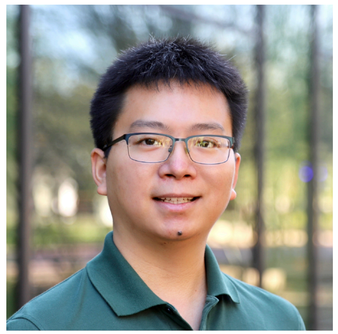
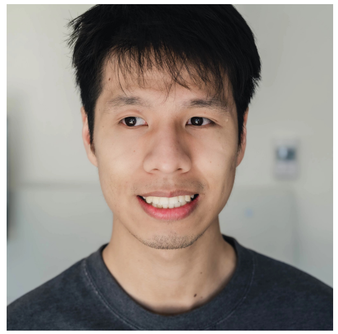
Langevin Monte Carlo for Contextual Bandits



Pan Xu



Hongkai Zheng

Eric Mazumdar

Kamyar Azizzadenesheli

Anima Anandkumar

Caltech

PURDUE UNIVERSITY

Contextual Bandit Problem

- At round t, receives an arm set $\mathcal{X}_t = \{x_{t,1}, \ldots, x_{t,K}\} \subseteq \mathbb{R}^d$
- Reward of selecting arm $x \in \mathcal{X}_t$: $r(x) = f(x, \theta^*)$+subGaussian noise
- Cumulative reward for T rounds: $\sum_{t=1}^{T} r(x_{t,a_t})$, $a_t \in \{1, \ldots, K\}$ is the arm chosen by the algorithm at round t
- **Goal**: **minimize the cumulative regret** due to not knowing the best arm at each round

$$R_\mu(T) = \mathbb{E}\left[\sum_{i=1}^{T} \max_{k=1,\ldots,K} f(x_{t,k}, \theta^*) - \sum_{t=1}^{T} r(x_{t,a_t})\right]$$

best arm at round t — the actually chosen arm

Exploration: construct an estimate $\hat{\theta}$ for the reward generating function parameter θ^*

Exploitation: choose the arm with the highest reward based on estimation $\hat{\theta}$

LinTS: Thompson Sampling with Laplace Approximation

- Maintain an estimate of the unknown weight parameter: $\hat{\theta} = V^{-1}b$

  V : feature gram/covariance matrix

  b : reward weighted sum of feature vectors

- Construct a Laplace Approximation to the posterior distribution: $\mathcal{N}(\hat{\theta}, cV^{-1})$

1: **for** $t = 1,2,\ldots,T$
2: sample $\theta \sim \mathcal{N}(\hat{\theta}, cV^{-1})$
3: choose arm $a = \arg\max_{k \in [K]} \theta^\top x_k$
4: pull arm a, receive noisy reward r
5: update $\hat{\theta} = V^{-1}b$, where $V \leftarrow V + x_a x_a^\top, \quad b \leftarrow b + r x_a$
6: **end**

- The Laplace approximation is not a good estimation for the posterior distribution when the reward distribution has more general forms than linearity.
- Sampling from a Gaussian distribution with general covariance matrix in high dimensional problems is computationally inefficient

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In International Conference on Machine Learning, pp. 127–135. PMLR, 2013.

LMC-TS: Langevin Monte Carlo Thompson Sampling

- Define a general loss function

$$L_t(\theta) = \sum_{i=1}^{t} \left(f(x_i, \theta) - r_i\right)^2 + \lambda \|\theta\|^2$$

  - f could be any bandit model

- Langevin Monte Carlo update:
  $\theta \leftarrow \theta - \eta \nabla L_t(\theta) + \sqrt{2\eta/\beta}\epsilon$
  - It approximately samples from $\pi_t \propto \exp(-\beta L_t(\theta))$
  - When f is linear $\pi_t = \mathcal{N}(\hat{\theta}, \beta^{-1} V^{-1})$

