Randomized Exploration for Linear Bandits via Absolute Perturbations

Toshinori Kitamura, Shuai Liu, Csaba Szepesvári

Keywords: Linear bandit, Randomized exploration

Summary

In stochastic linear bandits, the canonical Upper Confidence Bound (UCB) algorithm admits a simple frequentist regret analysis but can be computationally demanding, while Thompson Sampling (TS) is computationally attractive yet typically harder to analyze due to its non-optimistic nature. We propose Absolute Thompson Sampling (ATS), a simple modification of TS that ensures optimism in expectation by replacing the signed exploration noise with its absolute value. This preserves the computational efficiency of TS while avoiding the technically involved anti-concentration arguments common in TS analyses, enabling a simple UCB-style regret analysis. We show that ATS achieves $\widetilde{O}(d^{3/2}\sqrt{K})$ regret, matching existing bounds for TS in linear bandits. We further introduce Ensemble Absolute Thompson Sampling (EATS), which takes the maximum over multiple absolute perturbations with normalization by the ensemble size. As the ensemble size grows, EATS converges to the UCB objective, recovering UCB behavior in the limit. Experiments show that moderate ensemble sizes already yield strong performance. Our results point to a bridge between randomized exploration and deterministic optimism both in theory and practice.

Contribution(s)

1. We introduce Absolute Thompson Sampling (ATS) a simple modification of Thompson Sampling that ensures optimism in expectation while retaining the same computational structure as TS.
   Context: While standard TS analyses rely on anti-concentration arguments and explicit handling of non-optimistic rounds (Agrawal & Goyal, 2013; Abeille & Lazaric, 2017), ATS enables a regret analysis conceptually similar to that of UCB.
2. To address the $\sqrt{d}$ regret gap between ATS and UCB, we propose Ensemble Absolute Thompson Sampling (EATS), which replaces a single perturbation with the maximum over multiple perturbations to more closely approximate the UCB objective.
   Context: None

Randomized Exploration for Linear Bandits via Absolute Perturbations

Toshinori Kitamura[1,†], **Shuai Liu**[1,†], **Csaba Szepesvári**[1]
{toshinor,shuai14,szepesva}@ualberta.ca

[1]**Department of Computing Science, University of Alberta, Canada**

[†] Equal contribution

Abstract

In stochastic linear bandits, the canonical Upper Confidence Bound (UCB) algorithm admits a simple frequentist regret analysis but can be computationally demanding, while Thompson Sampling (TS) is computationally attractive yet typically harder to analyze due to its non-optimistic nature. We propose Absolute Thompson Sampling (ATS), a simple modification of TS that ensures optimism in expectation by replacing the signed exploration noise with its absolute value. This preserves the computational efficiency of TS while avoiding the technically involved anti-concentration arguments common in TS analyses, enabling a simple UCB-style regret analysis. We show that ATS achieves $\widetilde{O}(d^{3/2}\sqrt{K})$ regret, matching existing bounds for TS in linear bandits. We further introduce Ensemble Absolute Thompson Sampling (EATS), which takes the maximum over multiple absolute perturbations with normalization by the ensemble size. As the ensemble size grows, EATS converges to the UCB objective, recovering UCB behavior in the limit. Experiments show that moderate ensemble sizes already yield strong performance. Our results point to a bridge between randomized exploration and deterministic optimism both in theory and practice.

1 Introduction

We study stochastic linear bandit problems, where a learner sequentially selects arms from a given arm set and receives noisy linear rewards. Efficient exploration is the central challenge for achieving low regret. Two of the most widely used approaches are optimism-based algorithms, such as Upper Confidence Bound (UCB), and randomized algorithms, such as Thompson Sampling (TS).

The UCB algorithm selects arms according to the *Optimism in the Face of Uncertainty* (OFU) principle, maximizing an upper confidence bound on the estimated reward. This principle leads to a simple and elegant regret analysis based on confidence sets and the elliptical potential lemma (Abbasi-Yadkori et al., 2011). However, UCB can be computationally expensive, as it requires maximizing an elliptical bonus term over the arm set. In contrast, TS selects arms by sampling a parameter from a posterior distribution and acting greedily with respect to the sampled parameter. TS is computationally attractive since each round reduces to solving a linear maximization problem over the arm set. However, its frequentist regret analysis is more involved than that of UCB. Because the sampled parameter may underestimate the true reward, typical analyses require that TS produces an optimistic parameter with some probability and control the regret incurred during non-optimistic rounds (Agrawal & Goyal, 2013; Abeille & Lazaric, 2017).

Thus, UCB and TS have complementary computational and analytical properties. We bridge this gap by introducing a simple modification of TS that mimics the behavior of UCB. In Section 4, we introduce **Absolute Thompson Sampling (ATS)**, which replaces the signed exploration noise

in TS with its absolute value. This modification ensures *optimism in expectation*: the exploration bonus used in ATS matches the UCB bonus in expectation. As a result, the regret analysis becomes conceptually similar to that of UCB and avoids explicitly handling non-optimistic rounds. We show that ATS achieves $\widetilde{O}(d^{3/2}\sqrt{K})$ regret with high probability, matching the existing TS regret bound for linear bandits (Abeille & Lazaric, 2017). Moreover, ATS retains the computational efficiency of TS, requiring only one additional linear maximization per round.

While ATS achieves the same regret bound as TS, it incurs an additional $\sqrt{d}$ factor compared to UCB (Abbasi-Yadkori et al., 2011). To address this gap, we propose an extension of ATS that more closely resembles UCB. Motivated by extreme-value properties of Gaussian random variables (Proposition 5.1), we introduce **Ensemble Absolute Thompson Sampling (EATS)**, which replaces the single perturbation in ATS with the maximum over multiple perturbations. We show that, as the ensemble size grows, the EATS objective converges to the UCB objective, recovering UCB behavior in the limit. Empirically, moderate ensemble sizes already yield strong performance, suggesting a practical interpolation between randomized exploration and deterministic optimism (Section 6). Establishing theoretical guarantees for EATS with a finite ensemble size remains an interesting open problem (Remark 5.3).

Together, our results highlight a new perspective on randomized exploration: by carefully shaping the exploration term, one can design algorithms that retain the computational efficiency of TS while exhibiting analysis and behavior closely aligned with UCB. Our work opens new directions for designing simple and efficient randomized algorithms for bandit problems.

2 Related Work

For linear bandits, Agrawal & Goyal (2013) and Abeille & Lazaric (2017) provided the fundamental frequentist regret analyses of TS, establishing $\widetilde{O}(d^{3/2}\sqrt{K})$ bounds. A common strategy in these works is to show that TS samples an optimistic parameter with constant probability p, and then bound the regret by scaling the optimistic regret by $1/p$ (Janz et al., 2024). Similar proof techniques were later extended to multinomial logistic bandits (Oh & Iyengar, 2019), generalized linear bandits (Kveton et al., 2020), and linear MDPs (Zanette et al., 2020). Abeille et al. (2025) developed a tighter analysis that avoids relying on constant-probability optimism, achieving $\widetilde{O}(d\sqrt{K})$ regret, matching the lower bound for linear bandits (Dani et al., 2008; Rusmevichientong & Tsitsiklis, 2010). However, their result requires additional structural assumptions on the action set such as the L_2 unit ball.

While these works establish sublinear regret guarantees, existing linear TS analyses are technically involved, as they must carefully control the regret incurred during non-optimistic rounds. In contrast, by ensuring *optimism in expectation*, our approach avoids the need to separately analyze non-optimistic rounds, leading to a proof that is conceptually as simple as that of UCB (Abbasi-Yadkori et al., 2011). See Section 4 for more details of our approach.

Closer to our work, Zhang (2022) and Huix et al. (2023) modify TS to enforce stronger optimism. However, their approaches rely on sampling from skewed posterior distributions via Markov Chain Monte Carlo, which increases computational complexity and diminishes the practical appeal of TS. In contrast, Vaswani et al. (2020) modify UCB by introducing random perturbations. While they achieve $\widetilde{O}(d\sqrt{K})$ regret, their method requires maximizing an elliptical bonus term over the arm set, which can be computationally expensive in practice. Our method, by contrast, introduces only a minimal modification to TS and can be implemented using the same linear maximization oracle as standard TS.

3 Preliminary

Given two vectors $x, y \in \mathbb{R}^d$ and a positive definite matrix $A \in \mathbb{R}^{d\times d}$, we denote $\langle x, y\rangle = x^\top y$ and $\|x\|_A = \sqrt{x^\top A x}$. We denote by $\mathbb{S}^{d-1}$ the unit sphere in $\mathbb{R}^d$. All-zero and all-one vectors are

denoted by $\mathbf{0}$ and $\mathbf{1}$, respectively. For random variables X_n and X, we write $X_n \xrightarrow{\mathbb{P}} X$ to denote convergence in probability, i.e., for any $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$.

