Accelerating Approximate Thompson Sampling with Underdamped Langevin Monte Carlo

Haoyang Zheng*, Wei Deng†, Christian Moya*, Guang Lin*

*Purdue University, West Lafayette, IN 47906, USA

†Morgan Stanley, New York, NY 10036, USA

PURDUE MECHANICAL ENGINEERING

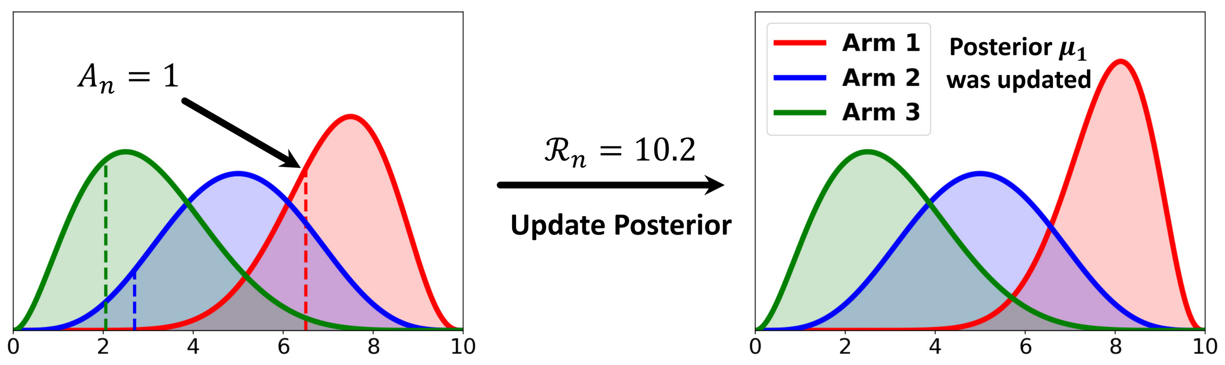
1. Introduction

Questions

- Can we avoid using Gaussian approximations for the posterior?
- How can we improve the performance of approximate Thompson Sampling?

Problem Formulations

- Time horizon N and indices $n \in \{1, 2, \ldots, N\}$
- The agent plays arm $A_n \in \mathcal{A} = \{1, 2, \ldots, K\}$ and receives rewards $\mathcal{R}_n$
- Update the posterior μ of arm A_n according to reward $\mathcal{R}_n$
- Total expected regret $\mathbb{E}[\Re(N)] = N \max_a \mathbb{E}[\mu_a] - \mathbb{E}\left[\sum_{n=1}^N \mathcal{R}_n\right]$
- Goal: find the optimal policy (arm) to minimize the regret



Contributions

- We proposed a computationally efficient **Thompson Sampling** algorithm with **underdamped Langevin Monte Carlo**
- We derive novel **posterior concentration** with a designed potential function
- With the novel posterior concentration rates, we prove it achieves $\tilde{O}\left(\frac{\log N}{\Delta_a}\right)$ regrets with $\tilde{O}(\sqrt{d})$ samples (previously with $\tilde{O}(d)$ samples)
- Both theoretical analysis and experimental results are provided

Assumptions

- Lipschitz Smooth and Strongly Convex on the log-likelihood functions $\log \mathbb{P}_a(\mathcal{R}|x)$, $x \in \mathbb{R}^d, \mathcal{R} \in \mathbb{R}$.
- Lipschitz Smooth and Strongly Convex on the reward distributions.
- Lipschitz Smooth on the priors $\pi_a(x)$.
- Joint Lipschitz Smooth on the log-likelihood functions (for Approximate Thompson Sampling).

2. Proposed Algorithms

Thompson Sampling Algorithm

1: **Input** Posteriors $\mu_a[\rho_a]$ and feature vectors α_a for $\forall a \in \mathcal{A}$.
2: **for** n=1 to N **do**
3: Sample $(x_{a,n}, v_{a,n}) \sim \mu_a[\rho_a]$ for $\forall a \in \mathcal{A}$.
4: Choose arm $A_n = \operatorname{argmax}_{a\in\mathcal{A}} \langle \alpha_a, x_{a,n}\rangle$.
5: Play arm A_n and receive reward $\mathcal{R}_n$.
6: Update posterior distribution of arm A_n: $\mu_{A_n}[\rho_{A_n}]$.
7: Calculate regret at round n: $\Re_n$.
8: **end for**
9: **Output** Total regrets $\Re(N) = \sum_{n=1}^N \Re_n$.