1: **for** $t = 1,2,\ldots,T$
2: update L_t based on historical data
3: **for** $m = 1,\ldots,M$
4: sample $\epsilon \sim \mathcal{N}(0,I)$
5: $\theta \leftarrow \theta - \eta \nabla L_t(\theta) + \sqrt{2\eta/\beta}\epsilon$
6: **end**
7: choose arm $a = \arg\max_{k \in [K]} \theta^\top x_k$
8: pull arm a, receive noisy reward r
9: **end**

- LMC-TS approximately samples from the true posterior distribution
- LMC-TS is computationally efficient due to
  - it only needs to sample from isotropic Gaussian $\mathcal{N}(0,I)$
  - it only needs to perform noisy gradient descent updates

Regret Analysis for Linear Contextual Bandits

Linear contextual bandits: $r(x) = \theta^{*\top} x + \xi$ for any arm feature $x \in \mathbb{R}^d$

- θ^* is an unknown weight parameter shared across all arms
- ξ is a sub-Gaussian noise

Theorem: For the linear contextual bandit model, if we choose parameters of **LMC-TS** as follows:

- step size $\eta = 1/(4\lambda_{\max}(V))$,
- inner loop length $M = \lambda_{\max}(V)/\lambda_{\min}(V)\log(dT\log(T^3/\delta))$,
- and temperature parameter $1/\beta = 4(d\log(T^3/\delta))^{1/2}$.

Then with probability at least $1-\delta$, the regret of LMC-TS satisfies

$$R(T) = O\Big(d\log(1/\delta)\sqrt{dT\log^3(1+T/(\lambda d))}\Big)$$

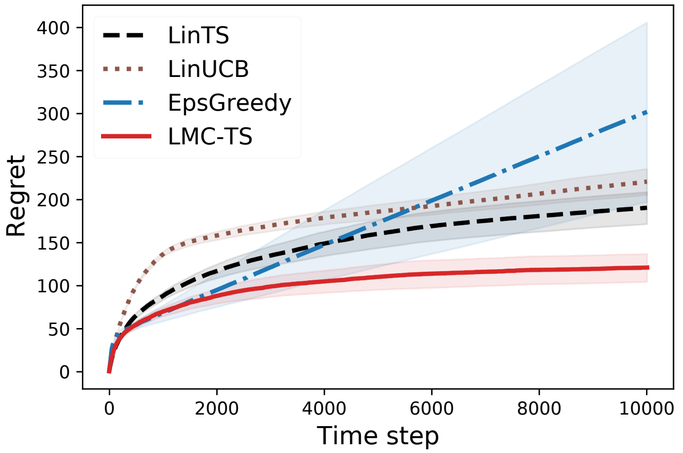
- The regret bound $\tilde{O}(d\sqrt{d})$ matches that of the best results for linear Thompson sampling (LinTS)

Experiments: Simulated Bandit Problems

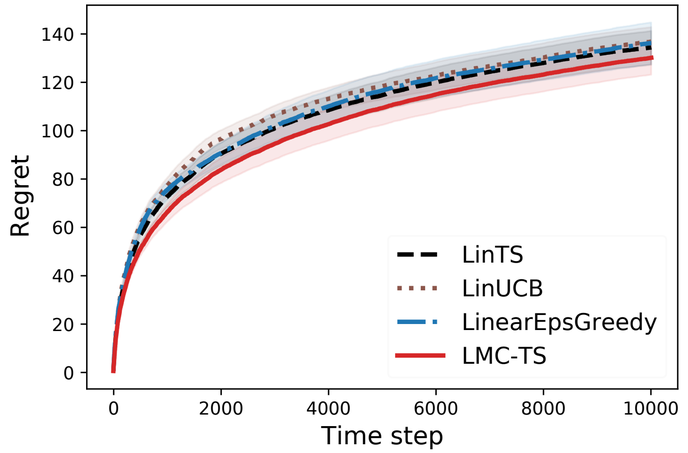
Dataset generation:

$r(x) = f(x, \theta^*) +$ Gaussian noise

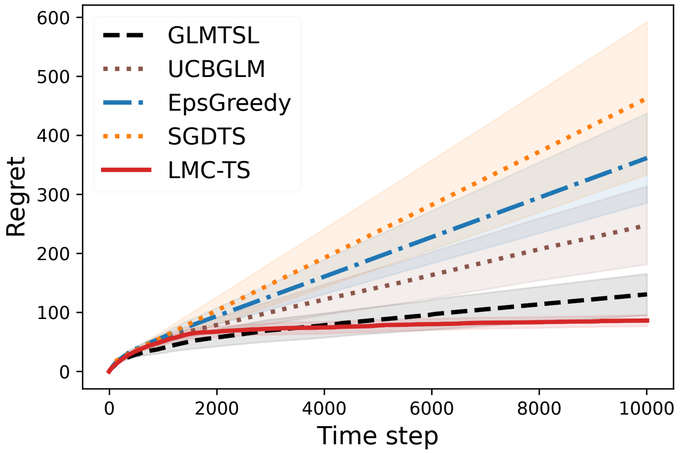
- *setting 1*: $f(x, \theta^*) = x^\top \theta^*$, arm set is fixed throughout the experiments
- *setting 2*: $f(x, \theta^*) = x^\top \theta^*$, arm set is varying at each round
- *setting 3*: $f(x, \theta^*) = 1/(1 + \exp(-x^\top \theta^*))$
- *setting 3*: $f(x, \theta^*) = 10(x^\top \theta^*)^2$, arm set is fixed throughout the experiments



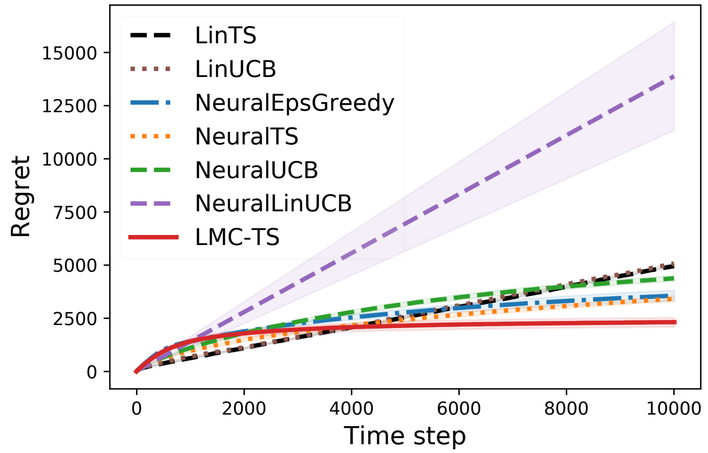
1. linear bandit (fixed arm set)



2. linear bandit (varying arm set)