3.1 Stochastic Linear Bandits

Let $\mathcal{A} \subset \mathbb{R}^d$ be the compact set of arms. Without loss of generality, we assume $\|\phi\|_2 \le 1$ for any $\phi \in \mathcal{A}$. At each round k, the agent pulls an arm $\phi_k \in \mathcal{A}$ and observes the reward $r_k = (\theta^\star)^\top \phi_k + \varepsilon_k$. Here, $\theta^\star \in \mathbb{R}^d$ is unknown to the agent and satisfies $\|\theta^\star\|_2 \le B$, and ε_k is zero-mean R-sub-Gaussian noise. Define $r(\phi) := (\theta^\star)^\top \phi$. The agent's goal is to achieve sublinear regret:

$$\mathrm{Reg}(K) := \textstyle\sum_{k=1}^{K} r^\star - r(\phi_k) = o(K) \tag{1}$$

where $r^\star = \max_{\phi\in\mathcal{A}} r(\phi)$. We denote $\phi^\star \in \arg\max_{\phi\in\mathcal{A}} r(\phi)$ as the optimal arm.

We denote Λ_k as the regularized Gram matrix and $\hat{\theta}_k$ as the ridge estimator of $\theta^\star$ at round k:

$$\Lambda_k = \lambda I + \sum_{i=1}^{k-1} \phi_i \phi_i^\top, \quad \hat{\theta}_k = \Lambda_k^{-1} \sum_{i=1}^{k-1} \phi_i r_i, \tag{2}$$

where $\lambda > 0$ is a regularization parameter.

3.2 Basic Algorithms

The two popular algorithms are Upper Confidence Bound (UCB) and Thompson Sampling (TS).

UCB algorithm. UCB constructs a confidence set for $\theta^\star$ and selects the arm with the largest upper confidence bound. The following lemma gives a confidence bound for the estimation error $\hat{\theta}_k - \theta^\star$.

Lemma 3.1 (Confidence bound; Abbasi-Yadkori et al., 2011)**.** *Let $\mathscr{E}$ be the event such that for all k,*

$$\left\|\hat{\theta}_k - \theta^\star\right\|_{\Lambda_k} \le \sqrt{\lambda} B + R\sqrt{d \ln\left(\frac{1 + k/\lambda}{\delta}\right)} =: \beta_k \,.$$

where $\delta > 0$. Then, for any $\delta > 0$, $\mathbb{P}(\mathscr{E}) \ge 1 - \delta$, and consequently, for any $\phi \in \mathcal{A}$ and k,

$$|\phi^\top \theta^\star - \phi^\top \hat{\theta}_k| \le \beta_k \|\phi\|_{\Lambda_k^{-1}} \,. \tag{3}$$

Using Eq. (3), UCB selects the arm with the largest upper confidence bound:

$$\text{(UCB)} \qquad \phi_k^{\mathrm{UCB}} \in \arg\max_{\phi\in\mathcal{A}} \ \phi^\top \hat{\theta}_k + \beta_k \|\phi\|_{\Lambda_k^{-1}}. \tag{4}$$

This rule embodies the *Optimism in the Face of Uncertainty* (OFU) principle: it favors arms with high estimated rewards and large uncertainty, thereby balancing exploitation and exploration. UCB has a regret of $\mathcal{O}(d\sqrt{K}\ln(K\delta^{-1}))$ with probability at least $1 - \delta$ (Abbasi-Yadkori et al., 2011).

Thompson sampling. TS maintains a posterior distribution over the unknown parameter $\theta^\star$ and acts greedily with respect to a sampled parameter. When Gaussian priors and likelihoods are used for estimating $\theta^\star$, the posterior distribution is also Gaussian, and TS selects the arm by:

$$\text{(TS)} \quad \phi_k^{\mathrm{TS}} \in \arg\max_{\phi\in\mathcal{A}} \ f(\phi, \xi_k) := \phi^\top \hat{\theta}_k + \beta_k \phi^\top \Lambda_k^{-1/2} \xi_k \quad \text{where} \quad \xi_k \sim \mathcal{N}(\mathbf{0}, I). \tag{5}$$

TS is often preferred in practice because it is computationally simpler and easier to implement. For a fixed Gaussian sample, Eq. (5) reduces to a linear maximization over $\mathcal{A}$, whereas UCB in Eq. (4) maximizes a linear term plus an ellipsoidal bonus.

However, regret analysis for TS is more challenging than for UCB. Unlike UCB, TS may select non-optimistic arms, which requires a more delicate analysis of non-optimistic rounds. A typical proof strategy first shows that TS samples an optimistic parameter with constant probability $p > 0$ via anti-concentration arguments (e.g., Lemma 2 in Agrawal & Goyal, 2013), and then scales the optimistic regret by $1/p$ (e.g., *master theorem* in Janz et al., 2024). This yields a regret bound of $\mathcal{O}(p^{-1}\sqrt{d^3K}\ln(K\delta^{-1}))$ (Agrawal & Goyal, 2013; Abeille & Lazaric, 2017).

4 Absolute Thompson Sampling

As explained in the previous section, UCB admits a simple regret analysis but can be computationally intensive, whereas TS is computationally efficient but requires a more involved analysis. To bridge this gap, we design a randomized algorithm that retains the computational efficiency of TS while allowing a regret analysis closer to that of UCB.

Our key idea is to make a minimal modification to TS so that it behaves more like UCB. Specifically, we use the following equality: For any symmetric positive definite matrix Λ and vector $\phi \in \mathbb{R}^d$,

$$\mathbb{E}_{\xi\sim\mathcal{N}(\mathbf{0},I)}\left[\left|\phi^\top \Lambda^{-1/2}\xi\right|\right] = \sqrt{\frac{2}{\pi}}\|\phi\|_{\Lambda^{-1}} \,. \tag{6}$$

Thus, for a fixed ϕ, the absolute inner-product term in TS, namely $|\phi^\top \Lambda_k^{-1/2}\xi_k|$, behaves like the UCB bonus term $\|\phi\|_{\Lambda_k^{-1}}$ in expectation.

Motivated by this observation, we replace the signed inner product in Eq. (5) with its absolute value.

$$\text{(ATS)} \quad \phi_k^{\mathrm{ATS}} \in \arg\max_{\phi\in\mathcal{A}} \ \phi^\top\hat{\theta}_k + \sqrt{\frac{\pi}{2}}\beta_k\left|\phi^\top\Lambda_k^{-1/2}\xi_k\right| \quad \text{where} \quad \xi_k \sim \mathcal{N}(\mathbf{0}, I). \tag{7}$$

We call this algorithm absolute Thompson sampling (ATS).

4.1 Properties of ATS

Let $\mathcal{F}_k := \sigma(\xi_1, \varepsilon_1, \ldots, \xi_k, \varepsilon_k)$ be the filtration such that ε_k and ξ_k are $\mathcal{F}_k$-measurable. By slightly abusing notation, we use the shorthand $\mathbb{E}_{\xi_k}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{k-1}]$.

Remark 4.1 (Optimism in expectation)**.** UCB pulls the arm with the largest upper confidence bound, which provides an upper bound on the optimal reward (see Eq. (3)). The ATS update rule is *more optimistic than UCB in expectation*, as the following lemma shows.

Lemma 4.2. *Suppose that the event $\mathscr{E}$ in Lemma 3.1 holds. Then, for any k,*

$$(\phi^\star)^\top\theta^\star \le \left(\phi_k^{\mathrm{UCB}}\right)^\top\hat{\theta}_k + \beta_k\left\|\left(\phi_k^{\mathrm{UCB}}\right)^\top\right\|_{\Lambda_k^{-1}} \le \mathbb{E}_{\xi_k}\left[\left(\phi_k^{\mathrm{ATS}}\right)^\top\hat{\theta}_k + \sqrt{\frac{\pi}{2}}\beta_k\left|\left(\phi_k^{\mathrm{ATS}}\right)^\top\Lambda_k^{-1/2}\xi_k\right|\right] .$$

Proof. It holds that

$$\begin{aligned}
0 &\overset{(a)}{\ge} (\phi^\star)^\top\theta^\star - \left[\left(\phi_k^{\mathrm{UCB}}\right)^\top\hat{\theta}_k + \beta_k\left\|\left(\phi_k^{\mathrm{UCB}}\right)^\top\right\|_{\Lambda_k^{-1}}\right] \\
&\overset{(b)}{=} (\phi^\star)^\top\theta^\star - \mathbb{E}_{\xi_k}\left[\left(\phi_k^{\mathrm{UCB}}\right)^\top\hat{\theta}_k + \sqrt{\frac{\pi}{2}}\beta_k\left|\left(\phi_k^{\mathrm{UCB}}\right)^\top\Lambda_k^{-1/2}\xi_k\right|\right] \\
&\overset{(c)}{\ge} (\phi^\star)^\top\theta^\star - \mathbb{E}_{\xi_k}\left[\left(\phi_k^{\mathrm{ATS}}\right)^\top\hat{\theta}_k + \sqrt{\frac{\pi}{2}}\beta_k\left|\left(\phi_k^{\mathrm{ATS}}\right)^\top\Lambda_k^{-1/2}\xi_k\right|\right] ,
\end{aligned}$$

where (a) follows from the definition of ϕ_k^{UCB} in Eq. (4) together with the reward confidence bound Eq. (3) under $\mathscr{E}$, (b) uses Eq. (6), and (c) follows from the arm-selection rule of ATS. □

Algorithm 1: Absolute Thompson Sampling (ATS)

for $k = 1, \dots, K$ **do**
  Update Gram matrix $\Lambda_k = \lambda I + \sum_{i=1}^{k-1} \phi_i \phi_i^\top$;
  Sample $\xi_k \sim \mathcal{N}(\mathbf{0}, I)$;
  Compute $\phi_k \in \arg\max_{\phi \in \mathcal{A}} \ \phi^\top \hat{\theta}_k + \sqrt{\frac{\pi}{2}} \beta_k \left| \phi^\top \Lambda_k^{-1/2} \xi_k \right|$. See Remark 4.3;
  Pull ϕ_k and observe r_k;
end