(Stochastic Gradient) Underdamped Langevin Monte Carlo

1: **Input** Data $\{\mathcal{R}_{a,1}, \cdots, \mathcal{R}_{a,\mathcal{L}_a(n-1)}\}$ and Sample $\left(x_{a,Ih^{(n-1)}}, v_{a,Ih^{(n-1)}}\right)$.
2: Initialize $x_0 = x_{a,Ih^{(n-1)}}$ and $v_0 = v_{a,Ih^{(n-1)}}$.
3: **for** $i = 0, 1, \cdots, I$ **do**
4: Uniformly subsample $\mathcal{S} \subseteq \{\mathcal{R}_{a,1}, \cdots, \mathcal{R}_{a,k}\}$.
5: Compute (stochastic) gradient $\nabla U(x_i)$.
6: Sample (x_{i+1}, v_{i+1}) based on $\nabla U(x_i)$.
7: **end for**
8: $x_{a,Ih^{(n)}} \sim \mathcal{N}\left(x_I, \frac{1}{nL_a\rho_a}\mathbf{I}_{d\times d}\right)$ and $v_{a,Ih^{(n)}} = v_I$
9: **Output** Sample $(x_{a,Ih^{(n)}}, v_{a,Ih^{(n)}})$ from current round.

Stochastic Differential Equations (SDEs)

- $dv_t = \underbrace{-\frac{\gamma}{2}v_t dt}_{\text{friction term}} \underbrace{-\frac{u}{2}\nabla f(x_t)\,dt}_{\text{log-likelihood}} \underbrace{-\frac{u}{2n}\nabla \log \pi(x_t) dt}_{\text{prior}} + \underbrace{\sqrt{\frac{\gamma u}{n\rho}}dB_t}_{\text{noise term}}.$
- $dx_t = v_t dt.$ **Notes:** $x_t \in \mathbb{R}^d$ are positions, and $v_t \in \mathbb{R}^d$ are velocities.

Discretization Scheme

- Sample $\begin{bmatrix} x_{i+1} \\ v_{i+1}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbb{E}[x_{i+1}] \\ \mathbb{E}[v_{i+1}]\end{bmatrix}, \begin{bmatrix}\mathbb{V}(x_{i+1}) & \mathbb{K}(x_{i+1}, v_{i+1}) \\ \mathbb{K}(v_{i+1}, x_{i+1}) & \mathbb{V}(v_{i+1})\end{bmatrix}\right)$ as follows:
- $\mathbb{E}[v_{i+1}] = v_i e^{-\gamma h} - \frac{u}{\gamma}(1 - e^{-\gamma h})\nabla U(x_i)$
- $\mathbb{E}[x_{i+1}] = x_i + \frac{1}{\gamma}(1 - e^{-\gamma h})v_i - \frac{u}{\gamma}\left(h - \frac{1}{\gamma}\left(1 - e^{-\gamma h}\right)\right)\nabla U(x_i)$
- $\mathbb{V}(x_{i+1}) = \frac{2u}{\gamma}\left[h - \frac{1}{2\gamma}e^{-2\gamma h} - \frac{3}{2\gamma} + \frac{2}{\gamma}e^{-\gamma h}\right] \cdot \mathbf{I}_{d\times d}$
- $\mathbb{V}(v_{i+1}) = u(1 - e^{-2\gamma h}) \cdot \mathbf{I}_{d\times d}$
- $\mathbb{K}(x_{i+1}, v_{i+1}) = \frac{u}{\gamma}\left[1 + e^{-2\gamma h} - 2e^{-\gamma h}\right] \cdot \mathbf{I}_{d\times d}$

Gradient Estimation

- Exact gradient: $\nabla U(x_i) = -\sum_j \nabla \log \mathbb{P}_a\left(\mathcal{R}_j|x_i\right) - \nabla \log \pi_a(x_i)$
- Stochastic gradient: $\nabla \tilde{U}(x_i) = -\frac{\mathcal{L}_a(n)}{|\mathcal{S}|}\sum_{\mathcal{R}\in\mathcal{S}} \nabla \log \mathbb{P}_a(\mathcal{R}_k|x_i) - \nabla \log \pi_a(x_i)$

