Neural Variance-aware Dueling Bandits with Deep Representation and Shallow Exploration

Youngmin Oh[*]
InfiniTree

Jinje Park
Samsung Electronics

Taejin Paik
Samsung Electronics

Abstract

We introduce the first variance-aware algorithms for contextual dueling bandits that leverage shallow exploration strategies with neural networks for nonlinear utility approximation. A key theoretical challenge is the absence of a closed-form estimator, which led prior work to require an extremely large network width m (i.e., $m = \widetilde{\Omega}(T^{14})$). We address this constraint with a novel analytical approach that combines iterative self-improvement with spectral analysis. Our analysis significantly reduces the network width requirement to $m = \widetilde{\Omega}(T^6)$, and shows that our algorithms achieve a sublinear regret of $\widetilde{\mathcal{O}}\left(d\sqrt{\sum_{t=1}^{T}\sigma_t^2} + \sqrt{dT}\right)$ under both UCB and TS frameworks. Empirical results show that the proposed algorithms are not only computationally efficient and exhibit sublinear regret in practical settings, but also achieve state-of-the-art performance on both synthetic and real-world tasks.[1]

1 INTRODUCTION

The contextual dueling bandits (CDBs) (Dudík et al., 2015; Yue et al., 2012) extend the standard contextual multi-armed bandit (CMAB) framework to handle pairwise preference feedback instead of numerical rewards. In each round, a learner observes a context, selects two arms to compare, and receives a preference outcome. This outcome is commonly sampled by a Bernoulli distribution, where the probability is a function of the arms' latent utilities under models such as Bradley–Terry–Luce (Hunter, 2004; Luce, 2005).

Many existing studies on the CDB framework assume a linear utility function (Bengs et al., 2022; Saha et al., 2021; Saha, 2021; Li et al., 2024; Bui et al., 2024). However, the linearity assumption imposes a significant limitation on its ability to model real-world preferences, as real-world user preferences often exhibit complex, nonlinear interactions among features that a simple linear model cannot fully capture.

To address this limitation, Verma et al. (2024) recently proposed neural dueling bandit algorithms based on the Upper Confidence Bound (UCB) and Thompson Sampling (TS) frameworks. These algorithms employ a neural network to approximate a general nonlinear utility function and are proven to achieve sublinear cumulative average regret when the network width is sufficiently large. Despite these theoretical advancements, their approach requires constructing a Gram matrix from the gradients of all learnable parameters, which incurs substantial computational overhead and poses practical challenges for large-scale applications.

In the CMAB literature, computational challenges have been mitigated by adopting shallow exploration strategies, which construct Gram matrices using only the gradients of the final layer (Bui et al., 2024; Xu et al., 2020). Separately, variance-aware methods have been actively studied in linear CMABs and dueling bandits (Di et al., 2023; Zhou and Gu, 2022; Zhou et al., 2021a; Zhang et al., 2021b, 2022) and were recently extended to neural bandits (Bui et al., 2024). However, to the best of our knowledge, the combination of variance-aware approaches with shallow exploration remains unexplored in the context of dueling bandits.

A key observation motivating variance-awareness in the dueling setting is that *harder comparisons produce noisier feedback*: the Bernoulli variance $g(\Delta u_t)(1-g(\Delta u_t))$ is maximized when $\Delta u_t \approx 0$—precisely when the comparison is hardest. Our variance-based weighting acts as an *information filter*, down-weighting uncertain ob-

[*]Corresponding author: youngmin0.oh@gmail.com

[1]Official code is available at https://github.com/youngmin0oh/NVLDB-AISTATS2026.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

Table 1: Comparison of Bandit Algorithms: Regret, Assumptions, and Explicit Form of θ_t

Name	**Regret** (without the term m)	**Assumption of m** for sublinear regrets	**Explicit form** of θ
Neural Bandits (Zhou et al., 2020)	$\widetilde{\mathcal{O}}(d\sqrt{T})$	$m = \widetilde{\Omega}(T^7)$	None
Neural Linear Bandits (Xu et al., 2020)	$\widetilde{\mathcal{O}}(d\sqrt{T})$	$m = \widetilde{\Omega}(T^3)$	$\hat{\theta}_{t-1} = \mathbf{V}_{t-1}^{-1}\mathbf{b}_{t-1}$
Neural-σ^2-LinearUCB (Bui et al., 2024)	$\widetilde{\mathcal{O}}\left(\sqrt{dT} + d\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right)$	$m = \widetilde{\Omega}(T^3)$	$\hat{\theta}_{t-1} = \mathbf{V}_{t-1}^{-1}\mathbf{b}_{t-1}$
Neural Dueling Bandits (Verma et al., 2024)	$\widetilde{\mathcal{O}}(d\sqrt{T})$	$m = \widetilde{\Omega}(T^{14})$	None
Our Method (variance-agnostic)	$\widetilde{\mathcal{O}}(d\sqrt{T})$	$m = \widetilde{\Omega}(T^6)$	None
Our Method (variance-aware)	$\widetilde{\mathcal{O}}\left(\sqrt{dT} + d\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right)$	$m = \widetilde{\Omega}(T^6)$	None

servations from small-margin pairs to achieve tighter, variance-dependent regret bounds.

Motivated by these gaps, we propose a method we term **Neural Variance-Aware Linear Dueling Bandits** (NVLDB). Our approach constructs the Gram matrix using only the gradients of the final layer's parameters, thereby improving computational efficiency compared to the method of Verma et al. (2024). To achieve variance-awareness, NVLDB estimates the variance of the underlying Bernoulli distribution, using the neural network as a nonlinear utility approximator. Furthermore, NVLDB supports various arm selection strategies—including one asymmetric and two symmetric variants—within both the UCB and TS frameworks.

Under a significantly weaker condition on the neural network width m than that required by previous work (Verma et al., 2024), we prove that, under standard assumptions, the expected cumulative average regret for each of our arm selection strategies is bounded by $\widetilde{\mathcal{O}}\left(d\sqrt{\sum_{t=1}^{T}\sigma_t^2} + d\sqrt{T}\right)$. Here, d is the contextual dimension, and σ_t^2 represents the variance of the Bernoulli preference model at round t. We also establish that variance-agnostic variants of our algorithm achieve a sublinear regret bound of $\widetilde{\mathcal{O}}\left(d\sqrt{T}\right)$ under the same standard assumptions.

A fundamental challenge that distinguishes the theoretical analysis of CDB from the CMAB setting is the **absence of a closed-form solution for the parameter estimate** $\hat{\theta}_{t-1}$. Analyses in the CMABs under shallow exploration literature (e.g., (Bui et al., 2024)) hinge on an explicit estimator, typically of the form $\hat{\theta}_{t-1} = \mathbf{V}_{t-1}^{-1}\mathbf{b}_{t-1}$. This closed-form results in a time-independent bound on the estimator's norm, $\|\hat{\theta}_{t-1}\|_2$, which is crucial for proving sublinear regret of neural bandits with a moderately sized neural network.

In the dueling bandit setting, however, no such solution exists, even with shallow exploration. The absence of this analytical expression forced prior work (Verma et al., 2024) to assume an extremely large neural network width of $m = \widetilde{\Omega}\left(T^{14}\right)$ to guarantee sublinear cumulative regret. Such a strong requirement is inconsistent with results from related bandit literature (Zhou et al., 2020; Bui et al., 2024; Xu et al., 2020).

In this paper, we overcome this critical limitation by introducing a novel analytical technique that combines a bootstrap argument (Gilbarg et al., 1977) (iterative self-improvement) with spectral analysis to prove that the estimator norm $\|\hat{\theta}_{t-1}\|_2$ is bounded by a time-independent constant, even without an explicit form for the estimator, with a network width of $m = \widetilde{\Omega}(T^6)$. Consequently, we prove **sublinear regret** for **both** Upper Confidence Bound (UCB) and Thompson Sampling (TS) arm selections under standard assumptions that are customary in the neural (linear) bandit literature (Xu et al., 2020; Bui et al., 2024). Although the width condition is a stronger assumption compared to neural bandits under shallow exploration (Bui et al., 2024; Xu et al., 2020), the gap is originated from the absence of the closed form of θ_{t-1} even under the shallow exploration in the case of dueling bandits. Nevertheless, this represents a significant improvement over the $m = \widetilde{\Omega}(T^{14})$ required by previous work (Verma et al., 2024). A brief comparison with prior results is presented in Table 1. We demonstrate that these results are obtained under standard assumptions that are customary in the neural (linear) bandit literature. We provide further discussion in Section 5, and defer the full proofs to the supplementary material.

Our contributions are summarized as follows:

1. We propose NVLDB, the first framework for neural variance-aware linear dueling bandits with a shallow exploration strategy that significantly reduces computational overhead compared to full-network approaches (Verma et al., 2024).

2. We provide a rigorous theoretical analysis proving that, under standard assumptions, NVLDB achieves sublinear cumulative regret. This guarantee holds for both its variance-aware and variance-agnostic variants.

3. Our analysis establishes a required network width of $m = \widetilde{\Omega}(T^6)$. This is a substantial improvement over the $m = \widetilde{\Omega}(T^{14})$ required by prior work (Verma et al., 2024) and brings the requirement in line with that of standard neural bandits with shallow exploration.

4. Extensive experiments on synthetic and real-world datasets demonstrate that NVLDB empirically achieves sublinear regret and consistently outperforms existing contextual dueling bandit methods.

2 RELATED RESULTS

Linear Contextual Dueling Bandits. CDBs with a linear utility function have been the subject of active study in recent years (Bengs et al., 2022; Saha, 2021; Di et al., 2023; Li et al., 2024). The UCB-based approaches remain effective in this formulation. Indeed, Bengs et al. (2022) and Saha (2021) proposed UCB-based methods and showed that the cumulative regret is bounded by $\widetilde{\mathcal{O}}(d\sqrt{T})$ and $\widetilde{\mathcal{O}}(\sqrt{dT})$. Di et al. (2023) introduced a variance-aware action-elimination method and showed that the cumulative regret remains bounded. Li et al. (2024) utilized Thompson Sampling and proposed an algorithm called FGTS.CDB, showing that the cumulative regret is bounded by $\widetilde{\mathcal{O}}(d\sqrt{T})$. In addition, there exist lower bounds of order $\widetilde{\mathcal{O}}(\sqrt{dT})$ (Saha, 2021) for both the cumulative average and weak regret.

Neural Bandits. To the best of our knowledge, neural networks were first employed as nonlinear reward estimators in MABs (Zhou et al., 2020; Zhang et al., 2021a; Deb et al., 2023; Ban et al., 2021). Ban et al. (2021) adds a second network for potential-gain estimation but must retain all past checkpoints, whereas Deb et al. (2023) rely on perturbed-prediction regression oracles that require several forward passes per round. Zhou et al. (2020) and Zhang et al. (2021a) proposed neural UCB and TS algorithms, respectively, both achieving sublinear cumulative average regret with a sufficiently large network width and bounded effective dimension. To address computational challenges of previous results, Xu et al. (2020) introduced shallow exploration approach that leverages the gradients on the last layer's parameters to construct a Gram matrix. Building on these pioneering results, neural networks have been widely adopted in various bandit settings, such as combinatorial bandits (Hwang et al., 2023) and federated bandits (Dai et al., 2022). Moreover, Verma et al. (2024) proposed both UCB and TS-based strategies utilizing neural networks to address contextual dueling bandits with nonlinear utility functions.

Variance-aware Linear Bandits. Variance-aware approaches for linear contextual bandits can be categorized into those that require knowledge of the true variance at each round (Zhou et al., 2021a,b) and those that do not (Zhao et al., 2023; Kim et al., 2022; Zhang et al., 2021b). Before the introduction of the method in Zhao et al. (2023), methods that did not rely on the true variance often exhibited worse regret bounds compared to their variance-known counterparts. In contrast, Zhao et al. (2023) achieves computational efficiency while matching the regret bounds of methods that assume variance knowledge, thus bridging the gap between these two lines of work. Such variance-aware approaches have also been applied to neural bandits (Bui et al., 2024) and to contextual linear dueling bandits (Di et al., 2023), yielding sublinear cumulative average regret. However, to the best of our knowledge, there has been no prior work on neural variance-aware dueling bandits.

3 PROBLEM SETTING

Contextual Dueling Bandits. A learner observes a context set $X_t = \{x_{t,1}, \cdots, x_{t,K}\} \subset \mathcal{X} \subset \mathbb{R}^d$ at each round $0 < t \leq T$, where T is the total number of rounds, K is the number of arms, and $\mathcal{X}$ is the overall context space. Then the learner chooses two contexts $\left(x_{t,k_{t,1}}, x_{t,k_{t,2}}\right)$ in X_t where $k_{t,1}, k_{t,2} \in [K]$, which is equivalent to pull a tuple of arms $(k_{t,1}, k_{t,2})$. As a result, the learner receives a binary preference feedback signal o_t. The feedback o_t is sampled from a Bernoulli distribution $\mathcal{B}(p_t)$, where p_t depends on the chosen pair of contexts. We assume that the probability p_t is determined by a link function $g : \mathbb{R} \to [0, 1]$, depending on latent utilities. The latent utility is defined by an unknown function $u : \mathbb{R}^d \to \mathbb{R}$, which takes a context as input. In other words, $p_t = \mathbb{P}\left(o_t = 1 \mid x_{t,k_{t,1}}, x_{t,k_{t,2}}\right) = g\left(\Delta u_t\right)$, where $\Delta u_t := u\left(x_{t,k_{t,1}}\right) - u\left(x_{t,k_{t,2}}\right)$. We assume that the link function belongs to the exponential family (Bengs et al., 2022). The exponential family contains various link functions, including the Bradley-Terry-Luce (BTL) model (Hunter, 2004; Luce, 2005), the Thurstone-Mosteller model (Thurstone, 2017), and the Exponential Noise model (Jain and Natarajan, 2016). Under this class, the Fisher Information and variance satisfy $I(\Delta u_t) = c \cdot \sigma_t^2$ for a bounded constant $c > 0$ (with $c = 1$ for BTL, Thurstone-Mosteller, and Exponential Noise models) (McCullagh, 2019), so our inverse-variance weighting is equivalent to the inverse

Fisher Information weighting (Amari and Douglas, 1998). The gradient of the log-likelihood loss within this family has a particular form, which is utilized in the theoretical analysis of cumulative regret.

Utility Function. The linear utility function u has been extensively studied in contextual dueling bandits (Saha, 2021; Bengs et al., 2022; Di et al., 2023; Li et al., 2024). However, real-world utility structures are often far more complex. To address this, recent studies have explored nonlinear utility functions by approximating them with neural networks (Verma et al., 2024). Building on this idea and following a similar approach to that of Xu et al. (2020), we assume that the utility function u can be expressed as an inner product between a feature map $\phi_* : \mathbb{R}^d \to \mathbb{R}^d$ and an unknown true parameter $\theta_* \in \mathbb{R}^d$, i.e.,

$$u\left(x_{t,k}\right) = \theta_*^\intercal \phi_*\left(x_{t,k}\right), \tag{1}$$

which generalizes the linear setting when the feature map is the identity function.

3.1 Objective

The two most common types of instantaneous regret are average regret r_t^a and weak regret r_t^w (Saha, 2021):

$$r_t^a = \frac{2u\left(x_{t,k_t^*}\right) - u\left(x_{t,k_{t,1}}\right) - u\left(x_{t,k_{t,2}}\right)}{2},$$
$$r_t^w = u\left(x_{t,k_t^*}\right) - \max\left\{u\left(x_{t,k_{t,1}}\right), u\left(x_{t,k_{t,2}}\right)\right\},$$

where $k_t^* = \arg\max_{k\in[K]} u\left(x_{t,k}\right)$. Then we define the cumulative average and weak regrets as $R_T^a = \sum_{t=1}^T r_t^a$ and $R_T^w = \sum_{t=1}^T r_t^w$, respectively. An efficient algorithm for the stochastic contextual dueling bandits guarantees the sublinear increase of R_t^a and R_t^w on T, e.g., $\widetilde{\mathcal{O}}(\sqrt{T})$. The weak and average cumulative regrets have individual advantages. For instance, the latter is more useful than the former when selecting the inferior action has significant side effects, e.g., clinical trials. The objective of this paper is to minimize the expected cumulative average regret R_T^a as (Bengs et al., 2022).

4 METHOD

4.1 Nonlinear Utility Function Estimator

To achieve a sublinear expected cumulative average regret, it is crucial to accurately estimate the unknown utility function u. While many existing works have relied on a linear structure (Bengs et al., 2022; Li et al., 2024; Saha, 2021; Di et al., 2023), i.e., $u\left(x_{t,k}\right) = \theta_*^\intercal x_{t,k}$, real-world utility functions are often inherently nonlinear. To approximate such nonlinear utility functions, we employ a fully connected neural network f with the ReLU activation function $\max\{x, 0\}$, as follows:

$$f\left(x;\theta,W\right) = \theta^\intercal \phi\left(x;W\right), \tag{2}$$
$$\phi\left(x;W\right) = \sqrt{m}\mathrm{ReLU}\left(W^L(\cdots \mathrm{ReLU}\left(W^1 x\right)\cdots)\right),$$

where $W = (\mathrm{vec}(W^1), \cdots, \mathrm{vec}(W^L))$ consists of learnable parameters as $W^1 \in \mathbb{R}^{m\times d}$, $W^l \in \mathbb{R}^{m\times m}$ for $2 \le l \le L-1$, and $W^L \in \mathbb{R}^{d\times m}$. The parameter $\theta \in \mathbb{R}^d$ is initialized as $\theta = \theta_0$, and the weight matrices are initialized as $W = W_0^i$ for $1 \le i \le L$, following the setup of prior neural bandit frameworks (Xu et al., 2020). In detail, we initialize them as

$$W_0^l = \begin{bmatrix} w & 0 \\ 0 & w \end{bmatrix}, W_0^L = \begin{pmatrix} \omega \\ -\omega \end{pmatrix}^\intercal, \theta_0 \sim \mathcal{N}\left(0, \frac{1}{d}\right), \tag{3}$$

for $1 \le l \le L-1$ by $w \sim \mathcal{N}\left(0, 4/m\right)$ and $\omega \sim \mathcal{N}\left(0, 2/m\right)$ where $\mathcal{N}$ denotes a Gaussian distribution. After receiving the binary signal o_t from the arm selection $(k_{t,1}, k_{t,2})$ in each round t, the neural network f is trained on historical data $\{x_{t,k_{i,1}}, x_{t,k_{i,2}}, o_i\}_{i=1}^t$. The parameters θ_t and W_t are updated by minimizing the regularized log-likelihood loss $\mathcal{L}_t$, defined as

$$\mathcal{L}_t\left(\theta, W\right) = -\sum_{i=1}^{t-1} \frac{\log\left(g\left((-1)^{1-o_i}\Delta f_i\right)\right)}{\zeta_i^2} + \frac{1}{2}\lambda\left\|\theta - \theta_0\right\|_2^2, \tag{4}$$

for some $\lambda > 0$, where $\Delta f_i = f\left(x_{i,k_{i,1}};\theta,W\right) - f\left(x_{i,k_{i,2}};\theta,W\right)$ and

$$\zeta_i = \begin{cases} \max\{\widehat{\sigma}_i, \epsilon\} & \text{variance-aware,} \\ 1 & \text{variance-agnostic,} \end{cases} \tag{5}$$

with the estimated variance $|\widehat{\sigma}_i|^2 = g(\Delta f_i)(1 - g(\Delta f_i))$ by (θ_t, W_t) and a positive constant $\epsilon > 0$.

Remark 1 (Role of Eqs. (4)–(5)). Eq. (4) is a weighted negative log-likelihood (Ueno et al., 2012; Dmochowski et al., 2010), the standard approach under heteroscedastic noise. The stability term ϵ in (5) ensures bounded weights for the concentration inequalities (Lemma 1) and prevents loss divergence when the estimated variance approaches zero. Setting $\zeta_i = 1$ reverts to a variance-agnostic approach with a looser $\widetilde{\mathcal{O}}(d\sqrt{T})$ bound (Corollary 1).

4.2 Arm Selection Methods.

We propose two types of methods: UCB-based (Abbasi-Yadkori et al., 2011) and TS-based (Zhang et al., 2021a) strategies. We assume access to a set of parameters (θ_{t-1}, W_{t-1}) learned from

the previous round $t-1$, as well as the Gram matrix V_{t-1}:

$$V_{t-1} = \sum_{i=1}^{t-1} \frac{\Delta\phi_{i,k_{i,1},k_{i,2}} \Delta\phi_{i,k_{i,1},k_{i,2}}^{\intercal}}{\zeta_i^2} + \lambda I, \quad (6)$$

where $\Delta\phi_{i,k,k'} := \Delta\phi_{i,k,k',i}$. Here, $\Delta\phi_{i,k,k',j} := \phi(x_{i,k}; W_{j-1}) - \phi(x_{i,k'}; W_{j-1})$ for any $k, k' \in [K]$ and $i, j \in [T]$. The value ζ_i in Eq. (6) is identical to that in Eq. (4). Based on the Gram matrix, we propose three arm selection strategies: one asymmetric arm selection and two symmetric arm selections, applicable to both UCB- and TS-based methods.

4.2.1 UCB-based Methods.

Asymmetric Arm Selection (UCB-ASYM). This strategy selects the first arm $k_{t,1}$ greedily based on (θ_{t-1}, W_{t-1}):

$$k_{t,1} = \arg\max_{k \in [K]} \theta_{t-1}^{\intercal} \phi(x_{t,k}; W_{t-1}). \quad (7)$$

The second arm $k_{t,2}$ is determined as the best competitor of $k_{t,1}$ with an exploration bonus, expressed as the upper confidence width $\left\|\Delta\phi_{t,k,k_{t,1}}\right\|_{V_{t-1}^{-1}}$ as $k_{t,2} = \arg\max_{k \in [K]} \theta_{t-1}^{\intercal} \phi(x_{t,k}; W_{t-1}) + a_t \left\|\Delta\phi_{t,k,k_{t,1}}\right\|_{V_{t-1}^{-1}}$, where $a_t > 0$ is a confidence width coefficient. The asymmetric arm selection strategy has been studied in (Bengs et al., 2022) and (Verma et al., 2024), in the contexts of linear and nonlinear utility functions, respectively.

Optimistic Symmetric Arm Selection (UCB-OSYM). In this strategy, the first and second arms, which maximize the sum of utilities including the exploration bonus, are selected as follows:

$$k_{t,1}, k_{t,2} = \arg\max_{k,k' \in [K] \times [K]} \theta_{t-1}^{\intercal} \phi_{t,k,k'} + a_t \left\|\Delta\phi_{t,k,k'}\right\|_{V_{t-1}^{-1}},$$

with the confidence width coefficient $a_t > 0$ and $\phi_{t,k,k'} = \phi_{t,k,k',t}$, where $\phi_{i,k,k',j} = \phi(x_{i,k}; W_{j-1}) + \phi(x_{i,k'}; W_{j-1})$ for $i, j \in [T]$ and $k, k' \in [K]$. This strategy has been studied in the linear setting, as in (Di et al., 2023).

Candidate-based Symmetric Arm Selection (UCB-CSYM). This strategy first constructs a confidence set $\mathcal{C}_t$ for each round t for strong candidates as potential optimal arms: for the confidence width coefficient $a_t > 0$,

$$\mathcal{C}_t = \left\{ k \mid a_t \left\|\Delta\phi_{t,k,k'}\right\|_{V_{t-1}^{-1}} > \theta_{t-1}^{\intercal} \Delta\phi_{t,k',k}, \forall k' \in [K] \right\}.$$

A tuple of arms with the maximal exploration bonus is then selected from the confidence set $\mathcal{C}_t$ as follows: $k_{t,1}, k_{t,2} = \arg\max_{k,k' \in \mathcal{C}_t \times \mathcal{C}_t} \left\|\Delta\phi_{t,k,k'}\right\|_{V_{t-1}^{-1}}$. This strategy type has been studied in the linear setting (Saha, 2021).

4.2.2 Thompson Sampling Methods.

Asymmetric Arm Selection (TS-ASYM). In this strategy, the first arm $k_{t,1}$ per round t is greedily selected as in Eq. (7). Then, we sample each arm's relative utility compared to the first arm from a Gaussian distribution with mean $\theta_{t-1}^{\intercal} \Delta\phi_{t,k,k_{t,1}}$ and variance $a_t^2 \left\|\Delta\phi_{t,k,k_{t,1}}\right\|_{V_{t-1}^{-1}}^2$, where $a_t > 0$ is the confidence width coefficient, i.e., $\widetilde{u}_{t,k_{t,1},k} \sim \mathcal{N}\left(\theta_{t-1}^{\intercal} \Delta\phi_{t,k,k_{t,1}}, a_t^2 \left\|\Delta\phi_{t,k,k_{t,1}}\right\|_{V_{t-1}^{-1}}^2\right)$. Then, the second arm $k_{t,2}$ in round t is selected as $k_{t,2} = \arg\max_{k \in [K]} \widetilde{u}_{t,k_{t,1},k}$. This strategy was studied in (Verma et al., 2024).

Optimistic Symmetric Arm Selection (TS-OSYM). This strategy constructs a Gaussian distribution, where the mean is defined as the sum of the estimated utilities of the pair of arms (k, k'), and the variance is determined by the exploration bonus for the pair. Then, a value $\widetilde{u}_{t,k,k'}$ is drawn from a Gaussian distribution as $\widetilde{u}_{t,k,k'} \sim \mathcal{N}\left(\theta_{t-1}^{\intercal} \phi_{t,k,k'}, a_t^2 \|\Delta\phi_{t,k,k'}\|_{V_{t-1}^{-1}}^2\right)$ for $k, k' \in [K]$ where $a_t > 0$ is the confidence width coefficient. Then, the arms $k_{t,1}$ and $k_{t,2}$ are finally selected as $k_{t,1}, k_{t,2} = \arg\max_{k,k' \in [K] \times [K]} \widetilde{u}_{t,k,k'}$.

Candidate-based Symmetric Arm Selection (TS-CSYM). In the same manner as the UCB-CSYM regime, a confidence set $\mathcal{C}_t$ is constructed with the confidence width coefficient $a_t > 0$. Then the pair $(k_{t,1}, k_{t,2})$ of arms is chosen from the set $\mathcal{C}_t$ as $(k_{t,1}, k_{t,2}) = \arg\max_{(i,j) \in C_t \times C_t} \widehat{\sigma}_{i,j}$, where

$$\widehat{\sigma}_{i,j} \sim \mathcal{N}\left(\|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}}^2, \frac{\|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}}^4}{4\log(Kt^2)} \right).$$

While TS-USYM aligns with the method in (Verma et al., 2024), our symmetric TS strategies (TS-OSYM and TS-CSYM) introduce a novel selection paradigm: instead of sequentially picking arms as in prior work (Verma et al., 2024; Wu and Liu, 2016), both arms are simultaneously drawn from the same posterior. This shift not only differentiates our approach conceptually but also complicates the analysis, requiring more elaborate event constructions and careful use of concentration tools such as Azuma–Hoeffding.

Shallow Exploration. Our approach constructs the Gram matrix V_{t-1} using only the last-layer gradients ($\theta \in \mathbb{R}^d$) instead of all parameters ($p = \widetilde{\mathcal{O}}(m^2 L)$), reducing Gram matrix inversion from $\widetilde{\mathcal{O}}(p^3)$ to $\widetilde{\mathcal{O}}(d^3)$

and yielding ∼28× speedup over full-gradient methods. While treating hidden layers as a fixed feature extractor introduces a representation bias ($U_{t-1}^{-1}M_t$), our iterative self-improvement technique controls $\|\theta_{t-1}\|_2$, ensuring this bias does not dominate the regret.

4.3 Neural Dueling Bandit Algorithms

Building on the proposed UCB- and TS-based arm selection strategies, we develop Algorithm 1, which employs a neural network as a nonlinear estimator of the unknown utility function. At each round t, the value ζ_t is defined exactly as in Eq. (5) and incorporated into the loss function as shown in Eq. (4). While our method primarily relies on the estimated Bernoulli variance $\widehat{\sigma}_t^2$, it can seamlessly accommodate the true variance σ_t^2 when available from an oracle.

Algorithm 1 Variance-aware Neural Dueling Bandits

1: Initialize the learnable parameters θ_0, W_0 of a neural network f defined in Eq. (3); the Gram matrix $V_0 = \lambda I$ for some $\lambda > 0$; the confidence interval coefficient $a_t > 0$.
2: Set the episode length $\mathcal{H}$ of f, the step size η, and the gradient steps n.
3: Determine one of the arm selection strategies in either UCB or Thompson sampling explained in Section 4.2.
4: **for** each round $t \in [T]$ **do**
5: Observe a context set $X_t = \{x_{t,1}, \cdots, x_{t,K}\} \subset \mathcal{X}$.
6: Choose $k_{t,1}, k_{t,2}$ with a_t, X_t, V_{t-1} via the determined arm selection.
7: Observe a preference (binary) signal o_t by dueling of $x_{t,k_{t,1}}$ and $x_{t,k_{t,2}}$.
8: **if** $t \equiv 0 \pmod{\mathcal{H}}$ **then**
9: Update (θ_t, W_t) by n gradient steps via the loss $L_t(\theta, W)$ in Eq. (4) with the step size η.
10: Update θ_t again such that θ_t minimizes Eq. (4) while keeping W_t fixed.
11: **end if**
12: Set ζ_t by Eq. (5) or by the oracle
13: Update V_t as

$$V_t = V_{t-1} + \frac{\Delta\phi_{t,k_{t,1},k_{t,2}} \Delta\phi_{t,k_{t,1},k_{t,2}}^{\intercal}}{\zeta_t^2}.$$

14: **end for**

5 REGRET ANALYSIS

In this section, we provide theoretical results for the upper bounds of cumulative average regret for each arm selection method. To facilitate the analysis of our algorithm, we assume the following structure for the context set $\mathcal{X}$, utility function u and the initial parameters (θ_0, W_0) initialized by Eq. (3).

Assumption 1. For all $x_{t,k}$, the context vector is bounded as $\|x_{t,k}\|_2 = 1$ and $[x_{t,k}]_j = [x_{t,k}]_{j+d/2}$ for simplicity. The link function g is continuously differentiable, takes values in $[-1, 1]$, and κ_g-Lipschitz, i.e., $|g(x) - g(x')| \leq \kappa_g |x - x'|$ for all $x, x' \in \mathbb{R}$. Moreover, since the utility function is bounded and the contexts are bounded, the input to the link function remains within a compact effective domain. Within this bounded domain, the derivative of g is strictly bounded away from zero: $0 < \beta_g \leq \dot{g}(x) \leq \alpha_g$, for some positive constants β_g and α_g. Finally, the utility function is bounded in absolute value by 1, i.e., $|u(x)| \leq 1$ for simplicity.

