Online Regret Minimization in Linear Bandits with Offline data.

Sushant Vijayan [1] **Arun S. Suggala** [2] **Karthikeyan Shanmugam** [2] **Soumyabrata Pal** [3]

Abstract

We study hybrid offline-to-online regret minimization in stochastic linear bandits, where an agent leverages prior offline logs to accelerate online adaptation. To safely and optimally incorporate this historical data, we introduce Offline-Online Phased Elimination (OOPE), an algorithm utilizing an extended D-optimal experimental design. We show OOPE achieves an online regret of $\tilde{O}(\sqrt{d_{\text{eff}} T \log(|\mathcal{A}|T)} + d^2)$, where the *effective dimension* $d_{\text{eff}}(\leq d)$ quantitatively captures the quality and coverage of the offline dataset via the eigenspectrum of its Gram matrix. This bound smoothly bridges the gap between purely online learning ($T_{\text{off}} = o(T), d_{\text{eff}} = d$) and regimes with abundant, well-explored offline data ($T \ll T_{\text{off}}, d_{\text{eff}} = o(d)$) where regret is substantially reduced. Furthermore, we derive the first minimax lower bounds for this setting that explicitly depend on offline data quality, establishing that OOPE is near-optimal in both well-explored and poorly-explored regimes. Finally, we propose a Frank-Wolfe variant (OOPE-FW) that strictly improves the additive $O(d^2)$ support term, yielding better performance when offline data provides moderate coverage.

1. Introduction

Bandit optimization requires significant exploration to identify optimal actions, limiting its application in high-stakes domains like personalized healthcare or education, where exploration carries ethical concerns or high costs Kapp (2006). For instance, when designing mobile health interventions to nudge physical activity, federal regulations severely constrain online experimentation. However, learners often have access to rich prior logs in these settings. A growing body of work studies how to leverage this historical data to reduce online exploration and accelerate adaptation Banerjee et al. (2022); Wagenmaker & Pacchiano (2022); Agrawal et al. (2023); Hao et al. (2023); Cheung & Lyu (2024).

We study this hybrid Offline to Online (OO) problem in the setting of linear stochastic bandits, where the expected reward of an action $a \in \mathcal{A}$ is a linear function of the action's feature vector: $\mathbb{E}[r_a] = \langle a, \theta \rangle$. Here $\theta \in \mathbb{R}^d$ is the unknown parameter vector, and with a slight overload of notation we let 'a' represent the feature vector of arm a. The learner is provided T_{off} many offline observations generated by a *non-adaptive* logging policy π_{off}, and must subsequently face T rounds of online interaction. Two natural questions that arise in this OO setting are:

1. *When does offline data contribute to a significant reduction in the online regret?*
2. *Can we design optimal algorithms for this setting?*

Prior research has predominantly focused on Multi-Armed Bandits (MABs) or proposed heuristic warm-starts for linear bandits (e.g., warm-started LinUCB, LinTS, or artificial replay) Shivaswamy & Joachims (2012); Banerjee et al. (2022); Cheung & Lyu (2024). However, direct extensions like warm-started LinUCB yield sub-optimal regret guarantees (see Appendix B.7), and Thompson Sampling carries a worst-case $\Omega(d^{3/2}\sqrt{T})$ bound even in purely online settings. Crucially, there is a notable absence of regret lower bounds for linear bandits that explicitly account for the *quality* of the offline logs. Thus, designing provably near-optimal algorithms for linear bandits with offline data remains a significant open question.

Current Work: We introduce the Offline-Online Phased Elimination (OOPE) algorithm. OOPE builds on purely online phased elimination Lattimore & Szepesvári (2020) but carefully incorporates historical logs using a novel *extended D-optimal* experimental design. OOPE operates in phases with increasing geometric lengths. In each phase, OOPE first runs an *exploration* sub-phase based on the extended D-optimal design and then subsequently eliminates under-performing arms in the *elimination* part. To ensure concentration holds while still enabling a simple analysis, OOPE partitions offline samples across phases. A carefully chosen weight α of the offline Gram matrix to balance the current-phase regret against future exploration needs.

[1] Plaksha University, Mohali [2]Google DeepMind, Bengaluru [3]Adobe Research, Bengaluru. Correspondence to: Sushant Vijayan <sushant.vijayan@plaksha.edu.in>.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning. Copyright 2026 by the author(s).

Our main contributions to the problem of online regret minimization in linear bandits for the OO setting are:

- *Effective offline data usage with OOPE:* We prove OOPE achieves an online regret of $\tilde{O}\left(\sqrt{d_{\text{eff}} T \log(|\mathcal{A}|T)} + d^2\right)$. The *effective dimension* d_{eff} precisely quantifies the offline data's quality via the eigenspectrum of its Gram matrix. It smoothly interpolates from pure online settings to settings with large offline data logs. OOPE shows theoretical and practical regret reduction as compared to pure online setting.
- *Minimax Lower Bounds in Linear Bandits OO setting:* We derive the first minimax lower bounds for this setting that explicitly depend on the offline Gram matrix's eigenspectrum. By constructing a novel hard instance that perturbs the hypercube[1] based on π_{off}, we establish that OOPE is near-optimal (up to logarithmic factors in T, T_{off}, d) in both well-explored and poorly-explored regimes.
- *Improved support term overhead via OOPE-FW:* In certain regimes, for example $d = \Omega(\sqrt{T})$, the support-size $O(d^2)$, of the exact D-optimal design dominates the OOPE regret. We introduce OOPE-FW, which utilizes a Frank-Wolfe (FW) approximation. By developing a novel dual feasibility relation for extended Minimum Volume Enclosing Ellipsoid (MVEE) problems, we prove OOPE-FW strictly reduces the support overhead, improving the regret to $\tilde{O}(\sqrt{d_{\text{eff}} T \log(|\mathcal{A}|T)} + d^2/d_{\text{eff}})$.

Organization of the paper: Section 2 reviews the relevant literature, while Section 3 introduces the problem, recaps some preliminaries and defines the notation. The description and analysis of OOPE and the minimax lower bounds are presented in Section 4. To improve the dimensional dependence on the regret we introduce and analyze OOPE-FWin Section 5. Experimental results are given in Section 6 . Section 7 concludes with suggestion for further work.

2. Related Work

We present only the most relevant works in this section and defer a more detailed review to Appendix A.

Regret Minimization in Online Linear Bandits. Dani et al. (2008) provided one of the earliest known regret bounds of $\tilde{O}(d\sqrt{T})$ for purely online regret minimization without any offline data. For the same problem, Abbasi-Yadkori et al. (2011) developed the OFUL algorithm which again achieves $\tilde{O}(d\sqrt{T})$ regret, but with improved log dependence. Although optimal for exponentially sized arm sets ($|\mathcal{A}| = \Omega(2^d)$), these regret bounds are sub-optimal in the sub-exponential arm setting. Subsequent research has developed algorithms to address this drawback. The Phased Elimination algorithm achieves a regret of $O(\sqrt{dT \log(|\mathcal{A}|T)} + d^2)$ Lattimore & Szepesvári (2020) which only has $\log(|\mathcal{A}|)$ dependence.OOPE is based on phased elimination, which has better regret guarantees than UCB style algorithms in the sub-exponential arm setting.

Incorporating Offline Data into Bandits. Several works have investigated regret minimization in OO setting Shivaswamy & Joachims (2012); Banerjee et al. (2022); Hao et al. (2023); Cheung & Lyu (2024) for MABs. Most of them study warm started algorithms like Upper Confidence Bound (UCB), Thompson Sampling (TS) and ϵ-greedy. (Banerjee et al., 2022) proposed artificial-replay algorithm that is shown to attain similar bounds like the warm-started approaches mentioned above. Bayesian regret bounds were obtained for warm started TS by Hao et al. (2023) in the MAB setting. They assume a particular stationary offline data generation that is subsumed in our framework. Furthermore, their regret does not vanish as the quantity and quality of offline data increase. (Cheung & Lyu, 2024) provide a lower bound which helps establish minimax optimality of the warm-started UCB in the OO setting for MABs. A similar lower bound for MABs is reported in (Sentenac et al., 2025), where, in addition to the online regret, the proposed algorithm (a mix of UCB for online settings and LCB for offline settings) also maintains sublinear regret with the offline data generation policy. Oetomo et al. (2023) considered warm-started approach to linear bandits directly and provided regret upper bounds but no lower bounds. We provide an example which shows (section B.8 and lower bound results for TS of (Hamidi & Bayati, 2020)) that warm-starting approaches with typical regret analysis in linear bandits can have sub-optimal dimension factors in the important case of well-explored and plentiful offline data.

3. Preliminaries & Problem Formulation

We consider a stochastic linear bandit problem over a finite action set $\mathcal{A} \subset \mathbb{R}^d$ that spans the entire space $\mathbb{R}^d$. At each step t, pulling arm $a_t \in \mathcal{A}$ yields a reward $y_{a_t} = \langle \theta^*, a_t \rangle + \eta_{a_t,t}$ where $\theta^* \in \mathbb{R}^d$ is the unknown parameter and $\eta_{a_t,t}$ is conditionally independent 1-sub-Gaussian noise. The optimal arm is $a^* := argmax_{a \in \mathcal{A}} \langle \theta^*, a \rangle$ (ties broken arbitrarily).

Notation. Let $\Delta(\mathcal{A})$ denote the probability simplex over $\mathcal{A}$. For $x \in \mathbb{R}^d$ and positive definite matrix B, define the norm $||x||_B := \sqrt{x^T B x}$. For a design $\pi \in \Delta(\mathcal{A})$, its Gram matrix is $V_\pi := \sum_{a \in \mathcal{A}} \pi(a) a a^t$. For a subset of arms

[1]The hard instance in pure online setting is a hypercube, see Chapter 24, Lattimore & Szepesvári (2020).

$\mathcal{B} \subseteq \mathcal{A}$, define $g_{\mathcal{B}} := \max_{a \in \mathcal{B}} ||a||^2_{V_\pi^{-1}}$, which bounds the maximum variance of the Ordinary Least Squares (OLS) estimator over $\mathcal{B}$. The sub-optimality gap for an arm a is $\Delta_a := \langle \theta^*, a^* \rangle - \langle \theta^*, a \rangle$, with $\Delta_{\max}$ and $\Delta_{\min}$ as the maximum and minimum sub-optimality gap respectively. For a positive definite matrix H, the ellipsoid $\xi(H, c)$ is defined as $\xi(H, c) := \{x \in \mathbb{R}^d \mid x^t H x \leq c\}$.

OO regret minimization. Prior to online interaction, the learner is given a dataset of T_{off} offline samples and is denoted as $\mathcal{D}_{\text{off}} := \{(a_t, y_{a_t})\}_{t=1}^{T_{\text{off}}}$.We assume this data was collected via a *fixed, non-adaptive* logging policy $\pi_{\text{off}} \in \Delta(\mathcal{A})$, and that the reward-generating distribution remains stationary between the offline and online phases. The learner interacts with the bandit model for an additional T online rounds. The learner's goal is to choose arms a_t, at each online round $t \in [T]$, to minimize the expected online regret, defined as:

$$\mathcal{R}(T, T_{\text{off}}, \mathcal{A}, \pi_{\text{off}}, \text{Alg}) := T\langle \theta^*, a^* \rangle - \sum_{t=1}^{T} \mathbb{E}[\langle \theta^*, a_t \rangle].$$

The expectation $\mathbb{E}[\cdot]$ is over the realizations of the offline and online noise, as well as any randomness in the learning algorithm. For ease of exposition, we will often write the online regret of an Algorithm Alg as $\mathcal{R}(\text{Alg})$ and suppress the dependence on $T, T_{\text{off}}, \mathcal{A}, \pi_{\text{off}}$.

Preliminaries in Optimal Design. A design is *D-optimal* Kiefer (1960) if it maximizes $\log\det(V_\pi)$ and *G-optimal* if it minimizes $g_{\mathcal{A}}(\pi)$. The celebrated Kiefer-Wolfowitz (KW) theorem establishes that D-optimal and G-optimal designs share the same optimizing distribution π^*, and crucially, that $g_{\mathcal{A}}(\pi) = d$. Furthermore, the dual of D-optimal design is related to finding the Minimum Volume Enclosing Ellipsoid (MVEE) for the set $\mathcal{A}$ Titterington (1975).

Coverage assumption. To ensure our algorithm does not request an impossible number of offline samples during execution, we make the technical assumption:

Assumption 3.1. For every $a \in supp(\pi_{\text{off}}) \subset \mathcal{A}$, we have:

$$\pi_{\text{off}}(a) T_{\text{off}} \geq 3 \lfloor \log_2 \sqrt{(T + T_{\text{off}})/4d \log(4|\mathcal{A}T|) + 1} \rfloor.$$

This assumption demands that for any logging design π_{off} each sampled arm has atleast $\tilde{\Omega}(\log(\sqrt{(T + T_{\text{off}})/d})$ samples and is crucial in establishing Proposition 4.2. We believe this assumption can be relaxed by tweaking OOPE but we have not been able to show this rigorously.

4. Offline-Online Phased Elimination (OOPE)

OOPE proceeds in distinct phases, each requiring geometrically increasing number of offline and online samples. Each phase consists of two parts: an *exploration part*, in which we carefully sample arms based on a well chosen design and an *elimination part* where we eliminate sub-optimal arms based on the observed rewards in the exploration part. At the end of the each phase l, we maintain a set of "live" arms $\mathcal{A}_l$, that is, arms that have survived elimination. The later phases have more stricter thresholds for eliminating arms and thus, necessitates the geometric increase in samples in the latter phases.

Effective dimension d_{eff}: The *effective* dimension d_{eff} is defined as:

$$d_{\text{eff}} := \min\left(\sum_{k=1}^{d} \frac{1}{1 + \frac{T_{\text{off}}}{T} \frac{\lambda_k(V_{\pi_{\text{off}}})}{\max_a ||a||^2}}, \frac{T}{T_{\text{off}}} g_{\mathcal{A}}(\pi_{\text{off}}) \right), \quad (1)$$

where $\lambda_k(V_{\pi_{\text{off}}})$ is the kth largest eigenvalue of the offline Gram matrix. The effective dimension captures the remaining uncertainty of θ^* after incorporating offline data. If the offline policy π_{off} is skewed (say pulling only a one arm) then $d_{\text{eff}} \approx d$, and conversely when offline data is well-explored (say all arms are uniformly explored) then $d_{\text{eff}} = o(d)$. Thus d_{eff} captures the offline data quality in a quantitative manner through the eigenspectrum of $V_{\pi_{\text{off}}}$.

Online Exploration Policy: In a phase l, we choose an exploration design $\pi^*_{l,\text{on}}$ which maximizes the information gain about the unknown parameter θ^*. Mathematically, we have:

$$\pi^*_{l,\text{on}} \in \operatorname*{argmax}_{\pi \in \Delta(\mathcal{A}_l)} \log\left(\det\left(V_{(1-\alpha)\pi + \alpha\pi_{\text{off}}}\right)\right). \quad (2)$$

Here $\alpha \in [0, 1]$ corresponds to the relative weight given to offline to online samples. This objective naturally extends the D-optimal design found in experimental design literature Fedorov (2013) to incorporate offline data and is interpreted as minimizing the volume of the confidence ellipsoid of an Ordinary Least Square (OLS) estimator for θ^* given the access to offline data. In each exploration part of phase l, we sample each arm $a \in \mathcal{A}_l$, $n^l_{\text{on}}(a)$ times:

$$n^l_{\text{on}}(a) := \left\lceil \frac{3 d_{\text{eff}} \pi^*_{l,\text{on}}(a) \log(4l^2|\mathcal{A}|T)}{\epsilon_l^2} \right\rceil. \quad (3)$$

This gives an ϵ_l accurate estimate of the unknown parameter θ, when we construct the OLS estimate $\hat{\theta}_l$.

Offline Data Allocation: In each phase l, we use $n^l_{\text{off}}(a)$ offline samples for arm $a \in \mathcal{A}_l$, where $n^l_{\text{off}}(a)$ is defined as:

$$n^l_{\text{off}}(a) := \left\lceil \frac{2\alpha \pi_{\text{off}}(a) g(\tilde{\pi}^\star_l) \log(4l^2|\mathcal{A}|T)}{\epsilon_l^2} \right\rceil. \quad (4)$$

Here, $g(\tilde{\pi}^\star_l) := g_{\mathcal{A}_l}((1-\alpha)\pi^*_{l,\text{on}} + \alpha\pi_{\text{off}})$. We set $\alpha = \frac{T_{\text{off}}}{T_{\text{off}} + T}$. This choice of α is *crucial* for the success of the algorithm. If α is set too low then we would unnecessarily

Algorithm 1 OFFLINE ONLINE PHASED ELIMINATION (OOPE).

Input: Horizon T, Action Set $\mathcal{A}$, Offline data $\mathcal{D}_{\text{off}}$, Offline Policy π_{off}, Offline Horizon T_{off}.
Initialize: Live arms $\mathcal{A}_1 \leftarrow \mathcal{A}$, phase $l \leftarrow 1$, online pulls $s \leftarrow 0$.
while $s < T$ **do**
  if $|\mathcal{A}_l| == 1$ **then**
    Pull the single arm $a \in \mathcal{A}_l$ for remaining $T - s$ rounds; **break**
  end if
  Set $\epsilon_l \leftarrow 2^{-l}$ and mixing weight $\alpha \leftarrow T_{\text{off}}/(T_{\text{off}} + T)$.
  Compute extended D-optimal design $\pi^*_{l,\text{on}} \in \Delta(\mathcal{A}_l)$ via Eq. (2).
  Set mixture design $\tilde{\pi}^\star_l \leftarrow (1-\alpha)\pi^*_{l,\text{on}} + \alpha\pi_{\text{off}}$.
  for $a \in \mathcal{A}$ **do**
    if $a \in \mathcal{A}_l$ **then**
      Compute target allocations $n^l_{\text{on}}(a)$.
      Pull arm a online $\tilde{n}_{\text{on}} = \min\left(n^l_{\text{on}}(a), T - s\right)$ times.
      Update online budget: $s \leftarrow s + \tilde{n}_{\text{on}}$.
    end if
    Compute offline target allocations $n^l_{\text{off}}(a)$.
    Fetch $\tilde{n}_{\text{off}} = \min\left(n^l_{\text{off}}(a), \text{unused } \mathcal{D}_{\text{off}} \text{ for } a\right)$ offline samples.
  end for
  if $s < T$ **then**
    Compute OLS estimate $\hat{\theta}_l$ using the gathered $\tilde{n}_{\text{on}} + \tilde{n}_{\text{off}}$ samples.
    $\mathcal{A}_{l+1} \leftarrow \left\{a \in \mathcal{A}_l : \max_{a' \in \mathcal{A}_l} \langle a' - a, \hat{\theta}_l\rangle < 2\epsilon_l\right\}$.
    $l \leftarrow l + 1$.
  end if
end while

use excess online samples where the offline samples would have sufficed and incur regret. If α is set too high then we consume too much offline samples for a given confidence width and hence have to use too many online samples in the latter phases.

Elimination: We utilize the offline samples $n^l_{\text{off}}(a)$ and online samples $n^l_{\text{on}}(a)$ collected in the exploration part to construct an OLS estimator $\hat{\theta}_l$. Then we eliminate those arms that are suboptimal wrt $\hat{\theta}_l$, i.e., eliminate arms a which satisfy the inequality $\max_{a' \in \mathcal{A}_l} \langle a', \hat{\theta}_l\rangle \geq \langle a, \hat{\theta}_l\rangle + 2\epsilon_l$. The elimination threshold $\epsilon_l = 2^{-l}$ becomes more stringent for latter phases and consequently the samples considered in each phase l increases geometrically.

4.1. Correctness of OOPE

From Algorithm 1 it is clear that the online and offline sample budgets are not violated for any arm a by OOPE. However, for deriving necessary concentration bounds we do require to ensure that the offline samples are exhausted, if at all, for every arm only in the very last phase. We do this by first deriving an upper bound on maximum number of phases and then showing that showing the offline samples are not exhausted until the penultimate phase wrt to this bound. The next lemma establishes the upper bound on l:

Lemma 4.1. *Denote the total number of phases upto and including the penultimate phase as l_M (the last phase is one in which we exhaust online samples). We define the function $H^{-1} : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{N} \cup \{0\}$ as: $H^{-1}(x) = \max\left\{n \mid \sum_{l=1}^{n} 4^l \log(4l^2|\mathcal{A}|T) \leq x, n \in \mathbb{N} \cup \{0\}\right\}$. Then we have that $l_M \leq H^{-1}\left(\frac{T}{3d_{\text{eff}}}\right)$.*

The proof of Lemma 4.1 is presented in Appendix B.1. Using this bound we can show the following property of OOPE:

Proposition 4.2. *For every arm $a \in supp(\pi_{\text{off}})$, the total number of offline samples requested by OOPE of arm a till phase l_M does not exceed $\pi_{\text{off}}(a)T_{\text{off}}$.*

The proof of Proposition 4.2 is presented in Appendix B.2. This property is useful in enabling us to derive tight regret bounds for OOPE.

4.2. Regret bound for OOPE

In OO setting, the KW theorem is not valid to estimate the key quantity $g(\tilde{\pi}^\star_l)$. The following lemma, which is key for our regret analysis, bounds $g(\tilde{\pi}^\star_l)$ by d_{eff}. The proof is presented in Appendix B.3.