3. Theoretical Analysis

Posterior Concentration

For $x \in \mathbb{R}^d$ and $\delta_1 \in \left(0, e^{-\frac{1}{2}}\right)$, the posterior distribution of SDEs satisfies:

$$\mathbb{P}_{x\sim\mu_a[\rho_a]}\left[\|x - x_*\|_2 \geq \sqrt{\frac{2e}{mn}\left(D + 2\Omega \log \frac{1}{\delta_1}\right)}\right] \leq \delta_1,$$

where $D = 8d/\rho + 2\log B$, $\Omega = 256/\rho + 16\kappa^2 d$.

Regrets of Exact Thompson Sampling

With $\rho_a = \kappa_a^{-3}(8d)^{-1}$, constants C_1 and C_a can upper-bound the expected regret after N rounds of exact Thompson Sampling:

$$\mathbb{E}[\Re(N)] \leq \sum_{a>1}\left[\frac{C_1}{\Delta_a}\sqrt{B_1}\left(\log B_1 + d^2\right) + \frac{C_a}{\Delta_a}\left(\log B_a + d^2 + d\log N\right) + 2\Delta_a\right].$$

(Approximate) Posterior Concentration

With the choice of step size $h^{(n)} = \tilde{O}\left(1/\sqrt{d}\right)$ and number of steps $N = \tilde{O}\left(\sqrt{d}\right)$ in ULMC, the following inequality holds when $x_{a,t} \in \mathbb{R}^d$ and $\delta_2 \in \left(0, e^{-\frac{1}{2}}\right)$:

$$\mathbb{P}_{x_{a,t}\sim\tilde{\mu}_a^{(n)}[\tilde{\rho}_a]}\left[\|x_{a,t} - x_*\|_2 \geq 6\sqrt{\frac{e}{m_a n}\left(D_a + 2\Omega_a \log\frac{1}{\delta_1} + 2\tilde{\Omega}_a \log\frac{1}{\delta_2}\right)}\right] \leq \delta_2,$$

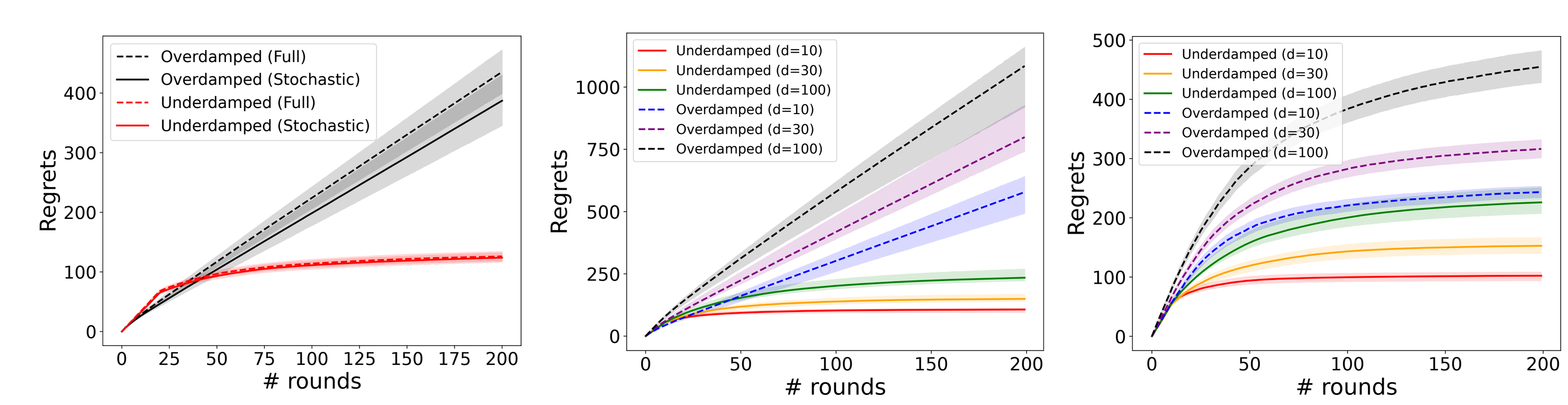
where $D_a = 2\log B_a + 8d$, $\Omega_a = 256 + 16d\kappa_a^2$, and $\tilde{\Omega}_a = 256 + 16d\kappa_a^2 + d/18\kappa_a\tilde{\rho}_a$.

