Adaptive Foundation Models for Online Decisions: HyperAgent with Fast Incremental Uncertainty Estimation

Yingru Li[12]*, **Jiawei Xu**[1] and **Zhi-Quan Luo**[12]
[1]The Chinese University of Hong Kong, Shenzhen, China
[2]Shenzhen Research Institute of Big Data

Abstract

Foundation models often struggle with uncertainty when faced with new situations in online decision-making, necessitating scalable and efficient exploration to resolve this uncertainty. We introduce GPT-HyperAgent, an augmentation of GPT with HyperAgent for uncertainty-aware, scalable exploration in contextual bandits, a fundamental online decision problem involving natural language input. We prove that HyperAgent achieves fast incremental uncertainty estimation with $\tilde{O}(\log T)$ per-step computational complexity over T periods under the linear realizable assumption. Our analysis demonstrates that HyperAgent's regret order matches that of exact Thompson sampling in linear contextual bandits, closing a significant theoretical gap in scalable exploration. Empirical results in real-world contextual bandit tasks, such as automated content moderation with human feedback, validate the practical effectiveness of GPT-HyperAgent for safety-critical decisions. Our code is open-sourced at https://github.com/szrlee/GPT-HyperAgent/.

1 Introduction

Real-world decision-making often faces uncertainty due to a lack of comprehensive information about the environment. Intelligent agents must not only understand this uncertainty but also actively gather information to resolve it. This task is particularly challenging for real-time online decisions involving foundation models—large-scale AI models pretrained on vast datasets that process unstructured inputs like text and images.

Content moderation on digital platforms, a real-world safety-critical task, exemplifies these challenges [Gorwa et al., 2020]. Traditionally, human reviewers detected violations of human value and community standards [Roberts, 2019], but the high volume of posts on platforms like Facebook [Meta, 2024], Twitter [Corp., 2024], and Reddit [Reddit, 2024] required automating content moderation. AI systems using foundation models [Weng et al., 2023] provide real-time capabilities and reduce human workload. However, pretrained on historical data, these models may struggle with uncertainty in online production traffic where new and rare situations exhibit, leading to errors [Markov et al., 2023]. Reliable content moderation requires real-time human feedback to correct AI errors, reduce uncertainty, and refine detection policies. This human-AI collaboration pipeline aims to minimize human intervention (by exploiting the current AI capability) while ensuring long-term reliability (by exploring uncertain content for human review to improve future ability), as illustrated in Fig. 1. To achieve these goals, AI systems using foundation models need to quickly adjust uncertainty estimates and refine policies as new data continually arrives, necessitating fast incremental uncertainty estimation and scalable solutions to balance exploration and exploitation.

*Corresponding to Yingru Li: szrlee@gmail.com or yingruli@link.cuhk.edu.cn.

Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024).

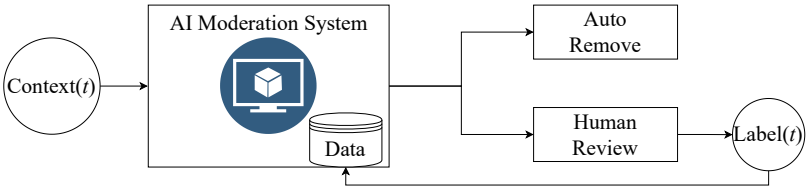


Figure 1: The Human-AI agile collaboration pipeline for risk oversight in an online production environment: at time $t \in \mathbb{N}$, AI moderation system receives post context(t) and decides whether it is auto-removed or published for human review. If reviewed, AI moderation system quickly integrates human feedback via label(t) to overturn AI decision and continually improve AI capability. This pipeline aims to save human workload while ensuring long-term reliability and safety.

These challenges can be framed within the contextual bandit problem [Wang et al., 2005, Langford and Zhang, 2007]—a fundamental online decision-making problem involving contextual information, including unstructured language and vision input, that affects decisions.

1.1 Key Contributions

We introduce GPT-HyperAgent, leveraging pretrained GPT for expressive feature embeddings and integrating HyperAgent [Li et al., 2024b] for scalable uncertainty-guided exploration in contextual bandits with unstructured language input. Vanilla HyperAgent [Li et al., 2024b] was designed based on hypermodel framework [Dwaracherla et al., 2020, Li et al., 2022] and achieves state-of-the-art computational and data efficiency for large-scale deep reinforcement learning benchmarks. Yet, the compatibility of vanilla HyperAgent with foundation models for contextual bandits has never been examined. More importantly, existing literature lack rigorous understanding on HyperAgent or hypermodel-type algorithm under function approximation, and thus cannot provide much guidance on the algorithmic configurations.

In this work, we provide an in-depth theoretical understanding that leads to practical advancement, and close a fundamental gap in the theory for scalable randomized exploration algorithms.

Theoretical Understanding

- Efficient and Scalable Uncertainty Estimation: We theoretically prove that HyperAgent achieves fast and scalable incremental uncertainty estimation with $\tilde{O}(\log T)$ per-step computational complexity over T periods under the linear realizable assumption. This enables real-time adaptation and efficient handling of increasing data volumes. The underlying mechanism is incrementally updating an approximate factor of the covariance matrix via the outer product of the feature vector and a random vector draw from perturbation distribution at each time period.
- Distribution-Dependent Regret: We develop a general regret analysis framework that leads to a regret bound dependent on the reference distribution. Certain continuous-support reference distributions (e.g., Gaussian) outperform ensemble sampling, which is a special case of HyperAgent with the reference distribution being a uniform distribution over coordinates. This distinction is due to the probability of optimism, proved via new anti-concentration bounds we developed. To highlight, with various continuous-support reference distributions, HyperAgent can match the frequentist regret of exact Thompson sampling (TS) in linear contextual bandits, closing a fundamental gap in the theory for scalable randomized exploration methods.

Practical Guidance and Performance

- Separated Distributions & Empirical Validation: We demonstrate that update and perturbation distributions in HyperAgent can be chosen separately, unconventional to existing literature. This allows dual benefits by using discrete-support update distributions for lower computation cost while maintaining the advantages of continuous-support reference distributions. This algorithmic insight is theoretically justified under linear realizable assumption and empirically validated in neural contextual bandit setups.

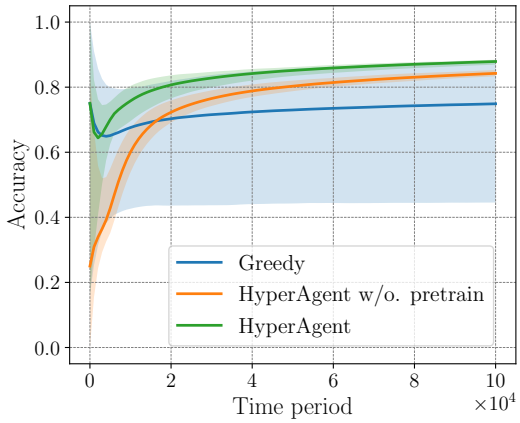


Figure 2: We simulate the human-AI pipeline for content moderation, focusing on hate speech detection, using multiple random seeds. The solid line indicates average performance, while the shaded area shows variation across simulations. The 'HyperAgent' uncertainty-aware moderation policy *reduces labeling effort by tenfold* and achieves higher detection accuracy than the uncertainty-agnostic 'Greedy' policy, which exhibits significant variance across simulations. Additionally, using pretrained GPT embeddings significantly enhances initial performance compared to random initialization.

- Foundation Model Online Decisions: By addressing the challenges of uncertainty estimation and scalable exploration, GPT-HyperAgent advances the state-of-the-art in online decisions with foundation models. This is crucial for applications like content moderation with human feedback, where the balance between reducing human workload and ensuring long-term safety is paramount. Fig. 2 demonstrates the practical effectiveness of GPT-HyperAgent.

2 Problem formulation and HyperAgent algorithm

2.1 Sequential decision-making under uncertainty

We consider a environment involving a set of actions $\mathcal{A}$ and a ground-truth real-valued functions $f^* : \mathcal{A} \mapsto \mathbb{R}$. We will define random variables with respect to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The agent is uncertain about the function f^* in the beginning. At each time t, the agent is presented with a possibly random subset $\mathcal{A}_t \subseteq \mathcal{A}$ and selects an action $A_t \in \mathcal{A}_t$, after which she observes a reward Y_t. We denote by $\mathcal{H}_t$ the σ-algebra generated from history $(\mathcal{A}_1, A_1, Y_1, \ldots, \mathcal{A}_{t-1}, A_{t-1}, Y_{t-1}, \mathcal{A}_t)$ of observations available to the agent when choosing an action A_t. The agent employs a policy $\pi = \{\pi_t \mid t \in \mathbb{N}\}$, which is a deterministic sequence of functions, each mapping the history $\mathcal{H}_t$ to a probability distribution over actions $\mathcal{A}$. For each realization of $H_t \in \mathcal{H}_t, \pi_t(H_t)$ is a distribution over $\mathcal{A}$ with support $\mathcal{A}_t$, though with some abuse of notation, we will often write this distribution as π_t. The action A_t is selected by sampling from the distribution π_t, so that $\mathbb{P}(A_t \in \cdot \mid \pi_t) = \mathbb{P}(A_t \in \cdot \mid H_t) = \pi_t(\cdot)$. We assume that $\mathbb{E}[Y_t \mid H_t, A_t] = f^*(A_t)$. In other words, the realized reward is the mean reward value corrupted by zero mean noise. We will also assume that for each $t \in \mathbb{N}$, $\arg\max_{a \in \mathcal{A}_t} f^*(a)$ is nonempty with probability one, though algorithms and results can be generalized to handle cases where this assumption does not hold. The T-period regret of a policy π is the random variable defined by $R(T) = \sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} f^*(a) - f^*(A_t)$.

Example 1 (Contextual Bandit Models)**.** *The contextual bandit model [Langford and Zhang, 2007, Wang et al., 2005] is a special case of the formulation presented above. In such a model, an exogenous Markov process X_t taking values in a set $\mathscr{X}$ influences rewards. In particular, the expected reward at time t is given by $f^*(a, X_t)$. However, this is mathematically equivalent to a problem with stochastic time-variant decision sets $\mathcal{A}_t$. In particular, one can define the set of actions to be the set of state-action pairs $\mathcal{A} := \{(x, a) : x \in \mathscr{X}, a \in \mathcal{A}(x)\}$, and the set of available actions to be $\mathcal{A}_t = \{(X_t, a) : a \in \mathcal{A}(X_t)\}$.*

Algorithm 1 Generic HyperAgent for bandits (time step t)

Require: Reference distribution P_ζ, Update distribution P_ξ, Perturbation distribution P_z; Index dim. M, Perturbation level σ, Regularization λ, Update ratio B, Initialization θ_0, Data buffer D

1: **Index Sampling**: sample index $\zeta_t \overset{i.i.d.}{\sim} P_\zeta$ and play $A_t = \arg\max_{a \in \mathcal{A}_t} f_{\theta_{t-1}}(a, \zeta_t)$.

2: Receive noisy feedback Y_t, sample perturbations $\mathbf{z}_t \overset{i.i.d.}{\sim} P_z$ and add $\{(A_t, Y_t, \mathbf{z}_t)\}$ to buffer D.

3: **Incremental update**: take B-step gradient descent w.r.t. $\tilde{L}(\theta; \tilde{\Xi}, \tilde{D})$ from θ_{t-1} to θ_t

$$\tilde{L}(\theta; \tilde{\Xi}, \tilde{D}) := \frac{1}{|\tilde{\Xi}||\tilde{D}|} \sum_{\xi \in \tilde{\Xi}, (A_s, Y_s, \mathbf{z}_s) \in \tilde{D}} (f_\theta(A_s, \xi) - Y_s - \sigma \mathbf{z}_s^\top \xi)^2 + \frac{\lambda}{|D|} \|\theta\|^2 \tag{1}$$

Within the *online automated content moderation* task as in Fig. 1, the context X_t consists of text, image or video content that the user submits to the platform while the moderator needs to choose an action $a \in \{\text{publish}, \text{remove}\}$.

Example 2 (Linear Realizable Rewards)**.** *We say the reward function f^* is linear realizable w.r.t a known feature map $\phi : \mathcal{A} \mapsto \mathbb{B}^d$ if there exists a vector $\theta^* \in \mathbb{B}^d$, such that $f^*(a) = \langle \phi(a), \theta^* \rangle$.*

These examples are widely studied in the literature of linear contextual bandits [Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011a, Dani et al., 2008].

2.2 HyperAgent, hypermodel and index sampling

Vanilla HyperAgent [Li et al., 2024b] was shown state-of-the-arts performance in large-scale deep RL benchmarks. Its success can be attributed to several key mechanisms: hypermodel [Dwaracherla et al., 2020, Li et al., 2022], incremental updates, and index-based approximate Thompson sampling, known as index sampling.

The hypermodel f_θ, parameterized by θ, is designed for uncertainty estimation. It takes an input $x \in \mathbb{R}^d$ and a random index ζ draw from a fixed reference distribution P_ζ, producing an *index sample* $f_\theta(x, \zeta)$, reflecting a predictive sample from a desired distribution. For instance, if we want to approximate a linear-Gaussian distribution $N(x^\top \mu, x^\top \boldsymbol{\Sigma} x)$, one can use linear hypermodel $f_\theta(x, \zeta) = \langle x, \mu + \boldsymbol{A}\zeta \rangle$, with $\theta = (\boldsymbol{A} \in \mathbb{R}^{d \times M}, \mu \in \mathbb{R}^d)$ and Gaussian reference distribution $P_\zeta = N(0, I_M)$. It essentially performs a Box-Muller transformation: when $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Sigma}$, the index sample $f_\theta(x, \zeta) \sim N(x^\top \mu, x^\top \boldsymbol{\Sigma} x)$. Another example is (symmetric) ensemble sampling, where the reference distribution $P_\zeta = \mathcal{U}(\pm e_1, \ldots, \pm e_M)$ is uniform over coordinates and $\theta = (\boldsymbol{A} := [\tilde{\theta}_1, \ldots, \tilde{\theta}_M] \in \mathbb{R}^{d \times M}, \mu \in \mathbb{R}^d)$ contains M ensemble models such that each $\tilde{\theta}_m \sim N(0, \boldsymbol{\Sigma})$. In this case, the variance of index sample $\mathrm{Var}_{\zeta \sim P_\zeta}(f(x, \zeta)) = \frac{1}{M} \sum_{i=1}^M [(\tilde{\theta}_i)^\top x]^2 \to x^\top \boldsymbol{\Sigma} x$ as M increases. In both cases, index samples can capture the variance of target distribution by varying the indices $\{\zeta\}$. HyperAgent incrementally adjust hypermodel's parameters θ over time, aiming to incrementally refine its uncertainty estimation as data accumulated. Specifically, [Li et al., 2024b] proves that HyperAgent indeed achieves incremental posterior update without the need for conjugacy in tabular reinforcement learning setups.

We present the general HyperAgent framework for bandit environments, as shown in Algorithm 1. At each time step t, HyperAgent takes an index sample $f_{\theta_{t-1}}(a, \zeta)$ and selects the greedy action accordingly. This procedure is called *index sampling* (IS). After taking action A_t, it receives feedback Y_t and generates a perturbation vector $\mathbf{z}_t$ drawn from P_z. The parameters are then updated incrementally via off-the-shelf stochastic gradient methods. The objective function $\tilde{L}(\theta; \tilde{\Xi}, \tilde{D})$ is a sample-average approximation of the true objective:

$$L(\theta; D) := \frac{1}{|D|} \mathbb{E}_{\xi \sim P_\xi} \sum_{(A_s, Y_s, \mathbf{z}_s) \in D} \left(f_\theta(A_s, \xi) - Y_s - \sigma \mathbf{z}_s^\top \xi\right)^2 + \frac{\lambda}{|D|} \|\theta\|^2, \tag{2}$$

where the indices $\xi \in \tilde{\Xi}$ are drawn i.i.d. from the update distribution P_ξ and $\tilde{D}$ is a sampled mini-batch from D. Existing literature for Ensemble+ [Osband et al., 2019], Hypermodel [Dwaracherla et al., 2020, Li et al., 2022], Epistemic Neural Networks (ENN) [Osband et al., 2023] and vanilla HyperAgent [Li et al., 2024b] when specified in bandit environments can be regarded as special cases of the

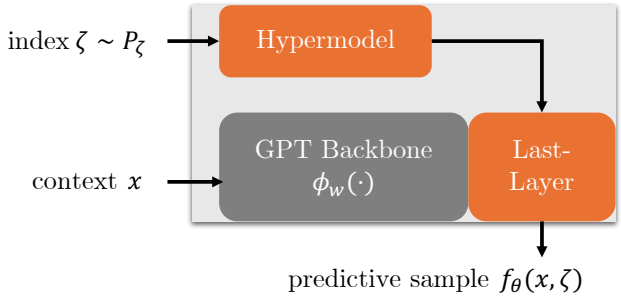


Figure 3: The network structure of HyperAgent using GPT pretrained feature embedding.

HyperAgent framework in Algorithm 1, as they all use identical reference and update distributions, i.e. $P_\zeta = P_\xi$. Specifically, Ensemble+ [Osband et al., 2019] coincides with a special case of Algorithm 1, with P_ξ and P_ζ both being uniform distributions over coordinate vectors [Dwaracherla et al., 2020, Li et al., 2024b], a discrete-support distribution. Despite repeated empirical claims [Dwaracherla et al., 2020, Li et al., 2022, Osband et al., 2023, Li et al., 2024b] suggesting that hypermodel with continuous-support reference distribution P_ζ (e.g. Gaussian) exhibits advantages over ensemble-based methods with discrete-support P_ζ, the reasons are not well explained. Additionally, as shown in Eq. (1), sample-based approximation on continuous-support distribution P_ξ requires a large size of $|\tilde{\Xi}|$, incurring high computational costs [Li et al., 2024b]. Existing works also provide no practical guidance on the configuration of several key distributions P_ζ, P_ξ and P_z due to limited theoretical understanding. Additionally, there is no rigorous justification for HyperAgent regarding uncertainty representation and regret-computation trade-offs, even with linear function approximation. In this work, we build a theoretical framework to address these questions and provide practical guidance, which is essential for integrate HyperAgent with foundation models that demands high computational power.

2.3 Integration with foundation models

Following the last-layer linear hypermodel construction [Li et al., 2024b], we use the pretrained backbone of GPT-2 as the feature extractor $\phi_w(\cdot)$, resulting in the context-aware GPT-HyperAgent: it takes the context x and a random index ζ as input and outputs a value for each action $a \in \mathcal{A}$

$$f_\theta(x, \zeta)[a] = \langle \phi_w(x), \mathbf{A}^a \zeta + b^a \rangle := f_\theta((x, a), \zeta),$$

where and $\mathbf{A}^a$ and b^a are action-specific parameters for each action a in the valid decision set $\mathcal{A}(x)$ associated with the context x. Here we use unified notations from the contextual bandit models in example 1.

