Exploration via Feature Perturbation in Contextual Bandits

Seouh-won Yi
Seoul National University
Seoul, South Korea
uniqueseouh@snu.ac.kr

Min-hwan Oh
Seoul National University
Seoul, South Korea
minoh@snu.ac.kr

Abstract

We propose *feature perturbation*, a simple yet effective exploration strategy for contextual bandits that injects randomness directly into feature inputs, instead of randomizing unknown parameters or adding noise to rewards. Remarkably, this algorithm achieves $\widetilde{\mathcal{O}}(d\sqrt{T})$ worst-case regret bound for generalized linear contextual bandits, while avoiding the $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret typical of existing randomized bandit algorithms. Because our algorithm eschews parameter sampling, it is both computationally efficient and naturally extends to non-parametric or neural network models. We verify these advantages through empirical evaluations, demonstrating that feature perturbation not only surpasses existing methods but also unifies strong practical performance with the near-optimal regret guarantees.

1 Introduction

Contextual bandits have long been a pivotal framework in sequential decision-making and online learning [6, 8, 31]. In each round, a decision-maker (the *agent*) observes contextual information and must select an action (arm) to maximize a reward. Beyond the classical multi-armed bandit setting, contextual bandits better reflect real-world scenarios where relevant context (such as features of users and items to be chosen) informs the choice of action.

A central model of contextual bandits is the *(generalized) linear contextual bandit*, where the expected reward is modeled either by a linear function or, more generally, by a generalized linear model (GLM) [1, 10, 15, 35, 36]. Within these settings, deterministic methods inspired by *optimism in the face of uncertainty* (OFU) [1, 6] and randomized approaches such as Thompson Sampling (TS) [2, 5, 9] or Perturbed History Exploration (PHE) [28, 29, 32] have been extensively studied. Notably, OFU-type algorithms achieve near-optimal regret of $\widetilde{\mathcal{O}}(d\sqrt{T})$ in linear contextual bandits (and likewise in GLM bandits), yet often underperform compared to TS and PHE in practice. In contrast, randomized exploration methods typically exhibit superior empirical performance but suffer from sub-optimal theoretical guarantees: standard analyses confirm a regret bound of $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ [2, 5] worst-case regret. [1] Crucially, this gap is not just an artifact of the analysis; Hamidi and Bayati [18] show that it reflects an inherent limitation of randomization in the (generalized) linear TS bandit algorithm.

This dichotomy prompts a natural question: *is it possible to close the gap between randomized exploration and $\widetilde{\mathcal{O}}(d\sqrt{T})$ worst-case regret?* If one adheres strictly to randomizing the unknown

[1] LinTS [2, 5] satisfies a regret of $\widetilde{\mathcal{O}}(\min(d^{3/2}\sqrt{T},\, d\sqrt{T\log K}))$. LinPHE was first shown to achieve $\widetilde{\mathcal{O}}(d\sqrt{T\log K})$ [28], but has recently been improved to $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ [32]. Since we consider large action spaces with $K > e^d$, our focus is on the $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regime for randomized (generalized) linear bandits.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

parameters or perturbing the observed rewards (as in PHE), there appears a fundamental barrier [18] preventing regret from achieving $\widetilde{\mathcal{O}}(d\sqrt{T})$. In this work, we propose a simple yet powerful alternative: instead of sampling parameters or rewards, we randomly perturb the *features* used in the reward model. By shifting the focus of exploration from parameter space to feature space, we circumvent the limitations that impose higher regret on existing randomized algorithms. Remarkably, our analysis shows that this new approach not only retains the empirical advantages of randomization but also achieves $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret in (generalized) linear bandit settings, with no additional dependence on the number of arms. Furthermore, our method avoids the computational overhead of sampling parameters, making it attractive for a wide range of real-world applications.

Beyond theoretical efficiency, feature perturbation can seamlessly extend to more flexible or non-parametric reward models, including neural networks. We demonstrate this empirically, showing that feature-based randomization can drive effective exploration even when parametric assumptions break down. By unifying strong theoretical guarantees with practical efficacy in both (generalized) linear and more complex models, our approach closes a longstanding gap in randomized exploration for contextual bandits. Our main contributions are summarized as follows:

- **Feature perturbation for contextual bandits.** We introduce a *new class of algorithms* for randomized exploration, termed *feature perturbation*, which focuses on perturbing feature inputs rather than parameters or rewards. This approach is straightforward to implement and conceptually distinct from existing randomization strategies.
- **Tight regret bounds.** To the best of our knowledge, our work is the first *randomized algorithm* for generalized linear contextual bandits that achieves: (i) a regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T})$, matching the best-known guarantees of deterministic (OFU-based) methods; and simultaneously (ii) benefiting from an instance-dependent constant κ. Notably, our algorithm's regret does not increase with the number of arms.
- **Empirical validation.** Through extensive experiments on both synthetic and real-world data, we show that feature perturbation not only performs competitively against existing randomized methods but also generalizes beyond parametric models (e.g., deep neural networks), demonstrating robustness even when linear assumptions do not hold.

2 Preliminaries

Notations. For vectors $x, y \in \mathbb{R}^d$, let $\|x\|$ denote the 2-norm and $\|x\|_A = \sqrt{x^\top A x}$ the weighted norm for a positive definite matrix $A \in \mathbb{R}^{d\times d}$. The inner product is $x^\top y = \langle x, y\rangle$, and the weighted version is $x^\top A y = \langle x, y\rangle_A$. The notation $\widetilde{\mathcal{O}}$ hides logarithmic factors in big-O notation, retaining instance-dependent constants. The set $\{1, \dots, K\}$ is abbreviated as $[K]$.

Generalized linear contextual bandit. A *generalized linear model* (GLM; [38]) describes a response $r \in \mathbb{R}$ drawn from an exponential-family distribution with mean $\mu(x^\top\theta^*)$, where $x \in \mathbb{R}^d$ is a feature vector and $\theta^* \in \mathbb{R}^d$ is an unknown parameter. Given differentiable functions g and h, and a base measure ν, the conditional density of r given x takes the form:

$$dp(r|x;\theta^*) = \exp\left(rx^\top\theta^* - g(x^\top\theta^*) + h(r)\right) d\nu,$$

where the derivative of g defines the *link function* μ. Let $\mathcal{H}_{t-1} := \sigma(\{(x_\tau, r_\tau)\}\, \tau = 1^{t-1})$ denote the filtration up to round $t-1$. We define $\mathbb{P}_t(\cdot) := \mathbb{P}(\cdot \mid \mathcal{H}_{t-1})$ and $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{H}_{t-1}]$. The negative log-likelihood and the *maximum likelihood estimate* (MLE) at round t are then given by:

$$L_t(\theta) = \sum_{\tau=1}^{t-1} \left(g(x_\tau^\top\theta) - r_\tau x_\tau^\top\theta\right), \quad \hat{\theta}_t := \operatorname*{argmin}_{\theta\in\Theta} L_t(\theta).$$

In the generalized linear contextual bandit (GLB) setting, the agent observes a context $c_t \in \mathcal{C}$ and a corresponding set of feature vectors $\mathcal{X}_t \subset \mathbb{R}^d$ representing each allowable arm $a \in \mathcal{A}(c_t)$ at each round. After selecting $x_t \in \mathcal{X}_t$, the reward is generated as $r_t = \mu(x_t^\top\theta^*) + \xi_t$, where the noise ξ_t is a *martingale difference sequence* adapted to $\mathcal{H}_{t-1}$ and is conditionally R-subgaussian. The learner aims to minimize the regret: $R(T) = \sum_{t=1}^{T}\left(\mu(x_{t*}^\top\theta^*) - \mu(x_t^\top\theta^*)\right)$, where $x_{t*} := \operatorname{argmax}_{x\in\mathcal{X}_t} \mu(x^\top\theta^*)$ is the optimal arm at round t, which depends on the context c_t.

The complexity of the GLB problem is fundamentally determined by the following quantities, which captures the degree of nonlinearity in the reward function:

$$\kappa_* := \frac{\sum_{t=1}^{T} \dot{\mu}(x_{t*}^\top \theta^*)}{T}, \quad \kappa := \min_{x \in \mathcal{X}_{[T]}, \theta \in \Theta} \dot{\mu}(x^\top \theta), \quad \text{where} \quad \mathcal{X}_{[T]} := \bigcup_{t=1}^{T} \mathcal{X}_t. \tag{1}$$

3 Algorithm: GLM-FP

Algorithm 1 GLM-FP: Feature Perturbation (FP) in Generalized Linear bandits

1: **Input:** Regularization parameter $\lambda > 0$, tuning parameter $\{c_t\}$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Compute $\hat{\theta}_t = \operatorname{argmin}_{\theta \in \mathbb{R}^d} L_t(\theta; \{x_\tau, r_\tau\}_{\tau=1}^{t-1})$
4: Sample $\zeta_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
5: Compute $\widetilde{x}_{ti} = x_{ti} + c_t \cdot \frac{\|x_{ti}\|_{\hat{H}_t^{-1}}}{\|\hat{\theta}_t\|} \cdot \zeta_t$ for all i
6: Choose $i_t = \operatorname{argmax}_{i \in [|\mathcal{X}_t|]} \mu(\widetilde{x}_{ti}^\top \hat{\theta}_t)$ and observe reward r_t ▷ Let $x_t := x_{t,i_t}$
7: **end for**

At each step t, given the history $\mathcal{H}_{t-1}$, the algorithm computes the MLE $\hat{\theta}_t$ via numerical methods such as SQP or IRLS ([12, 42]; line 3). Instead of perturbing rewards or parameters, we inject controlled randomness directly into feature vectors using a perturbing distribution, which by default is multivariate normal (line 4). Unlike prior GLB methods that rely on the vanilla Gram matrix—leading to a critical inverse dependence on κ [2, 36, 41]—we scale perturbations with the weighted Gram matrix $\hat{H}_t = \lambda I + \nabla^2 L_t(\hat{\theta}_t)$, modulated by a parameter c_t. This generates perturbed features $\{\widetilde{x}_{ti}\}$ (line 5), all coupled through a shared noise ζ_t, thereby removing explicit K-dependence and strengthening both theoretical guarantees and empirical performance. Finally, the algorithm selects the arm maximizing $\mu(\widetilde{x}_{ti}^\top \hat{\theta}_t)$ and updates the history with the observed reward r_t (line 6).

3.1 Intuition behind the algorithm

In contextual bandits, randomized algorithms typically fall into two categories: (i) model perturbation, where exploration is induced by perturbing parameters (e.g., PHE [28], TS [2, 5]), and (ii) reward perturbation, where deterministic models are trained but stochastic bonuses are added to arm values (e.g., RandUCB [41]). The former can be suboptimal even in linear settings [18], while the latter may break the inductive bias of the function class, since reward perturbations need not correspond to any realizable model.

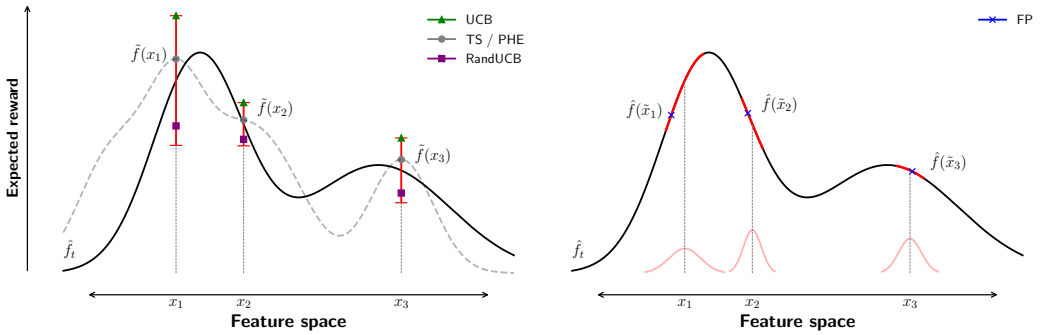


Figure 1: (Left) Model perturbation methods randomize rewards via off-estimated models $\widetilde{f}_t$. (Right) Feature Perturbation (FP) perturbs inputs and evaluates them with a fixed model $\hat{f}_t$.

We propose an alternative that retains the estimated model $\hat{f}$ and introduces randomness through input perturbations at decision time. By exploring in feature space rather than parameter or reward space, this preserves structural assumptions, remains effective in overparameterized regimes ($p \gg d$), and reflects real-world contexts where features are inherently noisy [7, 25, 26].

Our algorithm, GLM-FP, implements this principle by applying stochastic perturbations to contexts while keeping $\hat{f}$ fixed. Unlike TS, PHE, or RandUCB, which rely on modified models $\widetilde{f}_t$, GLM-FP strictly adheres to the learned model and shifts exploration to the input space, yielding both theoretical and practical benefits. In doing so, we develop and analyze feature-perturbation-based strategies within the GLB framework that extend naturally beyond model perturbation methods.

4 Regret analysis of GLM-FP

We now establish the regret guarantee for GLM-FP. Our analysis proceeds by controlling both parameter estimation error and the variability induced by perturbations, while ensuring a non-trivial probability of stochastic optimism. We begin by stating the standard boundedness assumption used in generalized linear bandit analyses [1–3, 5, 13, 15, 29, 34, 36, 44].

Assumption 1 (Boundedness)**.** *The feature space $\mathcal{X}$ and parameter space Θ are compact subsets of $\mathbb{R}^d$. For any $x \in \mathcal{X}_{[T]}$ and $\theta^* \in \Theta$, we have $\|x\| \leq 1$ and $\|\theta^*\| \leq 1$.*

4.1 High-probability events and stochastic optimism

With the confidnece width $\beta_t(\delta)$ defined in Proposition F.1 over ellipsoidal confidence sets around the MLE $\hat{\theta}_t$, we ensure that $\theta^* \in \Theta_t(\delta, \lambda)$ with high probability. In parallel, we require perturbed contexts $\widetilde{x}_{ti}$ to remain concentrated around their unperturbed counterparts.

Lemma 1. *Let $t \in [T]$ $\delta' = \delta/(4T)$, and $\gamma_t(\delta) = \beta_t(\delta')\sqrt{2\log(2/\delta)}$. Define the events*

$$\hat{E}_t := \left\{\forall \tau \leq t;\ \|\hat{\theta}_\tau - \theta^*\|_{\hat{H}_t} \leq \beta_t(\delta')\right\} \quad \textit{and} \quad \widetilde{E}_t := \left\{\forall \tau \leq t, x_{\tau i} \in \mathcal{X}_\tau;\ \widetilde{x}_{\tau i} \in \mathcal{E}_\tau(x_{\tau i})\right\},$$

where $\mathcal{E}_t(x) := \{\widetilde{x} \in \mathbb{R}^d \mid |\langle \widetilde{x} - x, \hat{\theta}_t\rangle| \leq \gamma_t(\delta')\|x\|_{\hat{H}_t^{-1}}\}$ represents a high probability region for the perturbed feature vector for each arm x. Under Assumption 1, with $c_t = \beta_t(\delta')$, we have

$$\mathbb{P}(\hat{E}_T \cap \widetilde{E}_T) \geq 1 - \tfrac{\delta}{2}, \quad \textit{and} \quad \mathbb{P}_t(\mu(\widetilde{x}_t^\top \hat{\theta}_t) \geq \mu(x_{t*}^\top \theta^*) \mid \hat{E}_t, \widetilde{E}_t) \geq \tfrac{1}{4\sqrt{e\pi}}.$$

Thus, with high probability, (i) parameter estimates remain close to θ^*, and (ii) perturbed features are both concentrated and stochastically optimistic. Unlike Thompson sampling, which incurs an additional $\sqrt{d}$ oversampling factor, our construction keeps $\beta_t, \gamma_t = \mathcal{O}(\sqrt{d})$, yielding tighter exploration terms. Further technical details are deferred to the Appendix H.

4.2 Regret bound of GLM-FP

Theorem 1. *For all $\delta \in (0, 1)$, define $\delta' = \delta/(4T)$. Under Assumption 1, with $c_t = \beta_t(\delta')$ and $\lambda = \mathcal{O}(d)$, the cumulative regret $R(T)$ is bounded with probability at least $1 - \delta$ as follows:*

$$R(T) = \widetilde{\mathcal{O}}\left(d\sqrt{\kappa_* T} + d^2/\kappa\right).$$

Discussion of Theorem 1. The leading term of the regret guarantee is $\widetilde{\mathcal{O}}(d\sqrt{\kappa_* T})$, which matches the minimax optimal regret bound in terms of the dimensionality d, the horizon T, and the instance-dependent constant κ_* [3, 34]. While RandUCB [41], another randomized algorithm, also achieves a regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T})$, it is penalized by its inverse dependence on κ, lacking adaptation to instance-dependent complexity. In contrast, to the best of our knowledge, our result is the first to show that a randomized algorithm achieves a regret bound with linear d-dependency, without additional dependence on the number of arms, benefiting form κ_* in GLB problems.

5 Experiments and Conclusion

Due to space constraints, detailed experimental results are deferred to the Appendix I. In this work, we introduced a novel randomized exploration strategy, *feature perturbation*, which improves the regret bound by a factor of $\sqrt{d}$ compared to prior randomized algorithms. As a result, our algorithm attains the theoretically optimal bound $\widetilde{\mathcal{O}}(d\sqrt{T})$ without inverse dependence on κ in contextual GLBs. These findings highlight the effectiveness of feature perturbation in contextual bandits and open new avenues for extending randomized exploration to broader online learning applications.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

[2] Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184. PMLR, PMLR, 20–22 Apr 2017.

[3] Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[4] Marc Abeille, David Janz, and Ciara Pike-Burke. When and why randomised exploration works (in linear bandits). In Gautam Kamath and Po-Ling Loh, editors, *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pages 4–22. PMLR, 24–27 Feb 2025.