Remark 4.3 (Computational efficiency)**.** The additional absolute value only minimally changes the implementation of TS. Indeed, Eq. (7) is equivalent to evaluating both signs of ξ_k and selecting the better one. Specifically, let

$$\phi_k^+ \in \arg\max_{\phi \in \mathcal{A}} f\left(\phi, \sqrt{\tfrac{\pi}{2}} \xi_k\right) \quad \text{and} \quad \phi_k^- \in \arg\max_{\phi \in \mathcal{A}} f\left(\phi, -\sqrt{\tfrac{\pi}{2}} \xi_k\right),$$

where $f(\phi, \xi)$ denotes the TS objective defined in Eq. (5). Then

$$\phi_k^{\mathrm{ATS}} = \begin{cases} \phi_k^+ & \text{if } f\left(\phi_k^+, \sqrt{\frac{\pi}{2}} \xi_k\right) \geq f\left(\phi_k^-, -\sqrt{\frac{\pi}{2}} \xi_k\right) \\ \phi_k^- & \text{otherwise} \end{cases}. \tag{8}$$

This implementation requires only one additional call to the optimization oracle compared to standard TS. The full procedure is summarized in Algorithm 1.

4.2 Regret Analysis

The minimal algorithmic modification by absolute value significantly simplifies the regret analysis of TS. This section proves that Algorithm 1 achieves sublinear regret with high probability.

Theorem 4.4. *For any $\delta > 0$, with probability at least $1 - \delta$, for any $K \geq 1$, Algorithm 1 achieves regret of* $\mathrm{Reg}(K) = \mathcal{O}\left(\sqrt{d^3 K} \ln(K \delta^{-1})\right)$.

Proof. The proof mirrors the regret analysis of UCB (Abbasi-Yadkori et al., 2011): Decompose the regret with optimism (**Step 1**) and then bound the remaining terms via the elliptical potential lemma (**Step 3**). Since the optimism holds in expectation, we add one step to remove the expectations over the samples ξ_k before applying the elliptical potential lemma (**Step 2**). Notably, unlike the standard analysis of TS (e.g., Agrawal & Goyal, 2013; Abeille & Lazaric, 2017), **we do not need to establish a constant probability of optimism or separately handle non-optimistic rounds**, which significantly simplifies the proof.

Step 1. Regret decomposition with optimism. Suppose that the event $\mathscr{E}$ in Lemma 3.1 holds. By adding and subtracting the expected objective of ATS $\mathbb{E}_{\xi_k}\left[\phi_k^\top \hat{\theta}_k + \sqrt{\frac{\pi}{2}} \beta_k \left| \phi_k^\top \Lambda_k^{-1/2} \xi_k \right|\right]$, we decompose the regret by:

$$\begin{aligned} \mathrm{Reg}(K) = & \sum_{k=1}^{K} \underbrace{(\phi^\star)^\top \theta^\star - \mathbb{E}_{\xi_k}\left[\phi_k^\top \hat{\theta}_k + \sqrt{\frac{\pi}{2}} \beta_k \left| \phi_k^\top \Lambda_k^{-1/2} \xi_k \right|\right]}_{\leq 0 \text{ by Lemma 4.2}} \\ & + \underbrace{\sum_{k=1}^{K} \left(\mathbb{E}_{\xi_k}\left[\phi_k^\top \hat{\theta}_k\right] - \phi_k^\top \theta^\star\right)}_{=:①} + \underbrace{\sqrt{\frac{\pi}{2}} \sum_{k=1}^{K} \beta_k \mathbb{E}_{\xi_k}\left[\left| \phi_k^\top \Lambda_k^{-1/2} \xi_k \right|\right]}_{=:②}. \end{aligned} \tag{9}$$

Therefore, it suffices to bound ① and ②. The term ② is bounded by

$$② \overset{(a)}{\leq} \sqrt{\frac{\pi}{2}}\beta_K \sum_{k=1}^{K} \sqrt{\mathbb{E}_{\xi_k}[\|\xi_k\|_2^2]}\sqrt{\mathbb{E}_{\xi_k}\left[\|\phi_k\|^2_{\Lambda_k^{-1}}\right]} \tag{10}$$
$$\overset{(b)}{=} \sqrt{\frac{\pi d}{2}}\beta_K \sum_{k=1}^{K} \sqrt{\mathbb{E}_{\xi_k}\left[\|\phi_k\|^2_{\Lambda_k^{-1}}\right]} \overset{(c)}{\leq} \sqrt{\frac{\pi d}{2}}\beta_K \sqrt{K\sum_{k=1}^{K} \mathbb{E}_{\xi_k}\left[\|\phi_k\|^2_{\Lambda_k^{-1}}\right]},$$

where (a) uses the Cauchy–Schwarz inequality: $\mathbb{E}[X^\top Y] \leq \mathbb{E}[\|X\|_2\|Y\|_2] \leq \sqrt{\mathbb{E}\|X\|_2^2}\sqrt{\mathbb{E}\|Y\|_2^2}$, (b) uses $\mathbb{E}_\xi[\|\xi\|_2^2] = d$ for $\xi \sim \mathcal{N}(\mathbf{0}, I)$, and (c) uses Cauchy–Schwarz again.

By inserting the bound of ② into Eq. (9),

$$\mathrm{Reg}(K) \leq \underbrace{\sum_{k=1}^{K} \mathbb{E}\left[\phi_k^\top \hat{\theta}_k \,\middle|\, \mathcal{F}_{k-1}\right] - \phi_k^\top \theta^\star}_{=①} + \underbrace{\sqrt{\frac{\pi d}{2}}\beta_K \sqrt{K\sum_{k=1}^{K} \mathbb{E}\left[\|\phi_k\|^2_{\Lambda_k^{-1}} \,\middle|\, \mathcal{F}_{k-1}\right]}}_{=:③}. \tag{11}$$

Step 2. Removing expectations. The confidence bound Eq. (3) suggests that ③ is on the order of $\sum_{k=1}^{K}\|\phi_k\|_{\Lambda_k^{-1}}$. This motivates us to bound ① and ③ using the elliptical potential lemma:

Lemma 4.5 (Elliptical potential; Abbasi-Yadkori et al., 2011)**.** *For all $K \geq 1$ and $\lambda \geq 1$,*

$$\sum_{k=1}^{K} \|\phi_k\|^2_{\Lambda_k^{-1}} \leq 2d \ln\left(1 + \frac{K}{\lambda d}\right) \implies \sum_{k=1}^{K} \|\phi_k\|_{\Lambda_k^{-1}} \leq \sqrt{2dK \ln\left(1 + \frac{K}{\lambda d}\right)}.$$

However, to use Lemma 4.5, we need to remove the expectations in ① and ③ over the samples ξ_k. We use the following lemma to convert expectations to concrete realizations.

Lemma 4.6. *Let $\mathscr{E}_1$ and $\mathscr{E}_2$ be the events such that, for any $K \geq 1$,*

$$\mathscr{E}_1: \quad \sum_{k=1}^{K} \mathbb{E}\left[\phi_k^\top \hat{\theta}_k \,\middle|\, \mathcal{F}_{k-1}\right] \leq \sum_{k=1}^{K} \phi_k^\top \hat{\theta}_k + 2\left(B + \frac{\beta_K}{\sqrt{\lambda}}\right)\sqrt{K \ln \frac{\pi^2 K^2}{6\delta}}, \tag{12}$$

$$\mathscr{E}_2: \quad \sum_{k=1}^{K} \mathbb{E}\left[\|\phi_k\|^2_{\Lambda_k^{-1}} \,\middle|\, \mathcal{F}_{k-1}\right] \leq 2\sum_{k=1}^{K} \|\phi_k\|^2_{\Lambda_k^{-1}} + \frac{4}{\lambda} \ln \frac{2K}{\delta}, \tag{13}$$

where $\delta > 0$. Then, for any $\delta > 0$, $\mathbb{P}(\mathscr{E} \cap \mathscr{E}_1^c) \leq \delta$ and $\mathbb{P}(\mathscr{E}_2^c) \leq \delta$.