3 Theoretical analysis

We start by providing a general analytical framework for agent, potentially randomized, operating in the generic bandit environments. Let us introduce a few necessary definitions to facilitate the understanding and analysis. The confidence bound is used for uncertainty estimation over the ture function f^* given the history $\mathcal{H}_t$.

Definition 1 (Confidence bounds)**.** *Confidence bounds are a sequence of real-valued $\mathcal{H}_t$-measurable functions $\mathrm{L}_t(\cdot)$ and $\mathrm{U}_t(\cdot)$ for $t \in [T]$ such that, w.p. at least $1-\delta$, the joint event $\mathcal{E} = \cap_{t\in[T]}\mathcal{E}_t$ holds, where $\mathcal{E}_t := \{f^*(a) \in [\mathrm{L}_t(a), \mathrm{U}_t(a)], \forall a \in \mathcal{A}_t\}$.*

The agent may not perform well unless it is well-behaved, defined by *reasonableness* and *optimism*. Intuitively, an agent that explores too much or too little will incur a high regret. Reasonableness and optimism are the mechanisms for controlling these potential flaws respectively.

Definition 2 (Reasonableness)**.** *Given confidence bounds $\mathrm{L}_t(\cdot)$ and $\mathrm{U}_t(\cdot)$ for $t \in [T]$, an (randomized) agent is called reasonable if it produces a sequence of functions $(\tilde{f}_t(\cdot), t \in [T])$ such that w.p. at least $1-\delta$, the joint event $\tilde{\mathcal{E}} = \cap_{t\in[T]}\tilde{\mathcal{E}}_t$ holds, where $\tilde{\mathcal{E}}_t := \{\tilde{f}_t(a) \in [\mathrm{L}_t(a), \mathrm{U}_t(a)], \forall a \in \mathcal{A}_t\}$.*

In short, *reasonableness* ensures that the chosen action according to $\tilde{f}_t$ is close to the best action which ensures agent does not explore actions unnecessarily. The following *optimism* guarantees the agent sufficient explores.

Definition 3 (p-optimism)**.** *Let* p *be a sequence of positive real number* $(p_t, t \in [T])$*. We say an (randomized) agent is* p*-optimistic when it produces a sequence of functions* $(\tilde{f}_t(\cdot), t \in [T])$ *such that for all* $t \in [T]$*,* $\tilde{f}_t(\cdot)$ *is* p_t*-optimistic, i.e.,* $\mathbb{P}(\max_{a\in\mathcal{A}_t} \tilde{f}_t(a) \geq \max_{a\in\mathcal{A}_t} f^*(a) \mid \mathcal{H}_t) \geq p_t$*.*

The generic agent satisfying the conditions on *reasonableness* and *optimism* has desired behavior.

Theorem 1 (General regret bound)**.** *Given confidence bound in Definition 1 and assume the agent is reasonable and optimistic (with parameter* $\mathrm{p} = (p_t, y \in [T])$*, then we have*

$$R(T) \leq \sum_{t=1}^{T} \frac{1}{p_t} \mathbb{E}[U_t(A_t) - L_t(A_t) \mid \mathcal{H}_t] + \sum_{t=1}^{T} U_t(A_t) - L_t(A_t) \tag{3}$$

Proof. Let $A_t = \max_{a\in\mathcal{A}_t} \tilde{f}_t(a)$ and $A_t^* = \max_{a\in\mathcal{A}_t} f^*(a)$. Let $B_t = \max_{a\in\mathcal{A}_t} L_t(a)$, which is $\mathcal{H}_t$-measurable. Conditioned on the event $\mathcal{E} \cap \tilde{\mathcal{E}}$, both $f^*(A_t^*) \geq B_t$ and $\tilde{f}_t(A_t) \geq B_t$ hold. By p-optimism and the fact $(f^*(A_t^*) - B_t)$ is $\mathcal{H}_t$-measurable and positive,

$$p_t \leq \mathbb{P}(f_t(A_t) - B_t \geq f^*(A_t^*) - B_t \mid \mathcal{H}_t) \overset{(*)}{\leq} \mathbb{E}[f_t(A_t) - B_t \mid \mathcal{H}_t]/(f^*(A_t^*) - B_t),$$

where $(*)$ is due to Markov inequality. Rearranging and using the additional fact $B_t \geq L_t(A_t)$ yield

$$f^*(A_t^*) - \tilde{f}_t(A_t) \leq f^*(A_t^*) - B_t \leq \frac{1}{p_t}\mathbb{E}[f_t(A_t) - B_t \mid \mathcal{H}_t] \leq \frac{1}{p_t}\mathbb{E}[U_t(A_t) - L_t(A_t) \mid \mathcal{H}_t]. \tag{4}$$

By the reasonableness, $\tilde{f}_t(A_t) \leq U_t(A_t)$. Then, from the definition of confidence bounds

$$\tilde{f}_t(A_t) - f^*(A_t) \leq U_t(A_t) - L_t(A_t) \tag{5}$$

Putting Eqs. (4) and (5) together and then summing over the time index t give the final results. □

A close inspection on Eq. (3) informs that the regret is sublinear as long as the confidence bounds are converging to the true function f^* as more information gathering in. In the following section, we illustrate these insights with linear setups and provide rigorous justifications for HyperAgent.

3.1 Insight from Linear HyperAgent

Consider the functional form for HyperAgent in a linear setup at time t:

$$\tilde{f}_t(a) := f_{\theta_{t-1}}(a, \zeta_t) = \langle \phi(a), \beta_t \mathbf{A}_{t-1}\zeta_t + \mu_{t-1} \rangle, \quad \forall a \in \mathcal{A}, \tag{6}$$

where β_t is an inflation coefficient defined later, $\phi(\cdot)$ is a feature map introduced in Example 2, and $\theta_t = (\mathbf{A}_t, \mu_t)$ are parameters representing uncertainty.

In the context of GPT-HyperAgent, if the underlying reward function f^* can be linearly approximated using GPT-2's pretrained feature embeddings, then we can freeze the GPT-2 torso during training Hyperagent, as shown in Figure 3. This approach, referred to as the frozen-GPT-torso method, is essentially an instance of linear HyperAgent and will be evaluated in Section 4. The assumption of linear realizability will be discussed formally in Assumption 1.

With the form of linear HyperAgent, we theoretically identify the general conditions for the update and perturbation distribution (P_ξ, P_z) that permit scalable uncertainty estimation via incremental posterior approximation. Then, we investigate the reasonableness and optimism condition through several reference distributions P_ζ.

Definition 4 (Isotropic)**.** *A distribution* P *over* $\mathbb{R}^M$ *is called isotropic if* $\mathbb{E}_{X\sim P}[X_i X_j] = \delta_{ij}$*, i.e.,* $\mathbb{E}_{X\sim P}[XX^\top] = I$*. Equivalently,* P *is isotropic if* $\mathbb{E}_{X\sim P}[\langle X, x\rangle^2] = \|x\|^2$*, for all* $x \in \mathbb{R}^M$*.*

Proposition 1. *If the update distribution* P_ξ *is zero-mean and isotropic, linear HyperAgent (Eq. (6)) with objective in Eq. (2) and full data buffer* $D = \mathcal{H}$ *permits closed-form incremental update, i.e.,*

$$\mathbf{A}_t = \boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_{t-1}^{-1}\mathbf{A}_{t-1} + \phi(A_t)\mathbf{z}_t^\top), \quad \mu_t = \boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_{t-1}^{-1}\mu_{t-1} + \phi(A_t)Y_t) \tag{7}$$

where $\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_{t-1}^{-1} + \phi(A_t)\phi(A_t)^\top$ *and* $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_{t-1} - \frac{\boldsymbol{\Sigma}_{t-1}\phi(A_t)\phi(A_t)^\top\boldsymbol{\Sigma}_{t-1}}{1+\phi(A_t)^\top\boldsymbol{\Sigma}_{t-1}\phi(A_t)}$*. The initialization is* $\mu_0 = 0$ *and* $\mathbf{A}_0 = \boldsymbol{\Sigma}_0^\top \mathbf{Z}_0$ *where* $\boldsymbol{\Sigma}_0^{-1} = \lambda I$ *and* $\mathbf{Z}_0 = (\mathbf{z}_{0,1}, \ldots, \mathbf{z}_{0,d})^\top$ *with each* $\mathbf{z}_{0,i} \sim P_z$*.*

The proof can be found in Appendix D. We discuss the 5 isotropic distributions in $\mathbb{R}^M$ including continuous-support (1) Gaussian $N(0, I)$, (2) (Spherical) Uniform over sphere $\sqrt{M}\mathcal{U}(\mathbb{S}^{M-1})$; and discrete support (1) (Cube) Uniform over cube $\mathcal{U}(\{1, -1\}^M)$, (2) (Coord) Uniform over coordinates $\mathcal{U}(\{\pm e_i\}_{i\in[M]})$ and (3) Sparse distribution in Appendix E. Next, we state the key lemma showing when perturbation distribution P_z allows fast incremental uncertainty estimation.

Lemma 1 (Incremental Uncertainty Estimation)**.** *For $t \in [T]$, let $\boldsymbol{\Sigma}_t$ and $\mathbf{A}_t$ be defined as in Proposition 1 and let the good event be*

$$\mathcal{G}_t := \{(1/2)x^\top \boldsymbol{\Sigma}_t x \le x^\top \mathbf{A}_t \mathbf{A}_t^\top x \le (3/2)x^\top \boldsymbol{\Sigma}_t x, \quad \forall x \in \mathbb{R}^d\}.$$

Let $s_{\min}^2 = \inf_{a\in\mathbb{S}^{d-1}} a^\top \boldsymbol{\Sigma}_0^{-1} a$ and $s_{\max}^2 = \sup_{a\in\mathbb{S}^{d-1}} a^\top \boldsymbol{\Sigma}_0^{-1} a$. When

$$M \ge 320\left(d\log\left(\left(1 + (48/s_{\min})\sqrt{s_{\max}^2 + T}\right)/\delta\right) + \log(1 + T/s_{\min}^2)\right) \simeq d\log T, \tag{8}$$

and P_z is $\sqrt{\frac{1}{M}}$-sub-Gaussian and unit-norm, the event $\mathcal{G} = \cap_{t\in[T]}\mathcal{G}_t$ happens w.p. at least $1-\delta$.

Remark 1. *Briefly speaking, the technical difficulty comes from the sequential dependence between a series of the perturbation random vectors and a serious high-dimensional random variables arise in the sequential decision processes. Our innovation is that we derive (1) a variance-aware discretization argument and (2) a reduction to sequential random projection [Li, 2024a]. Classical standard discretization arrives at a exponentially larger $M = O(dT^2 \log T)$, which is unacceptable for proving scalability. Our techniques effectively address the challenge arises in the high-dimensional sequential dependency structure. For the discussion on specific instances of perturbation distributions, see Remark 5. The proof of Lemma 1 can be found in Appendix B, where we clearly articulate the technical difficulties around Fig. 8 in Appendix B.2.*

Next, we are going to discuss the synergy effect of these distributions on the regret of HyperAgent. We first state the environment assumption under which we provide specific regret analysis.

Assumption 1. *The reward function f^* is linear realizable w.r.t. the feature mapping $\phi(\cdot)$ and the noisy feedback Y_t satisfies $\mathbb{E}\left[\exp\left\{s(Y_t - f^*(A_t))\right\} \mid \mathcal{H}_t, A_t\right] \le \exp\left\{s^2/2\right\}, \forall s \in \mathbb{R}$.*

Adapting the results from [Abbasi-Yadkori et al., 2011b, Abeille and Lazaric, 2017], we define $\beta_t = \sqrt{\lambda} + \sqrt{2\log(1/\delta) + \log\det(\boldsymbol{\Sigma}_{t-1}^{-1}/\lambda^d)}$. Under assumption 1, for the purpose of analysis, we define a slightly inflated confidence bounds as

$$L_t(\cdot; P_\zeta) = (\langle \mu_{t-1}, \phi(\cdot)\rangle - \beta_t \rho(P_\zeta)\|\phi(\cdot)\|_{\boldsymbol{\Sigma}_{t-1}}) \vee (-1),$$
$$U_t(\cdot; P_\zeta) = (\langle \mu_{t-1}, \phi(\cdot)\rangle + \beta_t \rho(P_\zeta)\|\phi(\cdot)\|_{\boldsymbol{\Sigma}_{t-1}}) \wedge 1.$$

$\rho(P_\zeta)$ will be defined via $\rho_1 = O(\sqrt{M\log(M/\delta)})$, $\rho_2 = O(\sqrt{M})$, and $\rho_3 = O(\sqrt{\log(|\mathcal{A}|/\delta)})$.

Proposition 2. *Under linear setups in Eq. (6) and Proposition 1, if Eq. (8) is satisfied, HyperAgent is reasonable, i.e., $\forall t \in [T], \tilde{f}_t(\cdot) = f_{\theta_{t-1}}(\cdot, \zeta_t) \in [L_t(\cdot; P_\zeta), U_t(\cdot; P_\zeta)]$ w.p. $1-\delta$.*

Proposition 3. *Under linear setups in Eq. (6) and Proposition 1, if Eq. (8) is satisfied, HyperAgent using reference distribution P_ζ is $p(P_\zeta)$-optimistic.*

P_ζ	Gaussian $N(0, I_M)$	Spherical $\sqrt{M}\mathcal{U}(\mathbb{S}^{M-1})$	Cube $\mathcal{U}(\{1,-1\}^M)$	Coord $\mathcal{U}(\{\pm e_i\}_{i\in[M]})$	Sparse
$\rho(P_\zeta)$	$\rho_1 \wedge \rho_3$	$\rho_2 \wedge \rho_3$	$\rho_2 \wedge \rho_3$	ρ_2	ρ_2
$p(P_\zeta)$	$\frac{1}{4\sqrt{e\pi}}$	$\frac{1}{2} - \frac{e^{1/12}}{\sqrt{2\pi}}$	$7/32$	$\frac{1}{2M}$	N/A

Table 1: The coefficient $\rho(P_\zeta)$ and $p(P_\zeta)$ related to the reasonableness and optimism condition.

The specific value of $\rho(P_\zeta)$ and $p(P_\zeta)$ is shown in Table 1, where the proof of Propositions 2 and 3 can be found in Appendix C. To highlight, the proof of optimism involves newly developed anti-concentration bounds in Appendix E. The reasonableness and optimism, together with the general regret bound in Theorem 1, have a direct consequence of the following specific regret bound.

Theorem 2 (Distribution-dependent regret bound). *Let $\beta = \max_{t\in[T]} \beta_t$. Consider assumption 1 and the linear setups in Eq. (6). If the update distribution P_ξ is zero-mean and isotropic, and Eq. (8) is satisfied, then HyperAgent(P_ζ) has T-period regret*

$$R(T) \le \frac{\rho(P_\zeta)}{p(P_\zeta)}\beta\left(\sqrt{dT\log\left(1+\frac{T}{\lambda d}\right)}+\sqrt{\frac{T}{\lambda}\log\left(\frac{T}{\lambda\delta}\right)}\right), \quad \textit{with probability} \quad 1-\delta.$$

Remark 2. *Theorem 2 provides a distribution-dependent bound characterized by the **ratio** $\rho(P_\zeta)/p(P_\zeta)$. This **ratio** implies that as long as the distribution P_ζ allows sufficient exploration (larger $p(P_\zeta)$) but not too much (smaller $\rho(P_\zeta)$), the regret is smaller. This explains why ensemble sampling (ES) methods perform relatively worse [Dwaracherla et al., 2020, Osband et al., 2023, Li et al., 2022, 2024b]. Specifically, ES (HyperAgent with Coord P_ζ) has a **ratio** of $O(M^{3/2})$, while other reference distributions have a **ratio** of $\tilde{O}(\sqrt{M}\wedge\sqrt{\log|\mathcal{A}|})$. For $M=\tilde{O}(d\log T)$, Theorem 2 suggests the regret of ES is $O((d\log T)^{5/2}\sqrt{T})$, matching the concurrent analysis of ES [Janz et al., 2023].*

*When using continuous-support reference distribution P_ζ, including Gaussian, Spherical, and Cube, our HyperAgent achieves a tighter bound, as shown in Table 2. Notably, our bound suggests that when M exceeds a threshold in the order of $O(d\log T)$, the regret bound of HyperAgent with continuous-support P_ζ has no dependence on M in the setting of finite decision sets where the **ratio** would be $\tilde{O}(\log|\mathcal{A}|)$. This theoretical finding matches the empirical observation in Appendix F.1, demonstrating the predictive power and practical guidance of distribution-dependent analysis.*

Remark 3. *Theorem 2 allows a fine-grained analysis for both finite and compact, and both time-variant and time-invariant decision sets due to the action-dependent nature of the general regret bound in Theorem 1. Existing Frequentist regret bounds for TS [Abeille and Lazaric, 2017] and approximate TS methods, including LMC [Xu et al., 2022] and ES [Janz et al., 2023], are specialized to compact decision sets. The Bayesian analysis of ES [Qin et al., 2022] applies only to time-invariant finite decision sets. We summarize the regret bound and corresponding per-step computation complexity in Table 2. Note that, except for ES [Qin et al., 2022], other upper bounds are all Frequentist regret.*

Decision Sets	Invariant & Compact	Variant & Compact	Invariant & Finite	Variant & Finite
Lower Bound	$\Omega(d\sqrt{T})$	$\Omega(d\sqrt{T\log T})$	$\Omega(\sqrt{dT\log\lvert\mathcal{A}\rvert})$	$\Omega(\sqrt{dT\log\lvert\mathcal{A}\rvert\log T})$
TS	$O(d^{\frac{3}{2}}\sqrt{T}\log T)$	$O(d^{\frac{3}{2}}\sqrt{T}\log T)$	$O(d\sqrt{T\log\lvert\mathcal{A}\rvert}\log T)$	$O(d\sqrt{T\log\lvert\mathcal{A}\rvert}\log T)$
ES[Qin]	N/A	N/A	$O(\sqrt{dT\log\lvert\mathcal{A}\rvert\log(\lvert\mathcal{A}\rvert T/d)})$	N/A
LMC[Xu]	$O((d\log T)^{\frac{3}{2}}\sqrt{T})$	$O((d\log T)^{\frac{3}{2}}\sqrt{T})$	N/A	N/A
ES[Janz]	$O((d\log T)^{\frac{5}{2}}\sqrt{T})$	$O((d\log T)^{\frac{5}{2}}\sqrt{T})$	N/A	N/A
(Ours)	$O(d^{\frac{3}{2}}\sqrt{T}(\log T)^{\frac{3}{2}})$	$O(d^{\frac{3}{2}}\sqrt{T}(\log T)^{\frac{3}{2}})$	$O(d\sqrt{T\log\lvert\mathcal{A}\rvert}\log T)$	$O(d\sqrt{T\log\lvert\mathcal{A}\rvert}\log T)$

Table 2: Regret lower and upper bounds under various decision set setups. The per-step computation complexity is $O(d^2+d|\mathcal{A}|T)$ for ES [Qin et al., 2022], $O(d^2T)$ for LMC [Xu et al., 2022], $O(d^3\log T)$ for ES [Janz et al., 2023], and $O(d^3\log T)$ for our HyperAgent. The lower bounds for the setups of (1) invariant and compact, (2) variant and compact, (3) invariant and finite, and (4) variant and finite decision sets are from the following sources, respectively: [Rusmevichientong and Tsitsiklis, 2010, Li et al., 2024a, Zhou, 2019, Li et al., 2024a].