[5] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[6] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[7] Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Oper. Res.*, 68(1):276–294, January 2020. ISSN 0030-364X.

[8] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012. ISSN 1935-8237.

[9] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

[10] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[12] Jean Bosco Etoa Etoa. A globally convergent sequential linear programming algorithm for mathematical programs with linear complementarity constraints. *Journal of Information & Optimization Sciences*, 31, 09 2010.

[13] Louis Faury, Marc Abeille, Clement Calauzenes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3052–3060. PMLR, 13–18 Jul 2020.

[14] Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clement Calauzenes. Jointly efficient and optimal algorithms for logistic bandits. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 546–580. PMLR, 28–30 Mar 2022.

[15] Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, page 586–594, Red Hook, NY, USA, 2010. Curran Associates Inc.

[16] Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy networks for exploration. In *International Conference on Learning Representations*, 2018.

[17] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[18] Nima Hamidi and Mohsen Bayati. On frequentist regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*, 2020.

[19] Osama A. Hanna, Lin F. Yang, and Christina Fragouli. Efficient batched algorithm for contextual linear bandits with large action space via soft elimination. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[20] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[21] Osband Ian, Van Roy Benjamin, and Russo Daniel. (more) efficient reinforcement learning via posterior sampling. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3003–3011, Red Hook, NY, USA, 2013. Curran Associates Inc.

[22] David Janz, Shuai Liu, Alex Ayoub, and Csaba Szepesvári. Exploration via linearly perturbed loss minimisation. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 721–729. PMLR, 02–04 May 2024.

[23] Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. *Advances in Neural Information Processing Systems*, 30, 2017.

[24] Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5148–5157. PMLR, 18–24 Jul 2021.

[25] Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 2231–2241, Red Hook, NY, USA, 2018. Curran Associates Inc.

[26] Jung-Hun Kim, Se-Young Yun, Minchan Jeong, Junhyun Nam, Jinwoo Shin, and Richard Combes. Contextual linear bandits under noisy features: Towards bayesian oracles. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 1624–1645. PMLR, 25–27 Apr 2023.

[27] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[28] Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. In *Uncertainty in Artificial Intelligence*, pages 530–540. PMLR, 2020.

[29] 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.

[30] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

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

[32] Harin Lee and Min-hwan Oh. Improved regret of linear ensemble sampling. In *Advances in Neural Information Processing Systems*, volume 38, 2024.

[33] Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. Improved regret bounds of (multinomial) logistic bandits via regret-to-confidence-set conversion. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4474–4482. PMLR, 02–04 May 2024.

[34] Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[35] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[36] 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.

[37] Kolby Nottingham Markelle Kelly, Rachel Longjohn. The uci machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[38] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Monographs on statistics and applied probability. Champman and Hall/CRC, 2001.

[39] Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8626–8638, Red Hook, NY, USA, 2018. Curran Associates Inc.

[40] Connor Shorten and Taghi Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, 07 2019.

[41] Sharan Vaswani, Abbas Mehrabian, Audrey Durand, and Branislav Kveton. Old dog learns new tricks: Randomized ucb for bandit problems. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1988–1998. PMLR, 26–28 Aug 2020.

[42] R. Wolke and H. Schwetlick. Iteratively reweighted least squares: Algorithms, convergence analysis, and numerical comparisons. *SIAM Journal on Scientific and Statistical Computing*, 9 (5):907–921, 1988.

[43] Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural thompson sampling. In *International Conference on Learning Representation (ICLR)*, 2021.

[44] Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11492–11502. PMLR, 13–18 Jul 2020.

Appendix

Table of Contents

A Related works **8**

B Properties of FP distributions **10**
B.1 Perturbation in Thompson Sampling 10
B.2 Examples of FP distributions 11

C Application of FP algorithm **11**
C.1 Generalized version of `GLM-FP` 11
C.2 Application to the linear bandit problem 12

D Limitations **12**

E Table of notations **13**

F Proof of main theorem **14**
F.1 Confidence bound for the true parameter 14
F.2 Proof of Lemma 1-Stochastic Optimism 15
F.3 Proof of Lemma 1-Concentration 15
F.4 Proof of Theorem 1 16

G Proof supplement and Lemmas **24**
G.1 Supporting Lemmas for main proof 24
G.2 Auxiliary bounding inequalities 26

H Carving off the $\sqrt{d}$ factor compared to TS **26**

I Experiments **28**
I.1 Generalized linear bandits 28
I.2 Neural bandits 29
I.3 Additional experiments 29

A Related works

Contextual bandits have been extensively investigated under various modeling assumptions. In the *linear* bandit setting, deterministic methods based on OFU [1, 6] achieve near-optimal $\widetilde{O}(d\sqrt{T})$ regret, but often exhibit conservative exploration in practice. By contrast, *randomized* algorithms such as TS [2, 5, 9] and PHE [28, 29] typically show better empirical performance yet suffer from a higher $\widetilde{O}(d^{3/2}\sqrt{T})$ regret bound. Notably, Hamidi and Bayati [18] demonstrated that the extra $\sqrt{d}$ inflation in TS-type algorithms is unavoidable in worst-case scenarios: eliminating this factor would lead to a linear dependence on T. Consequently, parameter-based randomization cannot, in general, achieve $\widetilde{O}(d\sqrt{T})$ regret without further modifications.

Generalized linear bandits (GLB; [15, 36]) extend linear bandits to settings where rewards follow a nonlinear link function. UCB- and TS-based approaches [2, 15, 29, 41] have also been applied here, displaying the same contrast between deterministic and randomized exploration. While UCB-type methods reach $\widetilde{O}(d\sqrt{T})$ regret, they tend to over-explore in practice; randomized strategies mitigate this over-exploration but retain an additional $\sqrt{d}$ penalty in the worst case. Like their linear counterparts, these methods rely on sampling the unknown parameter or perturbing rewards rather than altering the feature representation.

Table A.1: Representative GLB algorithms: regret bounds and source of stochasticity.

Type	Algorithm	Regret Upper Bound	Stochasticity
Deterministic	GLM-UCB [15]	$\widetilde{\mathcal{O}}(d\sqrt{T}/\kappa)$	–
	Logistic-UCB-2 [23]	$\widetilde{\mathcal{O}}(d\sqrt{T}+d^2/\kappa)$	–
	OFUGLB [34]	$\widetilde{\mathcal{O}}(d\sqrt{\kappa_* T}+d^2/\kappa)$	–
Randomized	LinTS [2]	$\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T}/\kappa)$	Parameter (θ)
	GLM-TSL [29]	$\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T}/\kappa)$	Parameter (θ)
	GLM-FPL [29]	$\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T}/\kappa)$	Reward (r)
	RandUCB [41]	$\widetilde{\mathcal{O}}(d\sqrt{T}/\kappa)$	Linear utility ($x^\top\theta$)
	GLM-FP (Ours)	$\boldsymbol{\widetilde{\mathcal{O}}\left(d\sqrt{\kappa_* T}+d^2/\kappa\right)}$	**Feature vector (x)**

By contrast, our work introduces a new class of *feature-perturbation* (FP) algorithms designed to circumvent the dimensional penalty inherent in standard randomized approaches. Instead of randomizing parameters or rewards, we propose to perturb the features directly. This perspective not only preserves the empirical robustness associated with randomized strategies but also achieves a tight regret bound in both linear and generalized linear settings—thereby reconciling the theoretical and practical advantages of contextual bandit exploration.

Landscape of GLB Algorithms. Generalized linear bandit (GLB) algorithms can be broadly divided into *deterministic OFU-type* and *randomized exploration*-based methods. OFU approaches such as GLM-UCB [15] and Logistic-UCB [23] achieve the tight $\widetilde{\mathcal{O}}(d\sqrt{T})$ or $\widetilde{\mathcal{O}}(d\sqrt{T}/\kappa)$ regret bound, and refinements further improve confidence construction [24], achieving $\widetilde{\mathcal{O}}(\sqrt{dT\log K})$ in the finite-K arm setting. Abeille et al. [3] provided an instance-dependent analysis showing that the regret can benefit from the curvature constant κ_*, achieving $\widetilde{\mathcal{O}}(d\sqrt{\kappa_* T})$. Subsequent works [33, 34] relaxed the dependence on S or improved computational efficiency [14] while preserving the same order. Randomized methods such as Thompson Sampling (TS) and Perturbed History Exploration (PHE) [2, 5, 28, 29], as well as more recent algorithms like EVILL [22] and RandUCB [41], typically achieve superior empirical performance but incur an additional $\sqrt{d}$ penalty in the worst case, yielding $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ for infinite arms or $\widetilde{\mathcal{O}}(d\sqrt{T\log K}/\kappa)$ for finite arms. Our proposed feature perturbation (FP) departs from parameter- or reward-perturbation by randomizing the *input features*, thereby closing this gap: as summarized in Table A.1, FP is the first randomized algorithm for GLBs to provably achieve $\widetilde{\mathcal{O}}(d\sqrt{\kappa_* T}+d^2/\kappa)$ regret with no dependence on K, unifying the tight guarantees of OFU-type methods with the empirical robustness of randomized exploration.

Comparison to RandUCB Algorithm. Our linear variant `LinFP` (see C.2) and RandUCB [41] coincide in the linear bandit setting. Both sample randomized scores $\widetilde{f}_t(x_i)\sim\mathcal{N}(x_{ti}^\top\hat{\theta}_t,\ \beta_t^2\|x_{ti}\|^2_{V_t^{-1}})$ and couple the arms identically, yielding matching $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret bounds. The difference lies only in the *source* of randomness: FP perturbs the input features, whereas RandUCB perturbs the reward estimate itself. These distinct mechanisms collapse to the same Gaussian rule under linear models, though they have been analyzed through different theoretical perspectives. Equivalently, one may view RandUCB as a special instance of FP with an identical perturbation distribution, differing only in interpretation and analytical framework.

In generalized linear bandits (GLBs), however, the two algorithms diverge fundamentally. RandUCB extends its linear recipe by linearizing the link function, introducing a multiplicative κ^{-1} penalty and yielding a regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T}/\kappa)$. In contrast, `GLM-FP` perturbs the inputs directly using a curvature-aware Gram matrix that weights past features by $\dot{\mu}(x^\top\hat{\theta}_t)$, enabling both anti-concentration (for exploration) and concentration (for confidence). This yields a tighter regret of $\widetilde{\mathcal{O}}(d\sqrt{\kappa_* T})$, linear in d and directly in κ_*. Conceptually, FP injects stochasticity *before* inference, so that each sampled reward $\hat{f}(\widetilde{x})$ remains within the hypothesis class—preserving inductive bias and reflecting epistemic uncertainty. RandUCB, in contrast, adds randomness *after* inference, producing post hoc scores that may not correspond to any $f\in\mathcal{F}$. As a result, in expressive models FP remains aligned with the

model structure, while RandUCB may misalign exploration incentives, leading to divergent empirical and theoretical behaviors.

Geometry and Scalability in Randomized Exploration. Recent studies have examined when randomized exploration can match the $\widetilde{\mathcal{O}}(d\sqrt{T})$ guarantees of optimistic approaches. Abeille et al. [4] identified a class of geometric conditions—absorbing, strongly convex, and smooth action sets—under which Thompson Sampling (TS) achieves optimal dependence on d. While these conditions provide valuable insights into the role of geometry, they often fail to hold in high-dimensional or unstructured settings. Subsequent works [18, 29] further clarified that posterior variance inflation can inherently introduce the extra $\sqrt{d}$ factor observed in randomized methods. In contrast, feature-level perturbation achieves similar statistical optimality under the standard boundedness assumption, bridging geometric optimality with more general feature-level regularity. From a computational standpoint, randomized exploration in large or continuous action spaces raises significant scalability challenges. Several strategies have been proposed to mitigate this issue, including lazy or delayed updates of the Gram matrix [1], two-stage candidate selection using approximate nearest neighbors, and optimization-oracle-based methods such as batched soft elimination [19]. These approaches highlight a trade-off between statistical tightness and computational efficiency: while algorithms like FP prioritize theoretical optimality in the online setting, batched or oracle-based techniques offer scalable alternatives for large-scale practical applications.

Connections to Feature Perturbation in Broader ML Feature perturbation is a common idea in other areas of machine learning, most notably in computer vision and natural language processing, where it is employed for robustness or regularization during training. Examples include data augmentation [40], adversarial training [17], or NoisyNets for exploration in deep reinforcement learning [16]. In these contexts, perturbations are introduced at training time to improve model generalization or robustness. In contrast, our FP algorithm introduces perturbations at *decision time* as a principled mechanism for exploration in online learning. This distinction highlights the novelty of FP: rather than making a static predictor robust, we leverage feature perturbations dynamically to induce stochasticity in action selection, enabling efficient exploration. The same principle applies naturally when contextual information comes from high-dimensional embeddings, such as ResNet or ViT features for images [20, 27] or BERT embeddings for language [11], where FP can induce semantically meaningful exploration by perturbing compact representations. Thus, FP not only closes a theoretical gap in contextual bandits but also suggests a unifying exploration paradigm that resonates with broader trends in modern ML. Finally, this perspective also provides a bridge to reinforcement learning (RL), where perturbing the state–action feature representation can serve as an efficient alternative to parameter-space randomization used in posterior sampling or Noisy Networks [21, 39]. Extending feature perturbation to structured settings such as Linear MDPs or value-function approximation is a promising direction for future work, potentially unifying exploration principles across bandit and reinforcement learning paradigms.

B Properties of FP distributions

B.1 Perturbation in Thompson Sampling

The perturbation distribution utilized in the TS algorithm to bring randomness to the parameter, as described by Abeille and Lazaric [2] is as followed:

Definition B.1 (Definition 1. in Abeille and Lazaric [2])**.** *$\mathcal{D}^{TS}$ is a multivariate distribution on $\mathbb{R}^d$, absolutely continuous with respect to the Lebesgue measure, and satisfies the following properties:*

1. (Anti-concentration) There exists a positive probability $p > 0$ such that for any unit vector $u \in \mathbb{R}^d$,

$$\mathbb{P}_{\zeta\sim\mathcal{D}^{TS}}(u^\top \zeta \geq 1) \geq p,$$

2. (Concentration) There exist positive constants c and c' such that for all $\delta \in (0, 1)$,

$$\mathbb{P}_{\zeta\sim\mathcal{D}^{TS}}\left(\|\zeta\| \leq \sqrt{cd\log(c'd/\delta)}\right) \geq 1 - \delta.$$

B.2 Examples of FP distributions

Below, we provide examples of distributions satisfying the anti-concentration and concentration properties.

Example 1: Gaussian distribution $\zeta \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ The concentration property comes directly from Lemma G.6, as the inner product of a standard multivariate normal random variable ζ and an arbitrary unit vector u follows a standard normal distribution. In the same manner, for a unit vector u,

$$\mathbb{P}_{\zeta\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left(u^\top\zeta \geq 1\right) = \mathbb{P}_{z\sim\mathcal{N}(0,1)}\left(z \geq 1\right) = \frac{1}{2}\text{erfc}(\frac{1}{\sqrt{2}}) \geq \frac{1}{4\sqrt{e\pi}}. \tag{B.1}$$

Thus, the standard Gaussian distribution satisfies the concentration property with $c = c' = 2$ and anti-concentration property with $p = \frac{1}{4\sqrt{e\pi}}$. Adjusting the scale of the covariance matrix, we can easily prove other variants satisfy the conditions.

Example 2: Uniform distribution $\zeta \sim \mathcal{U}_{B_d(\mathbf{0},\sqrt{d})}$ Let the random variable $\zeta = rv$, where $r = \|\zeta\| \in [0, \sqrt{d}]$ and $v = \zeta/\|\zeta\|$ is a unit vector. Then, $u^\top\zeta$ can be expressed as the product of two independent random variables, r and $u^\top v$ ($\sim$ Beta$(\frac{1}{2}, \frac{d-1}{2})$), as $r \cdot (u^\top v)$. These two random variables follow the distributions:

$$f_r(r) = \frac{dr^{d-1}}{\sqrt{d}^d}, \quad r \in [0, \sqrt{d}], \qquad \text{and} \qquad f_{u^\top v}(x) = \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{d-1}{2})}(1-x^2)^{\frac{d-3}{2}}, \quad x \in [-1, 1].$$

Based on these random variables, we can write:

$$f_{u^\top\zeta}(z) = \int_0^{\sqrt{d}}\int_{-1}^{1} \delta(z - rx) f_r(r) f_{u^\top v}(x) dx dr.$$

Using Monte Carlo simulations, we observe that $u\top\zeta$ has a lighter tail distribution compared a standard normal distribution. Accordingly, $c = c' = 2$ satisfies the concentration property. By proposition 9 and 10 in Abeille and Lazaric [2],

$$\mathbb{P}(u^\top\zeta \geq 1) = \frac{1}{2} I_{1-\frac{1}{d}}\left(\frac{d+1}{2}, \frac{1}{2}\right) \geq \frac{1}{16\sqrt{3\pi}},$$

where $I_x(a, b)$ is the incomplete regularized beta function. This suggests that the Uniform distribution satisfies the anti-concentration property with $p = \frac{1}{16\sqrt{3\pi}}$.

C Application of FP algorithm

C.1 Generalized version of `GLM-FP`

In Section 3, we introduced how FP can be applied to the contextual bandit settings in which the reward model extends beyond generalized linear models. We provide the general algorithmic framework below.