Proof. For the first claim, under the good event $\mathscr{E}$, we have

$$\left|\phi_k^\top \hat{\theta}_k\right| \leq \|\hat{\theta}_k\|_2 \leq \|\theta^\star\|_2 + \|\hat{\theta}_k - \theta^\star\|_2 \leq \|\theta^\star\|_2 + \frac{1}{\sqrt{\lambda}}\|\hat{\theta}_k - \theta^\star\|_{\Lambda_k} \leq B + \frac{\beta_K}{\sqrt{\lambda}}.$$

Then, the first claim is followed by applying a version of Azuma–Hoeffding (Lemma A.2) to the martingale difference sequence $\mathbb{E}\left[\phi_k^\top \hat{\theta}_k \,\middle|\, \mathcal{F}_{k-1}\right] - \phi_k^\top \hat{\theta}_k$. The second claim uses Lemma D.4 in Rosenberg et al. (2020) (see Lemma A.3) with the fact that $\|\phi_k\|^2_{\Lambda_k^{-1}} \leq \frac{1}{\lambda}$. □

Step 3. Applying elliptical potential lemma. Now suppose that the event $\mathscr{E} \cap \mathscr{E}_1 \cap \mathscr{E}_2$ holds. This happens with probability at least $1 - 3\delta$.[1] Then the first term ① in Eq. (11) is bounded by

$$① \overset{(a)}{\leq} \sum_{k=1}^{K} \phi_k^\top \hat{\theta}_k - \phi_k^\top \theta^\star + 2\left(B + \frac{\beta_K}{\sqrt{\lambda}}\right)\sqrt{K \ln \frac{\pi^2 K^2}{6\delta}} \overset{(b)}{\leq} \beta_K \sum_{k=1}^{K} \|\phi_k\|_{\Lambda_k^{-1}} + 2\left(B + \frac{\beta_K}{\sqrt{\lambda}}\right)\sqrt{K \ln \frac{\pi^2 K^2}{6\delta}},$$

[1] This is because $\mathbb{P}(\mathscr{E} \cap \mathscr{E}_1 \cap \mathscr{E}_2) = \mathbb{P}(\mathscr{E} \cap \mathscr{E}_1) - \mathbb{P}(\mathscr{E} \cap \mathscr{E}_1 \cap \mathscr{E}_2^c) \geq 1 - 2\delta - \mathbb{P}(\mathscr{E}_2^c) \geq 1 - 3\delta$.

where (a) uses $\mathscr{E}_1$ and (b) uses Eq. (3) under $\mathscr{E}$. Using $\mathscr{E}_2$, we have

$$③ \overset{(a)}{\leq} \beta_K \sqrt{\frac{\pi dK}{2}} \sqrt{2 \sum_{k=1}^{K} \|\phi_k\|^2_{\Lambda_k^{-1}} + \frac{4}{\lambda} \ln \frac{2K}{\delta}} \,.$$

Finally, by inserting the bounds for ① and ③ into Eq. (11),

$$\begin{aligned}
\mathrm{Reg}(K) \leq & \beta_K \sum_{k=1}^{K} \|\phi_k\|_{\Lambda_k^{-1}} + 2\left(B + \frac{\beta_K}{\sqrt{\lambda}}\right)\sqrt{K \ln \frac{\pi^2 K^2}{6\delta}} + \beta_K \sqrt{\frac{\pi dK}{2}} \sqrt{2 \sum_{k=1}^{K} \|\phi_k\|^2_{\Lambda_k^{-1}} + \frac{4}{\lambda} \ln \frac{2K}{\delta}} \\
\overset{(a)}{\leq} & \beta_K \sqrt{2dK \ln\left(1 + \frac{K}{\lambda d}\right)} + 2\left(B + \frac{\beta_K}{\sqrt{\lambda}}\right)\sqrt{K \ln \frac{\pi^2 K^2}{6\delta}} + \beta_K \sqrt{\frac{\pi dK}{2}} \sqrt{4d \ln\left(1 + \frac{K}{\lambda d}\right) + \frac{4}{\lambda} \ln \frac{2K}{\delta}} \\
\overset{(b)}{=} & \mathcal{O}\left(\sqrt{d^3 K} \ln(K\delta^{-1})\right) \,,
\end{aligned}$$

where (a) applies Lemma 4.5 to the red terms and (b) uses $\beta_K = \mathcal{O}(\sqrt{d \ln K\delta^{-1}})$. □

5 Ensemble Absolute Thompson Sampling

Although ATS is computationally efficient, its regret bound incurs an additional $\sqrt{d}$ factor compared to UCB. This looseness stems from the fact that the key identity (6) holds only for a fixed ϕ independent of the sample ξ, whereas in ATS the selected action ϕ_k^{ATS} depends on the sample ξ_k. Consequently, the ATS objective can be overly optimistic in expectation relative to UCB, as suggested by Lemma 4.2.

In this section, we explore a simple extension of ATS that can mitigate this $\sqrt{d}$ gap by more closely approximating the UCB objective. **Our analysis provides only asymptotic justification, and a non-asymptotic regret guarantee remains open.** We discuss the details in Remark 5.3.

The key idea is to leverage extreme-value properties of Gaussian random variables.

Proposition 5.1. *Let $\xi_1, \ldots, \xi_N \sim \mathcal{N}(\mathbf{0}, I)$ be i.i.d. Gaussian vectors. For fixed symmetric positive definite matrix $\Lambda \in \mathbb{R}^{d \times d}$ and vector $\phi \in \mathbb{R}^d$, we have*

$$\max_{i \in [N]} \frac{1}{\sqrt{2 \ln N}} \left| \phi^\top \Lambda^{-1/2} \xi_i \right| \xrightarrow{\mathbb{P}} \|\phi\|_{\Lambda^{-1}} \,. \tag{14}$$

The proof is provided in appendix B.1. Thus, for large N and fixed ϕ, the left hand side of Eq. (14) provides an estimator of the UCB bonus $\|\phi\|_{\Lambda^{-1}}$. This observation motivates the following algorithm which we call ensemble absolute Thompson sampling (EATS):

$$\text{(EATS)} \quad \phi_k^{\mathrm{EATS}} \in \arg\max_{\phi \in \mathcal{A}} \max_{i \leq N} \phi^\top \hat{\theta}_k + \frac{\beta_k}{\sqrt{2 \ln N}} |\phi^\top \Lambda_k^{-1/2} \xi_k^i| \quad \text{where} \quad \xi_k^i \sim \mathcal{N}(0, I). \tag{15}$$

5.1 Properties of EATS

Remark 5.2 (Computation of EATS)**.** Similar to ATS, EATS can be implemented using the same linear maximization oracle as standard TS. EATS requires $2N$ calls to the maximization oracle per round. For each $i \in [N]$, let

$$\phi_k^{+,i} \in \arg\max_{\phi \in \mathcal{A}} f(\phi, \tfrac{1}{\sqrt{2 \ln N}} \xi_k^i) \quad \text{and} \quad \phi_k^{-,i} \in \arg\max_{\phi \in \mathcal{A}} f(\phi, -\tfrac{1}{\sqrt{2 \ln N}} \xi_k^i),$$

Then, EATS can be implemented by picking the arm $\phi_k^{+,i}$ or $\phi_k^{-,i}$ that achieve the maximum value of $f(\phi, \pm \frac{1}{\sqrt{2 \ln N}} \xi_k^i)$ across all $i \in [N]$. Algorithm 2 summarizes the full procedure of EATS.

Algorithm 2: Ensemble Absolute Thompson Sampling (EATS)

for $k = 1, \ldots, K$ **do**

Update Gram matrix $\Lambda_k = \lambda I + \sum_{i=1}^{k-1} \phi_i \phi_i^\top$;

Sample $\xi_k^1, \ldots, \xi_k^N \sim \mathcal{N}(\mathbf{0}, I)$;

Compute $\phi_k \in \arg\max_{\phi \in \mathcal{A}} \max_{i \in [N]} \phi^\top \hat{\theta}_k + \frac{\beta_k}{\sqrt{2 \ln N}} \left| \phi^\top \Lambda_k^{-1/2} \xi_k^i \right|$. See Remark 5.2;

Pull ϕ_k and observe r_k;

end

Remark 5.3 (UCB in the limit)**.** From Proposition 5.1 and the compactness of $\mathcal{A}$, it is trivial that the EATS objective closely approximates the UCB objective when N is sufficiently large. Formally, we have the following result.

Proposition 5.4. *For any k, when $N \to \infty$, it holds that*

$$\max_{\phi \in \mathcal{A}} \max_{i \in [N]} \phi^\top \hat{\theta}_k + \frac{\beta_k}{\sqrt{2 \ln N}} |\phi^\top \Lambda_k^{-1/2} \xi_k^i| \xrightarrow{\mathbb{P}} \max_{\phi \in \mathcal{A}} \phi^\top \hat{\theta}_k + \beta_k \|\phi\|_{\Lambda_k^{-1}} \,. \tag{16}$$

For completeness, we provide the proof in appendix B.2. Since the objective of EATS matches that of UCB, EATS recovers UCB behavior in the limit of large N. As a direct consequence, when N is sufficiently large, EATS achieves the same regret bound as UCB, $\mathcal{O}\left(d\sqrt{K} \ln(K\delta^{-1})\right)$, which is $\sqrt{d}$ improvement over the ATS bound (Theorem 4.4).

However, the above argument is purely asymptotic and does not provide a non-asymptotic regret bound for EATS. If approximating UCB requires N to be exponentially large in d, the computational cost of EATS may become prohibitive. **For what finite size N EATS can achieve a UCB-matching regret bound remains an interesting open question.** Although we do not currently have a theoretical answer, our preliminary experiments in Section 6 suggest that moderate ensemble sizes already yield strong empirical performance, which is encouraging for the practical utility of EATS.

6 Experiments

This section presents the numerical results for the proposed algorithms. We first compare the empirical performance of ATS and TS, with the goal of showing that ATS achieves performance comparable to that of TS. We then study the empirical behavior of EATS to illustrate the potential of a new algorithmic design that may approach the statistical performance of UCB while substantially reducing computational complexity, provided that the ensemble size is chosen reasonably.

6.1 Comparison of ATS vs TS

We conduct experiments in three environments: the unit ball, the hypercube, and the two-lane shortest-path problem. Unless otherwise specified, all reported performance metrics are averaged over 100 independent random seeds.