Remark 4. *As shown in Table 2, we provide the first result on approximate TS that achieves both provable scalability – $O(d^3\log T)$ per-step computation – and near-optimal regret matching exact TS [Agrawal and Goyal, 2013, Abeille and Lazaric, 2017] across all decision set setups. Notably, we achieve an exponential improvement in the T-factor of per-step computation complexity compared to [Qin et al., 2022, Xu et al., 2022], and an $O(d(\log T)^2)$ multiplicative factor improvement in the regret bound compared to concurrent work on ES [Janz et al., 2023]. This closes the theoretical gap in scalable randomized exploration in terms of both computation and regret.*

The proof of Theorem 2 can be found in Appendix C.

4 Experiments

We first conducted synthetic experiments on both linear and nonlinear contextual bandits to validate our theoretical insights and offer practical guidance for using foundation models in online decisions.

Then, we augmented GPT-2 with HyperAgent to demonstrate its effectiveness in handling natural language tasks requiring exploration, such as automated content moderation with human feedback. GPT-HyperAgent significantly improves data efficiency in human-AI interactions for risk oversight in content moderation.

4.1 Synthetic Experiments

We performed experiments on synthetic bandit tasks with both linear and nonlinear reward functions to validate our theoretical insights and provide practical guidance for scenarios with frozen and unfrozen backbones.

Linear reward functions:

- **Setup**: Mimics a scenario where the foundation model backbone induced feature mapping is expressive enough and thus is frozen through the learning and decision processes. Studies the impact on various perturbation and reference distributions.
- **Results**: Demonstrates HyperAgent's superiority over Ensemble+ [Osband et al., 2019] in linear contextual bandits when M is relatively small.

Neural reward functions:

- **Setup**: Mimics a scenario where the foundation model backbone is not frozen. Studies the impact on various update distributions.
- **Results**: Validates the advantage of separating reference and update distributions in neural contextual bandits.

4.1.1 Linear Contextual Bandits

We begin by examining the advantages of HyperAgent in linear bandit, which can be understood as scenario where the foundation model backbone is fixed. In this experiment, we primarily focus on studying the impact of perturbation and reference distributions on HyperAgent.

Settings: We study the finite-action linear bandit guided by prior research [Russo and Van Roy, 2018] to evaluate HyperAgent. In this task, we construct the finite action set $\mathcal{A}$ by uniformly sampling from the range $[-1/\sqrt{5}, 1/\sqrt{5}]^d$ where d is the ambient dimension of the linear reward function perturbed by an additive Gaussian noise term. To ensure robust results, each experiment is executed a total of 1000 time steps and repeated 200 times. We provide a detailed implementation of the task in Appendix F.1.

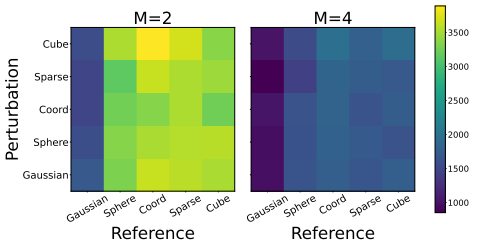


Figure 4: Analysis on the combinations of perturbation and reference distribution. A deeper color signifies lower accumulated regret and hence superior performance.

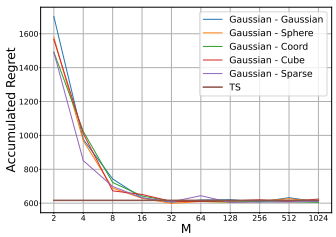


Figure 5: Study on regret under various index dimension M. The label $A - B$ indicates that HyperAgent uses A as the reference distribution and B as the perturbation distribution.

Results Analysis: This experiment involves all 25 combinations of perturbation and reference distribution as highlighted in Fig. 4, using accumulated regret as evaluative metric. Notably, the Gaussian reference demonstrates enhanced performance across all scenarios, while the Sphere reference shows better results with a larger index dimension M. However, the Coord reference distribution, as implemented in Ensemble+, failed to meet satisfactory performance. This underperformance aligns with the theoretical results presented in Theorem 2.

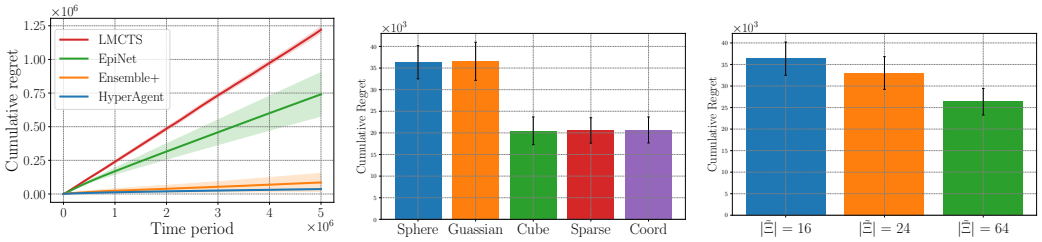
Computation Efficiency: We evaluate HyperAgent employing the Gaussian reference distribution across different index dimensions M, demonstrated in Fig. 5. It is apparent that larger M improves performance, implying that increasing the index dimension could enhance the approximation of the posterior covariance in the closed-form solution. Remarkably, HyperAgent shows performance on par with Thompson sampling with a minimal $M = 8$, highlighting its computational efficiency.

We carry out supplementary experiments on two linear bandit tasks with varying action sizes and dimensions, all of which display similar phenomena that convincingly reinforce most of our theoretical findings. For further information of experimental results, please refer to Appendix F.1, where we discuss when Theorem 2 perfect predicts the empirical results and when it does not.

4.1.2 Neural Contextual Bandits

We then extend HyperAgent to nonlinear tasks, using MLP networks as feature extractors for convenient analysis. In this experiment, we primarily focus on studying the impact of update distributions and demonstrating the advantages of HyperAgent compared to other algorithms that employ approximate posterior sampling.

Settings: We use the neural network to build the ground-truth nonlinear reward function f^*, which includes three fully connected layers, each consisting of 50 units, and connected by ReLU activation functions. The action dimension is set at 100, while the size of $\mathcal{A}$ designated as 1000.



(a) (b) (c)

Figure 6: Experimental results on Neural Bandit. (a) Comparison results with baselines. (b) The impact of various update distribution. (c) The impact of $|\tilde{\Xi}|$ when using Sphere update distribution.

Data Efficiency: We set the update, reference, and perturbation distributions with the same Sphere distribution for HyperAgent and juxtaposed it against several baselines that utilize approximate posterior sampling. These baselines include Ensemble+ [Osband et al., 2018], EpiNet [Osband et al., 2023], and LMCTS [Xu et al., 2022]. To maintain fairness of comparison, all methods employed the identical feature network $\phi_w(a)$, and the update ratio of LMCTS is uniformed with that of HyperAgent. We assessed the performance by using cumulative regret after completion of the training. As depicted in Fig. 6(a), HyperAgent remarkably exceeds other baselines, showcasing its superior data efficiency.

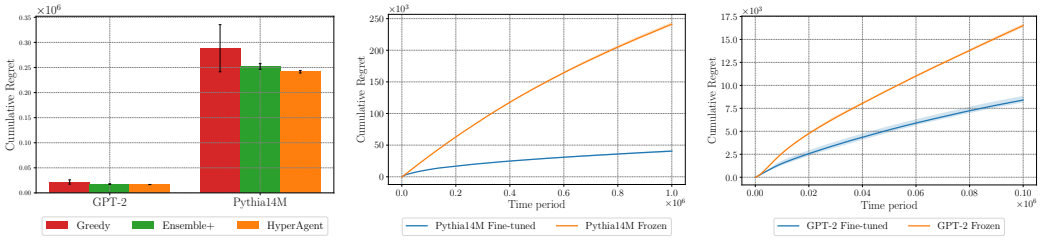
Computation Efficiency: We keep reference and perturbation with the Sphere distribution and evaluate different update distributions. As displayed in Fig. 6(b), the outcomes revealed that discrete update distributions consistently outperformed continuous ones. This is principally because we require a sampling-based approximation for expectation estimation with continuous options, but finite indices allow us to calculate the expectation precisely with discrete distributions as implied in Proposition 1. Notably, as depicted in Fig. 6(c), when employing Sphere update distribution, HyperAgent can further boost performance with larger $|\tilde{\Xi}|$ in Eq. (1). However, larger $|\tilde{\Xi}|$ significantly increases the computational costs. As a pragmatic solution, employing Coord for the update distribution strikes a favourable balance, achieving superior performance with reduced computational expenditure.

We also conduct comparisons involving differing reference and perturbation distributions and other ablation studies. We observe that continuous reference and perturbation distributions offered advantages, consistent with the findings from linear bandit. Please refer to Appendix F.2 for details.

4.2 Content Moderation with Human Feedback

In this experiment, we integrate HyperAgent with existing foundation models to enhance their decision-making capabilities. We utilize LLMs as the backbone to address language tasks that require exploration. Given the high computational cost associated with hyperparameter tuning in LLMs, we adopt the effective settings from the previous study on synthetic tasks. Specifically, we employ the Coord distribution for updates and use the Sphere distribution as both the reference and perturbation distribution.

Settings: We reinterpret the challenge of content moderation as a contextual bandit task, making use of the language dataset[2]. In this task, the agent must decide to either publish or block a content. The agent gains a reward of 0.5 by blocking any content. However, publishing a "free" content earns the agent a reward of 1 and publishing a "hate" content causes a penalty of -0.5.



(a) (b) (c)

Figure 7: Experimental results on content moderation. (a) Comparison results with other decision head with frozen LLM. (b) and (c) Empirical study on the fine-tune to LLM backbone.

Results Analysis: We use two LLMs - Pythia14m and GPT-2 as the feature extractor $\phi_w(x)$ in this experiemt. We firstly compare HyperAgent with Ensemble+ [Osband et al., 2018] and Greedy, a special case with fixed $\mathbf{A}^a = 0$ for all actions. As highlighted in Fig. 7(a), when freezing the feature extractor $\phi_w(x)$, HyperAgent surpasses other methods indicating its exploration efficiency. We also could observe larger model (GPT-2) with more prior knowledge could lead to better performance. To fully unleash the potential of LLMs, we conduct another experiment where the $\phi_w(x)$ are fine-tuned during the training process. As depicted in Fig. 7(b) and (c), this modification significantly reduce cumulative regret when employing HyperAgent.

5 Related works

Foundation Model Decision-making. Recent research applies foundation models, including large language models and generative models, to real-world decision-making applications [Yang et al., 2023, Shinn et al., 2024, Wang et al., 2023, Lee et al., 2023]. Although these systems grant agents decision-making and planning capabilities, they predominantly rely on pre-trained models using offline datasets and may struggle with actively seeking information and resolving uncertainties in an online decision environment that continually presents new, uncovered scenarios. Krishnamurthy et al. [2024] demonstrated that even the most advanced large language model, GPT-4 [OpenAI, 2023] with various advanced prompt design, is ineffective in making online decisions, as seen in a simple multi-armed bandit (MAB) problem, unless supplemented with an MAB-specific uncertain estimation method like external history summarization. More sophisticated algorithmic interventions, such as fine-tuning or dataset curation, might be necessary to enhance LLM-based decision-making agents in complex settings [Krishnamurthy et al., 2024]. Our work addresses this gap by integrating our uncertainty-aware HyperAgent techniques into foundation models, thereby enabling effective online decision-making in complex and uncertain environments.

Scalable Uncertainty Estimation & Exploration. Thompson Sampling (TS) is a favored exploration strategy in online sequential decision-making, balancing exploration and exploitation by sampling from the model's posterior distribution, a Bayesian uncertainty estimation principle. However, TS is computationally feasible primarily in straightforward scenarios where conjugacy permits efficient posterior updates as new data accumulates [Thompson, 1933, Russo et al., 2018]. In many

[2] https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech

practical applications, such as those involving unstructured inputs from language and vision, more complex models are necessary. For these models, exact Bayesian inference becomes computationally infeasible, and no conjugacy is available for Bayesian posterior updates. Various approaches, including Ensemble sampling (Ensemble+) and Langevin Monte-Carlo (LMC-TS), attempt to perform approximate posterior sampling without requiring conjugacy [Osband et al., 2018, 2019, Russo et al., 2018, Xu et al., 2022]. However, managing an ensemble of complex models can significantly increase computational demands, especially as the ensemble size needs to grow to accurately approximate complex posterior distributions [Dwaracherla et al., 2020, Osband et al., 2023, Li et al., 2022, Qin et al., 2022]. Likewise, the computational costs of LMC-based inference are prohibitive in large-scale deep learning systems [Osband et al., 2023]. Hypermodels, ENNs, and HyperAgent aim to efficiently estimate the uncertainty by tracking the approximate posterior distribution with bounded per-step computation, showcasing empirical computational and statistical benefits [Dwaracherla et al., 2020, Li et al., 2022, Osband et al., 2023, Li et al., 2024b]. Ensuring bounded per-step computational complexity is vital for scalability, as computational requirements that scale polynomially with the state-action space and interaction data can become unsustainable [Li et al., 2024b, Lu et al., 2023]. HyperAgent, in particular, exhibits state-of-the-art computational and data efficiency in deep RL benchmarks [Li et al., 2024b]. Nonetheless, the compatibility of HyperAgent with foundation models remains unexamined. Moreover, the existing literature lacks a rigorous understanding of HyperAgent under function approximation, providing limited guidance on algorithmic configurations necessary for scaling up exploration and online decisions with large foundation models that process various modalities, including language.

6 Conclusion and future directions

In this work, we introduced GPT-HyperAgent, a novel integration of foundation models with HyperAgent for online decision-making tasks, focusing on contextual bandits with natural language input. Our contributions include both theoretical insights and practical advancements. Specifically, we proved a regret bound for HyperAgent under linear setups, closing a gap in the theory for scalable exploration algorithms. We showed that perturbation and update distributions in HyperAgent can be chosen separately for computational benefits, providing practical guidance. Empirical results in an online content moderation task validated GPT-HyperAgent's superior online decision capability, demonstrating scalable and efficient performance in real-world safty-critical applications. Promising future directions include

- *Black-Box & Multi-Modal Foundation Models*: Extend GPT-HyperAgent to work with black-box foundation models accessed via APIs. This would allow leveraging powerful pretrained models without requiring access to their internal architectures, making the approach applicable to a wider range of commercial AI services. Specifically, current LLM/VLM APIs provide top-k token logits, text embedding or fine-tuning services by uploading private dataset. Additionally, integrate multi-modal inputs (e.g., combining vision, language, and audio) to tackle more diverse real-world scenarios and improve generalization capabilities.
- *Efficient Human-AI interplay & AI safety*: The AI moderation system, a safety-critical scenario, inevitably relies on human-AI collaboration pipeline. It needs further in-depth studies. Other important applications require human-AI interplay include reward modeling from human feedback by actively query informative data for human review. This is critical for reinforcement learning from human feedback (RLHF) and the efficiency in data-centric AI.
- *Theory beyond linear*: Theoretically understanding on the scalable uncertainty estimation and exploration capabilities of HyperAgent with neural network or general noninear function approximation would further bride the gap between theory and practice, beneficial for design practical solutions for real-world application.

Appendix

Contents

1 Introduction **1**
- 1.1 Key Contributions . . . 2

2 Problem formulation and HyperAgent algorithm **3**
- 2.1 Sequential decision-making under uncertainty . . . 3
- 2.2 HyperAgent, hypermodel and index sampling . . . 4
- 2.3 Integration with foundation models . . . 5

3 Theoretical analysis **5**
- 3.1 Insight from Linear HyperAgent . . . 6

4 Experiments **8**
- 4.1 Synthetic Experiments . . . 9
  - 4.1.1 Linear Contextual Bandits . . . 9
  - 4.1.2 Neural Contextual Bandits . . . 10
- 4.2 Content Moderation with Human Feedback . . . 11

5 Related works **11**

6 Conclusion and future directions **12**

A Additional related works **14**

B Incremental uncertainty estimation in Lemma 1 **14**
- B.1 Probability tools for sequential random projection . . . 14
- B.2 Reduce Lemma 1 to sequential random projection . . . 15

C Regret analysis **19**
- C.1 Proof of Proposition 2 . . . 19
- C.2 Proof of Proposition 3 . . . 20
- C.3 Proof of Theorem 2 . . . 20

D Derivation of the closed-form incremental update **20**
- D.1 Proof of Proposition 1 . . . 21

E Isotropy, Concentration and Anti-concentration **22**
- E.1 Sphere $P_\zeta = \mathcal{U}(\sqrt{M}\mathbb{S}^{M-1})$ . . . 22
- E.2 Cube $P_\zeta = \mathcal{U}(\{1,-1\}^M)$ . . . 24
- E.3 Gaussian $P_\zeta = N(0, I_M)$ . . . 24
- E.4 Coord $P_\zeta = \mathcal{U}(\sqrt{M}\{\pm e_1, \ldots, \pm e_M\})$ . . . 25
- E.5 Sparse distribution P_ζ . . . 25

F In-depth empirical and ablation studies **25**
- F.1 Linear Bandit Task . . . 25
- F.2 Nonlinear Bandit Task . . . 26

A Additional related works

Broad Capabilities and Diverse Applications of Foundation Models. Foundation models, pretrained on diverse datasets encompassing audio, vision, language, and other modalities, have demonstrated exceptional capabilities across a wide range of downstream tasks [Bommasani et al., 2021, Reid et al., 2024, OpenAI, 2023, Collaboration et al., 2024]. Their applications extend to real-world scenarios such as dialogue systems, autonomous driving, healthcare, robotics, and bio-engineering [OpenAI, 2023, Yan et al., 2024, Chen et al., 2023b, Zhang et al., 2023, Saab et al., 2024, Collaboration et al., 2024, Huang et al., 2024]. In these settings, foundation models encounter unique challenges like interacting with external environments, adapting to varied task modalities, and performing long-term reasoning and planning [Nakano et al., 2021, Yao et al., 2022].

Foundation Models in Sequential Decision-Making. Sequential decision-making, which includes domains such as reinforcement learning (RL) and bandit problems, has traditionally focused on task-specific settings with limited prior knowledge, achieving notable success in tasks such as board games, video games, and robotics manipulation [Sutton and Barto, 2018, Lattimore and Szepesvári, 2020, Schrittwieser et al., 2020, Li et al., 2024b, Kalashnikov et al., 2018]. However, these traditional methods, learning from scratch, often face issues with generalization and data efficiency.