Lemma 4.3. *The optimal mixture design $g(\tilde{\pi}^\star_l)$ of (2) satisfies the following relation:*

$$d = (1-\alpha)g(\tilde{\pi}^\star_l) + \alpha \sum_{a \in \mathcal{A}} \pi_{\text{off}}(a)\|a\|^2_{V^{-1}_{\tilde{\pi}^\star_l}}, \tag{5}$$

and for all $\pi \in \Delta(\mathcal{A}_l)$:

$$d \leq (1-\alpha)g_{\mathcal{A}_l}((1-\alpha)\pi + \alpha\pi_{\text{off}}) + \alpha \sum_{a \in \mathcal{A}} \pi_{\text{off}}(a)\|a\|^2_{V^{-1}_{\tilde{\pi}^\star_l}}. \tag{6}$$

Using these relations, we have that $(1-\alpha)g(\tilde{\pi}^\star_l) \leq d_{\text{eff}}$.

The lemma relates the maximum confidence width $g(\tilde{\pi}^\star_l)$ of the OLS estimator with the effective dimension d_{eff}. We now present our main theorem which provides a bound on the regret of OOPE.

Theorem 4.4 (Regret Bound)**.** *The OOPE algorithm satisfies the following regret bound with probability $1 - \frac{1}{T}$,*

$$\mathcal{R}(\text{OOPE}) \leq 16\sqrt{6d_{\text{eff}}T\log(4l^2_{max}|\mathcal{A}|T)} + 4d(d+1) \tag{7}$$

where $l_{max} = \left\lfloor \log_2 \sqrt{4 + \frac{T}{d_{\text{eff}}\log(4|\mathcal{A}|T)}} \right\rfloor$.

The proof of Theorem 4.4 is presented in the Appendix B.4. Note that in the pure online setting as there is no offline data we have $d_{\text{eff}} = d$ and we recover the online regret bound. The proof of Theorem 4.4 crucially relies on the confidence width $g(\tilde{\pi}_l^\star)$, with a tight bound on it made availble by Lemma 4.3.

Note that the regret analysis can be extended to compact continuum actions sets by approximating them by ϵ-nets. Additionally, OOPE can be made horizon free by incorporating the doubling trick with full restart as described in (Besson & Kaufmann, 2018).

4.3. Lower bound for minimax regret.

We informally define our problem class $\mathcal{P}^d_{v,T_{\text{off}},T}$ to be characterised by the parameters d, T, T_{off} and the d-dimensional vector v, where $v_i = \frac{\lambda_i(V_{\pi_{\text{off}}})}{\max_{a\in\mathcal{A}}||a||^2}$, is the vector of normalized eigenvalues of $V_{\pi_{\text{off}}}$. Intuitively the parameters T_{off} and v captures the amount and quality of the offline data respectively. For a more formal definition please see section B.5. We further assume that every bandit instance in $\mathcal{P}^d_{v,T_{\text{off}},T}$ is such that $|\mathcal{A}| \leq O(d^d)$. For this class, one defines the minimax regret $\mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T})$ informally as the best possible regret for the worst problem instance from $\mathcal{P}^d_{v,T_{\text{off}},T}$. [2]

Proposition 4.5. *For any problem class $\mathcal{P}^d_{v,T_{off},T}$ with $|\mathcal{A}| \leq O(d^d)$ we have that:*

$$\mathcal{R}_{minmax}(\mathcal{P}^d_{v,T_{off},T}) \geq \frac{\sqrt{T}exp(-2)}{8} \sup_{\substack{w\in\Delta_d \\ \forall i, w_i \geq v_i}} \sum_{i=1}^{d} \frac{1}{\sqrt{1+\frac{T_{off}}{T}\frac{v_i}{w_i}}}$$

where Δ_d is d-dimensional simplex.

The proof is provided in the Appendix B.5. The crux of the proof is finding a hard instance by perturbing the online hard instance of hypercubes (see Chapter 24 Lattimore & Szepesvári (2020)) based on the offline spectrum v_i's. A simple approach of rotating the hypercubes $\mathcal{A}$ and Θ by the eigenvectors of $V_{\pi_{\text{off}}}$ will become inconsistent due to the circular dependence of $V_{\pi_{\text{off}}}$ on $\mathcal{A}$. We carefully construct a hard instance that is consistent while perturbing the hypercube based on v_i's.

The lower bound is a concave optimization program. We are able to characterize its dual in Lemma B.10. Utilizing this dual, we can show (see Lemma B.11) that for $k \in [d]$, such that $v_i = 0$ for $i < k$ and $0 < v_i$ for $i \geq k$:

$$\frac{(1-\sum_i v_i)T_{\text{off}}}{2v_d(1+\frac{T_{\text{off}}}{T})^{3/2}T} + (k-1) + \frac{(d-k+1)}{\sqrt{1+\frac{T_{\text{off}}}{T}}} \geq \sup_{\substack{w\in\Delta_d \\ \forall i, w_i \geq v_i}} \sum_{i=1}^{d} \frac{1}{\sqrt{1+\frac{T_{\text{off}}}{T}\frac{v_i}{w_i}}} \geq (k-1) + \frac{(d-k+1)}{\sqrt{1+\frac{T_{\text{off}}(\sum_i v_i)}{T}}}. \tag{8}$$

Now consider the case when the offline data is well explored, i.e, $g(\pi_{\text{off}}) = O(d)$[3] or equally $v_i = \Omega(1/d)$ for all i, with $T_{\text{off}} \gg T$ then the effective dimension is $d_{\text{eff}} \leq O(dT/T_{\text{off}})$. Substituting $|A| = d^d$ in Theorem 4.4, gives us an upper bound $\mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T}) \leq O(dT\log(Td)/\sqrt{T+T_{\text{off}}} + d^2)$. Using the bounds in (8) (here $k = 1$) we obtain that $\mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T}) \geq \Omega(dT/\sqrt{T+T_{\text{off}}})$. OOPE is thus, minimax optimal up to logarithmic factors and an additive constant when large amount of offline data is well explored.

In the case where there are lots of poorly explored directions in offline data, that is $k = \Omega(d)$ in above, we get that $\mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T}) \geq \Omega(d\sqrt{T})$. Applying again, Theorem 4.4 the OOPE regret bound we get that $d_{\text{eff}} = \theta(d)$ and hence $\mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T}) \leq O(d\sqrt{T})$ and OOPE is again minimax optimal in this regime. We summarise the above discussion:

Corollary 4.6. *For problem classes $\mathcal{P}^d_{v,T_{off},T}$ with $T = \Omega(max(d\sqrt{T_{off}}, d^3))$, where either the offline data is well explored, that is $g(\pi_{off}) = \theta(d)$ or when the offline data is poorly explored, that is $k = \Omega(d)$ in (8), then OOPE is minimax optimal (modulo logarithmic factors and an additive constant) over these problem classes.*

A detailed calculation for the corollary is carried out in section B.7. In the case where there are only a few poorly explored directions, that is $k = o(d)$ in (8), there is a gap in the upper and lower bounds: $\tilde{O}(\sqrt{dkT}) \geq \mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T}) \geq \theta(k\sqrt{T})$. We believe this slack comes from weakening of the upper bound on the confidence width $g(\tilde{\pi}^*)$ in analysis of Theorem 4.4 in this regime. Table 1 summarizes the above discussion.

Offline Regime	$g(\pi_{\text{off}})$	$\mathcal{R}_{\text{OOPE}}$ (Upper)	Lower Bound
Well-explored ($k = 1$)	$\theta(d)$	$\tilde{O}(\frac{dT}{\sqrt{T+T_{\text{off}}}})$	$\Omega(\frac{dT}{\sqrt{T+T_{\text{off}}}})$
Moderately-explored ($k = o(d)$)	∞	$\tilde{O}(\sqrt{dkT})$	$\Omega(k\sqrt{T})$
Under-explored ($k = \Omega(d)$)	∞	$\tilde{O}(d\sqrt{T})$	$\Omega(d\sqrt{T})$

Table 1. Upper and lower bounds in various offline regimes. Here, k denotes the number of under-explored directions.

Remark 4.7 (Gap between OOPE's upper bound and the lower bound). We believe the gap is purely analytical and

[2] see eqn. (23) in section B.5.

[3] This happens when all directions are well explored in offline data, for *e.g.,* uniform offline arm pulls in orthogonal action sets.

essentially boils down to how we upper-bound the term $\lambda_d(V_{\pi^*_{l,on}})$ by $\max_a ||a||^2$ in proof of Lemma 4.3 in Appendix B.3. This bound does not reflect the dependence on the offline eigenspectrum and is just a uniform bound. This bound can be tight when the offline data has many underexplored directions, but it is quite weak in more well-explored settings. Improving this bound, to incorporate the offline spectrum in a more nuanced way, is challenging, and we leave it as further work.

Comparison with warm-started LinUCB: One approach to incorporate offline data into LinUCB is to create the initial confidence ellipsoid utilizing all the offline data, and pull the arms which maximize the UCB index in the subsequent online rounds. In Proposition B.7, we present the existing regret guarantees for warm started LinUCB. In Appendix B.8, we present a simple problem setting where this regret bound is $d^{1/2}$ worse than the regret bound of OOPE. The LinUCB bound has a dependence on $||\theta||_{V_0}$, where V_0 is the initial Gram matrix. When we "warm-start" LinUCB with $V_0 = T_{\text{off}} V_{\pi_{\text{off}}} + I$, $||\theta||_{V_0}$ on the hard instance of $\mathcal{A} = \{\pm 1\}^d, \Theta = \{\pm\sqrt{\frac{d}{(T_{\text{off}}+T)}}\}^d$ can become $\Omega(d)$. This is an instance where LinUCB, either by warm-starting or by cold-starting ($V_0 = I$) attains the same regret bound $O(d^{3/2}T/\sqrt{T_{\text{off}}})$ while in contrast OOPE gets $O(\frac{dT}{\sqrt{T_{\text{off}}}})$ rates.

Remark 4.8. In practice, OOPE has certain additional benefits over LinUCB. First is computational in nature. Each iteration of LinUCB requires computing the UCB index [4] for each sample. Whereas in OOPE, the D-optimal design problem is only solved $\log(T + T_{\text{off}})$ times at the beginning of each phase. This makes OOPE computationally more efficient than UCB and hence appealing in practice.

5. Improving the dimension dependence with Frank-Wolfe (FW)

The bound in Theorem 4.4 has an additional $O(d^2)$ term. This is due to the support of $\pi^*_{l,\text{on}}$, which is at most $d(d+1)/2$. In some scenarios, this term in the regret can be important, for instance, if $d_{\text{eff}} = \Omega(1)$, $|\mathcal{A}| = \Omega(d^2)$ and $T = O(d^2)$. Thus, this term can be a source of regret even if the typical dominant term $\tilde{O}(\sqrt{d_{\text{eff}}T})$ is small. To address this, we compute an ϵ-approximate solution to (2) using Frank-Wolfe that has $O(d/\epsilon \log\log(d))$ support points while ensuring the regret has only increased by at most $\sqrt{d\epsilon T}$ when using this approximate exploration schedule. We are trading the regret benefits of the optimal exploration for the smaller support size of the approximate solution.

To approximately solve the optimization in (2) we will use a version of Frank-Wolfe (FW) algorithm where at each update step at most one new arm is added. We will start with a carefully chosen initialization that has $O(d)$ support, and show that FW converges to an ϵ-approximate solution in $O(d\log\log(d) + d/\epsilon)$ steps ensuring the second term in the regret is $O(d\log\log(d) + d/\epsilon)$ rather than $O(d^2)$ as in the previous section. This yields the following improved bound in the regime when $d = \Omega(1)$ which we state informally.

Theorem (Informal Improved support regret Bound). The OOPE algorithm, where each phase l uses Frank-Wolfe iterations upto an accuracy $\epsilon = d_{\text{eff}}/d$, obtains a regret of $\tilde{O}(\sqrt{d_{\text{eff}}T\log(|\mathcal{A}|T)}) + \frac{d^2}{d_{\text{eff}}})$ with prob. $1 - \frac{1}{T}$.

When $d_{\text{eff}} = \Omega(1)$, the above regret bound is a strict improvement over the regret of OOPE.

Dual problem: An important tool in this analysis is the dual problem of (2). The dual of the maximization (2) is a minimum volume ellipsoid problem but the convex constraints are not the usual enclosing of arms condition that arise in the purely online setting (see chapter 2 of (Todd, 2016) for more details of the pure online case.). We recover the usual dual of Minimum Volume Enclosing Ellipsoid (MVEE) problem when we set $\alpha = 0$ in the following lemma (proof in Appendix C.1).

Lemma 5.1. *(Strong Duality) Consider the minimization problem*

$$\mathcal{P}(\mathcal{A}_l, \alpha) := \min_{H \succ 0} \quad -\log(\det(H))$$

such that for all $a \in \mathcal{A}_l$,

$$(1-\alpha)a^t H a + \alpha Tr(V_{\pi_{off}} H) \leq d, \tag{9}$$

This is dual to the optimization problem in (2) given by :

$$\mathcal{D}(\mathcal{A}_l, \alpha) := \max_{\pi \in \Delta(\mathcal{A}_l)} \log\left(\det\left((1-\alpha)V_\pi + \alpha V_{\pi_{off}}\right)\right),$$

that is, $\mathcal{P}(\mathcal{A}_l, \alpha) = \mathcal{D}(\mathcal{A}_l, \alpha)$.

Initialization and its Information Gain bound: The initialization of Frank-Wolfe is the construction given in (Kumar & Yildirim, 2005). The pseudocode for the initialization is in Appendix C.2. At a high level, the initialization choose d arms, each of which optimizes a well chosen linear function. Once the d support points (arms) are selected, we put a uniform measure on them and use it as our initialization $\pi_l^{(0)}$ for the Frank-Wolfe procedure. For any exploration phase l, define the *information gain* function:

$$d(\pi, \mathcal{A}_l, \alpha) := \log(\det(\alpha V_{\pi_{\text{off}}} + (1-\alpha)V_\pi), \tag{10}$$

where $\pi \in \Delta(\mathcal{A}_l)$. We have the following bound on $d(\pi_l^{(0)}, \mathcal{A}_l, \alpha)$:

[4] The rarely switching version (Theorem 4 of Abbasi-Yadkori et al. (2011)) requires $O(d\log(T + T_{\text{off}}))$ updates that has a d dependence that OOPE does not have.

Proposition 5.2. *For exploration phase l, let as before $\pi^*_{l,on}$ denote the optimal solution to (2), then with the above initialization $\pi_l^{(0)}$ we have: $d(\pi^*_{l,on}, \mathcal{A}_l, \alpha) - d(\pi_l^{(0)}, \mathcal{A}_l, \alpha) \leq d \log \frac{d^5}{(1-\alpha)}$.*

The proof of the proposition is in Appendix C.3. The proposition shows that the initialization is not too far from the optimal value with only $O(d)$ support points. The dual transforms the problem of maximizing the information gain into a geometric problem of minimizing the volume of an ellipsoid subject to certain constraints (9). The proof relates this geometric problem to the usual MVEE by means of a new feasibility relation amongst these two class of problems and scale invariance of the volume of MVEE (see Appendix C.7 and C.8). The use of feasibility relation and scale invariance is a novel addition to the typical online FW analysis on these types of problems (see (Todd, 2016)). Finally we bound this transformed MVEE using the following property of the initialization - $vol(conv(B)) \geq \frac{1}{d!} vol(conv(\mathcal{A}_l))$ (see Proposition C.3) where B is the support of initialization $\pi_l^{(0)}$ obtained from $\mathcal{A}_l$.

5.1. Frank-Wolfe (FW) iterations after initialization

In this section we show that performing $t = \tilde{O}(d)$ iterations starting from $\pi_n^{(0)}$ is enough to guarantee a good solution π_n^t with only $\tilde{O}(d)$ support. We first specify the FW updates. Then, we describe a potential function that FW update implicitly keeps tracks that directly translates to tighter regret bounds after $t = \tilde{O}(d)$ iterations.

Definition 5.3. If π is probability distribution on $\mathcal{V} \subseteq \mathcal{A}$ and if $(1-\alpha)\sum_a \pi(a)aa^t + \alpha V_{\pi_{\text{off}}}$ is non singular then define $H(\pi) := \left((1-\alpha)\sum_a \pi(a)aa^t + \alpha V_{\pi_{\text{off}}}\right)^{-1}$.

Frank-Wolfe Algorithm: We start from the initialization $\pi_l^{(0)}$ and apply Frank-Wolfe (FW) update specified in Algorithm 2.

The FW update adds atmost only one new arm in the support of the solution in each iteration. The arm added has the largest slack as defined in line 3 of Algorithm 2.

Slack in (6) as potential: In the previous section, we saw that the equality in Lemma 4.3 is crucial for establishing the regret bound in terms of d_{eff}. From the same Lemma, a sub-optimal online design $\pi \in \Delta(\mathcal{A}_l)$ would satisfy the inequality 6. Let $\tilde{\pi} = (1-\alpha)\pi + \alpha\pi_{\text{off}}$. In our FW updates, we track the *potential*:

$$\delta(\pi_l^{(t)}) = \frac{(1-\alpha)g(\tilde{\pi}_l^{(t)}) + \alpha \sum_{a\in\mathcal{A}} \pi_{\text{off}}(a)\|a\|^2_{V^{-1}_{\tilde{\pi}_l^{(t)}}}}{d} - 1,$$

which is precisely the slack in (6) for $\pi_l^{(t)}$. It can also be re-written as $\delta(\pi_l^{(t)}) = \frac{w_{a+}(\pi_n)}{d} - 1$ (Lines 3 or 8 of Algorithm 2). We next show that, if $\delta(\pi_l^{(t)})$ is large, then the FW update has a larger per iteration improvement in its information gain function (its objective) $d(\pi_l^{(t)}, \mathcal{A}_l, \alpha)$. Formally,

Algorithm 2 FRANK-WOLFE FOR OO SETTING

1: **Input:** $\epsilon, \mathcal{A}_l, \pi_l^{(0)}, V_{\pi_{\text{off}}}, \alpha$.
2: $t \leftarrow 0$.
3: $w \leftarrow \left(Tr\left(H(\pi_l^{(t)})\left((1-\alpha)aa^t + \alpha V_{\pi_{\text{off}}}\right)\right)\right)_{a\in\mathcal{A}_l}$;
$\delta(\pi_l^{(t)}) = \frac{\max_a w_a}{d} - 1$, $a_+ \leftarrow \underset{a}{argmax} \quad w_a$.
4: **while** $\epsilon < \delta(\pi_l^{(t)})$ **do**
5: $\quad \beta \leftarrow \frac{(w_{a_+} - d)}{(d-1)w_{a_+}}$, $\pi^{(+)} \leftarrow (1+\beta)^{-1}(\pi_l^{(t)} + \beta \mathbb{1}_{\{a_+\}})$.
6: $\quad t \leftarrow t+1$; $\pi_l^{(t)} \leftarrow \pi^{(+)}$.
7: $\quad w \leftarrow \left(Tr\left(H(\pi_l^{(t)})\left((1-\alpha)aa^t + \alpha V_{\pi_{\text{off}}}\right)\right)\right)_{a\in\mathcal{A}_l}$.
8: $\quad \delta(\pi_l^{(t)}) = \frac{\max_a w_a}{d} - 1$, $a_+ \leftarrow \underset{a}{argmax} \quad w_a$.
9: **end while**
10: **return:** $\pi_l^{(t)}$.

Lemma 5.4. *In the phase l, the per-iteration improvement in information gain of the FW update is given by:*

$$d(\pi_l^{(t+1)}, \mathcal{A}_l, \alpha) - d(\pi_l^{(t)}, \mathcal{A}_l, \alpha) \geq m(\delta(\pi_l^{(t)})) := \log(\delta(\pi_l^{(t)})) - \frac{\delta(\pi_l^{(t)})}{1+\delta(\pi_l^{(t)})}$$

The proof is presented in Appendix C.4. A novel step in proving this lemma is using a *reverse* Jensen inequality from (Merhav, 2022) to lower bound the FW update improvement in the presence of offline data unlike the typical analysis done in (Todd, 2016).

Bounding number of iterations to achieve slack $\delta(\pi_l^{(t)}) \sim O(1)$: Using the properties of the function $m(\delta)$ and Lemma 5.4, we bound the number of iterations of FW needed to get a small enough potential $\delta(\pi_l^{(t)})$ as follows:

Proposition 5.5. *The number of iterations required for the FW updates in Algorithm 2 to reach an iterate with slack $\delta(\pi_l^{(t)}) < \delta_0$ from $\pi_l^{(0)}$ is at most*

$$t = \frac{d}{1 - \frac{\delta_0}{(1+\delta_0)\log(1+\delta_0)}} \log\left(\frac{d}{\delta_0}\log\left(\frac{d^5}{1-\alpha}\right)\right).$$

The approximation $\pi_l^{(t)}$ has the property $|\text{supp}(\pi_l^{(t)})| \leq t + d$ and satisfies the following inequalities:

$$d \leq (1-\alpha)g(\tilde{\pi}_l^{(t)}) + \alpha \sum_{a\in\mathcal{A}} \pi_{off}(a)\|a\|^2_{V^{-1}_{\tilde{\pi}_l^{(t)}}} \leq d(1+\delta(\pi_l^{(t)})),$$

where $\tilde{\pi}_l^{(t)} = (1-\alpha)\pi_l^{(t)} + \alpha\pi_{off}$.