Across these three environments, in addition to UCB (Eq. (4)), TS (Eq. (5)), and ATS (Eq. (7)), we also compare the following variants:

$$\text{(TS-No-Inflation)} \quad \phi_k^{\text{TS-No-Inflation}} \in \arg\max_{\phi \in \mathcal{A}} \; \phi^\top \hat{\theta}_k + \phi^\top \Lambda_k^{-1/2} \xi_k \quad \text{where} \quad \xi_k \sim \mathcal{N}(\mathbf{0}, I) \,.$$

$$\text{(ATS-No-}\sqrt{\pi/2}\text{)} \quad \phi_k^{\text{ATS-No-}\sqrt{\pi/2}} \in \arg\max_{\phi \in \mathcal{A}} \; \phi^\top \hat{\theta}_k + \beta_k \left| \phi^\top \Lambda_k^{-1/2} \xi_k \right| \quad \text{where} \quad \xi_k \sim \mathcal{N}(\mathbf{0}, I) \,.$$

$$\text{(ATS-No-Inflation)} \quad \phi_k^{\text{ATS-No-Inflation}} \in \arg\max_{\phi \in \mathcal{A}} \; \phi^\top \hat{\theta}_k + \left| \phi^\top \Lambda_k^{-1/2} \xi_k \right| \quad \text{where} \quad \xi_k \sim \mathcal{N}(\mathbf{0}, I) \,.$$

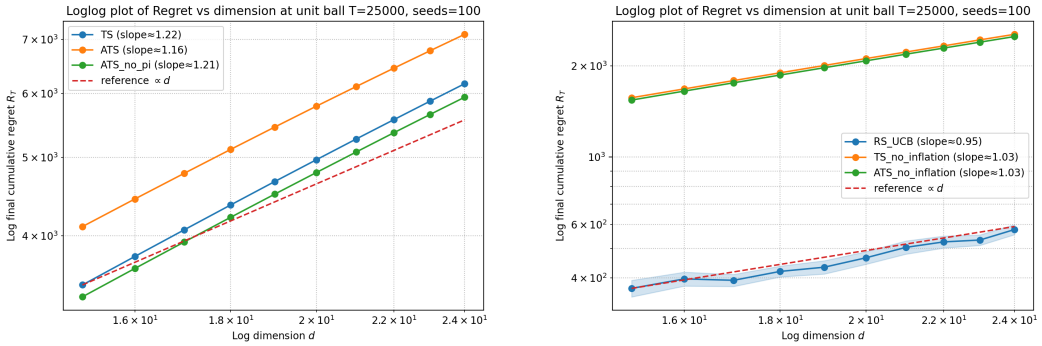

Figure 1: Sweep results for the linear bandit problem with unit-sphere arm set, $\|\theta^\star\| = 1$, and Gaussian noise $\eta_t \sim \mathcal{N}(0, 0.1)$. The parameter $\theta^\star$ is sampled uniformly from the unit sphere. Results are averaged over 100 independent random seeds. Both panels present log–log plots of the final regret of different algorithms, where the feature dimension ranges over $15, 16, \ldots, 24$ and the horizon is fixed at $T = 250000$: *(Left)* TS, ATS, and ATS-No-$\sqrt{\pi/2}$; *(Right)* RS-UCB, TS-No-Inflation, and ATS-No-Inflation.

In addition, because UCB can be computationally expensive, we use Rarely-Switching-UCB (RS-UCB) (Abbasi-Yadkori et al., 2011) to reduce the computational cost.

Unit Ball The unit ball environment is a natural benchmark in this setting, as information-theoretic lower bounds for linear bandits have been constructed with the arm set taken to be the unit ball (Lattimore & Szepesvári, 2020). In addition, recent work has shown that linear Thompson Sampling does not require forced exploration when the action set is smooth and convex (Abeille et al., 2025).

From figure 1, the left panel shows that ATS performs worse than TS by only a constant factor, while exhibiting the same dependence on the dimension d. After removing the $\sqrt{\pi/2}$ factor, as in ATS-No-$\sqrt{\pi/2}$, the resulting performance is comparable to that of TS, and is even slightly better in our experiments.

In the right panel, we further evaluate the no-inflation variants of both TS and ATS, following the algorithmic design in Abeille et al. (2025), where TS is shown to achieve a $\tilde{\mathcal{O}}(d\sqrt{T})$ regret guarantee on smooth and convex action sets such as the unit ball. Our experiments support this theoretical prediction and further suggest that ATS may also be able to match the performance of TS in the same setting considered by Abeille et al. (2025).

Hypercube The hypercube environment is defined as follows. For arm dimension $d \in \mathbb{Z}^+$, the action set is $\mathcal{A} = \{\pm 1\}^d/\sqrt{d}$, so that every arm $\phi \in \mathcal{A}$ satisfies $\|\phi\|_2 \leq 1$. The unknown parameter θ_* is sampled uniformly from the unit ℓ_2-sphere. We consider this environment for two reasons. First, information-theoretic lower bounds for linear bandits are known to arise from hypercube-type constructions (Lattimore & Szepesvári, 2020). Second, the action set has cardinality 2^d, making exact UCB optimization computationally expensive in this setting.

The hypercube is also a challenging experimental regime because the number of actions grows exponentially with d, and the asymptotic behavior appears only slowly. From figure 2, we observe that ATS performs worse than TS and ATS-No-$\sqrt{\pi/2}$ by an approximately constant factor. As in the unit ball setting, removing the $\sqrt{\pi/2}$ factor improves the empirical performance of ATS, making it slightly better than TS in our experiments. Moreover, the right panel suggests that these results should still be interpreted as transient behavior, since none of the observed regrets appears to have reached its asymptotic scaling by the end of the horizon. There are at least two possible explanations for why the observed d-dependence of TS, ATS, and ATS-No-$\sqrt{\pi/2}$ does not fully match the theoretical prediction. The first is that the dimensions considered may still be too small. In figure 2, for $d \in \{20, \ldots, 30\}$ and $T = 200{,}000$, all three algorithms still appear to be in the transient regime. Pushing to substantially larger d is computationally infeasible in our setting, since

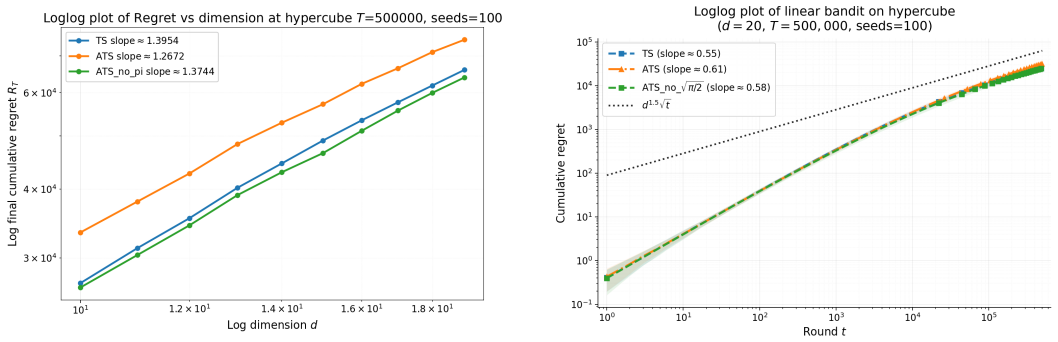

Figure 2: Sweep results for the linear bandit problem with the arm set is the hypercube, $\|\theta^\star\| = 1$, and noise $\eta_t \sim \mathcal{N}(0, 0.1)$. The parameter $\theta^\star$ is drawn uniformly at random from the unit sphere. Results are averaged over 100 independent random seeds. Both panels show plots of the final regret for different algorithms, with feature dimension varying over $10, 11, \ldots, 19$ and horizon fixed at $T = 500000$: *(Left)* Log–log plot of final regret of TS, ATS, and ATS-No-$\sqrt{\pi/2}$; *(Right)* TS, ATS, and ATS-No-$\sqrt{\pi/2}$.

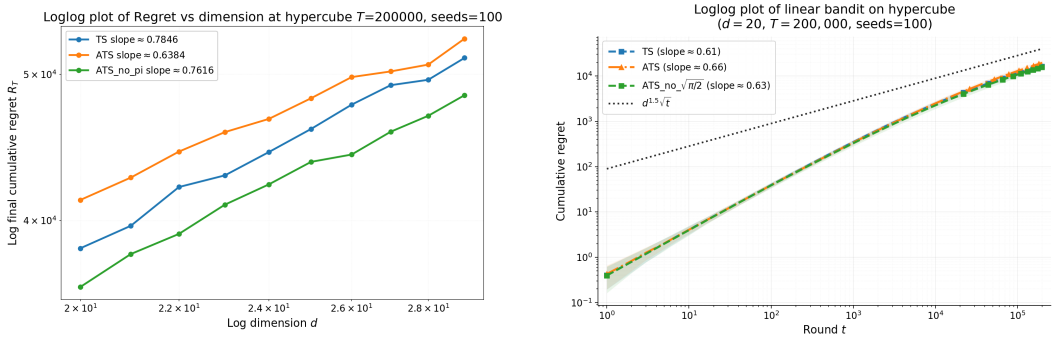

Figure 3: Numerical results for the linear bandit problem with the arm set is the hypercube, $\|\theta^\star\| = 1$, and noise $\eta_t \sim \mathcal{N}(0, 0.1)$. The parameter $\theta^\star$ is drawn uniformly at random from the unit sphere. Results are averaged over 100 independent random seeds. Both panels show plots of the final regret for different algorithms, with feature dimension varying over $20, 21, \ldots, 29$ and horizon fixed at $T = 200000$: *(Left)* Log–log plot of final regret of TS, ATS, and ATS-No-$\sqrt{\pi/2}$; *(Right)* TS, ATS, and ATS-No-$\sqrt{\pi/2}$.

even $d = 30$ already corresponds to roughly 2^{30} arms. The second possibility is that the time horizon is still not large enough. Indeed, the right panel of figure 2 indicates that all three methods remain in the transient regime even at the largest horizon we consider. However, horizons such as $T = 500{,}000$ are already substantial relative to the dimensions studied and are also near the limit of what we can simulate under our computational budget.