The integration of foundation models into this field represents a paradigm shift, giving rise to "foundation agents" that utilize extensive pretraining to solve a broader range of tasks more efficiently [Yang et al., 2023]. However, current foundation agents that use off-the-shelf models such as large language models (LLMs) and vision language models (VLMs) face significant challenges related to controllability, reproducibility, and efficiency [Chen et al., 2023a]. These models, not originally designed for decision-making tasks like action generation or self-evaluation, demonstrate limited capabilities in few-shot prompting and in-context learning, which do not effectively address the needs for exploration and exploitation in sequential decision-making [Brown et al., 2020, Krishnamurthy et al., 2024]. To address these shortcomings, (1) more robust algorithmic interventions such as fine-tuning or (2) bottom-up re-design of the foundation models for sequential decision-making may be necessary.

Bridging Theory and Practice. The transition from theoretical models to practical applications requires a deep understanding of both the capabilities and limitations of these advanced models. This understanding is essential for designing systems that are not only efficient but also scalable and adaptable to real-world complexities [Li et al., 2024b, Li, 2024a,b].

B Incremental uncertainty estimation in Lemma 1

Before proving Lemma 1, we state the preliminary tools of sequential random projection for completeness, which is adapted form [Li, 2024a]. This tool was used to prove incremental posterior approximation argument of HyperAgent in tabular RL setup [Li et al., 2024b]. As the tool in [Li, 2024a] works only for the scalar process, we need additional technical innovations to deal with high-dimensional vector process. We make a novel utilization of this tool in the linear function approximation setting for the first time, by a non-trivial discretization argument in Appendix B.2 and a reduction to the tool of sequential random projection in Appendix B.2.

B.1 Probability tools for sequential random projection

We define some important concept that would be useful in the analysis. Let $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t\in\mathbb{N}}, \mathbb{P})$ be a complete filtered probability space. We first consider the measurable properties within the filtered probability space.

Definition 5 (Adapted process)**.** *For an index set I of the form $\{t \in \mathbb{N} : t \geq t_0\}$ for some $t_0 \in \mathbb{N}$, we say a stochastic process $(X_t)_{t\in I}$ is adapted to the filtration $(\mathcal{F}_t)_{t\in I}$ if each X_t is $\mathcal{F}_t$-measurable.*

Definition 6 ((Conditionally) σ-sub-Gaussian)**.** *A random variable $X \in \mathbb{R}$ is σ-sub-Gaussian if*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right), \qquad \forall \lambda \in \mathbb{R}.$$

Let $(X_t)_{t\geq 1} \subset \mathbb{R}$ be a stochastic process adapted to filtration $(\mathcal{F}_t)_{t\geq 1}$. Let $\sigma = (\sigma_t)_{t\geq 0}$ be a stochastic process adapted to filtration $(\mathcal{F}_t)_{t\geq 0}$. We say the process is $(X_t)_{t\geq 1}$ is conditionally

σ-sub-Gaussian if

$$\mathbb{E}[\exp(\lambda X_t) \mid \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2 \sigma_{t-1}^2}{2}\right), \quad a.s. \quad \forall \lambda \in \mathbb{R}.$$

Specifically for the index $t+1$, we can say X_{t+1} is ($\mathcal{F}_t$-conditionally) σ_t-sub-Gaussian. If σ_t is a constant σ for all $t \geq 0$, then we just say (conditionally) σ-sub-Gaussian.

For a random vector $X \in \mathbb{R}^M$ or vector process $(X_t)_{t\geq 1} \subset \mathbb{R}^M$ in high-dimension, we say it is σ-sub-Gaussian is for every fixed $v \in \mathbb{S}^{M-1}$ if the random variable $\langle v, X\rangle$, or the scalarized process $(\langle v, X_t\rangle)_{t\geq 1}$ is σ-sub-Gaussian.

Definition 7 (Almost sure unit-norm)**.** *We say a random variable X is almost sure unit-norm if $\|X\|_2 = 1$ almost surely.*

Remark 5. *When talking about the perturbation distribution P_z, we scale all specific distribution discussed in Appendix E by $\sqrt{\frac{1}{M}}$. Then the spherical distribution $\mathcal{U}(\mathbb{S}^{M-1})$ and uniform over scaled cube $\mathcal{U}(\frac{1}{\sqrt{M}}\{1,-1\}^M)$ satisfy the sub-Gaussian condition in Definition 6 with parameter $\sigma = \frac{1}{\sqrt{M}}$ and also satisfy the unit-norm condition in Definition 7 according to the discussion in Appendix E.*

Additionally, we characterize the boundedness on the stochastic processes.

Definition 8 (Square-bounded process)**.** *For an index set I of the form $\{t \in \mathbb{N} : t \geq t_0\}$ for some $t_0 \in \mathbb{N}$, the stochastic process $(X_t)_{t\in I}$ is c-square-bounded if $X_t^2 \leq c$ almost surely for all $t \in I$.*

Now, we are ready to state the important tool that is fundamental to our analysis.

Theorem 3 (Sequential random projection in adaptive process [Li, 2024a])**.** *Let $\varepsilon \in (0,1)$ be fixed and $(\mathcal{F}_t)_{t\geq 0}$ be a filtration. Let $\mathbf{z}_0 \in \mathbb{R}^M$ be an $\mathcal{F}_0$-measurable random vector satisfies $\mathbb{E}[\|\mathbf{z}_0\|^2] = 1$ and $|\|\mathbf{z}_0\|^2 - 1| \leq (\varepsilon/2)$. Let $(\mathbf{z}_t)_{t\geq 1} \subset \mathbb{R}^M$ be a stochastic process adapted to filtration $(\mathcal{F}_t)_{t\geq 1}$ such that it is $\sqrt{c_0/M}$-sub-Gaussian and each $\mathbf{z}_t$ is unit-norm. Let $(x_t)_{t\geq 1} \subset \mathbb{R}$ be a stochastic process adapted to filtration $(\mathcal{F}_{t-1})_{t\geq 1}$ such that it is c_x-square-bounded. Here, c_0 and c_x are absolute constants. For any fixed $x_0 \in \mathbb{R}$, if the following condition is satisfied*

$$M \geq \frac{16c_0(1+\varepsilon)}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right) + \log\left(1 + \frac{c_x T}{x_0^2}\right)\right), \tag{9}$$

we have, with probability at least $1-\delta$

$$\forall t \in \mathcal{T}, \quad (1-\varepsilon)\left(\sum_{i=0}^{t} x_i^2\right) \leq \|\sum_{i=0}^{t} x_i \mathbf{z}_i\|^2 \leq (1+\varepsilon)\left(\sum_{i=0}^{t} x_i^2\right). \tag{10}$$

B.2 Reduce Lemma 1 to sequential random projection

Without loss of generality, let us consider the set $\mathbb{S}^{d-1}$. First, we define a fine-grained good event for desired approximation error $\varepsilon \in [0,1]$: the approximate posterior variance is ε-close to the true posterior variance for action a at time $t \in \mathcal{T} := \{0, 1, \ldots, T\}$.

$$\mathcal{G}_t(a,\varepsilon) = \left\{|a^\top \mathbf{A}_t \mathbf{A}_t^\top a - a^\top \boldsymbol{\Sigma}_t a| \leq \varepsilon a^\top \boldsymbol{\Sigma}_t a\right\}, \tag{11}$$

and corresponding joint event over the set $\mathbb{S}^{d-1}$,

$$\mathcal{G}_t(\varepsilon) = \cap_{a\in\mathbb{S}^{d-1}} \mathcal{G}_t(a,\varepsilon). \tag{12}$$

The good event defined in Lemma 1 is indeed $\mathcal{G}_t(1/2)$.

A reduction. To fully utilize the probability tool for sequential random projection in Theorem 3, we make use of the following reduction from vector process to scalar process. For a fixed $a \in \mathbb{S}^{d-1}$, we let $\mathbf{s}(a) = a^\top \boldsymbol{\Sigma}_0^{-1/2}\mathbf{Z}_0, s(a)^2 = a^\top \boldsymbol{\Sigma}_0^{-1} a$. Further define short notation $\mathbf{z}_0 := \mathbf{s}(a)/s(a)$ and $x_0 := s(a)$. and $x_t = a^\top \phi(A_t)$ for all $t \in [T]$, then we can relate the incremental update in Proposition 1

$$a^\top \boldsymbol{\Sigma}_t^{-1}\mathbf{A}_t = \underbrace{a^\top \boldsymbol{\Sigma}_0^{-1/2}\mathbf{Z}_0}_{\mathbf{s}(a)=\mathbf{z}_0 x_0} + \sum_{i=1}^{t} \underbrace{a^\top \phi(A_t)}_{x_i} \mathbf{z}_t^\top, \quad a^\top \boldsymbol{\Sigma}_t^{-1} a = \underbrace{a^\top \boldsymbol{\Sigma}_0^{-1} a}_{x_0^2} + \sum_{i=1}^{t} \underbrace{a^\top \phi(A_t)\phi(A_t) a}_{x_i^2}$$

to the scalar sequence $(x_t)_{t\geq 0}$ and the vector sequence $(\mathbf{z}_t)_{t\geq 0}$ that would be applied in Theorem 3.

Recall that $\mathcal{H}_t$ the σ-algebra generated from history $(\mathcal{A}_1, A_1, Y_1, \ldots, \mathcal{A}_{t-1}, A_{t-1}, Y_{t-1}, \mathcal{A}_t)$. Denote $\mathcal{Z}_1 = \sigma(\mathbf{Z}_0)$ and $\mathcal{Z}_t = \sigma(\mathbf{Z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{t-1})$ for $t \geq 2$. We observe the following statistical relationship, which is further demonstrated in Fig. 8

- $\mathbf{z}_t \perp\!\!\!\perp (\mathcal{H}_t, A_t, \mathcal{Z}_t)$, $\mathbf{x}_t$ is dependent on $\mathcal{H}_t, \mathcal{Z}_t$,
- $\mathbf{A}_{t-1} \in \sigma(\mathcal{H}_t, \mathcal{Z}_t)$,
- $\mu_{t-1}, \boldsymbol{\Sigma}_{t-1} \in \mathcal{H}_t$.

For all $t \geq \mathbb{N}$, let us define the sigma-algebra $\mathcal{F}_t = \sigma(\mathcal{H}_{t+1}, \mathcal{Z}_{t+1}, A_{t+1})$. We can verify $\mathcal{F}_k \subseteq \mathcal{F}_l$ for all $k \leq l$. Thus $\mathbb{F} = (\mathcal{F}_t)_{t\in\mathbb{N}}$ is a filtration. Now, we could verify $(\mathbf{z}_t)_{t\geq 0}$ is adapted to $(\mathcal{F}_t)_{t\geq 0}$ and $(x_t)_{t\geq 1}$ is adapted to $(\mathcal{F}_t)_{t\geq 0}$, satisfying the conditions in Theorem 3.

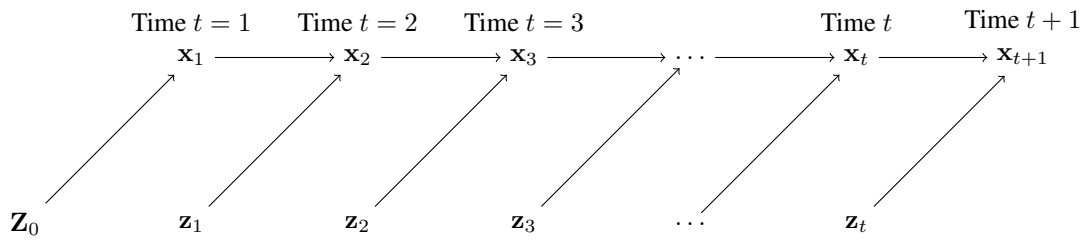


Figure 8: Sequential Dependence Structure in Index Sampling.

Prior approximation. First, we state a standard covering argument on sphere.

Lemma 2 (Covering number of a sphere)**.** *There exists a set $\mathcal{C}_\iota \subset \mathbb{S}^{d-1}$ with $|\mathcal{C}_\iota| \leq (1 + 2/\iota)^d$ such that for all $x \in \mathbb{S}^{d-1}$ there exists a $y \in \mathcal{C}_\iota$ with $\|x - y\|_2 \leq \iota$.*

Lemma 3 (Computing spectral norm on a covering set)**.** *Let $\boldsymbol{A}$ be a symmetric $d \times d$ matrix, and let $\mathcal{C}_\iota$ be the an ι-covering of $\mathbb{S}^{d-1}$ for some $\iota \in (0, 1)$. Then,*

$$\|\boldsymbol{A}\| = \sup_{x\in\mathbb{S}^{d-1}} |x^\top \boldsymbol{A} x| \leq (1 - 2\iota)^{-1} \sup_{x\in\mathcal{C}_\iota} |x^\top \boldsymbol{A} x|.$$

For compact set $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$, by standard covering argument in Lemma 3 and the distributional Johnson-Lindenstrauss lemma [Li, 2024b], when $M \geq 256\varepsilon^{-2}(d \log 9 + \log(2/\delta))$, the initial good event for prior approximation $G_0(\varepsilon/2)$ holds with probability $1 - \delta$.

Next, we are going to show that, under the event $G_0(\varepsilon/2)$, the initial condition on $|\|\mathbf{z}_0\|^2 - 1| \leq (\varepsilon/2)$ in Theorem 3 is satisfied. That is, under the event $G_0(\varepsilon/2)$

$$\begin{aligned}
& (1 + \varepsilon/2) a^\top \boldsymbol{\Sigma}_0 a \leq \|a^\top \boldsymbol{\Sigma}_0^{1/2} \mathbf{Z}_0\|^2 \leq (1 + \varepsilon/2) a^\top \boldsymbol{\Sigma}_0 a, \quad \forall a \in \mathbb{S}^{d-1} \\
\Leftrightarrow \quad & \|\mathbf{Z}_0 \mathbf{Z}_0^\top - \boldsymbol{I}\| \leq \varepsilon/2 \\
\Leftrightarrow \quad & (1 + \varepsilon/2) a^\top \boldsymbol{\Sigma}_0^{-1} a \leq \|a^\top \boldsymbol{\Sigma}_0^{-1/2} \mathbf{Z}_0\|^2 \leq (1 + \varepsilon/2) a^\top \boldsymbol{\Sigma}_0^{-1} a, \quad \forall a \in \mathbb{S}^{d-1}.
\end{aligned}$$

Recall the short notation $\mathbf{s}(a) = a^\top \boldsymbol{\Sigma}_0^{-1/2} \mathbf{Z}_0$ and $s(a)^2 = a^\top \boldsymbol{\Sigma}_0^{-1} a$, we have $\mathbf{z}_0 = \mathbf{s}(a)/s(a)$ satisfying $|\|\mathbf{z}_0\|^2 - 1| \leq (\varepsilon/2)$.

Posterior approximation. Notice that $x_0^2 = a^\top \boldsymbol{\Sigma}_0 a \geq \inf_{a\in\mathbb{S}^{d-1}} a^\top \boldsymbol{\Sigma}_0^{-1} a = s_{\min}^2$. As by the definition of the feature map $\phi(\cdot)$ in example 2, we can examine that $x_i^2 = (a^\top \phi(A_t))^2 \leq 1$, that is, the sequence $(a^\top \phi(A_t))_{t\geq 1}$ is 1-square-bounded for any $a \in \mathbb{S}^{d-1}$.

We could also check that $(\mathbf{z}_t)_{t\geq 1}$ is $1/\sqrt{M}$-sub-Gaussian and with unit-norm when the perturbation distribution P_z is Cube $\mathcal{U}(\{1, -1\}^M)$ or Sphere $\mathcal{U}(\mathbb{S}^{M-1})$.

Under the prior approximation event $\mathcal{G}_0(\varepsilon/2)$, we apply Theorem 3 to show that for any fixed $a \in \mathbb{S}^{d-1}$,

$$\forall t \in \mathcal{T}, E_t(a, \varepsilon) := \left\{ |a^\top \boldsymbol{\Sigma}_t^{-1} \mathbf{A}_t \mathbf{A}_t^\top \boldsymbol{\Sigma}_t^{-1} a - a^\top \boldsymbol{\Sigma}_t^{-1} a| \leq \varepsilon a^\top \boldsymbol{\Sigma}_t^{-1} a \right\} \tag{13}$$

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

$$M \geq \frac{16(1+\varepsilon)}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right)+\log\left(1+\frac{T}{s_{\min}^2}\right)\right). \tag{14}$$

We need discretization (covering) argument to relate the result in Eq. (13) to the desired good event defined in Eq. (12)

$$\mathcal{G}_t(\varepsilon)=\left\{\|\boldsymbol{\Sigma}_t^{-1/2}\mathbf{A}_t\mathbf{A}_t^\top\boldsymbol{\Sigma}_t^{-1/2}-\boldsymbol{I}\|\leq\varepsilon\right\}.$$

Standard discretization produces unacceptable results. Utilizing standard discretization for computing spectral norm in Lemma 3, let $\iota=1/4$, we can show that

$$\cap_{a\in\mathcal{C}_{1/4}}E_t(a,\varepsilon/2T)\subseteq\mathcal{G}_t(\varepsilon).$$

This is due to,

$$\begin{aligned}\|\boldsymbol{\Sigma}_t^{-1/2}\mathbf{A}_t\mathbf{A}_t^\top\boldsymbol{\Sigma}_t^{-1/2}-\boldsymbol{I}\| &= \sup_{x\in\mathbb{S}^{d-1}}\frac{|x^\top(\boldsymbol{\Sigma}_t^{-1}\mathbf{A}_t\mathbf{A}_t^\top\boldsymbol{\Sigma}_t^{-1}-\boldsymbol{\Sigma}_t^{-1})x|}{x^\top\boldsymbol{\Sigma}_t^{-1}x}\\ &\leq\frac{2}{\lambda_{\min}(\boldsymbol{\Sigma}_t^{-1})}\sup_{a\in\mathcal{C}_{1/4}}|a^\top(\boldsymbol{\Sigma}_t^{-1}\mathbf{A}_t\mathbf{A}_t^\top\boldsymbol{\Sigma}_t^{-1}-\boldsymbol{\Sigma}_t^{-1})a|\\ &\leq 2\varepsilon'\frac{\sup_{a\in\mathcal{C}_{1/4}}a^\top\boldsymbol{\Sigma}_t^{-1}a}{\lambda_{\min}(\boldsymbol{\Sigma}_t^{-1})}\leq 2\varepsilon'\cdot\kappa(\boldsymbol{\Sigma}_t^{-1})\leq 2T\varepsilon'.\end{aligned}$$

Then by union bound over $\mathcal{C}_{1/4}$, plugging in $\varepsilon/2T$ to Eq. (14), we require $M\geq\tilde{O}(dT^2\log T)$ to let $\cap_{a\in\mathcal{C}_{1/4}}E_t(a,\varepsilon/2T)$ hold with probability $1-\delta$. This result is not acceptable as the per-step computation complexity is growing unbounded polynimally with the interaction steps T. In the next section, we provide a non-trivial discretization to resolve this analytical problem.