Assumption 2. There exists a constant $\kappa_\phi > 0$ such that $\left\| \frac{\partial\phi(x;W_0)}{\partial W} - \frac{\partial\phi(x';W_0)}{\partial W} \right\|_2 \leq \kappa_\phi \|x - x'\|_2$, for all $x, x' \in \{x_{t,k}\}_{t\in[T],k\in[K]}$. The vector θ_* in Eq. (1) is bounded in L_2 norm as $\|\theta_*\|_2 \leq J$ for a constant $J > 0$.

To analyze the cumulative average regret, we define a neural tangent kernel (NTK) $\mathbf{H}$ as follows.

Definition 1. $\mathbf{H} = \{\mathbf{H}_{i,j}\}_{i,j=1}^{TK} \in \mathbb{R}^{TK\times TK}$ is the neural tangent kernel (NTK) matrix such that $\mathbf{H}_{i,j} = \frac{1}{2}\left(\widetilde{\Sigma}^L(x_{t,k}, x_{t',k'})\right)$, for all $t, t' \in [T]$ and $k, k' \in [K]$, where for any $x, x' \in \mathbb{R}^d$,

$$\begin{aligned}
&\widetilde{\Sigma}^0(x, x') = \Sigma^0(x, x') = x^{\intercal}x', \\
&\Lambda^l(x, x') = \begin{bmatrix} \Sigma^{l-1}(x, x) & \Sigma^{l-1}(x, x') \\ \Sigma^{l-1}(x', x) & \Sigma^{l-1}(x', x') \end{bmatrix}, \\
&\Sigma^l(x, x') = 2\mathbb{E}_{(u,v)\sim\mathcal{N}(\mathbf{0},\Lambda^{l-1}(x,x'))}[\sigma(u)\sigma(v)], \\
&\widetilde{\Sigma}^l(x, x') = 2\widetilde{\Sigma}^{l-1}(x', x')\,\mathbb{E}_{u,v}[\dot{\sigma}(u)\dot{\sigma}(v)] + \Sigma^l(x, x').
\end{aligned}$$

Assumption 3. We assume that the minimum spectrum $\lambda_m(\mathbf{H})$ is strictly positive.

Note that the assumptions listed above are standard assumptions commonly found in the literature on bandits (Verma et al., 2024; Bengs et al., 2022; Xu et al., 2020; Bui et al., 2024). We introduce the following notations: $\Delta\phi_{i,j} = \Delta\phi_{i,k_{i,1},k_{i,2},j}$ for all $i, j \in [T]$ and $\mathbf{u} = (u(x_{1,1}), \cdots, u(x_{T,K}))$, $\widetilde{\mathbf{u}} = (f(x_{1,1}; \theta_0, W_0)), \cdots, f(x_{T,K}; \theta_{T-1}, W_{T-1}))$.

We introduce two auxiliary quantities: the matrix U_{t-1}, a weighted Gram matrix using the link function derivative, and the vector M_t, capturing the bias from shallow exploration. The term $U_{t-1}^{-1}M_t$ explicitly accounts for this representation learning bias.

Let (θ_{t-1}, W_{t-1}) be the parameters updated at round

$t-1$ under the loss in Eq. (4) and

$$U_{t-1} = \sum_{i=1}^{t-1} \dot{g}(\mathcal{Z}_{i,i}(\overline{\theta}_{i,t})) \frac{\Delta\phi_{i,i}\Delta\phi_{i,i}^\intercal}{\zeta_i^2} + \lambda' I,$$
$$M_t = \sum_{i=1}^{t-1} \left(g(\theta_{t-1}^\intercal \Delta\phi_{i,t}) - g(\mathcal{Z}_{i,t}(\theta_{t-1}))\right) \frac{\Delta\phi_{i,i}}{\zeta_i^2},$$

where $\lambda' = \beta_g\lambda$ and $\overline{\theta}_{i,t} = c\theta_{t-1} + (1-c)\theta_*$ for some $c \in [0,1]$ satisfying $g(\mathcal{Z}_{i,i}(\theta_{t-1})) - g(\mathcal{Z}_{i,i}(\theta_*)) = \dot{g}(\mathcal{Z}_{i,i}(\overline{\theta}_{i,t}))\Delta\phi_{i,i}^\intercal(\theta_{t-1}-\theta_*)$. Using the notations above, we prove a lemma that provides an upper bound on the confidence interval between θ_{t-1} and θ_*, accounting for an additional bias term.

Lemma 1. *Let m be a width of neural network as $m = poly(\mathcal{H}, L, \beta_g, \alpha_g, \kappa_g, \lambda, \log(TK/\delta), 1/\delta, \|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}})$ such that*

$$m = \widetilde{\Omega}(T^6), \tag{8}$$

where $\widetilde{\Omega}$ is the asymptotic lower bound ignoring logarithm factors and all other parameters. Then the following concentration inequality holds:

$$A_t := \left\|\theta_{t-1} - \theta_* + U_{t-1}^{-1}M_t\right\|_{V_{t-1}} \tag{9}$$
$$= \widetilde{\mathcal{O}}\left(\sqrt{d} + \epsilon^{-1} + \left(\frac{d}{\epsilon^4} + d^3\right)\frac{\sqrt{t}}{m^{1/6}}\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right)$$

ignoring all logarithmic and other terms.

Intuitively, A_t measures how far our parameter estimate θ_{t-1} deviates from the true parameter θ_*, adjusted for the representation bias. The last term $U_{t-1}^{-1}M_t$ of A_t arises from the bias during representation learning. Furthermore, ϵ^{-1} cannot be made arbitrarily small while still ensuring reasonable bounds. Our proof employs a multi-step refinement strategy centered on the auxiliary function $\mathcal{G}_t(\theta)$: $\mathcal{G}_t(\theta) = \sum_{i=1}^{t-1}\left[g(\mathcal{Z}_{i,i}(\theta)) - g(\mathcal{Z}_{i,i}(\theta_*))\right]\frac{\Delta\phi_{i,i}}{\zeta_i^2} + \lambda(\theta - \theta_0)$, where $\mathcal{Z}_{i,j}(\theta) = \theta^\intercal\Delta\phi_{i,j} + \theta_0^\intercal\nabla\Delta\phi_{i,1}(W_* - W_0)$.

The analysis proceeds iteratively. **First**, starting from the total loss $\mathcal{L}_t$ in Eq. (4), we establish an initial loose estimate $\|\theta_{t-1}\|_2 = \mathcal{O}(\sqrt{t})$. Substituting this bound into the analysis of $\|\mathcal{G}_t(\theta_{t-1}) + M_t\|_{V_{t-1}^{-1}}$ yields an intermediate bound of order $\widetilde{\mathcal{O}}(t/m^{1/6})$. Consequently, the resulting concentration inequality for A_t is limited by a term that scales with $\sqrt{t}$ and cannot be improved by simply increasing the network width m. **Second**, to address this limitation, we leverage the loose estimate together with a sufficiently large width m to obtain a sharper parameter bound of $\|\theta_{t-1}\|_2$ depending only on d and ϵ. **Finally**, we substitute this improved estimate back into the analysis of $\|\mathcal{G}_t(\theta_{t-1}) + M_t\|_{V_{t-1}^{-1}}$ to derive a significantly tighter bound of order $\widetilde{\mathcal{O}}(\sqrt{t}/m^{1/6})$ for the main concentration inequality in Eq. (9). A high-level overview of the proof structure is illustrated in Figure 1, and the complete technical details are provided in Section B of the supplementary material.

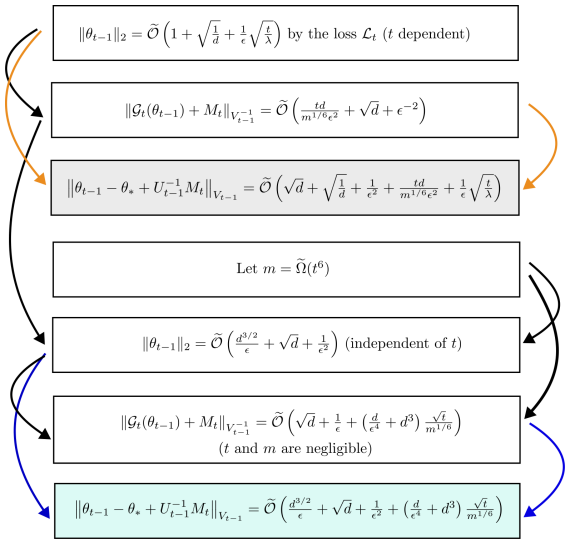


Figure 1: Illustration of the procedure in the proof of Lemma 1. A naive analysis (orange line) fails to tighten as m increases (as t increases). In contrast, our proposed analysis (blue line) establishes a bound by $\sqrt{t}/m^{1/6}$.

Next, we establish a bound for each arm selection strategy, e.g., in the case of UCB-ASYM:

$$\theta_{t-1}^\intercal(2\Delta\phi_{t,k_t^*,k_{t,1}} + \Delta\phi_{t,k_{t,1},k_{t,2}})$$
$$\leq a_t\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} - a_t\|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}} \tag{10}$$

where $a_t > 0$ is the confidence interval coefficient. To analyze the regret r_t^a for the UCB methods, we decompose the regret term and apply our main concentration inequality from Lemma 1. This inequality is used in conjunction with strategy-specific bounds like Eq. (10) and other theoretical tools, such as the elliptical potential lemma, to establish a final upper bound on the regret. The complete analysis is provided in Section C of the supplementary material.

As a result, we can obtain the following result for variance-aware UCB-based algorithms.

Theorem 1 (Variance-aware UCB methods)**.** *Let us assume Assumption 1-Assumption 3. Let the confidence width coefficient a_t for the arm selection methods and the step size η be chosen to satisfy the following:*

$$a_t = \frac{8}{\beta_g}\left(\sqrt{d\log(a_t')\log(4t^2/\delta)} + \frac{1}{\epsilon}\log(4t^2/\delta)\right)$$
$$+2\sqrt{\lambda}\left(J + 2\left(2 + \sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right)\right),$$
$$\eta \leq C\left(d^2mnT^6L^6\log(TK/\delta)\right)^{-1}$$

for some $C > 0$ with

$$a'_t = \left(1 + t\left(2\frac{\sqrt{d\log n\log(TK/\delta)}}{\sqrt{d\lambda}\epsilon}\right)^2\right).$$

If the width m of f is sufficiently large as in Lemma 1 and $\epsilon = \mathcal{O}(1/\sqrt{d})$, then the cumulative average regret R_T^a under any UCB-based arm selection method in Section 4.2 is upper-bounded as

$$R_T^a = \widetilde{\mathcal{O}}\left(\sqrt{dT} + d\sqrt{\sum_{i=1}^{T}\sigma_i^2} + \frac{Td^3}{m^{1/6}}\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right). \tag{11}$$

ignoring all other hyperparameters with probability at least $1-\delta$ over the randomness of the initialization of the neural network f.

The following result is about variance-agnostic UCB-based algorithms.

Corollary 1 (Variance-agnostic UCB methods)**.** *Assume the conditions in Theorem 1 hold, but we fix $\zeta_t = 1$ (i.e., $\epsilon = 1$) for all round $0 \le t \le T$. Then the cumulative regret R_t^a under any UCB-based arm selection method in Section 4.2 is bounded as*

$$R_T^a = \widetilde{\mathcal{O}}\left(d\sqrt{T} + \frac{Td^3}{m^{1/6}}\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right). \tag{12}$$

ignoring all other hyperparameters with probability at least $1-\delta$ over the randomness of learnable parameters in the neural network f

For our TS-based methods, we again leverage the concentration inequality from Lemma 1 to define good event sets. The construction for the TS-ASYM variant follows the established approach of (Verma et al., 2024). However, our other proposed TS methods introduce a new analytical challenge, as they involve the simultaneous stochastic selection of two arms. This feature, distinct from prior work like (Zhang et al., 2021a; Verma et al., 2024), requires a more intricate construction of the event sets and a more delicate application of the Azuma-Hoeffding inequality. Despite this added complexity, our analysis successfully yields regret bounds for the TS algorithms that are comparable to their UCB-based counterparts. A detailed derivation is provided in Section D of the supplementary material.

Theorem 2 (Variance-aware TS methods)**.** *Assume the conditions in Theorem 1 hold. Then the cumulative regret R_t^a under any TS-based arm selection method in Section 4.2 is bounded as Eq. (11) with probability at least $1-\delta$ over the randomness of learnable parameters in the neural network f.*

In the case of variance-agnostic TS algorithms, we have the following result.

Corollary 2 (Variance-agnostic TS methods)**.** *Assume the conditions in Corollary 1 hold. Then the cumulative regret R_t^a under any TS-based arm selection method in Section 4.2 is bounded as (12) with probability at least $1-\delta$ over the randomness of learnable parameters in the neural network f.*

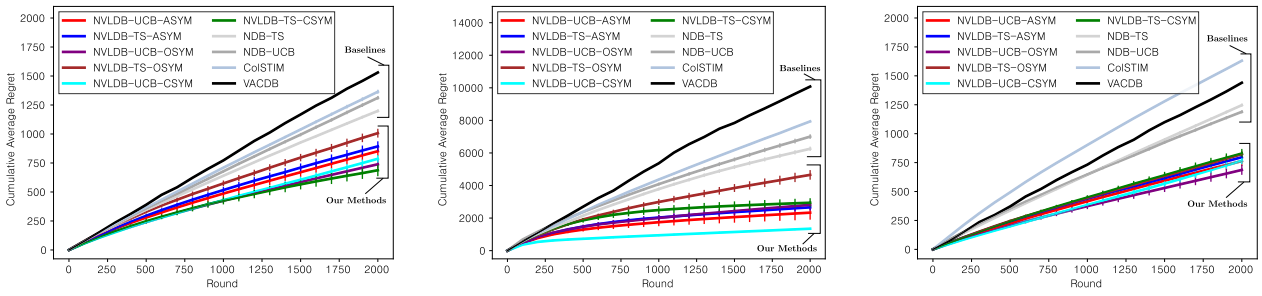
Theorem 1–Corollary 2 establish that our UCB and TS methods achieve the same order of cumulative regret, up to logarithmic factors. To specialize these results, we adopt the common assumption that $\|\widetilde{\mathbf{u}} - \mathbf{u}\|_{\mathbf{H}^{-1}} = \mathcal{O}(1)$, which is theoretically justified in (Xu et al., 2020; Bui et al., 2024). Under this assumption and with a network width m satisfying Eq. (8), we have two key settings: In the **variance-aware** setting, R_T^a is bounded by $\mathcal{O}\left(\sqrt{dT} + d\sqrt{\sum_{i=1}^{T}\sigma_i^2}\right)$, which matches the bound established by Bui et al. (2024). In the **variance-agnostic** setting, R_T^a is bounded by $\mathcal{O}(d\sqrt{T})$, aligning with the result derived by Xu et al. (2020).

A key distinction arises between the linear and neural bandit settings. In the linear case, where representation learning is unnecessary, variance-aware methods can achieve cumulative regret on the order of $\mathcal{O}(d)$ when variances are negligible (Zhou and Gu, 2022; Di et al., 2023). This is because the model parameters can be estimated without incurring additional representation errors. In contrast, since the neural setting fundamentally relies on representation learning, accumulated estimation errors are unavoidable. These errors appear as an additional bias term in the regret analysis, preventing the achievement of $\widetilde{\mathcal{O}}(d)$ regret. Consequently, while neural networks provide greater expressive power than linear models, they also introduce a fundamental challenge that does not arise in the linear setting.

If $\|\theta_{t-1}\|_2 \le C$ could be established without additional assumptions, the width requirement would reduce to $\widetilde{\Omega}(T^3)$, matching neural CMABs (Xu et al., 2020; Bui et al., 2024). Our experiments confirm robust performance with moderate widths ($m = 100$) and scalability with larger widths and higher dimensions (see Sections E.4 and E.5).

6 EXPERIMENTS

To empirically validate NVLDB, we consider the following synthetic nonlinear utility functions—cosine, square, and matrix—as shown in Figure 2, which have been widely studied in prior work (Verma et al., 2024; Zhang et al., 2021a; Zhou et al., 2020; Dai et al., 2022).



(a): $\cos(3\Theta^\intercal x)$ (b): $10(\Theta^\intercal x)^2$ (c): $x^\intercal \Theta\Theta^\intercal x$

Figure 2: Comparison of cumulative average regret under (a)-(c): synthetic tasks with a context dimension of $d = 5$ and a total of $K = 5$ arms. Each experiment is repeated across 20 random seeds.

We evaluate each arm-selection strategy introduced in Section 4.2 under NVLDB. We use a fully connected network with depth $L = 3$, width $m = 100$, learning rate 10^{-3}, regularization $\lambda = 1.0$, Adam optimizer, gradient steps $n = 20$, and episode length $\mathcal{H} = 1$. Full details are in the supplementary material.

To benchmark their performance, we compare them with the following baselines: VACDB (Di et al., 2023), ColSTIM (Bengs et al., 2022)—linear dueling bandits with and without variance-awareness, respectively—and Neural Dueling Bandits (NDB) (Verma et al., 2024) with both UCB and TS methods. We used the official implementations without modification.

Figure 2 shows that the proposed NVLDB variants–under both UCB and TS methods–consistently outperform the baseline algorithms in cumulative average regret across all three tasks (with context dimension $d = 5$ and $K = 5$) while exhibiting sublinear cumulative regret under nonlinear utility functions.

Scalability and Robustness. To verify scalability, we conducted additional experiments with higher dimensions ($d = 100$, $m = 100$, square utility) and wider networks ($m \in \{500, 1000\}$, $d = 5$, square utility) over $T = 10{,}000$ rounds. All NVLDB variants maintain sublinear regret (decreasing $R(T)/T$), confirming robustness well beyond the settings in prior work (Verma et al., 2024; Bui et al., 2024). For instance, with $d = 100$, NVLDB-UCB-ASYM achieves $R(10000)/10000 = 6.05 \pm 0.74$; with wider networks ($m = 500$), NVLDB-UCB-OSYM achieves 0.39 ± 0.04. Detailed results are provided in Section E.4 and Section E.5.

Variance-Aware vs. Variance-Agnostic. On the UCI Statlog dataset under a deterministic preference model ($P(i \succ j) = 1$ for $i > j$, negligible variance), NVLDB variants significantly outperform their variance-agnostic counterparts (NLDB), achieving near-zero cumulative average regret (e.g., NVLDB-UCB-OSYM: 0.017 vs. NLDB-UCB-OSYM: 0.199 at $T = 200$), confirming the theoretical advantage of variance-aware confidence sets.

Additional results including real-world tasks and variance-agnostic comparisons are in the supplementary material.

7 CONCLUSION

We introduce NVLDB, a family of neural variance-aware linear dueling bandit algorithms. Our approach employs a neural network to approximate nonlinear utility functions and estimates uncertainty by constructing a Gram matrix from the gradients with respect to the learnable parameters of the last layer. These gradients are weighted by the estimated variance of the Bernoulli distribution, computed via exponential-family link functions. We theoretically prove that both NVLDB and its variance-agnostic variant achieve sublinear cumulative average regret across various arm selection strategies under both the UCB and TS frameworks. To the best of our knowledge, this is the first work to introduce variance-aware neural dueling bandits with shallow exploration. Furthermore, we empirically validate our method against existing baselines and demonstrate its superior performance on widely used synthetic and real-world tasks.

Our key technical contribution is a bootstrap argument with spectral analysis that bounds $\|\hat{\theta}_{t-1}\|_2$ by a time-independent constant, reducing the width requirement from $\widetilde{\Omega}(T^{14})$ (Verma et al., 2024) to $\widetilde{\Omega}(T^6)$. We also introduce novel TS-based symmetric strategies (TS-OSYM, TS-CSYM). Closing the gap to $\widetilde{\Omega}(T^3)$ (Xu et al., 2020; Bui et al., 2024) remains future work. Finally, our inverse-variance weighting is equivalent to inverse Fisher Information under the exponential family, validated by strong empirical gains in low-variance regimes.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Shun-Ichi Amari and Scott C Douglas. Why natural gradient? In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98*, volume 2, pages 1213–1216. IEEE, 1998.

Yikun Ban, Yuchen Yan, Arindam Banerjee, and Jingrui He. Ee-net: Exploitation-exploration neural networks in contextual bandits. *arXiv preprint arXiv:2110.03177*, 2021.

Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. Stochastic contextual dueling bandits under linear stochastic transitivity models. In *International Conference on Machine Learning*, pages 1764–1786. PMLR, 2022.

Ha Manh Bui, Enrique Mallada, and Anqi Liu. Variance-aware linear ucb with deep representation for neural contextual bandits. *arXiv preprint arXiv:2411.05979*, 2024.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Kani Chen, Inchi Hu, and Zhiliang Ying. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27(4):1155–1163, 1999.

Zhongxiang Dai, Yao Shu, Arun , Flint Xiaofeng Fan, Bryan Kian Hsiang Low, and Patrick Jaillet. Federated neural bandits. *arXiv preprint arXiv:2205.14309*, 2022.

Rohan Deb, Yikun Ban, Shiliang Zuo, Jingrui He, and Arindam Banerjee. Contextual bandits with online neural regression. *arXiv preprint arXiv:2312.07145*, 2023.

Qiwei Di, Tao Jin, Yue Wu, Heyang Zhao, Farzad Farnoud, and Quanquan Gu. Variance-aware regret bounds for stochastic contextual dueling bandits. In *The Twelfth International Conference on Learning Representations*, 2023.

Jacek P Dmochowski, Paul Sajda, and Lucas C Parra. Maximum likelihood in cost-sensitive learning: Model specification, approximations, and upper bounds. *Journal of Machine Learning Research*, 11 (12), 2010.

Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pages 563–587. PMLR, 2015.

David Gilbarg, Neil S Trudinger, David Gilbarg, and NS Trudinger. *Elliptic partial differential equations of second order*, volume 224. Springer, 1977.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

David R Hunter. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32 (1):384–406, 2004.

Taehyun Hwang, Kyuwook Chai, and Min-hwan Oh. Combinatorial neural bandits. In *International Conference on Machine Learning*, pages 14203–14236. PMLR, 2023.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Prateek Jain and Nagarajan Natarajan. Regret bounds for non-decomposable metrics with missing labels. *arXiv preprint arXiv:1606.02077*, 2016.

Yeoneung Kim, Insoon Yang, and Kwang-Sung Jun. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. *Advances in Neural Information Processing Systems*, 35:1060–1072, 2022.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

Xuheng Li, Heyang Zhao, and Quanquan Gu. Feel-good thompson sampling for contextual dueling bandits. *arXiv preprint arXiv:2404.06013*, 2024.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2005.

Peter McCullagh. *Generalized linear models*. Routledge, 2019.

Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

Aadirupa Saha. Optimal algorithms for stochastic contextual preference bandits. *Advances in Neural Information Processing Systems*, 34:30050–30062, 2021.

Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *International Conference on Machine Learning*, pages 9235–9244. PMLR, 2021.

Louis L Thurstone. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge, 2017.

Tsuyoshi Ueno, Shin-ichi Maeda, Motoaki Kawanabe, and Shin Ishii. Weighted likelihood policy search with model selection. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Sharan Vaswani, Abbas Mehrabian, Audrey Durand, and Branislav Kveton. Old dog learns new tricks: Randomized ucb for bandit problems. *arXiv preprint arXiv:1910.04928*, 2019.

Arun Verma, Zhongxiang Dai, Xiaoqiang Lin, Patrick Jaillet, and Bryan Kian Hsiang Low. Neural dueling bandits, 2024. URL https://arxiv.org/abs/2407.17112.

Huasen Wu and Xin Liu. Double thompson sampling for dueling bandits. *Advances in neural information processing systems*, 29, 2016.

Pan Xu, Zheng Wen, Handong Zhao, and Quanquan Gu. Neural contextual bandits with deep representation and shallow exploration, 2020. URL https://arxiv.org/abs/2012.01780.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Tong Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural thompson sampling, 2021a. URL https://arxiv.org/abs/2010.00827.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems*, 34:4342–4355, 2021b.

Zihan Zhang, Xiangyang Ji, and Simon Du. Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory*, pages 3858–3904. PMLR, 2022.

Heyang Zhao, Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4977–5020. PMLR, 2023.

Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *Advances in neural information processing systems*, 35:36337–36349, 2022.

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

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021a.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021b.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material:

Neural Variance-aware Dueling Bandits with Deep Representation and Shallow Exploration

ORGANIZATION OF THE SUPPLEMENTARY MATERIAL

This document serves as a technical appendix to the main paper. It provides additional technical details and further experimental results omitted from the main text to facilitate a deeper understanding of the proposed method. The remainder of this document is organized as follows.

Section A. We present theoretical results on the width of neural networks and several useful inequalities, which are instrumental in establishing the sublinear cumulative regret bounds of our method, NVLDB.

Section B. We provide a theoretical analysis of the confidence intervals for neural dueling bandits with shallow exploration and deep representation. The resulting bounds play a central role in both the UCB and Thompson Sampling (TS) frameworks.

Section C. We prove sublinear cumulative regret bounds under the UCB-based arm selection strategies. First, we analyze each arm selection strategy within the UCB regime and derive key inequalities. We then combine these with the concentration bounds from Section B to establish our main results under the UCB framework.

Section D. Applying the concentration bounds from Section B, we prove that the TS framework achieves sublinear cumulative regret bounds. Specifically, we demonstrate that this sublinearity holds for each TS-based arm selection strategy.

Section E. We describe the experimental setup, including pseudocode, and provide additional experimental results that further validate the effectiveness of NVLDB.

Supplementary Zip File Contents The supplementary zip file contains the following items:

- **supplementary-material.pdf**: Includes proofs of the theoretical results, detailed descriptions of the experimental setup, and additional experimental results (e.g., on real-world datasets).
- **full-paper.pdf**: A consolidated version of the manuscript, including the main paper, references, checklist, and supplementary material.
- **code.zip**: Contains the implementation of our proposed method, NVLDB.

A PRELIMINARIES

This section provides a collection of mathematical notations and preliminaries that form the foundation of our theoretical analysis. For clarity and to ensure a self-contained exposition, we explicitly state all necessary definitions and results—some newly derived, others drawn from existing literature—even if they repeat material from the main manuscript. These components, though some may seem elementary, are instrumental in constructing our proofs and establishing the final cumulative average regret bounds.

Throughout this paper, we suppress constant factors and dependencies on all hyperparameters within the $\widetilde{\mathcal{O}}(\cdot)$ notation, explicitly indicating only the dependencies on the key parameters of interest, such as T, d, m, and $\frac{1}{\delta}$. Likewise, our asymptotic lower-bound notation, $\widetilde{\Omega}(\cdot)$, omits these dependencies when appropriate.

In the remaining sections, we will frequently use the notation $\phi(x; W)$ under weights W for $x \in \mathcal{X}$, as the deep neural network representation. Here, $\mathcal{X}$ is a context space. Then a neural network f can be expressed as

$$f(x; \theta, W) = \theta^{\intercal} \phi(x; W). \tag{A.1}$$

with the last layer's parameter $\theta \in \mathbb{R}^d$. We also define notations for ϕ already used in the main manuscript:

$$\begin{aligned}
\phi_{t,k,i} &= \phi\left(x_{t,k}; W_{i-1}\right) \\
\Delta\phi_{i,k,k',j} &= \phi\left(x_{i,k}; W_{j-1}\right) - \phi\left(x_{i,k'}; W_{j-1}\right), \\
\Delta\phi_{i,k,k'} := \Delta\phi_{i,k,k',i} &= \phi\left(x_{i,k}; W_{i-1}\right) - \phi\left(x_{i,k'}; W_{i-1}\right), \\
\Delta\phi_{i,j} := \Delta\phi_{i,k_{i,1},k_{i,2},j} &= \phi\left(x_{i,k_{i,1}}; W_{j-1}\right) - \phi\left(x_{i,k_{i,2}}; W_{j-1}\right), \\
\Delta\phi_{i} := \Delta\phi_{i,k_{i,1},k_{i,2},i} &= \phi\left(x_{i,k_{i,1}}; W_{i-1}\right) - \phi(x_{i,k_{i,2}}; W_{i-1}),
\end{aligned} \tag{A.2}$$

for $t, i, j \in [T] = \{1, \ldots, T\}$ and $k, k' \in [K]$, where K is the total number of arms, W_{i-1} and W_{j-1} denote the parameters obtained at rounds $i-1$ and $j-1$, respectively. Moreover, $(k_{i,1}, k_{i,2}) \in [K] \times [K]$ denotes the pair of arms selected by the learner at round i. We use ∇ to denote the gradient with respect to W, i.e., $\nabla = \nabla_W$. As we mentioned in the main manuscript, we denote W_0 and θ_0 as the initialization of W and θ, respectively. In detail, for $1 \le l \le L-1$, where L is the depth of f, we initialize the learnable parameters as

$$W_0^l = \begin{bmatrix} w & 0 \\ 0 & w \end{bmatrix}, \quad W_0^L = \begin{pmatrix} \omega^\intercal \\ -\omega^\intercal \end{pmatrix}^\intercal, \quad \theta_0 \sim \mathcal{N}\left(0, \frac{1}{d}\right), \tag{A.3}$$

where $w \sim \mathcal{N}(0, 4/m)$ and $\omega \sim \mathcal{N}(0, 2/m)$. We also redefine a utility function u assuming

$$u\left(x_{t,k}\right) = \theta_*^\intercal \phi_*\left(x_{t,k}\right), \tag{A.4}$$

where $\theta_* \in \mathbb{R}^D$ is the true parameter vector and $\phi_* : \mathcal{X} \to \mathbb{R}^{\,dim}$ is the true representation function, with d being the dimension of the representation space. Finally, a link function g belongs to an exponential family (Bengs et al., 2022), e.g., the Bradley–Terry model (Hunter, 2004; Luce, 2005): $g(x) = \frac{1}{1+\exp(-x)}$.