The proof is given in Appendix C.5. The convergence rate does slow down as δ_0 approaches zero as shown in the following lemma:

Theorem (Lemma 3.9, (Todd, 2016)). The number of iterations to reduce the slack from $0 < \delta_0 < 1$ to $\delta_0/2$ is at most $\frac{14d}{\delta_0}$.

Although proof in (Todd, 2016) is for the pure online case, it is straightforward to modify it to obtain the same result in the OO setting and is omitted. One can thus start with the (Kumar & Yildirim, 2005) initialization (Algorithm 3) and run the FW Algorithm 2. We can split the upper bound analysis of number of iterations into two phases : one where $\delta(\pi_l^{(t)}) \geq 1$ and the second where $\delta(\pi_l^{(t)}) < 1$. In the first phase from Proposition 5.5 (we set $\delta_0 = 1$) we get the iterations is atmost $4d\log(d\log(d^5/(1-\alpha))) + d$. In the second phase the number of iterations is bounded by $14d(1 + 2 + 2^2 + \cdots + 2^k)$ where $k = \log_2(1/\epsilon)$ to get $28d/\epsilon$ iterations. We have shown that:

Corollary 5.6. *The total number of iteration FW takes with (Kumar & Yildirim, 2005) initialization to attain a slack of $\epsilon < 1$ is $4d\log(d\log(d^5/(1-\alpha))) + d + \frac{28d}{\epsilon}$.*

5.2. OOPE-FW and its Regret with improved dependence on d.

OOPE-FW Algorithm: We modify the OOPE Algorithm 1, where in each iteration instead of solving the optimization 2 exactly we solve it approximately in each phase l. We use the (Kumar & Yildirim, 2005) initialization (Algorithm 3) on the live arms $\mathcal{A}_l$ and then run the FW updates (Algorithm 2) from this initialization till a slack $\delta(\pi_l^{(t)}) \leq \frac{d_{\text{eff}}}{d}$ is reached. One has the following bound for the confidence-width $g_{\mathcal{A}_l}(\tilde{\pi}_l^{(t)})$ when we use the initialization 3 with FW (Algorithm 1) updates:

Proposition 5.7. *When OOPE-FW with FW iterations run till $\delta(\pi_l^{(t)}) \leq \frac{d_{\text{eff}}}{d}$ we have that:*

$$(1-\alpha)g_{\mathcal{A}_l}(\tilde{\pi}_l^{(t)}) \leq 2d_{\text{eff}} \tag{11}$$

where $\tilde{\pi}_l^{(t)} = (1-\alpha)\pi_l^{(t)} + \alpha\pi_{\text{off}}$.

The proof is presented in Appendix C.6. In each phase l of OOPE-FW we use online samples:

$$n^l_{\text{on},fw}(a) = \left\lceil \frac{6d_{\text{eff}}\pi_l^{(t)}(a)\log(4l^2|\mathcal{A}|T)}{\epsilon_l^2} \right\rceil, \tag{12}$$

and offline samples:

$$n^l_{\text{off},fw}(a) = \left\lceil \frac{2\alpha\pi_{\text{off}}(a)g(\tilde{\pi}_l^{(t)})\log(4l^2|\mathcal{A}|T)}{\epsilon_l^2} \right\rceil. \tag{13}$$

These particular number of samples ensure the requisite concentration inequalities hold. The OOPE-FW algorithm does not demand excess offline samples and online samples just like OOPE. We can derive a similar result to Proposition 4.2 for OOPE-FW, which implies offline sample for each arm a does not exhaust until the last phase (when the online samples are exhausted).

We present the regret bound of OOPE-FW:

Theorem 5.8 (OOPE-FW Regret Bound)**.** *The OOPE-FW algorithm has the following regret bound hold with probability $1 - \frac{1}{T}$,*

$$\mathcal{R}(\text{OOPE-FW}) \leq 32\sqrt{3d_{\text{eff}}T\log(4l_{max}^2|\mathcal{A}|T)} + 8d + \frac{224d^2}{d_{\text{eff}}} + 32d\log\left(d\log\left(\left(\frac{d^5(T+T_{\text{off}})}{T}\right)\right)\right)$$

where d_{eff} is the effective dimension defined in (1) and l_{max} is the same as defined in Theorem 4.4.

The proof is given in Appendix C.7. The first term $\tilde{O}(\sqrt{d_{\text{eff}}T})$ has an additional constant $\sqrt{2}$ in OOPE-FW compared to OOPE. However, when $d_{\text{eff}} = \Omega(1)$, the above the bound on the second term is a strict improvement over the regret of OOPE. One can of course choose the version of phased elimination for the problem parameters (like $d, T, T_{\text{off}}, \pi_{\text{off}}$) and choose between OOPE and OOPE-FW. We record it as a corollary:

Corollary 5.9. *Using the appropriate phased elimination variant in OO setting, we can obtain with probability at least $1 - \frac{1}{T}$ a regret of $\tilde{O}\left(\sqrt{d_{\text{eff}}T} + \min\{d^2, \frac{d^2}{d_{\text{eff}}}\}\right)$.*

6. Experiments

Simulation Setting. Across all experiments, the unknown parameter θ^* and the action set $\mathcal{A}$ are sampled uniformly from the unit sphere in $\mathbb{R}^d$. Reward noise is drawn from a standard normal distribution $\mathcal{N}(0,1)$. The offline dataset $\mathcal{D}_{\text{off}}$ is generated by uniformly partitioning T_{off} samples across a randomly chosen subset of arms (denoted as n_{support}, where $2 \leq n_{\text{support}} \leq \min\{|\mathcal{A}|, T_{\text{off}}\}$). All reported metrics are averaged over 50 independent runs, with 95% confidence intervals shaded.

Improved Performance with Offline Data. First, we evaluate the baseline performance of OOPE as the volume of offline data increases. We set $d = 20, |\mathcal{A}| = 40, T = 10^4$, and $n_{\text{support}} = 40$. We fix the offline logging distribution π_{off} (generated uniformly with $T_{\text{off}} = 10^5$) and scale T_{off} down for shorter horizons to maintain the same π_{off}.

Figure 1 shows that OOPE performs better, i.e. incurs lower regret, as the number of offline samples increases. As a basic baseline, we compare with pure online Phased Elmination (dashed red). For this experiment we first sample a random partition uniformly with $T_{\text{off}} = 10^5$ and

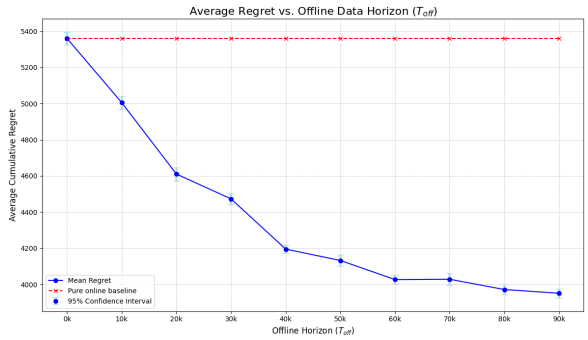

Figure 1. (**Improved performance with increasing offline data**) Plot showing lower regret with increasing offline samples for a fixed π_{off}. A purely online baseline is presented for comparison.

compute π_{off}. We regenerate offline data again for the shorter offline horizons holding the π_{off} as fixed.

Comparison against Warm-Started Baselines. Next, we compare OOPE against standard heuristics: warm-started LinUCB and warm-started Thompson Sampling (LinTS). For LinUCB, the Gram matrix is initialized as $V_0 = T_{\text{off}} V_{\pi_{\text{off}}} + I$. For LinTS, the Gaussian prior is initialized using the OLS estimate of θ and the variance-covariance matrix from $\mathcal{D}_{\text{off}}$. We use $T_{\text{off}} = 10^5$. As shown in Figure 2, OOPE significantly outperforms both warm-started LinUCB and LinTS, validating our theoretical claim that simple warm-starting yields sub-optimal exploration trade-offs in linear regimes.

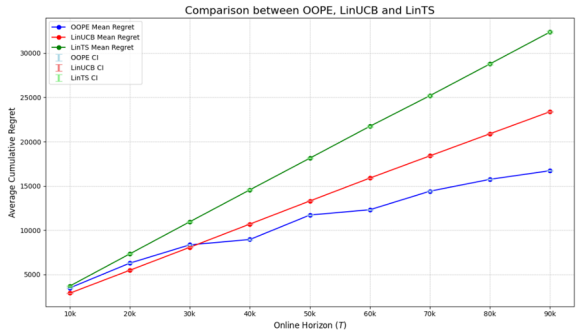

Figure 2. (**Comparison of OOPE versus warm-started LinUCB and LinTS.**) Plot showing better performance of OOPE.

High-Dimensional Efficiency: OOPE vs. OOPE-FW. As discussed in Section 5, the $O(d^2)$ support size of the exact D-optimal design can dominate regret in settings with large $\mathcal{A}$, moderate effective dimension d_{eff}, and longer horizons. To evaluate the efficiency of our Frank-Wolfe approximation, we systematically vary d_{eff} (by varying T_{off}) for a fixed offline eigenspectrum spread ($\kappa = \lambda_{\max}/\lambda_{\min} \approx 22.6$). We plot the performance gap, defined as $\Delta\text{Regret} = \mathcal{R}(\texttt{OOPE}) - \mathcal{R}(\texttt{OOPE-FW})$.

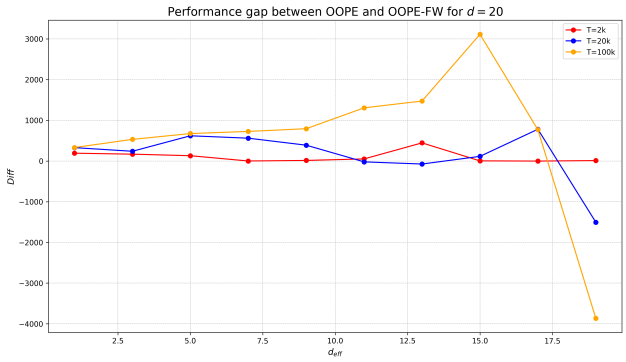

Figure 3. **Performance Gap (Δ Regret).** Positive values indicate OOPE-FW outperforms OOPE. OOPE-FW excels for moderate d_{eff} and longer horizons, while OOPE is better suited for highly under-explored (large d_{eff}) offline data.

Figure 3 ($d = 20, |\mathcal{A}| = 400, n_{\text{support}} = 60$) shows this gap across three online horizons: short ($T = 2,000$), medium ($T = 20,000$), and long ($T = 100,000$). For moderate d_{eff}, OOPE-FW strictly outperforms OOPE (positive Δ Regret), as the d^2/d_{eff} support scaling is vastly superior. Conversely, when d_{eff} is very large (approaching d), the exact design of OOPE wins out because the overhead constant of the FW approximation begins to dominate. This perfectly aligns with our theoretical bounds, confirming that OOPE-FW is the algorithm of choice for high-dimensional tasks with moderate offline data quality.

7. Conclusion and Future Work

We propose a phased elimination algorithm OOPE, based on a generalized notion of D-optimal design, which achieves significantly lower regret in OO setting. We identify an *effective* dimension (d_{eff}) based on the offline Gram eigenspectrum that quantitatively captures this. Consequently we obtain an improved regret of $O(\sqrt{d_{\text{eff}} T \log(|\mathcal{A}|)} + d^2)$. This matches a novel minimax lower bound upto $\log(dT)$ factors and additive constants in *well & poorly* explored offline data regimes. In settings with small T, d_{eff} and large number of arms, the $O(d^2)$ support term might dominate the $\tilde{O}(\sqrt{d_{\text{eff}} T})$. To overcome this, we propose a Frank-Wolfe variant of OOPE, called OOPE-FW. Our theoretical insights are further validated with synthetic numerical experiments.

Future Work. The current work assumes a stochastic offline data generation process. It will be useful to relax this assumption and study the case when offline data comes from adaptive policies. In many practical situations the online data comes from a slightly shifted θ^* vis-a-vis the offline data, with a certain shift budget known a priori (see for example Cheung & Lyu (2024)). It will be important to study extensions of OO setting with such drift.

References

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

Agnihotri, A., Jain, R., Ramachandran, D., and Wen, Z. Online bandit learning with offline preference data. *arXiv preprint arXiv:2406.09574*, 2024.

Agrawal, S., Juneja, S., Shanmugam, K., and Suggala, A. S. Optimal best-arm identification in bandits with access to offline data. *arXiv preprint arXiv:2306.09048*, 2023.

Balcan, M.-F., Harris, K., Khodak, M., and Wu, Z. S. Meta-learning adversarial bandits. *arXiv preprint arXiv:2205.14128*, 2022.

Ball, P. J., Smith, L., Kostrikov, I., and Levine, S. Efficient online reinforcement learning with offline data. *arXiv preprint arXiv:2302.02948*, 2023.

Banerjee, S., Sinclair, S. R., Tambe, M., Xu, L., and Yu, C. L. Artificial replay: a meta-algorithm for harnessing historical data in bandits. *arXiv preprint arXiv:2210.00025*, 2022.

Besson, L. and Kaufmann, E. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Betke, U. and Henk, M. Approximating the volume of convex bodies. *Discrete & Computational Geometry*, 10: 15–21, 1993.

Bu, J., Simchi-Levi, D., and Xu, Y. Online pricing with offline data: Phase transition and inverse square law. In *International Conference on Machine Learning*, pp. 1202–1210. PMLR, 2020.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pp. 41–1. JMLR Workshop and Conference Proceedings, 2012.

Cai, C., Cai, T. T., and Li, H. Transfer learning for contextual multi-armed bandits. *arXiv preprint arXiv:2211.12612*, 2022.

Cheng, C.-A., Xie, T., Jiang, N., and Agarwal, A. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pp. 3852–3878. PMLR, 2022.

Cheung, W. C. and Lyu, L. Leveraging (biased) information: Multi-armed bandits with offline data. *arXiv preprint arXiv:2405.02594*, 2024.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Cutkosky, A., Dann, C., Das, A., and Zhang, Q. Leveraging initial hints for free in stochastic linear bandits. In *International Conference on Algorithmic Learning Theory*, pp. 282–318. PMLR, 2022.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. 2008.

Fedorov, V. V. *Theory of optimal experiments*. Elsevier, 2013.

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

Hamidi, N. and Bayati, M. On worst-case regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*, 2020.

Hao, B., Jain, R., Lattimore, T., Van Roy, B., and Wen, Z. Leveraging demonstrations to improve online learning: Quality matters. In *International Conference on Machine Learning*, pp. 12527–12545. PMLR, 2023.

Kapp, M. B. Ethical and legal issues in research involving human subjects: do you want a piece of me? *Journal of clinical pathology*, 59(4):335, 2006.

Kausik, C., Tan, K., and Tewari, A. Leveraging offline data in linear latent bandits. *arXiv preprint arXiv:2405.17324*, 2024.

Kiefer, J. Optimum experimental designs v, with applications to systematic and rotatable designs. In *Proc. 4th Berkeley Symp*, volume 1, pp. 381–405, 1960.

Korda, N., Szorenyi, B., and Li, S. Distributed clustering of linear bandits in peer to peer networks. In *International conference on machine learning*, pp. 1301–1309. PMLR, 2016.

Kumar, P. and Yildirim, E. A. Minimum-volume enclosing ellipsoids and core sets. *Journal of Optimization Theory and applications*, 126(1):1–21, 2005.

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

Lazaric, A., Brunskill, E., et al. Sequential transfer in multi-armed bandit with finite set of models. *Advances in Neural Information Processing Systems*, 26, 2013.

Li, G., Zhan, W., Lee, J. D., Chi, Y., and Chen, Y. Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. *arXiv preprint arXiv:2305.10282*, 2023.

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

Mannor, S. and Shamir, O. From bandits to experts: On the value of side-observations. *Advances in neural information processing systems*, 24, 2011.

Merhav, N. Reversing jensen's inequality for information-theoretic analyses. *Information*, 13(1):39, 2022.

Oetomo, B., Perera, R. M., Borovica-Gajic, R., and Rubinstein, B. I. Cutting to the chase with warm-start contextual bandits. *Knowledge and Information Systems*, pp. 1–33, 2023.

Osadchiy, I., Levy, K. Y., and Meir, R. Online meta-learning in adversarial multi-armed bandits. *arXiv preprint arXiv:2205.15921*, 2022.

Rajaraman, N., Yang, L., Jiao, J., and Ramchandran, K. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33: 2914–2924, 2020.

Rakhlin, A. and Sridharan, K. Online learning with predictable sequences. In *Conference on Learning Theory*, pp. 993–1019. PMLR, 2013.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Sentenac, F., Lee, I., and Szepesvari, C. Balancing optimism and pessimism in offline-to-online learning. *arXiv preprint arXiv:2502.08259*, 2025.

Sharma, N., Basu, S., Shanmugam, K., and Shakkottai, S. Warm starting bandits with side information from confounded data. *arXiv preprint arXiv:2002.08405*, 2020.

Shivaswamy, P. and Joachims, T. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*, pp. 1046–1054. PMLR, 2012.

Soare, M., Alsharif, O., Lazaric, A., and Pineau, J. Multi-task linear bandits. In *NIPS2014 workshop on transfer and multi-task learning: theory meets practice*, 2014.

Song, Y., Zhou, Y., Sekhari, A., Bagnell, J. A., Krishnamurthy, A., and Sun, W. Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*, 2022.

Steinhardt, J. and Liang, P. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International conference on machine learning*, pp. 1593–1601. PMLR, 2014.

Tennenholtz, G., Shalit, U., Mannor, S., and Efroni, Y. Bandits with partially observable confounded data. In *Uncertainty in Artificial Intelligence*, pp. 430–439. PMLR, 2021.

Titterington, D. Optimal design: some geometrical aspects of d-optimality. *Biometrika*, 62(2):313–320, 1975.

Todd, M. J. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

Valko, M., Munos, R., Kveton, B., and Kocák, T. Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*, pp. 46–54. PMLR, 2014.

Wagenmaker, A. and Pacchiano, A. Leveraging offline data in online reinforcement learning. *arXiv preprint arXiv:2211.04974*, 2022.

Wang, Z., Zhang, C., Singh, M. K., Riek, L., and Chaudhuri, K. Multitask bandit learning through heterogeneous feedback aggregation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1531–1539. PMLR, 2021.

Wei, C.-Y. and Luo, H. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pp. 1263–1291. PMLR, 2018.

Wei, C.-Y., Luo, H., and Agarwal, A. Taking a hint: How to leverage loss predictors in contextual bandits? In *Conference on Learning Theory*, pp. 3583–3634. PMLR, 2020.

Xiao, C., Wu, Y., Mei, J., Dai, B., Lattimore, T., Li, L., Szepesvari, C., and Schuurmans, D. On the optimality of batch policy optimization algorithms. In *International Conference on Machine Learning*, pp. 11362–11371. PMLR, 2021.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

Yang, L., Tan, V. Y., and Cheung, W. C. Best arm identification with possibly biased offline data. *arXiv preprint arXiv:2505.23165*, 2025.

Zhang, C., Agarwal, A., Daumé III, H., Langford, J., and Negahban, S. N. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. *arXiv preprint arXiv:1901.00301*, 2019.

Zhou, Y., Sekhari, A., Song, Y., and Sun, W. Offline data enhanced on-policy policy gradient with provable guarantees. *arXiv preprint arXiv:2311.08384*, 2023.

A. Detailed Literature Review

We present here additional works that are relevant to regret minimization in linear bandits for the ○○ setting.

Linear Bandits with Dynamic Action sets: When the set of arms is dynamic and the losses chosen by an adversary, Chu et al. (2011) proposed the SUPLINUCB algorithm which achieves $O(\sqrt{dT\log^3|\mathcal{A}|T})$ regret. The EXP2 algorithm (with John's exploration) of Bubeck et al. (2012) achieves a regret of $O(\sqrt{dT\log|\mathcal{A}|T})$, for $T \geq d^2$. (Li et al., 2019) improved the analysis of SUPLINUCB and provided nearly matching minimax lower bounds.

Incorporating offline data in Contextual Bandits and Reinforcement Learning: (Cai et al., 2022) and (Kausik et al., 2024) consider regret minimization in contextual bandits for the MAB and linear bandit settings respectively. (Cai et al., 2022) studied the problem under a potential co-variate shift in the offline data and provided an algorithm that achieves the minimax optimal regret guarantees. This work, when specialized to the case of MAB in ○○ setting gives rates which are only tight for well explored offline data. (Kausik et al., 2024) studied the problem with a latent linear bandit setting and provide regret bounds that have improved dimensionality factors but the regret bound does not vanish with increasing quantity and quality of offline data. Warm-starting contexutal bandits with potentially shifted data was considered in (Zhang et al., 2019). However, as shown in Appendix A.1 of (Cheung & Lyu, 2024) the proposed algorithm in (Zhang et al., 2019) can have regret of $\tilde{O}(T^{2/3})$ in the ○○ setting. Agrawal et al. (2023) developed algorithms for Best Arm Identification (BAI) in fixed confidence setting with offline data for MABs while (Yang et al., 2025) studied the same problem with a potentially biased offline data.