Variance-aware discretization. The key contribution here is that we choose a variance weighted norm to measure discretization error. This variance-awareness, together with specific choice on a $O(1/\sqrt{T})$-discretization error and a constant approximation error ε, eventually arrives at $O(d\log T)$ log covering number and $M=\tilde{O}(d\log T)$ in Lemma 1.

Let $\mathbf{S}_t=\boldsymbol{\Sigma}_t^{-1}\mathbf{A}_t=\mathbf{X}_t^\top\mathbf{Z}_t$ and $\boldsymbol{\Gamma}_t=\boldsymbol{\Sigma}_t^{1/2}\mathbf{S}_t=\boldsymbol{\Sigma}_t^{-1/2}\mathbf{A}_t$. Notice that, from Eq. (13), the event holds with probability at least $1-\delta'$

$$\forall t\in\mathcal{T},E_t(a,\varepsilon')=\left\{\frac{|a^\top\mathbf{S}_t\mathbf{S}_t^\top a-a^\top\boldsymbol{\Sigma}_t^{-1}a|}{a^\top\boldsymbol{\Sigma}_t^{-1}a}\leq\varepsilon'\right\}$$

when

$$M\geq\frac{16(1+\varepsilon')}{(\varepsilon')^2}\left(\log\left(\frac{1}{\delta'}\right)+\log\left(1+\frac{T}{s_{\min}^2}\right)\right).$$

Let $\mathcal{C}_\iota\subset\mathbb{S}^{d-1}$ be the ι-covering set in Lemma 2 and the event $\cap_{a\in\mathcal{C}_\iota}E_t(a,\varepsilon')$ holds. Let $x\in\mathbb{S}^{d-1}$ and $y\in\mathcal{C}_\iota$ such that $\|x-y\|\leq\iota$. Define short notation $u=\boldsymbol{\Sigma}_t^{-1/2}x$, $v=\boldsymbol{\Sigma}_t^{-1/2}y$.

$$\begin{aligned}&\frac{|x^\top\mathbf{S}_t\mathbf{S}_t^\top x-x^\top\boldsymbol{\Sigma}_t^{-1}x|}{x^\top\boldsymbol{\Sigma}_t^{-1}x}-\frac{|y^\top\mathbf{S}_t\mathbf{S}_t^\top y-y^\top\boldsymbol{\Sigma}_t^{-1}y|}{y^\top\boldsymbol{\Sigma}_t^{-1}y}\\ &=\frac{|u^\top\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t^\top u-u^\top u|}{u^\top u}-\frac{|v^\top\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t^\top v-v^\top v|}{v^\top v}=\frac{|\|\boldsymbol{\Gamma}_t u\|^2-\|u\|^2|}{\|u\|^2}-\frac{|\|\boldsymbol{\Gamma}_t v\|^2-\|v\|^2|}{\|v\|^2}\\ &\leq\left|\frac{\|\boldsymbol{\Gamma}_t u\|^2}{\|u\|^2}-\frac{\|\boldsymbol{\Gamma}_t v\|^2}{\|v\|^2}\right|=\underbrace{\left|\frac{\|\boldsymbol{\Gamma}_t u\|^2-\|\boldsymbol{\Gamma}_t v\|^2}{\|u\|^2}\right|}_{(I)}+\underbrace{\|\boldsymbol{\Gamma}_t v\|^2\left|\frac{1}{\|u\|^2}-\frac{1}{\|v\|^2}\right|}_{(II)}.\end{aligned}$$

We bound (I) and (II) separately. W.L.O.G, assume $\|u\|\geq\|v\|$. Recall $s_{\max}^2\geq a^\top\boldsymbol{\Sigma}_0^{-1}a\geq s_{\min}^2$ for all $a\in\mathbb{S}^{d-1}$. Since $\|u\|=x^\top\boldsymbol{\Sigma}_t^{-1}x=x^\top(\boldsymbol{\Sigma}_0^{-1}+\sum_{s=1}^t\mathbf{x}_s\mathbf{x}_s^\top)x$, we have $s_{\min}^2\leq\|u\|\leq$

$s_{\max}^2 + t$. For (I), we have

$$(I) \leq \frac{(\|\mathbf{\Gamma}_t u\| - \|\mathbf{\Gamma}_t v\|)(\|\mathbf{\Gamma}_t u\| + \|\mathbf{\Gamma}_t v\|)}{\|u\|^2} \leq \frac{\|\mathbf{\Gamma}_t(u-v)\|}{s_{\min}} \left(\frac{\|\mathbf{\Gamma}_t u\|}{\|u\|} + \frac{\|\mathbf{\Gamma}_t v\|}{\|v\|} \right)$$
$$\leq \frac{\|\mathbf{\Gamma}_t\| \|u-v\|}{s_{\min}} (2\|\mathbf{\Gamma}_t\|) \leq \frac{2\|\mathbf{\Gamma}_t\|^2 \|\mathbf{\Sigma}_t^{-1/2}\| \iota}{s_{\min}} \leq \frac{2\|\mathbf{\Gamma}_t\|^2 \iota \sqrt{s_{\max}^2 + t}}{s_{\min}}.$$

For (II), we have

$$(II) \leq \frac{\|\mathbf{\Gamma}_t v\|^2}{\|v\|^2} \frac{\|u\|^2 - \|v\|^2}{\|u\|^2} \leq \|\mathbf{\Gamma}_t\|^2 \frac{\|u\|^2 - \|v\|^2}{\|u\|^2} \leq \|\mathbf{\Gamma}_t\|^2 \frac{(\|u\| - \|v\|)(\|u\| + \|v\|)}{\|u\|^2}$$
$$\leq \frac{2\|\mathbf{\Gamma}_t\|^2 \|u-v\|}{s_{\min}} \leq \frac{2\|\mathbf{\Gamma}_t\|^2 \|\mathbf{\Sigma}_t^{-1/2}\| \iota}{s_{\min}} \leq \frac{2\|\mathbf{\Gamma}_t\|^2 \iota \sqrt{s_{\max}^2 + t}}{s_{\min}}.$$

Then, putting (I) and (II) together, by the variance-aware discretization argument, we have the spectral norm

$$\begin{aligned}
\|\mathbf{\Sigma}_t^{-1/2} \mathbf{A}_t \mathbf{A}_t^\top \mathbf{\Sigma}_t^{-1/2} - \boldsymbol{I}\| &= \sup_{x \in \mathbb{S}^{d-1}} \frac{|x^\top (\mathbf{\Sigma}_t^{-1} \mathbf{A}_t \mathbf{A}_t^\top \mathbf{\Sigma}_t^{-1} - \mathbf{\Sigma}_t^{-1}) x|}{x^\top \mathbf{\Sigma}_t^{-1} x} \\
&\leq \frac{4\|\mathbf{\Gamma}_t\|^2 \iota \sqrt{s_{\max}^2 + t}}{s_{\min}} + \sup_{y \in \mathcal{C}_\iota} \frac{|y^\top (\mathbf{\Sigma}_t^{-1} \mathbf{A}_t \mathbf{A}_t^\top \mathbf{\Sigma}_t^{-1} - \mathbf{\Sigma}_t^{-1}) y|}{y^\top \mathbf{\Sigma}_t^{-1} y} \\
&\leq \frac{4\|\mathbf{\Gamma}_t\|^2 \iota \sqrt{s_{\max}^2 + t}}{s_{\min}} + \varepsilon'. \qquad (15)
\end{aligned}$$

Let

$$\iota = \frac{\alpha s_{\min}}{4\sqrt{s_{\max}^2 + T}},$$

where α to be determined. Equivalent formulation of the norm is $\|\mathbf{\Gamma}_t\|^2 = \lambda_{\max}(\mathbf{\Gamma}_t \mathbf{\Gamma}_t^\top)$ and

$$\|\mathbf{\Sigma}_t^{-1/2} \mathbf{A}_t \mathbf{A}_t^\top \mathbf{\Sigma}_t^{-1/2} - \boldsymbol{I}\| = \max\{\lambda_{\max}(\mathbf{\Gamma}_t \mathbf{\Gamma}_t^\top) - 1, 1 - \lambda_{\min}(\mathbf{\Gamma}_t \mathbf{\Gamma}_t^\top)\}.$$

Thus, we derive from Eq. (15),

$$\lambda_{\max}(\mathbf{\Gamma}_t \mathbf{\Gamma}_t^\top) \leq \frac{1+\varepsilon'}{1-\alpha}, \quad \lambda_{\min}(\mathbf{\Gamma}_t \mathbf{\Gamma}_t^\top) \geq 1 - \varepsilon' - \alpha \lambda_{\max}(\mathbf{\Gamma}_t \mathbf{\Gamma}_t^\top) \geq 1 - \varepsilon' - \frac{\alpha(1+\varepsilon')}{1-\alpha}.$$

Claim 1. *If* $\frac{1+\varepsilon'}{1-\alpha} = 1 + \varepsilon$ *and* $\varepsilon' + \frac{\alpha(1+\varepsilon')}{1-\alpha} = \varepsilon$*, then*

$$1 - \varepsilon \leq \lambda_{\min}(\mathbf{\Gamma}_t \mathbf{\Gamma}_t^\top) \leq \lambda_{\max}(\mathbf{\Gamma}_t \mathbf{\Gamma}_t^\top) \leq 1 + \varepsilon.$$

Let $\varepsilon = 1/2$, then $(\varepsilon', \alpha) = (1/4, 1/6)$ suffices for the Claim 1. That is to say the following configuration for discretization error ι suffices,

$$\iota = \frac{s_{\min}}{24\sqrt{s_{\max}^2 + T}}.$$

The covering number is $|\mathcal{C}_\iota| \leq (1 + 2/\iota)^d \leq (1 + (48/s_{\min})\sqrt{s_{\max}^2 + T})^d$.

Then, let $\delta' = \delta/(1 + (48/s_{\min})\sqrt{s_{\max}^2 + T})^d$ and by union bound, when

$$M \geq \frac{16(5/4)}{(1/4)^2} \left(d \log \left(\frac{1 + (48/s_{\min})\sqrt{s_{\max}^2 + T}}{\delta} \right) + \log \left(1 + \frac{T}{s_{\min}^2} \right) \right),$$

we have with probability $1 - \delta$, the event $\cap_{t \in \mathcal{T}} \mathcal{G}_t(1/2)$ holds.

P_ζ	Gaussian $N(0, I_M)$	Sphere $\sqrt{M}\mathcal{U}(\mathbb{S}^{M-1})$	Cube $\mathcal{U}(\{1,-1\}^M)$	Coord $\mathcal{U}(\{\pm e_i\}_{i\in[M]})$	Sparse
$\rho(P_\zeta)$	$\rho_1 \wedge \rho_3$	$\rho_2 \wedge \rho_3$	$\rho_2 \wedge \rho_3$	ρ_2	ρ_2
$p(P_\zeta)$	$\frac{1}{4\sqrt{e\pi}}$	$\frac{1}{2} - \frac{e^{1/12}}{\sqrt{2\pi}}$	$7/32$	$\frac{1}{2M}$	N/A

Table 3: (Restate of Table Table 1) The coefficient $\rho(P_\zeta)$ and $p(P_\zeta)$ related to reasonableness and optimism condition.

C Regret analysis

To make the proof easy to access, we restate the core results and a few notations that is needed for the proof of the propositions.

Adapting the results from [Abbasi-Yadkori et al., 2011b, Abeille and Lazaric, 2017], let $\beta_t = \sqrt{\lambda} + \sqrt{2\log(1/\delta) + \log\det(\boldsymbol{\Sigma}_{t-1}^{-1}/\lambda^d)}$. Under assumption 1, we define the confidence bound as

$$L_t(\cdot) = (-1) \vee (\langle \mu_{t-1}, \phi(\cdot)\rangle - \beta_t \|\phi(\cdot)\|_{\boldsymbol{\Sigma}_{t-1}}), U_t(\cdot) = 1 \wedge (\langle \mu_{t-1}, \phi(\cdot)\rangle + \beta_t \|\phi(\cdot)\|_{\boldsymbol{\Sigma}_{t-1}})$$

For the purpose of analysis within various reference distribution, we define a slightly inflated confidence bounds as

$$\begin{aligned} L_t(\cdot; P_\zeta) &= (\langle \mu_{t-1}, \phi(\cdot)\rangle - \beta_t \rho(P_\zeta) \|\phi(\cdot)\|_{\boldsymbol{\Sigma}_{t-1}}) \vee (-1), \\ U_t(\cdot; P_\zeta) &= (\langle \mu_{t-1}, \phi(\cdot)\rangle + \beta_t \rho(P_\zeta) \|\phi(\cdot)\|_{\boldsymbol{\Sigma}_{t-1}}) \wedge 1. \end{aligned}$$

$\rho(P_\zeta)$ is defined via $\rho_1 = O(\sqrt{M\log(M/\delta)})$, $\rho_2 = O(\sqrt{M})$, and $\rho_3 = O(\sqrt{\log(|\mathcal{A}|/\delta)})$ and Table 1. An immediate observation is that $[L_t(\cdot), U_t(\cdot)] \subset [L_t(\cdot; P_\zeta), U_t(\cdot; P_\zeta)]$. Thus, $L_t(\cdot; P_\zeta)$ and $U_t(\cdot; P_\zeta)]$ are also confidence bounds. We consider the the following functional form for HyperAgent under linear setup: for time t,

$$\tilde{f}_t(a) := f_{\theta_{t-1}}(a, \zeta_t) = \langle \phi(a), \beta_t \mathbf{A}_{t-1}\zeta_t + \mu_{t-1}\rangle, \quad \forall a \in \mathcal{A},$$

The condition on the propositions and theorem for regret analysis is when Eq. (9) is satisfied, that is when $M = O(d\log T)$, the Lemma 1 implies that with high probability, the good events $\mathcal{G} = \cap_{t\in\mathcal{T}}\mathcal{G}_t$ hold jointly, where

$$\mathcal{G}_t := \left\{ \frac{1}{2} x^\top \boldsymbol{\Sigma}_t x \le x^\top \mathbf{A}_t \mathbf{A}_t^\top x \le \frac{3}{2} x^\top \boldsymbol{\Sigma}_t x, \quad \forall x \in \mathbb{R}^d \right\}.$$

In the following section, we discuss the proof conditioned on the joint event $\mathcal{G}$ and also the confidence event that $f^*(a) \in [L_t(a), U_t(a)]$ for all $t \in [T]$ and $a \in \mathcal{A}$.

C.1 Proof of Proposition 2

Notice that from Eq. (6), we derive

$$\begin{aligned} |\tilde{f}_t(a) - \langle \mu_{t-1}, \phi(a)\rangle| &= |\langle \phi(a), \beta_t \mathbf{A}_{t-1}\zeta_t\rangle| \\ &= \beta_t \sqrt{\phi(a)^\top \mathbf{A}_{t-1}\mathbf{A}_{t-1}^\top \phi(a)} \left| \left\langle \frac{\phi(a)^\top \mathbf{A}_{t-1}}{\|\phi(a)^\top \mathbf{A}_{t-1}\|}, \zeta_t \right\rangle \right| \\ &\le (3/2)\beta_t \sqrt{\phi(a)^\top \boldsymbol{\Sigma}_{t-1}\phi(a)} \left| \left\langle \frac{\phi(a)^\top \mathbf{A}_{t-1}}{\|\phi(a)^\top \mathbf{A}_{t-1}\|}, \zeta_t \right\rangle \right|, \end{aligned}$$

where the last inequality is due to the good event $\mathcal{G}$. For compact action set, we use Cauchy–Schwarz inequality,

$$\left| \left\langle \frac{\phi(a)^\top \mathbf{A}_{t-1}}{\|\phi(a)^\top \mathbf{A}_{t-1}\|}, \zeta_t \right\rangle \right| \le \|\zeta_t\|.$$

Using the concentration properties of P_ζ in Appendix E to upper bound $\|\zeta_t\|$ yields part of the results. For finite action set $\mathcal{A}$, also taking the advantages of the concentration properties of several reference distributions P_ζ in Appendix E to bound the conditionally probability

$$\mathbb{P}\left(\left| \left\langle \frac{\phi(a)^\top \mathbf{A}_{t-1}}{\|\phi(a)^\top \mathbf{A}_{t-1}\|}, \zeta_t \right\rangle \right| \le \sqrt{\log \frac{2|\mathcal{A}|}{\delta}} \mid \mathcal{H}_t, \mathcal{Z}_t \right) \ge 1 - \delta,$$

as ξ_t is independent of the history $\mathcal{H}_t, \mathcal{Z}_t$. Finally, the inflated coefficient $\rho(P_\zeta)$ defined in Table 1 suffices to make $\tilde{f}_t(\cdot) \in [L_t(\cdot; P_\zeta), U_t(\cdot; P_\zeta)]$ reasonable.

C.2 Proof of Proposition 3

Let $A_t = \max_{a \in \mathcal{A}_t} \tilde{f}_t(a)$ and $A_t^* = \max_{a \in \mathcal{A}_t} f^*(a)$. Conditioned on $\mathcal{G}$ and confidence event,

$$\begin{aligned}
\tilde{f}_t(A_t) - f^*(A_t^*) \geq \tilde{f}_t(A_t^*) - f^*(A_t^*) &\geq \tilde{f}_t(A_t^*) - U^*(A_t^*) \\
&= \langle \phi(A_t^*), 2\beta_t \mathbf{A}_{t-1}\zeta_t \rangle - \beta_t \|A_t^*\|_{\boldsymbol{\Sigma}_{t-1}} \\
&= 2\beta_t \sqrt{\phi(A_t^*)^\top \mathbf{A}_{t-1}\mathbf{A}_{t-1}^\top \phi(A_t^*)} \left\langle \frac{\phi(A_t^*)^\top \mathbf{A}_{t-1}}{\|\phi(A_t^*)^\top \mathbf{A}_{t-1}\|}, \zeta_t \right\rangle - \beta_t \|\phi(A_t^*)\|_{\boldsymbol{\Sigma}_{t-1}} \\
&\geq \beta_t \|\phi(A_t^*)\|_{\boldsymbol{\Sigma}_{t-1}} \left(\left\langle \frac{\phi(A_t^*)^\top \mathbf{A}_{t-1}}{\|\phi(A_t^*)^\top \mathbf{A}_{t-1}\|}, \zeta_t \right\rangle - 1 \right).
\end{aligned}$$

We consider the conditional probability,

$$\begin{aligned}
\mathbb{P}(\tilde{f}_t(A_t) \geq f^*(A_t^*) \mid \mathcal{H}_t, \boldsymbol{Z}_t) &\geq \mathbb{P}\left(\beta_t \|\phi(A_t^*)\|_{\boldsymbol{\Sigma}_{t-1}} \left(\left\langle \frac{\phi(A_t^*)^\top \mathbf{A}_{t-1}}{\|\phi(A_t^*)^\top \mathbf{A}_{t-1}\|}, \zeta_t \right\rangle - 1 \right) \mid \mathcal{H}_t, \mathcal{Z}_t \right) \\
&= \mathbb{P}(\langle v, \zeta_t \rangle \geq 1), \qquad (16)
\end{aligned}$$

where v is a fixed unit vector in $\mathbb{R}^M$. The final probability bound in $Eq.$ (16) for each reference distribution P_ζ is essentially the anti-concentration bounds. Please find the anti-concentration results for each distribution in Appendix E, resulting in the Table 1.