We assume that (θ_{t-1}, W_{t-1}) denotes the learned parameters from the previous round $t-1$ by the loss:

$$\mathcal{L}_t\left(\theta, W\right) = -\sum_{i=1}^{t-1} \frac{\log\left(g\left((-1)^{1-o_i}\,\Delta f_i\right)\right)}{\zeta_i^2} + \frac{1}{2}\lambda \left\|\theta - \theta_0\right\|_2^2, \tag{A.5}$$

for some $\lambda > 0$ and $\Delta f_i = f(x_{i,k_{i,1}}; \theta, W) - f(x_{i,k_{i,2}}; \theta, W)$, as well as the Gram matrix V_{t-1}:

$$V_{t-1} = \sum_{i=1}^{t-1} \frac{\Delta\phi_{i,k_{i,1},k_{i,2}} \Delta\phi_{i,k_{i,1},k_{i,2}}^\intercal}{\zeta_i^2} + \lambda I. \tag{A.6}$$

where

$$\zeta_i = \begin{cases} \max\{\widehat{\sigma}_i, \epsilon\} & \text{variance-aware,} \\ 1 & \text{variance-agnostic,} \end{cases} \tag{A.7}$$

We restate our assumptions stated in the main manuscript for self-contained supplementary material.

Assumption A.1. For all $x_{t,k}$, the context vector is bounded as $\|x_{t,k}\|_2 = 1$ and $[x_{t,k}]_j = [x_{t,k}]_{j+d/2}$ for simplicity. The link function g is continuously differentiable, takes values in $[-1, 1]$, and κ_g-Lipschitz, i.e., $|g(x) - g(x')| \le \kappa_g |x - x'|$ for all $x, x' \in \mathbb{R}$. Moreover, for all $x \in \mathcal{X}$, the derivative of g is bounded as $0 < \beta_g \le \dot{g}(x) \le \alpha_g$, for some positive constants β_g and α_g. Finally, the utility function is bounded in absolute value by 1, i.e., $|u(x)| \le 1$ for simplicity.

Assumption A.2. There exists a constant $\kappa_\phi > 0$ such that $\left\| \frac{\partial\phi(x;W_0)}{\partial W} - \frac{\partial\phi(x';W_0)}{\partial W} \right\|_2 \le \kappa_\phi \left\|x - x'\right\|_2$, for all $x, x' \in \{x_{t,k}\}_{t\in[T],k\in[K]}$. The vector θ_* in Eq. (A.4) is bounded in L_2 norm as $\|\theta_*\|_2 \le J$ for a constant $J > 0$.

To analyze the cumulative average regret, we define a neural tangent kernel (NTK) $\mathbf{H}$ as follows.

Definition A.1. $\mathbf{H} = \{\mathbf{H}_{i,j}\}_{i,j=1}^{TK} \in \mathbb{R}^{TK \times TK}$ is the neural tangent kernel (NTK) matrix such that $\mathbf{H}_{i,j} =$

$\frac{1}{2}\left(\widetilde{\Sigma}^L\left(x_{t,k}, x_{t',k'}\right)\right)$, for all $t, t' \in [T]$ and $k, k' \in [K]$, where for any $x, x' \in \mathbb{R}^d$,

$$\begin{aligned}
\widetilde{\Sigma}^0(x, x') &= \Sigma^0(x, x') = x^\intercal x', \\
\Lambda^l(x, x') &= \begin{bmatrix} \Sigma^{l-1}(x, x) & \Sigma^{l-1}(x, x') \\ \Sigma^{l-1}(x', x) & \Sigma^{l-1}(x', x') \end{bmatrix}, \\
\Sigma^l(x, x') &= 2\mathbb{E}_{(u,v)\sim\mathcal{N}(\mathbf{0}, \Lambda^{l-1}(x,x'))}\left[\sigma(u)\sigma(v)\right], \\
\widetilde{\Sigma}^l(x, x') &= 2\widetilde{\Sigma}^{l-1}(x', x')\,\mathbb{E}_{u,v}\left[\dot{\sigma}(u)\dot{\sigma}(v)\right] + \Sigma^l(x, x').
\end{aligned}$$

Assumption A.3. We assume that the minimum spectrum $\lambda_m(\mathbf{H})$ is strictly positive.

Due to Eq. (A.3), L_2-boundedness for θ_0 is obtained.

Lemma A.1. *Let $\theta_0 \sim \mathcal{N}(0, 1/d)$ as in (A.3). Then θ_0 is L_2-bounded as follows:*

$$\|\theta_0\|_2 \leq 2\left(2 + \sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right) \tag{A.8}$$

with probability at least $1 - \delta$.

Combining Assumption A.2 and (A.8), it is also obvious that

$$\|\theta_* - \theta_0\|_{V_{t-1}^{-1}} \leq \frac{1}{\sqrt{\lambda}}\left(J + 2\left(2 + \sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right)\right), \tag{A.9}$$

for V_{t-1} defined in Eq. (A.6), with probability at least $1 - \delta$.

Key neural network lemmas. The following three lemmas (from prior work) are the building blocks for our analysis: Lemma A.2 establishes that the utility function can be linearized around W_0 with a controlled residual (W_* exists close to W_0); Lemma A.3 bounds the parameter drift $\|W_t - W_0\|_2$ under gradient descent, ensuring the NTK regime holds; and Lemma A.4 quantifies the linearization error of the representation ϕ.

We have the following three lemmas for a sufficiently wide network as Eq. (A.1).

Lemma A.2 (Xu et al. (2020))**.** *Assume that Assumption A.3 holds. Let $\mathbf{H}$ be the neural tangent kernel (Jacot et al., 2018) and*

$$\begin{aligned}
\mathbf{u} &= \left(u(x_{1,1}), \cdots, u(x_{T,K})\right) \in \mathbb{R}^{TK} \\
\widetilde{\mathbf{u}} &= \left(f(x_{1,1}; \theta_0, W_0), \cdots, f(x_{T,K}; \theta_{T-1}, W_{T-1})\right) \in \mathbb{R}^{TK}.
\end{aligned}$$

Then there exists W_* such that

$$\|W_* - W_0\|_2 \leq \frac{1}{\sqrt{m}}\sqrt{(\mathbf{u} - \widetilde{\mathbf{u}})^T \mathbf{H}^{-1}(\mathbf{u} - \widetilde{\mathbf{u}})} = \frac{1}{\sqrt{m}}\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}},$$

and

$$u(x_{t,k}) = \theta_*^\intercal \phi(x_{t,k}; W_{t-1}) + \theta_0^\intercal\left(\nabla\phi_{t,k,1}\right)(W_* - W_0),$$

for all $k \in [K]$ and $t \in [T]$.

Lemma A.3 (Xu et al. (2020))**.** *Assume that Assumption A.1 holds. For any round index $t \in [T]$ in the q-th epoch, i.e., $t = (q-1)\mathcal{H} + i$ for some $i \in [\mathcal{H}]$ where $\mathcal{H}$ is the episode length. If the step size η satisfies*

$$\eta \leq \frac{C_0}{d^2 m n T^6 L^6 \log(TK/\delta)},$$

for some $C_0 > 0$ and the width m of the neural network satisfies

$$m \geq \max\{L\log(TK/\delta), dL^2\log(m/\delta), \delta^{-6}\mathcal{H}^{18}L^{16}\log^3(TK)\},$$

then, the following inequalities hold with probability at least $1-\delta$*:*

$$\begin{aligned}\|W_t - W_0\|_2 &\leq \frac{\delta^{3/2}}{m^{1/2}Tn^{9/2}L^6\log^3(m)},\\ \|\nabla\phi(x_{t,k};W_0)\|_F &\leq C_1\sqrt{dLm},\\ \|\phi(x;W_t)\|_2 &\leq \sqrt{d\log(n)\log(TK/\delta)}, \quad \forall x\in\mathcal{X},\end{aligned}$$

for all $t\in[T]$ *and* $k\in[K]$.

Remark 2. Lemma A.3 imposes a stricter requirement of $\eta=\widetilde{\mathcal{O}}(T^{-6})$, in contrast to the $\eta=\widetilde{\mathcal{O}}(T^{-5.5})$ stated in the original paper (Xu et al., 2020) (ignoring all other parameters). This necessity arises from the fact that the parameter estimate $\widehat{\theta}_{t-1}$ lacks a closed-form solution in the contextual dueling bandit setting. This is in contrast to the multi-armed bandit case, where a linear utility model typically admits such a closed form.

Lemma A.4 (Cao and Gu (2019))**.** *Let* $W, W'\in B(W_0,\omega)$ *for some radius* $\omega>0$*. Let the neural network* f *be given as in* (A.1) *and let the width* m *and the radius* ω *satisfy*

$$\begin{aligned}m &\geq C_0\max\{dL^2\log(m/\delta), \omega^{-4/3}L^{-8/3}\log(TK)\log(m/(\omega\delta))\},\\ \omega &\leq C_1L^{-5}(\log m)^{-3/2},\end{aligned}$$

for some $C_0>0, C_1>0$*. Then the following inequality holds:*

$$\left|\phi(x;W)-\widehat{\phi}(x;W)\right| \leq C_2\omega^{4/3}L^3d^{-1/2}\sqrt{m\log m},$$

for all $x\in\{x_{t,k}\}_{t\in[T],k\in[K]}$ *with probability at least* $1-\delta$*, where* $\widehat{\phi}(x;W)$ *is the linearization of* $\phi(x;W)$ *at* W' *as follows:*

$$\hat{\phi}(x;W)=\phi(x;W')+\langle\nabla\phi(x;W'),W-W'\rangle\,.$$

We now estimate the difference of $\phi(x;W)-\phi(x;W')$ for some W and W' using the results we introduced.

Lemma A.5. *Under the same conditions in Lemma A.4, there exist constants* $C_1, C_2>0$ *such that*

$$\begin{aligned}\|\phi(x;W_{t-1})-\phi(x;W_{i-1})\|_2 &\leq C_1\left(m^{-1/6}L^3d^{1/2}\sqrt{\log m}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}+\frac{\sqrt{d}\delta^{3/2}}{Tn^{9/2}L^{11/2}\log^3 m}\right),\\ \|\phi(x;W_{t-1})-\phi(x;W_*)\|_2 &\leq C_2\left(\left(m^{-1/6}L^3d^{1/2}\sqrt{\log m}+d^{1/2}L^{1/2}\right)\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right),\end{aligned}$$

with probability at least $1-\delta$.

Proof. Let $\widehat{\phi}(x;W)$ be the linearization of $\phi(x;W)$ at W' resulting from Lemma A.4 as

$$\widehat{\phi}(x;W)=\phi(x;W')+\langle\nabla_W\phi(x;W'),W-W'\rangle.$$

Since $\phi(x;W_0)\equiv 0$ for all $x\in\mathcal{X}$, we can compute $\phi(x;W_{t-1})-\phi(x;W_{i-1})$ as follows:

$$\begin{aligned}\phi(x;W_{t-1})-\phi(x;W_{i-1}) &= \phi(x;W_{t-1})-\phi(x;W_0)-(\phi(x;W_{i-1})-\phi(x;W_0))\\ &= \phi(x;W_{t-1})-\widehat{\phi}(x;W_{t-1})-(\phi(x;W_{i-1})-\widehat{\phi}(x;W_{i-1}))\\ &\quad+\langle\nabla\phi(x;W_0),W_{t-1}-W_0\rangle-\langle\nabla\phi(x;W_0),W_{i-1}-W_0\rangle.\end{aligned}\tag{A.10}$$

Therefore, using Eq. (A.10), we can show that there exist constants $C, C'>0$ such that

$$\begin{aligned}\|\phi(x;W_{t-1})-\phi(x;W_{i-1})\|_2 &\leq C\left(w^{4/3}L^3d^{1/2}\sqrt{m\log m}+\frac{\sqrt{d}\delta^{3/2}}{Tn^{9/2}L^{11/2}\log^3 m}\right)\\ &\leq C'\left(m^{-1/6}L^3d^{1/2}\sqrt{\log m}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}+\frac{\sqrt{d}\delta^{3/2}}{Tn^{9/2}L^{11/2}\log^3 m}\right),\end{aligned}$$

where the first inequality follows from applying Lemma A.3 and Lemma A.4 together with the Cauchy–Schwarz inequality. The second inequality is derived using the estimate $w = \widetilde{\mathcal{O}}\left(m^{-\frac{1}{2}} \|u - \widetilde{u}\|_{\mathbf{H}^{-1}}\right)$. Similarly, we can show that there exists $C'' > 0$ such that

$$\|\phi(x; W_{t-1}) - \phi(x; W_*)\|_2 \leq C'' \left(\left(m^{-1/6} L^3 d^{1/2} \sqrt{\log m} + d^{1/2} L^{1/2}\right) \|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right).$$

This completes the proof. ■

We also estimate the difference of $\phi(x; W) - \phi(x'; W)$.

Lemma A.6. *Under the same conditions in Lemma A.4, there exist constants $C_1, C_2 > 0$ such that*

$$\begin{aligned}
&\|\phi(x; W_{t-1}) - \phi(x'; W_{t-1})\|_2 \\
&\leq C_1 \left(\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} L^3 d^{1/2} \sqrt{\log m}/(m^{1/6}) + 2G\kappa_\phi \frac{\delta^{3/2}}{\sqrt{m} T n^{9/2} \log^3 m}\right) \qquad \text{(A.11)} \\
&\|\phi(x; W_*) - \phi(x'; W_*)\|_2 \\
&\leq C_2 \left(\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} L^3 d^{1/2} \sqrt{\log m}/(m^{1/6}) + 2G\kappa_\phi \frac{\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}}{\sqrt{m}}\right) \qquad \text{(A.12)}
\end{aligned}$$

Proof. Using Lemma A.4, $\phi(x; W_{t-1}) - \phi(x'; W_{t-1})$ can be decomposed into

$$\begin{aligned}
\phi(x; W_{t-1}) - \phi(x'; W_{t-1}) &= \phi(x; W_{t-1}) - \phi(x; W_0) - (\phi(x'; W_{t-1}) - \phi(x'; W_0)) \\
&= \phi(x; W_{t-1}) - \widehat{\phi}(x; W_{t-1}) - (\phi(x'; W_{i-1}) - \widehat{\phi}(x'; W_{i-1})) \\
&\quad + \langle \nabla(\phi(x; W_0) - \phi(x'; W_0)), W_{t-1} - W_0 \rangle.
\end{aligned}$$

due to $\phi(x; W_0) = 0$ for any $x \in \mathcal{X}$. Using Assumption A.2 and Lemma A.3, we have

$$\langle \nabla(\phi(x; W_0) - \phi(x'; W_0)), W_{t-1} - W_0 \rangle \leq 2G\kappa_\phi \frac{\delta^{3/2}}{\sqrt{m} T n^{9/2} \log^3 m}.$$

Furthermore, using Lemma A.4 to estimate the difference of $\phi(x; W_{t-1}) - \widehat{\phi}(x; W_{t-1}))$ and $\phi(x'; W_{t-1}) - \widehat{\phi}(x'; W_{t-1}))$, we can show that

$$\|\phi(x; W_{t-1}) - \phi(x'; W_{t-1})\|_2 \leq w^{4/3} L^3 d^{1/2} \sqrt{m \log m} + 2G\kappa_\phi \frac{\delta^{3/2}}{\sqrt{m} T n^{9/2} \log^3 m}. \qquad \text{(A.13)}$$

By substituting $w = \widetilde{\mathcal{O}}\left(\|\frac{\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}}{\sqrt{m}}\right)$ into Eq. (A.13), we can show Eq. (A.11). The case of W_*, i.e., Eq. (A.12), is shown in a similar manner. This completes the proof. ■

We also provide the following elliptic potential lemma.

Lemma A.7 (Lattimore and Szepesvári (2020))**.** *Let $\{y_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$ and $\lambda > 0$. Assume $\|y_t\|_2 \leq B$ for all t and $\lambda \geq \max\{1, B^2\}$ for some $B > 0$. Let $A_t = \lambda I + \sum_{s=1}^{t} y_s y_s^\intercal$. Then, the following inequalities hold:*

$$\begin{aligned}
\det(A_t) &\leq (\lambda + tB^2/d)^d, \\
\sum_{t=1}^{T} \|y_t\|_{A_t^{-1}}^2 &\leq 2 \log \frac{\det(A_T)}{\det(\lambda I)} \leq 2d \log\left(1 + \frac{TB^2}{\lambda d}\right).
\end{aligned}$$

We define the filtration that is used to apply the Bernstein-type inequality in martingale regime.

Definition A.2 (Filtration)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *filtration* is a family $\{\mathcal{F}_t\}_{t \in I}$ of sub-σ-algebras of $\mathcal{F}$ indexed by a totally ordered set $\mathcal{T}$ (often time), such that for all $s, t \in \mathcal{T}$ with $s \leq t$, we have

$$\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}.$$

The sub-σ-algebra $\mathcal{F}_t$ can be interpreted as the information available up to time t, and $\mathcal{F}_s \subseteq \mathcal{F}_t$ reflects the fact that one cannot lose information over time. Using the filtration, we can obtain the following Bernstein-type inequality.

Lemma A.8 (Zhou et al. (2021a)). *Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtration, and $\{y_t, r_t\}_{t\geq 1}$ a stochastic process such that $y_t \in \mathbb{R}^d$ is $\mathcal{F}_t$-measurable and $r_t \in \mathbb{R}$ is $\mathcal{F}_{t+1}$-measurable. Fix $R, G, \sigma, \lambda > 0$, and $\theta^* \in \mathbb{R}^d$. For $t \geq 1$, suppose that r_t, y_t satisfy*

$$\begin{aligned} |r_t| &\leq R, \\ \mathbb{E}[r_t \mid \mathcal{F}_t] &= 0, \\ \mathbb{E}[r_t^2 \mid \mathcal{F}_t] &\leq \sigma^2, \\ \|y_t\|_2 &\leq G. \end{aligned}$$

Then, the following inequality holds with probability at least $1-\delta$:

$$\left\|\sum_{i=1}^{t} y_i r_i\right\|_{A_t^{-1}} \leq 8\sigma\sqrt{d\log\left(1+\frac{tG^2}{d\lambda}\right)\log\left(\frac{4t^2}{\delta}\right)} + 4R\log\left(\frac{4t^2}{\delta}\right)$$

for any $t \geq 1$, where $A_t = \lambda I + \sum_{i=1}^{t-1} y_i y_i^\intercal$.

The following lemma provides an inequality for the product of L_2 boundedness for the product of a Gram matrix and extra term.

Lemma A.9 (Xu et al. (2020)). *Let $\{y_t\}_{t=1,2,\ldots}$ be a real-valued sequence such that $|y_t| \leq D$ for some constant $D > 0$. Let $A_t = \lambda I + \sum_{s=1}^{t} y_s y_s^\intercal$ such that $y_t \in \mathbb{R}^d$ and $\|y_t\|_2 \leq B$ for all $t \geq 1$ and some constants $\lambda, B > 0$. Then the following inequality holds:*

$$\left\|A_t^{-1}\sum_{s=1}^{t}\phi_s\zeta_s\right\|_2 \leq 2Dd,$$

for all $t \geq 1$.

B PROOF OF CONFIDENCE INTERVAL LEMMA

Let (θ_{t-1}, W_{t-1}) be the parameters updated at round $t-1$ under the loss in Eq. (A.5) and

$$\begin{aligned} U_{t-1} &= \sum_{i=1}^{t-1} \dot{g}(\mathcal{Z}_{i,i}(\overline{\theta}_{i,t}))\frac{\Delta\phi_{i,i}\Delta\phi_{i,i}^\intercal}{\zeta_i^2} + \lambda' I, \\ M_t &= \sum_{i=1}^{t-1}\left(g(\theta_{t-1}^\intercal\Delta\phi_{i,t}) - g(\mathcal{Z}_{i,t}(\theta_{t-1}))\right)\frac{\Delta\phi_{i,i}}{\zeta_i^2}, \end{aligned} \tag{B.1}$$

where $\lambda' = \beta_g\lambda$ and $\overline{\theta}_{i,t} = c\theta_{t-1} + (1-c)\theta_*$ for some $c \in [0,1]$ satisfying $g(\mathcal{Z}_{i,i}(\theta_{t-1})) - g(\mathcal{Z}_{i,i}(\theta_*)) = \dot{g}(\mathcal{Z}_{i,i}(\overline{\theta}_{i,t}))\Delta\phi_{i,i}^\intercal(\theta_{t-1}-\theta_*)$. Using the notations above, we restate a lemma that provides an upper bound on the confidence interval between θ_{t-1} and θ_*, accounting for an additional bias term.

Lemma B.1. *Let m be a width of neural network as $m = poly(\mathcal{H}, L, \beta_g, \alpha_g, \kappa_g, \lambda, \log(TK/\delta), 1/\delta, \|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}})$ such that*

$$m = \widetilde{\Omega}(T^6), \tag{B.2}$$

where $\widetilde{\Omega}$ is the asymptotic lower bound ignoring logarithm factors and all other parameters. Then the following concentration inequality holds:

$$\begin{aligned} A_t &:= \left\|\theta_{t-1} - \theta_* + U_{t-1}^{-1}M_t\right\|_{V_{t-1}} \\ &= \widetilde{\mathcal{O}}\left(\sqrt{d} + \epsilon^{-1} + \left(\frac{d}{\epsilon^4} + d^3\right)\frac{\sqrt{t}}{m^{1/6}}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right) \end{aligned} \tag{B.3}$$

ignoring all logarithmic and other terms.

The existence of $\overline{\theta}_{i,t}$ satisfying Eq. (B.1) in Lemma B.1 is true due to the mean-value theorem (Rudin et al., 1964). This lemma is identical to the lemma introduced in the main manuscript. To prove Lemma B.1, we estimate the left-hand side by showing a loose upper bound for $\|\theta_{t-1}\|_2$.

Proof roadmap (Bootstrap / Iterative Self-Improvement). The proof of Lemma B.1 proceeds in three stages:

1. **Loose bound (Lemma B.2):** Using only the loss function, we obtain $\|\theta_{t-1}\|_2 = \mathcal{O}(\sqrt{t/\lambda}/\epsilon)$. This bound grows with time and is insufficient for sublinear regret.
2. **First refinement (Lemma B.5):** Substituting the loose bound into the decomposition of $\|\mathcal{G}_t(\theta_{t-1}) + M_t\|_{V_{t-1}^{-1}}$ yields an intermediate bound of order $\widetilde{\mathcal{O}}(td/(\epsilon^2 m^{1/6}))$. For a sufficiently large width $m = \widetilde{\Omega}(t^6)$, this allows us to derive a *time-independent* bound $\|\theta_{t-1}\|_2 = \widetilde{\mathcal{O}}(d^{3/2} + \epsilon^{-2})$ (Eq. B.27).
3. **Final refinement (Lemma B.8):** Substituting the improved time-independent bound back yields the tight result $\|\mathcal{G}_t(\theta_{t-1}) + M_t\|_{V_{t-1}^{-1}} = \widetilde{\mathcal{O}}(\sqrt{t}/m^{1/6})$, which gives the final concentration inequality.

This iterative strategy is necessary because dueling bandits lack a closed-form estimator—the key distinction from standard neural bandits where $\hat{\theta}_{t-1} = V_{t-1}^{-1} b_{t-1}$ trivially gives a time-independent norm bound.

Stage 1: Loose bound on $\|\theta_{t-1}\|_2$. The following lemma provides an initial, time-dependent bound by comparing the loss at (θ_{t-1}, W_{t-1}) with the loss at the initialization (θ_0, W_0). Since the initialized representation $\phi(x; W_0) = 0$, the initial loss reduces to $\frac{t}{\epsilon^2}\log 2$, giving a simple upper bound on $\|\theta_{t-1} - \theta_0\|_2$.

Lemma B.2. *For the given $\theta_0 \sim \mathcal{N}(0, 1/d)$ and θ_{t-1} updated by (A.5), the following holds:*

$$\|\theta_{t-1}\|_2 \le 2\left(2 + \sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right) + \frac{1}{\epsilon}\sqrt{\frac{t}{\lambda}\log 4}. \tag{B.4}$$

Proof. First, note that for any x, we have $\phi\big(x; W_0\big) = 0$ by Eq. (A.3). Hence,

$$g\big(\theta_0^\intercal \Delta\phi_{t,1}\big) \;=\; g(0) \;=\; \frac{1}{2},$$

for a link function g. By substituting the above into (A.5), we obtain

$$\frac{\lambda}{2}\|\theta_{t-1} - \theta_0\|_2^2 \le \mathcal{L}_t\big(\theta_{t-1}, W_{t-1}\big) \le \mathcal{L}_t\big(\theta_0, W_0\big) \le \frac{t}{\epsilon^2}\log 2.$$

It follows that

$$\begin{aligned}\|\theta_{t-1}\|_2 &\le \|\theta_0\|_2 + \frac{1}{\epsilon}\sqrt{\frac{t}{\lambda}\log 4} \\ &\le 2\left(2 + \sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right) + \frac{1}{\epsilon}\sqrt{\frac{t}{\lambda}\log 4}.\end{aligned}$$

where the second inequality uses Lemma A.1. This completes the proof. ∎

Auxiliary functions. We now introduce the auxiliary function $\mathcal{Z}_{i,j}$ and $\mathcal{G}_t$ that decompose the confidence interval into manageable components. The function $\mathcal{Z}_{i,j}(\theta)$ linearizes the utility around the initial weights W_0, while $\mathcal{G}_t(\theta)$ aggregates the link function differences across all rounds—it plays the role of a "gradient-like" quantity whose norm controls the concentration inequality.

We define the function $\mathcal{Z}_{i,j} : \mathbb{R}^d \to \mathbb{R}$ for $i, j \in [T]$:

$$\mathcal{Z}_{i,j}\left(\theta\right) = \theta^\intercal \Delta\phi_{i,j} + \theta_0^\intercal \nabla\Delta\phi_{i,1}\big(W_* - W_0\big), \tag{B.5}$$

where $\Delta\phi_{i,1}$ and $\Delta\phi_{i,j}$ are given in Eq. (A.2). Using $\mathcal{Z}_{i,j}$, we additionally define the following auxiliary function $\mathcal{G}_t : \mathbb{R}^d \to \mathbb{R}^d$:

$$\mathcal{G}_t(\theta) = \sum_{i=1}^{t-1} g\big(\mathcal{Z}_{i,i}(\theta)\big) - g\big(\mathcal{Z}_{i,i}(\theta_*)\big)\frac{\Delta\phi_{i,i}}{\zeta_i^2} + \lambda(\theta - \theta_0).$$

We establish the relationship between this confidence interval and the function $\mathcal{G}_t$.

Key decomposition. The following lemma shows that the confidence interval $\|\theta_{t-1} - \theta_* + U_{t-1}^{-1}M_t\|_{V_{t-1}}$ is controlled by $\|\mathcal{G}_t(\theta_{t-1}) + M_t\|_{V_{t-1}^{-1}}$. The core idea is that $\mathcal{G}_t(\theta_{t-1}) - \mathcal{G}_t(\theta_*) + M_t = U_{t-1}(\theta_{t-1} - \theta_* + U_{t-1}^{-1}M_t)$ via the mean-value theorem, and $U_{t-1} \geq \beta_g V_{t-1}$, so the V_{t-1}-norm of the left side lower-bounds the V_{t-1}-norm of the right side.

Lemma B.3. *Let θ_* be the parameter defined in Eq. (A.4), and let (θ_{t-1}, W_{t-1}) be the parameters updated at round $t-1$ under the loss in Eq. (A.5). Then the following inequality holds:*

$$\begin{aligned}\left\|\theta_{t-1} - \theta_* + U_{t-1}^{-1}M_t\right\|_{V_{t-1}} &\leq \frac{1}{\beta_g}\left\|\mathcal{G}_t(\theta_{t-1}) + M_t - \lambda(\theta_* - \theta_0)\right\|_{V_{t-1}^{-1}} \\ &\leq \frac{1}{\beta_g}\left(\lambda\|\theta_* - \theta_0\|_{V_{t-1}^{-1}} + \|\mathcal{G}_t(\theta_{t-1}) + M_t\|_{V_{t-1}^{-1}}\right).\end{aligned}$$

Proof. By using the mean value theorem (Rudin et al., 1964), there exists $\overline{\theta}_{i,t} = c\theta_{t-1} + (1-c)\theta_*$ for some $c \in [0,1]$ such that

$$g\left(\mathcal{Z}_{i,i}(\theta_{t-1})\right) - g\left(\mathcal{Z}_{i,i}(\theta_*)\right) = \dot{g}\left(\mathcal{Z}_{i,i}\left(\overline{\theta}_{i,t}\right)\right)\Delta\phi_{i,i}^{\intercal}\left(\theta_{t-1} - \theta_*\right).$$

As a result, $\mathcal{G}_t(\theta_{t-1})$ can be expressed as

$$\mathcal{G}_t(\theta_{t-1}) = \sum_{i=1}^{t-1}\left[\dot{g}(\mathcal{Z}_{i,i}(\overline{\theta}_{i,t}))\right]\frac{\Delta\phi_{i,i}\Delta\phi_{i,i}^{\intercal}}{\zeta_i^2}(\theta_{t-1} - \theta_*) + \lambda'(\theta_{t-1} - \theta_0). \tag{B.6}$$

It is obvious that

$$\mathcal{G}_t(\theta_*) = \lambda'(\theta_* - \theta_0), \tag{B.7}$$

and

$$\frac{d\mathcal{Z}_{s,t}(\theta)}{d\theta} = \Delta\phi_{s,t},$$

for $s, t \in [T]$. Therefore, we can obtain

$$\begin{aligned}\mathcal{G}_t(\theta_{t-1}) - \mathcal{G}_t(\theta_*) + M_t &= \mathcal{G}_t(\theta_{t-1}) - \lambda'(\theta_* - \theta_0) + M_t \\ &= \sum_{i=1}^{t-1}\dot{g}(\mathcal{Z}_{i,i}(\overline{\theta}_{i,t}))\frac{\Delta\phi_{i,i}\Delta\phi_{i,i}^{\intercal}}{\zeta_i^2}(\theta_{t-1} - \theta_*) + \lambda'(\theta_{t-1} - \theta_*) + M_t \\ &= U_{t-1}\left(\theta_{t-1} - \theta_* + U_{t-1}^{-1}M_t\right).\end{aligned} \tag{B.8}$$

where the first and the second equalities result from Eq. (B.7) and Eq. (B.6), respectively. Since V_{t-1} is the Gram matrix in (A.6), we can compute

$$U_{t-1} \geq \beta_g V_{t-1} = \beta_g\left(\sum_{i=1}^{t-1}\frac{\Delta\phi_{i,i}\Delta\phi_{i,i}^{\intercal}}{\zeta_i^2} + \lambda I\right) \geq \lambda' I. \tag{B.9}$$

As a result, using Eqs. (B.8) and (B.9), we can show the following result:

$$\begin{aligned}&\|\mathcal{G}_t(\theta_{t-1}) - \lambda'(\theta_* - \theta_0) + M_t\|_{V_{t-1}^{-1}}^2 \\ &= \left(\theta_{t-1} - \theta_* + U_{t-1}^{-1}M_t\right)^{\intercal}U_{t-1}V_{t-1}^{-1}U_{t-1}\left(\theta_{t-1} - \theta_* + U_{t-1}^{-1}M_t\right) \\ &\geq \beta_g^2\left(\theta_{t-1} - \theta_* + U_{t-1}^{-1}M_t\right)^{\intercal}V_{t-1}\left(\theta_{t-1} - \theta_* + U_{t-1}^{-1}M_t\right) \\ &= \beta_g^2\|\theta_{t-1} - \theta_* + U_{t-1}^{-1}M_t\|_{V_{t-1}}^2.\end{aligned} \tag{B.10}$$

This completes the proof. ∎

Due to the mean-value theorem (Rudin et al., 1964), M_t can also be expressed as

$$M_t = -\sum_{i=1}^{t-1} \dot{g}(\widetilde{\mathcal{Z}}_{i,t})(\theta_0^\intercal \nabla \Delta\phi_{i,1}(W_* - W_0))\frac{\Delta\phi_{i,i}}{\zeta_i^2}, \tag{B.11}$$

where $\widetilde{\mathcal{Z}}_{i,t} = c\mathcal{Z}_{i,t}(\theta_{t-1}) + (1-c)\theta_{t-1}^\intercal \Delta\phi_{i,t}$ for some $0 \le c \le 1$.