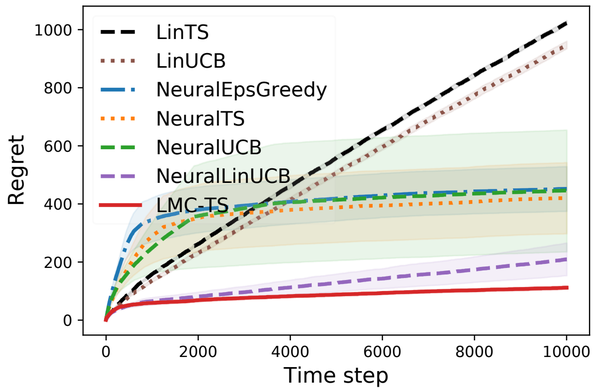
3. logistic (generalized linear) bandit



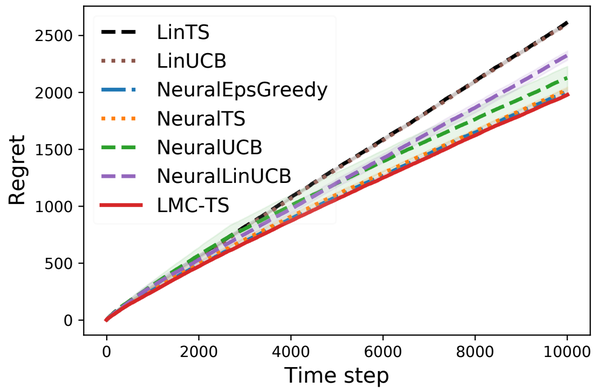
4. quadratic bandit

Experiments: Real-world Datasets

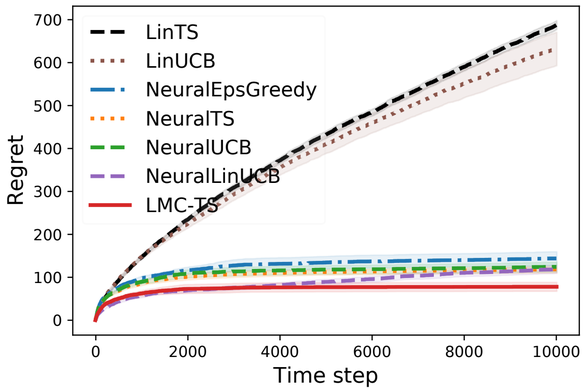
- UCI machine learning datasets: Shuttle, Magic, Mushroom, Covertype
- High dimensional image dataset: CIFAR10



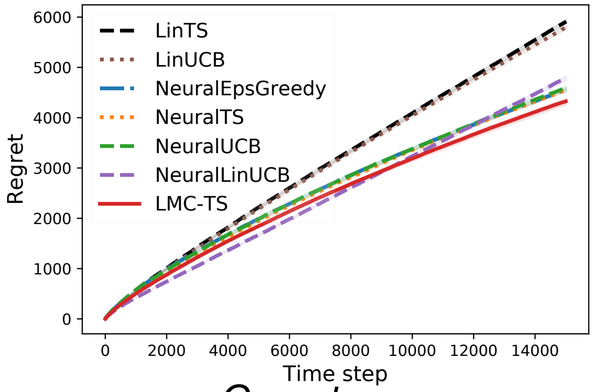
Shuttle



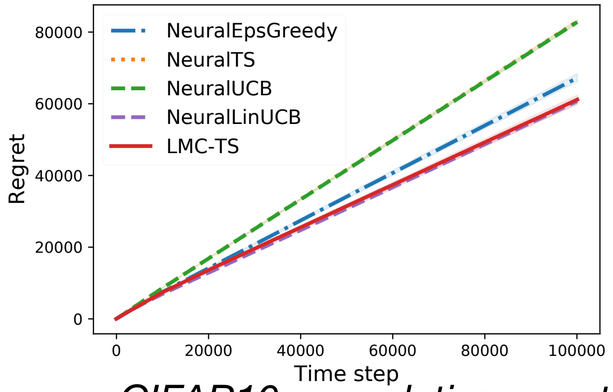
Magic



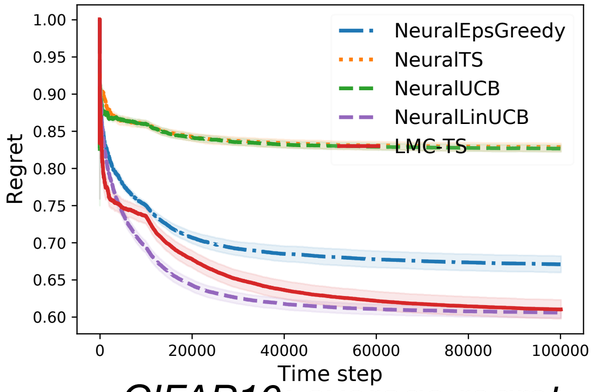
Mushroom



Covertype



CIFAR10: cumulative regret



CIFAR10: average regret

Summary

ICML
International Conference
On Machine Learning

- We propose LMC-TS which only needs to perform noisy gradient descent updates to approximately sample from the data posterior distribution.
- We prove LMC-TS achieves the same regret guarantee as the best Thompson sampling algorithm LinTS for linear contextual bandits.
- We conduct experiments to show that one algorithm (LMC-TS) is enough for learning a broad class of bandit models including linear contextual bandits, generalized bandits, and neural contextual bandits.