C.3 Proof of Theorem 2

The Theorem 2 follows directly from Propositions 2 and 3 and Theorem 1. Additionally, it requires the Azuma's inequality for the sum of bounded martingale difference

$$\sum_{t \in [T]} \mathbb{E}[(U_t(A_t) - L_t(A_t)) \mid \mathcal{H}_t] - (U_t(A_t) - L_t(A_t)) \leq O(\sqrt{T \log(1/\delta)}), \quad w.p. \;\; 1 - \delta$$

as $U_t(\cdot) - L_t(\cdot) \leq 2$. Then, it suffices to bound the summation of width between upper and lower confidence bounds

$$\sum_{t \in [T]} (U_t(A_t) - L_t(A_t)).$$

Under linear bandit setups in example 2, we use the elliptical potential lemma (e.g. Lemma 19.4 in [Lattimore and Szepesvári, 2020] and [Abbasi-Yadkori et al., 2011a]) to bound this summation.

D Derivation of the closed-form incremental update

Before proving Proposition 1, we provide useful technical lemmas for isotropic distributions in Definition 4.

Lemma 4. *For any isotropic distribution P over M-dim vector space, we have for any fixed vector $a \in \mathbb{R}^M$ and $x \in \mathbb{R}^d$, $\mathbb{E}_{X \sim P}[a^\top X x X^\top] = \mathbb{E}_{X \sim P}[X^\top a x X^\top] = x a^\top$.*

Proof. The (i, j)-th entry of the matrix is

$$[\mathbb{E}_X[X^\top a x X^\top]]_{ij} = \mathbb{E}_X[[X^\top a x X^\top]_{ij}] = \mathbb{E}_X[(\sum_{k=1}^M a_k X_k) x_i X_j] = x_i \sum_{k=1}^M a_k \mathbb{E}_X[X_k X_j] = x_i a_j$$

□

Lemma 5. *For any isotropic distribution P over M-dim vector space, we have for any fixed matrix $A \in \mathbb{R}^{d \times M}$ and any fixed vector $x \in \mathbb{R}^d$, $\mathbb{E}_{X \sim P}[X^\top A^\top x X^\top] = x^\top A$ and symmetrically, $\mathbb{E}_{X \sim P}[X x^\top A X] = A^\top x$.*

Proof. Let $A = (a_1, \ldots, a_M)$ where $a_i \in \mathbb{R}^d$ for $i = 1, 2, \ldots, M$.

$$AX = \sum_{k=1}^{M} X_i a_i,$$

and

$$(AX)^\top x = \sum_{k=1}^{M} X_i a_i^\top x.$$

Note $\mathbb{E}_X[X^\top A^\top x X^\top] \in \mathbb{R}^{1\times M}$. Then, the j-th entry of the row vector is

$$[\mathbb{E}_X[X^\top A^\top x X^\top]]_j = \mathbb{E}_X[X_j \sum_{k=1}^{M} X_i a_i^\top x] = a_j^\top x = x^\top a_j$$

□

Lemma 6. *For any isotropic distribution P over M-dim vector space, we have for any fixed matrix $B \in \mathbb{R}^{M\times M}$ and any fixed vector $x \in \mathbb{R}^d$, $\mathbb{E}_{X\sim P}[X^\top B X] = \mathrm{tr} B$.*

Proof.

$$X^\top A X = \sum_{i,j} X_i X_j B_{ij}$$

By taking the expectation,

$$\mathbb{E}[X^\top A X] = \sum_i B_{ii} = \mathrm{tr} B$$

□

Remark 6. *For any distribution P over M-dim vector space such that $\mathbb{E}_{X\sim P}[X_i X_j] = \delta_{ij}$, we have for any fixed vector $A \in \mathbb{R}^{d\times M}$,*

$$\mathbb{E}_X[AXX^\top] = A$$

Definition 9 (Zero-mean)**.** *We say a distribution P over M-dim vector space is a zero-mean distribution if $\mathbb{E}_{X\sim P}[X] = 0$.*

D.1 Proof of Proposition 1

In this section, we will derive the closed-form incremental update if the update distribution P_ξ is zero-mean and isotropic. We use short notation x for the feature $\phi(A)$ of some action $A \in \mathcal{A}$.

Let the trainable parameters be $\theta = (\boldsymbol{A}, b)$ and the linear hypermodel be

$$\begin{aligned} f_\theta(x,\xi) &= \underbrace{\langle \mathbf{A}\xi + b, x\rangle}_{\textbf{Learnable } f_\theta^L(x,\xi)} + \underbrace{\langle \boldsymbol{\Sigma}_0^{1/2}\mathbf{Z}_0\xi + \mu_0, x\rangle}_{\textbf{Fixed prior } f^P(x,\xi)} \\ &= \langle (b+\mu_0), x\rangle + \langle (\boldsymbol{A} + \boldsymbol{\Sigma}_0^{1/2}\mathbf{Z}_0)\xi, x\rangle \end{aligned} \tag{17}$$

Let $\|\theta\|^2 = \|\boldsymbol{A}\|_F^2 + \|b\|_2^2$ in Eq. (2) and $\boldsymbol{\Sigma}_0 = \frac{1}{\lambda}\boldsymbol{I}$.

By writing done the gradient w.r.t. b and taking a look at the stationary point,

$$\begin{aligned} \frac{\partial L(\theta; D_t)}{\partial b} &= 2\mathbb{E}_{\xi\sim P_\xi}[\sum_{s=1}^{t}(\langle(\boldsymbol{A} + \boldsymbol{\Sigma}_0^{1/2}\mathbf{Z}_0)\xi + (b+\mu_0), x_s\rangle - y_s - \sigma \mathbf{z}^\top \xi)x_s] + \boldsymbol{\Sigma}_0^{-1} b \\ &= \sum_{s=1}^{t} x_s x_s^\top (b+\mu_0) - x_s y_s + \boldsymbol{\Sigma}_0^{-1} b \qquad \text{(if } P_\xi \text{ is zero-mean.)} \end{aligned}$$

Let $\tilde{b}_t$ be the stationary point. Definition 9 implies posterior mean matching.

$$\mu_t := (\tilde{b}_t + \mu_0) = (\sum_{s=1}^{t} x_s x_s^\top + \boldsymbol{\Sigma}_0^{-1})^{-1}(\sum_{s=1}^{t} x_t y_t + \boldsymbol{\Sigma}_0^{-1}\mu_0) = \boldsymbol{\Sigma}_t^{-1}(\sum_{s=1}^{t} x_t y_t + \boldsymbol{\Sigma}_0^{-1}\mu_0) \tag{18}$$

By writing done the gradient w.r.t. $\mathbf{A}$ and taking a look at the stationary point,

$$\begin{aligned}\frac{\partial L(\theta; D_t)}{\partial \boldsymbol{A}} &= \mathbb{E}_{\xi\sim P_\xi}[\sum_{s=1}^{t}(\langle(\boldsymbol{A}+\boldsymbol{\Sigma}_0^{1/2}\mathbf{Z}_0)\xi + (b+\mu_0), x_s\rangle - y_s - \sigma\mathbf{z}_s^\top\xi)x_s\xi^\top] + 2\boldsymbol{\Sigma}_0^{-1}\boldsymbol{A} \\ &= \sum_{s=1}^{t}(x_s x_s^\top(\boldsymbol{A}+\boldsymbol{\Sigma}_0^{1/2}\mathbf{Z}_0) - \sigma x_s\mathbf{z}_s^\top) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{A},\end{aligned}$$

where the last equality holds if P_ξ is isotropic and zero-mean and we use Lemmas 4 to 6. Let $\tilde{\boldsymbol{A}}_t$ be the stationary point. We have

$$\boldsymbol{A}_t := \tilde{\boldsymbol{A}}_t + \boldsymbol{\Sigma}_0^{1/2}\mathbf{Z}_0 = \boldsymbol{\Sigma}_t\left(\boldsymbol{\Sigma}_0^{-1/2}\mathbf{Z}_0 + \frac{1}{\sigma}\sum_{s=1}^{t}x_s\mathbf{z}_s^\top\right) \tag{19}$$

From the observation of Eqs. (18) and (19), the solution μ_t and $\mathbf{A}_t$ can be recursively updated from μ_{t-1} and $\mathbf{A}_{t-1}$ respectively as more data gathering in.

E Isotropy, Concentration and Anti-concentration

Isotropy property is used for update distribution and proving the Proposition 1. The sub-Gaussianness in concentration property is used for perturbation distributions and proving Lemma 1. The concentration and anti-concentration properties are used for reference distributions and discussion on the resonableness and optimism condition for HyperAgent. Let us discuss each distribution case by case.

E.1 Sphere $P_\zeta = \mathcal{U}(\sqrt{M}\mathbb{S}^{M-1})$

Isotropy. By the rotational invariance of sphere distribution, we know for any fixed orthogonal matrix Q,

$$\langle\zeta, x\rangle \sim \langle Q\zeta, x\rangle = \langle\zeta, Q^\top x\rangle, \quad \forall x\in\mathbb{R}^d.$$

Then, for any fixed x, we select M orthogonal matrix $Q_1,\dots,Q_M$ to rotate x such that $Q_i^\top x = \|x\|e_i$ where e_i is the i-th coordinate vector. With this construction, for any fixed x,

$$M\mathbb{E}[\langle\zeta,x\rangle^2] = \mathbb{E}[\sum_{i=1}^{M}\langle\zeta,x_i\rangle^2] = \mathbb{E}[\|x\|^2\sum_{i=1}^{M}\zeta_i^2] = M\|x\|^2$$

and hence $\mathbb{E}[\langle\zeta,x\rangle^2] = \|x\|^2$, which is the definition of isotropic random vector.

Concentration. By definition, $\|\zeta\| = \sqrt{M}$. For a random variable $\zeta\sim\mathcal{U}(\mathbb{S}^{M-1})$ and any fixed $v\in\mathbb{S}^{M-1}$, the inner product follows the transformed Beta distribution

$$\langle\zeta, v\rangle \sim 2\,\mathrm{Beta}(\frac{M-1}{2},\frac{M-1}{2}) - 1.$$

Evidenced by [Skorski, 2023, Li, 2024a], $P_\zeta = \mathcal{U}(\sqrt{M}\mathbb{S}^{M-1})$ is 1-sub-Gaussian. For finite action set $\mathcal{A}$, using the concentration of Beta random variables with union bound, we have

$$\mathbb{P}\left(\forall a\in\mathcal{A}, \langle\zeta,\phi(a)\rangle \le \|\phi(a)\|\sqrt{\log\frac{2|\mathcal{A}|}{\delta}}\right) \ge 1-\delta,$$

Anti-concentration. Let's start by rewriting the problem in terms of the incomplete Beta function:

Given:

$$X \sim \mathrm{Beta}\left(\frac{M-1}{2},\frac{M-1}{2}\right)$$

We want to find:

$$\mathrm{P}\left(\langle\zeta,v\rangle \ge 1\right) = \mathrm{P}\left(2X-1 > \frac{1}{\sqrt{M}}\right) = \mathrm{P}\left(X > \frac{1}{2} + \frac{1}{2\sqrt{M}}\right).$$

Theorem 4. *For all $d \geq 2$, the random variable $X \sim$ Beta $\left(\frac{d-1}{2}, \frac{d-1}{2}\right)$ has the following anti-concentration behavior*

$$P\left(X > \frac{1}{2} + \frac{1}{2\sqrt{d}}\right) \geq \frac{1}{2} - \frac{e^{1/12}}{\sqrt{2\pi}}.$$

Remark 7. *We did not find any literature that can help derive such anti-concentration results for Beta distribution.*

Proof. Using the incomplete Beta function $I_x(a,b)$, this probability can be expressed as:

$$\mathrm{P}\left(X > \frac{1}{2} + \frac{1}{2\sqrt{d}}\right) = 1 - I_{\left(\frac{1}{2}+\frac{1}{2\sqrt{d}}\right)}\left(\frac{d-1}{2}, \frac{d-1}{2}\right)$$

To compute $I_{\left(\frac{1}{2}+\frac{1}{2\sqrt{d}}\right)}\left(\frac{d-1}{2}, \frac{d-1}{2}\right)$, we will use the following relationship for the regularized incomplete Beta function $I_x(a,b)$:

$$I_x(a,b) = \frac{B(x;a,b)}{B(a,b)}$$

where $B(x;a,b)$ is the incomplete Beta function and $B(a,b) := B(1;a,b)$ is the complete Beta function.

For $a = b = \frac{d-1}{2}$, the complete Beta function is:

$$B\left(\frac{d-1}{2}, \frac{d-1}{2}\right) = \frac{\Gamma\left(\frac{d-1}{2}\right)\Gamma\left(\frac{d-1}{2}\right)}{\Gamma(d-1)}$$

Using the property of the Gamma function:

$$\Gamma(n+1) = n\Gamma(n).$$

Let's compute the incomplete Beta function for $x = \frac{1}{2} + \frac{1}{2\sqrt{d}}$ and $a = b = \frac{d-1}{2}$:

1. Calculate the incomplete Beta function $B\left(x; \frac{d-1}{2}, \frac{d-1}{2}\right)$:

$$B\left(\frac{1}{2} + \frac{1}{2\sqrt{d}}; \frac{d-1}{2}, \frac{d-1}{2}\right) = \int_0^{\frac{1}{2}+\frac{1}{2\sqrt{d}}} t^{\frac{d-3}{2}}(1-t)^{\frac{d-3}{2}}\, dt$$

As $f(t) = t^{\frac{d-3}{2}}(1-t)^{\frac{d-3}{2}}$ is symmetric at $t = 1/2$ in the interval $[0,1]$,

$$B\left(\frac{1}{2} + \frac{1}{2\sqrt{d}}; \frac{d-1}{2}, \frac{d-1}{2}\right) = \frac{1}{2}B\left(\frac{d-1}{2}, \frac{d-1}{2}\right) + \int_{\frac{1}{2}}^{\frac{1}{2}+\frac{1}{2\sqrt{d}}} t^{\frac{d-3}{2}}(1-t)^{\frac{d-3}{2}}\, dt.$$

2. Calculate the regularized incomplete Beta function $I_x(a,b)$:

$$I_{\left(\frac{1}{2}+\frac{1}{2\sqrt{d}}\right)}\left(\frac{d-1}{2}, \frac{d-1}{2}\right) = \frac{B\left(\frac{1}{2} + \frac{1}{2\sqrt{d}}; \frac{d-1}{2}, \frac{d-1}{2}\right)}{B\left(\frac{d-1}{2}, \frac{d-1}{2}\right)}$$

As the function $f(t) = t^{\frac{d-3}{2}}(1-t)^{\frac{d-3}{2}}$ achieves the maximum at $t = 1/2$, we could upper bound the incomplete Beta function by

$$\int_{\frac{1}{2}}^{\frac{1}{2}+\frac{1}{2\sqrt{d}}} t^{\frac{d-3}{2}}(1-t)^{\frac{d-3}{2}}\, dt \leq \left(\frac{1}{4}\right)^{\frac{d-3}{2}}\left(\frac{1}{2\sqrt{d}}\right) = \left(\frac{1}{2}\right)^{d-3}\left(\frac{1}{2\sqrt{d}}\right). \tag{20}$$

The complete Beta function can be expressed as

$$B\left(\frac{d-1}{2}, \frac{d-1}{2}\right) = \frac{\Gamma\left(\frac{d-1}{2}\right)\Gamma\left(\frac{d-1}{2}\right)}{\Gamma(d-1)},$$

where $\Gamma(\cdot)$ is the Gamma function. We use the Stirling's Approximation on Gamma function which could provide strict lower bound[Nemes, 2015]

$$\Gamma(z) \geq \sqrt{2\pi} z^{z-\frac{1}{2}} e^{-z},$$

and upper bound [Gronwall, 1918]

$$\Gamma(z) \leq \sqrt{2\pi} z^{z-\frac{1}{2}} e^{-z+\frac{1}{12z}}$$

for all $z > 0$. Immediately, the lower bound of the complete Beta function is

$$B\left(\frac{d-1}{2}, \frac{d-1}{2}\right) \geq \frac{\sqrt{2\pi}((d-1)/2)^{d-2} e^{-(d-1)}}{(d-1)^{d-\frac{3}{2}} e^{-d+1+\frac{1}{12(d-1)}}} = \sqrt{2\pi}\left(\frac{1}{2}\right)^{d-2} (d-1)^{-1/2} e^{-\frac{1}{12(d-1)}}.$$

As $e^{-\frac{1}{12(d-1)}} \geq e^{-1/12}$ whenever $d \geq 2$, we further lower bound

$$B\left(\frac{d-1}{2}, \frac{d-1}{2}\right) \geq \sqrt{2\pi} e^{-1/12}\left(\frac{1}{2}\right)^{d-2} \frac{1}{\sqrt{d}}. \tag{21}$$

Finally, combining Eqs. (20) and (21) yields

$$\begin{aligned} I_{\left(\frac{1}{2}+\frac{1}{2\sqrt{d}}\right)}\left(\frac{d-1}{2}, \frac{d-1}{2}\right) &\leq \frac{1}{2} + \frac{2e^{1/12}(\frac{1}{2\sqrt{d}})}{\sqrt{2\pi}\frac{1}{\sqrt{d}}} \\ &\leq \frac{1}{2} + \frac{e^{1/12}}{\sqrt{2\pi}}, \end{aligned}$$

and

$$P(X > \frac{1}{2} + \frac{1}{2\sqrt{d}}) \geq \frac{1}{2} - \frac{e^{1/12}}{\sqrt{2\pi}} \approx 0.0668.$$

□

E.2 Cube $P_\zeta = \mathcal{U}(\{1,-1\}^M)$

Isotropy. Easy to verify by definition.