Bias term $U_{t-1}^{-1}M_t$. The term M_t captures the representation learning bias—the discrepancy between using the current features $\phi_{i,i}$ and the linearized features $\phi_{i,t}$. The following lemma bounds $\|U_{t-1}^{-1}M_t\|_2$, showing it scales as $\widetilde{\mathcal{O}}(d^{3/2}\epsilon^{-1}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3})$ and vanishes for large m.

We estimate $U_{t-1}^{-1}M_t$ in the following lemma.

Lemma B.4. *With probability at least $1-\delta$, the following inequality holds:*

$$\begin{aligned}\|U_{t-1}^{-1}M_t\|_2 &\le 2d\kappa_\phi G\frac{\alpha_u}{\epsilon}\left(2+\sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right)\sqrt{dL}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}. \\ &= \widetilde{\mathcal{O}}(d^{3/2}\epsilon^{-1}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}).\end{aligned} \tag{B.12}$$

Proof. M_t in Eq. (B.11) can be expressed as

$$M_t = \sum_{i=1}^{t-1}\frac{d_{i,t}}{\zeta_i}\frac{\Delta\phi_{i,i}}{\zeta_i},$$

where

$$d_{i,t} = -\dot{g}(\widetilde{\mathcal{Z}}_{i,t})(\theta_0^\intercal \nabla\Delta\phi_{i,k_{i,1},k_{i,2},1}(W_* - W_0)).$$

Notice that

$$\beta_g V_{t-1} \le U_{t-1},$$

by Eq. (B.9). Thus,

$$\|U_{t-1}^{-1}M_t\|_2 \le \frac{1}{\beta_g}\|V_{t-1}^{-1}M_t\|_2.$$

We need to show that $d_{i,t}$ is bounded. Indeed, we can show that

$$\begin{aligned}|d_{i,t}| &= \left|\dot{g}(\widetilde{\mathcal{Z}}_{i,t})(\theta_0^\intercal\nabla\Delta\phi_{i,k_{i,1},k_{i,2},1}(W_*-W_0))\right| \\ &\le \alpha_g\|\theta_0\|_2\|\nabla\Delta\phi_{i,k_{i,1},k_{i,2},1}\|_F\|W_*-W_0\|_2 \\ &\le \alpha_g\left(2+\sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right)\|\nabla\Delta\phi_{i,k_{i,1},k_{i,2},1}\|_F\frac{\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}}{\sqrt{m}} \\ &\le 2\kappa_\phi G\alpha_u\left(2+\sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right)\sqrt{dL}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}.\end{aligned}$$

Here, the first inequality follows from the matrix inequality and the existence of an upper bound α_g for $\dot{g}$. The second inequality uses Lemma A.2 and Lemma A.1, and the third follows from Lemma A.3. Finally, applying Lemma A.9, we can complete the proof. ∎

Stage 2: Decomposition of $\mathcal{G}_t(\theta_{t-1}) + M_t$. We now decompose $\mathcal{G}_t(\theta_{t-1}) + M_t$ into six sub-terms $\mathcal{G}_{t,1,1}$ through $\mathcal{G}_{t,1,6}$, plus $\mathcal{G}_{t,2,1}$, $\mathcal{G}_{t,2,2}$, and $\mathcal{G}_{t,3}$. The key insight is that the first-order optimality condition $\frac{\partial\mathcal{L}_t}{\partial\theta}(\theta_{t-1}, W_{t-1}) = 0$ causes $\mathcal{G}_{t,1,1} + \mathcal{G}_{t,2,1} = 0$ (Eq. B.15) and $\mathcal{G}_{t,1,2} + M_t = 0$ (Eq. B.16), so these terms vanish exactly. The remaining terms arise from (i) the feature drift $\Delta\phi_{i,t} - \Delta\phi_{i,i}$ between different rounds, (ii) the link function evaluated

at different parameter values, and (iii) the martingale noise ε_i. Each is bounded using the neural network approximation guarantees and the elliptic potential lemma.

Now we estimate $\|\mathcal{G}_t(\theta_{t-1}) + M_t\|_{V_{t-1}^{-1}}$. By Lemma A.4 and Eq. (B.5), we can obtain

$$g\left(\mathcal{Z}_{i,i}\left(\theta_*\right)\right) = g\left(u\left(x_{i,k_{i,1}}\right) - u\left(x_{i,k_{i,2}}\right)\right) = o_i - \varepsilon_i, \tag{B.13}$$

where o_i is a binary signal and ε_i is the remaining term at round i. Using Eq. (B.13), we can decompose $\mathcal{G}_t(\theta_t) + M_t$ as follows:

$$\begin{aligned}\mathcal{G}_t(\theta_{t-1}) + M_t &= \sum_{i=1}^{t-1} u\left(\mathcal{Z}_{i,i}\left(\theta_{t-1}\right)\right) - u\left(\mathcal{Z}_{i,i}\left(\theta_*\right)\right) \frac{\Delta\phi_{i,i}}{\zeta_i^2} + \lambda'\left(\theta_{t-1} - \theta_0\right) + M_t \\ &= \mathcal{G}_{t,1} + \mathcal{G}_{t,2} + \mathcal{G}_{t,3} + M_t\end{aligned}$$

where

$$\begin{aligned}\mathcal{G}_{t,1} &= \sum_{i=1}^{t-1} u\left(\mathcal{Z}_{i,i}\left(\theta_{t-1}\right)\right) \frac{\Delta\phi_{i,i}}{\zeta_i^2} + \lambda'\left(\theta_{t-1} - \theta_0\right), \\ \mathcal{G}_{t,2} &= \sum_{i=1}^{t-1} \left(-o_i\right) \frac{\Delta\phi_{i,i}}{\zeta_i^2}, \\ \mathcal{G}_{t,3} &= \sum_{i=1}^{t-1} \frac{\varepsilon_i}{\zeta_i} \frac{\Delta\phi_{i,i}}{\zeta_i}.\end{aligned} \tag{B.14}$$

We can further decompose $\mathcal{G}_{t,1}, \mathcal{G}_{t,2}, \mathcal{G}_{t,3}$ into

$$\begin{aligned}\mathcal{G}_{t,1} &= \mathcal{G}_{t,1,1} + \mathcal{G}_{t,1,2} + \mathcal{G}_{t,1,3} + \mathcal{G}_{t,1,4} + \mathcal{G}_{t,1,5} + \mathcal{G}_{t,1,6}, \\ \mathcal{G}_{t,2} &= \sum_{i=1}^{t-1} \frac{-o_i}{\zeta_i^2} \Delta\phi_{i,t} + \sum_{i=1}^{t-1} \frac{o_i}{\zeta_i^2} \left(\Delta\phi_{i,t} - \Delta\phi_{i,i}\right) = \mathcal{G}_{t,2,1} + \mathcal{G}_{t,2,2}.\end{aligned}$$

where

$$\begin{aligned}\mathcal{G}_{t,1,1} &= \sum_{i=1}^{t-1} g(\theta_{t-1}^\intercal \Delta\phi_{i,t}) \frac{\Delta\phi_{i,t}}{\zeta_i^2} + \lambda\left(\theta_{t-1} - \theta_0\right), \\ \mathcal{G}_{t,1,2} &= \sum_{i=1}^{t-1} g(\mathcal{Z}_{i,t}(\theta_{t-1})) - g(\theta_{t-1}^\intercal \Delta\phi_{i,t}) \frac{\Delta\phi_{i,i}}{\zeta_i^2}, \\ \mathcal{G}_{t,1,3} &= \sum_{i=1}^{t-1} g(\mathcal{Z}_{i,t}(\theta_{t-1})) - g(\theta_{t-1}^\intercal \Delta\phi_{i,t}) \frac{\left(\Delta\phi_{i,t} - \Delta\phi_{i,i}\right)}{\zeta_i^2}, \\ \mathcal{G}_{t,1,4} &= -\sum_{i=1}^{t-1} g\left(\mathcal{Z}_{i,t}\left(\theta_{t-1}\right)\right) - g\left(\mathcal{Z}_{i,i}\left(\theta_{t-1}\right)\right) \frac{\Delta\phi_{i,i}}{\zeta_i^2}, \\ \mathcal{G}_{t,1,5} &= -\sum_{i=1}^{t-1} g\left(\mathcal{Z}_{i,t}\left(\theta_{t-1}\right)\right) - g\left(\mathcal{Z}_{i,i}\left(\theta_{t-1}\right)\right) \frac{\left(\Delta\phi_{i,t} - \Delta\phi_{i,i}\right)}{\zeta_i^2}, \\ \mathcal{G}_{t,1,6} &= -\sum_{i=1}^{t-1} g\left(\mathcal{Z}_{i,i}\left(\theta_{t-1}\right)\right) \frac{\left(\Delta\phi_{i,t} - \Delta\phi_{i,i}\right)}{\zeta_i^2}.\end{aligned}$$

Notice that θ_{t-1} and W_{t-1} are updated to satisfy

$$\frac{\partial \mathcal{L}_t}{\partial \theta}(\theta_{t-1}, W_{t-1}) = 0,$$

where $\mathcal{L}_t$ is defined in Eq. (A.5). Then we can obtain

$$\frac{\partial \mathcal{L}_t}{\partial \theta}(\theta_{t-1}, W_{t-1}) = \mathcal{G}_{t,1,1} + \mathcal{G}_{t,2,1} = 0 \tag{B.15}$$

Moreover, by the definition of M_t, it holds that

$$\mathcal{G}_{t,1,2} + M_t = 0. \tag{B.16}$$

Thanks to Eq. (B.15) and Eq. (B.16), we can obtain

$$\|\mathcal{G}_{t,1,1} + \mathcal{G}_{t,2,1}\|_{V_{t-1}^{-1}} = 0, \quad \|\mathcal{G}_{t,1,2} + M_t\|_{V_{t-1}^{-1}} = 0. \tag{B.17}$$

Now, we estimate the terms $\|\mathcal{G}_{t,1,3} + \mathcal{G}_{t,1,5} + \mathcal{G}_{t,1,6} + \mathcal{G}_{t,2,2}\|_{V_{t-1}^{-1}}$. We can rearrange $\mathcal{G}_{t,1,3} + \mathcal{G}_{t,1,5} + \mathcal{G}_{t,1,6} + \mathcal{G}_{t,2,2}$ as

$$\begin{aligned}
\mathcal{G}_{t,1,3} + \mathcal{G}_{t,1,5} + \mathcal{G}_{t,1,6} + \mathcal{G}_{t,2,2} &= \sum_{i=1}^{t-1} o_i \frac{(\Delta\phi_{i,t} - \Delta\phi_{i,i})}{\zeta_i^2} - \sum_{i=1}^{t-1} g(\theta_{t-1}^\intercal \Delta\phi_{i,t}) \frac{(\Delta\phi_{i,t} - \Delta\phi_{i,i})}{\zeta_i^2} \\
&= \sum_{i=1}^{t-1} \left(o_i - g(\theta_{t-1}^\intercal \Delta\phi_{i,t})\right) \frac{(\Delta\phi_{i,t} - \Delta\phi_{i,i})}{\zeta_i^2}.
\end{aligned}$$

It is clear that $0 < g(x) < 1$, and $0 \le o_i \le 1$. Therefore, we can obtain

$$\|\mathcal{G}_{t,1,3} + \mathcal{G}_{t,1,5} + \mathcal{G}_{t,1,6} + \mathcal{G}_{t,2,2}\|_2 \le \frac{1}{\epsilon^2} \sum_{i=1}^{t-1} \|\Delta\phi_{i,t} - \Delta\phi_{i,i}\|_2,$$

due to the definition of ζ_i (in Eq. (A.7)). Thus, applying Lemma A.5 to $\|\Delta\phi_{i,t} - \Delta\phi_{i,i}\|_2$, we can show the following inequality:

$$\begin{aligned}
&\|\mathcal{G}_{t,1,3} + \mathcal{G}_{t,1,5} + \mathcal{G}_{t,1,6} + \mathcal{G}_{t,2,2}\|_{V_{t-1}^{-1}} \le \sqrt{\frac{1}{\lambda}} \left(\|\mathcal{G}_{t,1,3} + \mathcal{G}_{t,1,5} + \mathcal{G}_{t,1,6} + \mathcal{G}_{t,2,2}\|_2\right) \\
&\le \frac{C}{\epsilon^2} \sqrt{\frac{1}{\lambda}} t \left(\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} L^3 d^{1/2} \sqrt{\log m}/(m^{1/6}) + 2G\kappa_\phi \frac{\delta^{3/2}}{\sqrt{m} T n^{9/2} \log^3 m} \right) \\
&= \widetilde{\mathcal{O}} \left(\frac{d^{1/2} t \|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}}{m^{1/6} \epsilon^2} \right)
\end{aligned} \tag{B.18}$$

for some constant $C > 0$.

Now we estimate $\|\mathcal{G}_{t,1,4}\|_{V_{t-1}^{-1}}$. By the mean-value theorem (Rudin et al., 1964), the following holds:

$$\begin{aligned}
|g(\mathcal{Z}_{i,t}(\theta_{t-1})) - g(\mathcal{Z}_{i,i}(\theta_{t-1}))| &= |\dot{g}(\overline{\mathcal{Z}}_{i,t})(\mathcal{Z}_{i,t}(\theta_{t-1}) - \mathcal{Z}_{i,i}(\theta_{t-1}))| \\
&\le \alpha_g |\mathcal{Z}_{i,t}(\theta_{t-1}) - \mathcal{Z}_{i,i}(\theta_{t-1})|,
\end{aligned}$$

where $\overline{\mathcal{Z}}_{i,t} = c\mathcal{Z}_{i,t}(\theta_{t-1}) + (1-c)\mathcal{Z}_{i,i}(\theta_{t-1})$ for some $c \in [0,1]$. Further, we compute that there exists a constant $C > 0$ such that

$$\begin{aligned}
&|\mathcal{Z}_{i,t}(\theta_{t-1}) - \mathcal{Z}_{i,i}(\theta_{t-1})| = \left|\theta_{t-1}^\intercal \Delta\phi_{i,i} - \theta_{t-1}^\intercal \Delta\phi_{i,t}\right| \\
&\le 2C\|\theta_{t-1}\|_2 \left(\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} L^3 d^{1/2} \sqrt{\log m}/(m^{1/6}) + 2G\kappa_\phi \frac{\delta^{3/2}}{\sqrt{m} T n^{9/2} \log^3 m} \right) \\
&\le 2C \left(\sqrt{\frac{2t}{\lambda}} + 2\left(2 + \sqrt{d^{-1} \log\left(\frac{1}{\delta}\right)}\right) \right) \\
&\quad \times \left(\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} L^3 d^{1/2} \sqrt{\log m}/(m^{1/6}) + 2G\kappa_\phi \frac{\delta^{3/2}}{\sqrt{m} T n^{9/2} \log^3 m} \right).
\end{aligned} \tag{B.19}$$

with probability at least $1-\delta$, where the first equality follows from the definition of $\mathcal{Z}_{i,t}$ in Eq. (B.5), the second inequality uses the Cauchy–Schwarz inequality and Lemma A.5, and the third inequality results from Lemma B.2. Therefore, there exists a constant $C > 0$ such that

$$\begin{aligned}\|\mathcal{G}_{t,1,4}\|_{V_{t-1}^{-1}} &\leq \frac{\alpha_g}{\epsilon}\sqrt{\frac{1}{\lambda}}\sum_{i=1}^{t-1}|\mathcal{Z}_{i,t}(\theta_{t-1}) - \mathcal{Z}_{i,i}(\theta_{t-1})|\|\Delta\phi_{i,i}/\zeta_i\|_{V_{t-1}^{-1}} \\ &\leq C\sqrt{td}\frac{\alpha_g}{\epsilon}\sqrt{\frac{1}{\lambda}}\left(\sqrt{\frac{2t}{\lambda}} + 2\left(2 + \sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right)\right)\sqrt{\log n \log(TK/\delta)} \\ &\quad \times \left(\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} L^3 d^{1/2}\sqrt{\log m}/(m^{1/6}) + 2G\kappa_\phi \frac{\delta^{3/2}}{\sqrt{m}Tn^{9/2}\log^3 m}\right) \\ &= \widetilde{\mathcal{O}}\left(\frac{dt}{m^{-1/6}\epsilon}\left(\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right)\right)\end{aligned} \tag{B.20}$$

where the second inequality is driven by Eq. (B.19) and the elliptic potential lemma in Lemma A.7 to estimate the summation of $\|\Delta\phi_{i,i}/\zeta_i\|_{V_{t-1}^{-1}}$.

Martingale noise term $\mathcal{G}_{t,3}$. This term captures the stochastic noise from the Bernoulli preference feedback. Since ε_i/ζ_i forms a bounded martingale difference sequence, we apply a Bernstein-type inequality to obtain a $\widetilde{\mathcal{O}}(\sqrt{d} + \epsilon^{-1})$ bound—this is the variance-aware component that exploits the inverse-variance weighting.

Finally, we estimate $\mathcal{G}_{t,3}$. Since $|\varepsilon_i/\zeta_i| \leq \frac{1}{\epsilon}$, $\mathbb{E}\left[\varepsilon_i/\zeta_i \mid \mathcal{F}_t\right] = 0$, and $\mathbb{E}\left[(\varepsilon_i/\zeta_i)^2 \mid \mathcal{F}_t\right] \leq 1$, we can obtain the following bound by Lemma A.8:

$$\begin{aligned}\|\mathcal{G}_{t,3}\|_{V_{t-1}^{-1}} &\leq \frac{4}{\epsilon}\log(4t^2/\delta) \\ &\quad + 4\sqrt{d\log\left(1 + t\left(2\frac{\sqrt{d\log n\log(TK/\delta)}}{\epsilon}\right)^2/(d\lambda)\right)\log(4t^2/\delta)} \\ &= \widetilde{\mathcal{O}}(\sqrt{d} + \epsilon^{-1})\end{aligned} \tag{B.21}$$

by hiding the dependence on all other hyperparameters within the $\widetilde{\mathcal{O}}$ notation by Lemma A.7. Let us define

$$\mathcal{J}_1 = \|\mathcal{G}_{t,3}\|_{V_{t-1}^{-1}} \tag{B.22}$$

$$\mathcal{K}_1(t) = \|\mathcal{G}_{t,1,1} + \mathcal{G}_{t,2,1} + \mathcal{G}_{t,1,1} + M_t + \mathcal{G}_{t,1,3} + \mathcal{G}_{t,1,5} + \mathcal{G}_{t,1,6} + \mathcal{G}_{t,2,2} + \mathcal{G}_{t,1,4}\|_{V_{t-1}^{-1}}. \tag{B.23}$$

Combining the inequalities in Eqs. (B.17), (B.18), (B.20) and (B.21), we can derive the following lemma.

Lemma B.5. *Let the conditions in Lemma A.2, Lemma A.3, and Lemma A.4 be satisfied. Under Eq. (B.4) for θ_{t-1}, we can obtain*

$$\|\mathcal{G}_t(\theta_{t-1}) + M_t\|_{V_t^{-1}} \leq \mathcal{J}_1 + \mathcal{K}_1(t),$$

where

$$\begin{aligned}
\mathcal{J}_1 &\leq \frac{4}{\epsilon}\log(4t^2/\delta) \\
&\quad +4\sqrt{d\log\left(1+t\left(2\frac{\sqrt{d\log n\log(TK/\delta)}}{\epsilon}\right)^2/(d\lambda)\right)\log(4t^2/\delta)} \\
&= \widetilde{\mathcal{O}}\left(\sqrt{d}+\epsilon^{-1}\right), \\
\mathcal{K}_1(t) &\leq \frac{4}{\epsilon^2}\sqrt{\frac{1}{\lambda}}t\left(\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}L^3d^{1/2}\sqrt{\log m}/(m^{1/6})+2G\kappa_\phi\frac{\delta^{3/2}}{\sqrt{m}Tn^{9/2}\log^3 m}\right) \\
&\quad +C\sqrt{t}\frac{\alpha_g}{\epsilon^2}\sqrt{\frac{1}{\lambda}}\left(\sqrt{\frac{2t}{\lambda}}+2\left(2+\sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right)\right)\sqrt{d\log n\log(TK/\delta)} \\
&\quad \times\left(\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}L^3d^{1/2}\sqrt{\log m}/(m^{1/6})+2G\kappa_\phi\frac{\delta^{3/2}}{\sqrt{m}Tn^{9/2}\log^3 m}\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{td}{m^{1/6}\epsilon^2}\left(\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right)\right)
\end{aligned}$$

by hiding the dependence on all other hyperparameters within the $\widetilde{\mathcal{O}}$ notation for some constant $C>0$ with probability at least $1-\delta$.

To conclude, we can obtain

$$\|\mathcal{G}_t(\theta_{t-1})+M_t\|_{V_t^{-1}} \leq \widetilde{\mathcal{O}}\left(\frac{t}{m^{1/6}\epsilon^2}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}+\sqrt{d}+\epsilon^{-2}\right). \tag{B.24}$$

Width condition $m=\widetilde{\Omega}(t^6)$. We now choose m large enough so that the t-dependent term $\frac{t}{m^{1/6}\epsilon^2}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}$ becomes negligible compared to $\sqrt{d}+\epsilon^{-2}$. This requires $m^{1/6}\gtrsim t\cdot(\mathrm{poly}(d,\epsilon,\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}))$, i.e., $m=\widetilde{\Omega}(t^6)$.

Now we determine the neural network width m such that Eq. (B.24) can be expressed as

$$\|\mathcal{G}_t(\theta_{t-1})+M_t\|_{V_t^{-1}} \leq \widetilde{\mathcal{O}}\left(\sqrt{d}+\epsilon^{-2}\right). \tag{B.25}$$

Eq. (B.25) is indeed true if $m\geq Ct^6$ for some $C=C(\lambda,\alpha_g,L,1/\epsilon,\sqrt{d},\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}})$, i.e.,

$$m=\widetilde{\Omega}(t^6), \tag{B.26}$$

ignoring all other parameters. Importantly, the required network width m in our analysis scales more favorably with respect to t compared to the condition in Eq. (8) of Verma et al. (2024). While their regret bound holds only under a strong overparameterization assumption where m must grow at least on the order of t^{14}, our result remains valid under a substantially milder requirement, with m only needing to grow on the order of t^6, thereby highlighting the theoretical advantage of our approach.

Because $\lambda_m(V_{t-1})\geq\lambda$, it is obvious that

$$\begin{aligned}
&\lambda^{1/2}\|\theta_{t-1}-\theta_*+U_{t-1}^{-1}M_t\|_2 \leq \|\theta_{t-1}-\theta_*+U_{t-1}^{-1}M_t\|_{V_{t-1}} \\
&\leq \frac{1}{\beta_g}\left(\lambda\|\theta_*-\theta_0\|_{V_{t-1}^{-1}}+\|\mathcal{G}_t(\theta_{t-1})+M_t\|_{V_{t-1}^{-1}}\right) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{d}+\epsilon^{-2}\right)
\end{aligned}$$

if Eq. (B.26) is satisfied. The second inequality results from Lemma B.3. Then we can obtain

$$\begin{aligned}
\|\theta_{t-1}\|_2 &\leq \|\theta_*\|_2+\|U_{t-1}^{-1}M_t\|_2 \\
&\quad +\frac{1}{\beta_g}\lambda^{-1/2}\left(J+2\left(2+\sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right)\right)+\widetilde{\mathcal{O}}\left(\sqrt{d}+\epsilon^{-2}\right) \\
&\leq \widetilde{\mathcal{O}}(d^{3/2}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}+\sqrt{d}+\epsilon^{-2})
\end{aligned} \tag{B.27}$$

if Eq. (B.26) is satisfied, where the last inequality results from Lemma B.4 and the L_2-boundedness of θ_*. Eq. (B.27) denotes that for a sufficiently large width m, θ_{t-1} can be bounded by L_2 norm.

Stage 3: Spectral perturbation analysis for the final refinement. Having obtained the time-independent bound $\|\theta_{t-1}\|_2 = \widetilde{\mathcal{O}}(d^{3/2}+\epsilon^{-2})$, we now refine the bounds on the remaining terms. The key tools are the Courant–Fischer theorem (Lemma B.6) and a Gram matrix perturbation bound (Lemma B.7). These allow us to relate the V_{t-1}^{-1}-norm to a perturbed Gram matrix $\widetilde{U}_{t-1}^{-1}$-norm, absorbing the feature drift into a $(1+\widetilde{\mathcal{O}}(tm^{-1/3}/\lambda))$ multiplicative factor. This spectral technique is what ultimately improves the bound from $\widetilde{\mathcal{O}}(t/m^{1/6})$ to $\widetilde{\mathcal{O}}(\sqrt{t}/m^{1/6})$.

For a more refined estimation, we provide the following results based on spectral perturbation analysis.

Lemma B.6. *Let $V, W \in \mathbb{R}^{d\times d}$ be symmetric positive definite (SPD) matrices. Then for all $x \in \mathbb{R}^d$, the following inequality holds:*

$$\lambda_m(W^{-1}V)\|x\|_W^2 \le \|x\|_V^2 \le \lambda_M(W^{-1}V)\|x\|_W^2.$$

Proof. Since V and W are SPD, the matrix $W^{-1}V$ is also SPD and hence diagonalizable with positive real eigenvalues. Consider the generalized Rayleigh quotient:

$$R(x) := \frac{x^\intercal Vx}{x^\intercal Wx} = \frac{\|x\|_V^2}{\|x\|_W^2}.$$

By the Courant–Fischer Minimax theorem (Horn and Johnson, 2012), the following inequality holds for all nonzero $x \in \mathbb{R}^d$:

$$\lambda_m(W^{-1}V) \le R(x) \le \lambda_M(W^{-1}V).$$

Multiplying both sides by $\|x\|_W^2 = x^\intercal Wx$ gives:

$$\lambda_m(W^{-1}V)\cdot\|x\|_W^2 \le \|x\|_V^2 \le \lambda_M(W^{-1}V)\cdot\|x\|_W^2.$$

Taking square roots (since all terms are positive), we obtain:

$$\sqrt{\lambda_m(W^{-1}V)}\cdot\|x\|_W \le \|x\|_V \le \sqrt{\lambda_M(W^{-1}V)}\cdot\|x\|_W.$$

■

Lemma B.7. *Let $y_1, \dots, y_t \in \mathbb{R}^d$ be vectors satisfying $\|y_i\|_2 \le B$. Define $z_i := y_t - y_i$, and assume that $\|z_i\|_2 \le \widetilde{\mathcal{O}}(m^{-1/6})$ for all i. Let the regularized Gram matrices be*

$$V := \lambda I + \sum_{i=1}^n y_i y_i^\intercal, \qquad W := \lambda I + \sum_{i=1}^n z_i z_i^\intercal,$$

with $\lambda > 0$. Then the following bound holds:

$$\|WV^{-1}\| = \sup_{x\neq 0}\frac{x^\intercal Wx}{x^\intercal Vx} \le 1 + \widetilde{\mathcal{O}}\left(\frac{tm^{-1/3}}{\lambda}\right).$$

Proof. Since $\|z_i\|_2 \le Cm^{-1/6}$, we have

$$\|z_i z_i^\intercal\| = \|z_i\|_2^2 \le C^2 m^{-1/3}.$$

Therefore,

$$\left\|\sum_{i=1}^t x_i x_i^\intercal\right\| \le \sum_{i=1}^t \|x_i x_i^\intercal\| \le tC^2 m^{-1/3}.$$

Let $A := \sum_{i=1}^t y_i y_i^\intercal$ and $B := \sum_{i=1}^t z_i z_i^\intercal$. Then we write

$$WV^{-1} = (\lambda I + B)^{-1}(\lambda I + A).$$

We consider the Rayleigh quotient:

$$\lambda_M(WV^{-1}) = \max_{\|x\|=1} \frac{x^\intercal(\lambda I + B)z}{x^\intercal(\lambda I + A)x}.$$

Now, for all unit vectors $x \in \mathbb{R}^d$,

$$x^\intercal(\lambda I + B)x \le \lambda + \|B\| \le \lambda + tC^2 m^{-1/3},$$

for some constant $C > 0$ and

$$x^\intercal(\lambda I + A)x \ge \lambda.$$

Therefore,

$$\frac{x^\intercal(\lambda I + B)x}{x^\intercal(\lambda I + A)x} \le \frac{\lambda + tC^2 m^{-1/3}}{\lambda} = 1 + \frac{tC^2}{\lambda m^{1/3}}.$$

We conclude that

$$\lambda_M(W^{-1}V) \le 1 + \frac{tC^2}{\lambda m^{1/3}},$$

which gives the bound:

$$\lambda_M(W^{-1}V) = 1 + \mathcal{O}\left(\frac{t}{m^{1/3}}\right).$$

This completes the proof. ■

Refined bound with the time-independent $\|\theta_{t-1}\|_2$. The following lemma replaces the loose $\sqrt{t}$-dependent parameter norm in Lemma B.5 with the time-independent bound from Eq. (B.27). This substitution, combined with the spectral perturbation bounds from Lemmas B.6–B.7, reduces $\mathcal{K}_1(t)$ from $\widetilde{\mathcal{O}}(td/(\epsilon^2 m^{1/6}))$ to $\widetilde{\mathcal{O}}(\sqrt{t}d^2/(\epsilon^2 m^{1/6}))$—the critical improvement that yields sublinear regret.