Algorithm C.1 Feature Perturbation in Bandit Problems

Input: Regularization parameter $\lambda > 0$, tuning parameter $\{c_t\}$
for $t = 1, 2, \ldots, T$ **do**
  Compute $\hat{f} = \text{argmin}_{f\in\mathcal{F}} \sum_{\tau=1}^{t-1}(f(x_{\tau,i_\tau}) - r_\tau)^2$ via a least squares oracle
  Sample $\widetilde{x}_{ti} \sim \mathcal{D}(x_{ti}, \Sigma_{ti})$ for all i ▷ e.g., $\mathcal{D}(x_{ti}, \Sigma_{ti}) = \mathcal{N}(x_{ti}, \boldsymbol{I}/t)$
  Select arm $i_t = \arg\max_{i\in[|\mathcal{X}_t|]} \hat{f}(\widetilde{x}_{ti})$ ▷ either i.i.d. or via shared perturbation
  Observe reward $r_t = f^*(x_{t,i_t}) + \xi_t$
end for

C.2 Application to the linear bandit problem

While line 4 in Algorithm C.1 merely defines the sampling distribution for each arm, in practice—mirroring the design of GLM-FP—one may introduce a shared perturbing factor that is first sampled and then applied to all arms. This construction induces dependencies across the perturbed arms and can serve as the basis for the arm selection mechanism.

Algorithm C.2 LinFP: Feature Perturbation in Linear bandits

1: **Input:** Regularization parameter $\lambda > 0$, tuning parameter $\{c_t\}$
2: **Initialize:** $V_1 \leftarrow \lambda \boldsymbol{I}$, $b_1 \leftarrow \mathbf{0}_d$
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Compute $\hat{\theta}_t = V_t^{-1} b_t$
5: Sample $\zeta_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ ▷ Shared perturbing factor
6: Compute $\widetilde{x}_{ti} = x_{ti} + c_t \cdot \frac{\|x_{ti}\|_{V_t^{-1}}}{\|\hat{\theta}_t\|} \cdot \zeta_t$ for all i
7: Choose $i_t = \arg\max_{i \in [|\mathcal{X}_t|]} \widetilde{x}_{ti}^\top \hat{\theta}_t$ ▷ $x_{ti}^\top \widetilde{\theta}_t \sim \mathcal{N}(x_{ti}^\top \hat{\theta}_t, c_t^2 \|x_{ti}\|_{V_t^{-1}}^2)$ for all i.
8: Observe reward $r_t = x_{t,i_t}^\top \theta^* + \xi_t$
9: Update$V_{t+1} = V_t + x_{t,i_t} x_{t,i_t}^\top, \quad b_{t+1} = b_t + x_{t,i_t} r_t$
10: **end for**

In Appendix H, we compare LinFP, the linear variant of our approach, to LinTS [2, 5]. The primary algorithmic difference lies in lines 6–7 of Algorithm C.2. In our method, the shared random vector ζ_t is used to perturb each feature vector. In contrast, LinTS perturbs the model parameter as $\widetilde{\theta}_t = \hat{\theta}_t + c_t \cdot V_t^{-1/2} \zeta_t$, and selects the arm maximizing $x_{ti}^\top \widetilde{\theta}_t$. For LinTS, we have:

$$\mathbb{E}[x_{ti}^\top \widetilde{\theta}_t] = x_{ti}^\top \hat{\theta}_t + c_t \cdot x_{ti}^\top V_t^{-1/2} \mathbb{E}[\zeta_t] = x_{ti}^\top \hat{\theta}_t$$
$$\mathrm{Var}[x_{ti}^\top \widetilde{\theta}_t] = c_t^2 \cdot \mathrm{Var}[x_{ti}^\top V_t^{-1/2} \zeta_t] = c_t^2 \cdot x_{ti}^\top V_t^{-1} x_{ti} = c_t^2 \|x_{ti}\|_{V_t^{-1}}^2$$

While the marginal distribution of $x_{ti}^\top \widetilde{\theta}_t$ under both methods is identical, the use of a shared perturbation ζ_t in LinFP induces algorithmic coupling across arms. This distinction is further discussed in Appendix H.

D Limitations

Our analysis focuses on structured settings such as generalized linear bandits (GLBs), where feature perturbation achieves both strong empirical performance and provable regret guarantees. While the same principle shows promise in more flexible or non-linear models, theoretical guarantees in these broader settings remain open. The current formulation, though practically effective, is heuristic outside the GLM framework and lacks formal justification under complex function classes. Extending the theory to overparameterized or general Lipschitz models represents an important direction for future work, where feature-level stochasticity may offer a stable alternative to parameter perturbation.

E Table of notations

This section introduces additional notations and concepts essential for the analysis. For ease of reference, Table E.1 summarizes the key notations defined in this paper, along with newly introduced notations. Conventional concepts such as d, T, $\mathcal{A}$, $\mathcal{C}$, $\mathcal{X}$, K or r are omitted here. The concepts will be reintroduced as needed in subsequent sections.

Table E.1: Notations and concepts for the analysis of the main theorem

Notation	**Definition**
M_μ	Self-concordance constant
$\mathcal{L}_\mu$	Lipschitz constant of the link function μ
$\mathcal{L}_t$	Lipshitz constant of the negative log-likelihood function
$\Theta_t(\delta, \lambda)$	$1-\delta$ probability ellipsoidal relaxed confidence set with regularization λ for the true parameter θ^*
$\beta_t(\delta)$	$\sqrt{4S^2\lambda + 2(1+SM_\mu)(\log(1/\delta) + d\log(2e\mathcal{L}_t/d))} = \widetilde{\mathcal{O}}(\sqrt{d})$
$\gamma_t(\delta)$	$\beta_t(\delta/(4T))\sqrt{c\log(4c'T/\delta)}$ (c, c': constant satisfying concentration property)
$\mathcal{E}_t(x)$	$\left\{\widetilde{x} \in \mathbb{R}^d \middle\| \|\langle \widetilde{x} - x, \hat{\theta}_t\rangle\| \le \gamma_t(\delta/(4T))\|x\|_{\hat{H}_t^{-1}}\right\}$
κ_*	Average derivative of link function at the true optimal arm over T rounds
κ	Minimum reachable derivative of link function
Warm-up stage	
I_T	$\left\{t \in [T] : \left(\left\|\sqrt{\dot{\mu}(x_t^\top \bar{\theta}_t)}x_t\right\|_{\bar{H}_t^{-1}} \ge 1\right) \vee \left(\|x_t\|_{V_t^{-1}} \ge 1\right)\right\}$
Taylor remainder term	
$\bar{\alpha}_t(x)$	$\int_0^1 (1-u)\dot{\mu}(x_t^\top \hat{\theta}_t + u(x^\top \hat{\theta}_t - x_t^\top \hat{\theta}_t))du$
$\bar{\alpha}_t(\theta, \nu)$	$\int_0^1 (1-u)\dot{\mu}\left(x_t^\top \theta + u(x_t^\top \nu - x_t^\top \theta)\right) du$
Matrices	
V_t	$\sum_{\tau=1}^{t-1} x_\tau x_\tau^\top + \lambda \boldsymbol{I}$
$\bar{V}_t$	$\sum_{\tau=1}^{t-1} x_\tau x_\tau^\top + \lambda \boldsymbol{I}/\kappa$
$\hat{H}_t$	$\sum_{\tau=1}^{t-1} \dot{\mu}(x_\tau^\top \hat{\theta}_t) x_\tau x_\tau^\top + \lambda \boldsymbol{I}$
$\bar{H}_t$	$\sum_{\tau=1}^{t-1} \dot{\mu}(x_\tau^\top \bar{\theta}_\tau) x_\tau x_\tau^\top + \lambda \boldsymbol{I}$
$\widetilde{H}_t(\theta, \nu)$	$\sum_{\tau=1}^{t-1} \bar{\alpha}_\tau(\theta, \nu) x_\tau x_\tau^\top$
Other notations	
R^*	$\max_{x \in \mathcal{X}} \lvert\mu(x^\top \theta^*)\rvert$
$\bar{\theta}_t$	$\operatorname{argmin}_{\theta \in \cup_{\tau \in [t,T]} \Theta_\tau(\delta, \lambda)} \dot{\mu}(x_t^\top \theta)$
$(\tau(t), \omega_t)$	$\operatorname{argmax}_{\tau \in [t,T], \theta \in \Theta_\tau(\delta, \lambda)} \left\lvert\mu(x_t^\top \theta) - \mu(x_t^\top \hat{\theta}_\tau)\right\rvert$
$(p, c_{\delta'})$	constants related to standard normal distribution ($p = 1/(4\sqrt{e\pi})$, $c_{\delta'} = \sqrt{2\log(2/\delta')}$)

F Proof of main theorem

The first step in our proof is to derive the high-probability confidence bound for the estimate $\hat{\theta}_t$. Using this bound, we ensure that both concentration and anti-concentration events occur with high probability. Finally, we compute the regret bound of our algorithm under these events.

F.1 Confidence bound for the true parameter

An important step in analyzing the regret bound of the algorithm is to establish a confidence set for the underlying parameter θ^*. This involves constructing a region that reliably contains θ^* throughout the learning process. To obtain a practical and tighter bound, we adopt confidence sets derived from the log-likelihood function using an *ellipsoidal relaxation*.

Proposition F.1 (Adapted from Theorem 3.2. in Lee et al. [34])**.** *Let $\mathcal{L}_t := \max_{\theta\in\Theta} \|\nabla L_t(\theta)\|$ be the Lipschitz constant of the loss function, and $M_\mu := \sup_{x\in\mathcal{X},\theta\in\Theta} |\ddot{\mu}(\langle x,\theta\rangle)| / \dot{\mu}(\langle x,\theta\rangle)$. For any $\lambda > 0$, define the regularized Hessian $\hat{H}_t := \nabla^2 L_t(\hat{\theta}_t) + \lambda I$. Then, with probability at least $1-\delta$, for all $t \geq 1$, it holds that $\theta^* \in \Theta_t(\delta,\lambda) := \{\theta \in \mathbb{R}^d \mid \|\theta - \hat{\theta}_t\|_{\hat{H}_t} \leq \beta_t(\delta)\}$, where*

$$\beta_t(\delta) = \sqrt{4\lambda + 2(1+M_\mu)\left(\log\frac{1}{\delta} + d\log\left(\frac{2e\mathcal{L}_t}{d}\right)\right)}.^2$$

Proof. We rederived Theorem 3.2 of Lee et al. [34] to obtain a tighter bound with improved dependence on the regularization parameter λ, such that the λ term no longer scales with M_μ.

By Theorem 3.1. in [34], with probability at least $1-\delta$, for all $t \geq 1$, the following inequality holds:

$$L_t(\theta^*) - L_t(\hat{\theta}_t) \leq \log\frac{1}{\delta} + d\log\left(\frac{2eS\mathcal{L}_t}{d}\right) := \mathcal{W}_t(\delta)^2$$

Then we observe:

$$\begin{aligned}\int_0^1 (1-u)\nabla^2 L_t(\hat{\theta}_t + u(\theta^* - \hat{\theta}_t))du &= \int_0^1 (1-u)\sum_{\tau=1}^{t-1} \dot{\mu}(x_\tau^\top(\hat{\theta}_t + u(\theta^* - \hat{\theta}_t))x_\tau x_\tau^\top du \\ &= \sum_{\tau=1}^{t-1} \underbrace{\left(\int_0^1 (1-u)\dot{\mu}(x_\tau^\top(\hat{\theta}_t + u(\theta^* - \hat{\theta}_t))du\right)}_{\bar{\alpha}_\tau(\hat{\theta}_t,\theta^*)} x_\tau x_\tau^\top = \widetilde{H}_t(\hat{\theta}_t,\theta^*),\end{aligned}$$

where the second equality follows from Fubini's theorem, where the order of the integral and the summation can be switched. Using Taylor's theorem with an integral remainder (Proposition G.2), we can further deduce that, with probability $1-\delta$:

$$\begin{aligned}\mathcal{W}_t(\delta)^2 &\geq L_t(\theta^*) - L_t(\hat{\theta}_t) \\ &= \langle \nabla L_t(\hat{\theta}_t), \theta^* - \hat{\theta}_t\rangle + (\theta^* - \hat{\theta}_t)^\top \left(\int_0^1 (1-u)\nabla^2 L_t(\hat{\theta}_t + u(\theta^* - \hat{\theta}_t))du\right)(\theta^* - \hat{\theta}_t).\end{aligned}$$

As the optimality condition at $\hat{\theta}_t$ infers that $\langle \nabla L_t(\hat{\theta}_t), \theta^* - \hat{\theta}_t\rangle \geq 0$ and by equation (Proposition G.1), we have that

$$\mathcal{W}_t(\delta)^2 \geq \left\|\theta^* - \hat{\theta}_t\right\|^2_{\widetilde{H}_t(\hat{\theta}_t,\theta^*)} \geq \frac{1}{2+2SM_\mu}\left\|\theta^* - \hat{\theta}_t\right\|^2_{\hat{H}_t - \lambda\boldsymbol{I}}, \qquad \text{(Proposition G.1)}$$

where the last inequality holds, since

$$\widetilde{H}_t(\hat{\theta}_t,\theta^*) \succeq \sum_{\tau=1}^{t-1}\left(\frac{\dot{\mu}(x_\tau^\top\hat{\theta}_t)}{2+2SM_\mu}\right) x_\tau x_\tau^\top = \frac{1}{2+2SM\mu}(\hat{H}_t - \lambda\boldsymbol{I}).$$

Accordingly,

$$\left\|\theta^* - \hat{\theta}_t\right\|^2_{\hat{H}_t} \leq \left\|\theta^* - \hat{\theta}_t\right\|^2_{\hat{H}_t - \lambda\boldsymbol{I}} + \lambda\left\|\theta^* - \hat{\theta}_t\right\|^2 \leq 4S^2\lambda + 2(1+SM_\mu)\mathcal{W}_t(\delta)^2 = \beta_t(\delta)^2,$$

and by Assumption 1, we finish a proof. □

[2] It has been shown by Lee et al. [34] that $\mathcal{L}_t = \mathcal{O}(t)$ for linear, logistic, and Poisson bandit instances.

F.2 Proof of Lemma 1-Stochastic Optimism

Let the event $\ddot{E}_t$ be defined as $\ddot{E}_t := \left\{\mu(\widetilde{x}_t^\top \hat{\theta}_t) \geq \mu(x_{t*}^\top \theta^*)\right\}$. This event corresponds to the case where the chosen perturbed feature vector yields an optimistic expected reward relative to the true optimal arm at step t. To bound the anti-concentration probability as required in Lemma 1, we aim to lower bound $\mathbb{P}_t(\ddot{E}_t \mid \hat{E}_t, \widetilde{E}_t)$, conditioned on the events $\hat{E}_t$ and $\widetilde{E}_t$. For any $t \in [T]$, we have:

$$
\begin{aligned}
\mathbb{P}_t\left(\ddot{E}_t \mid \hat{E}_t, \widetilde{E}_t\right) &= \mathbb{P}_t\left(\mu(\widetilde{x}_t^\top \hat{\theta}_t) \geq \mu(x_{t*}^\top \theta^*) \mid \hat{E}_t, \widetilde{E}_t\right) \\
&= \mathbb{P}_t\left(\widetilde{x}_t^\top \hat{\theta}_t \geq x_{t*}^\top \theta^* \mid \hat{E}_t, \widetilde{E}_t\right) && (\because \mu \text{ is strictly increasing}) \\
&\geq \mathbb{P}_t\left(\widetilde{x}_{t*}^\top \hat{\theta}_t - x_{t*}^\top \hat{\theta}_t \geq x_{t*}^\top \theta^* - x_{t*}^\top \hat{\theta}_t \mid \hat{E}_t, \widetilde{E}_t\right) && (\because x_t = \operatorname*{argmax}_{i \in [|\mathcal{X}_t|]} x_{ti}^\top \hat{\theta}_t) \\
&\geq \mathbb{P}_t\left(\langle \widetilde{x}_{t*} - x_{t*}, \hat{\theta}_t \rangle \geq \left|\langle x_{t*}, \theta^* - \hat{\theta}_t \rangle\right| \;\middle|\; \hat{E}_t, \widetilde{E}_t\right) \\
&\geq \mathbb{P}_t\left(\left(\beta_t(\delta') \frac{\|x_{t*}\|_{\hat{H}_t^{-1}}}{\|\hat{\theta}_t\|} \zeta_t\right)^\top \hat{\theta}_t \geq \|x_{t*}\|_{\hat{H}_t^{-1}} \|\hat{\theta}_t - \theta^*\|_{\hat{H}_t} \;\middle|\; \hat{E}_t, \widetilde{E}_t\right) \\
&\geq \mathbb{P}\left(\beta_t(\delta') \|x_{t*}\|_{\hat{H}_t^{-1}} \cdot \langle \zeta_t, u_t \rangle \geq \beta_t(\delta') \|x_{t*}\|_{\hat{H}_t^{-1}} \mid \hat{E}_t, \widetilde{E}_t\right) \\
&= \mathbb{P}\left(\langle \zeta_t, u_t \rangle \geq 1\right) \geq \frac{1}{4\sqrt{e\pi}} := p,
\end{aligned}
$$

where the third inequality follows from the Cauchy-Schwarz inequality, and the fourth from the assumption that under the event $\hat{E}_t$, we have $\|\theta^* - \hat{\theta}_t\|_{\hat{H}_t} \leq \beta_t(\delta')$. The final inequality follows from the anti-concentration property of the standard normal distribution, as detailed in (B.1). For simplicity, we henceforth fix $p := 1/(4\sqrt{e\pi})$ as the corresponding lower bound on this probability.

F.3 Proof of Lemma 1-Concentration

We now proceed to establish how the probability of each event defined in Lemma 1 can be ensured using the confidence bound $\beta_t(\delta)$ derived in Appendix F.1. Each event is analyzed and bounded individually, and the results are then combined to complete the proof of the lemma.