Nevertheless, the current experiments consistently support the conclusion that ATS is worse than TS and ATS-No-$\sqrt{\pi/2}$ by an approximately constant factor, that removing the $\sqrt{\pi/2}$ term improves ATS and can even make it slightly outperform TS empirically, and that these methods appear to exhibit broadly similar dependence on d. It is therefore reasonable to conjecture that this qualitative picture persists in larger-d and larger-T regimes.

Two-Lane Shortest Path In this section, we study a shortest-path bandit instance on a two-lane layered graph. The learner's goal is to identify a minimum-cost path from a source node s to a sink node t. The graph has $m \in \mathbb{Z}^+$ stages, each containing two nodes representing the upper and lower lanes. From each node at stage i, the learner may either stay in the same lane or switch to the other lane at stage $i + 1$. Hence, each inter-stage transition admits four possible edges: $\{\text{upper, lower}\} \to \{\text{upper, lower}\}$.

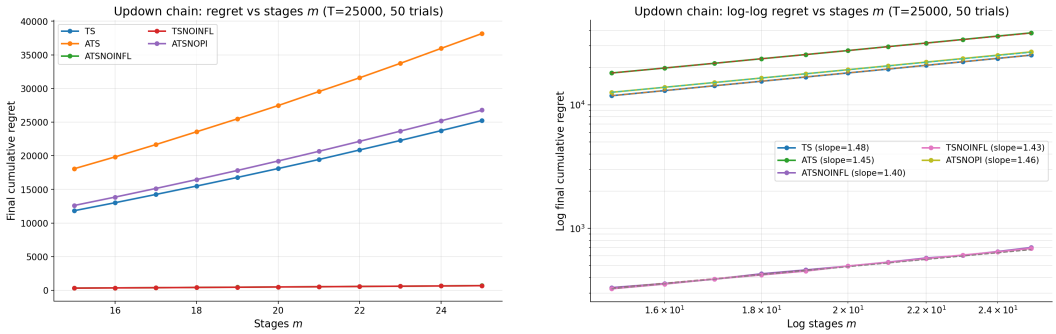

Figure 4: Sweep results for the two-lane shortest-path problem with Gaussian noise $\varepsilon_t \sim \mathcal{N}(0, 0.2)$, averaged over 50 independent random seeds. Both panels report the final regret of TS, TS-No-Inflation, ATS, ATS-No-$\sqrt{\pi/2}$, and ATS-No-Inflation as the feature dimension ranges over $\{15, 16, \ldots, 25\}$, with horizon fixed at $T = 250000$: *(Left)* final regret versus d; *(Right)* log-log plot of final regret versus d.

We assign cost 1 to every switching transition, cost 1 to upper $\to$ upper, and cost 0 to lower $\to$ lower. Consequently, the optimal path is obtained by staying in the lower lane at every stage, yielding total cost 0. The resulting graph contains 2^m distinct s-t paths and $d = 2 + 4(m-1) + 2 = 4m$ edges. See figure 9 as an example.

For any path x, let $\phi(x) \in \{0,1\}^d$ denote its edge-incidence feature vector. The observed cost is given by $\tilde{C}(x) = \phi(x)^\top \theta_* + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, R^2)$. The edge-parameter vector θ_* assigns value 0 to lower$\to$lower edges and value 1 to all other edges, that is, $(\theta_*)_i = 1 - \mathbb{I}(e_i = \text{lower} \to \text{lower})$. From figure 4, we see that ATS is worse than TS and ATS-No-$\sqrt{\pi/2}$ only by a constant factor, while all three appear to share the same scaling with the dimension d. Consistent with the unit ball experiments, removing the $\sqrt{\pi/2}$ factor yields a constant-factor improvement in empirical performance, and ATS-No-$\sqrt{\pi/2}$ becomes comparable to TS.

Moreover, TS-No-Inflation and ATS-No-Inflation continue to outperform the other algorithms by a significant margin. However, their empirical dependence on d appears to be close to $d^{1.5}$, in contrast to the behavior observed in the unit ball setting.

6.2 Empirical evidences for EATS

In this section, we present empirical evidence that EATS has the potential to satisfy a $\tilde{O}(d\sqrt{K})$ regret guarantee when $N = \text{poly}(d)$. To this end, we study the dependence of cumulative regret on the dimension d through a dimension sweep. In all EATS experiments reported in this section, we choose $N = d$. The settings of the environments in this section are also the same as the ones in Section 6.1.

Unit Ball From figure 5, we see that the regret of EATS exhibits an approximately linear dependence on d, while remaining worse than UCB by a substantial constant factor. At the same time, its empirical performance is comparable to that of TS-No-Inflation, for which Abeille et al. (2025) established a regret bound with linear dependence on d.

Hypercube From the left panel of figure 6, we observe that the regret of EATS scales linearly with d. The right panel further suggests that EATS has already reached its asymptotic regime by $d = 30$ and $T = 200{,}000$, which in turn indicates that the experiments for $d \in \{20, \ldots, 30\}$ are likely also operating in the asymptotic regime. In contrast, figure 3 shows that TS and ATS still remain in their transient regimes over the same range, which explains why their observed dependence on d is smaller.

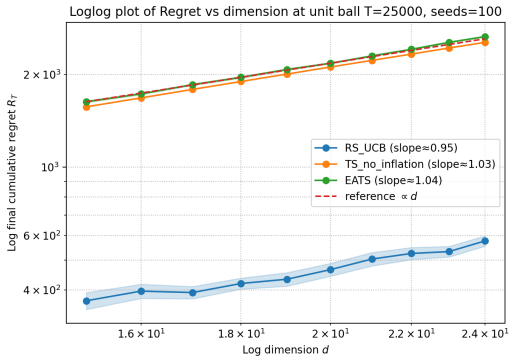

Figure 5: Sweep results for the linear bandit problem with unit-sphere arm set. Same setting as figure 1. Results are averaged over 100 independent random seeds. Log–log plot of regret vs dimension for EATS, RS-UCB, and TS-No-Inflation.

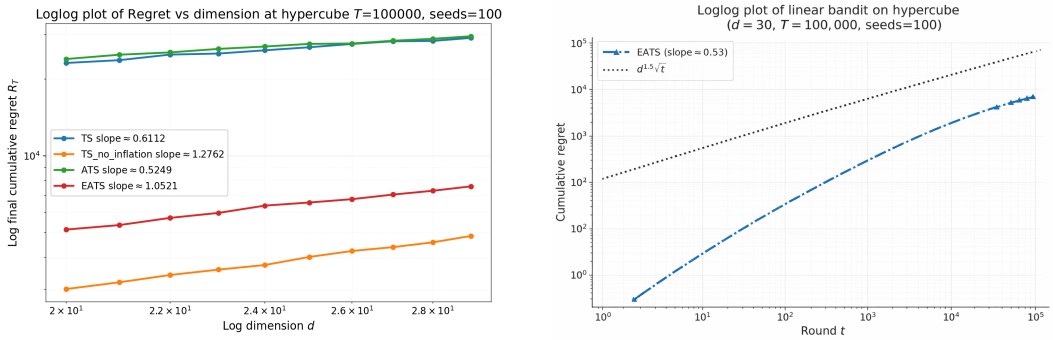

Figure 6: Sweep results for the linear bandit problem with the arm set is the hypercube. Same setting as figure 2. *(Left)* Log–log plot of final regret of EATS, TS, ATS, and TS-No-Inflation; *(Right)* The behavior of EATS at $d = 30$ and $T = 100,000$.

Overall, the fact that EATS consistently displays linear scaling across both environments, including one in which the asymptotic regime appears to be visible, suggests that a $\tilde{O}(d\sqrt{K})$-type guarantee for EATS is plausible.

7 Conclusion

We introduced Absolute Thompson Sampling (ATS), which replaces the signed exploration noise in TS with its absolute value. This ensures optimism in expectation, enabling a UCB-style regret analysis without the technically involved anti-concentration arguments typically required for TS. ATS achieves $\widetilde{O}(d^{3/2}\sqrt{K})$ regret while retaining the computational efficiency of randomized exploration.

We also proposed Ensemble Absolute Thompson Sampling (EATS), which approximates the UCB objective by taking the maximum over multiple absolute perturbations with normalization by the ensemble size. As the ensemble size grows, EATS recovers UCB behavior in the limit. Experiments suggest that moderate ensemble sizes can yield strong performance. Establishing non-asymptotic guarantees for EATS with finite ensemble sizes remains an interesting direction for future work.

