write $S_{\varphi_1}(P,\lambda) = \Phi_1^T D_P \Phi_1 + \lambda I_N$ and $S_{\varphi_2}(P,\lambda) = \Phi_2^T D_P \Phi_2 + \lambda I_N$. Using the Sylvester's determinant identity $\det(AB + I) = \det(BA + I)$, we get

$$\begin{aligned}
\log\det S_{\varphi_1}(P,\lambda) &= \log\det(\Phi_1^T D_P \Phi_1 + \lambda I_{p_1}) \\
&= \log\det(\Phi_1\Phi_1^T D_P + \lambda I_N) \\
&= \log\det(\Phi_2\Phi_2^T D_P + \lambda I_N) \\
&= \log\det(\Phi_2^T D_P \Phi_2 + \lambda I_{p_2}) \\
&= \log\det S_{\varphi_2}(P,\lambda)
\end{aligned}$$

which completes the proof. ■

J Additional experiments

In this section, we provide additional experimental results under a simulated environment with the reward function $r_t(x) = 0.8\cos(3x^T\theta + \phi(t))$ where the action x and the parameter θ are randomly sampled from the unit sphere in $\mathbb{R}^d$, and $\phi(t)$ denotes the phase over time. We use the parameters tuned in Section 6 for all the experiments in this section.

J.1 Algorithm Tuning

We tune SW-GPUCB, WGPUCB, ADA-OPKB and ADA-OPNN algorithms under the single switch environment. For SW-GPUCB, we do a grid search for λ over the range $\{0.01, 0.02, 0.05, 0.1, \ldots, 100\}$, the UCB scale parameter v over $[0.001, 1]$, and the window size over $\{100, 200, 500, 1000, \ldots, 10000\}$. See Algorithm 8 for the definition of λ. For WGPUCB, we do a grid search for λ over the range $\{0.01, 0.02, 0.05, 0.1, \ldots, 100\}$, the UCB scale parameter over $\{0.001, 0.002, 0.005, 0.01, \ldots, 1\}$, and the discounting factor over $\{0.99, 0.995, 0.999, 0.9995, 0.9999\}$. See Algorithm 8 for the definition of λ. For ADA-OPKB and ADA-OPNN, we do a grid search for σ over

$\{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}$ and c_0, c_1, c_2, c_3, c_4 over $\{0.001, 0.002, 0.005, 0.01, \ldots, 100\}$. For ADA-OPNN, we do a grid search for the learning rate η over $\{10^{-9}, 10^{-8}, 10^{-7}\}$, training steps J over $\{100, 1000, 10000\}$ and regularization parameter λ over $\{1, 10, 100, 1000\}$. We use a neural network of depth $L = 3$ and width $m = 2048$.

J.2 Stationary cosine bandits

We perform an experiment to demonstrate that OPNN benefits from dynamically adapting the feature mapping. We use the cosine bandits described earlier with the phase fixed at $\phi(t) = 0$. For a comparison, we run the algorithm OPNN0 that does not train the neural network for updating the feature mapping and uses the feature mapping induced by the initial weight of the neural network for all blocks.

The cumulative regrets averaged over 50 random seeds are shown in plot (b) of Figure 2. Error bars indicates standard errors of the means. OPNN outperforms OPNN0, suggesting that updating feature mapping by training the neural network with observed data is beneficial. Also, note that the performance of OPNN is comparable to GPUCB and OPKB.

J.3 Slowly-varying cosine bandits

We perform an experiment on slowly-varying bandits to demonstrate that our change detection based algorithms ADA-OPKB and ADA-OPNN adapt to slowly-varying environments. We use the cosine bandit described earlier with varying phase $\phi(t)$. We keep $\phi(t) = 0$ from time 0 to 1000, then let it grow from 0 to π linearly from time 1000 to 3000. From time 4000 to 6000, we let $\phi(t)$ grow again from π to 2π linearly, and then keep $\phi(t) = 2\pi$ until the end of the simulation.

The cumulative regrets averaged over 25 random seeds under the slowly-varying cosine environment are shown in plot(b) of Figure 2. Error bars indicate standard errors of the means. Note that SW-GPUCB with window size 3000, which is the best tuned parameter for the switching environment in Section 6, is outperformed by the change detection based algorithms ADA-OPKB and ADA-OPNN. If we tune SW-GPUCB again and use SW-GPUCB with window size 1000, SW-GPUCB performs the best. Similarly, the best tuned WGPUCB under the single switching environment in Section 6 is outperformed by ADA-OPKB and ADA-OPNN in the slowly varying environment.