Concentration. By definition, $\|\xi\| = \sqrt{M}$. Also notice that we could sample the random vector ζ by sample each entry independently from $\zeta_i \sim \mathcal{U}(\{1,-1\})$ for $i \in [M]$. Then, for any $v \in \mathbb{S}^{M-1}$, by independence,

$$\mathbb{E}[\exp(\lambda\langle v, \zeta\rangle)] = \prod_{i=1}^{m} \mathbb{E}[\exp(\lambda v_i \mathrm{z}_i)] \leq \prod_{i=1}^{m} \exp(\lambda^2 v_i^2) = \exp(\lambda^2 \sum_i v_i^2).$$

The inequality is due to MGF of rademacher distribution (e.g. Example 2.3 in [Wainwright, 2019]). Then we confirm that $P_\zeta = \mathcal{U}(\{1,-1\}^M)$ is 1-sub-Gaussian. For finite action set $\mathcal{A}$, we have from sub-Gaussian property

$$\mathbb{P}\left(\forall a \in \mathcal{A}, \langle \zeta, \phi(a)\rangle \leq \|\phi(a)\|\sqrt{\log\frac{2|\mathcal{A}|}{\delta}}\right) \geq 1-\delta.$$

Anti-concentration. Using the anti-concentration result from [Hollom and Portier, 2023], we have for any fixed unit vector v in $\mathbb{R}^M$

$$P(\langle \zeta, v\rangle) \geq 7/32 \approx 0.21875.$$

E.3 Gaussian $P_\zeta = N(0, I_M)$

Isotropy. Easy to verify by definition.

Concentration. The concentration property comes directly from the Chernoff bound for standard Gaussian random variable together with union bound argument. For any $\alpha > 0$, we have

$$\mathbb{P}(\|\zeta\| \leq \alpha\sqrt{M}) \geq \mathbb{P}\left(\forall 1 \leq i \leq M, |\zeta_i| \leq \alpha\right) \geq 1 - M\mathbb{P}\left(|\zeta_i| \geq \alpha\right).$$

Standard concentration inequality for Gaussian random variable gives, $\forall \alpha > 0$,

$$\mathbb{P}\left(|\zeta_i| \geq \alpha\right) \leq 2e^{-\alpha^2/2}.$$

Plugging everything together with $\alpha = \sqrt{2\log\frac{2M}{\delta}}$ gives the desired result, which is

$$\|\zeta\| \leq \sqrt{2M\log\frac{2M}{\delta}}, \quad w.p.\ 1-\delta.$$

For the case of finite action set $\mathcal{A}$,

$$\mathbb{P}\left(\forall a \in \mathcal{A}, \langle \zeta, \phi(a)\rangle \leq \|\phi(a)\|\sqrt{\log\frac{2|\mathcal{A}|}{\delta}}\right) \geq 1-\delta.$$

Anti-concentration. Here $\langle \zeta, v\rangle \sim N(0,1)$ for for any fixed unit vector v in $\mathbb{R}^M$.

$$P(N(0,1) \geq 1) = \frac{1}{2}\operatorname{erfc}\left(\frac{1}{\sqrt{2}}\right) \geq \frac{1}{4\sqrt{e\pi}} \approx 0.0856$$

E.4 Coord $P_\zeta = \mathcal{U}(\sqrt{M}\{\pm e_1, \ldots, \pm e_M\})$

Isotropy. Easy to verify by definition.

Concentration. By definition, $\|\zeta\| = \sqrt{M}$.

Anti-concentration.

$$P(\langle \zeta, v\rangle \geq 1) = \frac{1}{2M}\sum_{j\in[M]}(\mathbb{1}_{v_j \geq \frac{1}{\sqrt{M}}} + \mathbb{1}_{-v_j \geq \frac{1}{\sqrt{M}}}) = \frac{1}{2M}\sum_{j\in[M]}(\mathbb{1}_{|v_j| \geq \frac{1}{\sqrt{M}}}) \geq \frac{1}{2M},$$

where the last inequality is due to a simple fact that for any fixed $v \in \mathbb{R}^M$ with unit norm $\|v\| = 1$, there always exists an entry $j \in [M]$ with $|v_j| \geq \frac{1}{\sqrt{M}}$.

E.5 Sparse distribution P_ζ

Definition 10 (s-sparse distribution)**.** *The sparse vector is in the form $\zeta = \sqrt{\frac{M}{s}}\eta \odot \omega$ where $P_\omega := \mathcal{U}(\{1,-1\}^M)$, and η is independently and uniformly sampled from all possible s-hot vectors, where s-hot vectors is with exactly s non-zero entries with number 1. This construction is introduced by [Kane and Nelson, 2014].*

Isotropy. By definition,

$$\mathbb{E}[\zeta_j\zeta_k] = \frac{M}{s}\mathbb{E}[\eta_j\eta_k]\mathbb{E}[\omega_j\omega_k] = \frac{M}{s}\delta_{jk}\mathbb{E}[\omega_j] = \delta_{hk}. \tag{22}$$

Therefore, the sparse distribution in Definition 10 is indeed isotropic distribution.

Concentration. $\|\zeta\| = \sqrt{M}$.

Anti-concentration. Not clear.

F In-depth empirical and ablation studies

In this section, we dive into the intricacies of each evaluation testbed. Through a comprehensive set of empirical results, we'll further illuminate the benefits afforded by HyperAgent. All experiments are conducted on P40 GPUs to maintain processing standardization.

F.1 Linear Bandit Task

We consider the linear bandit tasks to study the impact pf perturbation and reference distribution indicated by the theoretical analysis.

Task Settings: We use the action feature set $\mathcal{X}$ to denote the set of features $\phi(a) : a \in \mathcal{A}$ induced by action set $\mathcal{A}$ and feature mapping $\phi(\cdot)$. We build two linear bandit tasks with different action distribution as follow:

- **Finite-action Linear Bandit**: We construct the finite set $\mathcal{X}$ by uniformly sampling a set of action features from the range $[-1/\sqrt{5}, 1/\sqrt{5}]^d$ where d is the ambient dimension of the linear reward function. This task builds upon prior research Russo and Van Roy [2018]. We vary the action size $|\mathcal{X}|$ over a set of $\{100, 1000, 10000\}$, and the ambient dimension across $\{10, 50\}$.
- **Compact-action Linear Bandit**: Let the action feature set $\mathcal{X} = \mathbb{S}^{d-1}$ be the unit sphere. In this task, we vary the ambient dimension d over a set of $\{10, 50, 100\}$.

In both tasks, the reward of each feature $X_t \in \mathbb{R}^d$ is computed as $r_t = X_t^\top \theta + \epsilon$, where $\theta \sim \mathcal{N}(0, 10\mathrm{I})$ is drawn from the multivariate Gaussian prior distribution, and $\epsilon \sim \mathcal{N}(0, 1)$ is an independent additive Gaussian noise term. At every step t, only the reward from the chosen feature X_t is discernible. To ensure robust results, each experiment is executed a total of 1000 time steps and repeated 200 times.

Analysis of Results: We investigated all 25 combinations of perturbation and reference distribution under different scales of the linear bandit tasks and numerous index dimensions denoted as M. As depicted in Figs. 9 to 11, the outcomes across diverse problem scales corroborate each other. The use of a Gaussian reference distribution significantly enhances performance when the index dimension M is relatively small, such as when M is 2 or 4. As the index dimension M grows, all combinations show an analogous performance under varying problem scales. However, it is worth noting that for extremely large index dimensions, such as 512 or 1024, combinations involving the Coord perturbation and Coord reference distribution significantly underperform compared to other combinations. Given that Coord distributions are used in the ensemble-based methods, the results prompt a compelling argument. HyperAgent equipped with a continuous reference distribution presents a superior performance, suggesting its potential for surpassing traditional ensemble-based methods. These findings strongly support the superior advantage of our index sampling method, validating our theoretical analysis.

Analysis of Computational Efficiency: We delve deeper into the effects of varying the index dimension M within the HyperAgent. We assessed its performance across distinctive combinations of perturbation and reference distributions over an assortment of index dimensions $M \in \{4, 8, 16, 32, 64, 128, 256, 512, 1024\}$. The outcomes, visualized in Figs. 12 and 13, are congruent with findings illustrated in Figs. 9 to 11. We observed that for large index dimensions M, the Coord perturbation and Coord reference distributions degrade the performance. This implies that the index sampling method employed by ensemble-based methods lacks efficiency. When HyperAgent utilizes Gaussian or Sphere reference distributions, it achieves satisfactory performance, comparable with Thompson sampling with small M. These results significantly showcase the computational efficiency of HyperAgent, which attains satisfactory performance with small index dimensions M.

Remark 8 (Limitation of Theorem 2.)**.** *Notice that Theorem 2 suggest that when $M \geq O(d \log T)$, the regret bound of Ensemble sampling would increase with factor $M^{3/2}$, which contradicts with our empirical evidence in Figs. 9 to 13.*

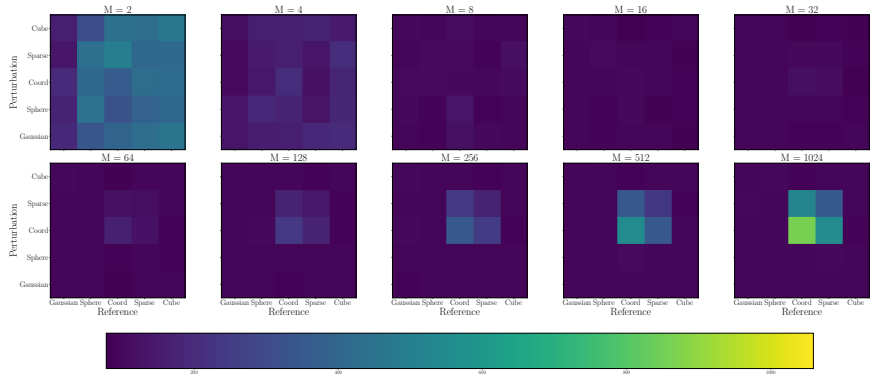
Remark 9 (Good prediction of Theorem 2.)**.** *Our empirical evidence in Figs. 9 to 13 confirms the Theorem 2 in finite decision set setting for continuous-support reference distributions: when M is larger then a threshold $O(d \log T)$, the regret has no dependence on M.*

F.2 Nonlinear Bandit Task

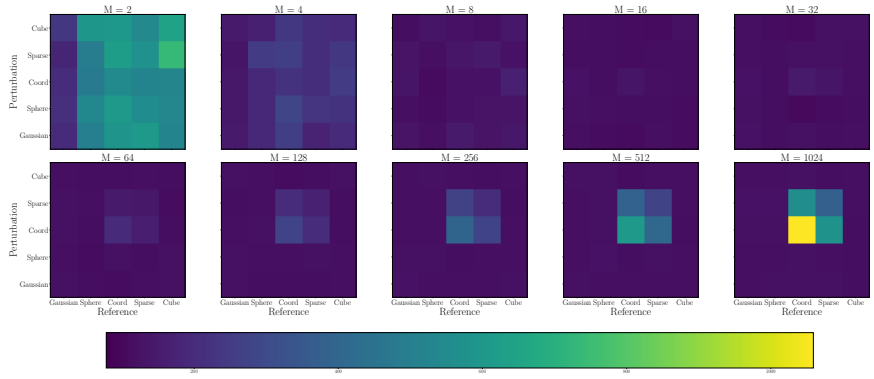
We utilize the nonlinear bandit tasks to further study the impact of update distribution in HyperAgent.

Task Settings: We We formulated two nonlinear contextual bandit tasks, with rewards generated by nonlinear functions in each.

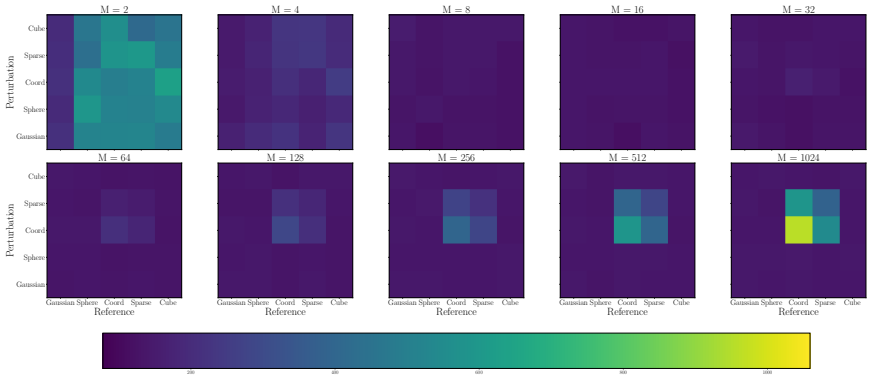
- **Neural Bandit**: It employs a nonlinear neural model denoted as $f_1(a)$ in reward generation. This model features three fully connected layers, each consisting of 50 units, and connected by ReLU activation functions.



(a) $d = 10 \quad |\mathcal{X}| = 100$

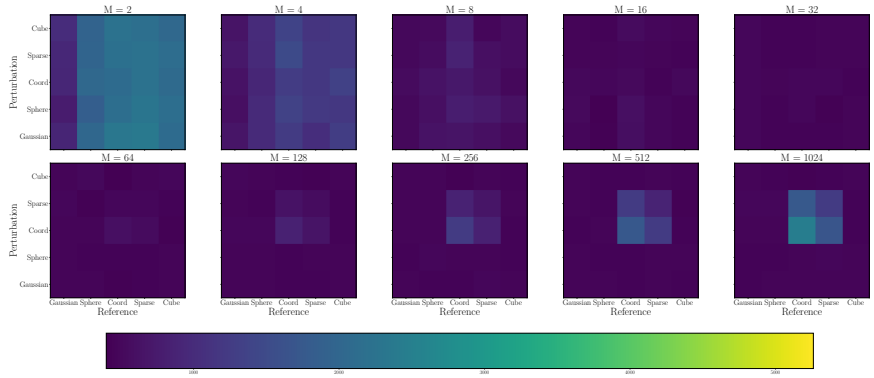


(b) $d = 10 \quad |\mathcal{X}| = 1000$

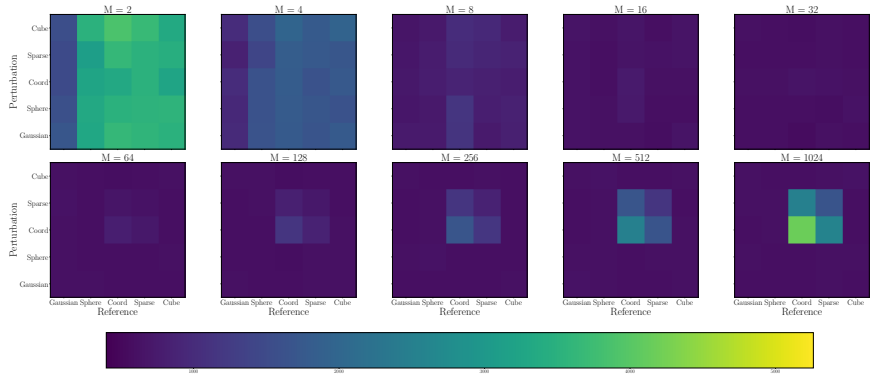


(c) $d = 10 \quad |\mathcal{X}| = 10000$

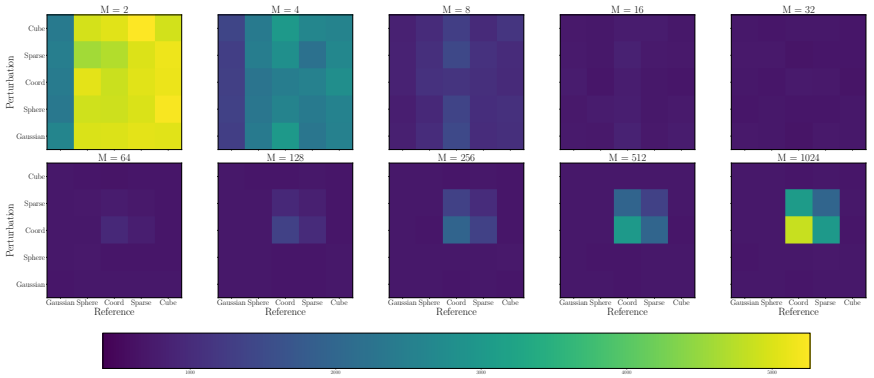
Figure 9: Results on the combinations of perturbation and reference distribution in Finite-action Linear Bandit under action dimension $d = 10$. A deeper color signifies lower accumulated regret and hence superior performance. Gaussian reference distribution significantly enhances performance.



(a) $d = 50 \quad |\mathcal{X}| = 100$

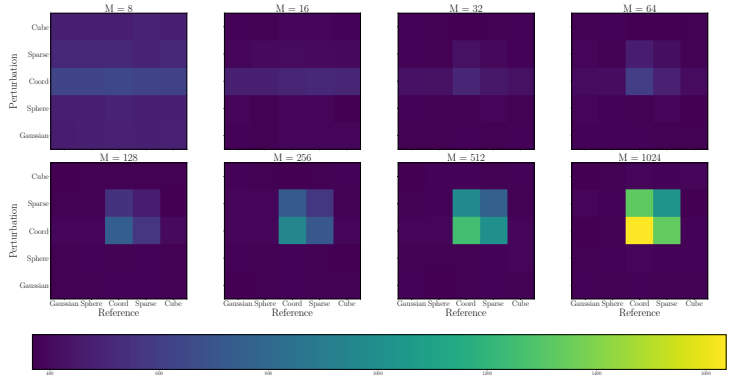


(b) $d = 50 \quad |\mathcal{X}| = 1000$

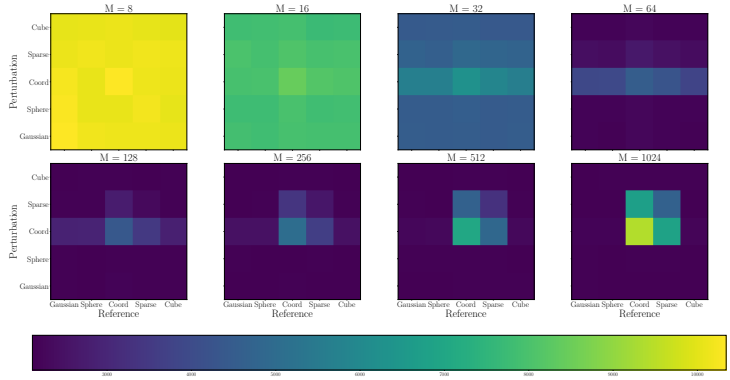


(c) $d = 50 \quad |\mathcal{X}| = 10000$

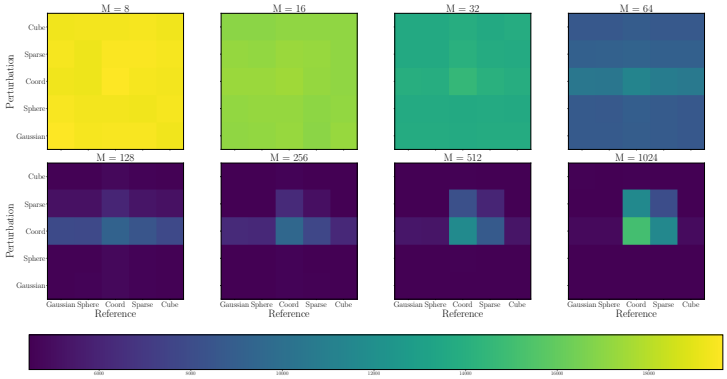
Figure 10: Results on the combinations of perturbation and reference distribution in Finite-action Linear Bandit under action dimension $d = 50$. A deeper color signifies lower accumulated regret and hence superior performance. Gaussian reference distribution significantly enhances performance.