The problem of leveraging offline data to improve the performance in the online phase has also been studied intensely in the Reinforcement Learning (RL) literature in the past few years under the name of hybrid-RL. Xie et al. (2021); Song et al. (2022); Wagenmaker & Pacchiano (2022); Ball et al. (2023); Li et al. (2023) used offline data for finding the optimal policy in online RL using as few samples as possible (similar to BAI in fixed confidence setting). In contrast, Zhou et al. (2023) derive high probability regret bounds under boundedness assumptions on coverage, transfer and concentration coefficients that are hard to verify in practice. Bu et al. (2020) analyzed the related dynamic pricing problem with offline data using a stylized demand model. Agnihotri et al. (2024), a follow-up work of (Hao et al., 2023), studies regret minimization where the offline data contains user preferences rather than reward feedback. They analyze a warm started posterior sampling and provide its Bayesian regret.

Offline Learning. In this problem, the learner is given an offline dataset consisting of trajectories of a Markov Decision Process (MDP) and has to identify the optimal policy with high accuracy. Unlike the ○○ setting, in offline learning there is no adaptive online learning phase. The goal is to optimize identification measures like simple regret and error probability of misidentifying the optimal policy. A number of works like Rajaraman et al. (2020); Rashidinejad et al. (2021); Xiao et al. (2021); Cheng et al. (2022) have shown that pessimistic algorithms (like Lower Confidence Bound (LCB)) have good guarantees in this setting.

Online Learning with Advice. Many works have considered online learning with additional information [5] in the full (for e.g., Rakhlin & Sridharan (2013); Steinhardt & Liang (2014)) and bandit (for e.g., Mannor & Shamir (2011); Wei & Luo (2018); Wei et al. (2020); Sharma et al. (2020); Tennenholtz et al. (2021); Cutkosky et al. (2022)) feedback settings respectively. In this problem before each arm-pull, the learner also has additional information about the unknown system that can be incorporated into its decision. The regret guarantees in the above works are typically weaker than prior work in ○○ setting (like Shivaswamy & Joachims (2012); Cheung & Lyu (2024); Sentenac et al. (2025)) when the offline data is well-explored and plentiful[6]. This is due to the additional information typically being less informative than good quality offline data.

Multi-task Bandit Learning. This is a related line of work on solving multiple bandit problems jointly by aggregating observed rewards where the underlying unknown parameters are shared or are similar in certain statistical sense. (Wang et al., 2021) solved the multi-armed bandit problem where the arm means are close. (Gentile et al., 2014; Korda et al., 2016) extended the problem to linear bandits and assume a cluster structure on parameters. (Lazaric et al., 2013; Soare et al., 2014; Osadchiy et al., 2022; Balcan et al., 2022) considered a sequential arrival of tasks in the context of multi-armed bandits and linear bandits. For a particular task, the samples from previous tasks can be represented as offline data but note that all

[5] variously called as advice, side-information or hints in the literature.

[6] Roughly, this is what we mean by *good quality.* offline data

samples have the flexibility of being chosen by the agent. In contrast, in our setting, we do not have any control over how the offline samples are being generated.

B. Proof of results in Section 4

B.1. Proof of Lemma 4.1.

From the definition of l_M and the number of online samples required ((3)) in each phase we have that:

$$\begin{aligned}
T &\geq \sum_{l=1}^{l_M} \sum_{a \in \mathcal{A}_l} n^l_{\text{on}}(a) \\
&\geq \sum_{l=1}^{l_M} \sum_{a \in \mathcal{A}_l} \frac{3d_{\text{eff}} \pi^*_{l,\text{on}}(a)}{\epsilon_l^2} \log(4l^2|\mathcal{A}|T) \\
&= \sum_{l=1}^{l_M} \frac{3d_{\text{eff}}}{\epsilon_l^2} \log(4l^2|\mathcal{A}|T) \\
&= 3d_{\text{eff}} \sum_{l=1}^{l_M} 4^l \log(4l^2|\mathcal{A}|T).
\end{aligned}$$

From this inequality and definition of the H^{-1} we have that:

$$l_M \leq H^{-1}\left(\frac{T}{3d_{\text{eff}}}\right).$$

B.2. Proof of Proposition 4.2.

It is useful to define the following function $H : \mathbb{N} \cup \{0\} \longrightarrow \mathbb{R}_{\geq 0}$ as:

$$H(n) = \sum_{l=1}^{n} 4^l \log(4l^2|\mathcal{A}|T).$$

Now for a given arm $a \in supp(\pi_{\text{off}})$ the number of samples used upto phase l_M is given by:

$$\begin{aligned}
\sum_{l=1}^{l_M} n^l_{\text{off}}(a) &= \sum_{l=1}^{l_M} \left\lceil \frac{2\alpha \pi_{\text{off}}(a) g(\tilde{\pi}^\star_l) \log(4l^2|\mathcal{A}|/\delta)}{\epsilon_l^2} \right\rceil \\
&\leq \sum_{l=1}^{l_M} \frac{2\alpha \pi_{\text{off}}(a) d_e \log(4l^2|\mathcal{A}|T)}{(1-\alpha)\epsilon_l^2} + l_M \\
&= \sum_{l=1}^{l_M} \frac{2T_{\text{off}} \pi_{\text{off}}(a) d_{\text{eff}} \log(4l^2|\mathcal{A}|T)}{T\epsilon_l^2} + l_M \\
&= \frac{2T_{\text{off}} \pi_{\text{off}}(a) d_{\text{eff}}}{T} \sum_{l=1}^{l_M} 4^l \log(4l^2|\mathcal{A}|T) + l_M \\
&= \frac{2T_{\text{off}} \pi_{\text{off}}(a) d_{\text{eff}}}{T} H(l_M) + l_M \\
&\leq \frac{2T_{\text{off}} \pi_{\text{off}}(a) d_{\text{eff}}}{T} \frac{T}{3d_{\text{eff}}} + l_M \\
&= \frac{2}{3} \pi_{\text{off}}(a) T_{\text{off}} + l_M.
\end{aligned}$$

where the last inequality uses Lemma 4.1. We next derive a bound on l_M in a similar way to Lemma 4.1.

$$\begin{aligned}
T &\geq \sum_{l=1}^{l_M} \sum_{a \in \mathcal{A}_l} n^l_{\text{on}}(a) \\
&\geq \sum_{l=1}^{l_M} \sum_{a \in \mathcal{A}_l} \frac{3 d_{\text{eff}} \pi^*_{l,\text{on}}(a)}{\epsilon_l^2} \log(4 l^2 |\mathcal{A}| T) \\
&\geq 3 d_{\text{eff}} \log(4|\mathcal{A}|T) \sum_{l=1}^{l_M} 4^l \\
&= 4 d_{\text{eff}} \log(4|\mathcal{A}|T)(4^{l_M} - 1).
\end{aligned}$$

Next we lower bound d_{eff} using its definition 1:

$$\sum_{k=1}^{d} \frac{1}{1 + \frac{T_{\text{off}}}{T} \frac{\lambda_k(V_{\pi_{\text{off}}})}{\max_a ||a||^2}} \geq \sum_{k=1}^{d} \frac{1}{1 + \frac{T_{\text{off}}}{T}} = \frac{dT}{T + T_{\text{off}}}$$

where we have used the fact that $\lambda_k(V_{\pi_{\text{off}}}) \leq \max_a ||a||^2$. Similarly:

$$\frac{T}{T_{\text{off}}} g_{\mathcal{A}}(\pi_{\text{off}}) = \frac{T}{T_{\text{off}}} \max_{a \in \mathcal{A}} \sum_{i=1}^{d} \frac{a_i^2}{\lambda_i(V_{\pi_{\text{off}}})} \geq \frac{T}{T_{\text{off}}},$$

where the inequality again follows from $\lambda_k(V_{\pi_{\text{off}}}) \leq \max_a ||a||^2$. Combining the above two inequalities with definition 1 we have that:

$$d_{\text{eff}} \geq \min\left\{ \frac{dT}{T + T_{\text{off}}}, \frac{T}{T_{\text{off}}} \right\}.$$

This implies that:

$$\left\lfloor \log_2 \sqrt{\frac{\max\{\frac{T + T_{\text{off}}}{d}, T_{\text{off}}\}}{4 \log(4|\mathcal{A}|T)} + 1} \right\rfloor \geq l_M$$

But from Coverage Assumption we know that the LHS is less than $\frac{\pi_{\text{off}}(a) T_{\text{off}}}{3}$. Thus one concludes that:

$$l_M \leq \frac{\pi_{\text{off}}(a) T_{\text{off}}}{3}.$$

Using this we have that :

$$\sum_{l=1}^{l_M} n^l_{\text{off}}(a) \leq \frac{2}{3} \pi_{\text{off}}(a) T_{\text{off}} + l_M \leq \frac{2}{3} \pi_{\text{off}}(a) T_{\text{off}} + \frac{\pi_{\text{off}}(a) T_{\text{off}}}{3} = \pi_{\text{off}}(a) T_{\text{off}}.$$

This concludes the proof of the proposition.

B.3. Proof of Lemma 4.3.

For any $\pi = (1 - \alpha)\pi_{l,\text{on}} + \alpha \pi_{\text{off}}$, with $\pi_{l,\text{on}} \in \Delta(\mathcal{A}_l)$ we have:

$$\begin{aligned}
\sum_{a \in \mathcal{A}} \pi(a) ||a||^2_{V_\pi^{-1}} &= \sum_{a \in \mathcal{A}} \pi(a) a^t V_\pi^{-1} a \\
&= \sum_{a \in \mathcal{A}} \pi(a) Tr(a a^t V_\pi^{-1}) \\
&= Tr\left(\left(\sum_{a \in \mathcal{A}} \pi(a) a a^t \right) V_\pi^{-1} \right) \\
&= Tr\left(V_\pi V_\pi^{-1} \right) \\
&= d.
\end{aligned}$$

But by definition of $g_{\mathcal{A}_l}(\pi) := \max_{a\in\mathcal{A}_l} ||a||^2_{V_\pi^{-1}}$ we have

$$d = \sum_{a\in\mathcal{A}} \pi(a)||a||^2_{V_\pi^{-1}} = (1-\alpha)\sum_{a\in\mathcal{A}_l} \pi_{l,\text{on}}(a)||a||^2_{V_\pi^{-1}} + \alpha \sum_{a\in\mathcal{A}} \pi_{\text{off}}(a)||a||^2_{V_\pi^{-1}}$$
$$d \le (1-\alpha) g_{\mathcal{A}_l}(\pi) + \alpha \sum_{a\in\mathcal{A}} \pi_{\text{off}}(a)||a||^2_{V_\pi^{-1}}.$$

To show (5) we first observe that the optimization in (2) :

$$\max_{\pi_{l,\text{on}}\in\Delta(\mathcal{A}_l)} \log\left(\det\left((1-\alpha)\sum_a \pi_{l,\text{on}}(a)aa^t + \alpha\sum_a \pi_{\text{off}}(a)aa^t\right)\right)$$

is concave in $\pi_{l,\text{on}}$. Next, using the fact that the derivative of the objective wrt $\pi_{l,\text{on}}(a)$ from standard matrix calculus is $(1-\alpha)||a||^2_{V_\pi^{-1}}$ and using first-order optimality conditions for a concave maximization we have:

$$0 \ge \sum_{a\in\mathcal{A}_l} (1-\alpha)||a||^2_{V^{-1}_{\tilde{\pi}_l^*}}(\pi_{l,\text{on}}(a) - \pi^*_{l,\text{on}}(a)).$$

But this implies that since $\tilde{\pi}_l^\star = (1-\alpha)\pi^*_{l,\text{on}} + \alpha\pi_{\text{off}}$ we have:

$$\alpha\sum_{a\in\mathcal{A}} \pi_{\text{off}}(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^*}} + (1-\alpha)\sum_{a\in\mathcal{A}_l} \pi^*_{l,\text{on}}(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^*}} \ge \alpha\sum_{a\in\mathcal{A}} \pi_{\text{off}}(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^*}} + (1-\alpha)\sum_{a\in\mathcal{A}_l} \pi_{l,\text{on}}(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^*}}$$
$$\sum_{a\in\mathcal{A}} \tilde{\pi}^*(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^*}} \ge \alpha\sum_{a\in\mathcal{A}} \pi_{\text{off}}(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^*}} + (1-\alpha)\sum_{a\in\mathcal{A}_l} \pi_{l,\text{on}}(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^*}}$$
$$d \ge \alpha\sum_{a\in\mathcal{A}} \pi_{\text{off}}(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^*}} + (1-\alpha)\sum_{a\in\mathcal{A}_l} \pi_{l,\text{on}}(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^*}}.$$

But since $\pi_{l,\text{on}}$ is arbitrary element from $\Delta(\mathcal{A}_l)$ we have that

$$d \ge \alpha\sum_{a\in\mathcal{A}} \pi_{\text{off}}(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^*}} + (1-\alpha)g(\tilde{\pi}_l^\star). \tag{14}$$

Combining (14) with the earlier inequality gives us (5):

$$d = \alpha\sum_{a\in\mathcal{A}} \pi_{\text{off}}(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^*}} + (1-\alpha)g(\tilde{\pi}_l^\star).$$

Applying to this the Woodbury Matrix identity $(A+B)^{-1} = A^{-1} - A^{-1}(I + AB^\dagger)^{-1}$ to this, where $B^\dagger$ represents the pseudoinverse, we get:

$$(1-\alpha)g(\tilde{\pi}_l^\star) = Tr\left(\left(I + \frac{\alpha}{1-\alpha}V_{\pi_{\text{off}}}V^\dagger_{\pi^*_{l,\text{on}}}\right)^{-1}\right).$$

Now using the following linear algebra result for any two PSD matrices A, B:

$$\lambda_k(A)\lambda_1(B) \le \lambda_k(AB) \le \lambda_k(A)\lambda_d(B), \tag{15}$$

we then have that (we set $B = I$, $A = I + \frac{\alpha}{1-\alpha}V_{\pi_{\text{off}}}V^\dagger_{\pi^*_{l,\text{on}}}$)

$$1 + \frac{\alpha}{1-\alpha}\frac{\lambda_k(V_{\pi_{\text{off}}})}{\lambda_d(V_{\pi^*_{l,\text{on}}})} \le \lambda_k\left(I + \frac{\alpha}{1-\alpha}V_{\pi_{\text{off}}}V^\dagger_{\pi^*_{l,\text{on}}}\right). \tag{16}$$

Using (16) in the representation above we get :

$$(1-\alpha)g(\tilde{\pi}_l^\star) \le \sum_{k=1}^d \frac{1}{1 + \frac{\alpha}{1-\alpha}\frac{\lambda_k(V_{\pi_{\text{off}}})}{\lambda_d(V_{\pi^*_{l,\text{on}}})}}.$$

Using the fact that $\lambda_d(V_{\pi^*_{l,\text{on}}}) \leq Tr(V_{\pi^*_{l,\text{on}}}) \leq \max_a ||a||^2$ and $\alpha = \frac{T_{\text{off}}}{T+T_{\text{off}}}$, we finally have:

$$(1-\alpha)g(\tilde{\pi}^\star_l) \leq \sum_{k=1}^{d} \frac{1}{1+\frac{T_{\text{off}}}{T}\frac{\lambda_k(V_{\pi_{\text{off}}})}{\max_a ||a||^2}}. \tag{17}$$

A simple but alternate useful bound for the confidence width is:

$$g(\tilde{\pi}^\star_l) \leq g(\pi_{\text{off}})/\alpha \tag{18}$$

where $g(\pi_{\text{off}}) \triangleq \max_{a\in\mathcal{A}} ||a|^2_{V^{-1}_{\pi_{\text{off}}}}$ and follows from the fact that $V_{\tilde{\pi}^*} \succeq \alpha V_{\pi_{\text{off}}}$. Note that this bound is useful in the regime where the offline is well explored, i.e, $g(\pi_{\text{off}}) = \theta(d)$, but can be pretty loose when the offline data is not uniformly well explored in some directions (in particular $g(\pi_{\text{off}}) = \infty$). Combining (18) with (16) and using the definition of d_{eff} ((1)) finally gives us:

$$(1-\alpha)g(\tilde{\pi}^\star_l) \leq d_{\text{eff}}.$$

B.4. Proof of Theorem 4.4.

We divide the proof into the following key steps:

Step 1: Concentration result for "clean" execution of `OOPE`.

Let us define the event $\xi_{a,l}(\mathcal{V})$ for any arm a and subset $\mathcal{V} \subseteq \mathcal{A}$ as

$$\xi_{a,l}(\mathcal{V}) := \{|\langle a, \theta^* - \hat{\theta}_l\rangle|\} \leq \epsilon_l\}$$

where $\hat{\theta}_l$ is the OLS estimator constructed from the offline samples and online samples in phase l. We assume the set of "live" arms is $\mathcal{V}$ and the online samples are taken only from this collection of live arms. We then have that:

$$\begin{aligned}\mathbb{P}\Big(\bigcup_{l=1}^{\infty}\bigcup_{a\in\mathcal{A}_l}\{\xi^c_{a,l}(\mathcal{A}_l)\}\Big) &\leq \sum_{l=1}^{\infty}\mathbb{P}\Big(\bigcup_{a\in\mathcal{A}_l}\{\xi^c_{a,l}(\mathcal{A}_l)\}\Big)\\ &= \sum_{l=1}^{\infty}\sum_{\mathcal{V}\subseteq\mathcal{A}}\mathbb{P}\Big(\bigcup_{a\in\mathcal{V}}\{\xi^c_{a,l}(\mathcal{V}), \mathcal{A}_l = \mathcal{V}\}\Big).\end{aligned}$$

We note that the estimate $\hat{\theta}_l$ is independent from the event $\{\mathcal{A}_l = \mathcal{V}\}$ as the offline samples are only ever used in one phase and never thereafter in estimating $\hat{\theta}_l$. This ensures no dependency is created between these two events. Thus

$$\begin{aligned}\mathbb{P}\Big(\bigcup_{l=1}^{\infty}\bigcup_{a\in\mathcal{A}_l}\{\xi^c_{a,l}(\mathcal{A}_l)\}\Big) &\leq \sum_{l=1}^{\infty}\sum_{\mathcal{V}\subseteq\mathcal{A}}\mathbb{P}\Big(\bigcup_{a\in\mathcal{V}}\{\xi^c_{a,l}(\mathcal{V})\}\Big)\mathbb{P}\Big(\{\mathcal{A}_l = \mathcal{V}\}\Big)\\ &\leq \sum_{l=1}^{\infty}\sum_{\mathcal{V}\subseteq\mathcal{A}}\sum_{a\in\mathcal{V}}\mathbb{P}\Big(\{\xi^c_{a,l}(\mathcal{V})\}\Big)\mathbb{P}\Big(\{\mathcal{A}_l = \mathcal{V}\}\Big)\end{aligned} \tag{19}$$

Now we state a useful sub-gaussian concentration result:

Lemma B.1. *Given the offline samples $n^l_{on}(a)$ and online samples $n^l_{off}(a)$ are as defined in (3) and (4) respectively, we have that*

$$\mathbb{P}(|\langle a, \theta^* - \hat{\theta}_l\rangle| \geq \epsilon_l) \leq \frac{1}{2Tl^2|\mathcal{A}|}. \tag{20}$$

Proof of Lemma B.1. Define $n^l(a) = n^l_{\text{on}}(a) + n^l_{\text{off}}(a)$, where

$$n^l_{\text{on}}(a) = \left\lceil \frac{3d_{\text{eff}}\pi^*_{l,\text{on}}(a)}{\epsilon_l^2}\log(4l^2|\mathcal{A}|T)\right\rceil$$

and

$$n^l_{\text{off}}(a) = \left\lceil \frac{2\alpha\pi_{\text{off}}(a)g(\tilde{\pi}^\star_l)\log(4l^2|\mathcal{A}|T)}{\epsilon_l^2}\right\rceil.$$

are the online and offline samples, respectively, of arm a utilized in phase l to construct $\hat{\theta}_l$. Further, define the positive definite matrix $V := \sum_a n^l(a)aa^t$. Then, we have that:

$$\begin{aligned} V &\succeq \left(\alpha \frac{2g(\tilde{\pi}_l^\star)\log(4l^2|\mathcal{A}|T)}{\epsilon_l^2}\sum_a \pi_{\text{off}}(a)aa^t + (1-\alpha)\frac{3d_{\text{eff}}\pi^*_{l,\text{on}}(a)\log(4l^2|\mathcal{A}|T)}{\epsilon_l^2}\sum_a \pi^*_{l,\text{on}}(a)aa^t\right) \\ &\succeq \frac{2g(\tilde{\pi}_l^\star)\log(4l^2|\mathcal{A}|T)}{\epsilon_l^2}\left(\alpha\sum_a \pi_{\text{off}}(a)aa^t + (1-\alpha)\sum_a \pi^*_{l,\text{on}}(a)aa^t\right) \\ &= \frac{2g(\tilde{\pi}_l^\star)\log(4l^2|\mathcal{A}|T)}{\epsilon_l^2}V_{\tilde{\pi}^*}. \end{aligned}$$

Here $\succeq$ refers to the standard partial ordering of positive definiteness on the space of $d \times d$ matrices. The second inequality above is obtained from $2(1-\alpha)g(\tilde{\pi}_l^\star) \leq 3d_{\text{eff}}$ (see Lemma 4.3). Consequently, for all $a \in \mathcal{A}$ we have:

$$\sqrt{2||a||^2_{V^{-1}}\log(4l^2|\mathcal{A}|T)} \leq \epsilon_l\sqrt{\frac{||a||^2_{V^{-1}_{\tilde{\pi}_l^*}}}{g(\tilde{\pi}_l^*)}} \leq \epsilon_l.$$

Then, utilising the standard sub-gaussain concentration result for OLS estimator (see Chapter 20 in (Lattimore & Szepesvári, 2020) for a derivation):

$$\mathbb{P}(|\langle a, \theta^* - \hat{\theta}_l\rangle| \geq \sqrt{2||a||^2_{V^{-1}}\log(1/\epsilon_0)}) \leq 2\epsilon_0$$

we have that

$$\begin{aligned} \mathbb{P}(|\langle a, \theta^* - \hat{\theta}_l\rangle| \geq \epsilon_l) &\leq \mathbb{P}\left(|\langle a, \theta^* - \hat{\theta}_l\rangle| \geq \sqrt{2||a||^2_{V^{-1}}\log(4l^2|\mathcal{A}|T)}\right) \\ &\leq \frac{1}{2Tl^2|\mathcal{A}|}. \end{aligned}$$

This concludes the proof of lemma. □

Substituting the bound (20) into (19) we get:

$$\begin{aligned} \mathbb{P}\left(\bigcup_{l=1}^{\infty}\bigcup_{a\in\mathcal{A}_l}\{\xi^c_{a,l}(\mathcal{A}_l)\}\right) &\leq \sum_{l=1}^{\infty}\sum_{\mathcal{V}\subseteq\mathcal{A}}\sum_{a\in\mathcal{V}}\frac{1}{2Tl^2|\mathcal{A}|}\mathbb{P}\left(\{\mathcal{A}_l = \mathcal{V}\}\right) \\ &\leq \sum_{l=1}^{\infty}\frac{1}{2Tl^2} \\ &\leq \frac{1}{T}. \end{aligned}$$

Thus, for the rest of the regret bound proof we will work within the "clean" execution event: $\bigcap_{l=1}^{\infty}\bigcap_{a\in\mathcal{A}_l}\{\xi_{a,l}(\mathcal{A}_l)\}$.