Bounding $\hat{E}_T$ Let $\delta' = \delta/(4T)$. By the choice of $\beta_t(\delta)$ in Proposition F.1, we have that

$$
\begin{aligned}
\forall 1 \leq t \leq T, \quad & \mathbb{P}\left(\left\|\hat{\theta}_t - \theta^*\right\|_{\hat{H}_t} \leq \beta_t(\delta')\right) \geq 1 - \delta' \\
\text{from union bound,} \quad & \mathbb{P}\left(\bigcap_{t=1}^{T} \left\{\left\|\hat{\theta}_t - \theta^*\right\|_{\hat{H}_t} \leq \beta_t(\delta')\right\}\right) \geq 1 - \sum_{t=1}^{T} \mathbb{P}\left(\left\|\hat{\theta}_t - \theta^*\right\|_{\hat{H}_t} \geq \beta_t(\delta')\right) \\
\implies \quad & \mathbb{P}\left(\bigcap_{t=1}^{T} \left\{\left\|\hat{\theta}_t - \theta^*\right\|_{\hat{H}_t} \leq \beta_t(\delta')\right\}\right) \geq 1 - \sum_{t=1}^{T} \delta' \\
\implies \quad & \mathbb{P}(\hat{E}_T) \geq 1 - T\delta' = 1 - \frac{\delta}{4}.
\end{aligned}
$$

Bounding $\widetilde{E}_T$ The expression for the perturbed feature vector $\widetilde{x}_{ti}$ is given as the expression $\widetilde{x}_{ti} = x_{ti} + \beta_t(\delta') \frac{\|x_{ti}\|_{\hat{H}_t^{-1}}}{\|\hat{\theta}_t\|} \zeta_t$ with the choice of $c_t = \beta_t(\delta')$, where ζ_t is drawn $i.i.d.$ from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Note that the constants c and c' using standard multivariate normal distribution are both 2, from now on for simplicity, we let $c_{\delta'} := \sqrt{2\log(2/\delta')}$. Since all arms are coupled[3] with same ζ_t, we can write

[3]Uncoupled sampling means ζ_{ti}'s are sampled for each arm respectively and it results in extra $\log K$ term in the regret bound because of the union bound over the number of arms at each step.

$$\forall 1 \le t \le T, \quad \mathbb{P}\left(\forall x_{ti} \in \mathcal{X}_t; \quad \left|\langle \widetilde{x}_{ti} - x_{ti}, \hat{\theta}_t \rangle\right| \le \gamma_t(\delta')\|x_{ti}\|_{\hat{H}_t^{-1}}\right) \tag{F.1}$$

$$= \mathbb{P}\left(\forall x_{ti} \in \mathcal{X}_t; \quad \beta_t(\delta')\|x_{ti}\|_{\hat{H}_t^{-1}} \left|\langle \zeta_t, \frac{\hat{\theta}_t}{\|\hat{\theta}_t\|} \rangle\right| \le c_{\delta'} \cdot \beta_t(\delta')\|x_{ti}\|_{\hat{H}_t^{-1}}\right) \tag{F.2}$$

$$= \mathbb{P}\left(|\langle \zeta_t, u_t \rangle| \le c_{\delta'}\right) \ge 1 - \delta', \tag{F.3}$$

where u_t is a unit vector. The first equality holds by the definition of $c_{\delta'}$ and Lemma 1, and the inequality follows from the concentration property. The cancellation in the second equality plays a critical role in removing arm-wise dependence in GLB setting. A union bound over T rounds yields

$$\mathbb{P}(\widetilde{E}_T) \ge 1 - T\delta' = 1 - \frac{\delta}{4}.$$

Finally, applying the union bound across the events $\hat{E}_T$ and $\widetilde{E}_T$, we have that

$$\mathbb{P}(\hat{E}_T \cap \widetilde{E}_T) \ge 1 - \frac{\delta}{2}.$$

Remark. To guarantee the same probability level $1 - \delta'$ as in equations (F.1)–(F.3), which bound the randomness arising from perturbations to the feature vectors, Thompson Sampling requires the confidence parameters β and γ to be inflated by an additional factor of $\sqrt{d}$. This inflation arises due to the right-hand side of the concentration bound in Definition B.1, which scales with $\sqrt{d}$. Such adjustment is necessary to control the deviation in the perturbed estimated expected reward, which takes the form $x^\top(\widetilde{\theta} - \hat{\theta})$. This is consistent with the reasoning discussed in Appendix H.

F.4 Proof of Theorem 1

In this section, we establish the regret guarantee for our algorithm. Given the complexity of the analysis, we divide the proof into multiple steps. Supporting lemmas and their proofs are deferred to Appendix G.

Step 1 (Warm-up) We begin by partitioning the T rounds into a "warm-up" stage and the primary stage. The set of time steps corresponding to the primary stage is defined as:

$$I_T := \left\{ t \in [T] : \left(\left\| \sqrt{\dot{\mu}(x_t^\top \bar{\theta}_t)} x_t \right\|_{\bar{H}_t^{-1}} \le 1 \right) \wedge \left(\|x_t\|_{\bar{V}_t^{-1}} \le 1 \right) \right\},$$

where $\bar{H}_t, \bar{V}_t$, and $\bar{\theta}_t$ are given by:

$$\bar{H}_t := \lambda \boldsymbol{I} + \sum_{\tau=1}^{t-1} \dot{\mu}(x_\tau^\top \bar{\theta}_\tau) x_\tau x_\tau^\top, \qquad \bar{V}_t := \lambda \boldsymbol{I}/\kappa + \sum_{\tau=1}^{t-1} x_\tau x_\tau^\top, \qquad \bar{\theta}_t := \operatorname*{argmin}_{\theta \in \cup_{\tau \in [t,T]} \Theta_\tau(\delta,\lambda)} \dot{\mu}(x_t^\top \theta).$$

The introduction of $\bar{H}_t$ is crucial because $\hat{H}_t = \lambda \boldsymbol{I} + \sum_{\tau=1}^{t-1} \dot{\mu}(x_\tau^\top \hat{\theta}_t) x_\tau x_\tau^\top$ depends on t, which prevents direct application of the Elliptical Potential Lemma (EPL; Lemma G.1), as discussed in Section 4. To address this, we leverage $\bar{H}_t$, which incorporates the minimum derivative of μ within future confidence sets ($\bar{\theta}_\tau$), ensuring it serves as a smaller Gram matrix suitable for bounding the regret. Similarly, $\bar{V}_t$ is introduced to directly apply EPL.

Next, we bound each weighted 2-norm using the Elliptical Potential Count Lemma (EPCL; Lemma G.2), which guarantees that the regret incurred during the warm-up phase remains manageable. Consequently, the cumulative regret over T rounds is decomposed as follows:

$$\begin{aligned} R(T) &= \underbrace{\sum_{t \in I_T} \left\{\mu(x_{t*}^\top \theta^*) - \mu(x_t^\top \theta^*)\right\}}_{\text{Reg}(T)} + \underbrace{\sum_{t \notin I_T} \left\{\mu(x_{t*}^\top \theta^*) - \mu(x_t^\top \theta^*)\right\}}_{\text{warm-up regret}} \\ &\le \text{Reg}(T) + 2R^* \left(\sum_{t=1}^{T} \mathbb{1}\left\{ \left\| \sqrt{\dot{\mu}(x_t^\top \bar{\theta}_t)} x_t \right\|_{\bar{H}_t^{-1}} \ge 1 \right\} + \sum_{t=1}^{T} \mathbb{1}\left\{ \|x_t\|_{\bar{V}_t^{-1}} \ge 1 \right\} \right) \\ &\le \text{Reg}(T) + \frac{4dR^*}{\log 2} \left\{ \log\left(1 + \frac{\mathcal{L}_\mu}{\lambda \log 2}\right) + \log\left(1 + \frac{\kappa}{\lambda \log 2}\right) \right\}, \end{aligned}$$

where $R^* := \max_{x\in\mathcal{X}} \left|\mu(x^\top\theta^*)\right|$ is the maximum expected reward achievable under the underlying model.The first and the second inequalities hold from the definition of I_T and by EPCL (Lemma G.2), respectively. Note that the warm-up regret is $\widetilde{\mathcal{O}}(d)$, which is independent of T.

Step 2-1 (Decomposition) We decompose the cumulative regret for the primary stage into three components:

$$\mathrm{Reg}(T) = \sum_{t\in I_T} \Bigg(\underbrace{\left\{\mu(x_{t*}^\top\theta^*) - \mu(\widetilde{x}_t^\top\hat{\theta}_t)\right\}}_{A_t} + \underbrace{\left\{\mu(\widetilde{x}_t^\top\hat{\theta}_t) - \mu(x_t^\top\hat{\theta}_t)\right\}}_{B_t} + \underbrace{\left\{\mu(x_t^\top\hat{\theta}_t) - \mu(x_t^\top\theta^*)\right\}}_{C_t} \Bigg).$$

Here, A_t and B_t relate to the perturbations' effect on the estimated reward, while C_t concerns the closeness of $\hat{\theta}_t$ to θ^*. We will bound each term under the events $\hat{E}_t$ and $\widetilde{E}_t$.

Bounding C_t Bounding C_t is straightforward. Using the confidence set $\Theta_t(\delta,\lambda)$, abbreviated as Θ_t, we define $(\tau(t),\omega_t)$ as the pair maximizing the confidence width computed on the selected action at round t, x_t, after round t: $\mathrm{argmax}_{\tau\in[t,T],\theta\in\Theta_\tau} \left|\mu(x_t^\top\theta) - \mu(x_t^\top\hat{\theta}_\tau)\right|$. Under the event $\hat{E}_t$, we know that $\theta^*\in\Theta_\tau$ for all $t\le\tau\le T$. Thus,

$$C_t\mathbb{1}\left\{\hat{E}_t\cap\widetilde{E}_t\right\} = \left(\mu(x_t^\top\hat{\theta}_t) - \mu(x_t^\top\theta^*)\right)\mathbb{1}\left\{\hat{E}_t\cap\widetilde{E}_t\right\} \le \left|\mu(x_t^\top\omega_t) - \mu(x_t^\top\hat{\theta}_{\tau(t)})\right|\mathbb{1}\left\{\hat{E}_t\cap\widetilde{E}_t\right\}.$$

Bounding B_t Let $\widetilde{x}_t^* := \mathrm{argmax}_{x\in\mathcal{E}_t(x_t)} \left|\mu(x^\top\hat{\theta}_t) - \mu(x_t^\top\hat{\theta}_t)\right|$. Under the event $\widetilde{E}_t$, $\widetilde{x}_t\in\mathcal{E}_t(x_t)$ holds and we can write:

$$B_t\mathbb{1}\left\{\hat{E}_t\cap\widetilde{E}_t\right\} = \left(\mu(\widetilde{x}_t^\top\hat{\theta}_t) - \mu(x_t^\top\hat{\theta}_t)\right)\mathbb{1}\left\{\hat{E}_t\cap\widetilde{E}_t\right\} \le \left|\mu(\widetilde{x}_t^{*\top}\hat{\theta}_t) - \mu(x_t^\top\hat{\theta}_t)\right|\mathbb{1}\left\{\hat{E}_t\cap\widetilde{E}_t\right\}.$$

Before proceeding, note that for $\widetilde{x}_t\in\mathcal{E}_t(x_t)$, we can derive an upper bound using Taylor's theorem with an integral remainder (Proposition G.2). Define $\bar{\alpha}_t(x) = \int_0^1(1-u)\dot{\mu}(x_t^\top\hat{\theta}_t + u(x^\top\hat{\theta}_t - x_t^\top\hat{\theta}_t))du$, which accounts for higher-order terms based on the estimated parameter $\hat{\theta}_t$ and feature vectors x and x_t. The difference $\left|\mu(\widetilde{x}_t^\top\hat{\theta}_t) - \mu(x_t^\top\hat{\theta}_t)\right|$ then can be bounded as:

$$\begin{aligned}
\left|\mu(\widetilde{x}_t^\top\hat{\theta}_t) - \mu(x_t^\top\hat{\theta}_t)\right| &= \left|\dot{\mu}(x_t^\top\hat{\theta}_t)\langle\widetilde{x}_t - x_t,\hat{\theta}_t\rangle + \int_{x_t^\top\hat{\theta}_t}^{\widetilde{x}_t^\top\hat{\theta}_t}(\mu(\widetilde{x}_t^\top\hat{\theta}_t) - z)\ddot{\mu}(z)dz\right| \\
&\le \dot{\mu}(x_t^\top\hat{\theta}_t)\left|\langle\widetilde{x}_t - x_t,\hat{\theta}_t\rangle\right| + \langle\widetilde{x}_t - x_t,\hat{\theta}_t\rangle^2\int_0^1(1-u)\left|\ddot{\mu}\left(x_t^\top\hat{\theta}_t + u(\widetilde{x}_t^\top\hat{\theta}_t - x_t^\top\hat{\theta}_t)\right)\right|du \\
&\le \dot{\mu}(x_t^\top\hat{\theta}_t)\left|\langle\widetilde{x}_t - x_t,\hat{\theta}_t\rangle\right| + M_\mu\langle\widetilde{x}_t - x_t,\hat{\theta}_t\rangle^2\underbrace{\int_0^1(1-u)\dot{\mu}\left(x_t^\top\hat{\theta}_t + u(\widetilde{x}_t^\top\hat{\theta}_t - x_t^\top\hat{\theta}_t)\right)du}_{=\bar{\alpha}_t(\widetilde{x}_t)} \\
&\le \dot{\mu}(x_t^\top\hat{\theta}_t)\left|\langle\widetilde{x}_t - x_t,\hat{\theta}_t\rangle\right| + M_\mu\bar{\alpha}_t(\widetilde{x}_t)\langle\widetilde{x}_t - x_t,\hat{\theta}_t\rangle^2 \\
&\le \dot{\mu}(x_t^\top\hat{\theta}_t)\gamma_t(\delta')\|x_t\|_{\hat{H}_t^{-1}} + M_\mu\bar{\alpha}_t(\widetilde{x}_t)\gamma_t(\delta')^2\|x_t\|_{\hat{H}_t^{-1}}^2,
\end{aligned}$$

where the second and the last inequalities hold from the definition of M_μ and $\mathcal{E}_t(x_t)$ in Lemma 1. This bound captures both the linear and higher-order contributions to the regret from perturbations in the feature vectors.

Bounding A_t With $\ddot{E}_t := \left\{\mu(\widetilde{x}_t^\top \hat{\theta}_t) \geq \mu(x_{t*}^\top \theta^*)\right\}$ and $\widetilde{x}_t^*$ defined above, we write:

$$
\begin{aligned}
A_t \mathbb{1}\left\{\hat{E}_t \cap \widetilde{E}_t\right\} &= \left(\mu(x_{t*}^\top \theta^*) - \mu(\widetilde{x}_t^\top \hat{\theta}_t)\right) \mathbb{1}\left\{\hat{E}_t \cap \widetilde{E}_t\right\} \\
&\leq \left(\mu(x_{t*}^\top \theta^*) - \inf_{\widetilde{\widetilde{x}}_t \in \mathcal{E}_t(x_t)} \mu(\widetilde{\widetilde{x}}_t^\top \hat{\theta}_t)\right) \mathbb{1}\left\{\hat{E}_t \cap \widetilde{E}_t\right\} \\
&\leq \mathbb{E}_t\left[\left(\mu(\widetilde{x}_t^\top \hat{\theta}_t) - \inf_{\widetilde{\widetilde{x}}_t \in \mathcal{E}_t(x_t)} \mu(\widetilde{\widetilde{x}}_t^\top \hat{\theta}_t)\right) \mathbb{1}\left\{\hat{E}_t \cap \widetilde{E}_t\right\} \middle| \ddot{E}_t\right] \\
&= \mathbb{E}_t\left[\left(\mu(\widetilde{x}_t^\top \hat{\theta}_t) - \mu(x_t^\top \hat{\theta}_t)\right) + \left(\mu(x_t^\top \hat{\theta}_t) - \inf_{\widetilde{\widetilde{x}}_t \in \mathcal{E}_t(x_t)} \mu(\widetilde{\widetilde{x}}_t^\top \hat{\theta}_t)\right) \middle| \hat{E}_t, \widetilde{E}_t, \ddot{E}_t\right] \mathbb{P}(\hat{E}_t \cap \widetilde{E}_t) \\
&\leq 2\mathbb{E}_t\left[\left(\sup_{\widetilde{\widetilde{x}}_t \in \mathcal{E}_t(x_t)} \left|\mu(\widetilde{\widetilde{x}}_t^\top \hat{\theta}_t) - \mu(x_t^\top \hat{\theta}_t)\right|\right) \middle| \hat{E}_t, \widetilde{E}_t, \ddot{E}_t\right] \mathbb{P}(\hat{E}_t \cap \widetilde{E}_t) \\
&\leq \frac{2}{p}\mathbb{E}_t\left[\left|\mu(\widetilde{x}_t^{*\top} \hat{\theta}_t) - \mu(x_t^\top \hat{\theta}_t)\right| \mathbb{1}\left\{\hat{E}_t \cap \widetilde{E}_t\right\}\right].
\end{aligned}
$$

We justify the second inequality under the specified event, and the final inequality follows from the following reasoning: we use the bound $C \leq \mathbb{E}[Z \mid Z \geq C]$, and compensate for introducing the conditional expectation by incorporating the inverse of the probability of the conditioning event. Specifically, define $C := (\mu(x_{t*}^\top \theta^*) - \inf_x \mu(x^\top \hat{\theta}_t) \cdot \mathbb{1}\{(\} \hat{E}_t \cap \widetilde{E}_t)$ and $Z := (\mu(\widetilde{x}_t^\top \hat{\theta}_t) - \inf_x \mu(x^\top \hat{\theta}_t) \cdot \mathbb{1}\{(\} \hat{E}_t \cap \widetilde{E}_t)$. Then the second inequality holds. To compensate for conditioning on the favorable event, we use the following logic:

$$
\begin{aligned}
\mathbb{E}_t\left\{\cdot \mid \hat{E}_t, \widetilde{E}_t\right\} &\geq \mathbb{E}_t\left\{\cdot \mid \hat{E}_t, \widetilde{E}_t, \ddot{E}_t\right\} \mathbb{P}_t(\ddot{E}_t \mid \hat{E}_t, \widetilde{E}_t) \\
&\geq \mathbb{E}_t\left\{\cdot \mid \hat{E}_t, \widetilde{E}_t, \ddot{E}_t\right\} \cdot p. \qquad \text{(Lemma 1)}
\end{aligned}
$$

The upper bound for this term is similar to the previous part (B_t), differing only by a constant and the inclusion of the expectation over the filtration. However, since our goal is to bound the cumulative sum over T rounds rather than the expectation itself, directly handling the expectation complicates the application of the Elliptical Potential Lemma (EPL). To address this, we eliminate the expectation at the cost of introducing a concentration error, which we control using Azuma-Hoeffding's inequality.

