Simple Regret Minimization for Contextual Bandits

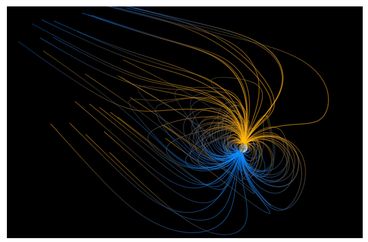
Aniket Deshmukh, Srinagesh Sharma, James W. Cutler, Mark Moldwin, Clayton Scott

University of Michigan, Ann Arbor

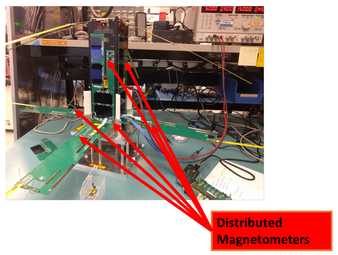
Introduction

There are two variants of the classical multi-armed bandit (MAB) problem that have received considerable attention from machine learning researchers in recent years: contextual bandits and simple regret minimization. Contextual bandits are a sub-class of MABs where, at every time step, the learner has access to side information that is predictive of the best arm. Simple regret minimization assumes that the learner only incurs regret after a pure exploration phase. In this work, we study simple regret minimization for contextual bandits. Our experiments examine a novel application to adaptive sensor selection for magnetic field estimation in interplanetary spacecraft, and demonstrate considerable improvement over algorithms designed to minimize the cumulative regret.

Magnetometer Interference Cancellation



(a) Scientific Measurement (Credit: NASA/Goddard Scientific Visualization Studio)



(b) Small Satellite with Multiple Magnetometers (TBEX)

Figure: Interplanetary Magnetic Fields

Adaptive Sensor Selection

- Due to power constraints, can only record from a sensor at at time
- Accuracy is reflected by a loss (More on loss later)
- Best sensor depends on state of satellite: Contextual bandits
- Simple regret: difference of loss incurred and optimal loss at a particular time
- Goal is to minimize the simple regret

Exploration and Exploitation Phases

- Challenge: Computation of loss requires knowledge of B_{true} as

$$\text{Loss} = \|B_{true} - B_{sensor}\|^2$$

- B_{true} known in certain portions of a satellites orbit
- Revised problem statement:
  - Exploration: B_{true} known, no regret incurred
  - Exploitation: B_{true} not known, regret incurred

Goal: Optimize strategy during exploration phase to minimizing regret during exploitation phase

Contextual Bandits

1: **for** $t = 1, ..., T$ **do**
2: Observe context $x_{a,t} \in \mathbb{R}^d$ for all arms $a \in [N]$, where $[N] = \{1, ...N\}$
3: Choose an arm $a_t \in [N]$
4: Receive a reward $r_{a_t,t} \in \mathbb{R}$ s.t. $\mathbb{E}[r_{a_t,t}|x_{a_t}] = f_{a_t}(x_{a_t})$
5: Improve arm selection strategy based on new observation $(x_{a_t,t}, a_t, r_{a_t,t})$
6: **end for**

Cumulative Regret Goal: Minimize the T-trial regret, $R(T) = \sum_{t=1}^{T} f_{a_t^*}(x_{a_t^*}) - \sum_{t=1}^{T} f_{a_t}(x_{a_t})$.
Simple Regret Goal: Minimize the regret at T, $r(T) = f_{a_T^*}(x_{a_T^*}) - f_{a_T}(x_{a_T})$.

Contribution

	Cumulative Regret	Simple Regret
Multi-armed Bandits	Auer et al. 2002 [1]	Hoffman et al. 2014 [3]
Contextual Bandits	Chu et al. 2011 [2]	**This work**

Our contribution: An algorithm with performance guarantee and experimental results on multiple datasets.

Proposed Algorithm: Contextual Gap

1: Number of arms A, Time Steps T, parameter β, regularization parameter λ, burn-in phase constant N_λ.
2: **for** $t = 1, ..., AN_\lambda$ **do**
3: Observe context x_t.
4: Choose $a_t = t \mod A$.
5: Receive reward $r_t \in \mathbb{R}$
6: **end for**
7: **for** $t = AN_\lambda + 1, \ldots, T$ **do**
8: Observe context x_t.
9: Learn reward estimators $\hat{f}_{a,t}(x_t)$ and confidence estimators $s_{a,t}(x_t)$ based on history.
10: $U_{a,t}(x_t) = \hat{f}_{a,t}(x_t) + \frac{s_{a,t}(x_t)}{2}$, $L_{a,t}(x_t) = \hat{f}_{a,t}(x_t) - \frac{s_{a,t}(x_t)}{2}$.
11: $B_{a,t}(x_t) = \max_{i \neq a} U_{i,t}(x_t) - L_{a,t}(x_t)$.
12: $J_t(x_t) = \text{argmin}_a B_{a,t}(x_t)$, $j_t(x_t) = \text{argmax}_{a \neq J_t(x_t)} U_{a,t}(x_t)$.
13: Choose $a_t = \text{argmax}_{a \in \{j_t(x_t), J_t(x_t)\}} s_{a,t}(x_t)$.
14: Receive reward $r_t \in \mathbb{R}$.
15: **end for**

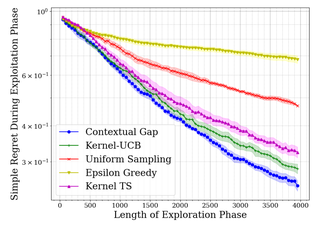
Theoretical Gurantees

Assumptions:

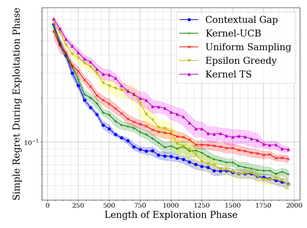
- Contexts x_t are bounded and depend on previous contexts through a filtration.
- Feature map $\phi(x_t)$ maps x_t to a higher dimensional space such that $\mathbb{E}_{t-1}[\phi(x_t)\phi(x_t)^T]$ lies in a finite dimensional space, where $\mathbb{E}_{t-1}[\cdot] := \mathbb{E}[\cdot|\phi(x_1), \phi(x_2), \cdots, \phi(x_{t-1})]$
- Loss/reward is a non-linear function (from RKHS) of context.
- Exploration phase is from time step 1 to T.

Theorem 1 *With high probability, simple regret converges to zero as the number of time steps T during exploration phase increase.*

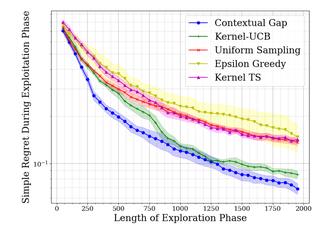
Experimental Results



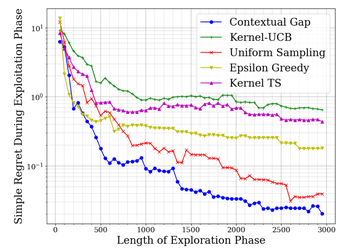
(a) Letter dataset



(b) USPS dataset



(c) MNIST dataset



(d) Spacecraft dataset

Figure: Simple Regret evaluation

References

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[2] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

[3] M. Hoffman, B. Shahriari, and N. Freitas. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*, pages 365–374, 2014.