Utilizing Lemma B.6 and Lemma B.7, we establish the following lemma, which holds for a sufficiently large width m satisfying Eq. (B.26), **yielding a finer estimation result:**

Lemma B.8. *Let the conditions in Lemma B.5 be satisfied. Additionally, if the width m satisfies Eq. (B.26), we can obtain*

$$\|\mathcal{G}_t(\theta_{t-1}) + M_t\|_{V_t^{-1}} \le \mathcal{J}_1 + \mathcal{K}_1(t),$$

where $\mathcal{J}_1$ and $\mathcal{K}_1(t)$ are defined in Eq. (B.22) and Eq. (B.23), respectively, such that

$$\mathcal{J}_1 = \widetilde{\mathcal{O}}\left(\sqrt{d} + \epsilon^{-1}\right),$$
$$\mathcal{K}_1(t) = \widetilde{\mathcal{O}}\left(t^{1/2} m^{-1/6}\left(\frac{d^2}{\epsilon^2} + d^3\right)\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} + d^{1/2}\right),$$

with probability at least $1 - \delta$.

Proof. It suffices to consider the terms $\mathcal{G}_{t,1,4}$ and

$$\mathcal{B}_t := \mathcal{G}_{t,1,3} + \mathcal{G}_{t,1,5} + \mathcal{G}_{t,1,6} + \mathcal{G}_{t,2,2}$$

which are appeared in Lemma B.5. Using Eq. (B.27) by Eq. (B.26), we can replace the result in Eq. (B.19) as follows:

$$\begin{aligned}
&|\mathcal{Z}_{i,t}(\theta_{t-1}) - \mathcal{Z}_{i,i}(\theta_{t-1})| = \left|\theta_{t-1}^\intercal \Delta\phi_{i,i} + \theta_{t-1}^\intercal \Delta\phi_{i,t}\right| \\
&\le C\|\theta_{t-1}\|_2 \left(\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} L^3 d^{1/2}\sqrt{\log m}/(m^{1/6}) + 2G\kappa_\phi \frac{\delta^{3/2}}{\sqrt{m}Tn^{9/2}\log^3 m}\right) \\
&= \widetilde{\mathcal{O}}\left((d^2 + d\epsilon^{-2})\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} m^{-1/6}\right),
\end{aligned}$$

for some constant $C > 0$ with probability at least $1 - \delta$. Therefore, we can compute

$$\begin{aligned}\|\mathcal{G}_{t,1,4}\|_{V_{t-1}^{-1}} &\leq \frac{\alpha_g}{\epsilon}\sqrt{\frac{1}{\lambda}}\sum_{i=1}^{t-1}|\mathcal{Z}_{i,t}(\theta_{t-1}) - \mathcal{Z}_{i,i}(\theta_{t-1})|\|\Delta\phi_{i,i}/\zeta_i\|_{V_{t-1}^{-1}} \\ &= \widetilde{\mathcal{O}}\left(\sqrt{td}\left((d^2 + d\epsilon^{-2})\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} m^{-1/6}\right)\right)\end{aligned} \tag{B.28}$$

with probability at least $1 - \delta$, where the last inequality is driven by Eq. (B.19) and the elliptic potential lemma in Lemma A.7 to estimate the summation of $\|\Delta\phi_{i,i}\|_{V_{t-1}^{-1}}$. Besides, note that

$$\mathcal{B}_t = \sum_{i=1}^{t-1}\left(o_i - g(\theta_{t-1}^\intercal \Delta\phi_{i,t})\right)\frac{(\Delta\phi_{i,t} - \Delta\phi_{i,i})}{\zeta_i^2}.$$

Combining Lemma B.6 and Lemma B.7, we can obtain

$$\|\mathcal{B}_t\|_{V_{t-1}^{-1}} \leq \widetilde{\mathcal{O}}\left(\left(1 + \left(\frac{t}{\epsilon^4 m^{1/3}}\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right)\right)\|\mathcal{B}_t\|_{\widetilde{U}_{t-1}^{-1}}\right)$$

where

$$\widetilde{U}_{t-1} = \lambda I + \sum_{i=1}^{t-1}\frac{(\Delta\phi_{i,t} - \Delta\phi_{i,i})}{\zeta_i^2}\frac{(\Delta\phi_{i,t} - \Delta\phi_{i,i})^\intercal}{\zeta_i^2}.$$

Since

$$\begin{aligned}\mathbb{E}\left[\left(o_i - g(\theta_{t-1}^\intercal \Delta\phi_{i,t})\right) \mid \mathcal{F}_t\right] &= 0, \\ \mathbb{E}\left[\left(o_i - g(\theta_{t-1}^\intercal \Delta\phi_{i,t})\right)^2 \mid \mathcal{F}_t\right] &\leq 1,\end{aligned}$$

we can apply Lemma A.8, resulting in

$$\|\mathcal{B}_t\|_{\widetilde{U}_{t-1}^{-1}} = \widetilde{\mathcal{O}}\left(\left(1 + \left(\frac{t}{\epsilon^4 m^{1/3}}\right)\right)d^{1/2}\right). \tag{B.29}$$

Combining Eqs. (B.28) and (B.29) and the other estimation results in Lemma B.5, we can obtain

$$\mathcal{K}_1(t) = \widetilde{\mathcal{O}}\left(\left(d\left(\frac{d}{\epsilon^2} + d^2\right)\frac{t^{1/2}}{m^{1/6}} + \frac{d^{1/2}}{\epsilon^4}\frac{t}{m^3}\right)\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} + d^{1/2}\right). \tag{B.30}$$

Notice that

$$\frac{t}{m^{1/3}} = \left(\frac{t^{1/2}}{m^{1/6}}\right)^2 = \left(\frac{t}{m^{1/6}}\right)\left(\frac{t^{1/2}}{m^{1/6}}\right)\frac{1}{\sqrt{t}} = \widetilde{\mathcal{O}}\left(\frac{t^{1/2}}{m^{1/6}}\right) \tag{B.31}$$

due to Eq. (B.2). Using Eq. (B.31), we can complete the proof.

■

Proof of Lemma B.1. Thanks to Lemma B.3, the following holds:

$$\begin{aligned}\left\|\theta_{t-1} - \theta_* + U_{t-1}^{-1}M_t\right\|_{V_{t-1}} &\leq \frac{1}{\beta_g}\left(\lambda\|\theta_* - \theta_0\|_{V_{t-1}^{-1}} + \|\mathcal{G}_t(\theta_{t-1}) + M_t\|_{V_{t-1}^{-1}}\right) \\ &\leq \mathcal{J}_1' + \mathcal{K}_1'(t)\end{aligned}$$

where

$$\mathcal{J}_1' = \mathcal{J}_1\frac{1}{\beta_g} + \|\theta_* - \theta_0\|_{V_{t-1}^{-1}} \tag{B.32}$$

$$\mathcal{K}_1' = \frac{1}{\beta_g}\mathcal{K}_1. \tag{B.33}$$

Here, $\mathcal{J}_1$ and $\mathcal{K}_1(t)$ are defined in Eq. (B.22) and Eq. (B.23). Due to Eq. (B.26), we can apply Lemma B.8 to estimate $\|\mathcal{G}_t(\theta_{t-1}) + M_t\|_{V_{t-1}^{-1}}$. Since $\|\theta_* - \theta_0\|_{V_{t-1}^{-1}}$ is bounded independently of T, m, ϵ and $\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}$, it is obvious that

$$\mathcal{J}_1' = \widetilde{\mathcal{O}}\left(d^{1/2} + \epsilon^{-1}\right), \tag{B.34}$$

$$\mathcal{K}_1'(t) = \widetilde{\mathcal{O}}\left(\left(\frac{d^2}{\epsilon^2} + d^3\right)\frac{t^{1/2}}{m^{1/6}}\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} + d^{1/2}\right). \tag{B.35}$$

This completes the proof. ■

Although there exist several theoretical results on concentration inequalities for generalized linear bandits (Kveton et al., 2020; Vaswani et al., 2019; Chen et al., 1999; Li et al., 2017) and contextual linear dueling bandits (Bengs et al., 2022; Saha, 2021), to the best of our knowledge, this is the first work that combines spectral perturbation analysis (the Courant–Fischer Minimax theorem) with a bootstrap argument. This refined approach is crucial for achieving a tight regret bound. A standard analysis that combines a concentration inequality with the elliptic potential lemma typically introduces an extra $\sqrt{T}$ factor, resulting in a suboptimal regret term on the order of $\widetilde{\mathcal{O}}(T^{3/2}m^{-1/6})$. In contrast, our analysis circumvents this issue, yielding a tighter and more favorable bound that scales as $\widetilde{\mathcal{O}}(Tm^{-1/6})$.

C THEORETICAL ANALYSIS OF UCB FRAMEWORK

Overview. This section proves sublinear cumulative regret for all UCB-based arm selection strategies. The proof structure is: (1) derive strategy-specific bounds on $\theta_{t-1}^\intercal(\cdot)$ relating the estimated utility gap to the exploration bonus (Section C.1); (2) decompose the instantaneous regret r_t^a into representation bias, estimated utility gap, and parameter estimation error terms; (3) apply the concentration inequality from Lemma B.1 and the elliptic potential lemma to sum the per-round bounds across T rounds.

C.1 Arm Selection Strategies

In this section, we consider three arm-selection strategies: Asymmetric Arm Selection (UCB-ASYM), Optimistic Symmetric Arm Selection (UCB-OSYM), and Candidate-Based Symmetric Arm Selection (UCB-CSYM).

We begin by analyzing the asymmetric strategy (UCB-ASYM):

Lemma C.1 (Asymmetric Arm Selection (UCB-ASYM))**.** *Let the asymmetric arm selection strategy be defined as*

$$k_{t,1} = \arg\max_{i\in[K]} \theta_{t-1}^\intercal \phi(x_{t,i}; W_{t-1}), \tag{C.1}$$

$$k_{t,2} = \arg\max_{i\in[K]} \left(\theta_{t-1}^\intercal \phi(x_{t,i}; W_{t-1}) + a_t\|\Delta\phi_{t,i,k_{t,1}}\|_{V_{t-1}^{-1}}\right), \tag{C.2}$$

for some $a_t > 0$. Then we can obtain

$$\theta_{t-1}^\intercal(2\Delta\phi_{t,k_t^*,k_{t,1}} + \Delta\phi_{t,k_{t,1},k_{t,2}}) \le a_t\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} - a_t\|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}}. \tag{C.3}$$

Proof. Notice that

$$\theta_{t-1}^\intercal\left(\Delta\phi_{t,k_t^*,k_{t,1}} + \Delta\phi_{t,k_t^*,k_{t,2}}\right) = 2\theta_{t-1}^\intercal(\Delta\phi_{t,k_t^*,k_{t,1}}) + \theta_{t-1}^\intercal(\Delta\phi_{t,k_{t,1},k_{t,2}}).$$

Eq. (C.1) results in

$$\theta_{t-1}^\intercal\Delta\phi_{t,k_t^*,k_{t,1}} \le 0. \tag{C.4}$$

In addition, Eq. (C.2) results in

$$\begin{aligned}&\theta_{t-1}^\intercal\phi\left(x_{t,k_t^*}; W_{t-1}\right) + a_t\|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}} \\ &\le \theta_{t-1}^\intercal\phi\left(x_{t,k_{t,2}}; W_{t-1}\right) + a_t\|\Delta\phi_{t,k_{t,2},k_{t,1}}\|_{V_{t-1}^{-1}}.\end{aligned} \tag{C.5}$$

Adding Eq. (C.4) and Eq. (C.5) leads to Eq. (C.3). This completes the proof. ■

Now we analyze two symmetric arm selection strategies as follows:

Lemma C.2 (Optimistic Symmetric Arm Selection (UCB-OSYM)). *Let the optimistic symmetric arm selection strategy be defined as*

$$(k_{t,1}, k_{t,2}) = \underset{(i,j)\in[K]\times[K]}{\arg\max} \left[\theta_{t-1}^\intercal \phi(x_{t,i}; W_{t-1}) + \theta_{t-1}^\intercal \phi(x_{t,j}; W_{t-1}) + a_t \|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}}\right], \tag{C.6}$$

for some $a_t > 0$. Then,

$$\begin{aligned} &2\theta_{t-1}^\intercal \phi(x_{t,k_t^*}; W_{t-1}) - \theta_{t-1}^\intercal \phi(x_{t,k_{t,1}}; W_{t-1}) - \theta_{t-1}^\intercal \phi(x_{t,k_{t,2}}; W_{t-1}) \\ &\le -a_t \|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}} - a_t \|\Delta\phi_{t,k_t^*,k_{t,2}}\|_{V_{t-1}^{-1}} + 2a_t \|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}. \end{aligned} \tag{C.7}$$

Proof. Using the strategy as in Eq. (C.6), we can obtain

$$\begin{aligned} &\theta_{t-1}^\intercal \phi(x_{t,k_t^*}; W_{t-1}) + \theta_{t-1}^\intercal \phi(x_{t,k_{t,1}}; W_{t-1}) + a_t \|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}} \\ &\qquad \le \theta_{t-1}^\intercal \phi(x_{t,k_{t,1}}; W_{t-1}) + \theta_{t-1}^\intercal \phi(x_{t,k_{t,2}}; W_{t-1}) + a_t \|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}, \\ &\theta_{t-1}^\intercal \phi(x_{t,k_t^*}; W_{t-1}) + \theta_{t-1}^\intercal \phi(x_{t,k_{t,2}}; W_{t-1}) + a_t \|\Delta\phi_{t,k_t^*,k_{t,2}}\|_{V_{t-1}^{-1}} \qquad \text{(C.8)} \\ &\qquad \le \theta_{t-1}^\intercal \phi(x_{t,k_{t,1}}; W_{t-1}) + \theta_{t-1}^\intercal \phi(x_{t,k_{t,2}}; W_{t-1}) + a_t \|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}.. \qquad \text{(C.9)} \end{aligned}$$

Adding Eq. (C.8) and Eq. (C.9), we can derive Eq. (C.7). This completes the proof. ■

Lemma C.3 (Candidate-based Symmetric Arm Selection (UCB-CSYM)). *Let the confidence set be defined as*

$$\mathcal{C}_t = \left\{ i \in [K] : \theta_{t-1}^\intercal \phi(x_{t,i}; W_{t-1}) - \theta_{t-1}^\intercal \phi(x_{t,j}; W_{t-1}) + a_t \|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}} \ge 0, \forall j \in [K] \right\}.$$

for some a_t and let two arms be chosen from $\mathcal{C}_t$ as

$$(k_{t,1}, k_{t,2}) = \underset{(i,j)\in\mathcal{C}_t\times\mathcal{C}_t}{\arg\max} \|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}}. \tag{C.10}$$

Then,

$$\begin{aligned} &2\theta_{t-1}^\intercal \phi(x_{t,k_t^*}; W_{t-1}) - \theta_{t-1}^\intercal \phi(x_{t,k_{t,1}}; W_{t-1}) - \theta_{t-1}^\intercal \phi(x_{t,k_{t,2}}; W_{t-1}) && \text{(C.11)} \\ &\le -a_t \|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}} - a_t \|\Delta\phi_{t,k_t^*,k_{t,2}}\|_{V_{t-1}^{-1}} + 4a_t \|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}. && \text{(C.12)} \end{aligned}$$

Proof. Thanks to Eq. (C.10), for $k_{t,1}, k_{t,2} \in C_t$, we can get

$$\begin{aligned} &\theta_{t-1}^\intercal \left(2\phi\left(x_{t,k_t^*}; W_{t-1}\right) - \phi\left(x_{t,k_{t,1}}; W_{t-1}\right) - \phi\left(x_{t,k_{t,2}}; W_{t-1}\right)\right) && \text{(C.13)} \\ &\le a_t \|\Delta\phi_{t,k_{t,1},k_t^*}\|_{V_{t-1}^{-1}} + a_t \|\Delta\phi_{t,k_{t,2},k_t^*}\|_{V_{t-1}^{-1}}. && \text{(C.14)} \end{aligned}$$

Due to Eq. (C.10),

$$\begin{aligned} &\|\Delta\phi_{t,k_{t,1},k_t^*}\|_{V_{t-1}^{-1}} \le 2\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} - \|\Delta\phi_{t,k_{t,1},k_t^*}\|_{V_{t-1}^{-1}}, && \text{(C.15)} \\ &\|\Delta\phi_{t,k_{t,2},k_t^*}\|_{V_{t-1}^{-1}} \le 2\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} - \|\Delta\phi_{t,k_{t,2},k_t^*}\|_{V_{t-1}^{-1}}. && \text{(C.16)} \end{aligned}$$

Substituting Eq. (C.15) and Eq. (C.16) to Eq. (C.13) leads to Eq. (C.11), thus we can complete the proof. ■

C.2 Theoretical Cumulative Average Regret Analysis under UCB Regime

We restate the theoretical results under Upper Confidence Bound (UCB) regime, already stated in the main manuscript.

Theorem C.1 (Variance-aware UCB methods)**.** *Let us assume Assumption A.1-Assumption A.3. Let the confidence width coefficient a_t for the arm selection methods and the step size η be chosen to satisfy the following:*

$$\begin{aligned} a_t &= \frac{8}{\beta_g}\left(\sqrt{d\log(a_t')\log(4t^2/\delta)} + \frac{1}{\epsilon}\log(4t^2/\delta)\right) \\ &\quad +2\sqrt{\lambda}\left(J+2\left(2+\sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right)\right), \\ \eta &\le C\left(d^2mnT^6L^6\log(TK/\delta)\right)^{-1} \end{aligned}$$

for some $C>0$ with

$$a_t' = \left(1+t\left(2\frac{\sqrt{d\log n\log(TK/\delta)}}{\sqrt{d\lambda}\epsilon}\right)^2\right).$$

If the width m of f is sufficiently large as in Lemma 1 and $\epsilon = \mathcal{O}(1/\sqrt{d})$, then the cumulative average regret R_T^a under any of our UCB-based arm selection methods is upper-bounded as

$$R_T^a = \widetilde{\mathcal{O}}\left(\sqrt{dT} + d\sqrt{\sum_{i=1}^{T}\sigma_i^2} + \frac{Td^3}{m^{1/6}}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right). \tag{C.17}$$

ignoring all other hyperparameters with probability at least $1-\delta$ over the randomness of the initialization of the neural network f.

The following result is about variance-agnostic UCB-based algorithms.

Corollary C.1 (Variance-agnostic UCB methods)**.** *Assume the conditions in Theorem 1 hold, but we fix $\zeta_t = 1$ (i.e., $\epsilon = 1$) for all round $0 \le t \le T$. Then the cumulative regret R_t^a under any of our UCB-based arm selection methods is bounded as*

$$R_T^a = \widetilde{\mathcal{O}}\left(d\sqrt{T} + \frac{Td^3}{m^{1/6}}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right). \tag{C.18}$$

ignoring all other hyperparameters with probability at least $1-\delta$ over the randomness of learnable parameters in the neural network f

Regret decomposition. The instantaneous regret $2r_t^a$ is decomposed into three pairs of terms: $\mathcal{A}_{t,1,\cdot}$ captures the *representation learning bias* (difference between the true utility and the linearized approximation around W_0); $\mathcal{A}_{t,2,\cdot}$ captures the *estimated utility gap* (controlled by the arm selection strategy bounds above); and $\mathcal{A}_{t,3,\cdot}$ captures the *parameter estimation error* $(\theta_* - \theta_{t-1})$, which is bounded by our concentration inequality. The bias terms $\mathcal{A}_{t,1,\cdot}$ and the inner product terms $\mathcal{K}_3$ vanish as $m \to \infty$, while $\mathcal{A}_{t,2,\cdot}$ and $\mathcal{A}_{t,3,\cdot}$ yield the exploration bonus terms that are summed via the elliptic potential lemma.

To prove the results above, we decompose the average regret as follows by Lemma A.2 with the notations in Eq. (A.2):

$$2r_t^a = \mathcal{A}_{t,1,1} + \mathcal{A}_{t,1,2} + \mathcal{A}_{t,2,1} + \mathcal{A}_{t,2,2} + \mathcal{A}_{t,3,1} + \mathcal{A}_{t,3,2}$$

where

$$\begin{aligned} \mathcal{A}_{t,1,1} &= \theta_0^\intercal\left(\nabla\Delta\phi_{t,k_t^*,k_{t,1},1}\right)(W_* - W_0), \\ \mathcal{A}_{t,1,2} &= \theta_0^\intercal\left(\nabla\Delta\phi_{t,k_t^*,k_{t,2},1}\right)(W_* - W_0), \\ \mathcal{A}_{t,2,1} &= \theta_{t-1}^\intercal\Delta\phi_{t,k_t^*,k_{t,1}}, \\ \mathcal{A}_{t,2,2} &= \theta_{t-1}^\intercal\Delta\phi_{t,k_t^*,k_{t,2}}, \\ \mathcal{A}_{t,3,1} &= (\theta_* - \theta_{t-1})^\intercal\Delta\phi_{t,k_t^*,k_{t,1}}, \\ \mathcal{A}_{t,3,2} &= (\theta_* - \theta_{t-1})^\intercal\Delta\phi_{t,k_t^*,k_{t,2}}. \end{aligned}$$

Lemma C.4. *With probability at least $1-\delta$, the following holds:*

$$\begin{aligned}\mathcal{A}_{t,1,1}+\mathcal{A}_{t,1,2} &\leq \kappa_\phi\|\theta_0\|_2\|x_{t,k_{t,1}}-x_{t,k_{t,2}}\|_2\|W_*-W_0\|_2\\ &\leq \mathcal{K}_2(m)\end{aligned} \tag{C.19}$$

where

$$\mathcal{K}_2 = 4G\kappa_\phi\left(2+\sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right)\frac{\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}}{\sqrt{m}} = \widetilde{\mathcal{O}}\left(\frac{\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}}{\sqrt{m}}\right). \tag{C.20}$$

Proof. It is obvious

$$\begin{aligned}|\mathcal{A}_{t,1,1}+\mathcal{A}_{t,1,2}| \leq{}& \|\theta_0\|_2\left\|\left(\nabla\phi\left(x_{t,k_t^*};W_0\right)-\nabla\phi\left(x_{t,k_{t,1}};W_0\right)\right)(W_*-W_0)\right\|_2\\ &+\|\theta_0\|_2\left\|\left(\nabla\phi\left(x_{t,a_t^*};W_0\right)-\nabla\phi\left(x_{t,k_{t,2}};W_0\right)\right)(W_*-W_0)\right\|_2\end{aligned}$$

Notice that

$$\|Mx\|_2 \leq \|M\|_2\|x\|_2$$

for a matrix M and vector x. By Eq. (A.8), Assumption A.1, Assumption A.2 and Lemma A.2, we can show Eq. (C.19). This completes the proof. ■

$\mathcal{A}_{t,2,1}+\mathcal{A}_{t,2,2}$ can be expressed as

$$\begin{aligned}&\mathcal{A}_{t,2,1}+\mathcal{A}_{t,2,2}\\ &=\theta_{t-1}^\intercal\left(2\phi\left(x_{t,k_t^*};W_{t-1}\right)-\phi\left(x_{t,k_{t,1}};W_{t-1}\right)-\phi\left(x_{t,k_{t,2}};W_{t-1}\right)\right)\\ &=\theta_{t-1}^\intercal\left(2\left(\phi\left(x_{t,k_t^*};W_{t-1}\right)-\phi\left(x_{t,k_{t,1}};W_{t-1}\right)\right)\right)\\ &\quad+\theta_{t-1}^\intercal\left(\phi\left(x_{t,k_{t,1}};W_{t-1}\right)-\phi\left(x_{t,k_{t,2}};W_{t-1}\right)\right),\end{aligned}$$

where the first equality applies to symmetric arm selections, while the second equality pertains to the asymmetric arm selection. Using the results in Section C.1, the following holds:

$$\begin{aligned}&\text{(UCB-ASYM):}\quad \mathcal{A}_{t,2,1}+\mathcal{A}_{t,2,2} \leq 2a_t\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} - 2a_t\|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}}.\\ &\text{(UCB-OSYM):}\quad \mathcal{A}_{t,2,1}+\mathcal{A}_{t,2,2} \leq 2a_t\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}\\ &\qquad\qquad -a_t\|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}} - a_t\|\Delta\phi_{t,k_t^*,k_{t,2}}\|_{V_{t-1}^{-1}}.\\ &\text{(UCB-CSYM):}\quad \mathcal{A}_{t,2,1}+\mathcal{A}_{t,2,2} \leq 4a_t\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}\\ &\qquad\qquad -a_t\|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}} - a_t\|\Delta\phi_{t,k_t^*,k_{t,2}}\|_{V_{t-1}^{-1}}.\end{aligned}$$

$\mathcal{A}_{t,3,1}+\mathcal{A}_{t,3,2}$ can be decomposed into

$$\begin{aligned}\mathcal{A}_{t,3,1}+\mathcal{A}_{t,3,2} ={}& \left(\theta_*-\theta_{t-1}+U_{t-1}^{-1}M_t\right)^\intercal\Delta\phi_{t,k_t^*,k_{t,1}}+\left(\theta_*-\theta_{t-1}+U_{t-1}^{-1}M_t\right)^\intercal\Delta\phi_{t,k_t^*,k_{t,2}}\\ &-(U_{t-1}^{-1}M_t)^\intercal\Delta\phi_{t,k_t^*,k_{t,1}}-(U_{t-1}^{-1}M_t)^\intercal\Delta\phi_{t,k_t^*,k_{t,2}}\\ ={}& 2\left(\theta_*-\theta_{t-1}+U_{t-1}^{-1}M_t\right)^\intercal\Delta\phi_{t,k_t^*,k_{t,1}}+\left(\theta_*-\theta_{t-1}+U_{t-1}^{-1}M_t\right)^\intercal\Delta\phi_{t,k_{t,1},k_{t,2}}\\ &-2(U_{t-1}^{-1}M_t)^\intercal\Delta\phi_{t,k_t^*,k_{t,1}}-(U_{t-1}^{-1}M_t)^\intercal\Delta\phi_{t,k_{t,1},k_{t,2}}.\end{aligned}$$

The first inequality is used for symmetric arm selections, while the second equality pertains to the asymmetric arm selection. The last two terms are estimated using the following lemma.

Lemma C.5. *With probability at least $1-\delta$, the following holds:*

$$|(U_{t-1}^{-1}M_t)^\intercal\Delta\phi_{t,i,j}| \leq \mathcal{K}_3,$$

where

$$\begin{aligned}\mathcal{K}_3 &= \left(\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} L^3 d^{1/2}\sqrt{\log m}/(m^{1/6}) + 2G\kappa_\phi \frac{\delta^{3/2}}{\sqrt{m}n^{9/2}\log^3 m} \right) \\ &\quad \times 2d\kappa_\phi G\frac{\alpha_g}{\epsilon}\left(2+\sqrt{d^{-1}\log\left(\frac{1}{\delta}\right)}\right)\sqrt{dL}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}} \\ &= \widetilde{\mathcal{O}}\left(m^{-1/6}\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} d^{3/2}\epsilon^{-1}\right).\end{aligned} \tag{C.21}$$

Proof. Note that

$$|(U_{t-1}^{-1}M_t)^\intercal \Delta\phi_{t,i,j}| \le \|U_{t-1}^{-1}M_t\|_2 \|\Delta\phi_{t,i,j}\|_2.$$

by the Cauchy–Schwarz inequality. By applying Lemma B.4 to $\|U_{t-1}^{-1}M_t\|_2$ and Lemma A.6 to $\|\Delta\phi_{t,i,j}\|_2$, we can complete the proof. ∎

Proof of Corollary C.1. Under the asymmetric arm selection,

$$2r_t^a \le \mathcal{K}_2 + 2\mathcal{K}_3 + a_t\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} - a_t\|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}} \tag{C.22}$$

$$+ (\mathcal{K}_1'(t)+\mathcal{J}_1')\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} + (\mathcal{K}_1'(t)+\mathcal{J}_1')\|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}}, \tag{C.23}$$

where the last two terms result from Lemma B.1. Therefore, choosing $a_t = \mathcal{K}_1'(t)+\mathcal{J}_1'$, we can obtain

$$2r_t^a = \widetilde{\mathcal{O}}\left(\mathcal{K}_2 + \mathcal{K}_3 + (\mathcal{K}_1'(t)+\mathcal{J}_1')\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}\right). \tag{C.24}$$

where $\mathcal{J}_1', \mathcal{K}_1', \mathcal{K}_2, \mathcal{K}_3$ are defined in (B.32), (B.33), (C.20), (C.21).