Step 2-2 (Azuma-Hoeffding's inequality) Unless otherwise specified, we now analyze the regret bound under the assumption that events $\hat{E}_T$ and $\widetilde{E}_T$ hold.

$$
\begin{aligned}
&\frac{p}{2}\sum_{t \in I_T} A_t \leq \sum_{t \in I_T} \mathbb{E}_t\left[\left|\mu(\widetilde{x}_t^{*\top} \hat{\theta}_t) - \mu(x_t^\top \hat{\theta}_t)\right|\right] \\
&\leq \gamma_T(\delta') \sum_{t \in I_T} \dot{\mu}(x_t^\top \hat{\theta}_t)\|x_t\|_{\hat{H}_t^{-1}} + \gamma_T(\delta') \underbrace{\sum_{t \in I_T}\left(\mathbb{E}_t\left[\dot{\mu}(x_t^\top \hat{\theta}_t)\|x_t\|_{\hat{H}_t^{-1}}\right] - \dot{\mu}(x_t^\top \hat{\theta}_t)\|x_t\|_{\hat{H}_t^{-1}}\right)}_{R_1} \\
&\quad + M_\mu \gamma_T(\delta')^2 \sum_{t \in I_T} \bar{\alpha}_t(\widetilde{x}_t)\|x_t\|^2_{\hat{H}_t^{-1}} + M_\mu \gamma_T(\delta')^2 \underbrace{\sum_{t \in I_T}\left(\mathbb{E}_t\left[\bar{\alpha}_t(\widetilde{x}_t)\|x_t\|^2_{\hat{H}_t^{-1}}\right] - \bar{\alpha}_t(\widetilde{x}_t)\|x_t\|^2_{\hat{H}_t^{-1}}\right)}_{R_2}
\end{aligned}
$$

Note that R_1 and R_2 are constructed as martingales. Since the norm of each feature vector satisfies $\|x_t\| \leq 1$, and given $\hat{H}_t^{-1} \preceq \hat{H}_0^{-1} = \boldsymbol{I}/\lambda$ and $\dot{\mu}(x_t^\top \hat{\theta}_t) \leq \mathcal{L}_\mu$, the following bounds hold:

$$
0 \leq \dot{\mu}(x_t^\top \hat{\theta}_t)\|x_t\|_{\hat{H}_t^{-1}} \leq \mathcal{L}_\mu \sqrt{x_t^\top \hat{H}_t^{-1} x_t} \leq \mathcal{L}_\mu \sqrt{\frac{1}{\lambda}\|x_t\|^2} \leq \frac{\mathcal{L}_\mu}{\sqrt{\lambda}}.
$$

This provides an upper bound for each instantaneous element of R_1 as $\mathcal{L}_\mu/\sqrt{\lambda}$. Applying Azuma-Hoeffding's inequality (Proposition G.3), with probability at least $1 - \delta/4$, we obtain:

$$
R_1 \leq \sqrt{\frac{2T\mathcal{L}_\mu^2}{\lambda} \log \frac{8}{\delta}}.
$$

Due to the convexity of Θ_t, the term $\dot{\mu}$ in $\bar{\alpha}_t(x_t)$ is bounded by $\mathcal{L}_\mu$. Consequently, we obtain:

$$\bar{\alpha}_t(x) := \int_0^1 (1-u)\dot{\mu}\left(x_t^\top \hat{\theta}_t + u(x^\top \hat{\theta}_t - x_t^\top \hat{\theta}_t)\right) du \leq \int_0^1 (1-u)\mathcal{L}_\mu du = \mathcal{L}_\mu/2 \tag{F.4}$$

Similarly, we can show that

$$0 \leq \bar{\alpha}_t(\widetilde{x}_t)\|x_t\|^2_{\hat{H}_t^{-1}} \leq \frac{\mathcal{L}_\mu}{2}\left(x_t^\top \hat{H}_t^{-1} x_t\right) \leq \frac{\mathcal{L}_\mu}{2\lambda}\|x_t\|^2 \leq \frac{\mathcal{L}_\mu}{2\lambda}.$$

This provides an upper bound for each instantaneous element of R_2 as $\mathcal{L}_\mu/(2\lambda)$. Applying Proposition G.3, with probability at least $1-\delta/4$, we have:

$$R_2 \leq \sqrt{\frac{2T\mathcal{L}_\mu^2}{4\lambda^2}\log\frac{8}{\delta}}.$$

By applying a union bound, we conclude that with probability at least $1-\delta/2$, both R_1 and R_2 are bounded. Therefore, the regret term $\sum_{t\in I_T} A_t$ can be bounded by:

$$\sum_{t\in I_T} A_t \leq \frac{2}{p}\left(\gamma_T(\delta')\sum_{t\in I_T}\dot{\mu}(x_t^\top\hat{\theta}_t)\|x_t\|_{\hat{H}_t^{-1}} + M_\mu\gamma_T(\delta')^2\sum_{t\in I_T}\bar{\alpha}_t(\widetilde{x}_t)\|x_t\|^2_{\hat{H}_t^{-1}}\right) + \varepsilon,$$

where ε is a function of T, δ, and λ, defined as:

$$\varepsilon = \varepsilon(T,\delta,\lambda) := \frac{2}{p}\left(\gamma_t(\delta')\sqrt{\frac{2T\mathcal{L}_\mu^2}{\lambda}\log\frac{8}{\delta}} + M_\mu\gamma_T(\delta')^2\sqrt{\frac{2T\mathcal{L}_\mu^2}{4\lambda^2}\log\frac{8}{\delta}}\right) = \widetilde{\mathcal{O}}\left(d\sqrt{\frac{T}{\lambda^2}}\right).$$

In summary, combining this result with Lemma 1 and the bounds for B_t and C_t, the total regret $\mathrm{Reg}(T)$ can be bounded with probability at least $1-\delta$ as:

$$\begin{aligned}\mathrm{Reg}(T) \leq \left(\frac{2}{p}+1\right)&\underbrace{\sum_{t\in I_T}\left(\gamma_T(\delta')\dot{\mu}(x_t^\top\hat{\theta}_t)\|x_t\|_{\hat{H}_t^{-1}} + M_\mu\gamma_T(\delta')^2\bar{\alpha}_t(\widetilde{x}_t)\|x_t\|^2_{\hat{H}_t^{-1}}\right)}_{\mathrm{Reg}_{\mathrm{FP}}}\\ &+\underbrace{\sum_{t\in I_T}\left|\mu(x_t^\top\omega_t) - \mu(x_t^\top\hat{\theta}_{\tau(t)})\right|}_{\mathrm{Reg}_{\mathrm{EST}}} + \varepsilon\end{aligned}$$

Here, $\mathrm{Reg}_{\mathrm{FP}}$ captures the regret arising from perturbing the feature vectors, while $\mathrm{Reg}_{\mathrm{EST}}$ accounts for the estimation error of $\hat{\theta}_t$ compared to the true parameter θ^*.

Step 3 (Bounding $\mathrm{Reg}_{\mathrm{FP}}$) The presence of $\hat{H}_t^{-1}$ in the weighted norm makes it challenging to directly apply the Elliptical Potential Lemma (EPL; Lemma G.1). To address this, we introduce $\bar{H}_t$, which allows us to leverage EPL by splitting $\dot{\mu}(x_t^\top\hat{\theta}_t)$ into two components: a leading term based on $\bar{\theta}_t$ (used in defining $\bar{H}_t$) and a transient term that accounts for deviations from $\hat{\theta}_t$.

$$\begin{aligned}\mathrm{Reg}_{\mathrm{FP}} &= \gamma_T(\delta')\sum_{t\in I_T}\dot{\mu}(x_t^\top\hat{\theta}_t)\|x_t\|_{\hat{H}_t^{-1}} + M_\mu\gamma_T(\delta')^2\sum_{t\in I_T}\bar{\alpha}_t(\widetilde{x}_t)\|x_t\|^2_{\hat{H}_t^{-1}}\\ &\leq \gamma_T(\delta')\sum_{t\in I_T}\left(\dot{\mu}(x_t^\top\bar{\theta}_t) + \left|\dot{\mu}(x_t^\top\bar{\theta}_t) - \dot{\mu}(x_t^\top\hat{\theta}_t)\right|\right)\|x_t\|_{\hat{H}_t^{-1}} + M_\mu\gamma_T(\delta')^2\sum_{t\in I_T}\bar{\alpha}_t(\widetilde{x}_t)\|x_t\|^2_{\hat{H}_t^{-1}}\\ &\leq \sum_{t\in I_T}\underbrace{\dot{\mu}(x_t^\top\bar{\theta}_t)\gamma_T(\delta')\|x_t\|_{\hat{H}_t^{-1}}}_{D_t^{\mathrm{FP}}} + \sum_{t\in I_T}\underbrace{\left|\dot{\mu}(x_t^\top\bar{\theta}_t) - \dot{\mu}(x_t^\top\hat{\theta}_t)\right|\gamma_T(\delta')\|x_t\|_{\hat{H}_t^{-1}}}_{E_t^{\mathrm{FP}}} + \sum_{t\in I_T}\underbrace{M_\mu\bar{\alpha}_t(\widetilde{x}_t)\gamma_T(\delta')^2\|x_t\|^2_{\hat{H}_t^{-1}}}_{F_t^{\mathrm{FP}}}.\end{aligned}$$

Bounding $\sum_t D_t^{\mathrm{FP}}$ We define $\bar{H}_t := \lambda\boldsymbol{I} + \sum_{\tau=1}^{t-1}\dot{\mu}(x_\tau^\top\bar{\theta}_\tau)x_\tau x_\tau^\top$. Then for all $\tau \leq t$, as the equation $\dot{\mu}(x_\tau^\top\bar{\theta}_\tau) \leq \dot{\mu}(x_\tau^\top\hat{\theta}_\tau)$ holds, we write:

$$\hat{H}_t = \lambda\boldsymbol{I} + \nabla^2\mathcal{L}_t(\hat{\theta}_t) = \lambda\boldsymbol{I} + \sum_{\tau=1}^{t-1}\dot{\mu}(x_\tau^\top\hat{\theta}_t)x_\tau x_\tau^\top \succeq \lambda\boldsymbol{I} + \sum_{\tau=1}^{t-1}\dot{\mu}(x_\tau^\top\bar{\theta}_\tau)x_\tau x_\tau^\top \succeq \bar{H}_t. \tag{F.5}$$

Then, we can bound $\sum_t D_t^{\text{FP}}$ as following:

$$\begin{aligned}
\sum_{t\in I_T} D_t^{\text{FP}} &\le \gamma_T(\delta') \sum_{t\in I_T} \dot{\mu}(x_t^\top \bar{\theta}_t)\|x_t\|_{\hat{H}_t^{-1}} \\
&\le \gamma_T(\delta') \sqrt{\sum_{t\in I_T} \dot{\mu}(x_t^\top \bar{\theta}_t)} \sqrt{\sum_{t\in I_T} \dot{\mu}(x_t^\top \bar{\theta}_t)\|x_t\|^2_{\hat{H}_t^{-1}}} && \text{(Cauchy-Schwartz inequality)} \\
&\le \gamma_T(\delta') \sqrt{\sum_{t\in I_T} \dot{\mu}(x_t^\top \bar{\theta}_t)} \sqrt{\sum_{t\in I_T} \dot{\mu}(x_t^\top \bar{\theta}_t)\|x_t\|^2_{\bar{H}_t^{-1}}} && \text{(By (F.5))} \\
&\le \gamma_T(\delta') \sqrt{\sum_{t\in I_T} \dot{\mu}(x_t^\top \bar{\theta}_t)} \sqrt{\sum_{t=1}^{T} \min\left\{1, \|\sqrt{\dot{\mu}(x_t^\top \bar{\theta}_t)}x_t\|^2_{\bar{H}_t^{-1}}\right\}} && \text{(Definition of } I_T)
\end{aligned}$$

Applying Lemma G.1, the second square-root term is bounded by

$$\sqrt{\sum_{t=1}^{T} \min\left\{1, \|\sqrt{\dot{\mu}(x_t^\top \bar{\theta}_t)}x_t\|^2_{\bar{H}_t^{-1}}\right\}} \le \sqrt{2d\log\left(1+\frac{\mathcal{L}_\mu T}{d\lambda}\right)}.$$

To handle the first term in the square root, we decompose it as:

$$\sqrt{\sum_{t\in I_T} \dot{\mu}(x_t^\top \bar{\theta}_t)} \le \sqrt{\sum_{t=1}^{T} \dot{\mu}(x_{t*}^\top \theta^*) + \sum_{t\in I_T} \left\{\dot{\mu}(x_t^\top \bar{\theta}_t) - \dot{\mu}(x_{t*}^\top \theta^*)\right\}} = \sqrt{\kappa_* T + \sum_{t\in I_T} \left\{\dot{\mu}(x_t^\top \bar{\theta}_t) - \dot{\mu}(x_{t*}^\top \theta^*)\right\}}.$$

Note that $\bar{\theta}_t := \operatorname{argmin}_{\theta\in\cup_{\tau\in[t,T]}\Theta_\tau(\delta,\lambda)} \dot{\mu}(x_t^\top \theta)$. Let τ' be an arbitrary τ whose $\Theta_\tau(\delta,\lambda)$ contains $\bar{\theta}_t$. Then, the latter term in the square root is then bounded as follows:

$$\begin{aligned}
&\sum_{t\in I_T} \left\{\dot{\mu}(x_t^\top \bar{\theta}_t) - \dot{\mu}(x_{t*}^\top \theta^*)\right\} = \sum_{t\in I_T} \left\{\dot{\mu}(x_t^\top \bar{\theta}_t) - \dot{\mu}(x_t^\top \theta^*)\right\} + \sum_{t\in I_T} \left\{\dot{\mu}(x_t^\top \theta^*) - \dot{\mu}(x_{t*}^\top \theta^*)\right\} \\
&\le M_\mu \left\{\sum_{t\in I_T} \left|\mu(x_t^\top \bar{\theta}_t) - \mu(x_t^\top \theta^*)\right| + \sum_{t\in I_T} \left|\mu(x_t^\top \theta^*) - \mu(x_{t*}^\top \theta^*)\right|\right\} && \text{(Lemma G.3)} \\
&\le M_\mu \left\{\sum_{t\in I_T} \left|\mu(x_t^\top \bar{\theta}_t) - \mu(x_t^\top \hat{\theta}_{\tau'})\right| + \sum_{t\in I_T} \left|\mu(x_t^\top \hat{\theta}_{\tau'}) - \mu(x_t^\top \theta^*)\right| + \sum_{t\in I_T} \left\{\mu(x_{t*}^\top \theta^*) - \mu(x_t^\top \theta^*)\right\}\right\} \\
&\le M_\mu \left\{2\sum_{t\in I_T} \left|\mu(x_t^\top \omega_t) - \mu(x_t^\top \hat{\theta}_{\tau(t)})\right| + \text{Reg}(T)\right\} \le M_\mu \left\{2\text{Reg}_{\text{EST}} + \text{Reg}(T)\right\},
\end{aligned}$$

where the second inequality holds from triangular inequality. For the last inequality, we leverage the definition of $\bar{\theta}_t$ and the pair$(\tau(t), \omega_t)$, as shown below:

$$\begin{aligned}
\left|\mu(x_t^\top \bar{\theta}_t) - \mu(x_t^\top \hat{\theta}_{\tau'})\right| &\le \max_{\theta\in\Theta_{\tau'}(\delta,\lambda)} \left|\mu(x_t^\top \theta) - \mu(x_t^\top \hat{\theta}_{\tau'})\right| \\
&\le \max_{\tau\in[t,T],\theta\in\Theta_\tau(\delta,\lambda)} \left|\mu(x_t^\top \theta) - \mu(x_t^\top \hat{\theta}_{\tau})\right| = \left|\mu(x_t^\top \omega_t) - \mu(x_t^\top \hat{\theta}_{\tau(t)})\right|,
\end{aligned}$$

and that under the event $\hat{E}_T$, as $\theta^* \in \cup_{\tau\in[t,T]}\Theta_\tau(\delta,\lambda)$, we can easily show that the second term $|\mu(x_t^\top \hat{\theta}_{\tau'}) - \mu(x_t^\top \theta^*)|$ can be upper bounded by $|\mu(x_t^\top \omega_t) - \mu(x_t^\top \hat{\theta}_{\tau(t)})|$. Note that after decomposing $\text{Reg}(T)$, we are currently in the process of bounding its individual components. However, during this process, $\text{Reg}(T)$ itself appears in the upper bound. This issue will be addressed in Step 5, where we will provide a strategy to effectively resolve this recursive dependency.