Step 2: In clean execution suboptimal arms are eliminated while the optimal arms aren't.

If an optimal arm a^* is in $\mathcal{A}_l$ then

$$\begin{aligned} \langle a - a^*, \hat{\theta}_l\rangle &= \langle a, \hat{\theta}_l - \theta^*\rangle - \langle a^*, \hat{\theta}_l - \theta^*\rangle + \langle a - a^*, \theta^*\rangle \\ &\leq 2\epsilon_l. \end{aligned}$$

This means that a^* is also in $\mathcal{A}_{l+1}$ in the good set. Thus by induction it is clear that in the good set the best arm is not eliminated.

For a such that $\langle a^* - a, \theta^*\rangle > 4\epsilon_l$ we have that

$$\begin{aligned} \max_{a'\in\mathcal{A}_l}\langle a' - a, \hat{\theta}_l\rangle &\geq \langle a^* - a, \hat{\theta}_l\rangle \\ &= \langle a^*, \hat{\theta}_l - \theta^*\rangle - \langle a, \hat{\theta}_l - \theta^*\rangle + \langle a^* - a, \theta^*\rangle \\ &> 2\epsilon_l, \end{aligned}$$

and hence get eliminated after phase l in the good set. This means in the next phase $l+1$ we have the property that for all $a \in \mathcal{A}_{l+1}$ we have that $\langle a^* - a, \theta^* \rangle \leq 4\epsilon_l = 8\epsilon_{l+1}$. Thus, in phase l, all arms a with a sub-optimality gap greater than $4\epsilon_l$ is eliminated.

A further consequence of the above result is that for $l \geq \log_2(8\Delta_{min}^{-1})$ we have $\mathcal{A}_l \subseteq \mathcal{A}^*$, i.e, only optimal arms survives.

Step 3: Upper bounding the regret with the confidence width d_{eff} and $|\text{supp}(\pi^*_{l,\text{on}})|$.

Let $n_{n,a}$ denote number of times arm a has been pulled within the online horizon T. Then, within the good set the regret is upper bounded by:

$$\begin{aligned}
\mathcal{R}(\texttt{OOPE}) &= \sum_{a \in \mathcal{A}\setminus\{a^*\}:\Delta_a \leq v} \Delta_a n_{n,a} + \sum_{a \in \mathcal{A}\setminus\{a^*\}:\Delta_a > v} \Delta_a n_{n,a} \\
&\leq vT + \sum_{a \in \mathcal{A}\setminus\{a^*\}:\Delta_a > v} \Delta_a n_{n,a} \\
&\leq vT + \sum_{l=1}^{l_M} \sum_{a \in \mathcal{A}\setminus\{a^*\}:\Delta_a > v} \Delta_a n^l_{\text{on}}(a) \mathbb{1}_{\{a \in \mathcal{A}_l\}}.
\end{aligned}$$

where $l_M = \min(l_{max}, \log_2(8v^{-1}))$. This is because from Step 2 we know that any suboptimal arms that survive into phase l has a sub-optimality gap of atmost $8\epsilon_l$, and $l \leq \log_2(8v^{-1})$. The other bound l_{max} denotes an upper bound for the very last phase in which the online samples are exhausted. From Proposition 4.2 we know that the offline samples if they exhaust will happen only in the last phase as well. Thus, we can proceed by assuming the concentration result from Step 1 holds in all but the last phase. For the last phase, since there is no elimination, we can bound the regret from the last phase by assuming excess online and offline samples as per (3), even though in reality they would have been exhausted. This is so because the online regret can only increase with more online samples and the concentration does not matter in the last round since we don't have elimination in the last round.

Substituting the (3) for $n^l_{\text{on}}(a)$ we get

$$\begin{aligned}
\mathcal{R}(\texttt{OOPE}) &\leq vT + \sum_{l=1}^{l_M} \sum_{a \in \mathcal{A}\setminus\{a^*\}:\Delta_a > v} 8\epsilon_l \left\lceil \frac{3d_{\text{eff}}\pi^*_{l,\text{on}}(a)}{\epsilon_l^2} \log(4l^2|\mathcal{A}|T) \right\rceil \\
&\leq vT + \sum_{l=1}^{l_M} \frac{24 d_{\text{eff}}}{\epsilon_l} \log(4l^2|\mathcal{A}|T) + \sum_{l=1}^{l_M} 8\epsilon_l |\text{supp}(\pi^*_{l,\text{on}})| \\
&\leq vT + 24 d_{\text{eff}} \log(4l^2_{max}|\mathcal{A}|T) \left(\sum_{l=1}^{l_M} 2^l \right) + \sum_{l=1}^{l_M} 8\epsilon_l |\text{supp}(\pi^*_{l,\text{on}})|
\end{aligned}$$

Step 4: Upper bound on $|\text{supp}(\pi^*_{l,\text{on}})|$.

Now we shall show, just as in Kiefer-Wolfowitz theorem (see Theorem 21.1 in (Lattimore & Szepesvári, 2020)), there exists an optimizer $\pi^*_{l,\text{on}}$ such that $|supp(\pi^*_{l,\text{on}})| \leq \frac{d(d+1)}{2}$. The Lagrangian for the concave optimization problem (2) is

$$\mathcal{L} = \log\left(\det\left((1-\alpha)\sum_a \pi_{l,\text{on}}(a) aa^t + \alpha \sum_a \pi_{\text{off}}(a) aa^t\right)\right) - \mu(\sum_a \pi_{l,\text{on}}(a) - 1) - \sum_a \lambda_a \pi_{l,\text{on}}(a).$$

where μ, λ_a are the multipliers. The Karush-Kuhn-Tucker conditions which are sufficient and necessary give:

$$\begin{aligned}
&||a||^2_{V^{-1}_{\tilde{\pi}}} - \mu - \lambda_a = 0, \\
&\lambda_a \pi_{l,\text{on}}(a) = 0, \\
&\sum_a \pi_{l,\text{on}}(a) = 1.
\end{aligned}$$

Here $\tilde{\pi}(a) = (1-\alpha_\ell)\pi_{l,\text{on}}(a) + \alpha_\ell \pi_{\text{off}}(a)$. We have that if $a \in supp(\pi^*_{l,\text{on}})$ then $\lambda_a = 0$. Therefore, we observe that $\forall a \in supp(\pi_{l,\text{on}})$, we have

$$||a||^2_{V^{-1}_{\tilde{\pi}^*_l}} = \mu > 0.$$

Now suppose $|supp(\pi^*_{l,\text{on}})| > \frac{d(d+1)}{2}$ then as the dimension of space of symmetric matrices is only $\frac{d(d+1)}{2}$ we know there exists a v such that

$$\sum_{a \in supp(\pi^*_{l,\text{on}})} v(a)aa^t = 0.$$

Then from the earlier observation we have

$$\mu \sum_{a \in supp(\pi^*_{l,\text{on}})} v(a) = \sum_{a \in supp(\pi^*_{l,\text{on}})} v(a)||a||^2_{V^{-1}_{\tilde{\pi^*_l}}} = Tr(V^{-1}_{\tilde{\pi^*_l}} \sum_{a \in supp(\pi^*_{l,\text{on}})} v(a)aa^t) = 0.$$

Thus $\sum_{a \in supp(\pi^*_{l,\text{on}})} v(a) = 0$. Then we notice that there exists a perturbation of π^*_n with v, i.e, $\widehat{\pi^*_{l,\text{on}}} = \pi^*_{l,\text{on}} + tv$ such that some of support points have zero mass and still have an unchanged optimal objective value. And hence from $\pi^*_{l,\text{on}}$ we have produced a new optimal solution $\widehat{\pi^*_{l,\text{on}}}$ that has strictly smaller subset as the support. By induction on the support size we conclude that there exists a $\pi^*_{l,\text{on}}$ such that $|supp(\pi^*_{l,\text{on}})| \leq \frac{d(d+1)}{2}$.

Step 5: The final bound.

Using the result in Step 4 we get a regret bound of

$$\begin{aligned}\mathcal{R}(\texttt{OOPE}) &\leq vT + 24d_{\text{eff}} \log(4l^2_{max}|\mathcal{A}|T) \left(\sum_{l=1}^{l_M} 2^l\right) + \sum_{l=1}^{l_M} 8\epsilon_l \frac{d(d+1)}{2} \\ &\leq vT + 48d_{\text{eff}} \log(4l^2_{max}|\mathcal{A}|T) \min\left\{\frac{8}{v}, \sqrt{4 + \frac{T}{d_{\text{eff}} \log(4|\mathcal{A}|T)}}\right\} + 4d(d+1).\end{aligned} \tag{21}$$

Optimizing over v gives the bound:

$$\mathcal{R}(\texttt{OOPE}) \leq 16\sqrt{6d_{\text{eff}} T \log(4l^2_{max}|\mathcal{A}|T)} + 4d(d+1), \tag{22}$$

under clean execution. But as we know from Step 1 that set arises with probability atleast $1 - \frac{1}{T}$ and thus the proof of Theorem 4.4 is complete.

B.5. Minimax regret and lower bounds for it in presence of offline data.

A *problem instance* is an ordered quintuple $p := (\pi_{\text{off}}, \mathcal{A}, \theta, T_{\text{off}}, T)$ where π_{off} is a measure on $\mathcal{A}$, $\mathcal{A} \subset R^d$, $\theta \in R^d$, and $T_{\text{off}}, T \in \mathbb{N}$. We define a *problem class* $\mathcal{P}$ to be a set of such problem instances p. In this work, we consider problem classes where all problem instances p are such that the ratio of eigenvalues of the offline Gram matrix $V_{\pi_{\text{off}}}$ to $\max_{a \in \mathcal{A}} ||a||^2$ is held fixed, that is,

$$v_i = \frac{\lambda_i(V_{\pi_{\text{off}}})}{\max_{a \in \mathcal{A}} ||a||^2}$$

for $i \in [d]$ is fixed. Further we impose the condition that $|\mathcal{A}| \leq (2d)^d$ for every instance in this class. Thus our problem classes are parametrized by $(d+3)$ parameters-$(d, v := (v_1, v_2, \ldots, v_d), T_{\text{off}}, T)$. We assume the *consistency condition* -

$$\sum_{i=1}^{d} v_i \leq 1$$

is satisfied by the parameter vector v since:

$$\begin{aligned}&\sum_{i=1}^{d} \lambda_i(V_{\pi_{\text{off}}}) = Tr(V_{\pi_{\text{off}}}) = \sum_{a \in \mathcal{A}} \pi_{\text{off}}(a)||a||^2 \leq \max_{a \in \mathcal{A}} ||a||^2 \\ \implies &\sum_{i=1}^{d} v_i \leq 1.\end{aligned}$$

We will denote such consistent problem classes as $\mathcal{P}^d_{v,T_{\text{off}},T}$.

We remark here that for any consistent set of parameters, the corresponding class $\mathcal{P}^d_{v,T_{\text{off}},T}$ is non-empty. We can just take the an action set consisting of scaled standard basis to see this.

Let us define the minimax regret value of a problem class as follows:

$$\mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T}) := \inf_{S\in\mathcal{S}} \sup_{p\in\mathcal{P}^d_{v,T_{\text{off}},T}} \mathcal{R}_p(T,S), \tag{23}$$

where S is an adaptive regret minimization algorithm for horizon T which takes T_{off} offline samples as input from a class $\mathcal{S}$ of such adaptive algorithms.

Hard instance that lower bounds $\mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T})$ and proof of Proposition 4.5.

In this section we will construct a hard instance $p_0 \in \mathcal{P}^d_{v,T_{\text{off}},T}$ such that we can derive a lower bound to $\mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T})$.

In this construction we assume that π_{off} is such that $|supp(\pi_{\text{off}})| \leq (2d)^d$. We will assume our action set is of the form:

$$\mathcal{A} = \left\{ a \in R^d \middle| a_i \in \{\pm c_{1,i}, \pm c_{2,i}, \ldots, \pm c_{d,i}\}, \ \ \forall i \in [d] \right\} \tag{24}$$

where $c_{k,i} \in R$. We collect these $c_{k,i}$'s in a column vector and denote it as C_i for each $i \in [d]$. The C_i's will be chosen to optimize the lower bound. Let the set of possible θ (the unknown parameter) be:

$$\Theta = \{\theta \in R^d | \theta_i \in \{\pm\alpha_i\}\}$$

where we will choose α_i later on. We assume the noise is standard iid gaussian $\mathcal{N}(0,1)$.

Define the following sets for $k, i \in [d]$:

$$\mathcal{A}^+_{c_{k,i}} = \{a \in \mathcal{A} | a_i = |c_{k,i}|\}$$
$$\mathcal{A}^-_{c_{k,i}} = \{a \in \mathcal{A} | a_i = -|c_{k,i}|\}$$

Remark B.2. We make the following observation about the sets $\mathcal{A}^+_{c_{k,i}}, \mathcal{A}^-_{c_{k,i}}$:

1. In case for some $k, i \in [d]$ if $c_{k,i} = 0$ then we have $\mathcal{A}^+_{c_{k,i}} = \mathcal{A}^-_{c_{k,i}}$.

2. If $c_{k,i} \neq 0$, then we have that the sets $\mathcal{A}^+_{c_{k,i}}, \mathcal{A}^-_{c_{k,i}}$ are disjoint.

Now let us define a $d(d-1)/2$ collection A^{ij} of $d \times d$ matrices indexed by the ordered pair $(i,j), i,j \in [d]$ with $i < j$ as follows:

$$A^{ij}_{k,l} = \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}} \cap \mathcal{A}^+_{c_{l,j}}) + \pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}} \cap \mathcal{A}^-_{c_{l,j}}) - \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}} \cap \mathcal{A}^-_{c_{l,j}}) - \pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}} \cap \mathcal{A}^+_{c_{l,j}}).$$

for $k, l \in [d]$.
Now we state a useful lemma:

Lemma B.3. *Given that $|supp(\pi_{off})| \leq (2d)^d$ over an action set of the form (24), we have that the following are equivalent:*

- *The offline Gram matrix V_{π_o} (which is a function of both π_{off} and $\mathcal{A}$) has the standard basis as its eigenvectors.*
- *The set of column vectors C_i, $i \in [d]$ is such that:*

$$C_i^t A^{ij} C_j = 0 \tag{25}$$

for all $i, j \in [d]$ and $i < j$.

Proof. A necessary and sufficient condition for standard basis to be the eigenvectors is that offdiagonal entries of V_{π_o} are zero, that is, for all $1 \leq i < j \leq d$:

$$\begin{aligned}
0 &= (\sum_{\mathcal{A}} \pi_{\text{off}}(a) a a^t)_{i,j} \\
&= \sum_{\mathcal{A}} \pi_{\text{off}}(a)(aa^t)_{ij} \\
&= \sum_{k,l \in [d]} \sum_{a \in \mathcal{A}: |a_i| = |c_{k,i}|, |a_j| = |c_{l,j}|} \pi_{\text{off}}(a) a_i a_j \\
&\overset{(a)}{=} \sum_{k,l \in [d]} c_{k,i} c_{l,j} (\pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}} \cap \mathcal{A}^+_{c_{l,j}}) + \pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}} \cap \mathcal{A}^-_{c_{l,j}}) - \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}} \cap \mathcal{A}^-_{c_{l,j}}) - \pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}} \cap \mathcal{A}^+_{c_{l,j}})) \\
&= \sum_{k,l \in [d]} c_{k,i} c_{l,j} A^{ij}_{k,l} \\
&= C_i^t A^{ij} C_j
\end{aligned}$$

where (a) follows from Remark B.2. This concludes the proof. □

Remark B.4. If $c_{k,i} = 0$ then it is clear that $\forall j > i$ and $l \in [d]$ that $A^{ij}_{kl} = 0$.

Lemma B.5. *Suppose we have that $A \sim \pi_{off}$ and π_{off} is such that it satisfies:*

1. *(Coordinate Independence). $\forall i, j, k, l \in [d]$*

$$\pi_{off}(A_i = c_{k,i}, A_j = c_{l,j}) = \pi_{off}(A_i = c_{k,i}) \pi_{off}(A_j = c_{l,j})$$

2. *($\pm$ Symmetricity). $\forall i, k \in [d]$*

$$\pi_{off}(A_i = c_{k,i}) = \pi_{off}(A_i = -c_{k,i})$$

, that is, π_{off} is independent co-ordinate wise and symmetrically distributed wrt to positive & negative values of a component $c_{k,i}$, then we have that $A^{ij} = 0$ for all $i < j$.

Proof. From coordinate independence we have that:

$$\begin{aligned}
\pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}} \cap \mathcal{A}^+_{c_{l,j}}) &= \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}}) \pi_{\text{off}}(\mathcal{A}^+_{c_{l,j}}) \\
\pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}} \cap \mathcal{A}^-_{c_{l,j}}) &= \pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}}) \pi_{\text{off}}(\mathcal{A}^-_{c_{l,j}}) \\
\pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}} \cap \mathcal{A}^-_{c_{l,j}}) &= \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}}) \pi_{\text{off}}(\mathcal{A}^-_{c_{l,j}}) \\
\pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}} \cap \mathcal{A}^+_{c_{l,j}}) &= \pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}}) \pi_{\text{off}}(\mathcal{A}^+_{c_{l,j}})
\end{aligned}$$

for all $i, j, k, l \in [d]$. From $\pm$ symmetricity we have that:

$$\pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}}) = \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}})$$

for all $i, k \in [d]$. Using these relations we get that:

$$\begin{aligned}
A^{ij}_{k,l} &= \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}} \cap \mathcal{A}^+_{c_{l,j}}) + \pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}} \cap \mathcal{A}^-_{c_{l,j}}) - \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}} \cap \mathcal{A}^-_{c_{l,j}}) - \pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}} \cap \mathcal{A}^+_{c_{l,j}}) \\
&= \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}}) \pi_{\text{off}}(\mathcal{A}^+_{c_{l,j}}) + \pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}}) \pi_{\text{off}}(\mathcal{A}^-_{c_{l,j}}) - \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}}) \pi_{\text{off}}(\mathcal{A}^-_{c_{l,j}}) - \pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}}) \pi_{\text{off}}(\mathcal{A}^+_{c_{l,j}}) \\
&= \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}}) \pi_{\text{off}}(\mathcal{A}^+_{c_{l,j}}) + \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}}) \pi_{\text{off}}(\mathcal{A}^+_{c_{l,j}}) - \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}}) \pi_{\text{off}}(\mathcal{A}^+_{c_{l,j}}) - \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}}) \pi_{\text{off}}(\mathcal{A}^+_{c_{l,j}}) \\
&= 0,
\end{aligned}$$

for every $i < j$ and $k, l \in [d]$. This shows that A^{ij} is the zero matrix for this choice of π_{off}. □

The upshot of the previous lemma is that if a measure π_{off} satisfies the coordinate independence and $\pm$ symmetricity then by Lemma B.3 we know any choice of C_i ensure that the eigenvectors of V_{π_o} are the standard basis.