Overall, our results suggest a new perspective on randomized exploration: by carefully shaping the exploration term, one can design algorithms that retain the computational efficiency of TS while exhibiting analysis and behavior closely aligned with UCB.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, 2011.

Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In *Artificial Intelligence and Statistics*, 2017.

Marc Abeille, David Janz, and Ciara Pike-Burke. When and why randomised exploration works (in linear bandits). In *International Conference on Algorithmic Learning Theory*, 2025.

Shipra Agrawal and Navin Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *International conference on Machine Learning*, 2013.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic Linear Optimization Under Bandit Feedback. In *Annual Conference on Learning Theory*, 2008.

Tom Huix, Matthew Zhang, and Alain Durmus. Tight Regret and Complexity Bounds for Thompson Sampling via Langevin Monte Carlo. In *International Conference on Artificial Intelligence and Statistics*, 2023.

David Janz, Alexander E Litvak, and Csaba Szepesvári. Ensemble Sampling for Linear Bandits: Small Ensembles Suffice. In *Advances in Neural Information Processing Systems*, 2024.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized Exploration in Generalized Linear Bandits. In *International Conference on Artificial Intelligence and Statistics*, 2020.

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

Min-hwan Oh and Garud Iyengar. Thompson Sampling for Multinomial Logit Contextual Bandits. In *Advances in Neural Information Processing Systems*, 2019.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-Optimal Regret Bounds for Stochastic Shortest Path. In *International Conference on Machine Learning*, 2020.

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

Sharan Vaswani, Abbas Mehrabian, Audrey Durand, and Branislav Kveton. Old Dog Learns New Tricks: Randomized UCB for Bandit Problems. In *Artificial Intelligence and Statistics*, 2020.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist Regret Bounds for Randomized Least-Squares Value Iteration. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Tong Zhang. Feel-Good Thompson Sampling for Contextual Bandits and Reinforcement Learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Useful Lemmas

Lemma A.1 (Azuma-Hoeffding)**.** *Let $(X_k, \mathcal{F}_k)_{k=1}^K$ be a martingale difference sequence, i.e., X_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] = 0$. Assume $X_k \in [l_k, u_k]$ almost surely. Then, for any $\delta \in (0,1)$,*

$$\mathbb{P}\left(\sum_{k=1}^K X_k \geq \sqrt{\sum_{k=1}^K \frac{(u_k - l_k)^2}{2} \log \frac{1}{\delta}}\right) \leq \delta \,.$$

Lemma A.2 (Azuma-Hoeffding with good events)**.** *Let $(Y_k)_{k=1}^\infty$ be $\mathcal{F}_k$-measurable and integrable random variables. Let $\mathscr{E}_k$ be $\mathcal{F}_{k-1}$-measurable events and let $\mathscr{E}$ be an event such that $\mathscr{E} \subseteq \bigcap_{k=1}^\infty \mathscr{E}_k$. Assume that on $\mathscr{E}_k$, we have $Y_k \in [l_k, u_k]$ almost surely. Then, for any $\delta \in (0,1)$,*

$$\mathbb{P}\left(\mathscr{E} \cap \left\{\exists K \geq 1 : \sum_{k=1}^K (Y_k - \mathbb{E}[Y_k \mid \mathcal{F}_{k-1}]) \geq \sqrt{\sum_{k=1}^K \frac{(u_k - l_k)^2}{2} \log \frac{\pi^2 K^2}{6\delta}}\right\}\right) \leq \delta \,.$$

Proof. Define $\tilde{Y}_k \coloneqq \mathbb{1}[\mathscr{E}_k] Y_k$ which returns Y_k if $\mathscr{E}_k$ holds and 0 otherwise. Since $\mathscr{E}_k \in \mathcal{F}_{k-1}$, the process $\tilde{X}_k \coloneqq \tilde{Y}_k - \mathbb{E}[\tilde{Y}_k \mid \mathcal{F}_{k-1}]$ is $\mathcal{F}_k$-measurable and a martingale difference sequence. Moreover, $\tilde{X}_k$ lies in an interval of length at most $u_k - l_k$ almost surely. Define $\delta_K \coloneqq 6\delta/\pi^2 K^2$. By Lemma A.1, and taking a union bound over $K \geq 1$,

$$\mathbb{P}\left(\exists K \geq 1 : \sum_{k=1}^K \tilde{X}_k \geq \sqrt{\sum_{k=1}^K \frac{(u_k - l_k)^2}{2} \log \frac{1}{\delta_K}}\right) \leq \sum_{K=1}^\infty \delta_K = \delta \,.$$

On the event $\mathscr{E}$, we have $\tilde{Y}_k = Y_k$ for all k. Therefore,

$$\mathbb{P}\left(\mathscr{E} \cap \left\{\exists K \geq 1 : \sum_{k=1}^K (Y_k - \mathbb{E}[Y_k \mid \mathcal{F}_{k-1}]) \geq \sqrt{\sum_{k=1}^K \frac{(u_k - l_k)^2}{2} \log \frac{1}{\delta_K}}\right\}\right) \leq \delta \,.$$

□

Lemma A.3 (Lemma D.4 in Rosenberg et al. (2020)**).** *Let $(X_k)_{k=1}^\infty$ be random variables with expectation adapted to the filtration $(\mathcal{F}_k)_{k=0}^\infty$. Suppose that $0 \leq X_k \leq C$ almost surely. Then, with probability at least $1-\delta$, the following holds for all $K \geq 1$:*

$$\sum_{k=1}^K \mathbb{E}\left[X_k \mid \mathcal{F}_{k-1}\right] \leq 4C \ln \frac{2K}{\delta} + 2\sum_{k=1}^K X_k \,.$$

B Proofs for Section 5

B.1 Proof of Proposition 5.1

Lemma B.1 (Restatement of Proposition 5.1)**.** *Let $\xi_1, \ldots, \xi_N \sim \mathcal{N}(\mathbf{0}, I)$ be i.i.d. Gaussian vectors. For fixed symmetric positive definite matrix $\Lambda \in \mathbb{R}^{d \times d}$ and vector $\phi \in \mathbb{R}^d$, we have*

$$\max_{i \in [N]} \frac{1}{\sqrt{2 \ln N}} \left|\phi^\top \Lambda^{-1/2} \xi_i\right| \xrightarrow{\mathbb{P}} \|\phi\|_{\Lambda^{-1}} \,.$$

Proof. Let $Z_1, \dots, Z_N \sim \mathcal{N}(0,1)$ be i.i.d. random variables. Then,

$$\max_{i\in[N]} \frac{1}{\sqrt{2\ln N}} \left|\phi^\top \Lambda^{-1/2} \xi_i\right| = \|\phi\|_{\Lambda^{-1}} \frac{1}{\sqrt{2\ln N}} \max_{i\in[N]} |Z_i| \,.$$

Thus, it suffices to show that $\frac{1}{\sqrt{2\ln N}} \max_{i\in[N]} |Z_i| \xrightarrow{\mathbb{P}} 1$.

We use the shorthand $u_N \coloneqq \sqrt{2\ln N}$. We prove the claim by showing that, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{\max_{i\in[N]}|Z_i|}{u_N} - 1\right| > \varepsilon\right) \le \mathbb{P}\left(\max_{i\in[N]} |Z_i| > (1+\varepsilon)u_N\right) + \mathbb{P}\left(\max_{i\in[N]} |Z_i| \le (1-\varepsilon)u_N\right) \underset{N\to\infty}{\longrightarrow} 0. \tag{17}$$

We show that both the upper tail and the lower tail vanish as $N \to \infty$.

Upper tail bound. Fix $\varepsilon > 0$. Let $Z \sim \mathcal{N}(0,1)$. Using $\mathbb{P}(|Z| > t) \le 2\exp(-t^2/2)$ for $t \ge 0$,

$$\begin{aligned}\mathbb{P}\left(\max_{i\in[N]} |Z_i| > (1+\varepsilon)u_N\right) &\le N\mathbb{P}\left(|Z| > (1+\varepsilon)u_N\right) \le 2N\exp\left(-\frac{(1+\varepsilon)^2 u_N^2}{2}\right) \\ &= 2N\exp\left(-(1+\varepsilon)^2 \ln N\right) = 2N^{1-(1+\varepsilon)^2} = 2N^{-\left(2\varepsilon+\varepsilon^2\right)} \underset{N\to\infty}{\longrightarrow} 0.\end{aligned}$$

Lower tail bound. For the lower tail, we have

$$\mathbb{P}\left(\max_{i\le N} |Z_i| \le t\right) = (1 - \mathbb{P}(|Z| > t))^N \le \exp(-N\mathbb{P}(|Z| > t)) \,. \tag{18}$$

We use the following Gaussian tail lower bound:

$$\mathbb{P}(|Z| > t) \ge \frac{2}{\sqrt{2\pi}} \frac{t}{1+t^2} e^{-t^2/2}$$

Take $t = (1-\varepsilon)u_N$. When $N \to \infty$, we have $t \to \infty$ and $\frac{t}{1+t^2} \asymp \frac{1}{t}$. So for large N,

$$\mathbb{P}(|Z| > t) \ge c\frac{1}{u_N} \exp\left(-\frac{(1-\varepsilon)^2 u_N^2}{2}\right) = c\frac{1}{u_N} N^{-(1-\varepsilon)^2} \,.$$