In a similar manner to the asymmetric arm selection, we can obtain equivalent inequalities under the two symmetric arm selections. In other words, there exists $C>0$ such that

$$2r_t^a = \widetilde{\mathcal{O}}\left(\left(\mathcal{K}_2 + \mathcal{K}_3 + (\mathcal{K}_1'(t)+\mathcal{J}_1')\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}\right)\right), \tag{C.25}$$

independently of the arm selection strategies, if $a_t = \mathcal{K}_1'(t)+\mathcal{J}_1'$. Note that $\mathcal{J}_1' = \widetilde{\mathcal{O}}(\sqrt{d})$. Therefore, if we set $\zeta_t = 1$, which is the variance-agnostic case. By summing up to T, and applying Lemma A.7, we can show that

$$\begin{aligned}2\sum_{t=1}^T r_t^a &= 2\sqrt{T}\left(\sum_{t=1}^T |r_t^a|^2\right)^{1/2} \\ &= \widetilde{\mathcal{O}}\left(\sqrt{T}\left(d + d^3\left(T^{1/2}M^{-1/6}\right)\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right)\right) \\ &= \widetilde{\mathcal{O}}\left(d\sqrt{T} + d^3\left(TM^{-1/6}\right)\|\mathbf{u}-\widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right).\end{aligned} \tag{C.26}$$

Thus, we can prove Corollary 1. ∎

Proof of Theorem C.1. We now set $\epsilon = \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{d}}\right)$. Since

$$\Delta\phi_{t,k_{t,1},k_{t,2}} = \zeta_t\left(\frac{\Delta\phi_{t,k_{t,1},k_{t,2}}}{\zeta_t}\right),$$

summing both sides of Eq. (C.25) from $t=1$ to T, we apply the Cauchy–Schwarz inequality and subsequently use the elliptic potential lemma in Lemma A.7 to obtain the following bound:

$$
\begin{aligned}
2\sum_{t=1}^{T} r_t^a &\leq C\left(\mathcal{J}_1' + \mathcal{K}_1'\right)\left(\sum_{t=1}^{T}\left\|\frac{\Delta\phi_{t,k_{t,1},k_{t,2}}}{\zeta_t}\right\|^2_{V_{t-1}^{-1}}\right)^{1/2}\left(\sum_{t=1}^{T}\zeta_t^2\right)^{1/2} + \mathcal{R}_1(m,T) \\
&\leq C\sqrt{d}\sqrt{2d\log\left(1+T\left(d\log n\log(TK/\delta)\right)^2/\lambda\right)}\left(\sum_{t=1}^{T}\zeta_t^2\right)^{1/2} + \mathcal{R}_1(m,T) \\
&= \widetilde{\mathcal{O}}\left(d\left(\sum_{t=1}^{T}\zeta_t^2\right)^{1/2} + \mathcal{R}_1(m,T)\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\mathcal{R}_1(m,T) &= \widetilde{\mathcal{O}}\left(d^3\left(TM^{-1/6} + \left(TM^{-1/6}\right)^2/(\sqrt{T})\right)\|\mathbf{u}-\widetilde{\mathbf{u}}\|^{4/3}_{\mathbf{H}^{-1}}\right) \\
&= \widetilde{\mathcal{O}}\left(d^3\left(TM^{-1/6}\right)\|\mathbf{u}-\widetilde{\mathbf{u}}\|^{4/3}_{\mathbf{H}^{-1}}\right)
\end{aligned}
$$

by hiding the dependence on all other hyperparameters within the $\widetilde{\mathcal{O}}$ notation. The last term results from Eq. (B.26). Note that $\mathcal{J}_1' = \widetilde{\mathcal{O}}(\sqrt{d})$. It is obvious that

$$
\zeta_i^2 = \max\left\{\epsilon^2, \widehat{\sigma}_i^2\right\} \leq \epsilon^2 + \widehat{\sigma}_i^2.
$$

$\widehat{\sigma}_i^2$ is a estimation of variance by (θ_i, W_i). Since $\epsilon = \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{d}}\right)$ and

$$
\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}
$$

for $a, b > 0$, we can estimate

$$
\begin{aligned}
\left(\sum_{i=1}^{T}\zeta_i^2\right)^{\frac{1}{2}} &\leq \epsilon\sqrt{T} + \sqrt{\sum_{i=1}^{T}\widehat{\sigma}_i^2} \\
&= \widetilde{\mathcal{O}}\left(\sqrt{\frac{T}{d}} + \sqrt{\sum_{i=1}^{T}\widehat{\sigma}_i^2}\right)
\end{aligned} \tag{C.27}
$$

As a result, utilizing Eq. (C.27), we can estimate cumulative average regret as

$$
\begin{aligned}
2\sum_{i=1}^{T} r_t^a &= \widetilde{\mathcal{O}}\left(\sqrt{dT} + d\sqrt{\sum_{i=1}^{T}\widehat{\sigma}_i^2}\right) + \mathcal{R}_1(m,T) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{dT} + d\sqrt{\sum_{i=1}^{T}\widehat{\sigma}_i^2 + T^2 m^{-1/6} d^3 \|\mathbf{u}-\widetilde{\mathbf{u}}\|^{4/3}_{\mathbf{H}^{-1}}}\right).
\end{aligned}
$$

Now we estimate the difference between σ_t^2 a real variance with respect to $x_{t,k_{t,1}}$ and $x_{t,k_{t,2}}$, and the estimated variance $\widehat{\sigma}_t^2$. Let

$$
\begin{aligned}
\widehat{g}_t &= g(\theta_{t-1}^{\intercal}\Delta\phi_t) \\
g_t &= g(\theta_*^{\intercal}\Delta\phi_t + \theta_0^{\intercal}\nabla\Delta\phi_{t,1}(W_* - W_0)).
\end{aligned}
$$

We can compute that

$$
\begin{aligned}
|\sigma_t^2 - \widehat{\sigma}_t^2| &= |g_t(1-g_t) - \widehat{g}_t(1-\widehat{g}_t)| = |g_t - \widehat{g}_t - g_t^2 + \widehat{g}_t^2| \\
&= |(g_t - \widehat{g}_t)(1-(g_t+\widehat{g}_t))|.
\end{aligned}
$$

Note that the mean-value theorem leads to

$$|g_t - \widehat{g}_t| \leq \alpha_u |(\theta_{t-1} - \theta_*)^\intercal \Delta\phi_t + \theta_0^\intercal \nabla \Delta\phi_{t,1}(W_* - W_0)|. \tag{C.28}$$

Due to Assumption A.2 and Eq. (B.4),

$$\|\theta_{t-1} - \theta_*\|_2 = \widetilde{\mathcal{O}}\left(d^{3/2}\left(1 + \|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}\right)\right)$$

since Eq. (B.26) is satisfied. Moreover, due to (A.11),

$$\|\Delta\phi_t\|_2 = \widetilde{\mathcal{O}}\left(\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3} d^{1/2} m^{-1/6}\right). \tag{C.29}$$

Besides, since θ_0 is bounded with probability at least $1-\delta$ by Lemma A.1 and the Lipschitz condition of $\nabla_W \phi(x, W_0)$ by Assumption 2, we can have

$$|\theta_0^\intercal \nabla \Delta\phi_{t,1}(W_* - W_0)| = \widetilde{\mathcal{O}}(\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}} m^{-1/2}) \tag{C.30}$$

as

$$|W_* - W_0| \leq \frac{\|\widetilde{\mathbf{u}} - \mathbf{u}\|_{\mathbf{H}^{-1}}}{\sqrt{m}}.$$

Thus, using Eqs. (C.28) to (C.30), we can show that

$$|\sigma_t^2 - \widehat{\sigma}_t^2| \leq \mathcal{R}_2(T, m) \tag{C.31}$$

where

$$\mathcal{R}_2(m, T) = \widetilde{\mathcal{O}}(d^2 m^{-1/6} \|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}) \tag{C.32}$$

by hiding the dependence on all other hyperparameters within the $\widetilde{\mathcal{O}}$ notation for sufficiently large m such that Eq. (B.26) is satisfied. As a result, since $\epsilon = \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{d}}\right)$, we can obtain

$$\begin{aligned}
2\sum_{i=1}^{T} r_t^a &= \widetilde{\mathcal{O}}\left(d\epsilon\sqrt{T} + d\sqrt{\sum_{i=1}^{T} \widehat{\sigma}_i^2}\right) + \mathcal{R}_1(m, T) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{dT} + d\sqrt{\sum_{i=1}^{T} \sigma_i^2}\right) + \mathcal{R}_1(m, T) + T\mathcal{R}_2(m, T) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{dT} + d\sqrt{\sum_{i=1}^{T} \sigma_i^2 + m^{-1/6} T d^3 \|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}}\right).
\end{aligned}$$

This completes the proof of Theorem 1. ■

D THEORETICAL CUMULATIVE AVERAGE REGRET ANALYSIS UNDER TS REGIME

Overview. The TS proof strategy differs from UCB. Instead of directly bounding r_t^a, we construct a super-martingale $Y_t = \sum_{i=1}^{t}(r_i^a I_{E_i} - c_t)$ and apply the Azuma–Hoeffding inequality. The key steps for each arm selection strategy are: (1) define "good events" E_t (concentration holds) and F_t (Gaussian samples are well-behaved); (2) define a "saturation set" S_t of arm pairs whose utility gap is large enough that they are unlikely to be selected; (3) show that the selected arms $(k_{t,1}, k_{t,2})$ fall outside S_t with positive probability p; (4) use this to bound $\mathbb{E}[r_t^a | \mathcal{F}_{t-1}]$ by exploration bonus terms. The TS-OSYM and TS-CSYM strategies are more complex because both arms are drawn simultaneously from a joint posterior, requiring careful probability bounds via independence arguments.

We restate theoretical results for TS algorithms, already explained in the main manuscript, as follows.

Theorem D.1 (Variance-aware TS methods). *Assume the conditions in Theorem C.1 hold. Then the cumulative regret R_t^a under any of TS-based arm selection methods is bounded as Eq. (C.17) with probability at least $1-\delta$ over the randomness of learnable parameters in the neural network f.*

In the case of variance-agnostic TS algorithms, we have the following result.

Corollary D.1 (Variance-agnostic TS methods). *Assume the conditions in Corollary C.1 hold. Then the cumulative regret R_t^a under any of our TS-based arm selection methods is bounded as (C.18) with probability at least $1-\delta$ over the randomness of learnable parameters in the neural network f.*

Let the difference of the utilities resulting from $x_{t,k}$ and $x_{t,k'}$ for $k, k' \in [K]$ as

$$\Delta u_{t,k,k'} = u(x_{t,k}) - u(x_{t,k'}).$$

Due to Lemma B.1 and Lemma B.4, we have the following inequality:

$$\begin{aligned}
|\Delta u_{t,k,k'} - \theta_{t-1}^\intercal \Delta\phi_{t,k,k'}| &= |(\theta_* - \theta_{t-1})^\intercal \Delta\phi_{t,k,k'} + \theta_0^\intercal \nabla\Delta\phi_{t,k,k',1}(W_* - W_0)| \\
&\le \|\theta_* - \theta_{t-1} + U_{t-1}^{-1} M_t\|_{V_{t-1}} \|\Delta\phi_{t,k,k'}\|_{V_{t-1}^{-1}} \\
&\quad + |(U_{t-1}^{-1} M_t)^\intercal \Delta\phi_{t,k,k'}| + |\theta_0^\intercal \nabla\Delta\phi_{t,k,k',1}(W_* - W_0)| \\
&\le (\mathcal{J}_1' + \mathcal{K}_1')\|\Delta\phi_{t,k,k'}\|_{V_{t-1}^{-1}} + \mathcal{K}_3 + 2CG\frac{\|\widetilde{\mathbf{u}} - \mathbf{u}\|_{\mathbf{H}^{-1}}}{\sqrt{m}} \\
&= (\mathcal{J}_1' + \mathcal{K}_1')\|\Delta\phi_{t,k,k'}\|_{V_{t-1}^{-1}} + \mathcal{R}(m),
\end{aligned} \tag{D.1}$$

for some constant $C > 0$ with probability at least $1-\delta$, where $\mathcal{J}_1'$ is defined in Eq. (B.32), $\mathcal{K}_1'$ is defined in Eq. (B.33), $\mathcal{K}_3$ defined in Eq. (C.21). Note that

$$\mathcal{R}(m) = \widetilde{\mathcal{O}}\left(d^2 m^{-1/6}\|\widetilde{\mathbf{u}} - \mathbf{u}\|_{\mathbf{H}^{-1}}^{4/3}\right)$$

with probability at least $1-\delta$. Using Lemma A.6 for $\|\Delta\phi_{t,t}\|_2$, we can show that

$$|\Delta u_{t,t} - \theta_{t-1}^\intercal \Delta\phi_{t,t}| = \widetilde{\mathcal{O}}(dm^{-1/6}\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}^{4/3}).$$

In other words, for sufficiently large m, $|\Delta u_{t,t} - \theta_{t-1}^\intercal \Delta\phi_{t,t}|$ can be bounded.

We need the following lemma for the theoretical analysis of each arm selection strategy.

Lemma D.1. *Suppose that Eq. (B.26) is satisfied. Let $\delta \in [0,1]$ and X_t be a random variable of the form:*

$$X_t = r_t^a I_{E_t} - c_t$$

where E_t is a set of events such that $\mathbb{P}(E_t) \ge 1-\delta$ and c_t is of the form:

$$c_t \le C\left(c_t'\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} + \mathcal{R}(m) + \frac{1}{t^2}\right)$$

for some $c_t' > 0$ such that $c_t' = \widetilde{\mathcal{O}}(\sqrt{d})$. If the network width m satisfies Eq. (B.26), and $Y_t = \sum_{i=1}^t X_i$ is a super-martingale with respect to some filtration set $\{\mathcal{F}_t\}$, the following holds:

$$R_T^a = \widetilde{\mathcal{O}}\left(d\epsilon\sqrt{T} + d^3 T m^{-1/6}\right) \quad \text{for } \epsilon = \widetilde{\mathcal{O}}(1/\sqrt{d}) \text{ (variance-aware) in Eq. (A.7)},$$

$$R_T^a = \widetilde{\mathcal{O}}\left(d\epsilon\sqrt{T} + d\sqrt{\sum_{i=1}^T \sigma_i^2} + d^3 T m^{-1/6}\right) \quad \text{for } \epsilon = 1 \text{ (variance-agnostic) in Eq. (A.7)}.$$

Proof. We apply the Azuma-Hoeffding inequality (Zhang, 2023) to Y_t. With probability at least $1-\delta$, the following holds:

$$Y_t - Y_0 = \sum_{t=1}^T r_t^a I_{E_t} - c_t \le \sqrt{\log(1/\delta)\sum_{i=1}^t |Y_i - Y_{i-1}|}.$$

Notice that $|Y_i - Y_{i-1}| = |r_t^a| + |c_t| = \widetilde{\mathcal{O}}(1)$ due to the definition of $\mathcal{R}(m)$ to c_t with Eq. (B.26). Therefore, there exists a constant C such that $|Y_i - Y_{i-1}| \leq C$ and

$$R_T^a = \sum_{t=1}^{T} r_t \leq \sum_{t=1}^{T} c_t + \sqrt{\log(1/\delta)TC}.$$

with probability at least $1-\delta$. By the definition of c_t, we can compute the following:

$$\begin{aligned} R_T^a = \sum_{t=1}^{T} r_t &\leq C_1 \left(\sum_{t=1}^{T} c_t' \|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} + T\mathcal{R}(m) + \sum_{t=1}^{T} \frac{1}{t^2} \right) \\ &\leq C_2 \left(\sum_{t=1}^{T} c_t' \|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} + T\mathcal{R}(m) + \pi^2/6 \right) \\ &\leq C_2 \left(\max_{t\in[0,T]} c_t' \left(\sum_{t=1}^{T} \|\Delta\phi_{t,k_{t,1},k_{t,2}}\|^2_{V_{t-1}^{-1}} \right) + T\mathcal{R}(m) + \pi^2/6 \right) \end{aligned}$$

for some constants $C_1, C_2 > 0$, where the second inequality results from $\sum_{t=1}^{\infty} \frac{1}{t^2} = \pi/6$ and the third inequality results from the Cauchy–Schwarz inequality. Notice that $c_t' = \widetilde{\mathcal{O}}(\sqrt{d})$. In a similar manner to the case of the UCB framework (Section C), applying Lemma A.7, we can obtain the following inequality in both the variance-aware

$$R_T^a = \widetilde{\mathcal{O}} \left(d\epsilon\sqrt{T} + d\sqrt{\sum_{i=1}^{T} \sigma_i^2} + m^{-1/6} d^3 T \right), \tag{D.2}$$

and variance-agnostic cases:

$$R_T^a = \widetilde{\mathcal{O}} \left(d\epsilon\sqrt{T} + m^{-1/6} d^3 T \right), \tag{D.3}$$

with probability at least $1-\delta$. This completes the proof. ■

Accordingly, for each TS-based arm selection, we will show that Lemma D.1 is satisfied. Then Theorem D.1 and Corollary D.1 are proved.

Proof structure for each TS strategy. For each arm selection method below, the proof follows the same template: (1) define the good events E_t and F_t; (2) define the saturation set S_t; (3) show the selected arms avoid S_t with positive probability; (4) bound the conditional regret $\mathbb{E}[r_t^a | \mathcal{F}_{t-1}]$; (5) construct the super-martingale Y_t and apply Lemma D.1.

D.1 Asymmetry Arm Selection Strategy:

We consider the following strategy:

$$\begin{aligned} k_{t,1} &= \arg\max_{k\in[K]} \theta_{t-1}^{\intercal} \phi(x_{t,k}; W_{t-1}), \\ k_{t,2} &= \arg\max_{k\in[K]} \Delta\widetilde{u}_{t,k,k_{t,1}} \end{aligned}$$

where $\Delta\widetilde{u}_{t,k,k_{t,1}} \sim \mathcal{N}(\theta_{t-1}^{\intercal}\Delta\phi_{t,k,k_{t,1}}, a_t^2 \|\Delta\phi_{t,k,k_{t,1}}\|^2_{V_{t-1}^{-1}})$ with the Gaussian distribution $\mathcal{N}$. We define E_t, and F_t as

$$E_t = \{|\Delta u_{t,k,k_{t,1}} - \theta_{t-1}^{\intercal}\Delta\phi_{t,k,k_{t,1}}| \leq a_t \|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}} + \mathcal{R}(m)\} \tag{D.4}$$

$$F_t = \{|\Delta\widetilde{u}_{t,k,k_{t,1}} - \theta_{t-1}^{\intercal}\Delta\phi_{t,k,k_{t,1}}| \leq a_t\sqrt{2\log(Kt^2)} \|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}}\}. \tag{D.5}$$

Moreover, let $b_t = a_t \left(1 + \sqrt{2\log(Kt^2)}\right)$. We need to show that E_t and F_t in Eqs. (D.4) and (D.5) are good events.

Lemma D.2. *The following inequalities are satisfied:*

$$\mathbb{P}\left(E_t\right) \geq 1-\delta, \tag{D.6}$$

$$\mathbb{P}\left(F_t\right) \geq 1-\frac{1}{t^2}. \tag{D.7}$$

Proof. Eq. (D.6) is obtained by Eq. (D.1). Besides, by the property of the Gaussian distribution, Eq. (D.7) is obtained by Hoeffding's inequality (Lattimore and Szepesvári, 2020) with choosing the confidence radius as $a_t\sqrt{2\log(Kt^2)}\|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}}$ with the sample size 1. ■

Lemma D.3. *For any filtration $\mathcal{F}_{t-1}$ and the condition E_t, the following inequality is satisfied:*

$$\mathbb{P}(\Delta\widetilde{u}_{t,k,k_{t,1}} + \mathcal{R}(m) > \Delta u_{t,k,k_{t,1}} | \mathcal{F}_{t-1}) \geq \frac{1}{4e\sqrt{\pi}}.$$

Proof. It is straightforward due to the following inequalities:

$$\begin{aligned}
&\mathbb{P}(\Delta\widetilde{u}_{t,k,k_{t,1}} + \mathcal{R}(m) > \Delta u_{t,k,k_{t,1}} | \mathcal{F}_{t-1}) \\
&= \mathbb{P}\left(\frac{\Delta\widetilde{u}_{t,k,k_{t,1}} + \mathcal{R}(m) - \theta_{t-1}^\intercal \Delta\phi_{t,k,k_{t,1}}}{a_t\|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}}} > \frac{\Delta u_{t,k,k_{t,1}} - \theta_{t-1}^\intercal \Delta\phi_{t,k,k_{t,1}}}{a_t\|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}}} \middle| \mathcal{F}_{t-1}\right) \\
&\geq \mathbb{P}\left(\frac{\Delta\widetilde{u}_{t,k,k_{t,1}} - \theta_{t-1}^\intercal \Delta\phi_{t,k,k_{t,1}}}{a_t\|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}}} > \frac{|\Delta u_{t,k,k_{t,1}} - \theta_{t-1}^\intercal \Delta\phi_{t,k,k_{t,1}}| - \mathcal{R}(m)}{a_t\|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}}} \middle| \mathcal{F}_{t-1}\right) \\
&\geq \mathbb{P}\left(\frac{\Delta\widetilde{u}_{t,k,k_{t,1}} - \theta_{t-1}^\intercal \Delta\phi_{t,k,k_{t,1}}}{a_t\|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}}} > 1 \middle| \mathcal{F}_{t-1}\right) \geq \frac{1}{4e\sqrt{\pi}}.
\end{aligned}$$

The last inequality is obtained by the definition of E_t. This completes the proof. ■

Saturation set. The saturation set S_t contains arms that are "clearly suboptimal"—their utility gap to the optimal arm exceeds the exploration bonus. Arms in S_t are unlikely to be selected by TS because the Gaussian perturbation is insufficient to make them appear optimal. The key argument is that the selected arm $k_{t,2}$ falls outside S_t with positive probability, so the instantaneous regret can be bounded by the exploration bonus of the non-saturated arm.

We define a set of saturated points.

$$S_t = \left\{k : \Delta u_{t,k_t^*,k} > b_t\|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}} + 2\mathcal{R}(m)\right\}. \tag{D.8}$$

Then we prove the following lemma.

Lemma D.4. *Under any filtration $\mathcal{F}_{t-1}$ and E_t,*

$$\mathbb{P}\left(k_{t,2} \in [K] \setminus S_t | \mathcal{F}_{t-1}\right) \geq p - 1/t^2.$$

where $[K] \setminus S_t$ is a set of elements belonging to $[K]$ but not in S_t.

Proof. We first show that

$$\mathbb{P}(k_{t,2} \in [K] \setminus S_t | \mathcal{F}_{t-1}) \geq \mathbb{P}(\Delta\widetilde{u}_{t,k_t^*,k_{t,1}} > \Delta\widetilde{u}_{t,k,k_{t,1}},\ \forall k \in S_t | \mathcal{F}_{t-1}).$$

Let us suppose that $\widetilde{u}_{t,k_t^*} > \widetilde{u}_{t,k}$ for all $k \in S_t$. Note that the definition of $k_{t,2} = \arg\max_{k\in[K]} \Delta\widetilde{u}_{t,k,k_{t,1}}$. Then $k_{t,2} \notin S_t$ due to the property of S_t in Eq. (D.8). Accordingly, under F_t, for all $k \in S_t$, we have

$$\begin{aligned}
\Delta\widetilde{u}_{t,k,k_{t,1}} &\leq \Delta u_{t,k,k_{t,1}} + b_t\|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}} + 2\mathcal{R}(m) - \mathcal{R}(m) \\
&\leq \Delta u_{t,k,k_{t,1}} + \Delta u_{t,k_t^*,k} - \mathcal{R}(m) \text{ by Eq. (D.8)} \\
&= \Delta u_{t,k_t^*,k_{t,1}} - \mathcal{R}(m).
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{P}(k_{t,2} \in [K] \setminus S_t | \mathcal{F}_{t-1}) &\geq \mathbb{P}(\Delta\widetilde{u}_{t,k_t^*,k_{t,1}} > \Delta\widetilde{u}_{t,k,k_{t,1}} \forall k \in S_t | \mathcal{F}_{t-1}) \\ &\geq \mathbb{P}(\Delta\widetilde{u}_{t,k_t^*,k_{t,1}} > \Delta u_{t,k_t^*,k_{t,1}} - \mathcal{R}(m) | \mathcal{F}_{t-1}) - \mathbb{P}(F_t^c | \mathcal{F}_{t-1}) \\ &\geq \frac{1}{4e\sqrt{\pi}} - \frac{1}{t^2},\end{aligned}$$

where F_t^c is the complement of F_t. This completes the proof. ■

Lemma D.5. *Under any filtration $\mathcal{F}_{t-1}$ and E_t,*

$$\mathbb{E}[2r_t^a | \mathcal{F}_{t-1}] = \mathbb{E}\left[\|\Delta\phi_{t,k_{t,2},k_{t,1}}\|_{V_{t-1}^{-1}} | \mathcal{F}_{t-1}\right](4b_t/(p - 1/t^2) + 3b_t) + 9\mathcal{R}(m) + \frac{4}{t^2}.$$

Proof. Let us define $\overline{k}_t = \arg\min_{k \in [K] \setminus S_t} \|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}}$. Notice that a regret can be decomposed as follows.

$$r_t^a = 2\Delta u_{t,k_t^*,k_{t,2}} + \Delta u_{t,k_{t,2},k_{t,1}}.$$

As a result, under $E_t \bigcap F_t$,

$$\begin{aligned}\Delta u_{t,k_t^*,k_{t,2}} &= \Delta u_{t,k_t^*,\overline{k}_t} + \Delta u_{t,\overline{k}_t,k_{t,1}} + \Delta u_{t,k_{t,1},k_{t,2}} \\ &= \Delta u_{t,k_t^*,\overline{k}_t} + \Delta u_{t,\overline{k}_t,k_{t,1}} - \Delta u_{t,k_{t,2},k_{t,1}} \\ &\leq \Delta u_{t,k_t^*,\overline{k}_t} + \Delta\widetilde{u}_{t,\overline{k}_t,k_{t,1}} + b_t\|\Delta\phi_{t,\overline{k}_t,k_{t,1}}\|_{V_{t-1}^{-1}} + \mathcal{R}(m) \\ &\quad -\Delta\widetilde{u}_{t,k_{t,2},k_{t,1}} + b_t\|\Delta\phi_{t,k_{t,2},k_{t,1}}\|_{V_{t-1}^{-1}} + \mathcal{R}(m) \\ &\leq \Delta u_{t,k_t^*,\overline{k}_t} + b_t\|\Delta\phi_{t,\overline{k}_t,k_{t,1}}\|_{V_{t-1}^{-1}} + b_t\|\Delta\phi_{t,k_{t,2},k_{t,1}}\|_{V_{t-1}^{-1}} + 2\mathcal{R}(m).\end{aligned} \tag{D.9}$$

The first inequality is obtained by the definition of E_t and F_t in Eq. (D.4) and Eq. (D.5), respectively. The asymmetric arm selection to choose $(k_{t,1}, k_{t,2})$ is used to get the second inequality. Since $\overline{k}_t$ is not saturated,

$$\Delta u_{t,k_t^*,\overline{k}_t} \leq b_t\|\Delta\phi_{t,\overline{k}_t,k_{t,1}}\|_{V_{t-1}^{-1}} + \mathcal{R}(m). \tag{D.10}$$

Thus, combining Eq. (D.9) and Eq. (D.10), we can estimate the following inequality:

$$\Delta u_{t,k_t^*,k_{t,2}} \leq 2b_t\|\Delta\phi_{t,\overline{k}_t,k_{t,1}}\|_{V_{t-1}^{-1}} + b_t\|\Delta\phi_{t,k_{t,2},k_{t,1}}\|_{V_{t-1}^{-1}} + 3\mathcal{R}(m).$$

Furthermore, by the definition of E_t,

$$\begin{aligned}\Delta u_{t,k_{t,2},k_{t,1}} &\leq \theta_{t-1}^\intercal\Delta\phi_{t,k_{t,2},k_{t,1}} + a_t\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} + \mathcal{R}(m) \\ &\leq a_t\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} + \mathcal{R}(m).\end{aligned}$$

The second inequality results from

$$\theta_{t-1}^\intercal\Delta\phi_{t,k_{t,2},k_{t,1}} \leq 0,$$

by the asymmetry arm strategy, i.e., $k_{t,1} = \arg\max_{k \in [K]} \theta_{t-1}^\intercal\phi(x_{t,k}; W_{t-1})$. Thus,

$$\begin{aligned}r_t^a \leq 2\Big(&2b_t\|\Delta\phi_{t,\overline{k}_t,k_{t,1}}\|_{V_{t-1}^{-1}} + b_t\|\Delta\phi_{t,k_{t,2},k_{t,1}}\|_{V_{t-1}^{-1}} + 4\mathcal{R}(m)\Big) \\ &+a_t\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} + \mathcal{R}(m).\end{aligned}$$

Notice that

$$\begin{aligned}&\mathbb{E}[\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} | \mathcal{F}_{t-1}] \\ &\geq \mathbb{E}[\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} | \mathcal{F}_{t-1}, k_{t,2} \in [K] \setminus S_t] \times \mathbb{P}(k_{t,2} \in [K] \setminus S_t | \mathcal{F}_{t-1}) \\ &\geq \|\Delta\phi_{t,k_{t,1},\overline{k}_t}\|_{V_{t-1}^{-1}}(p - 1/t^2).\end{aligned}$$

Thus, we can estimate

$$\mathbb{E}[2r_t^a | \mathcal{F}_{t-1}] \leq \mathbb{E}\left[\|\Delta\phi_{t,k_{t,2},k_{t,1}}\|_{V_{t-1}^{-1}} | \mathcal{F}_{t-1}\right](4b_t/(p - 1/t^2) + 3b_t) + 9\mathcal{R}(m) + \frac{4}{t^2}. \tag{D.11}$$

due to $b_t \geq a_t$. This completes the proof. ■

Finally, we can show the following:

Lemma D.6. *Let* $Y_t = \sum_{i=1}^{t} X_i$ *where* $X_t = r_t^a I_{E_t} - c_t$ *with*

$$c_t = \frac{7b_t}{2p}\|\Delta\phi_{t,k_{t,2},k_{t,1}}\|_{V_{t-1}^{-1}} + \frac{9}{2}\mathcal{R}(m) + \frac{2}{t^2}. \tag{D.12}$$

Then Y_t *is a super-martingale with respect to the filtration* $\mathcal{F}_t$.

Proof. It is obvious that

$$\mathbb{E}[Y_t - Y_{t-1}|\mathcal{F}_{t-1}] = \mathbb{E}[X_t|\mathcal{F}_{t-1}] \leq 0, \tag{D.13}$$

since $\mathbb{E}[r(t)|\mathcal{F}_{t-1}] \leq c_t$ by Eq. (D.11). Due to Eq. (D.13), Y_t is a super-martingale under the filtration $\mathcal{F}_t$, thus this completes the proof. ■

Finally, by applying Lemma D.1 with the help of Lemma D.6, we complete the proofs of Theorem D.1 and Corollary D.1 under the asymmetric arm selection strategy.

TS-OSYM: Simultaneous arm selection. Unlike the asymmetric case where $k_{t,1}$ is chosen greedily and only $k_{t,2}$ is sampled, the optimistic symmetric strategy draws *both* arms simultaneously from a joint Gaussian posterior. This introduces a key technical challenge: we must show that the pair $(k_{t,1}, k_{t,2})$ avoids the saturation set $S_t \subset [K]^2$ with positive probability, which requires bounding a product of Gaussian tail probabilities. The proof exploits the independence of the two Gaussian draws conditioned on $\mathcal{F}_{t-1}$.