(a) $d = 10$

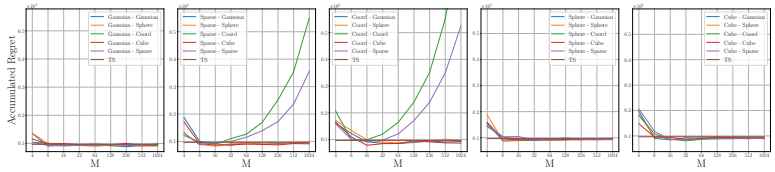


(b) $d = 50$

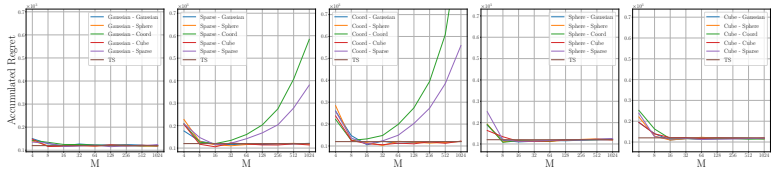


(c) $d = 100$

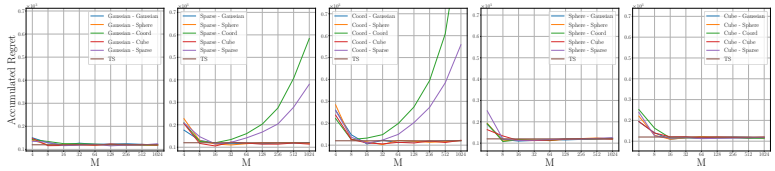
Figure 11: Results on the combinations of perturbation and reference distribution in Compact-action Linear Bandit. A deeper color signifies lower accumulated regret and hence superior performance. Gaussian reference distribution significantly enhances performance.



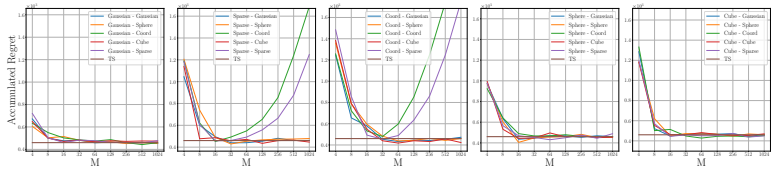
(a) $d = 10 \quad |\mathcal{X}| = 100$



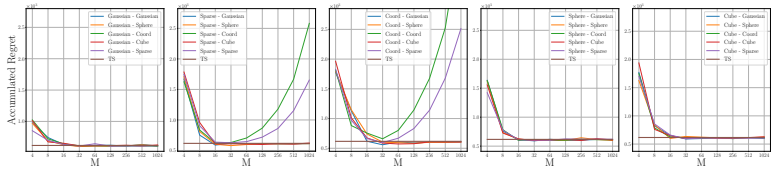
(b) $d = 10 \quad |\mathcal{X}| = 1000$



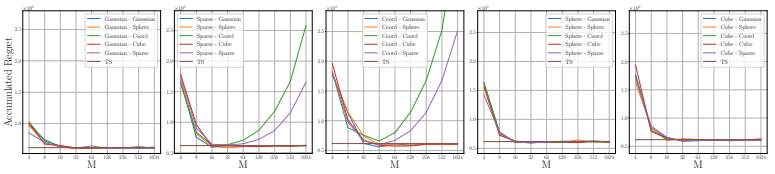
(c) $d = 10 \quad |\mathcal{X}| = 10000$



(d) $d = 50 \quad |\mathcal{X}| = 100$

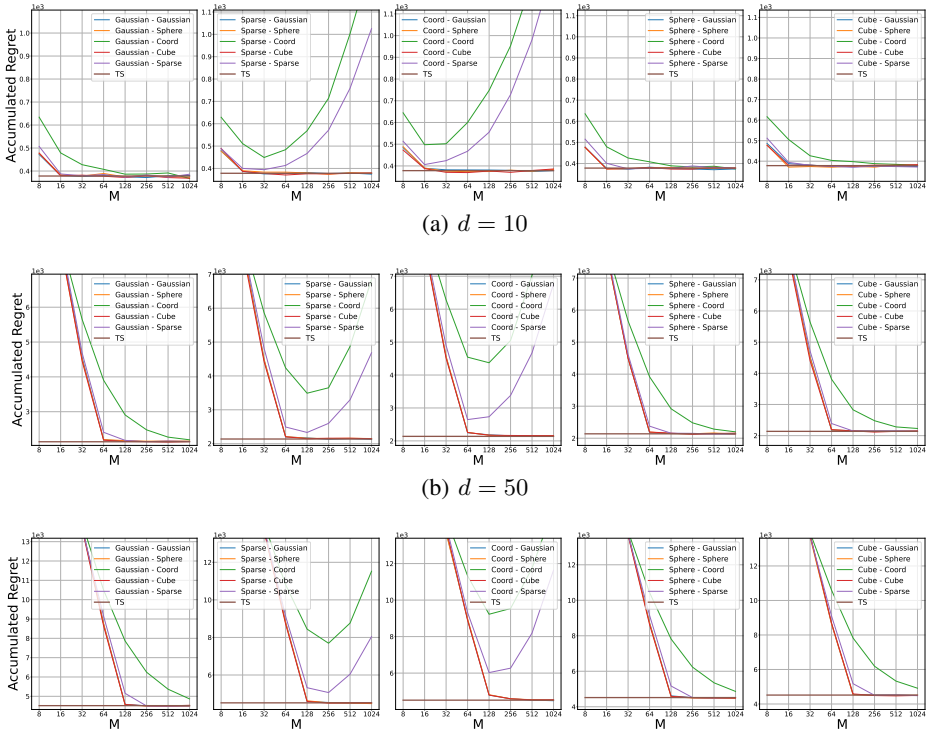


(e) $d = 50 \quad |\mathcal{X}| = 1000$



(f) $d = 50 \quad |\mathcal{X}| = 10000$

Figure 12: Results on regret under various index dimension M in Finite-action Linear Bandit. The label $A - B$ indicates that HyperAgent uses A as the reference distribution and B as the perturbation distribution. HyperAgent with Gaussian or Sphere reference distribution could achieve comparable performance with that of Thompson sampling under small M.



(a) $d = 10$

(b) $d = 50$

(c) $d = 100$

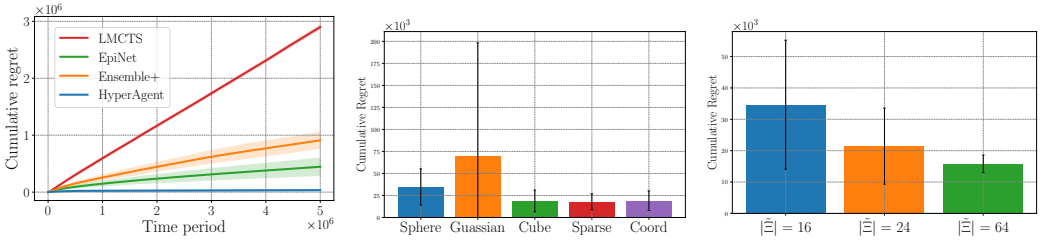
Figure 13: Results on regret under various index dimension M in Compact-action Linear Bandit. The label $A - B$ indicates that HyperAgent uses A as the reference distribution and B as the perturbation distribution. HyperAgent with Gaussian or Sphere reference distribution could achieve comparable performance with that of Thompson sampling under small M.

- **Quadratic Bandit:** Its reward generation mechanism built on a quadratic function, expressed as $f_2(a) = 10^{-2}(a^\top \Theta \Theta^\top a)$. Here, $a \in \mathbb{R}^d$ stands for the action while $\Theta \in \mathbb{R}^{d \times d}$ is a matrix filled with random variables originating from $\mathcal{N}(0, 1)$.

In both tasks, the original reward r is disrupted by Gaussian noise ϵ drawn from $\mathcal{N}(0, 0.1)$. We set the action dimension d at 100 and the size of action set $\mathcal{X}$ to 1000. Each experiment was run over 5 million time steps and repeated with 10 distinct random seeds to secure robust results

Results on Quadratic Bandit: We have conducted a comprehensive analysis of the results from the Neural Bandit task as depicted in Fig. 6, demonstrating the distinct advantages of HyperAgent. The outcomes from the Quadratic Bandit task, represented in Fig. 14, resonate with the observations from the Neural Bandit task. HyperAgent consistently outperforms other baselines and exhibits enhanced performance when implementing a discrete update distribution (i.e., Sparse, Cube, and Coord). The quantity of indices $|\tilde{\Xi}|$, in accordance with Eq. (1), continues to influence the performance of HyperAgent when the Sphere update distribution is applied. The uniformity of results across both Neural and Quadratic Bandit tasks highlights the superior generalization and efficient data and computational capability of HyperAgent in decision-making tasks.

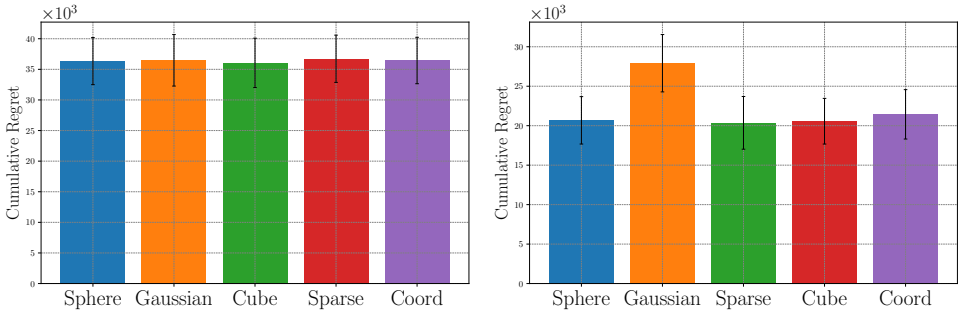
Ablation Study on Neural Bandit: We performed comparisons experiment on different reference and perturbation distributions on Neural Bandit. Firstly, we assigned the same Sphere to both the update and reference distributions, aiming to observe the impact of different perturbation distributions. From Fig. 15(a), it's evident that various perturbations resulted in near-identical performances. This observation aligns with the findings reported in the linear bandit task, suggesting that perturbations do not significantly influence the performance of the model. We then explored the impact of different reference distributions by allocating the Coord update distribution and Sphere perturbation distribution for HyperAgent. The results in Fig. 15(b) reveal a slight divergence from the linear bandit task, with Gaussian reference distribution showing suboptimal performance in the Neural Bandit. Nevertheless, the Sphere reference distribution demonstrated consistent efficacy, in line with the theoretical analysis



(a) (b) (c)

Figure 14: Experimental results on Quadratic Bandit. (a) Comparison results with baselines. (b) The impact of various update distribution. (c)The impact of $|\tilde{\Xi}|$ when using Sphere update distribution.

presented in Theorem 2. While the sub-par results under the Gaussian reference distribution set the stage for future investigations.



(a) Ablation study on perturbation distribution. (b) Ablation study on reference distribution.

Figure 15: Ablation studies on the Neural Bandit.

Additional Results: To ensure fair comparison between HyperAgent and LMCTS, we have meticulously adjusted LMCTS's update ratio to match that of HyperAgent's in Figs. 6 and 14. To further investigate the influence of the update ratio, we also evaluated LMCTS under its original settings, referred to as *LMCTS (original)*. In this setting, the update ratio initially follows a linearly increasing trend before stabilizing at 100. As demonstrated in Fig. 16, despite the incremental computational cost, *LMCTS (original)* falls short in attaining satisfactory performance in comparison to HyperAgent.

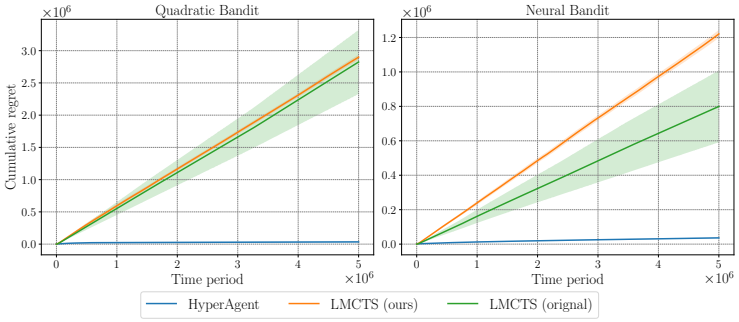


Figure 16: Comparison results between HyperAgent and LMCTS.

References

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

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011b.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2), 2017.

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

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. Fireact: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*, 2023a.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, et al. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*, 2023b.

Open X-Embodiment Collaboration, Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Hao Su, Hao-Shu Fang, Haochen Shi, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jaehyung Kim, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jiajun Wu, Jialin Wu, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jitendra Malik, Jonathan Tompson, Jonathan Yang, Joseph J. Lim, João Silvério, Junhyek Han, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Zhang, Keyvan Majd, Krishan Rana, Krishnan Srinivasan, Lawrence Yunliang Chen, Lerrel Pinto, Liam Tan, Lionel Ott, Lisa Lee, Masayoshi Tomizuka, Maximilian Du, Michael Ahn, Mingtong Zhang, Mingyu Ding, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Pannag R Sanketi, Paul Wohlhart, Peng Xu, Pierre Sermanet, Priya Sundaresan, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Sherry Moore, Shikhar Bahl, Shivin Dass, Shuran Song, Sichun Xu, Siddhant Haldar, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Sudeep Dasari, Suneel Belkhale, Takayuki Osa, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xuanlin Li, Yao Lu, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yueh hua Wu, Yujin Tang, Yuke Zhu, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zhuo Xu, and Zichen Jeff Cui. Open X-Embodiment: Robotic learning datasets and RT-X models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, 2024.

X Corp. The x rules: Safety, privacy, authenticity, and more. https://help.twitter.com/en/rules-and-policies/x-rules, 2024. Accessed: 2024-07-09.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In Rocco A. Servedio and Tong Zhang, editors, *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 355–366. Omnipress, 2008. URL http://colt2008.cs.helsinki.fi/papers/80-Dani.pdf.

Vikranth Dwaracherla, Xiuyuan Lu, Morteza Ibrahimi, Ian Osband, Zheng Wen, and Benjamin Van Roy. Hypermodels for exploration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryx6WgStPB.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1): 2053951719897945, 2020.

T. H. Gronwall. The gamma function in the integral calculus. *Annals of Mathematics*, 20(2):35–124, 1918. ISSN 0003486X. URL http://www.jstor.org/stable/1967180.

Lawrence Hollom and Julien Portier. Tight lower bounds for anti-concentration of rademacher sums and tomaszewski's counterpart problem. *arXiv preprint arXiv:2306.07811*, 2023.

Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. Crispr-gpt: An llm agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021*, 2024.

David Janz, Alexander E Litvak, and Csaba Szepesvári. Ensemble sampling for linear bandits: small ensembles suffice. *arXiv preprint arXiv:2311.08376*, 2023.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, pages 651–673. PMLR, 2018.

Daniel M Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):1–23, 2014.

Akshay Krishnamurthy, Keegan Harris, Dylan J. Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context?, 2024.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

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

Peter Lee, Carey Goldberg, and Isaac Kohane. *The AI revolution in medicine: GPT-4 and beyond*. Pearson, 2023.

Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. *IEEE Transactions on Information Theory*, 70(1):372–388, 2024a. doi: 10.1109/TIT.2023.3267732.

Yingru Li. Probability tools for sequential random projection, 2024a. URL https://arxiv.org/abs/2402.14026.

Yingru Li. Simple, unified analysis of johnson-lindenstrauss with applications, 2024b. URL https://arxiv.org/abs/2402.10232.

Yingru Li, Jiawei Xu, Lei Han, and Zhi-Quan Luo. Q-Star Meets Scalable Posterior Sampling: Bridging Theory and Practice via HyperAgent. In *Forty-first International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2024b. URL https://arxiv.org/abs/2402.10228.

Ziniu Li, Yingru Li, Yushun Zhang, Tong Zhang, and Zhi-Quan Luo. HyperDQN: A randomized exploration method for deep reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=X0nrKAXu7g-.

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, Zheng Wen, et al. Reinforcement learning, bit by bit. *Foundations and Trends® in Machine Learning*, 16(6): 733–865, 2023.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018, 2023.

Meta. Facebook community standards. https://transparency.meta.com/policies/community-standards/, 2024. Accessed: 2024-07-09.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Gergő Nemes. Error bounds and exponential improvements for the asymptotic expansions of the gamma function and its reciprocal. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 145(3):571–596, 2015. doi: 10.1017/S0308210513001558.

OpenAI. Hello gpt-4o, 2023. URL https://openai.com/index/hello-gpt-4o/. Accessed: 15-May-2024.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019. URL http://jmlr.org/papers/v20/18-339.html.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=dZqcC1qCmB.

Chao Qin, Zheng Wen, Xiuyuan Lu, and Benjamin Van Roy. An analysis of ensemble sampling. *Advances in Neural Information Processing Systems*, 35:21602–21614, 2022.

Reddit. Automoderator guide. https://www.reddit.com/r/reddit.com/wiki/automoderator/, 2024. Accessed: 2024-07-09.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Sarah T Roberts. *Behind the screen*. Yale University Press, 2019.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Maciej Skorski. Bernstein-type bounds for beta distribution. *Modern Stochastics: Theory and Applications*, 10(2):211–228, 2023.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Chih-Chun Wang, Sanjeev R Kulkarni, and H Vincent Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355, 2005.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Lilian Weng, Vik Goel, and Andrea Vallone. Using gpt-4 for content moderation. August 2023. URL https://openai.com/index/using-gpt-4-for-content-moderation/. OpenAI.

Pan Xu, Hongkai Zheng, Eric V Mazumdar, Kamyar Azizzadenesheli, and Animashree Anandkumar. Langevin monte carlo for contextual bandits. In *International Conference on Machine Learning*, pages 24830–24850. PMLR, 2022.

Xu Yan, Haiming Zhang, Yingjie Cai, Jingming Guo, Weichao Qiu, Bin Gao, Kaiqiang Zhou, Yue Zhao, Huan Jin, Jiantao Gao, et al. Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities. *arXiv preprint arXiv:2401.08045*, 2024.

Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. Foundation models for decision making: Problems, methods, and opportunities. *arXiv preprint arXiv:2303.04129*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. Huatuogpt, towards taming language model to be a doctor, 2023.

Yuan Zhou. Lecture 14: Lower bounds for linear bandits. Lecture Notes in IE498, 2019. URL https://yuanz.web.illinois.edu/teaching/IE498fa19/lec_14.pdf. Accessed: Jan 17th, 2024.