Bounding $\sum_t E_t^{\text{FP}}$ We define the Gram matrix $\bar{V}_t := \lambda \boldsymbol{I}/\kappa + \sum_{\tau=1}^{t-1} x_\tau x_\tau^\top$. The introduction of this matrix is to apply EPL as in the previous section. As the weight $\dot{\mu}(x_\tau^\top \hat{\theta}_t)$ on the weighted Gram matrix $\hat{H}_t$ is greater than or equal to $\kappa := \min_{x\in\mathcal{X}_{[T]},\theta\in\Theta} \dot{\mu}(x^\top \theta)$, we have:

$$\hat{H}_t = \lambda \boldsymbol{I} + \sum_{\tau=1}^{t-1} \dot{\mu}(x_\tau^\top \hat{\theta}_t) x_\tau x_\tau^\top \succeq \kappa \left(\frac{\lambda}{\kappa}\boldsymbol{I} + \sum_{\tau=1}^{t-1} x_\tau x_\tau^\top\right) \succeq \kappa \bar{V}_t. \tag{F.6}$$

Then, we can bound $\sum_t E_t^{\text{FP}}$ as following:

$$\begin{aligned}
\sum_{t\in I_T} E_t^{\text{FP}} &\le \gamma_T(\delta') \sum_{t\in I_T} \left|\dot{\mu}(x_t^\top \bar{\theta}_t) - \dot{\mu}(x_t^\top \hat{\theta}_t)\right| \|x_t\|_{\hat{H}_t^{-1}} \\
&\le M_\mu \gamma_T(\delta') \sum_{t\in I_T} \left|\mu(x_t^\top \bar{\theta}_t) - \mu(x_t^\top \hat{\theta}_t)\right| \|x_t\|_{\hat{H}_t^{-1}} && \text{(Lemma G.3)} \\
&\le M_\mu \gamma_T(\delta') \sqrt{\frac{1}{\kappa}} \sum_{t\in I_T} \left|\mu(x_t^\top \omega_t) - \mu(x_t^\top \hat{\theta}_{\tau(t)})\right| \|x_t\|_{\bar{V}_t^{-1}} \\
&\le \frac{2}{\kappa} M_\mu \mathcal{L}_\mu \gamma_T(\delta') \beta_T(\delta') \sum_{t\in I_T} \|x_t\|^2_{\bar{V}_t^{-1}}, && \text{(Lemma G.4)}
\end{aligned}$$

where the third inequality holds from the definition of the pair $(\tau(t), \omega_t)$, and by (F.6) Here, Lemma G.4 plays a crucial role by introducing the weighted norm $\|x_t\|_{\bar{V}_t^{-1}}$, which enables the application of Lemma G.1. Utilizing this lemma, we further proceed to bound as:

$$\begin{aligned}
\sum_{t\in I_T} E_t^{\text{FP}} &\le \frac{2}{\kappa} M_\mu \mathcal{L}_\mu \gamma_T(\delta') \beta_T(\delta') \sum_{t=1}^{T} \min\left\{1, \|x_t\|^2_{\bar{V}_t^{-1}}\right\} && \text{(Definition of } I_T) \\
&\le \frac{4}{\kappa} d M_\mu \mathcal{L}_\mu \gamma_T(\delta')^2 \log\left(1 + \frac{\kappa T}{d\lambda}\right), && \text{(Lemma G.1)}
\end{aligned}$$

where the final inequality use the fact that $\gamma_T(\delta') = c_{\delta'} \cdot \beta_T(\delta') \ge \beta_T(\delta')$, as $c_{\delta'} \ge 1$.

Bounding $\sum_t F_t^{\text{FP}}$ The process closely resembles that of bounding $\sum_t E_t^{\text{FP}}$. we have:

$$\begin{aligned}
\sum_{t\in I_T} F_t^{\text{FP}} &\le M_\mu \gamma_T(\delta')^2 \sum_{t\in I_T} \bar{\alpha}_t(\widetilde{x}_t) \|x_t\|^2_{\hat{H}_t^{-1}} \\
&\le \frac{\mathcal{L}_\mu}{2\kappa} M_\mu \gamma_T(\delta')^2 \sum_{t\in I_T} \|x_t\|^2_{\bar{V}_t^{-1}} && \text{(By (F.4) and (F.6))} \\
&\le \frac{\mathcal{L}_\mu}{2\kappa} M_\mu \gamma_T(\delta')^2 \sum_{t=1}^{T} \min\left\{1, \|x_t\|^2_{\bar{V}_t^{-1}}\right\} && \text{(Definition of } I_T) \\
&\le \frac{d}{\kappa} M_\mu \mathcal{L}_\mu \gamma_T(\delta')^2 \log\left(1 + \frac{\kappa T}{d\lambda}\right). && \text{(Lemma G.1)}
\end{aligned}$$

Combining all three terms, we derive an upper bound for $\text{Reg}_{\text{FP}} \le \sum_t \left\{D_t^{\text{FP}} + E_t^{\text{FP}} + F_t^{\text{FP}}\right\}$ as follows:

$$\begin{aligned}
\text{Reg}_{\text{FP}} \le\ & \gamma_T(\delta') \sqrt{2d \log\left(1 + \frac{\mathcal{L}_\mu T}{d\lambda}\right)} \sqrt{\kappa_* T + M_\mu \left\{2\text{Reg}_{\text{EST}} + \text{Reg}(T)\right\}} \\
&+ \frac{5d}{\kappa} M_\mu \mathcal{L}_\mu \gamma_T(\delta')^2 \log\left(1 + \frac{\kappa T}{d\lambda}\right). && \text{(F.7)}
\end{aligned}$$

Step 4 (Bounding Reg_{EST}) Next, we proceed to bound Reg_{EST}. The overall process closely mirrors that of Reg_{FP}. To begin, we decompose each instantaneous regret term. Let $\bar{\alpha}_t(\theta_1, \theta_2) = \int_0^1 (1-u)\dot{\mu}(x_t^\top \theta_1 + u(x_t^\top \theta_2 - x_t^\top \theta_1))du$, which is derived from Taylor's theorem when estimating the regret for the selected arm x_t using two parameter vectors, θ_1 and θ_2. Using Proposition G.2, we

expand the instantaneous regret term as:

$$
\begin{aligned}
&\left|\mu(x_t^\top \omega_t) - \mu(x_t^\top \hat{\theta}_{\tau(t)})\right| = \left|\dot{\mu}(x_t^\top \hat{\theta}_{\tau(t)})\langle x_t, \omega_t - \hat{\theta}_{\tau(t)}\rangle + \int_{x_t^\top \hat{\theta}_{\tau(t)}}^{x_t^\top \omega_t} (\mu(x_t^\top \omega_t) - z)\ddot{\mu}(z)dz\right| \\
&\leq \dot{\mu}(x_t^\top \hat{\theta}_{\tau(t)}) \left|\langle x_t, \omega_t - \hat{\theta}_{\tau(t)}\rangle\right| + \langle x_t, \omega_t - \hat{\theta}_{\tau(t)}\rangle^2 \int_0^1 (1-u) \left|\ddot{\mu}\left(x_t^\top \hat{\theta}_{\tau(t)} + u(x_t^\top \omega_t - x_t^\top \hat{\theta}_{\tau(t)})\right)\right| du \\
&\leq \dot{\mu}(x_t^\top \hat{\theta}_t) \left|\langle x_t, \omega_t - \hat{\theta}_{\tau(t)}\rangle\right| + M_\mu \langle x_t, \omega_t - \hat{\theta}_{\tau(t)}\rangle^2 \underbrace{\int_0^1 (1-u)\dot{\mu}\left(x_t^\top \hat{\theta}_{\tau(t)} + u(x_t^\top \omega_t - x_t^\top \hat{\theta}_{\tau(t)})\right) du}_{=\bar{\alpha}_t(\hat{\theta}_{\tau(t)}, \omega_t)} \\
&\leq \left(\dot{\mu}(x_t^\top \bar{\theta}_t) + \left|\dot{\mu}(x_t^\top \bar{\theta}_t) - \dot{\mu}(x_t^\top \hat{\theta}_{\tau(t)})\right|\right) \left|\langle x_t, \omega_t - \hat{\theta}_{\tau(t)}\rangle\right| + M_\mu \bar{\alpha}_t(\hat{\theta}_{\tau(t)}, \omega_t)\langle x_t, \omega_t - \hat{\theta}_{\tau(t)}\rangle^2 \\
&\leq \left(\dot{\mu}(x_t^\top \bar{\theta}_t) + \left|\dot{\mu}(x_t^\top \bar{\theta}_t) - \dot{\mu}(x_t^\top \hat{\theta}_{\tau(t)})\right|\right) \left|\langle x_t, \omega_t - \hat{\theta}_{\tau(t)}\rangle\right| + \underbrace{M_\mu \bar{\alpha}_t(\hat{\theta}_{\tau(t)}, \omega_t)\beta_{\tau(t)}(\delta')^2 \|x_t\|^2_{\hat{H}^{-1}_{\tau(t)}}}_{F_t^{\text{EST}}},
\end{aligned}
$$

where the second inequality follows from the definition of M_μ, and the last inequality results from the Cauchy-Schwartz inequality. The key distinction between bounding Reg_{FP} and Reg_{EST} lies is the use of $\beta(\delta')$ instead of $\gamma(\delta')$. The first term then can be upper bounded by the following terms, using triangular inequality:

$$
\begin{aligned}
&\left(\dot{\mu}(x_t^\top \bar{\theta}_t) + \left|\dot{\mu}(x_t^\top \bar{\theta}_t) - \dot{\mu}(x_t^\top \hat{\theta}_{\tau(t)})\right|\right) \|x_t\|_{\hat{H}^{-1}_{\tau(t)}} \|\omega_t - \hat{\theta}_{\tau(t)}\|_{\hat{H}_{\tau(t)}} \\
&\qquad \leq \underbrace{\dot{\mu}(x_t^\top \bar{\theta}_t)\beta_{\tau(t)}(\delta')\|x_t\|_{\hat{H}^{-1}_{\tau(t)}}}_{D_t^{\text{EST}}} + \underbrace{\left|\dot{\mu}(x_t^\top \bar{\theta}_t) - \dot{\mu}(x_t^\top \hat{\theta}_t)\right| \beta_{\tau(t)}(\delta')\|x_t\|_{\hat{H}^{-1}_{\tau(t)}}}_{E_t^{\text{EST}}}.
\end{aligned}
$$

Bounding $\sum_t D_t^{\text{EST}}$ By the equation(F.5), instead of t-dependent $\hat{H}_t$, we use τ-only dependent $\bar{H}_t$ to upper bound each term. We have:

$$
\begin{aligned}
\sum_{t \in I_T} D_t^{\text{EST}} &\leq \beta_T(\delta') \sum_{t \in I_T} \dot{\mu}(x_t^\top \bar{\theta}_t)\|x_t\|_{\hat{H}_t^{-1}} \qquad (t \leq \tau(t) \leq T) \\
&\leq \beta_T(\delta')\sqrt{2d\log\left(1 + \frac{\mathcal{L}_\mu T}{d\lambda}\right)}\sqrt{\sum_{t=1}^T \dot{\mu}(x_{t*}^\top \theta^*) + \sum_{t \in I_T}\left\{\dot{\mu}(x_t^\top \bar{\theta}_t) - \dot{\mu}(x_{t*}^\top \theta^*)\right\}} \\
&\leq \beta_T(\delta')\sqrt{2d\log\left(1 + \frac{\mathcal{L}_\mu T}{d\lambda}\right)}\sqrt{\kappa_* T + M_\mu\left\{2\text{Reg}_{\text{EST}} + \text{Reg}(T)\right\}}.
\end{aligned}
$$

Since the derivation proceeds identically to the case of $\sum_t D_t^{\text{FP}}$, with the only difference being the use of β_T in place of γ_T, we omit the detailed derivation here for brevity.

Bounding $\sum_t E_t^{\text{EST}}$ Following the bounding process of $\sum_t E_t^{\text{FP}}$, and using the equation (F.6),

$$
\begin{aligned}
\sum_{t \in I_T} E_t^{\text{EST}} &\leq \beta_T(\delta') \sum_{t \in I_T} \left|\dot{\mu}(x_t^\top \bar{\theta}_t) - \dot{\mu}(x_t^\top \hat{\theta}_t)\right| \|x_t\|_{\hat{H}_t^{-1}} \qquad (t \leq \tau(t) \leq T) \\
&\leq \frac{4d}{\kappa} M_\mu \mathcal{L}_\mu \beta_T(\delta')^2 \log\left(1 + \frac{\kappa T}{d\lambda}\right).
\end{aligned}
$$

Bounding $\sum_t F_t^{\text{EST}}$ For $\bar{\alpha}_t(\hat{\theta}_{\tau(t)}, \omega_t) := \int_0^1 (1-u)\dot{\mu}(x_t^\top \hat{\theta}_{\tau(t)} + u(x_t^\top \omega_t - x_t^\top \hat{\theta}_{\tau(t)}))du$, as shown in (F.4), it follows that $\bar{\alpha}_t(\hat{\theta}_{\tau(t)}, \omega_t) \leq \mathcal{L}_\mu/2$. Following the bounding process of $\sum_t F_t^{\text{FP}}$,

$$
\begin{aligned}
\sum_{t \in I_T} F_t^{\text{EST}} &\leq M_\mu \beta_T(\delta')^2 \sum_{t \in I_T} \bar{\alpha}_t(\widetilde{x}_t)\|x_t\|^2_{\hat{H}_t^{-1}} \qquad (t \leq \tau(t) \leq T) \\
&\leq \frac{d}{\kappa} M_\mu \mathcal{L}_\mu \beta_T(\delta')^2 \log\left(1 + \frac{\kappa T}{d\lambda}\right).
\end{aligned}
$$

In summary, combining three regret components, we bound $\text{Reg}_{\text{EST}} \leq \sum_t \left\{ D_t^{\text{EST}} + E_t^{\text{EST}} + F_t^{\text{EST}} \right\}$ as follows:

$$\begin{aligned}\text{Reg}_{\text{FP}} \leq \beta_T(\delta') &\sqrt{2d \log\left(1 + \frac{\mathcal{L}_\mu T}{d\lambda}\right)} \sqrt{\kappa_* T + M_\mu \left\{2\text{Reg}_{\text{EST}} + \text{Reg}(T)\right\}} \\ &+ \frac{5d}{\kappa} M_\mu \mathcal{L}_\mu \beta_T(\delta')^2 \log\left(1 + \frac{\kappa T}{d\lambda}\right). \end{aligned} \tag{F.8}$$

Step 5 (Solving equation) The equation (F.8) is indeed $c_{\delta'} > 1$ times the equation (F.7). Therefore, both Reg_{FP} and Reg_{EST} can be bounded using the regret bound of Reg_{FP}. Let $\text{Reg}_{\max} = \max\left\{\text{Reg}_{\text{FP}}, \text{Reg}_{\text{EST}}\right\}$, and we can bound $\text{Reg}_{\max}$ with the equation (F.7). Then we have,

$$\text{Reg}(T) \leq \left(\frac{2}{p} + 1\right) \text{Reg}_{\text{FP}} + \text{Reg}_{\text{EST}} + \varepsilon \leq \left(\frac{2}{p} + 2\right) \text{Reg}_{\max} + \varepsilon,$$

and accordingly, we get the following inequality:

$$\begin{aligned}\text{Reg}_{\max} \leq \gamma_T(\delta') &\sqrt{2d \log\left(1 + \frac{\mathcal{L}_\mu T}{d\lambda}\right)} \sqrt{\kappa_* T + M_\mu \left\{\left(\frac{2}{p} + 4\right) \text{Reg}_{\max} + \varepsilon\right\}} \\ &+ \frac{5d}{\kappa} M_\mu \mathcal{L}_\mu \gamma_T(\delta')^2 \log\left(1 + \frac{\kappa T}{d\lambda}\right)\end{aligned}$$

Now, The bound takes the form of $\text{Reg}_{\max} \leq A\sqrt{B + C\text{Reg}_{\max}} + D$, where the following quantities are defined:

$$\begin{aligned}A &:= \gamma_T(\delta') \sqrt{2d \log\left(1 + \frac{\mathcal{L}_\mu T}{d\lambda}\right)} = \widetilde{\mathcal{O}}\left(d\right) \\ B &:= \kappa_* T + M_\mu \varepsilon = \widetilde{\mathcal{O}}\left(\kappa_* T + d\sqrt{\frac{T}{\lambda^2}}\right) \\ C &:= M_\mu \left(\frac{2}{p} + 4\right) = \mathcal{O}(1) \\ D &:= \frac{5d}{\kappa} M_\mu \mathcal{L}_\mu \gamma_T(\delta')^2 \log\left(1 + \frac{\kappa T}{d\lambda}\right) = \widetilde{\mathcal{O}}\left(\frac{d^2}{\kappa}\right),\end{aligned}$$

focusing on the terms involving d,T, and κ. With the choice of $\lambda = \mathcal{O}(d)$ and applying Lemma G.5, the upper bound for $\text{Reg}_{\max}$ simplifies to:

$$\text{Reg}_{\max} = \widetilde{\mathcal{O}}\left(d\sqrt{\kappa_* T} + d^2 + \frac{d^2}{\kappa}\right) = \widetilde{\mathcal{O}}\left(d\sqrt{\kappa_* T} + \frac{d^2}{\kappa}\right).$$

Now, combining all terms, the cumulative regret $R(T)$ can be bounded as:

$$\begin{aligned}R(T) &\leq \text{Reg}(T) + \frac{4dR^*}{\log 2}\left\{\log\left(1 + \frac{\mathcal{L}_\mu}{\lambda \log 2}\right) + \log\left(1 + \frac{\kappa}{\lambda \log 2}\right)\right\} \\ &\lesssim \left(\frac{2}{p} + 2\right) \text{Reg}_{\max} + \varepsilon \\ &= \widetilde{\mathcal{O}}\left(d\sqrt{\kappa_* T} + \frac{d^2}{\kappa} + d\sqrt{\frac{T}{\lambda^2}}\right) = \widetilde{\mathcal{O}}\left(d\sqrt{\kappa_* T} + \frac{d^2}{\kappa}\right). \qquad \text{(Choose } \lambda = \mathcal{O}(d)\text{)}\end{aligned}$$

Our derived regret bound, $\widetilde{\mathcal{O}}\left(d\sqrt{\kappa_* T} + d^2/\kappa\right)$, aligns with the state-of-the-art regret guarantee [3, 13, 33, 34].