Remark B.6. It is straightforward to construct a π_{off} which is coordinate independent and $\pm$ symmetric. Consider any d mutually independent distributions on $[d]$. Denote them as π_j, $j \in [d]$ and $\pi_j \in \Delta_d$. Simply set $\pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}} \cup \mathcal{A}^-_{c_{k,i}}) = \pi_i(k)$ for each $i, k \in [d]$. If it is the case that $c_{k,i} \neq 0$, then one further sets: $\pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}}) = \pi_{\text{off}}(\mathcal{A}^-_{c_{k,i}}) = \pi_i(k)/2$ to ensure $\pm$ symmetricity. We note that coordinate wise independence and $\pm$ symmetricity are only partially restrictive of our choice of π_{off}, We still could make any arbitrary d-collection of π_j's in the above construction.

Now we want the π_{off} and C_i's should satisfy for each $i \in [d]$:

$$v_i = \frac{\lambda_i}{\max_{a \in \mathcal{A}} ||a||^2}. \tag{26}$$

We note that

$$\lambda_i = \sum_{k=1}^{d} \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}} \cup \mathcal{A}^-_{c_{k,i}})|c_{k,i}|^2$$

and

$$\max_{a \in \mathcal{A}} ||a||^2 = \sum_{i=1}^{d} |c_{d,i}|^2$$

where we assume WLOG that $|c_{d,i}| \geq |c_{k,i}|$ for all $1 \leq k \leq (d-1)$ and $i \in [d]$. Re-writing the desired condition using the above equations we obtain:

$$\begin{aligned} v_i &= \frac{\lambda_i}{\max_{a \in \mathcal{A}} ||a||^2} \\ &= \frac{\sum_{k=1}^{d} \pi_{\text{off}}(\mathcal{A}^+_{c_{k,i}} \cup \mathcal{A}^-_{c_{k,i}})|c_{k,i}|^2}{\sum_{i=1}^{d} |c_{d,i}|^2} \\ &= \frac{\sum_{k=1}^{d} \pi_i(k)|c_{k,i}|^2}{\sum_{i=1}^{d} |c_{d,i}|^2}. \end{aligned}$$

Now for any $w_i = \frac{|c_{d,i}|^2}{\sum_{i=1}^{d} |c_{d,i}|^2} \geq v_i$, it is clear from the above equation that there will always exist a $\pi_i \in \Delta_d$ and choice of $|c_{k,i}|^2 \leq |c_{d,i}|^2$ for $1 \leq k \leq (d-1)$ that satisfy condition (26).

Let $P_\theta, P_{\theta'}$ be the distribution on the offline+online samples for parameters θ and θ'. Then we can decompose the KL divergence as:

$$\begin{aligned} D(P_\theta, P_{\theta'}) &= \mathbb{E}_\theta[\sum_{t=-T_{\text{off}}}^{0} D(\mathcal{N}(\langle A_t, \theta \rangle, 1), \mathcal{N}(\langle A_t, \theta' \rangle, 1))] + \mathbb{E}_\theta[\sum_{t=1}^{T} D(\mathcal{N}(\langle A_t, \theta \rangle, 1), \mathcal{N}(\langle A_t, \theta \rangle, 1))] \\ &= \frac{1}{2} \sum_{t=-T_{\text{off}}}^{0} (\langle A_t, \theta - \theta' \rangle)^2 + \frac{1}{2} \sum_{t=1}^{T} \mathbb{E}_\theta[(\langle A_t, \theta - \theta' \rangle)^2] \\ &= \frac{T_{\text{off}}}{2} ||\theta - \theta'||^2_{V_{\pi_o}} + \frac{1}{2} \sum_{t=1}^{T} \mathbb{E}_\theta\left[||\theta - \theta'||^2_{A_t A_t^T}\right] \end{aligned}$$

where A_t refers to the arm pull sequence and the randomness in the last equation is due to the sampling strategy S.

For $i \in [d]$ and $\theta \in \Theta$ we define

$$p_{\theta_i} := \mathbb{P}_\theta\left(\sum_{t=1}^{T} \mathbb{1}\left\{sgn(A_{ti} \neq sgn(\theta_i)\right\} \geq \frac{T}{2}\right).$$

Now let for every θ define $\tilde{\theta}_i \in \Theta$ such that $\tilde{\theta}_i = -\theta_i$ and $\forall j \neq i$, $\tilde{\theta}_j = \theta_j$.

We now describe our choice for α_i:

$$\alpha_i = \frac{1}{|c_{d,i}|\sqrt{T + T_{\text{off}}\frac{v_i}{w_i}}}.$$

We apply Bretagnoulle-Huber lemma and use the definitions of α_i and the KL decomposition to get :

$$\begin{aligned}
p_{\theta_i} + p_{\tilde{\theta}_i} &\geq \frac{1}{2}exp\Big(- \frac{T_{\text{off}}}{2}||\theta - \tilde{\theta}_i||^2_{V_{\pi_{\text{off}}}} - \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}_\theta\big[||\theta - \tilde{\theta}_i||^2_{A_tA_t^T}\big]\Big) \\
&= \frac{1}{2}exp\Big(- \frac{4\alpha_i^2 T_{\text{off}}\lambda_i(V_{\pi_{\text{off}}})}{2} - \frac{4\alpha_i^2}{2}\sum_{t=1}^{T}\mathbb{E}_\theta\big[A_{ti}^2\big]\Big) \\
&\overset{(a)}{\geq} \frac{1}{2}exp\Big(- \frac{4\alpha_i^2 T_{\text{off}}\lambda_i(V_{\pi_{\text{off}}})}{2} - \frac{4T|c_{d,i}|^2\alpha_i^2}{2}\Big) \\
&\overset{(b)}{=} \frac{1}{2}exp\Big(- 2\alpha_i^2|c_{d,i}|^2\big(T + T_{\text{off}}\frac{v_i}{w_i}\big)\Big) \\
&= \frac{1}{2}exp(-2)
\end{aligned}$$

where (a) follows because $|c_{d,i}| \geq |c_{k,i}|$ for $1 \leq k \leq (d-1)$ and (b) follows because $w_i = \frac{|c_{d,i}|^2}{\max_{a\in\mathcal{A}}||a||^2}$.

Then we have for any $b \in R_+^d$ that

$$\sum_{\theta\in\Theta}\frac{1}{|\Theta|}\sum_{i=1}^{d} b_i p_{\theta_i} = \frac{1}{|\Theta|}\sum_{i=1}^{d} b_i \sum_{\theta\in\Theta} p_{\theta_i} \geq \frac{(\sum_{i=1}^{d} b_i)}{4}exp(-2).$$

In what follows we choose:

$$b_i = \frac{1}{\sqrt{T + T_{\text{off}}\frac{v_i}{w_i}}}$$

This guarantees the existence of a θ_0 such that $\sum_{i=1}^{d} b_i p_{\theta_{0_i}} \geq \frac{(\sum_{i=1}^{d} b_i)}{4}exp(-2)$. Finally, we choose our problem instance $p_0 = (\pi_{\text{off}}, \mathcal{A}, \theta_0, T_{\text{off}}, T)$ where $\mathcal{A}$ is given by (24) and π_{off} is a measure on this $\mathcal{A}$ that has coordinate independence and $\pm$ symmetricity. For this p_0 we have

$$\begin{aligned}
\mathcal{R}_{p_0}(T,S) &\overset{(a)}{=} \mathbb{E}_{\theta_0}[\sum_{t=1}^{T}\sum_{i=1}^{d}(|c_{d,i}|sgn(\theta_{0_i}) - A_{t,i})\theta_{0_i}] \\
&= \sum_{i=1}^{d}\alpha_i\mathbb{E}_{\theta_0}\Big[\sum_{t=1}^{T}(|c_{d,i}| + |A_{t,i}|)\mathbb{1}\big\{sgn(A_{ti}) \neq sgn(\theta_i)\big\} + (|c_{d,i}| - |A_{t,i}|)\mathbb{1}\big\{sgn(A_{ti}) = sgn(\theta_i)\big\}\Big] \\
&\geq \sum_{i=1}^{d}\frac{|c_{d,i}|}{|c_{d,i}|\sqrt{T + T_{\text{off}}\frac{v_i}{w_i}}}\mathbb{E}_{\theta_0}\Big[\sum_{t=1}^{T}\mathbb{1}\big\{sgn(A_{ti}) \neq sgn(\theta_i)\big\}\Big] \\
&\geq \frac{T}{2}\sum_{i=1}^{d}\frac{1}{\sqrt{T + T_{\text{off}}\frac{v_i}{w_i}}}\mathbb{P}_{\theta_0}\Big(\sum_{t=1}^{T}\mathbb{1}\big\{sgn(A_{ti} \neq sgn(\theta_i)\big\} \geq \frac{T}{2}\Big) \\
&\geq \frac{\sqrt{T}exp(-2)}{8}\sum_{i=1}^{d}\frac{1}{\sqrt{1 + \frac{T_{\text{off}}}{T}\frac{v_i}{w_i}}},
\end{aligned}$$

where $w_i \geq v_i$ for each i. (a) follows from the fact the optimal arm has the same sign as θ_0 in each coordinate. Now we know this construction works for every $w \in \Delta_d$ such that $w_i \geq v_i$. This implies:

$$\mathcal{R}_{p_0}(T,S) \geq \frac{\sqrt{T}exp(-2)}{8}\sup_{\substack{w\in\Delta_d \\ \forall i, w_i \geq v_i}}\sum_{i=1}^{d}\frac{1}{\sqrt{1 + \frac{T_{\text{off}}}{T}\frac{v_i}{w_i}}}.$$

This gives a lower bound for the minimax value:

$$\begin{aligned}\mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T}) &\geq \frac{\sqrt{T}exp(-2)}{8} \sup_{\substack{w\in\Delta_d \\ \forall i, w_i\geq v_i}} \sum_{i=1}^{d} \frac{1}{\sqrt{1+\frac{T_{\text{off}}}{T}\frac{v_i}{w_i}}} \\ &= \theta\left(\sqrt{T} \sup_{\substack{w\in\Delta_d \\ \forall i, w_i\geq v_i}} \sum_{i=1}^{d} \frac{1}{\sqrt{1+\frac{T_{\text{off}}}{T}\frac{v_i}{w_i}}}\right)\end{aligned} \tag{27}$$

B.6. Regret bound for Warm Started LinUCB

In warm-started LinUCB with offline data we have the UCB-index defined as:

$$UCB_t(a) = \underset{\tilde{\theta}\in\mathcal{C}_t}{argmax} \ \langle\tilde{\theta}, a\rangle$$

where $\mathcal{C}_t = \{\tilde{\theta} : ||\tilde{\theta} - \hat{\theta}_t||_{V_t} \leq \beta_t\}$ is confidence interval. This usual elliptical confidence interval is warm started with offline data $V_t = \gamma I + T_{\text{off}} V_{\pi_{\text{off}}} + \sum_{s=1}^{t} a_s a_s^t$ and the algorithm plays the arm:

$$A_t \in \underset{a\in\mathcal{A}}{argmax} \ UCB_t(a).$$

Proposition B.7. *Assuming that for all $\theta \in \Theta$, we have $||\theta||_{V_0} \leq m$, where m is some known constant, then regret for warm started LinUCB is given by :*

$$\mathcal{R}(\texttt{UCB}) \leq \sqrt{8T\beta_T d_e^U \log\left(1 + \frac{T\max||a||^2}{\lambda_1(V_0)}\right)}$$

where

$$\sqrt{\beta_t} = m + \sqrt{2\log(T) + 2d_e^U \log\left(1 + \frac{T\max||a||^2}{\lambda_1(V_0)}\right)}$$

and

$$d_e^U := \max\left\{i \in [d] : \quad (i-1)\lambda_i(V_0) \leq \frac{T\max||a||^2}{\log\left(1 + \frac{T\max||a||^2}{\lambda_1(V_0)}\right)}\right\}$$

Here V_0 refers to Gram matrix warm started using the entire offline data.

Remark B.8. The proof utilizes a result of (Valko et al., 2014), where we set $V_0 = T_{\text{off}} V_{\pi_{\text{off}}}$.

Proof. Let $V_t = V_0 + \sum_{s=1}^{t} A_s A_s^t$. Let the eigenvalues be $\delta_i + \nu_i$ and ν_i, of V_t and V_0 respectively. Then

$$\begin{aligned}\log\left(\frac{det(V_t)}{det(V_0)}\right) &= \sum_{i=1}^{d} \log\left(1 + \frac{\delta_i}{\nu_i}\right) \\ &\leq \sum_{i=1}^{d_e^U} \log\left(1 + \frac{T\max||a||^2}{\nu_1}\right) + \frac{T\max||a||^2}{\nu_{d_e^U+1}} \\ &\leq 2\sum_{i=1}^{d_e^U} \log\left(1 + \frac{T\max||a||^2}{\nu_1}\right)\end{aligned}$$

The first inequality follows from definition of d_e^U and the fact that $\sum_i \delta_i \leq T\max||a||^2$. The result then is established from the standard analysis of LinUCB (see Theorem 19.2, Lemma 19.4 and Theorem 20.4 in (Lattimore & Szepesvári, 2020)). □

Properties of the lower bound (27).

Let us define the following quantity:

$$d_e^{lb} := \sup_{\substack{w\in\Delta_d \\ \forall i, w_i\geq v_i}} \sum_{i=1}^{d} \frac{1}{\sqrt{1+\frac{T_{\text{off}}}{T}\frac{v_i}{w_i}}}. \tag{28}$$

Then the lower bound (27) is just $\theta(\sqrt{T}d_e^{lb})$. The following are simple properties of d_e^{lb} and the proof is omitted:

Lemma B.9. *We have that:*

1. *The optimization defining d_e^{lb} in (28) is a concave program in w.*
2. $d_e^{lb} \leq d$.
3. *If $\sum_{i=1}^{d} v_i = 1$, $v_1 = v_2 = \cdots = v_{k-1} = 0$ and $0 < v_k \leq v_{k+1} \leq \cdots \leq v_d$, for some $k \in [d]$ then:*

$$d_e^{lb} = (k-1) + \frac{(d-k+1)}{\sqrt{1+\frac{T_{off}}{T}}}$$

We provide a dual representation of d_e^{lb} in the next lemma:

Lemma B.10 (Dual representation of d_e^{lb})**.** *For simplicity assume that $0 < v_1 \leq \cdots \leq v_d$. Define for each $i \in [d]$, the following functions:*

$$\beta_i(x) = \begin{cases} v_i, & \text{for } x \leq \frac{-T_{off}}{2Tv_i(1+\frac{T_{off}}{T})^{3/2}} \\ z, & \text{for } \frac{-T_{off}}{2Tv_i(1+\frac{T_{off}}{T})^{3/2}} \leq x \leq \frac{-T_{off}v_i}{2T(1+\frac{T_{off}v_i}{T})^{3/2}} \\ 1, & \text{for } \frac{-T_{off}v_i}{2T(1+\frac{T_{off}v_i}{T})^{3/2}} \leq x \end{cases}$$

where z is the unique positive real root (which is guaranteed to exist exist under the condition on x) to the quartic polynomial:

$$z\left(z+\frac{T_{off}v_i}{T}\right)^3 - \left(\frac{T_{off}v_i}{2Tx}\right)^2 = 0.$$

We note that each β_i are non-decreasing function of x. Then we have that:

$$d_e^{lb} = \min_{x\in\mathbb{R}}\left\{\sum_{i=1}^{d}\left(\beta_i(x)x + \frac{1}{\sqrt{1+\frac{T_{off}v_i}{T\beta_i(x)}}}\right) - x\right\}. \tag{29}$$

Furthermore, the minimizer x^* (it need not be unique) for the dual problem is characterised by the necessary and sufficient condition:

$$\sum_{i=1}^{d}\beta_i(x^*) = 1,$$

with the unique primal optimizer $w^* = (\beta_1(x^*), \beta_2(x^*), \ldots, \beta_d(x^*))$.

Using primal and dual forms of d_e^{lb} one may derive the following upper and lower bounds:

Lemma B.11. *For $k \in [d]$, such that $v_i = 0$ for $i < k$ and $0 < v_i$ for $i \geq k$, then we have:*

$$\frac{(1-\sum_i v_i)T_{off}}{2v_d(1+\frac{T_{off}}{T})^{3/2}T} + (k-1) + \frac{(d-k+1)}{\sqrt{1+\frac{T_{off}}{T}}} \geq d_e^{lb} \geq (k-1) + \frac{(d-k+1)}{\sqrt{1+\frac{T_{off}(\sum_i v_i)}{T}}}$$

Note that these bounds are tight if $\sum_i v_i \approx 1$ but can be loose if $\sum_i v_i << 1$.

B.7. Upper and Lower bounds for $\mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T})$.

Using Theorem 4.4 we see that the OOPE regret upper bound when $|\mathcal{A}| = (2d)^d$ and $T = \Omega(max(d\sqrt{T_{\text{off}}}, d^3))$[7] becomes

$$\mathcal{R}(\texttt{OOPE}) = \tilde{O}(\sqrt{d_{eff}dT})$$

where $\tilde{O}$ suppresses logarithmic factors in d, T, T_{off}. To use this as an upper bound on $\mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T})$ we will first bound d_{eff} in terms of v:

$$\begin{aligned}
d_{eff} &= \min\left(\sum_{i=1}^{d} \frac{1}{1+\frac{T_{\text{off}}}{T}\frac{\lambda_i(V_{\pi_{\text{off}}})}{\max_a ||a||^2}}, \frac{T}{T_{\text{off}}} g(\pi_{\text{off}})\right) \\
&= \min\left(\sum_{i=1}^{d} \frac{1}{1+\frac{T_{\text{off}}v_i}{T}}, \frac{T}{T_{\text{off}}} g(\pi_{\text{off}})\right) \\
&= \min\left(\sum_{i=1}^{d} \frac{1}{1+\frac{T_{\text{off}}v_i}{T}}, \frac{T}{T_{\text{off}}} \max_{a\in\mathcal{A}} \sum_{i=1}^{d} \frac{a_i^2}{\lambda_i(V_{\pi_o})}\right) \\
&\leq \min\left(\sum_{i=1}^{d} \frac{1}{1+\frac{T_{\text{off}}v_i}{T}}, \frac{T}{T_{\text{off}}} \max_{w\in\Delta_d} \sum_{i=1}^{d} \frac{w_i}{v_i}\right) \\
&= \min\left(\sum_{i=1}^{d} \frac{1}{1+\frac{T_{\text{off}}v_i}{T}}, \frac{T}{T_{\text{off}}v_1}\right)
\end{aligned}$$

Thus we have that

$$\tilde{O}\left(\sqrt{dT\min\left(\sum_{i=1}^{d} \frac{1}{1+\frac{T_{\text{off}}v_i}{T}}, \frac{T}{T_{\text{off}}v_1}\right)}\right) \geq \mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T}) \geq \theta(d_e^{lb}\sqrt{T}) \tag{30}$$

In the case where $v_i \geq \frac{c}{d}$ where $c(<1)$ is a small constant, for all i we have from the above bounds that:

$$\tilde{O}(dT/\sqrt{T_{\text{off}}c}) \geq V(\mathcal{P}^d_{v,T_{\text{off}},T}) \geq \theta(dT/\sqrt{T+T_{\text{off}}})$$

which is tight upto logarithmic factors when $T = o(T_{\text{off}})$ and c is bounded away from zero.
In the case where $v_i = 0$ for all $i < k$ and $v_k > c/d$ for $i \geq k$, with $k = \Omega(d)$, $T = O(T_{\text{off}})$ and T_{off} is large, we have that:

$$\tilde{O}(d\sqrt{T}) = \tilde{O}(\sqrt{dT(k-1)}) \geq \mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T}) \geq \theta((k-1)\sqrt{T}) = \Omega(d\sqrt{T})$$

which is again tight upto logarithmic factors since $k \leq d$. Note that if $k = o(d)$ then:

$$\tilde{O}(\sqrt{dT(k-1)}) \geq \mathcal{R}_{\text{minmax}}(\mathcal{P}^d_{v,T_{\text{off}},T}) \geq \theta((k-1)\sqrt{T})$$

and hence there is a multiplicative gap of $\sqrt{d/k}$ between the upper and lower bounds.

Summary:The above calculations show that we are tight when all directions are well explored or if quite a large number of directions remain under-explored in the offline data. We remark that the there is a multiplicative gap of $O(\sqrt{d/k})$ between our current upper bound and lower bound in the regime where a few directions ($k = o(d)$ in the above example) are under-explored ($v_{k-1} = o(d^{-1})$) in the offline data.

B.8. Example where warm started LinUCB bound is weaker than OOPE.

Consider the setting where $\mathcal{A} = \{\pm 1\}^d$ and $\Theta = \{\pm\sqrt{\frac{d}{(T_{\text{off}}+T)}}\}^d$. Let each arm be uniformly pulled in the offline data, that is, $\pi_{\text{off}}(a) = \frac{1}{|\mathcal{A}|}$. Then $\lambda_k(V_{\pi_{\text{off}}}) = 1$ for all k. Let $V_0 = T_{\text{off}}V_{\pi_{\text{off}}} + \gamma I$ (we choose $\gamma = O(T)$). Now:

$$\begin{aligned}
||\theta||^2_{V_0} &= (T_{\text{off}}+\gamma)||\theta||_2^2 \\
&= \frac{d^2(T_{\text{off}}+\gamma)}{T_{\text{off}}+T} \\
&\leq d^2
\end{aligned}$$

[7] This ensures that $\tilde{O}(\sqrt{d_{eff}dT})$ is larger than d^2.

