Efficient Contextual Bandit Learning via Reward-Space Sampling and Online Optimization

Egor Suraveikin , Dastan Omirzak , Roman Sultimov , Yury Maximov

Abstract

The contextual multi-armed bandit problem underlies applications in recommendations, e-commerce, finance, and healthcare, where balancing exploration and exploitation is critical. While algorithms such as Upper Confidence Bound (UCB) and Thompson Sampling (TS) achieve strong theoretical guarantees, they often incur heavy computational cost from high-dimensional parameter estimation. We propose a new approach that combines *reward sampling* with online stochastic optimization. At each round, the algorithm samples hypothetical rewards for all actions and selects the action with the largest draw; the observed reward then updates the model via stochastic optimization. This design is both simple and efficient, preserving exploration while avoiding the pitfalls of greedy behavior on near-duplicate arms. Across synthetic and real-world datasets, our method attains near-optimal reward more quickly and with substantially lower computation than TS and UCB, demonstrating that sampling directly in reward space can improve both statistical efficiency and scalability. Importance sampling can further boost scalability.

1 Introduction

In the contextual bandit problem, an agent repeatedly observes a context and selects an action (arm) to maximize cumulative reward. Only the chosen arm context-dependent reward is observed, highlighting the exploration–exploitation trade-off between *exploiting* the current best estimate and *exploring* potentially better alternatives. Contextual bandits power applications such as personalized recommendations and online advertising. Classical methods, UCB and Thompson Sampling (TS), offer strong guarantees but depend on posterior estimation or confidence bounds, which can be computationally expensive. The stochastic *linear* contextual bandit is well studied. LinUCB and linear TS achieve sublinear regret and strong empirical performance but scale poorly with dimension, limiting practicality in high-dimensions.

Novelty and contribution. We propose a sampling scheme that draws directly from the reward vector rather than estimating the full arm distribution. Combined with stochastic optimization, it enables efficient policy updates, speeds convergence, and avoids getting trapped in suboptimal arms. The gains are most pronounced in ill-posed problems common in practice, as demonstrated across multiple benchmarks.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

2 Problem Setup and Related Work

We consider the standard stochastic contextual bandit framework. At each round $t \in \{1, \dots, T\}$, the agent observes a context $x_t \in \mathcal{X} \subset \mathbb{R}^d$, selects an arm $a_t \in \{1, \dots, K\}$, and receives a reward $r_t \sim D_{(x_t, \theta_t)}$ depending on both x_t and an arm-specific latent parameter θ_t. Rewards are modeled as $r_t(x_t, a_t) = \mu(x_t, a_t) + \eta_t$, where μ is the expected reward function and η_t is zero-mean noise with bounded variance. A policy π maps each x to a distribution over arms, and the goal is to maximize the expected average reward $\bar{r}_T = T^{-1} \sum_{t=1}^{T} \mathbb{E}[r_t]$.

Linear bandits assume $\mu(x, a) = x^\top \theta$, while generalized linear bandits (GLMs) use $\mu(x, a) = f(\theta^\top x)$. GLM-UCB (Filippi et al. 2010) achieves $\tilde{O}(\sqrt{T})$ regret but requires costly maximum likelihood updates, and TS for GLMs relies on expensive posterior sampling. To improve scalability, SGD-TS (Ding et al. 2021) combines online SGD with TS, reducing per-round cost to $O(d)$. Other strategies include feature hashing (Jun et al. 2017), reduced exploration (Bastani, Bayati, and Khosravi 2020), and information-directed sampling (Russo and VanRoy 2014).

Nonlinear bandits allow richer models $r = f(x, \theta)$. NeuralUCB (Zhou et al. 2020) and NeuralTS (Zhang et al. 2021) established the first regret bounds for neural bandits, with extensions via neural tangent kernels (Kassraie et al. 2022). Neural methods remain computationally demanding. Recent work on neural-linear (Xu, Zhou, and Jin 2022), bootstrapping (Wan et al. 2023), batch TS (Shu et al. 2022) improves efficiency; balancing expressiveness vs scalability is open.