G Proof supplement and Lemmas

In this section, we provide key propositions, lemmas, and inequality bounds that are essential for the main proof. These results serve as the mathematical foundation for the regret analysis and other theoretical guarantees established in this work.

G.1 Supporting Lemmas for main proof

Proposition G.1 (Lemma D.1. of Lee et al. [34])**.** *Let μ be increasing and self-concordant with M_μ. Let $\mathcal{Z} \subseteq \mathcal{B}(S) := \{z \in \mathbb{R} \mid |z| \leq S\}$ in $\mathbb{R}$. Then for any $z_1, z_2 \in \mathcal{Z}$, the following holds:*

$$\int_0^1 (1-u)\dot{\mu}(z_1 + u(z_2 - z_1))du \geq \frac{\dot{\mu}(z_1)}{2 + 2SM_\mu}$$

This proposition establishes a lower bound for the integral involving a self-concordant, particularly useful for controlling weighted norms and derivatives in regret analysis. The following proposition provides the well-known Taylor's expansion expression with integral remainder. We highlight specific cases that are frequently used throughout the main proof.

Proposition G.2 (Taylor's Theorem with Integral Remainder Form)**.** *Let $n \geq 0$ be an integer and let the function $f : \mathbb{R} \to \mathbb{R}$ be $(n+1)$ times differentiable at the point $x_0 \in \mathbb{R}$. Let $f^{(n)}$ to denote its n-th derivatives, then $f(x)$ can be expressed as:*

$$f(x) = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!}(x - x_0)^i + \frac{1}{n!}\int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n dt$$

Especially for $n = 0$, by letting $t = x_0 + u(x - x_0)$, $f(x)$ can be expressed as:

$$f(x) = f(x_0) + \int_{x_0}^{x} f'(t)dt = f(x_0) + (x - x_0)\int_0^1 f'\left(x_0 + u(x - x_0)\right) du \tag{G.1}$$

Especially for $n = 1$, by letting $t = x_0 + u(x - x_0)$, $f(x)$ can be expressed as:

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x - x_0) + \int_{x_0}^{x} f''(t)(x-t)dt \\ &= f(x_0) + f'(x_0)(x - x_0) + \int_0^1 f''\left(x_0 + u(x - x_0)\right)\left((1-u)(x - x_0)\right) \cdot (x - x_0)du \\ &= f(x_0) + f'(x_0)(x - x_0) + (x - x_0)^2 \int_0^1 f''\left(x_0 + u(x - x_0)\right)(1-u)du \end{aligned} \tag{G.2}$$

Similarly, with multivariate function $f : \mathbb{R}^d \to \mathbb{R}$ and $n = 1$, we have that:

$$\begin{aligned} f(x) = f(x_0) &+ \sum_{i=1}^{d} \frac{\partial}{\partial x_i} f(x_0)(x - x_0)_i + \sum_{i=1}^{d}(x - x_0)_i^2 \int_0^1 (1-u)\frac{\partial^2}{\partial x_i^2} f(x_0 + u(x - x_0))du \\ &+ \sum_{i \neq j} 2(x - x_0)_i (x - x_0)_j \int_0^1 (1-u)\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} f(x_0 + u(x - x_0))du \\ = f(x_0) &+ \nabla f(x_0)^\top (x - x_0) + (x - x_0)^\top \left(\int_0^1 (1-u)\nabla^2 f(x_0 + u(x - x_0))du\right)(x - x_0). \end{aligned}$$

The following lemmas provide critical tools for analyzing regret bounds in the context of bandit algorithms. Specifically, they focus on bounding terms that involve weighted norms of feature vectors under Gram matrices. These results are pivotal in deriving efficient bounds on cumulative regret and exploration.

Lemma G.1 (Elliptical Potential Lemma)**.** *let $\{x_t\}_{t=1}^T$ be a sequence in $\mathbb{R}^d$ satisfying $\|x_t\| \leq R$ for all $t \leq T$. For a gram matrix $V_t := \lambda \boldsymbol{I} + \sum_{\tau=1}^{t-1} x_\tau x_\tau^\top$, we have that*

$$\sum_{t=1}^{T} \min\left\{1, \|x_t\|_{V_t^{-1}}^2\right\} \leq 2d \log\left(1 + \frac{R^2 T}{d\lambda}\right).$$

Lemma G.2 (Elliptical Potential Count Lemma)**.** *For $R, c > 0$, let $\{x_t\}_{t=1}^T$ be a sequence in $\mathbb{R}^d$ satisfying $\|x_t\| \le R$ for all $t \le T$. For a gram matrix $V_t := \lambda \boldsymbol{I} + \sum_{\tau=1}^{t-1} x_\tau x_\tau^\top$, the length of the sequence $\mathcal{N}_T := \left\{ t \in [T] \mid \|x_t\|_{V_t^{-1}} > c \right\}$ is bounded as:*

$$|\mathcal{N}_T| \le \frac{2d}{\log(1+c^2)} \log\left(1 + \frac{R^2}{\lambda \log(1+c^2)}\right)$$

The following connects the smoothness of $\dot{\mu}$ to the difference in μ.

Lemma G.3. *For $x, y \in \mathbb{R}$, $|\dot{\mu}(x) - \dot{\mu}(y)| \le M_\mu |\mu(x) - \mu(y)|$.*

Proof.

$$\begin{aligned}
|\dot{\mu}(x) - \dot{\mu}(y)| &= \left|(x-y)\int_0^1 \ddot{\mu}(y + u(x-y))du\right| && \text{(By (G.1) with } f = \dot{\mu}) \\
&\le |x-y| \int_0^1 |\ddot{\mu}(y + u(x-y))|\, du \\
&\le M_\mu |x-y| \int_0^1 \dot{\mu}(y + u(x-y))du \\
&= M_\mu \left|(x-y)\int_0^1 \dot{\mu}(y + u(x-y))du\right| \\
&= M_\mu |\mu(x) - \mu(y)| && \text{(By (G.1) with } f = \mu)
\end{aligned}$$

where the second equality holds since the integral term is positive as $\dot{\mu} \ge 0$. □

The following lemma bounds the difference in μ values, using linearization as in the previous works.

Lemma G.4. *For any $t \ge 1$ and $\theta_1, \theta_2 \in \Theta_{\tau(t)}$, we have the following:*

$$\left|\mu(x_t^\top \theta_1) - \mu(x_t^\top \theta_2)\right| \le 2\gamma_T(\delta)\mathcal{L}_\mu \sqrt{\frac{1}{\kappa}} \|x_t\|_{V_t^{-1}}$$

Proof.

$$\begin{aligned}
|\mu(x_t^\top \theta_1) - \mu(x_t^\top \theta_2)| &= \left|\langle x_t, \theta_1 - \theta_2\rangle \int_0^1 \dot{\mu}(x_t^\top \theta_2 + u(x_t^\top \theta_1 - x_t^\top \theta_2))du\right| && \text{(By (G.1))} \\
&\le \left\{\|x_t\|_{\hat{H}_{\tau(t)}^{-1}} \cdot \|\theta_1 - \theta_2\|_{\hat{H}_{\tau(t)}}\right\} \int_0^1 \mathcal{L}_\mu du && \text{(Cauchy-Schwartz, } \mathcal{L}_\mu\text{-Lipschitzness)} \\
&\le \mathcal{L}_\mu \|x_t\|_{\hat{H}_{\tau(t)}^{-1}} \left\{\|\theta_1 - \hat{\theta}_{\tau(t)}\|_{\hat{H}_{\tau(t)}} + \|\theta_2 - \hat{\theta}_{\tau(t)}\|_{\hat{H}_{\tau(t)}}\right\} && \text{(Triangle inequality)} \\
&\le 2\beta_{\tau(t)}(\delta')\mathcal{L}_\mu \|x_t\|_{\hat{H}_{\tau(t)}^{-1}} \\
&\le 2\beta_{\tau(t)}(\delta')\mathcal{L}_\mu \sqrt{\frac{1}{\kappa}} \|x_t\|_{V_{\tau(t)}^{-1}} && \text{(By (F.6))} \\
&\le 2\beta_T(\delta')\mathcal{L}_\mu \sqrt{\frac{1}{\kappa}} \|x_t\|_{V_t^{-1}} && (t \le \tau(t) \le T)
\end{aligned}$$

□

Lemma G.5. *Let $A, B, C, D, X \in \mathbb{R}^+$. The following implication holds:*

$$X \le A\sqrt{B + CX} + D \quad \Longrightarrow \quad X \le 2(A\sqrt{B} + A^2 C + D)$$

Proof. let $f : x \to x^2 - px - q$ for $p, q > 0$. Then the roots for $f(x) = 0$ are:

$$x_1, x_2 = \frac{p \pm \sqrt{p^2 + 4q}}{2}.$$

Now, as f is a convex function, $x^2 \le px + q$ implies:

$$x \le \max\{x_1, x_2\} \le \frac{p + \sqrt{p^2 + 4q}}{2} \le \frac{p + (p + 2\sqrt{q})}{2} = p + \sqrt{q}. \qquad \text{(triangle inequality)}$$

And accordingly, we have:

$$\begin{aligned} x \le p\sqrt{x} + q \quad &\Longrightarrow \quad \sqrt{x} \le p + \sqrt{q} \\ &\Longrightarrow \quad x \le (p + \sqrt{q})^2 \le 2p^2 + 2q \end{aligned} \tag{G.3}$$

where the inequality holds from $(x + y)^2 \le 2(x^2 + y^2)$. Then according to the equation (G.3),

$$\begin{aligned} X \le A\sqrt{B + CX} + D \quad &\Longrightarrow \quad X \le A\sqrt{C}\sqrt{X} + A\sqrt{B} + D \qquad \text{(triangle inequality)} \\ &\Longrightarrow \quad X \le 2(A\sqrt{C})^2 + 2(A\sqrt{B} + D) = 2(A\sqrt{B} + A^2C + D) \end{aligned}$$

□

G.2 Auxiliary bounding inequalities

We introduce key probabilistic inequalities and bounds frequently used in the analysis of randomized algorithms. These results provide tools to bound the probabilities of deviations and concentration of random variables, which are essential for deriving high-probability guarantees in the main analysis.

Proposition G.3 (Azuma's inequality)**.** *If a super-martingale $(X_t)_{t\ge 0}$ corresponding to a filtration $\mathcal{H}_{t-1}$ satisfies $|X_t - X_{t-1}| < c_t$ for some constant c_t for all $t = 1, \dots, T$ then for any $\alpha > 0$,*

$$\mathbb{P}(X_T - X_0 \ge \alpha) \le 2\exp\left(-\frac{\alpha^2}{2\sum_{t=1}^{T} c_t^2}\right).$$

Proposition G.4 (Chernoff bound)**.** *For a random variable X and its moment-generating function $M(t) = \mathbb{E}\left[e^{tX}\right]$,*

$$\mathbb{P}(X \ge \alpha) \le \inf_{t>0} M(t)e^{-t\alpha}.$$

Accordingly, for a random variable following a standard normal distribution (i.e. $X \sim \mathcal{N}(0, 1)$),

$$\mathbb{P}(X \ge \alpha) \le \inf_{t>0} \exp(\frac{t^2}{2} - t\alpha) = e^{-\frac{\alpha^2}{2}}.$$

Lemma G.6. *Let z be a random variable sampled from the standard Normal distribution. Then for all $\delta \in (0, 1)$,*

$$\mathbb{P}(|z| \le \sqrt{2\log(2/\delta)}) \ge 1 - \delta.$$

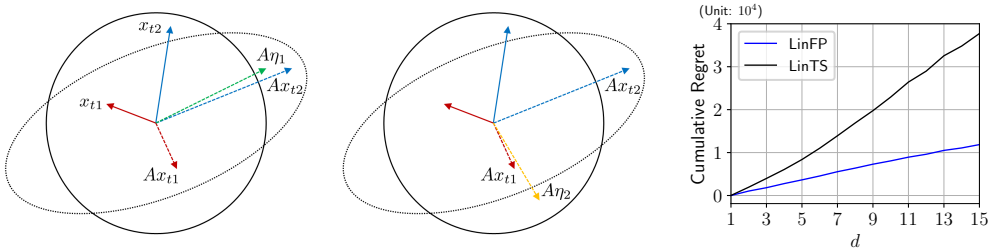
Proof. By proposition Proposition G.4, $\mathbb{P}(|z| > \alpha) = 2\mathbb{P}(z > \alpha) \le 2e^{-\frac{\alpha^2}{2}}$. Set the right-hand side as δ, and get $\alpha = \sqrt{2\log(2/\delta)}$. □

H Carving off the $\sqrt{d}$ factor compared to TS

The proposed algorithm, `GLM-FP`, adopts a novel exploration strategy by perturbing the input feature vectors, in contrast to conventional randomized algorithms such as Thompson Sampling (TS), which introduce randomness into the model parameter θ. This design yields a regret bound with linear dependence on d, whereas TS incurs a higher $\mathcal{O}(d^{3/2})$ dependence. We examine the origin of this discrepancy by comparing the linear variants of both algorithms (see Appendix C.2), highlighting how each introduces randomness to facilitate exploration.

The randomized evaluation score $\widetilde{f}_t(x_i)$ (either $x_{ti}^\top \hat{\theta}_t$ for TS or $\widetilde{x}_{ti}^\top \hat{\theta}_t$ for FP) for each arm used for action selection in both algorithms is straightforward to compute as follows:

$$\text{(TS)} \quad \widetilde{f}_t(x_i) = x_{ti}^\top \widetilde{\theta}_t = x_{ti}^\top \hat{\theta}_t + c_t\, x_{ti}^\top V_t^{-1/2} \zeta_t, \qquad \text{(FP)} \quad \widetilde{f}_t(x_i) = \widetilde{x}_{ti}^\top \hat{\theta}_t = x_{ti}^\top \hat{\theta}_t + c_t\, z_t \|x_{ti}\|_{V_t^{-1}},$$



(a) Illustration of the transformation of the feature vectors in TS. (b) Cumulative regret vs. d.

Figure H.1: **(a)** Transformation of the well-explored arm x_{t1} and under-explored arm x_{t2} using $A = D^{1/2}P^\top$, where $V_t^{-1/2} = PDP^\top$. Left: ζ_1 induces a proper bonus. Right: ζ_2 reverses the effect. **(b)** Average terminal regret $R(T)$ over 100 runs with $T = 200{,}000$, $K = 50$, and varying d.

where $\zeta_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $z_t \sim \mathcal{N}(0, 1)$. Thus, for each arm individually, both methods induce the same Gaussian distribution,[4] $\widetilde{f}_t(x_i) \sim \mathcal{N}(x_{ti}^\top \hat{\theta}_t,\, c_t^2 \|x_{ti}\|^2_{V_t^{-1}})$. However, the *object of perturbation*—parameter in TS versus feature in FP—fundamentally alters how exploration bonuses are assigned and how arm comparisons are coupled at each timestep.

In TS, the bonus $\langle x_{ti}, \zeta_t \rangle_{V_t^{-1/2}}$ projects each arm onto a shared random direction ζ_t in the $V_t^{-1/2}$-transformed space. As conceptually illustrated in Figure H.1a, this shared dependence can produce counterintuitive effects: well-explored arms may occasionally receive large bonuses simply due to alignment with ζ_t, while under-explored ones may be neglected. Because the same ζ_t governs all arms, the analysis must ensure uniform reliability of exploration across directions, requiring high-probability control of the d-dimensional Gaussian vector ζ_t. Applying a union bound over d coordinates introduces an additional $\sqrt{d}$ factor into the regret bound.

FP, in contrast, decouples exploration from directional uncertainty. Its bonus $z_t \cdot \|x_{ti}\|_{V_t^{-1}}$ scales directly with per-arm uncertainty, ensuring under-explored arms systematically receive larger bonuses. Since randomness enters only through the scalar z_t, the analysis reduces to bounding a one-dimensional Gaussian projection $u^\top \zeta_t$ for some fixed unit vector u.