Regrets of Approximate Thompson Sampling

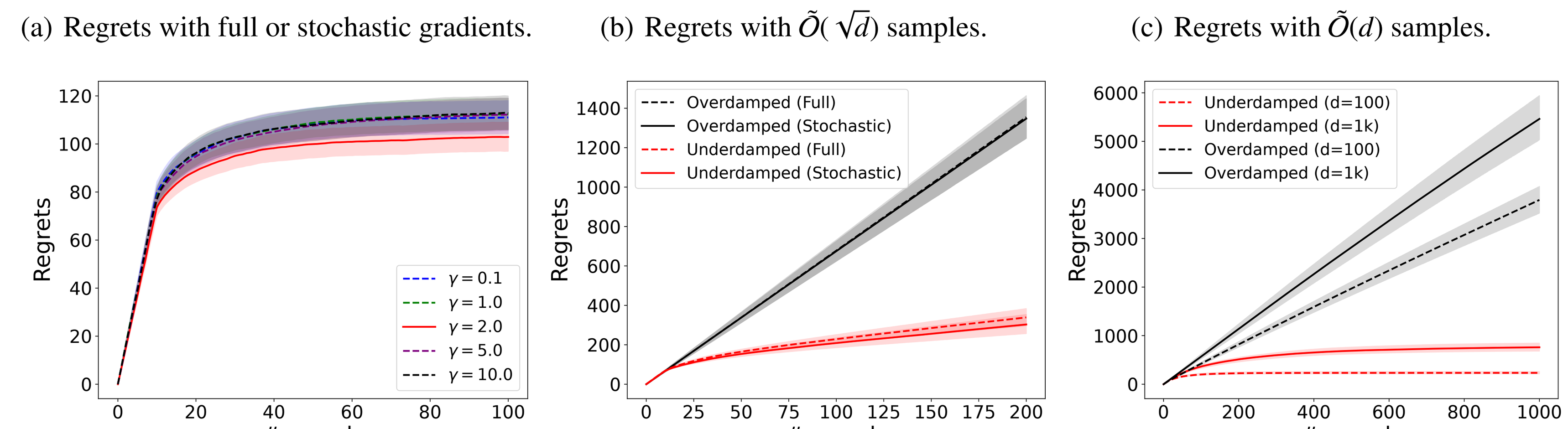
Given $\hat{\rho}_a = (8\kappa_a\Omega_a)^{-1}$, the total expected regrets after N rounds of approximate Thompson sampling are capped by constants $\tilde{C}_1$ and $\tilde{C}_a$:

$$\mathbb{E}[\Re(N)] \leq \sum_{a>1}\left[\frac{\tilde{C}_1\sqrt{B_1}}{\Delta_a}\left(\log B_1 + d^2\kappa_1^2 \log N\right) + \frac{\tilde{C}_a}{\Delta_a}\left(\log B_a + d^2\kappa_a^2 \log N\right) + 4\Delta_a\right].$$

4. Experimental Results



(a) Regrets with full or stochastic gradients. (b) Regrets with $\tilde{O}(\sqrt{d})$ samples. (c) Regrets with $\tilde{O}(d)$ samples.



(d) Regrets with different momentums γ. (e) Regrets with flat priors. (f) Regrets in high dimensional cases.

Take-away

- With the proper choice of batch size, approximate Thompson Sampling can achieve sub-linear regrets with stochastic gradients.
- Thompson Sampling with underdamped Langevin algorithms attains sub-linear regrets at $\tilde{O}\left(\sqrt{d}\right)$ samples, while previous requires $\tilde{O}(d)$ samples.
- Selecting $\gamma = 2.0$ achieves the lowest regrets (consistent with our analysis).
- The improvements become prominent as d increases.

5. Related Works

- Cheng, Xiang, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. *Underdamped Langevin MCMC: A non-asymptotic analysis*, Conference on learning theory, pp. 300-323. PMLR, 2018.
- Mazumdar, Eric, Aldo Pacchiano, Yian Ma, Michael Jordan, and Peter Bartlett. *On approximate Thompson sampling with Langevin algorithms*, International Conference on Machine Learning, pp. 6797-6807. PMLR, 2020.