Therefore,

$$N\mathbb{P}(|Z| > t) \ge c\frac{1}{u_N} N^{1-(1-\varepsilon)^2} = c\frac{1}{\sqrt{2\ln N}} N^{2\varepsilon-\varepsilon^2} \underset{N\to\infty}{\longrightarrow} \infty,$$

since $2\varepsilon - \varepsilon^2 > 0$. By Eq. (18), we have

$$\mathbb{P}\left(\max_{i\le N} |Z_i| \le (1-\varepsilon)u_N\right) \underset{N\to\infty}{\longrightarrow} 0 \,.$$

By substituting the upper and lower tail bounds into Eq. (17), we obtain the desired result. □

B.2 Proof of Proposition 5.4

Proposition B.2 (Restatement of Proposition 5.4)**.** *For any k, when $N \to \infty$, it holds that*

$$\max_{\phi\in\mathcal{A}} \max_{i\in[N]} \phi^\top \hat{\theta}_k + \frac{\beta_k}{\sqrt{2\ln N}} |\phi^\top \Lambda_k^{-1/2} \xi_k^i| \xrightarrow{\mathbb{P}} \max_{\phi\in\mathcal{A}} \phi^\top \hat{\theta}_k + \beta_k \|\phi\|_{\Lambda_k^{-1}} \,. \tag{19}$$

Proof. We condition on $\mathcal{F}_{k-1}$. Without loss of generality, we prove the following claim: For fixed symmetric positive definite matrix $\Lambda \in \mathbb{R}^{d\times d}$, vector $\theta \in \mathbb{R}^d$ and $\beta > 0$,

$$\max_{\phi\in\mathcal{A}} \phi^\top \theta + \underbrace{\beta \max_{i\in[N]} \frac{1}{\sqrt{2\ln N}} \left|\phi^\top \Lambda^{-1/2} \xi^i\right|}_{=: M_N(\phi)} \xrightarrow{\mathbb{P}} \max_{\phi\in\mathcal{A}} \phi^\top \theta + \beta\|\phi\|_{\Lambda^{-1}} \,. \tag{20}$$

Since $\mathcal{A}$ is compact, pick a finite ε-net $\mathcal{N}_\varepsilon$ of $\mathcal{A}$ such that, for every $\phi \in \mathcal{A}$, there is $\psi \in \mathcal{A}_\varepsilon$ with $\|\phi - \psi\|_2 \le \varepsilon$. From Proposition 5.1 and union bound over $\psi \in \mathcal{A}_\varepsilon$, we have

$$\max_{\psi \in \mathcal{A}_\varepsilon} |M_N(\psi) - \beta \|\psi\|_{\Lambda^{-1}}| \xrightarrow{\mathbb{P}} 0 \,. \tag{21}$$

We have the following Lipschitz continuity of M_N: for any $\phi_1, \phi_2 \in \mathcal{A}$,

$$|M_N(\phi_1) - M_N(\phi_2)| \le \frac{\beta}{\sqrt{2 \ln N}} \max_{i \in [N]} \left| (\phi_1 - \phi_2)^\top \Lambda^{-1/2} \xi^i \right| \le \|\phi_1 - \phi_2\|_2 L_N \,,$$

where we defined

$$L_N \coloneqq \frac{\beta}{\sqrt{2 \ln N}} \max_{i \in [N]} \left\| \Lambda^{-1/2} \xi^i \right\|_2 \le \beta \left\| \Lambda^{-1/2} \right\|_{\mathrm{op}} \frac{1}{\sqrt{2 \ln N}} \max_{i \in [N]} \left\| \xi^i \right\|_2 \xrightarrow{\mathbb{P}} \beta \left\| \Lambda^{-1/2} \right\|_{\mathrm{op}} .$$

Here, $\| \cdot \|_{\mathrm{op}}$ denotes the operator norm and the convergence in the last step follows from Proposition 5.1. Additionally, $\phi \mapsto \|\phi\|_{\Lambda^{-1}}$ is $\|\Lambda^{-1/2}\|_{\mathrm{op}}$-Lipschitz continuous. Therefore, by combining the above Lipschitz continuity results with Eq. (21), we have

$$\max_{\phi \in \mathcal{A}} |M_N(\phi) - \beta \|\phi\|_{\Lambda^{-1}}| \le \max_{\psi \in \mathcal{A}_\varepsilon} |M_N(\psi) - \beta \|\psi\|_{\Lambda^{-1}}| + O(\varepsilon) \xrightarrow{\mathbb{P}} O(\varepsilon) \,.$$

Finally, define $F_N(\phi) \coloneqq \phi^\top \theta + M_N(\phi)$ and $F(\phi) \coloneqq \phi^\top \theta + \beta \|\phi\|_{\Lambda^{-1}}$. We have

$$\left| \max_{\phi \in \mathcal{A}} F_N(\phi) - \max_{\phi \in \mathcal{A}} F(\phi) \right| \le \max_{\phi \in \mathcal{A}} |F_N(\phi) - F(\phi)| \le \max_{\psi \in \mathcal{A}_\varepsilon} |M_N(\psi) - \beta \|\psi\|_{\Lambda^{-1}}| + O(\varepsilon) \xrightarrow{\mathbb{P}} O(\varepsilon) \,.$$

By letting $\varepsilon \to 0$, we obtain the desired result. □

C Additional Experimental results

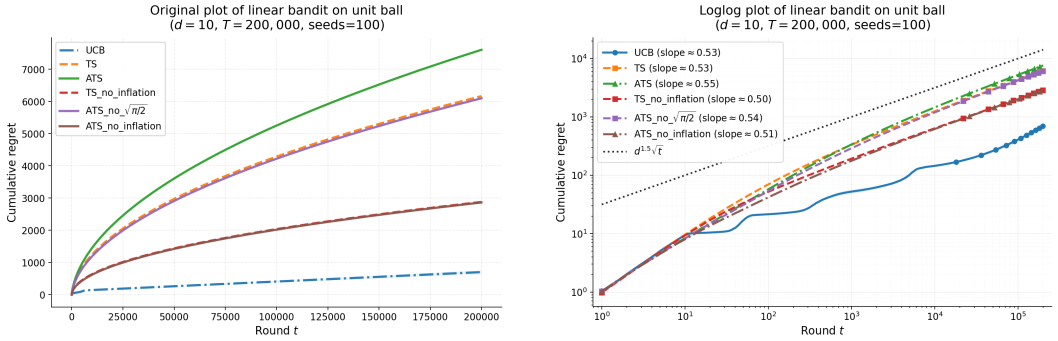

Figure 7: Numerical results for linear bandits on the unit ball with feature dimension $d = 10$, horizon $T = 200000$, $\|\theta^\star\| = 1$, and noise $\eta_t \sim \mathcal{N}(0, 0.1)$. The parameter $\theta^\star$ is sampled uniformly from the unit sphere, and regret is averaged over 100 independent seeds. *(Left)* Regret curves of UCB, TS, TS with no inflation, ATS, ATS with no $\sqrt{\pi/2}$ factor and ATS and the right column shows log-log plots of cumulative regret. In the right column, the black dotted line is $R_{\text{ref}}(t) = d^{1.5}\sqrt{t}$ for reference.

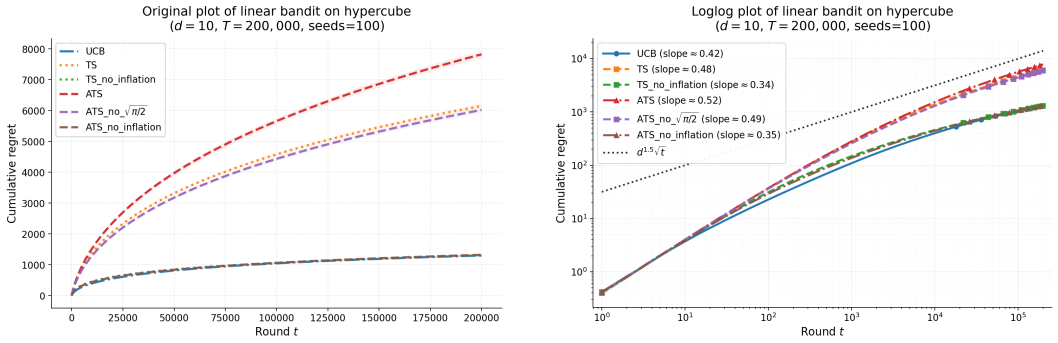

Figure 8: Numerical results for linear bandits on the hypercube with feature dimension $d = 10$, horizon $T = 200000$, $\|\theta^\star\| = 1$, and noise $\eta_t \sim \mathcal{N}(0, 0.1)$. The parameter $\theta^\star$ is sampled uniformly from the unit sphere, and regret is averaged over 100 independent seeds. *(Left)* Regret curves of UCB, TS, TS-no-inflation, ATS, ATS-no-inflation, ATS-no-$\sqrt{\pi/2}$. *(Right)* Log-log plots of the regret curves. The black dotted line is $R_{\text{ref}}(t) = d^{1.5}\sqrt{t}$ for reference.

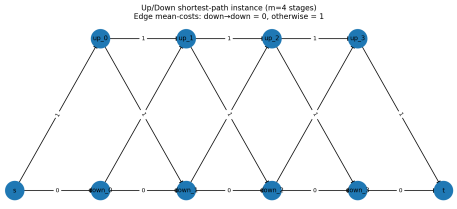

Figure 9: Example of the two-lane shortest path problem with $m = 4$