From an equation-level viewpoint, the regret of TS also admits the decomposition of ours [2]. For both algorithms, the estimation component $\mathrm{Reg}_{\mathrm{EST}}$ is bounded in the same manner, and the anti-concentration event occurs with the same probability p. Consequently, the optimism-driven term in the regret scales with the concentration width divided by p. The essential difference therefore lies in the concentration width—or, equivalently, in how each algorithm controls the perturbation magnitude that also determines the second part of the decomposition. We express the perturbation term as

$$\text{(TS)} \quad \left|x_t^\top(\widetilde{\theta}_t - \hat{\theta}_t)\right| = c_t \left|x_{ti}^\top V_t^{-1/2} \zeta_t\right| \le c_t \|x_{ti}\|_{V_t^{-1}} \cdot \|V_t^{-1/2}\zeta_t\|_{V_t} = c_t \|x_{ti}\|_{V_t^{-1}} \cdot \left|\zeta_t\right|,$$

$$\text{(FP)} \quad \left|(\widetilde{x}_t - x_t)^\top \hat{\theta}_t\right| = c_t \left|\|x_t\|_{V_t^{-1}} \frac{\hat{\theta}_t^\top \zeta_t}{|\hat{\theta}_t|_2}\right| = c_t \|x_{ti}\|_{V_t^{-1}} \cdot \left|u^\top \zeta_t\right|.$$

The concentration width in TS depends on $\|\zeta_t\|_2$, the norm of a d-dimensional Gaussian vector, whereas that of FP scales with the one-dimensional projection $|u^\top \zeta_t|$. To obtain a uniform high-probability guarantee across all coordinates, a union bound over the d-dimensional perturbation space introduces an additional $\mathcal{O}(\sqrt{d})$ factor for TS. Thus, despite having identical marginal distributions for individual arms, the two algorithms differ fundamentally in how their perturbations couple across arms: TS requires concentration over all directions in $\mathbb{R}^d$, whereas FP relies on a single scalar randomization. This structural decoupling eliminates the extraneous $\sqrt{d}$ factor, yielding linear $\mathcal{O}(d)$ dependence in the regret bound and clarifying the geometric origin of FP's improvement over TS.

[4] In the linear bandit setting, this distribution also matches RandUCB [41], though its derivation is conceptually distinct and diverges beyond the GLB case.

I Experiments

We evaluate our proposed algorithms in two contextual bandit settings: (i) generalized linear bandits (GLBs), including linear and logistic models, and (ii) nonlinear contextual bandits based on neural networks. All results are averaged over 100 runs (GLBs) or 5 runs (neural bandits) to ensure robustness. Experiments were conducted on modest compute resources: GLBs on a CPU server (Intel Xeon Silver 4210R, 40 threads), and neural bandits on a single NVIDIA RTX 3090 GPU. Each GLB run completes within minutes, while neural runs finish in under an hour.

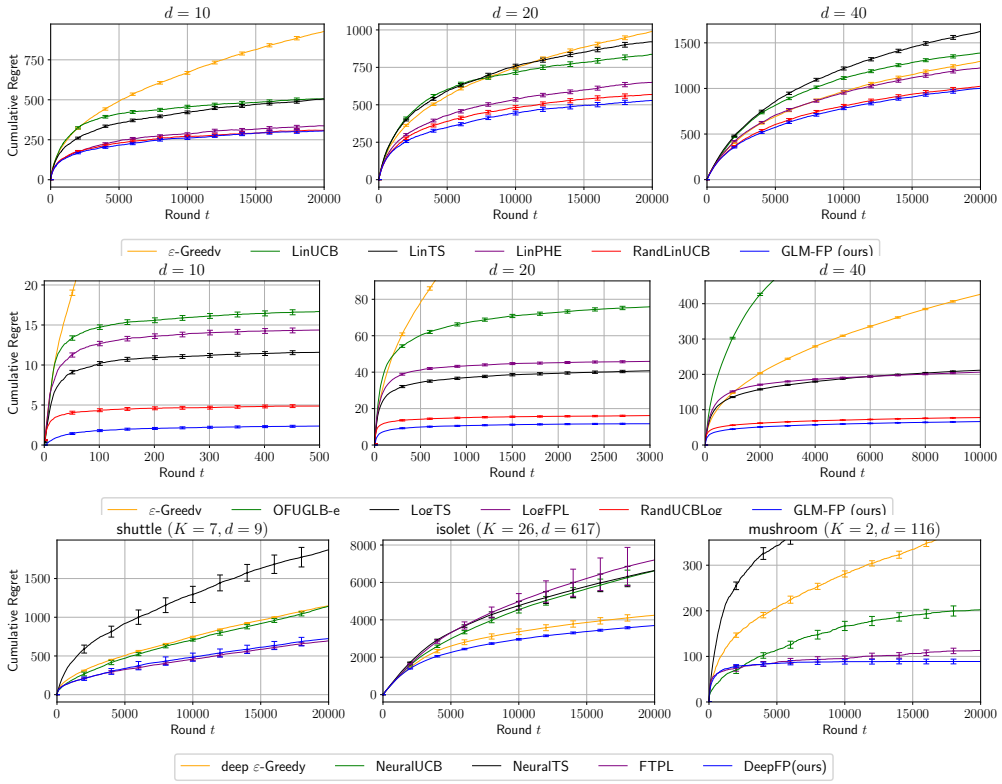


Figure I.1: Comparison of cumulative regret across contextual bandit algorithms: linear (top), logistic (middle), and neural (bottom).

I.1 Generalized linear bandits

We evaluate `GLM-FP` in both linear and logistic bandit settings, comparing against ε-greedy, UCB, TS, PHE, and RandUCB. Parameter estimation is performed via regularized weighted least squares (linear) or IRLS (logistic). We consider $K = 100$ arms per round over horizons $T = 20{,}000$ (linear) and $T = 10{,}000$ (logistic). Contexts and θ^* are sampled from a standard Gaussian and normalized. For logistic bandits, we use the sigmoid link and constrain $\|\theta^*\| \leq 4$ to keep logits in $[-4, 4]$; linear rewards include Gaussian noise $\mathcal{N}(0, 1)$. Feature dimension d is varied over $\{10, 20, 40\}$, with confidence level $\delta = 1/T$ and regularization $\lambda = 10^{-4}$. Parameter estimation is performed via regularized weighted least squares (WLS) in linear bandit setting or IRLS in logistic bandit setting.

Baselines and tuning. For linear bandits, baselines include ε-greedy [30], LinUCB [1], LinTS [5], LinPHE [28], and RandLinUCB [41]. For logistic bandits, we include ε-greedy, OFUL-GLB-e [34], LogTS, LogPHE, and RandUCBLog [29]. Hyperparameters follow the original papers: TS-based methods and ours use inflation $c_t = 1$; ε-greedy uses $\varepsilon_t = \varepsilon\sqrt{T/t}$ with $\varepsilon = 0.05$.

Results. As shown in Figure I.1 (top and middle), `GLM-FP` consistently achieves the lowest regret across all tested dimensions. While RandLinUCB performs competitively in the linear case, `GLM-FP` exhibits superior robustness, particularly in fixed-arm settings or environments with non-stationary arm sets. In the logistic setting, `GLM-FP` outperforms all baselines across all tested configurations, demonstrating its effectiveness even when the utility scale (i.e., $x^\top\theta^*$, the inner product term inside

$\mu(\cdot)$) or the reward noise variance is varied. These results highlight the reliability and adaptability of our approach across diverse GLB scenarios.

I.2 Neural bandits

We extend the feature perturbation framework to neural contextual bandits through a simple and scalable algorithm, `DeepFP`. At each round, `DeepFP` perturbs inputs as $\widetilde{x}_{ti} = x_{ti} + \zeta_{ti}$, where $\zeta_{ti} \sim \mathcal{N}(0, I/t)$. The variance decay $1/t$ reflects the diminishing feature uncertainty over time, analogous to the confidence scaling $\|x\|_{\hat{H}_t^{-1}}$ in GLBs. This enables exploration without requiring posterior sampling or gradient-based confidence intervals. Unlike NeuralTS, which perturbs predicted rewards (essentially a randomized UCB), or FTPL, which perturbs past rewards, `DeepFP` directly perturbs the *input features*. This structural distinction simplifies implementation, preserves the inductive bias of the function class, and remains stable in high-capacity models where $p \gg d$, since it acts in feature space rather than parameter space. Consequently, `DeepFP` provides a principled and computationally efficient approach to exploration in neural bandits.

Experimental setup. Following Zhang et al. [43], we evaluate neural contextual bandits using classification tasks on three UCI benchmark datasets [37]: `shuttle` (7 classes, 9 features), `isolet` (26 classes, 617 features), and `mushroom` (binary, 112 features). These tasks are converted into multi-armed bandit problems via a disjoint feature mapping [35]: for a k-class classification problem with input dimension d, we construct a kd-dimensional feature vector by embedding the input x into the slot corresponding to each class. Each x_i is treated as the feature of arm i, and a reward of 1 is given if the chosen arm matches the true label, and 0 otherwise. Regret is measured as cumulative classification errors over $T = 10{,}000$, with each experiment repeated five times on shuffled data.

Neural architectures. To examine the effect of capacity, we consider both shallow and deep neural networks. The shallow network consists of a single hidden layer of size $d \times 100$ with ReLU activation, followed by a fully connected layer of size 100×1 and a softmax output, following the design in Zhang et al. [43]. The deep network instead uses two hidden layers: the first of size $d \times 50$ with ReLU, and the second of size 50×50 with ReLU, followed by a 50×1 output with softmax. Both models are trained online using the Adam optimizer (learning rate 0.001, batch size 32), updating on the most recent 32 observed samples at each round. Figure I.1(bottom) reports results for the deep model.

Baselines and tuning. We compare against ε-greedy, NeuralUCB [44], NeuralTS [43], and FTPL [29]. For NeuralUCB and NeuralTS, the confidence matrix inverse is approximated diagonally for efficiency. In `DeepFP`, instead of perturbing the entire kd-dimensional parameter space, we perturb only the d-dimensional subspace corresponding to the chosen arm. This masking avoids interference across arms and ensures more accurate value estimation.

Results. As shown in Figure I.1(bottom), `DeepFP` consistently outperforms all baselines across the three datasets. It achieves strong performance without relying on posterior approximations or confidence intervals, and remains effective even in complex, high-dimensional settings. These results demonstrate that simple feature perturbation can provide a robust, scalable, and principled approach to exploration in neural contextual bandits.

I.3 Additional experiments

We further explore robustness by varying the parameter norm S and context set size $|\mathcal{C}|$ (see figure captions). For neural bandits, we also evaluate on MNIST and Fashion-MNIST datasets. Each 28×28 image forms a K-armed context ($K = 10$), where the i-th arm corresponds to placing the image in the i-th channel. The architecture of the model is as follows: two convolutional layers are applied with ReLU activation and 2×2 max pooling after each. The first layer uses 32 filters and the second 64 filters, both with kernel size 3×3 and stride 1. The resulting $64 \times 7 \times 7$ feature map is flattened and passed through two fully connected layers with hidden size 128, followed by ReLU and a final output scalar. This design enables efficient learning of visual features while maintaining compatibility with the bandit framework through arm-wise shared representation.

Results in Figures I.2 to I.8 confirm the adaptability of feature perturbation across linear, logistic, and neural settings, including vision-based bandit tasks.

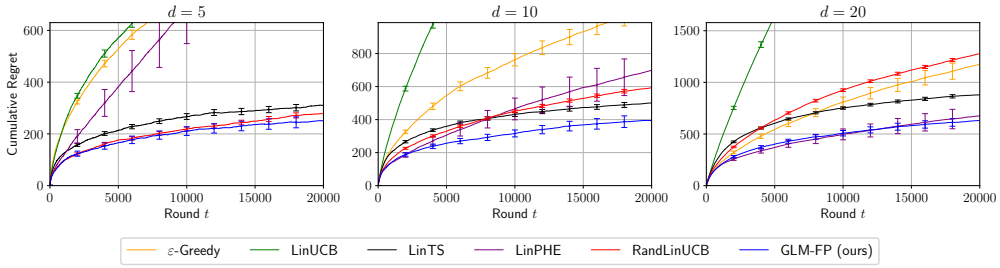


Figure I.2: Linear Bandit. $|\mathcal{C}| = 1$, $d = \{5, 10, 20\}$, $K = 100$, $S = 1$.

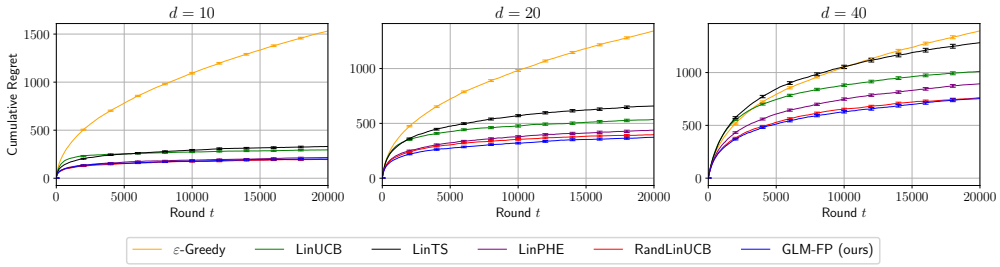


Figure I.3: Linear Bandit. $|\mathcal{C}| = T$, $d = \{10, 20, 40\}$, $K = 100$, $S = 2$.

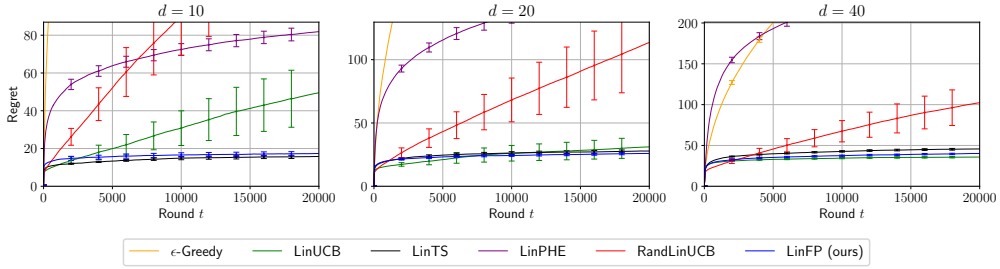


Figure I.4: Linear Bandit with noise $\mathcal{N}(0, 0.1^2)$. $|\mathcal{C}| = T$, $d = \{10, 20, 40\}$, $K = 100$, $S = 2$.

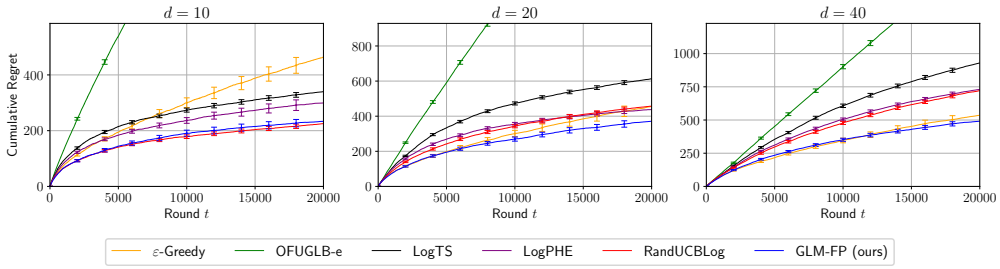


Figure I.5: Logistic Bandit. $|\mathcal{C}| = 1$, $d = \{5, 10, 20\}$, $K = 100$, $S = 1$.

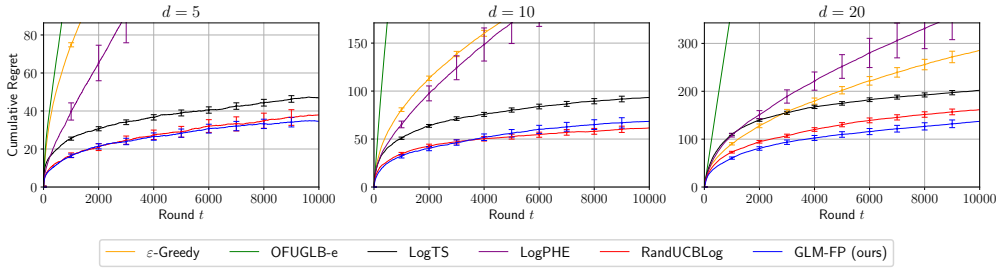


Figure I.6: Logistic Bandit. $|\mathcal{C}| = 1$, $d = \{5, 10, 20\}$, $K = 100$, $S = 4$.

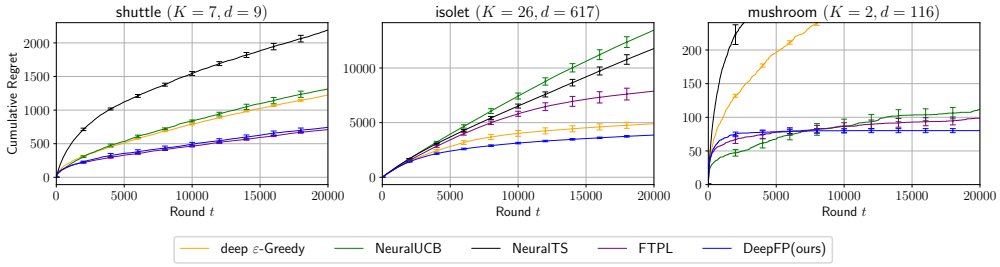


Figure I.7: Neural Bandit. Multi-layer perceptron model with one hidden layer and output layer.

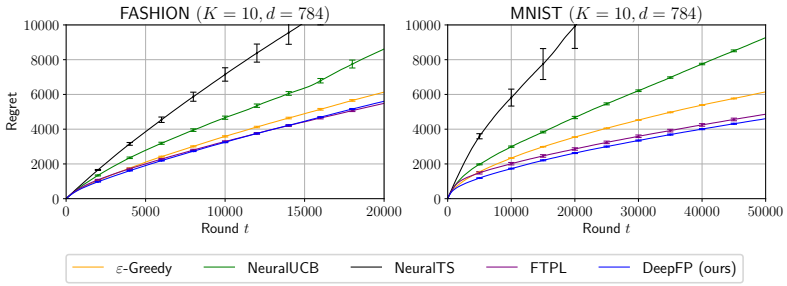


Figure I.8: Neural Bandit. Experiments on MNIST dataset with a CNN model.