D.2 Optimistic Symmetric Arm Selection Strategy

Now, we consider the following strategy:

$$k_{t,1}, k_{t,2} = \underset{i,j\in[K]}{\arg\max}\left(\widetilde{u}_{t,i,j}\right). \tag{D.14}$$

Here,

$$\widetilde{u}_{t,i,j} \sim \mathcal{N}\left(\theta_{t-1}^{\intercal}\phi_{t,i,j}, a_t^2\|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}}^2\right),$$

where $\phi_{t,i,j} = \phi(x_{t,i}; W_{t-1}) + \phi(x_{t,j}; W_{t-1})$ as in Eq. (A.2). Let us define

$$\widetilde{u}_{t,i,j,1} = \widetilde{u}_{t,i,j} - \theta_{t-1}^{\intercal}\phi(x_{t,j}, W_{t-1}),$$
$$\widetilde{u}_{t,i,j,2} = \widetilde{u}_{t,i,j} - \theta_{t-1}^{\intercal}\phi(x_{t,i}, W_{t-1}).$$

We define the following sets:

$$S_t = \left\{(i,j) : \Delta u_{t,k_t^*,i} + \Delta u_{t,k_t^*,j} > b_t\|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}} + \mathcal{R}(m)\right\}, \tag{D.15}$$

$$E_t = \left\{(k_1,k_2) \in [K]^2 : |\Delta u_{t,k_1,k_2} - \theta_{t-1}^{\intercal}\Delta\phi_{t,k_1,k_2}| \leq a_t\|\Delta\phi_{t,k_1,k_2}\|_{V_{t-1}^{-1}} + \mathcal{R}(m)\right\}, \tag{D.16}$$

$$F_t = \Big\{x : x \sim \mathcal{N}\left(\theta_{t-1}^{\intercal}\phi_{t,i,j}, a_t^2\|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}}\right), \tag{D.17}$$
$$\text{such that } |x - \mathbb{E}[x]| \leq a_t\sqrt{2\log(Kt^2)}\|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}}) \text{ for all } i,j \in [K] \Big\},$$

with $b_t = a_t\sqrt{2\log(Kt^2)}$ and $\mathcal{R}(m)$ defined in Eq. (D.1).

Notice that in the saturation set S_t in Eq. (D.15), $(k_t^*, k_t^*) \notin S_t$ and $(k,k) \in S_t$ such that $k \neq k_t^*$ at round t for sufficiently large m since $\mathcal{R}(m)$ approaches to 0 as m increases. We need to show that E_t and F_t in Eq. (D.16) and Eq. (D.17) are good events.

Lemma D.7. *The following inequalities are satisfied:*

$$\mathbb{P}(E_t) \geq 1 - \delta, \tag{D.18}$$

$$\mathbb{P}(F_t) \geq 1 - \frac{1}{t^2}. \tag{D.19}$$

Proof. Eq. (D.18) is obtained by Eq. (D.1). In addition, since

$$\mathbb{P}(\widehat{\mu} - \mu \geq \epsilon) \leq \frac{1}{2}\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right),$$

by setting $\epsilon = a_t\sqrt{2\log(Kt^2)}\|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}}$, we can show that Eq. (D.19). ■

Then we prove the following lemma.

Lemma D.8. *For any filtration $\mathcal{F}_{t-1}$, under the condition F_t, there exists a strictly positive p such that*

$$\mathbb{P}((k_{t,1}, k_{t,2}) \in [K]^2 \setminus S_t | \mathcal{F}_{t-1}) \geq p - 1/t^2.$$

Proof. Due to the definition of $\widetilde{u}_{i,k_{t,1},k_{t,2}}$, the following is satisfied:

$$\mathbb{P}((k_{t,1}, k_{t,2}) \in ([K] \setminus S_t)^2 | \mathcal{F}_{t-1}) \geq \mathbb{P}\left(\widetilde{u}_{t,k_t^*,i} + \widetilde{u}_{t,k_t^*,j} \geq 2\widetilde{u}_{t,i,j}, \forall (i,j) \in S_t | \mathcal{F}_{t-1}\right). \tag{D.20}$$

We can rewrite Eq. (D.20) as

$$\begin{aligned}
&\mathbb{P}\left(\widetilde{u}_{t,k_t^*,i} + \widetilde{u}_{t,k_t^*,j} \geq 2\widetilde{u}_{t,i,j}, \forall (i,j) \in S_t | \mathcal{F}_{t-1}\right) \\
&= \mathbb{P}\left(\widetilde{u}_{t,k_t^*,i} + \widetilde{u}_{t,k_t^*,j} - 2\theta_{t-1}^\intercal \Delta\phi_{t,i,j} \geq 2\widetilde{u}_{t,i,j} - 2\theta_{t-1}^\intercal \Delta\phi_{t,i,j}, \forall (i,j) \in S_t | \mathcal{F}_{t-1}\right).
\end{aligned}$$

and

$$\begin{aligned}
\widetilde{u}_{t,i,j} - \theta_{t-1}^\intercal \phi_{t,i,j} \leq b_t\|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}} \leq \frac{\Delta u_{t,k_t^*,i} + \Delta u_{t,k_t^*,j}}{2} \text{ by Eq. (D.17) and Eq. (D.15)}, \\
\widetilde{u}_{t,k_t^*,i} + \widetilde{u}_{t,k_t^*,j} - 2\theta_{t-1}^\intercal \phi_{t,i,j} \sim \mathcal{N}(\theta_{t-1}^\intercal\left(\Delta\phi_{t,k_t^*,i} + \Delta\phi_{t,k_t^*,j}\right), \|\Delta\phi_{t,k_t^*,i}\|_{V_{t-1}^{-1}}^2 + \|\Delta\phi_{t,k_t^*,j}\|_{V_{t-1}^{-1}}^2),
\end{aligned}$$

under F_t. Thus, in a similar manner to Lemma D.3, we can show that there is a strictly positive probability $p > 0$, e.g. $\left(\frac{1}{4e\sqrt{\pi}}\right)^2$, such that

$$\begin{aligned}
&\mathbb{P}\left(\widetilde{u}_{t,k_t^*,i} + \widetilde{u}_{t,k_t^*,j} \geq 2\widetilde{u}_{t,i,j}, \forall (i,j) \in S_t | \mathcal{F}_{t-1}\right) \\
&\geq \mathbb{P}^2\left(x \geq \mu : x \sim \mathcal{N}(\mu, a_t^2(\|\Delta\phi_{t,k_t^*,i}\|_{V_{t-1}^{-1}}^2 + \|\Delta\phi_{t,k_t^*,j}\|_{V_{t-1}^{-1}}^2))\right) \geq p > 0,
\end{aligned}$$

where $\mu := \theta_{t-1}^\intercal(\Delta\phi_{t,k_t^*,i} + \Delta\phi_{t,k_t^*,j})$. As a result,

$$\begin{aligned}
\mathbb{P}((k_{t,1}, k_{t,2}) \in [K] \times [K] \setminus S_t | \mathcal{F}_{t-1}) &\geq \mathbb{P}\left(\widetilde{u}_{t,k_t^*,i} + \widetilde{u}_{t,k_t^*,j} \geq 2\widetilde{u}_{t,i,j}, \forall (i,j) \in S_t | \mathcal{F}_{t-1}\right) \\
&\quad - \mathbb{P}(\overline{F_t} | \mathcal{F}_{t-1}) \\
&\geq p - \frac{1}{t^2}.
\end{aligned}$$

This completes the proof. ■

Lemma D.9. *For any filtration $\mathcal{F}_{t-1}$,*

$$\mathbb{E}[2r_t | \mathcal{F}_{t-1}] = b_t(6p+3)\mathbb{E}\left[\|\Delta\phi_{k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}} \mid \mathcal{F}_{t-1}\right] + 4\mathcal{R}(m) + \frac{4}{t^2}.$$

Proof. Notice that

$$2r_t = 2h_{t,k_t^*} - h_{t,k_{t,1}} - h_{t,k_{t,2}}. \tag{D.21}$$

Let $(\overline{k}_1, \overline{k}_2) \in [K]^2 \setminus (S_t)$ such that

$$\begin{aligned}
k_1 = \arg\max_{k \in [K]} \|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}} \leq \|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}, \\
k_2 = \arg\max_{k \in [K]} \|\Delta\phi_{t,k,k_{t,2}}\|_{V_{t-1}^{-1}} \leq \|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}[\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}|\mathcal{F}_{t-1}] \\
&\geq \mathbb{E}\left[\frac{\|\Delta\phi_{k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}+\|\Delta\phi_{k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}}{2}|\mathcal{F}_{t-1},(k_{t,1},k_{t,2})\in[K]\setminus S_t\right] \\
&\quad\times\mathbb{P}\left[(k_{t,1},k_{t,2})\in[K]\setminus S_t|\mathcal{F}_{t-1}\right] \\
&\geq (\|\Delta\phi_{t,\overline{k}_1,k_{t,1}}\|_{V_{t-1}^{-1}}+\|\Delta\phi_{t,\overline{k}_2,k_{t,2}}\|_{V_{t-1}^{-1}})\left(p-\frac{1}{t^2}\right) \qquad \text{(D.22)}
\end{aligned}$$

Note that a tuple (k_1,k_2) exists in $[K]^2\setminus S_t$ since $(k_t^*,k_t^*)\in[K]^2\setminus S_t\neq\emptyset$. We can decompose Eq. (D.21) as

$$\begin{aligned}
2r_t = h_{t,k_t^*}-h_{t,\overline{k}_2}+h_{t,\overline{k}_2}-h_{t,k_{t,1}} \\
+h_{t,k_t^*}-h_{t,\overline{k}_1}+h_{t,\overline{k}_1}-h_{t,k_{t,2}}.
\end{aligned}$$

The fact $(\overline{k}_1,\overline{k}_2)\in[K]^2\setminus S_t$ results in

$$|h_{t,k_t^*}-h_{t,\overline{k}_1}+h_{t,k_t^*}-h_{t,\overline{k}_2}|\leq 2b_t\|\Delta\phi_{\overline{k}_1,\overline{k}_2}\|_{V_{t-1}^{-1}}+2\mathcal{R}(m).$$

Furthermore,

$$\begin{aligned}
&\Delta u_{t,\overline{k}_1,k_{t,1}}+\Delta u_{t,\overline{k}_2,k_{t,2}} \\
&\leq \theta_{t-1}^{\intercal}\Delta\phi_{t,\overline{k}_1,k_{t,1}}+\theta_{t-1}^{\intercal}\Delta\phi_{t,\overline{k}_2,k_{t,2}}+2b_t\|\Delta\phi_{\overline{k}_1,k_{t,1}}\|_{V_{t-1}^{-1}}+\|\Delta\phi_{\overline{k}_2,k_{t,2}}\|_{V_{t-1}^{-1}} \\
&\leq \widetilde{u}_{t,\overline{k}_1,\overline{k}_2}-\widetilde{u}_{t,k_{t,1},k_{t,2}} \\
&\quad+2b_t\left(\|\Delta\phi_{\overline{k}_1,k_{t,1}}\|_{V_{t-1}^{-1}}+\|\Delta\phi_{\overline{k}_2,k_{t,2}}\|_{V_{t-1}^{-1}}+\|\Delta\phi_{k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}+\|\Delta\phi_{\overline{k}_1,\overline{k}_2}\|_{V_{t-1}^{-1}}\right)+2\mathcal{R}(m).
\end{aligned}$$

Then we can derive the following by the triangular inequality:

$$\|\Delta\phi_{\overline{k}_1,\overline{k}_2}\|_{V_{t-1}^{-1}}\leq\|\Delta\phi_{\overline{k}_1,k_{t,1}}\|_{V_{t-1}^{-1}}+\|\Delta\phi_{k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}+\|\Delta\phi_{k_{t,2},\overline{k}_2}\|_{V_{t-1}^{-1}}.$$

By the definition of $k_{t,1},k_{t,2}$,

$$\widetilde{u}_{t,\overline{k}_1,\overline{k}_2}-\widetilde{u}_{t,k_{t,1},k_{t,2}}<0.$$

Thus, we can obtain

$$2r_t\leq b_t\left(6\left(\|\Delta\phi_{\overline{k}_1,k_{t,1}}\|_{V_{t-1}^{-1}}+\|\Delta\phi_{\overline{k}_2,k_{t,2}}\|_{V_{t-1}^{-1}}\right)+3\|\Delta\phi_{k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}\right)+4\mathcal{R}(m).$$

Using Eq. (D.22), we can conclude that under Assumption 1,

$$\begin{aligned}
2\mathbb{E}[r_t|\mathcal{F}_{t-1}] &\leq b_t\left(6\left(p-\frac{1}{t^2}\right)^{-1}+3\right)\mathbb{E}\left[\|\Delta\phi_{k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}\mid\mathcal{F}_{t-1}\right]+4\mathcal{R}(m)+\frac{4}{t^4} \qquad \text{(D.23)} \\
&\leq b_t(6p+3)\mathbb{E}\left[\|\Delta\phi_{k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}\mid\mathcal{F}_{t-1}\right]+4\mathcal{R}(m)+\frac{4}{t^2}.
\end{aligned}$$

This completes the proof. ■

After the above, as the same manner to the case of Lemma D.6 using Eq. (D.23), we can obtain the following lemma.

Lemma D.10. *Let $Y_t=\sum_{i=1}^t X_i$ where $X_t=r_tI_E-c_t$ with*

$$c_t=\frac{b_t(6p+3)}{2}\|\Delta\phi_{k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}+2\mathcal{R}(m)+\frac{2}{t^2}$$

Then (Y_t) is a super-martingale with respect to the filtration $\mathcal{F}_t$.

Combining Lemma D.1 and Lemma D.10, we complete the proofs of Theorem D.1 and Corollary D.1 under the optimistic symmetric arm selection strategy.

TS-CSYM: Confidence set filtering + stochastic selection. This strategy first constructs a confidence set $\mathcal{C}_t$ of "plausibly optimal" arms, then selects the pair $(k_{t,1}, k_{t,2})$ from $\mathcal{C}_t \times \mathcal{C}_t$ by sampling exploration bonuses from Gaussian distributions. The proof follows the same template as TS-OSYM but restricts attention to arms within $\mathcal{C}_t$, leveraging the property that the optimal arm $k_t^* \in \mathcal{C}_t$ with high probability (when a_t is chosen appropriately).

D.3 Candidate-based Symmetric Arm Selection Strategy

Let us define

$$\mathcal{C}_t = \left\{ i : \theta_{t-1}^\intercal \phi(x_{t,i}; W_{t-1}) - \theta_{t-1}^\intercal \phi(x_{t,j}; W_{t-1}) + a_t \|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}} \geq 0, \forall j \in [K] \right\}.$$

We choose $(k_{t,1}, k_{t,2})$ as follows:

$$\begin{aligned}(k_{t,1}, k_{t,2}) = &\operatorname*{arg\,max}_{(i,j)\in\mathcal{C}_t\times\mathcal{C}_t} \widehat{\sigma}_{i,j}, \\ \widehat{\sigma}_{i,j} \sim &\mathcal{N}\left(\|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}}^2, \frac{\|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}}^4}{4\log(Kt^2)} \right).\end{aligned}$$

We define

$$\begin{aligned}\widetilde{u}_{t,i,j} &= \theta_{t-1}\phi(x_{t,i}; W_{t-1}) + \theta_{t-1}\phi(x_{t,j}; W_{t-1}) + a_t\widehat{\sigma}_{i,j}, \\ \widetilde{u}_{t,i,j,1} &= \widetilde{u}_{t,i,j} - \theta_{t-1}\phi(x_{t,j}; W_{t-1}), \\ \widetilde{u}_{t,i,j,2} &= \widetilde{u}_{t,i,j} - \theta_{t-1}\phi(x_{t,i}; W_{t-1}),\end{aligned}$$

and define

$$\begin{aligned}S_t &= \{i \in \mathcal{C}_t : \Delta u_{t,k_t^*,i} > b_t \|\Delta\phi_{t,k_t^*,i}\|_{V_{t-1}^{-1}} + \mathcal{R}(m)\}, \\ E_t &= \{(k_1,k_2) \in [K]^2 : |\Delta u_{t,k_1,k_2} - \theta_{t-1}^\intercal \Delta\phi_{t,k_1,k_2}| \leq a_t \|\Delta\phi_{t,k_1,k_2}\|_{V_{t-1}^{-1}} + \mathcal{R}(m)\}, \\ F_t &= \{x : x \sim \mathcal{N}(\mu, a^2\|\Delta\phi_{i,j}\|_{V_{t-1}^{-1}}), i,j \in [K] : |x - \mathbb{E}[x]| \leq a_t\sqrt{2\log(Kt^2)}\|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}})\},\end{aligned}$$

where $b_t = a_t(1 + \sqrt{2\log(Kt^2)})$. Applying the Hoeffding inequality (Lattimore and Szepesvári, 2020), in F_t, we can show that the following holds:

$$\frac{1}{2}\|\Delta\phi_{t,i,j}\|_{V_{t-1}^{-1}} \leq \frac{3}{2}\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}. \tag{D.24}$$

We show that E_t and F_t are good events.

Lemma D.11. *The following inequalities are satisfied:*

$$\mathbb{P}(E_t) \geq 1 - \delta, \tag{D.25}$$

$$\mathbb{P}(F_t) \geq 1 - \frac{1}{t^2}. \tag{D.26}$$

Proof. Eq. (D.25) is obtained by Eq. (D.1). Since

$$\mathbb{P}(\widehat{\mu} - \mu \geq \epsilon) \leq \frac{1}{2}\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right), \tag{D.27}$$

we get Eq. (D.26) in a similar manner to Lemma D.3 by setting the confidence radius as $a_t\sqrt{2\log(Kt^2)}\|\Delta\phi_{t,k,k_{t,1}}\|_{V_{t-1}^{-1}}\}$ with the sample size 1. ∎

Lemma D.12. *For any filtration $\mathcal{F}_{t-1}$, under F,*

$$\mathbb{E}[2r_t|\mathcal{F}_{t-1}] = C(b_t\mathbb{E}[\|\Delta\phi_{t,t}\|_{V_{t-1}^{-1}}|\mathcal{F}_{t-1}] + \mathcal{R}(m) + \frac{1}{t^2}) \tag{D.28}$$

for some constant $C > 0$.

Proof. Notice that

$$2r_t = \Delta u_{t,k_t^*,k_{t,1}} + \Delta u_{t,k_t^*,k_{t,2}}.$$

We can estimate that

$$\Delta\phi_{t,k_t^*,k_{t,1}} = \theta_{t-1}\Delta\phi_{t,k_t^*,k_{t,1}} + b_t\|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}} + \mathcal{R}(m), \tag{D.29}$$

$$\Delta\phi_{t,k_t^*,k_{t,2}} = \theta_{t-1}\Delta\phi_{t,k_t^*,k_{t,2}} + b_t\|\Delta\phi_{t,k_t^*,k_{t,2}}\|_{V_{t-1}^{-1}} + \mathcal{R}(m), \tag{D.30}$$

in E_t and F_t. Since $k_{t,1}, k_{t,2} \in \mathcal{C}_t$,

$$\theta_{t-1}\Delta\phi_{t,k_t^*,k_{t,1}} \le b_t\|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}} + \mathcal{R}(m),$$
$$\theta_{t-1}\Delta\phi_{t,k_t^*,k_{t,2}} \le b_t\|\Delta\phi_{t,k_t^*,k_{t,2}}\|_{V_{t-1}^{-1}} + \mathcal{R}(m).$$

in E_t and F_t. Thus, under Assumption 1, we have

$$\begin{aligned}\mathbb{E}[2r_t|\mathcal{F}_{t-1}] &\le \frac{4}{t^2} + 4\mathcal{R}(m) + 2b_t\mathbb{E}[\|\Delta\phi_{t,k_t^*,k_{t,1}}\|_{V_{t-1}^{-1}} \mid \mathcal{F}_t, E_t \bigcap F_t] \\ &\quad + 2b_t\mathbb{E}[\|\Delta\phi_{t,k_t^*,k_{t,2}}\|_{V_{t-1}^{-1}}|\mathcal{F}_t, E_t \bigcap F_t] \\ &\le \frac{4}{t^2} + 4\mathcal{R}(m) + 12b_t\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}.\end{aligned} \tag{D.31}$$

The last inequality is obtained by Eq. (D.24). This completes the proof. ■

Utilizing Eq. (D.31) above, we can show the following lemma as Lemma D.6:

Lemma D.13. *Let $Y_t = \sum_{i=1}^t X_i$ where $X_t = r_t I_E - c_t$ with*

$$2c_t = \frac{4}{t^2} + 4\mathcal{R}(m) + 12b_t\|\Delta\phi_{t,k_{t,1},k_{t,2}}\|_{V_{t-1}^{-1}}.$$

Then (Y_t) is a super-martingale with respect to the filtration $\mathcal{F}_t$.

Therefore, combining Lemma D.13 and Lemma D.1, we complete the proofs of Theorem D.1 and Corollary D.1 under the candidate-based symmetric arm selection strategy.

E ADDITIONAL EXPERIMENTS AND IMPLEMENTATION DETAILS

E.1 Datasets

E.1.1 Synthetic Dataset

We consider the following three synthetic tasks having nonlinear utilities u:

- **Cosine:** $u(\mathbf{x}) = \cos(3\boldsymbol{\Theta}^\intercal\mathbf{x})$
- **Square:** $u(\mathbf{x}) = 10(\boldsymbol{\Theta}^\intercal\mathbf{x})^2$
- **Quadratic:** $u(\mathbf{x}) = \mathbf{x}^\intercal(\boldsymbol{\Theta}\boldsymbol{\Theta}^\intercal)\mathbf{x}$

with unknown parameters $\mathbf{\Theta} \in \mathbb{R}^d$.

These tasks are commonly used as synthetic datasets in both dueling bandits and multi-armed bandits (Verma et al., 2024; Zhang et al., 2021a; Zhou et al., 2020; Dai et al., 2022). For each task, the environment first draws a parameter vector $\mathbf{\Theta}$ from the uniform distribution $\mathcal{U}([-1,1]^d)$. Then, in each round t, it samples a set of context vectors

$$\mathcal{X}_t = \{\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,K}\},$$

where each $\mathbf{x}_{t,k}$ is drawn independently from $\mathcal{U}([-1,1]^d)$ for $k \in [K]$. Finally, each $\mathbf{x}_{t,k}$ is normalized by its Euclidean norm, $\|\mathbf{x}_{t,k}\|_2$. The following code provides how to generate $\left\{\{\mathbf{x}_{t,k}\}_{k=1}^K\right\}_{t=1}^T$.

```
context_arms_list = []
for _ in range(T):
    # Generate all context_arm pairs
    context_arms = np.random.uniform(
        low=-1, high=1, size=(self.arms, self.dim)
    )
    context_arms = context_arms / np.linalg.norm(
        context_arms, 2, axis=1
    ).reshape(-1, 1)
    context_arms_list.append(context_arms)
```

Listing 1: Pseudocode for constructing context sets X_t.

E.1.2 Real-World Dataset

To verify the effectiveness of our method, we consider real-world datasets provided by the **UCI Machine Learning Repository**[2]—Statlog, Magic, and Covertype—each consisting of feature vectors associated with one of multiple possible actions (arms) (see Table E.1).

To the best of our knowledge, there is no standard experimental setup for dueling bandits using UCI datasets (unlike for multi-armed bandits (Xu et al., 2020)); thus, we designed the tasks as follows. For the **Magic** and **Covertype** datasets, we defined the pairwise preference probability using a sigmoid function:

$$P(i \succ j) = \sigma(u_i - u_j) = \frac{1}{1 + e^{-(u_i - u_j)}}$$

where u_k represents the underlying true utility value corresponding to category index $k \in [K]$ in each round, and we set a baseline of $P(i \succ j) = 0.7$ for clarity, though the preference is derived from the utility difference $u_i - u_j$.

Note that variance-aware algorithms are not expected to consistently outperform variance-agnostic algorithms in normal tasks, as their advantages are primarily realized in scenarios with very low variances. Thus, to specifically evaluate performance under **low-variance conditions**, we set a deterministic preference model for the **Statlog** dataset:

$$P(i \succ j) = 1 \quad \text{for } i > j$$

This setup ensures the advantage of a higher-indexed arm is realized with certainty.

Table E.1: UCI benchmark datasets used for contextual dueling bandit experiments.

Dataset	Number of Attributes	Number of Arms	Number of Instances
Statlog	9	7	58,000
Magic	11	2	19,020
Covertype	54	7	581,012

[2]The UCI Machine Learning Repository, developed by the University of California, Irvine, is a widely used collection of datasets for empirical evaluation and benchmarking of machine learning algorithms. Available at https://archive.ics.uci.edu/.

E.2 Implementation Details

E.2.1 Hyperparameters.

We report the hyperparameters used in our main manuscript. To evaluate NVLDB, we employ a fully connected deep neural network (DNN) with ReLU activations. The architecture consists of $L = 2$ hidden layers, each with $W = 32$ units, and an output feature dimension of D.

We set the following training parameters:

- Regularization parameter: $\lambda = 1.0$
- Number of gradient steps per round: $n = 20$
- Number of episodes: $\mathcal{H} = 1$

We set the **confidence width coefficient** $a_t = 1.0$. Note that this constant value is adopted from previous works (Verma et al., 2024; Xu et al., 2020; Zhou et al., 2020; Bui et al., 2024) because computing the exact theoretical value of a_t is challenging in the context of neural networks. Specifically, this difficulty arises from the non-trivial computation of the differential term $\|\mathbf{u} - \widetilde{\mathbf{u}}\|_{\mathbf{H}^{-1}}$. Implementing an exact algorithm to find the parameters θ_t that satisfy the first-order optimality condition $\frac{\partial \mathcal{L}_t}{\partial \theta}(\theta_t, W_t) = 0$ is infeasible. Therefore, we approximate the parameter set (θ_t, W_t) via gradient descent on the loss function $\mathcal{L}_t$ using a learning rate of $\eta = 0.01$ at each round t. Optimization is performed using the Adam optimizer in PyTorch, utilizing its default parameters except for the specified learning rate. Additionally, we reinitialize the parameter set V_{t-1} to its initial state at the beginning of each round, as this was found not to incur significant computational overhead. Each experiment is run across **20 random seeds**. Unless otherwise stated, the remaining sections in this supplementary material adhere to the hyperparameters described above.

E.2.2 Pseudocodes of Arm Selection Strategies

The following are the pseudocodes for the asymmetric, optimistic symmetric, and candidate-based symmetric arm selection strategies, respectively.

```
utilities = neural_network(context_actions)

# Selecting the first arm
at_1 = torch.argmax(utilities).item()
at_2 = 1
max_score = -np.inf
for j in range(len(utilities)):
    # Difference of features of the selected arm and the current arm
    zt_j1 = grad_list[j] - grad_list[at_1]
    zt_j1 = zt_j1.to("cpu")
    zt_dot_U = torch.matmul(zt_j1, torch.inverse(self.V))
    zt_dot_U_zt = torch.matmul(zt_dot_U, zt_j1.t())
    conf_term = torch.clamp(zt_dot_U_zt, min=0)
    sigma = self.nu * torch.sqrt(conf_term)

    # Selecting the arm based on the strategy
    utilities_j1 = utilities[j]
    if self.strategy == "ts":
        utilities_j1 = utilities_j1.to("cpu")
        action_score = torch.normal(utilities_j1.view(-1), sigma.view(-1))

        # Alternative: np.random.normal(loc=utilities_j1.item(), scale=sigma.item())

    elif self.strategy == "ucb":
        action_score = utilities_j1.item() + sigma.item()

    else:
        raise RuntimeError("Exploration strategy not set")

    # Selecting the second best arm
```

```
    if action_score > max_score:
        max_score = action_score
        at_2 = j
# Keeping the selected arms for model update
return at_1, at_2
```

Listing 2: Pseudocode for the asymmetric arm selection in both UCB and TS frameworks as in (Verma et al., 2024)

```
utilities = neural_network(context_actions)
at_1, at_2 = 0, 0
max_val = -torch.inf
if self.strategy == "ucb":
    for i in range(len(utilities)):
        for j in range(len(utilities)):
            zt = grad_list[i] - grad_list[j]
            zt_dot_U = torch.matmul(zt, torch.inverse(self.V))
            zt_dot_U_zt = torch.matmul(zt_dot_U, zt.t())
            conf_term = torch.clamp(zt_dot_U_zt, min=0)
            sigma = self.nu * torch.sqrt(conf_term)
            estimated = utilities[i] + utilities[j] + sigma
            if max_val < estimated: # Find best optimistic arms
                at_1, at_2 = i, j
                max_val = estimated
elif self.strategy == "ts":
    mus = torch.zeros((len(utilities), len(utilities)))
    sigmas = torch.zeros((len(utilities), len(utilities)))
    for i in range(len(utilities)):
        for j in range(len(utilities)):
            mus[i, j] = utilities[i] + utilities[j]
            zt = grad_list[i] - grad_list[j]
            zt_dot_U = torch.matmul(zt, torch.inverse(self.V))
            zt_dot_U_zt = torch.matmul(zt_dot_U, zt.t())
            conf_term = torch.clamp(zt_dot_U_zt, min=0)
            sigma = self.nu * torch.sqrt(conf_term)
            sigmas[i, j] = sigma
    sigmas = sigmas + 1e-12
    mus = mus.detach().cpu().numpy()
    sigmas = sigmas.detach().cpu().numpy()

    # Sample best optimistic arms
    sample = np.random.normal(loc=mus, scale=sigmas)
    row_index, col_index = np.where(sample == np.max(sample))
    at_1, at_2 = row_index.item(), col_index.item()
```

Listing 3: Pseudocode for the optimistic symmetric arm selection in both UCB and TS frameworks