Hence we get that $m = d$ and $d_e^U \leq 2$. This implies $\sqrt{\beta_T} = O(d\sqrt{\log(T)})$. Therefore $\mathcal{R}(\texttt{UCB}) \leq O(d\sqrt{T\log(T)\log(1+\frac{Td}{T_{\text{off}}+\gamma})})$ which for $T_{\text{off}} \gg T$ simplifies to $\mathcal{R}(\texttt{UCB}) \leq O(d^{3/2}T/\sqrt{T_{\text{off}}+\gamma})$.

We have the well-explored setting of offline data with $g_{\mathcal{A}}(\pi_{\text{off}}) = d$ and as a result OOPE's regret bound becomes $\mathcal{R}(\texttt{OOPE}) \leq O(\frac{dT}{\sqrt{T_{\text{off}}}} + d^2)$ which improves over LinUCB rate by the multiplicative factor of $\sqrt{d}$ and is important in moderate to low dimension ($d \leq 50$) regime.

C. Proof of results in Section 5

Motivating the dual: We will adapt the analysis found in chapter 2 and 3 of (Todd, 2016). We introduce the *primal* optimization problem

$$\begin{aligned}\mathcal{P}(\mathcal{V},\alpha,c) :=& \min_{H\succ 0} \quad -\log(\det(H)) \\ & \text{s.t} \quad (1-\alpha)a^tHa + \alpha Tr(V_{\pi_{\text{off}}}H) \leq c, \quad \forall a \in \mathcal{V} \subseteq \mathcal{A}.\end{aligned}$$

We shall show this optimization has an optimizer and will be denoted by $H_\alpha^*(\mathcal{V},c)$. We observe that $\mathcal{P}(\mathcal{A}_l,0,d)$ is the standard MVEE problem for the set of "live" arms in phase l, with optimal solution denoted by $H_0^*(\mathcal{A}_l,d)$. Mostly we set $c=d$, in what follows. Let us try to motivate a duality for $\mathcal{P}(\mathcal{V},\alpha,d)$.
The Lagrangian for the minimization is

$$L(H,u) := -\log(\det(H)) + \sum_{a\in\mathcal{V}} u_a((1-\alpha)a^tHa + \alpha Tr(V_{\pi_{\text{off}}}H) - d)$$

for the multipliers $u_a \geq 0, \forall a \in X$. For any H feasible we have

$$L(H,u) \leq -\log(\det(H)).$$

Differentiating wrt to H we get

$$\nabla_H L(H,u) = -H^{-1} + (1-\alpha)\sum_{a\in\mathcal{V}} u_a aa^t + \alpha V_{\pi_{\text{off}}}(\sum_a u_a)$$

Setting this to zero we get

$$H_\alpha^*(\mathcal{V},d) = \left[(1-\alpha)\sum_{a\in\mathcal{V}} u_a aa^t + +\alpha V_{\pi_{\text{off}}}(\sum_a u_a)\right]^{-1}.$$

Substituting this back into the $L(H,u)$ we have

$$\min_H L(H,u) = \log\left(\det\left([(1-\alpha)\sum_{a\in\mathcal{V}} u_a aa^t + \alpha V_{\pi_{\text{off}}}(\sum_a u_a)]\right)\right) + d - (\sum_a u_a)d.$$

Lagrangian duality then tells us

$$\mathcal{P}(\mathcal{V},\alpha,d) = \max_{u\geq 0} \quad \log\left(\det\left([(1-\alpha)\sum_{a\in\mathcal{V}} u_a aa^t + \alpha V_{\pi_{\text{off}}}(\sum_a u_a)]\right)\right) + d - (\sum_a u_a)d.$$

Now let $u_a = (\sum_a u_a)\pi(a)$, where $\pi(a)$ is from the probability simplex on $\mathcal{V}$. Then letting $\sum_a u_a = t$ we have

$$\mathcal{P}(\mathcal{V},\alpha,d) = \max_{\pi\in\Delta(\mathcal{V})} \max_t \quad d(\log(t)-t) + d + \log\left(\det\left((1-\alpha)V_\pi + \alpha V_{\pi_{\text{off}}}\right)\right).$$

We observe $\log(t)-t$ is maximized at $t=1$ and this gives the dual:

$$\mathcal{P}(\mathcal{V},\alpha,d) = \max_{\pi\in\Delta(\mathcal{V})} \log\left(\det\left((1-\alpha)V_{\pi_n} + \alpha V_{\pi_{\text{off}}}\right)\right).$$

We make this rigorous in the next subsection.

C.1. Proof of Lemma 5.1 (Strong Duality in the OO setting).

Define the *dual* problem as:

$$\mathcal{D}(\mathcal{V},\alpha) := \max_{\pi\in\Delta(\mathcal{V})} \log\left(\det\left((1-\alpha)V_\pi + \alpha V_{\pi_{\text{off}}}\right)\right).$$

Let $d(\pi,\mathcal{V},\alpha)$ denote the dual value for any feasible π (or the information gain of a design π) in the dual problem and $p(H,\mathcal{V},\alpha,d)$ denote the value for any feasible H in the primal problem.

Proposition C.1 (Weak duality)**.** *For any primal feasible H and dual feasible π we have*

$$p(H,\mathcal{V},\alpha,d) \geq d(\pi,\mathcal{V},\alpha).$$

Proof. We have

$$\begin{aligned} p(H,\mathcal{V},\alpha,d) - d(\pi,\mathcal{V},\alpha) &= -\log(\det(H)) - \log(\det((1-\alpha)V_\pi + \alpha V_{\pi_{\text{off}}})) \\ &= -\log(\det(H((1-\alpha)V_\pi + \alpha V_{\pi_{\text{off}}}))) \end{aligned}$$

Now the matrix inside the $\log(\det)$ function has the same eigenspectrum as a corresponding PD matrix. Let λ_j denote its eigenvalues then

$$\begin{aligned} p(H,\mathcal{V},\alpha,d) - d(\pi,\mathcal{V},\alpha) &= -\log(\det(H((1-\alpha)V_\pi + \alpha V_{\pi_{\text{off}}}))) \\ &= -\log(\prod_{j=1}^{d}\lambda_j) \\ &= -d \quad \log((\prod_{j=1}^{d}\lambda_j)^{1/d}) \\ &\geq -d\log\left(\frac{\sum_j \lambda_j}{d}\right) \\ &\geq 0. \end{aligned}$$

The first inequality is AM-GM inequality while the second one is because

$$Tr(H((1-\alpha)V_\pi + \alpha V_{\pi_{\text{off}}})) = \sum_a \pi(a)((1-\alpha)a^t Ha + \alpha Tr(V_{\pi_{\text{off}}}H)) \leq d$$

and the primal feasibility of H. □

We prove the strong duality next, that is

$$p(H^*_\alpha(\mathcal{V},d),\mathcal{V},\alpha,d) = d(\pi^*,\mathcal{V},\alpha)$$

where π^* is the optimal solution to $\mathcal{D}(\mathcal{V},\alpha)$ and $H^*_\alpha(\mathcal{V},d)$ the optimal solution to the primal $\mathcal{P}(\mathcal{V},\alpha,d)$. Now we show the proof of strong duality.

Proof of Strong Duality (Lemma 5.1). For $\epsilon(>0)$ small enough we know that ϵI is primal feasible. Thus we can restrict H by adding the constraint $-\log(\det(H)) \leq -\log(\det(\epsilon I))$ without changing the optimization. This restriction means the H must be strictly positive definite and cannot be arbitrarily close to semi-definiteness. Also with this restriction the feasible set has become closed.
Now as $\mathcal{A}$ is assumed to span the entire space, we assume there exists $\mu_j > 0$ such that $\mu_j e_j$ is a convex combination of $\{\pm a\}$. Thus as the ellipsoid is symmetric between $\pm a$ we have:

$$\begin{aligned} &(1-\alpha)(\mu_j e_j)^t H(\mu_j e_j) + \alpha Tr(V_{\pi_{\text{off}}}H) \leq d \\ &h_{jj}(1-\alpha)\mu_j^2 + \alpha Tr(V_{\pi_{\text{off}}}H) \leq d \\ \implies &h_{jj} \leq \frac{d}{1-\alpha}. \end{aligned}$$

This implies there is a uniform bound on the trace of H and hence on the spectral norm. Thus the feasible region for $\mathcal{P}(\mathcal{V},\alpha,d)$ is bounded. Thus, it is compact. As the functional is continuous in H, the minima is attained.

The uniqueness of the optima follows from the strict convexity of the objective.
Now we apply the Karush-Kuhn-Tucker (KKT) conditions to get:

$$-\tau H^{-1} + (1-\alpha)\sum_a u_a a a^t + \alpha V_{\pi_{\text{off}}} = 0,$$
$$u_a((1-\alpha)a^t H a + \alpha Tr(V_{\pi_{\text{off}}} H) - d) = 0$$

for multipliers τ, u_a. Multiplying the first equation with H and taking trace we get:

$$-d\tau + (1-\alpha)\sum_a u_a a^t H a + \alpha \sum_a u_a Tr(V_{\pi_{\text{off}}} H) = 0$$

which with complementary slackness gives

$$-\tau d + d(\sum_a u_a) = 0. \implies \quad \tau = \sum_a u_a$$

Suppose $u_a = 0$ for all a. Then the KKT conditions imply that $\alpha V_{\pi_{\text{off}}} = 0$ which is impossible. Thus we can set $\tau = 1$ by suitably scaling the multipliers u_a. Thus these u_a are dual feasible. Further the KKT conditions imply

$$H^*_\alpha(\mathcal{V}, d) = \left((1-\alpha)\sum_a u_a a a^t + \alpha V_{\pi_{\text{off}}}\right)^{-1}$$

Further we note that

$$-\log(\det(H^*_\alpha(\mathcal{V}, d))) = \log\left(\det\left((1-\alpha)\sum_a u_a a a^t + \alpha V_{\pi_{\text{off}}}\right)\right)$$

that is the primal feasible $H^*_\alpha(\mathcal{V}, d)$ has the same primal objective value as the dual feasible u for the dual objective. By weak duality then there is no duality gap and strong duality holds. □

We characterize the optimality conditions for $\mathcal{P}(\mathcal{V}, \alpha, d)$ and $\mathcal{D}(\mathcal{V}, \alpha)$ in the following proposition

Proposition C.2 (Optimality conditions)**.** *Necessary and sufficient conditions for a PD matrix H and π to be optimal for $\mathcal{P}(\mathcal{V}, \alpha, d)$ and $\mathcal{D}(\mathcal{V}, \alpha)$ respectively are:*

- $\sum_a \pi(a) = 1$ *and* $(1-\alpha)a^t H a + \alpha V_{\pi_{off}} \leq d \; \forall a \in X$.
- $H = ((1-\alpha)\sum_a \pi(a) a a^t + \alpha V_{\pi_{off}})^{-1}$
- $(1-\alpha)a^t H a + \alpha V_{\pi_{off}} = d$ *whenever* $\pi(a) > 0$.

Proof. Condition (a) is just the primal and dual feasibility conditions. From strong duality we know that when H and π are optimal for the primal and dual respectively, necessarily and sufficiently only when there is no duality gap. But from the proof of weak duality we know this happens only when all the eigenvalues of $H((1-\alpha)\sum_a \pi(a) a a^t + \alpha V_{\pi_{\text{off}}})$ are equal and its trace is equal to d. This shows that $H((1-\alpha)\sum_a \pi(a) a a^t + \alpha V_{\pi_{\text{off}}}) = I$ and hence condition (b) follows.
Moreover from the identity below we have

$$Tr(H((1-\alpha)V_\pi + \alpha V_{\pi_{\text{off}}})) = \sum_a \pi(a)((1-\alpha)a^t H a + \alpha Tr(V_{\pi_{\text{off}}} H)) \leq d$$

But since H is feasible it must be the case that $(1-\alpha)a^t H a + \alpha V_{\pi_{\text{off}}} = d$ whenever $\pi(a) > 0$ and hence (c) is also true. □

C.2. Algorithm for $O(d)$-initialization for Frank-Wolfe.

The initializing procedure for the Frank-Wolfe (FW) approximation used in section 5 of the main paper was first suggested in (Betke & Henk, 1993) and later adapted to the *D*-optimal design setting by (Kumar & Yildirim, 2005).

Algorithm 3 O(d) initialization.

Input: $\mathcal{A}_l$.
$c \leftarrow e_1$, $B \leftarrow \emptyset$.
for $i \in 1 : d$ **do**
  $a \leftarrow \underset{a \in \mathcal{A}_l}{\text{argmax}} |\langle c, a \rangle|$.
  $B \leftarrow B \cup \{a\}$.
  $c \leftarrow$ non-zero vector from orthogonal complement of B.
end for
$\pi_l^{(0)} \leftarrow Unif(B)$.
return: $\pi_l^{(0)}$.

In the initialization procedure an arbitrary orthogonal direction c to the vectors in set B is chosen. Then the arm a is added to the set B such that it has the maximum projection (in absolute value) along the direction c. This inductive procedure then keeps adding arms to set B until all the possible d directions are exhausted.

Define the set $\underline{\text{B}} = \cup_{a \in B} \{a, -a\}$. Similarly define $\underline{\mathcal{A}_l}$. This construction ensures that the set $\text{conv}(\underline{\text{B}})$ is contained in the $\text{conv}(\underline{\mathcal{A}_l})$. Further, an induction argument shows the following result:

Proposition C.3 (Theorem 2 in (Betke & Henk, 1993))**.** *Under the initialization procedure above we have the following bounds:*

$$\text{vol}(\text{conv}(\underline{\mathcal{A}_l})) \geq \text{vol}(\text{conv}(\underline{B})) \geq \frac{1}{d!}\text{vol}(\text{conv}(\underline{\mathcal{A}_l})).$$

The above relation will be used in the proof of Proposition 5.2. Intuitively, the bound in Proposition C.3 is true because B is a representative subset of $\mathcal{A}_l$.

C.3. Proof of Initialization Gap (Proposition 5.2).

Recall from Definition 5.3 in section 2 that $H(\pi) = ((1-\alpha)\sum_a \pi(a)aa^t + \alpha V_{\pi_{\text{off}}})^{-1}$. We now introduce the notion of ϵ-feasibility that proves useful in analyzing Algorithm 2:

Definition C.4. A dual feasible π is said to ϵ-primal feasible if $H(\pi)$ satisfies, $\forall a \in \mathcal{V}$,

$$(1-\alpha)a^t H(\pi)a + \alpha Tr(H(\pi)V_{\pi_{\text{off}}}) \leq (1+\epsilon)d$$

and if moreover, $\forall a$ such that $\pi(a) > 0$, it satisfies

$$(1-\alpha)a^t H(\pi)a + \alpha Tr(H(\pi)V_{\pi_{\text{off}}}) \geq (1-\epsilon)d$$

we call π ϵ-approximately optimal.

Now we give the following simple bound

Proposition C.5 (Dual Bound)**.** *If π is ϵ-primal feasible then π is dual feasible and $(1+\epsilon)^{-1}H(\pi)$ is primal feasible and both are within $d\ \log(1+\epsilon)$ of their optimal value.*

Proof. We have

$$\begin{aligned} p\left((1+\epsilon)^{-1}H(\pi), \mathcal{V}, \alpha, d\right) - d(\pi, \mathcal{V}, \alpha) &= d\ \log(1+\epsilon) + \log(\det\left((1-\alpha)\sum_a \pi(a)aa^t + \alpha V_{\pi_{\text{off}}}\right) \\ &\quad - \log\left(\det((1-\alpha)\sum_a \pi(a)aa^t + \alpha V_{\pi_{\text{off}}})\right) \\ &= d\ \log(1+\epsilon) \end{aligned}$$

But as $p\left((1+\epsilon)^{-1}H(\pi), \mathcal{V}, \alpha, d\right) \geq p(H^*_\alpha(\mathcal{V}, d), \mathcal{V}, \alpha, d) = d(\pi^*, \mathcal{V}, \alpha) \geq d(\pi, \mathcal{V}, \alpha)$ from strong duality we get the desired result. □

We recall the following definition of slack given in section 5 where $\tilde{\pi} = (1-\alpha)\pi + \alpha\pi_{\text{off}}$:

$$\delta(\pi) = \frac{(1-\alpha)g_{\mathcal{A}_l}(\tilde{\pi}) + \alpha \sum_{a\in\mathcal{A}} \pi_{\text{off}}(a)\|a\|^2_{V_{\tilde{\pi}}^{-1}}}{d} - 1.$$

then we have the following proposition:

Proposition C.6. *If π is uniform over a subset of $\mathcal{V}$ with size m, then $\delta(\pi) \leq m-1$ and $d(\pi^*, \mathcal{V}, \alpha) - d(\pi, \mathcal{V}, \alpha) \leq d\ \log(m)$.*

Proof. As π is uniform we have $\sum_a \pi(a) w_a(\pi) = d$ (recall w_a, w_{a^+} were defined in Line 2,3 of Algorithm 2) and hence $w_{a_+} \leq dm$. As $\delta(\pi) = \frac{w_{a^+}}{d} - 1$, we have $\delta(\pi) \leq m-1$. Then the dual bound in Proposition C.5 gives that $d(\pi^*, \mathcal{V}, \alpha) - d(\pi, \mathcal{V}, \alpha) \leq d\ \log(m)$. □

We next define feasibility relation between the primal problem in the purely online setting and in the online with offline setting:

Lemma C.7 (Feasibility relation)**.** *We have the following two feasibility results when $\alpha \in [0,1)$:*

1. *The optimal solution $H^*_\alpha(\mathcal{V}, d)$ to the primal problem $\mathcal{P}(\mathcal{V}, \alpha, d)$ is feasible for the primal problem $\mathcal{P}\left(\mathcal{V}, 0, \frac{d-\alpha Tr(H^*_\alpha(\mathcal{V},d)V_{\pi_{off}})}{1-\alpha}\right)$.*
2. *The optimal solution $H^*_0(\mathcal{V}, d)$ to the primal problem $\mathcal{P}(\mathcal{V}, 0, d)$ is feasible for the primal problem $\mathcal{P}(\mathcal{V}, \alpha, d)$.*

Proof. (1) As $H^*_\alpha(\mathcal{V}, d)$ is feasible for $\mathcal{P}(\mathcal{V}, \alpha, d)$ we have that

$$(1-\alpha)a^t H^*_\alpha(\mathcal{V}, d)a + \alpha Tr(H^*_\alpha(\mathcal{V}, d)V_{\pi_{\text{off}}}) \leq d$$

for all $a \in \mathcal{V}$. Thus by simple manipulation of terms we have

$$a^t H^*_\alpha(\mathcal{V}, d)a \leq \frac{d - \alpha Tr(H^*_\alpha(\mathcal{V}, d)V_{\pi_{\text{off}}})}{1-\alpha}.$$

We observe that $d > \alpha Tr(H^*_\alpha(\mathcal{V}, d)V_{\pi_{\text{off}}})$ and $1-\alpha > 0$, and as $H^*_\alpha(\mathcal{V}, d)V_{\pi_{\text{off}}} \succ 0$ we see that $H^*_\alpha(\mathcal{V}, d)$ is feasible for $\mathcal{P}\left(\mathcal{V}, 0, \frac{d-\alpha Tr(H^*_\alpha(\mathcal{V},d)V_{\pi_{\text{off}}})}{1-\alpha}\right)$ and hence (1) is proved.

(2) We note that $\mathcal{P}(\mathcal{V}, 0, d)$ is the standard MVEE problem. Thus its optimal solution $H^*_0(\mathcal{V}, d)$ satisfies

$$a^t H^*_0(\mathcal{V}, d)a \leq d$$

for all $a \in \mathcal{V}$. Using this and the identity $Tr(H^*_0(\mathcal{V}, d)V_{\pi_{\text{off}}}) = \sum_a \pi_{\text{off}}(a) a^t H^*_0(\mathcal{V}, d)a$ we get that

$$\begin{aligned}
&(1-\alpha)a^t H^*_0(\mathcal{V}, d)a + \alpha Tr(H^*_0(\mathcal{V}, d)V_{\pi_{\text{off}}}) \\
\leq &(1-\alpha)d + \alpha \sum_a \pi_{\text{off}}(a) a^t H^*_0(\mathcal{V}, d)a \\
\leq &(1-\alpha)d + \alpha \sum_a \pi_{\text{off}}(a) d \\
= &\quad d
\end{aligned}$$

Using the fact that $H^*_0(\mathcal{V}, d) \succ 0$ we have that $H^*_0(\mathcal{V}, d)$ is feasible for $\mathcal{P}(\mathcal{V}, \alpha, d)$ and hence (2) is proved. □

We note that since $V_{\pi_{\text{off}}}$ is non-singular then the primal problem $\mathcal{P}(\mathcal{V}, 1, d)$ has well defined solution $H^*_1(\mathcal{V}, d) = (V_{\pi_{\text{off}}})^{-1}$. If it isn't then the optimal objective value is $-\infty$ with no optimizer. In the case where the optimizer is well-defined we can straightforwardly extend the above Lemma C.7 for the case $\alpha = 1$.