3 Algorithm

In practice, arms (context–arm space) can be *degenerate*: rarely observed, near–collinear, or coupled to saturated rewards. Classical optimism may prematurely overcommit, essentially converting TS and UCB to greedy algorithms. The result is poor coverage (self–reinforcing data imbalance), brittle confidence estimates, and unfair exposure.

Motivating example. Consider a two-arm linear bandit with features $\phi_1 = [1, \varepsilon]^\top$, $\phi_2 = [1, -\varepsilon]^\top$ for small $\varepsilon > 0$ and true parameter $\theta^\star = (\theta_1^\star, \theta_2^\star)^\top$ with $\theta_2^\star > 0$. The expected rewards are $\mathbb{E}[r|a = 1] = \theta_1^\star + \varepsilon\theta_2^\star$, $\mathbb{E}[r|a = 2] = \theta_1^\star - \varepsilon\theta_2^\star$, so arm 1 is optimal with gap $\Delta^\star = (\phi_1 - \phi_2)^\top \theta^\star = 2\varepsilon\,\theta_2^\star$.

Suppose the learner maintains a Gaussian posterior $\theta \sim$

Algorithm 1: GLM: Reward Sampling

Require: Number of rounds T, arms K, learning rate η

Initialize shared weights $\theta \in \mathbb{R}^{K \cdot d}$
for $t = 1$ **to** T **do**
  for each arm $a \in \{1, \dots, K\}$ **do**
    Get extended feature vector $\phi(x_{t,a}, a) \in \mathbb{R}^{K \cdot d}$
    Update mean: $\mu_{t,a} \leftarrow g\big(\phi(x_{t,a}, a)^\top \theta\big)$
    Sample reward: $\tilde{r}_{t,a} \sim \mathcal{N}(\mu_{t,a}, \sigma^2)$
  end for
  Select arm: $a_t \leftarrow \arg\max_a \tilde{r}_{t,a}$
  Observe reward r_t; update $\theta \leftarrow \theta - \eta \nabla L(\theta)$
end for

$\mathcal{N}(\mu, \Sigma)$. With TS, both arm scores are evaluated at the same sampled parameter $phi_a^\top \theta$, with $\theta \sim \mathcal{N}(\mu, \Sigma)$, so the score difference is Gaussian with variance $V_{\mathrm{TS}} = \phi_1^\top \Sigma \phi_1 + \phi_2^\top \Sigma \phi_2 - 2\,\phi_1^\top \Sigma \phi_2$. As $\varepsilon \to 0$, the features become collinear, $\phi_1 \approx \phi_2$, and thus $V_{\mathrm{TS}} \to 0$: TS behaves greedily and rarely reorders the arms if the current mean gap $\Delta = \phi_1^\top \mu - \phi_2^\top \mu$ has the wrong sign.

With *reward sampling (RS)*, each arm is drawn independently from its predictive distribution, $\tilde{r}_a \sim \mathcal{N}(\phi_a^\top \mu,\ \phi_a^\top \Sigma \phi_a + \sigma^2)$, independently for $a = 1, 2$, so the difference variance is $V_{\mathrm{RS}} = \phi_1^\top \Sigma \phi_1 + \phi_2^\top \Sigma \phi_2 + 2\sigma^2 > 0$ even as $\varepsilon \to 0$. Hence the probability of selecting arm 1 is $p_{\mathrm{RS}} = \Phi\big(\Delta / \sqrt{V_{\mathrm{RS}}}\big) \gg p_{\mathrm{TS}} = \Phi\big(\Delta / \sqrt{V_{\mathrm{TS}}}\big)$, and RS keeps exploring both arms until the gap is corrected. Similarly, UCB selects $\arg\max_a \phi_a^\top \mu + \beta \sqrt{\phi_a^\top \Sigma \phi_a}$. When $\phi_1 \approx \phi_2$ the confidence widths are nearly equal, so the index difference reduces to Δ, and UCB is fully greedy if $\Delta < 0$. In this bandit, RS strictly dominates TS and UCB in the near-duplicate regime by ensuring better per-arm exploration.