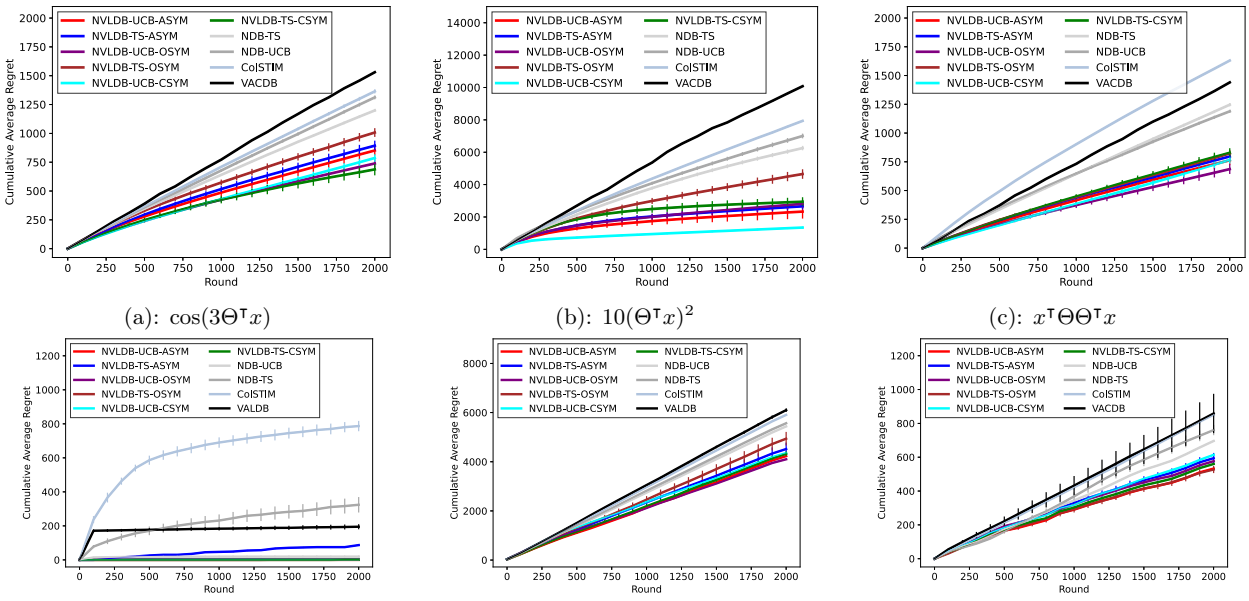
```
utilities = neural_network(context_actions)
at_1, at_2 = 0, 0
if self.strategy == "ucb":
    sigmas = np.zeros((len(context_actions), len(context_actions)))
    mask = np.zeros((len(context_actions), len(context_actions)))
    for i in range(len(utilities)):
        for j in range(len(utilities)):
            if i == j:
                mask[i, j] = 1
                sigmas[i, j] = 0
                continue
            zt = grad_list[i] - grad_list[j]
            zt_dot_U = torch.matmul(zt, torch.inverse(self.V))
            zt_dot_U_zt = torch.matmul(zt_dot_U, zt.t())
            conf_term = torch.clamp(zt_dot_U_zt, min=0)
            sigma = self.nu * torch.sqrt(conf_term)
            sigmas[i, j] = sigma
            if utilities[i] + sigma > utilities[j]:
                mask[i, j] = 1
    # Make a candidate set
```

```
    candidate_mask = mask.sum(axis=1) == len(utilities)
    candidate_mask = candidate_mask.reshape(-1, 1) * candidate_mask.reshape(1, -1)
    sigmas = sigmas + 1

    at_1, at_2 = np.unravel_index(
        np.argmax(candidate_mask * sigmas, axis=None),
        (len(context_actions), len(context_actions)),
    )
elif self.strategy == "ts":
    sigmas = torch.zeros((len(context_actions), len(context_actions)))
    mask = torch.zeros((len(context_actions), len(context_actions)))
    for i in range(len(utilities)):
        for j in range(len(utilities)):
            if i == j:
                mask[i, j] = 1
                sigmas[i, j] = 0
                continue
            zt = grad_list[i] - grad_list[j]
            zt_dot_U = torch.matmul(zt, torch.inverse(self.V))
            zt_dot_U_zt = torch.matmul(zt_dot_U, zt.t())
            conf_term = torch.clamp(zt_dot_U_zt, min=0)
            sigma = self.nu * torch.sqrt(conf_term)
            sigmas[i, j] = sigma
            if utilities[i] + sigma > utilities[j]:
                mask[i, j] = 1

    # Make a candidate set
    candidate_mask = mask.sum(axis=1) == len(utilities)
    candidate_mask = candidate_mask.reshape(-1, 1) * candidate_mask.reshape(1, -1)
    candidate_mask = candidate_mask.cpu().numpy()

    sigmas = sigmas.detach().cpu().numpy()
    mus = sigmas
    sigmas = sigmas * sigmas / 4 + 1e-12

    sample = np.random.normal(loc=mus, scale=sigmas)
    sample[candidate_mask == False] = -np.inf
    row_index, col_index = np.where(sample == np.max(sample))
    at_1, at_2 = row_index.item(), col_index.item()
```

Listing 4: Pseudocode for the candidate-based symmetric arm selection in both UCB and TS frameworks

E.3 Additional Experiments

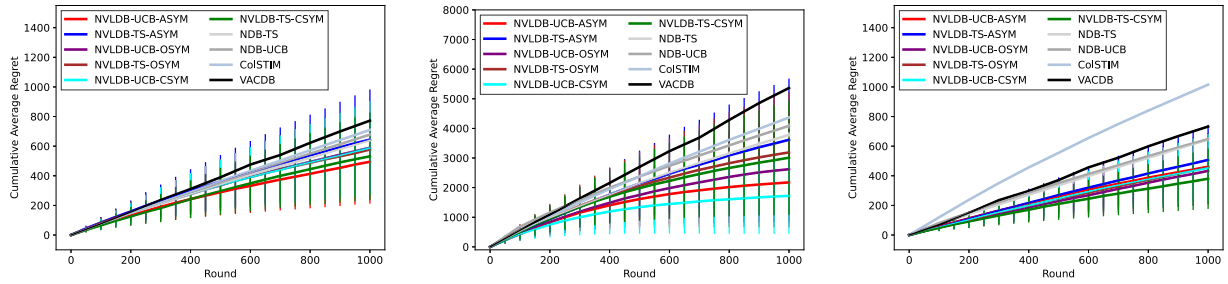
In this section, we conduct additional experiments supporting our proposed method NVLDB.



(a): $\cos(3\Theta^\intercal x)$ (b): $10(\Theta^\intercal x)^2$ (c): $x^\intercal\Theta\Theta^\intercal x$

(d): UCI-statlog (e): UCI-covertype (f): UCI-magic

Figure E.1: Comparison of cumulative average regret under (a)-(c): Synthetic tasks with context dimension $d = 5$ and $K = 5$ arms. (d)-(f): UC Irvine (UCI) machine learning repository data. All variants of NVLDB mostly outperform other baselines including NDB in both the UCB and TS frameworks. Each experiment is repeated across 20 random seeds.



(a): $\cos(3\Theta^\intercal x)$ (b): $10(\Theta^\intercal x)^2$ (c): $x^\intercal\Theta^\intercal\Theta x$

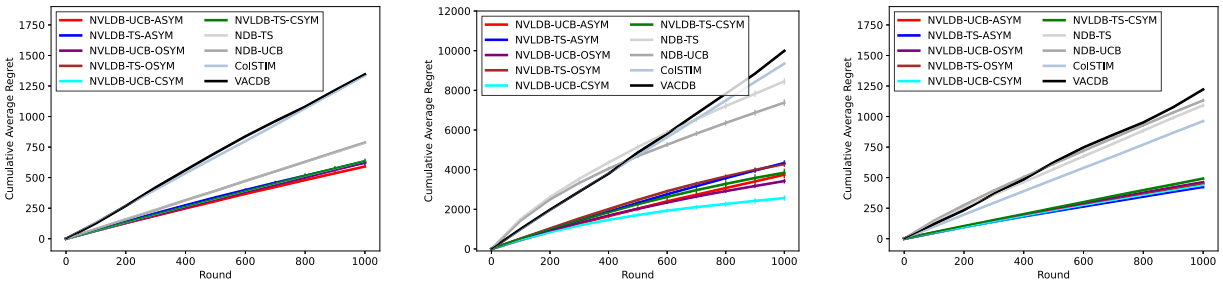
Figure E.2: Comparison of cumulative average regret with increased network width. The experiments utilize an increased neural network width of $m = 100$ (compared to the standard $m = 32$ reported elsewhere), while the dimension is set to $d = 5$ and the number of arms to $K = 5$ for all tasks. The results demonstrate that our methods consistently outperform NDB and achieve sublinear cumulative regret. Each experiment is repeated across 20 random seeds.

Comparison of Cumulative Average Regret on Synthetic and Real-World Datasets. While the main manuscript presents comparisons of cumulative average regret on synthetic datasets with nonlinear utility functions, we additionally provide results on real-world datasets from the UCI Machine Learning Repository, as described in Section E.1.2. Figure E.1 illustrates the cumulative average regret of NVLDB's variance-aware variants compared to baseline algorithms.In particular, when comparing our methods against established baselines, including linear dueling bandit algorithms (VALDB (Di et al., 2023) and ColSTIM (Bengs et al., 2022)) and the neural dueling bandit (NDB) algorithm (Verma et al., 2024), it is noteworthy that all our proposed arm selection strategies consistently outperform these baselines. Furthermore, our methods exhibit sublinear cumulative regrets across both the synthetic and real-world datasets. These empirical results strongly support our theoretical

findings, as established in Theorem C.1 and Theorem D.1. Each experiment was conducted using an average over 20 random seeds.

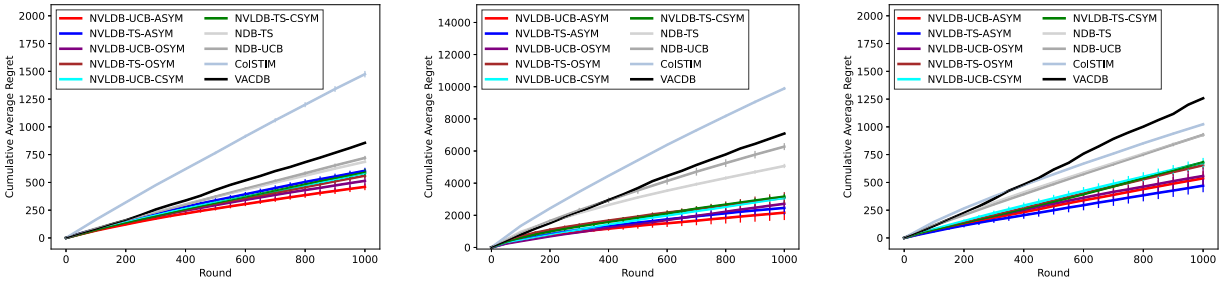
The Size of Neural Networks. To further demonstrate that NVLDB guarantees sublinear cumulative average regret for each task, we increase the neural network width from $m = 32$ to $m = 100$. Based on our theoretical results (Theorem C.1 and Theorem D.1), we anticipate that the average cumulative regret will remain sublinear for the wider network ($m = 100$), given that it already exhibited sublinear behavior at $m = 32$. Figure E.2 illustrates the cumulative average regrets of NVLDB under different arm selection strategies in comparison to the baselines. As theoretically and empirically expected, in all presented cases, the variants of NVLDB consistently outperform the baselines and continue to show sublinear regret behavior. Each experiment was conducted using an average over 20 random seeds.

The Environment Condition. As we stated in the main manuscript, Theorem C.1 and Theorem D.1 denote that the cumulative average regret should be sublinear even if the environment parameters K and d are changed. Inspired by this, we conduct experiments adjusting these parameters. First, we increase the context dimension d from 5 to 10 while keeping all other settings unchanged. Figure E.3 illustrates the performance on tasks with $d = 10$ and $K = 5$, revealing substantial gaps between the NVLDB variants and the baselines. Additionally, Figure E.4 demonstrates the performance on tasks with $d = 5$ and $K = 10$. These results indicate that NVLDB consistently outperforms the baselines across different task configurations, supporting our theoretical results. Each experiment was conducted using an average over 20 random seeds.



(a): $\cos(3\Theta^\intercal x)$ (b): $10(\Theta^\intercal x)^2$ (c): $x^\intercal \Theta \Theta^\intercal x$

Figure E.3: Comparison of cumulative average regret across our methods (NVLDB variants) and the other baselines. Each task uses $d = 10$ and $K = 5$. It is clear that our variants outperform the other baselines. Each experiment was conducted across 20 random seeds.



(a): $\cos(3\Theta^\intercal x)$ (b): $10(\Theta^\intercal x)^2$ (c): $x^\intercal \Theta \Theta^\intercal x$

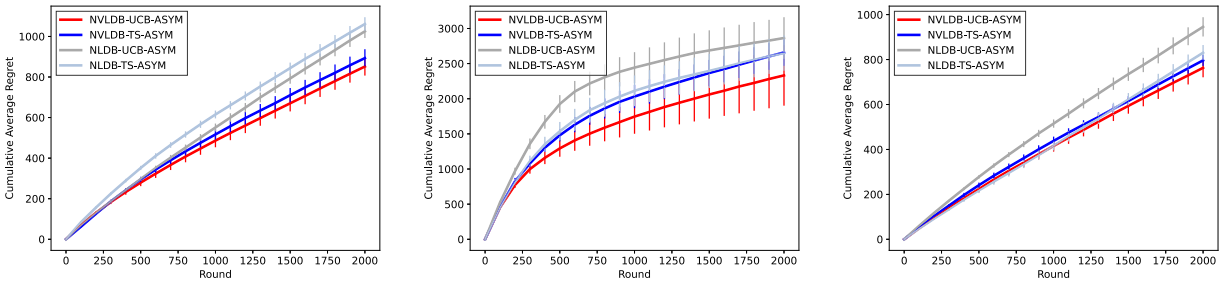
Figure E.4: Comparison of cumulative average regret across our methods (NVLDB variants) and the other baselines. Each task uses $d = 5$ and $K = 10$. It is clear that our variants outperform the other baselines. Each experiment was conducted across 20 random seeds.

Variance-agnostic Variants As stated in the main manuscript, Corollary C.1 and Corollary D.1 indicate that, even without incorporating variance-awareness, the cumulative average regret of our arm selection strategies remains theoretically sublinear. To validate these theoretical results empirically, we conduct experiments

comparing the performance of the variance-aware (NVLDB) and variance-agnostic variants (NLDB, which stands for *Neural Linear Dueling Bandit*) of our method. Figure E.5–Figure E.7 compare the cumulative average regrets of these two sets of variants across different synthetic datasets featuring nonlinear utility functions.

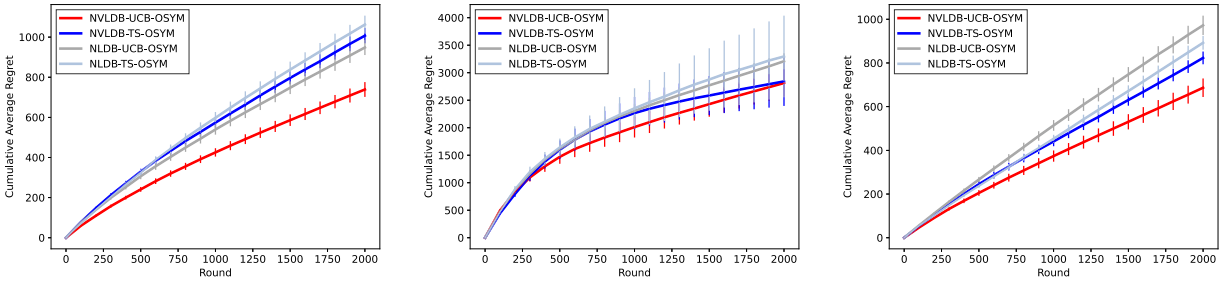
As stated in the main manuscript, Corollary C.1 and Corollary D.1 indicate that, without variance-awareness, the cumulative average regret of our arm selection strategies remains sublinear. To validate these theoretical results, we conduct experiments under variance-agnostic variants of NVLDB. Figure E.5–Figure E.7 compare the cumulative average regrets of the variance-aware (NVLDB) and variance-agnostic variants (NLDB) of NVLDB across different synthetic datasets with nonlinear utility functions.

We make two primary observations from these comparative experiments. First, the variance-agnostic NLDB (Neural Linear Dueling Bandit) methods exhibit sublinear cumulative average regret regardless of the arm selection strategy employed. This empirical finding is consistent with our theoretical results presented in Corollary C.1 and Corollary D.1, which suggest that a sublinear rate is attainable even without explicit variance-awareness. Second, the variance-aware variants generally outperform the variance-agnostic counterparts (NLDB) in terms of cumulative average regret, highlighting the practical benefit of incorporating variance information into the arm selection strategy. However, in some specific instances, the variance-agnostic variants (NLDB) are observed to perform better (see Figure E.7(c)). This observation is still consistent with the theory, as both the variance-aware (NVLDB) and variance-agnostic (NLDB) variants share the same theoretical upper bound on cumulative average regret, namely $\mathcal{O}(d\sqrt{T})$, provided that the variances are sufficiently large. Moreover, theoretical guarantees are typically derived under idealized assumptions that may not hold perfectly in practice. Thus, slight deviations in empirical performance between the two approaches are inevitable. **Crucially, these discrepancies do not undermine** the validity of the theoretical results but rather highlight the inherent gap between idealized theoretical modeling and real-world implementation.



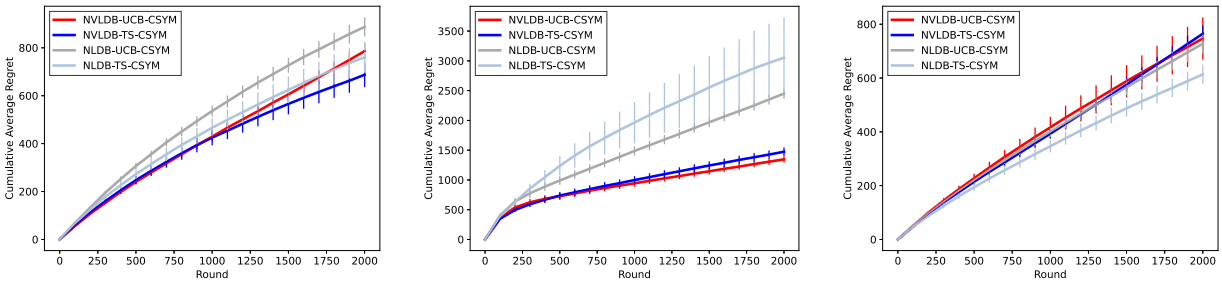
(a): $\cos(3\Theta^\intercal x)$ (b): $10(\Theta^\intercal x)^2$ (c): $x^\intercal\Theta\Theta^\intercal x$

Figure E.5: Comparison of cumulative average regret between variance-aware (NVLDB) and variance-agnostic (NLDB) variants of our proposed method under asymmetric (ASYM) arm selection on synthetic tasks with $d = 5$ and $K = 5$. Each experiment was conducted across 20 random seeds.



(a): $\cos(3\Theta^\intercal x)$ (b): $10(\Theta^\intercal x)^2$ (c): $x^\intercal\Theta\Theta^\intercal x$

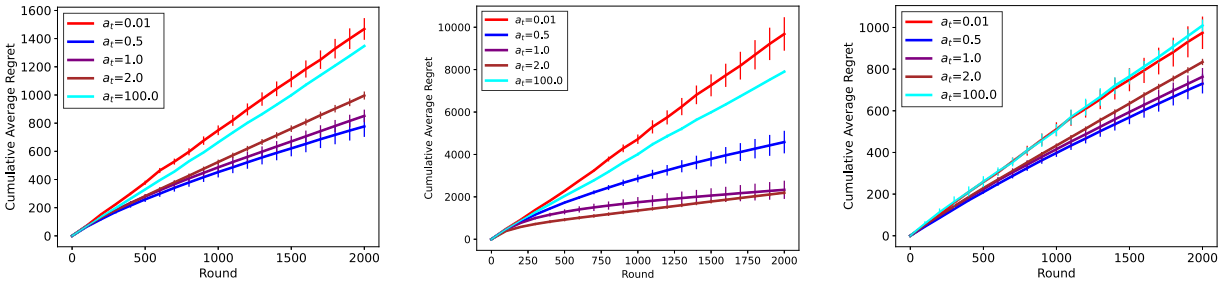
Figure E.6: Comparison of cumulative average regret between variance-aware (NVLDB) and variance-agnostic (NLDB) variants of our proposed method under optimistic symmetric (OSYM) arm selection on synthetic tasks with $d = 5$ and $K = 5$. Each experiment was conducted across 20 random seeds.



(a): $\cos(3\Theta^\intercal x)$ (b): $10(\Theta^\intercal x)^2$ (c): $x^\intercal \Theta \Theta^\intercal x$

Figure E.7: Comparison of cumulative average regret between variance-aware (NVLDB) and variance-agnostic (NLDB) variants of our proposed method under candidate-based symmetric (CSYM) arm selection on synthetic tasks with $d = 5$ and $K = 5$. Each experiment was conducted across 20 random seeds.

Different Confidence Interval Coefficients For practical reasons, we chose and fixed the confidence interval coefficient a_t. The selection of a fixed a_t has been widely adopted in similar neural bandit studies (Xu et al., 2020; Verma et al., 2024; Bui et al., 2024; Zhou et al., 2020), as previously stated. Intuitively, the parameter a_t controls the trade-off between exploration and exploitation. Both excessively small and excessively large values of a_t are detrimental to establishing sublinear cumulative average regret. A small a_t leads to weak exploration, potentially trapping the algorithm in suboptimal regions, while a large a_t conducts excessive and inefficient exploration. Indeed, excessively large or small values of this coefficient result in looser upper bounds during the proof of Theorem C.1 through Corollary D.1. We empirically evaluated the cumulative average regret for various values of β. Figure E.8 compares the cumulative average regrets of NVLDB using asymmetric arm selection under the UCB regime. The results show that a moderate value of a_t provides the best performance, and performance degrades when a_t is too small or too large. This observation aligns well with our theoretical analysis regarding the exploration-exploitation balance.



(a): $\cos(3\Theta^\intercal x)$ (b): $10(\Theta^\intercal x)^2$ (c): $x^\intercal \Theta \Theta^\intercal x$

Figure E.8: Comparison of cumulative average regret of NVLDB with asymmetric arm selection strategy arm selection under UCB regime across different confidence interval coefficient a_t on synthetic tasks with $d = 5$ and $K = 5$. Each experiment was conducted across 20 random seeds.

Computational Efficiency. We compared the wall-clock times of NVLDB-UCB-ASYM, which is one of our methods, and NDB (Verma et al., 2024) over 2,000 rounds across 20 random seeds. Measurements were taken on an NVIDIA GeForce RTX 3060 GPU paired with an AMD Ryzen 9 5950X 16-core CPU.

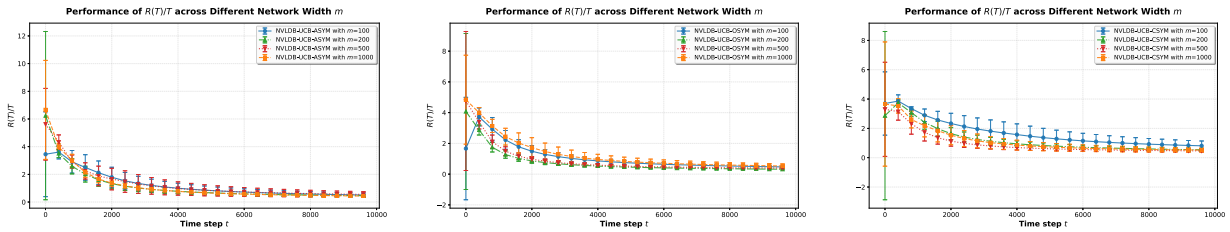
As a result, the proposed method significantly outperforms NDB in terms of computational efficiency. Specifically, **NVLDB-UCB-ASYM completes 2,000 rounds in an average of** 1.71 ± 0.78 **minutes, while NDB requires** 48.8 ± 15.5 **minutes under the same conditions.** This substantial disparity is justified by the underlying computational processes: NDB utilizes gradients with respect to the full set of learnable parameters in the neural network, whereas our method utilizes gradients with respect to only the learnable parameters of the last layer to construct the Gram matrix.

E.4 Network Width Sensitivity

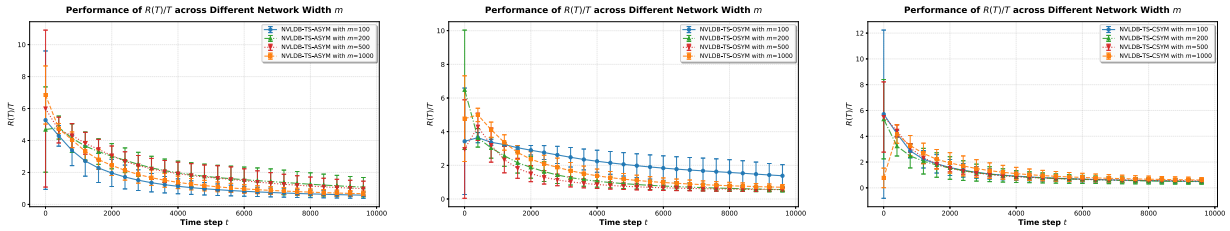
To verify scalability with respect to network width, we extended experiments to $m \in \{500, 1000\}$, confirming sublinear regrets over $T = 10{,}000$ rounds (square utility, $d = 5$, $K = 5$) for all NVLDB variants. Table E.2 summarizes the $R(T)/T$ values, and Figures E.9 and E.10 visualize the cumulative average and total regret curves, respectively.

Table E.2: $R(T)/T$ values across network width $m \in \{500, 1000\}$ on the square utility task ($d = 5$, $K = 5$). All tested algorithms exhibit steadily decreasing $R(T)/T$, following $\widetilde{\mathcal{O}}(T^{-1/2})$ convergence. Each experiment is repeated across 20 random seeds.

Algorithm	$R(2000)/2000$	$R(6000)/6000$	$R(10000)/10000$
NVLDB-UCB-ASYM ($m = 500$)	1.66 ± 0.78	0.82 ± 0.39	0.54 ± 0.19
NVLDB-UCB-OSYM ($m = 500$)	0.97 ± 0.17	0.51 ± 0.07	0.39 ± 0.04
NVLDB-UCB-CSYM ($m = 500$)	1.13 ± 0.35	0.61 ± 0.13	0.46 ± 0.07
NVLDB-UCB-ASYM ($m = 1000$)	1.37 ± 0.22	0.65 ± 0.09	0.42 ± 0.04
NVLDB-UCB-OSYM ($m = 1000$)	1.73 ± 0.65	0.81 ± 0.27	0.52 ± 0.13
NVLDB-UCB-CSYM ($m = 1000$)	1.54 ± 0.65	0.73 ± 0.27	0.51 ± 0.14



(a): UCB-ASYM (b): UCB-OSYM (c): UCB-CSYM

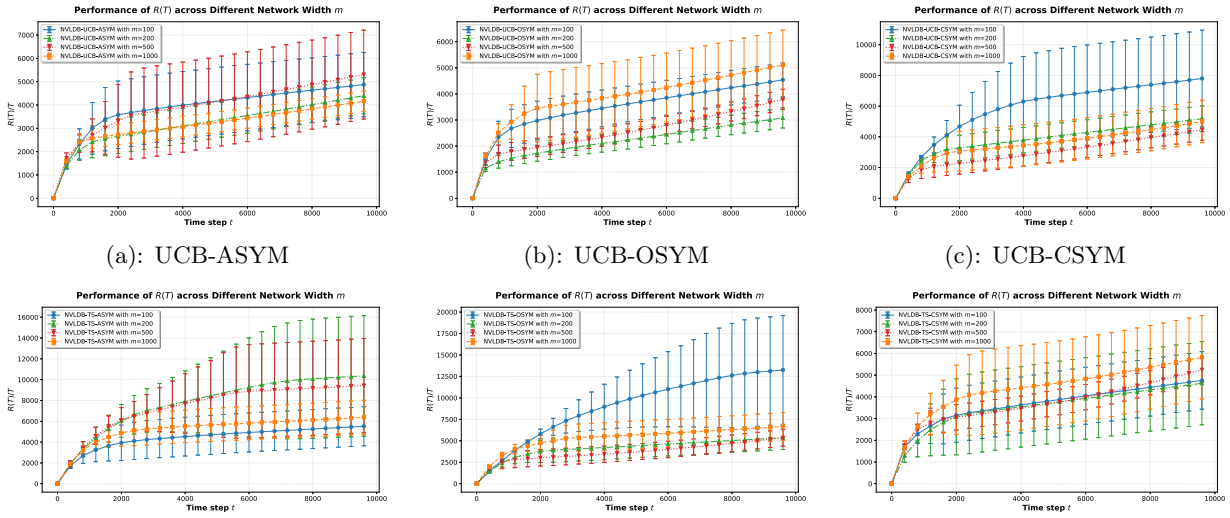


(d): TS-ASYM (e): TS-OSYM (f): TS-CSYM

Figure E.9: Cumulative average regret across different network widths ($m \in \{500, 1000\}$) on the square utility task with $d = 5$ and $K = 5$. All NVLDB variants maintain sublinear regret as the network width increases. Each experiment is repeated across 20 random seeds.

E.5 High-Dimensional Experiments

To evaluate scalability in higher dimensions, we conducted experiments with $d = 20$ and $K = 5$ over $T = 10{,}000$ rounds on the square utility task, significantly exceeding the dimensionality used in prior works (Verma et al., 2024; Bui et al., 2024). Table E.3 reports the $R(T)/T$ values, and Figure E.11 visualizes the cumulative average and total regret.



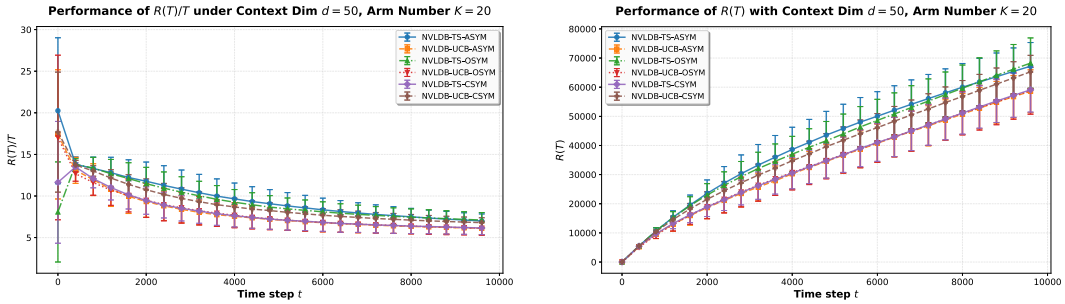
(a): UCB-ASYM (b): UCB-OSYM (c): UCB-CSYM

(d): TS-ASYM (e): TS-OSYM (f): TS-CSYM

Figure E.10: Cumulative total regret across different network widths ($m \in \{500, 1000\}$) on the square utility task with $d = 5$ and $K = 5$. Each experiment is repeated across 20 random seeds.

Table E.3: $R(T)/T$ values with $d = 20$ and $K = 5$ over $T = 10{,}000$ rounds on the square utility task. All tested algorithms exhibit steadily decreasing $R(T)/T$, following $\widetilde{\mathcal{O}}(T^{-1/2})$ convergence. Each experiment is repeated across 20 random seeds.

Algorithm	$R(2000)/2000$	$R(6000)/6000$	$R(10000)/10000$
NVLDB-UCB-ASYM	9.33 ± 1.89	6.79 ± 1.09	6.05 ± 0.74
NVLDB-UCB-OSYM	9.49 ± 2.05	6.82 ± 1.11	6.07 ± 0.83
NVLDB-UCB-CSYM	10.76 ± 1.29	7.69 ± 0.80	6.73 ± 0.58



(a): Cumulative average regret (b): Cumulative total regret

Figure E.11: Cumulative average and total regret with $d = 20$ and $K = 5$ on the square utility task over $T = 10{,}000$ rounds. All NVLDB variants maintain sublinear regret in the high-dimensional regime. Each experiment is repeated across 20 random seeds.