Let us show the invariance to scale c for the volume of MVEE

Lemma C.8 (Scale invariance of MVEEs)**.** *For all $c > 0$ we have that*

$$vol(\xi(H_0^*(\mathcal{V}, c), c)) = vol(\xi(H_0^*(\mathcal{V}, d), d).$$

Proof. We observe that $\frac{d}{c}H_0^*(\mathcal{V}, c)$ is feasible for $\mathcal{P}(\mathcal{V}, 0, d)$ and $\frac{c}{d}H_0^*(\mathcal{V}, d)$ is feasible for $\mathcal{P}(\mathcal{V}, 0, c)$. Thus we have

$$-\log(\det(\frac{d}{c}H_0^*(\mathcal{V}, c))) \geq -\log(\det(H_0^*(\mathcal{V}, d)))$$
$$-\log(\det(\frac{c}{d}H_0^*(\mathcal{V}, d))) \geq -\log(\det(H_0^*(\mathcal{V}, c))).$$

From these we get the inequalities,

$$\left(\frac{d}{c}\right)^d \det(H_0^*(\mathcal{V}, c)) \leq \det(H_0^*(\mathcal{V}, d))$$

and

$$\left(\frac{c}{d}\right)^d \det(H_0^*(\mathcal{V}, d)) \leq \det(H_0^*(\mathcal{V}, c)).$$

These then imply

$$\left(\frac{c}{d}\right)^d \det(H_0^*(\mathcal{V}, d)) = \det(H_0^*(\mathcal{V}, c)).$$

By the volume formula for ellipsoids we have :

$$\begin{aligned} vol(\xi(H_0^*(\mathcal{V}, c), c)) &= \frac{c^{d/2}B_d}{\sqrt{\det(H_0^*(\mathcal{V}, c))}} \\ &= \frac{c^{d/2}B_d}{\sqrt{\left(\frac{c}{d}\right)^d \det(H_0^*(\mathcal{V}, d))}} \\ &= \frac{d^{d/2}B_d}{\sqrt{\det(H_0^*(\mathcal{V}, d))}} \\ &= vol(\xi(H_0^*(\mathcal{V}, d), d). \end{aligned}$$

□

Now we are ready to give a proof of Proposition 5.2:

Proof of Proposition 5.2. In light of the formula for a volume of an ellipsoid $\xi(H, c)$,

$$\text{vol}(\xi(H, c)) = \frac{c^{d/2}B_d}{\sqrt{\det(H)}},$$

it is clear that the primal problem $\mathcal{P}(\mathcal{V}, \alpha, d)$ is equivalent to minimizing the volume of the ellipsoid $\xi(H, d)$ with constraint on H such that $(1-\alpha)a^t Ha + \alpha Tr(HV_{\pi_o}) \leq d$ for all $a \in \mathcal{V}$.
Setting $\mathcal{V} = B$, where B is the support of initialization procedure described in section C.2 for the set $\mathcal{A}_l$, from Lemma C.7 part (1) we get:

$$-\log(\det(H_\alpha^*(B, d))) \geq -\log\left(\det\left(H_0^*\left(B, \frac{d - \alpha Tr(H_\alpha^*(B, d)V_{\pi_{\text{off}}})}{1-\alpha}\right)\right)\right)$$

which implies then that

$$vol\left(\xi\left(H_\alpha^*(B, d), \frac{d - \alpha Tr(H_\alpha^*(B, d)V_{\pi_{\text{off}}})}{1-\alpha}\right)\right) \geq$$
$$vol\left(\xi\left(H_0^*\left(B, \frac{d - \alpha Tr(H_\alpha^*(B, d)V_{\pi_{\text{off}}})}{1-\alpha}\right), \frac{d - \alpha Tr(H_\alpha^*(B, d)V_{\pi_{\text{off}}})}{1-\alpha}\right)\right).$$

Now from scale invariance of Lemma C.8 we know that

$$vol\left(\xi\left(H_0^*\left(B, \frac{d-\alpha Tr(H_\alpha^*(B,d)V_{\pi_{\text{off}}})}{1-\alpha}\right), \frac{d-\alpha Tr(H_\alpha^*(B,d)V_{\pi_{\text{off}}})}{1-\alpha}\right)\right) = vol(\xi(H_0^*(B,d),d))$$

and from the volume formula for ellipsoids we have the fact that

$$vol\left(\xi\left(H_\alpha^*(B,d), \frac{d-\alpha Tr(H_\alpha^*(B,d)V_{\pi_{\text{off}}})}{1-\alpha}\right)\right) =$$
$$\left(\frac{d-\alpha Tr(H_\alpha^*(B,d)V_{\pi_{\text{off}}})}{d(1-\alpha)}\right)^{d/2} vol(\xi(H_\alpha^*(B,d),d))$$

using which we get the inequality

$$\left(\frac{d-\alpha Tr(H_\alpha^*(B,d)V_{\pi_{\text{off}}})}{d(1-\alpha)}\right)^{d/2} vol(\xi(H_\alpha^*(B,d),d)) \geq vol(\xi(H_0^*(B,d),d)).$$

Now as the $\xi(H_0^*(B,d),d)$ is the MVEE of B we have that $vol(\xi(H_0^*(B,d),d)) \geq vol(\text{conv}(\underline{B}))$, where $\underline{B}$ is as defined in Appendix C.2 and hence

$$\left(\frac{d-\alpha Tr(H_\alpha^*(B,d)V_{\pi_{\text{off}}})}{d(1-\alpha)}\right)^{d/2} vol(\xi(H_\alpha^*(B,d),d)) \geq vol(\text{conv}(\underline{B})).$$

As remarked in in Appendix C.2, the $O(d)$ initialization procedure there has the property $vol(\text{conv}(\underline{B})) \geq \frac{1}{d!} vol(\text{conv}(\underline{\mathcal{A}_l}))$ we get that

$$\left(\frac{d-\alpha Tr(H_\alpha^*(B,d)V_{\pi_{\text{off}}})}{d(1-\alpha)}\right)^{d/2} vol(\xi(H_\alpha^*(B,d),d)) \geq \frac{1}{d!} vol(\text{conv}(\underline{\mathcal{A}_l})).$$

Now John's theorem (see Theorem 1.1 in (Todd, 2016)) on MVEE states that for any finite set $\mathcal{C}$ we have $\frac{1}{d}\text{MVEE}(\underline{\mathcal{C}}) \subset \text{conv}(\underline{\mathcal{C}})$ and using the fact that $\text{MVEE}(\underline{\mathcal{C}}) = \text{MVEE}(\mathcal{C})$ gives us

$$\left(\frac{d-\alpha Tr(H_\alpha^*(B,d)V_{\pi_{\text{off}}})}{d(1-\alpha)}\right)^{d/2} vol(\xi(H_\alpha^*(B,d),d)) \geq \frac{1}{d!d^d} vol(\xi(H_0^*(\mathcal{A}_l,d),d)).$$

Now from Lemma C.7 part 2, with $\mathcal{V} = \mathcal{A}_l$, we have that

$$-\log(\det(H_0^*(\mathcal{A}_l,d))) \geq -\log(\det(H_\alpha^*(\mathcal{A}_l,d)))$$

and hence $vol(\xi(H_0^*(\mathcal{A}_l,d),d)) \geq vol(\xi(H_\alpha^*(\mathcal{A}_l,d),d))$. This give us

$$\left(\frac{d-\alpha Tr(H_\alpha^*(B,d)V_{\pi_{\text{off}}})}{d(1-\alpha)}\right)^{d/2} vol(\xi(H_\alpha^*(B,d),d)) \geq \frac{1}{d!d^d} vol(\xi(H_\alpha^*(\mathcal{A}_l,d),d)).$$

Now using the fact that $Tr(H_\alpha^*(B,d)V_{\pi_{\text{off}}}) \geq 0$ we get that

$$\frac{1}{(1-\alpha)^{d/2}} \geq \left(\frac{d-\alpha Tr(H_\alpha^*(B,d)V_{\pi_{\text{off}}})}{d(1-\alpha)}\right)^{d/2}.$$

Thus, we have

$$vol(\xi(H_\alpha^*(B,d),d)) \geq \frac{(1-\alpha)^{d/2}}{d!d^d} vol(\xi(H_\alpha^*(\mathcal{A}_l,d),d)).$$

We have finally managed to connect the optimal ellipsoids of the initialized set B and the overall set $\mathcal{A}_l$. We have:

$$d(\pi_{l,\text{on}}^*, \mathcal{A}_l, \alpha) - d(\pi_l^{(0)}, B, \alpha) = d(\pi_{l,\text{on}}^*, \mathcal{A}_l, \alpha) - d(\pi^*(B), B, \alpha) + d(\pi^*(B), B, \alpha) - d(\pi_l^{(0)}, B, \alpha).$$

where $\pi^*(B)$ is the optimizer to the dual problem $\mathcal{D}(B,\alpha)$ and that $\pi_{l,\text{on}}^*$ is the optimizer of $\mathcal{D}(\mathcal{A}_l,\alpha)$ (recall the definition of $\pi_{l,\text{on}}^*$ from (2)). Then by proposition C.6 we have that $d(\pi^*(B), B, \alpha) - d(\pi_l^{(0)}, B, \alpha) \leq d\, ln(d)$.

Next we try to bound $d(\pi^*_{l,\text{on}}, \mathcal{A}_l, \alpha) - d(\pi^*(B), B, \alpha)$. From the volume of an ellipsoid formula and strong duality we have the following identities:

$$d(\pi^*(B), B, \alpha) = 2\log(\text{vol}(\xi(H^*_\alpha(B,d),d))) - d\ \log(d) - 2\log(B_d)$$
$$d(\pi^*_{l,\text{on}}, \mathcal{A}_l, \alpha) = 2\log(\text{vol}(\xi(H^*_\alpha(\mathcal{A}_l,d),d))) - d\ \log(d) - 2\log(B_d).$$

where as before B_d is the volume of the unit ball in $\mathbb{R}^d$. Thus

$$d(\pi^*_{l,\text{on}}, \mathcal{A}_l, \alpha) - d(\pi^*(B), B, \alpha) = 2\log\left(\frac{vol(\xi(H^*_\alpha(\mathcal{A}_l,d),d))}{vol(\xi(H^*_\alpha(B,d),d))}\right).$$

But we know the ratio of volumes is upper bounded by $\frac{d^d d!}{(1-\alpha)^{d/2}}$ and thus we have

$$\begin{aligned} d(\pi^*_{l,\text{on}}, \mathcal{A}_l, \alpha) - d(\pi^*(B), B, \alpha) &\leq 2\log\left(\frac{d^d d!}{(1-\alpha)^{d/2}}\right) \\ &\leq 4d\ \log(d) - d\log(1-\alpha). \end{aligned}$$

Thus we have

$$\begin{aligned} d(\pi^*_{l,\text{on}}, \mathcal{A}_l, \alpha) - d(\pi^*(B), B, \alpha) &\leq (d\ \log(d)) + (4d\ \log(d) - d\log(1-\alpha)) \\ &= 5d\ \log(d) - d\ \log(1-\alpha). \end{aligned}$$

This concludes the proof. □

C.4. Proof of Lemma 5.4.

Proof. The update of FW in algorithm 2 is given by

$$\pi_l^{(t+1)} = (1+\beta)^{-1}(\pi_l^{(t)} + \beta \mathbb{1}_{\{a_+\}})$$

where $\beta = \frac{(w_{a_+} - d)}{(d-1)w_{a_+}}$, $a_+ = \underset{a}{argmax} \quad w_a$ and $w = \left(Tr\left(H(\pi_l^{(t)})\left((1-\alpha)aa^t + \alpha V_{\pi_{\text{off}}}\right)\right)\right)_{a\in\mathcal{A}_l}$.
Then by matrix determinant lemma we have

$$\begin{aligned} d(\pi_l^{(t+1)}, \mathcal{A}_l, \alpha) &= \log\ \det\left((1-\alpha)\frac{(V_{\pi_l^{(t)}} + \beta a_+ a_+^t)}{(1+\beta)} + \alpha V_{\pi_{\text{off}}}\right) \\ &= \log\left((1+\beta)^{-d}\det\left((1-\alpha)V_{\pi_l^{(t)}} + \alpha V_{\pi_{\text{off}}} + \beta((1-\alpha)a_+a_+^t + \alpha V_{\pi_{\text{off}}})\right)\right. \\ &= -d\log(1+\beta) + d(\pi_l^{(t)}, \mathcal{A}_l, \alpha) + \log(\det(I + \beta H(\pi_l^{(t)})((1-\alpha)a_+a_+^t + \alpha V_{\pi_{\text{off}}}))) \end{aligned}$$

Now the log-determinant in last term above can be re-written using the eigenvalues λ_k of $H(\pi_l^{(t)})((1-\alpha)a_+a_+^t + \alpha V_{\pi_{\text{off}}})$ as:

$$\log(\det(I + \beta H(\pi_l^{(t)})((1-\alpha)a_+a_+^t + \alpha V_{\pi_{\text{off}}}))) = \sum_{k=1}^d \log(1+\beta\lambda_k).$$

We observe that as the eigenspectrum of $H(\pi_l^{(t)})((1-\alpha)a_+a_+^t + \alpha V_{\pi_{\text{off}}})$ is the same as the positive-semidefinite matrix $H^{1/2}(\pi_l^{(t)})((1-\alpha)a_+a_+^t + \alpha V_{\pi_{\text{off}}})H^{1/2}(\pi_l^{(t)})$ we can conclude that $\lambda_k \geq 0$ for all $k \in [d]$.
Now using Lemma 1 in (Merhav, 2022) (we set $f(x) = ln(1+\beta x), a = \sum_k \lambda_k$ and $\mu = \frac{\sum_k \lambda_k}{d}$ in their Lemma 1), which is a *reverse* Jensen type inequality, to get

$$\begin{aligned} \log(\det(I + \beta H(\pi_l^{(t)})((1-\alpha)a_+a_+^t + \alpha V_{\pi_{\text{off}}}))) &= \sum_{k=1}^d \log(1+\beta\lambda_k) \\ &\geq \log\left(1 + \beta\sum_{k=1}^d \lambda_k\right). \end{aligned}$$

But as

$$\log\left(1+\beta\sum_{k=1}^{d}\lambda_k\right) = \log\left(1+\beta Tr\big(H(\pi_l^{(t)})((1-\alpha)a_+a_+^t+\alpha V_{\pi_{\text{off}}})\big)\right)$$
$$= \log(1+\beta w_{a_+}),$$

we have:

$$d(\pi_l^{(t+1)},\mathcal{A}_l,\alpha) - d(\pi_l^{(t)},\mathcal{A}_l,\alpha) \geq -d\log(1+\beta)+\log(1+\beta w_{a_+}).$$

Hence:

$$d(\pi_l^{(t+1)},\mathcal{A}_l,\alpha) - d(\pi_l^{(t)},\mathcal{A}_l,\alpha) \geq (d-1)\log\left(\frac{(d-1)w_{a_+}}{d(w_{a_+}-1)}\right)+\log\left(\frac{w_{a_+}}{d}\right). \tag{31}$$

Recall that:

$$\delta(\pi_l^{(t)}) := \frac{w_{a_+}(\pi_l^{(t)})}{d}-1.$$

Using this the inequality in equation (31) is re-written to get

$$d(\pi_l^{(t+1)},\mathcal{A}_l,\alpha) - d(\pi_l^{(t)},\mathcal{A}_l,\alpha) \geq \log(1+\delta(\pi_l^{(t)})) - (d-1)\log\left(1+\frac{\delta(\pi_l^{(t)})}{(d-1)(1+\delta(\pi_l^{(t)}))}\right)$$
$$\geq \log(1+\delta(\pi_l^{(t)})) - \frac{\delta(\pi_l^{(t)})}{1+\delta(\pi_l^{(t)})}$$
$$:= m(\delta(\pi_l^{(t)}))$$

This completes the proof of Lemma 5.4. □

Now $m(\delta)$ satisfies the following simple properties (see lemma 3.6 in (Todd, 2016) for a proof)

Lemma C.9. *We have*

1. *$m(\delta)$ is increasing if $\delta \geq 0$ and decreasing if $\delta < 0$.*
2. *for $\delta \geq \delta_0$, $m(\delta) \geq \left(1-\frac{\delta_0}{(1+\delta_0)(ln(1+\delta_0)}\right)ln(1+\delta)$.*
3. *for $|\delta| \leq 1/2$, $m(\delta) \geq 2/7\delta^2$.*

C.5. Proof of Proposition 5.5.

Proof. Consider an iterate $\pi_l^{(t)}$ such that $\delta(\pi_l^{(t)}) \geq 1$. Set $\gamma_t = d(\pi^*_{l,\text{on}},\mathcal{A}_l,\alpha) - d(\pi_l^{(t)},\mathcal{A}_l,\alpha)$. Then

$$\gamma_t - \gamma_{t+1} = d(\pi_l^{(t+1)},\mathcal{A}_l,\alpha) - d(\pi_l^{(t)},\mathcal{A}_l,\alpha)$$
$$\geq \log(1+\delta(\pi_l^{(t)})) - \frac{\delta(\pi_l^{(t)})}{1+\delta(\pi_l^{(t)})}$$
$$\geq \left(1-\frac{\delta_0}{(1+\delta_0)(ln(1+\delta_0)}\right)\log(1+\delta(\pi_l^{(t)}))$$
$$\geq \left(\frac{1}{d}-\frac{\delta_0}{d(1+\delta_0)(ln(1+\delta_0)}\right)\gamma.$$

The first inequality follows from Lemma 5.4, the second inequality from Lemma C.9 (2), and the third from proposition C.5. We set $k(\delta_0) := \left(1-\frac{\delta_0}{(1+\delta_0)(ln(1+\delta_0)}\right)$. We have:

$$\gamma_{t+1} \leq \left(1-\frac{k(\delta_0)}{d}\right)\gamma_t \leq exp\left(-\frac{k(\delta_0)}{d}\right)\gamma_t.$$

Since the initialization has an upper bound of $5d\ ln(d) - d\ ln(1-\alpha)$ then within

$$\frac{d}{k(\delta_0)} ln\left(\frac{d}{\delta_0} ln\left(\frac{d^5}{1-\alpha}\right)\right)$$

iterations γ_t is at most δ_0. □

C.6. Proof of Proposition 5.7.

Since $\pi_l^{(t)}$ satisfies $\delta(\pi_l^{(t)}) \leq \frac{d_{\text{eff}}}{d}$ we have by definition of δ that:

$$(1-\alpha)||a||^2_{V^{-1}_{\tilde{\pi}_l^{(t)}}} + \alpha \sum_a \pi_{\text{off}}(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^{(t)}}} \leq d(1 + \frac{d_{\text{eff}}}{d}).$$

for each $a \in \mathcal{A}_l$. We now observe that similar to the proof of Lemma 4.3 we have:

$$\begin{aligned} d - \alpha \sum_a \pi_{\text{off}}(a)||a||^2_{V^{-1}_{\tilde{\pi}_l^{(t)}}} &= Tr\left(\left(I + \frac{\alpha}{1-\alpha} V_{\pi_{\text{off}}} V^{\dagger}_{\pi_l^{(t)}}\right)^{-1}\right) \\ &\leq d_{\text{eff}}. \end{aligned}$$

This gives us the bound:

$$(1-\alpha)||a||^2_{V^{-1}_{\tilde{\pi}_l^{(t)}}} \leq 2d_{\text{eff}}$$

for each $a \in \mathcal{A}_l$ and completes the proof of the lemma.

C.7. Proof of Theorem 5.8.

Proof. The proof is quite similar to the proof of Theorem 4.4. The only difference is we allow FW to solve only up to $\delta = \frac{d_e}{d}$ instead of setting $\delta = 0$. The upper bound on the confidence width is supplied by Proposition 5.7 instead of Lemma 4.3. Now using much the same techniques to arrive at (21) we get the following regret bound:

$$\mathcal{R}(\texttt{OOPE-FW}) \leq vT + 48 d_{\text{eff}} \log(4 l_{max}^2 |\mathcal{A}| T) \left(\sum_{l=1}^{l_M} 2^l\right) + \sum_{l=1}^{l_M} 8\epsilon_l(|\text{supp}(\pi_l^{(t)})|) \tag{32}$$

Now from Theorem 5.5 we know that $|\text{supp}(\pi_l^{(t)})|$ is upper bounded by $4d\log(d\log(\left(\frac{d^5}{1-\alpha}\right))) + d + \frac{28d}{\delta}$. Using this bound instead of $d(d+1)/2$ bound we get in much the same way as Step 5 in Theorem 4.4's proof:

$$\begin{aligned} \mathcal{R}(\texttt{OOPE-FW}) \leq vT + 96 d_{\text{eff}} \log(4 l_{max}^2 |\mathcal{A}| T) \min\left\{\frac{8}{v}, \sqrt{4 + \frac{T}{d_{\text{eff}} \log(4|\mathcal{A}|T)}}\right\} + 8d + \frac{224 d^2}{d_{\text{eff}}} \\ + 32 d \log\left(d \log\left(\left(\frac{d^5(T + T_{\text{off}})}{T}\right)\right)\right) \end{aligned}$$

Now optimizing over v we get that:

$$\mathcal{R}(\texttt{OOPE-FW}) \leq 32\sqrt{3 d_{\text{eff}} T \log(4 l_{max}^2 |\mathcal{A}| T)} + 8d + \frac{224 d^2}{d_{\text{eff}}} + 32 d \log\left(d \log\left(\left(\frac{d^5(T + T_{\text{off}})}{T}\right)\right)\right)$$

and this concludes the proof. □