Scalability and time complexity. Coupled with reward sampling, stochastic gradient descent improve scalability of traditional methods requiring about $O(d)$ update time (Mei et al. 2023). Alg. 1 contains necessary details for generalized linear models. Similar approach work for linear and neural bandits.

4 Experiments

Synthetic data: linear case. We simulate a linear bandit with K arms and context dimension d, $K, d = 1'000$. The true parameter $\theta^* \in \mathbb{R}^d$ is drawn from $\mathcal{N}(0, I)$. Each context vector $x_{t,a}$ is drawn i.i.d. from $\mathcal{N}(0, I)$. Rewards are generated as $r_{t,a} = \theta^{*\top} x_{t,a} + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 0.01)$ added for noise. We run each algorithm for $T = 10'000$ rounds and average results over 20 random trials.

Real data: non-linear case. We use a set of 516 datasets from (Bietti, Agarwal, and Langford 2021) and follow their replay protocol to estimate cumulative bandit reward and regret. Algorithms are trained with reward sampling and online SGD.

Across datasets, Reward Sampling (RS) matches or outperforms TS and UCB, with largest gains when actions have near-duplicate features or when predictive uncertainties are highly correlated. Unlike UCB, which can act greedily under similar confidence widths, RS maintains exploration and achieves lower regret. Tab. 1 summarizes the empirical evaluation results.

Method/Regret at T, sec.	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1
TS Baseline	4.77	4.67	4.48	4.45	5.25
LinTS	6.12	4.90	5.06	4.01	3.40
Reward Sampling (Linear)	**3.97**	**3.78**	**1.82**	**0.73**	**0.67**
SGD-TS	5.59	4.44	**4.22**	4.07	3.86
Reward Sampling (TS)	**4.70**	4.62	4.72	**3.76**	**3.28**

Table 1: RS with SGD outperform (regret vs time, sec.) the state-of-the-art methods for both linear and non-linear cases.

References

Bastani, H.; Bayati, M.; and Khosravi, K. 2020. Mostly exploration-free algorithms for contextual bandits. *Management Science*.

Bietti, A.; Agarwal, A.; and Langford, J. 2021. A contextual bandit bake-off. *JMLR*, 22(133): 1–49.

Ding, W.; Qi, Y.; Lattimore, T.; Zou, J.; and Kpotufe, S. 2021. Provably Efficient Online Thompson Sampling with Linear Payoffs via Stochastic Gradient Descent.

Filippi, S.; Cappé, O.; Garivier, A.; and Szepesvári, C. 2010. Parametric bandits: The generalized linear case. In *NeurIPS*.

Jun, K. S.; Bhargava, A.; Nowak, R.; and Willett, R. 2017. Scalable generalized linear bandits: Online computation and hashing. In *NeurIPS*, 99–109.

Kassraie, P.; Kakade, S. M.; Lee, J. D.; and Ohannessian, M. I. 2022. Neural Contextual Bandits Without Regret. In *AISTATS*.

Mei, J.; Zhong, Z.; Dai, B.; Agarwal, A.; Szepesvari, C.; and Schuurmans, D. 2023. Stochastic gradient succeeds for bandits. In *ICML*, 24325–24360. PMLR.

Russo, D.; and VanRoy, B. 2014. Learning to optimize via information-directed sampling. In *NeurIPS*, 1583–1591.

Shu, Y.; Pan, Y.; Wang, Z.; Balandat, M.; and Eriksson, D. 2022. Thompson Sampling for Batch Deep Contextual Bandits. In *NeurIPS*.

Wan, R.; Li, J.; Liu, H.; Wang, L.; and Chen, W. 2023. Multiplier Bootstrap-based Exploration in Contextual Bandits. In *ICML*.

Xu, P.; Zhou, D.; and Jin, R. 2022. Neural Contextual Bandits with Deep Representation and Shallow Exploration. In *ICLR*.

Zhang, W.; Zhou, D.; Tang, Y.; and Jin, R. 2021. Neural Thompson Sampling. In *ICLR*.

Zhou, D.; Tang, Y.; Ren, Z.; Lyu, Q.; Wang, Y.; and Jin, R. 2020. Neural Contextual Bandits with UCB-Based Exploration. In *ICML*.