A Literature Overview

Existing works on contextual bandits can be broadly categorized into two categories [12]. The first category includes algorithms that focus on exploration problems while the second category pertains to using different base learners for contextual bandits. Literature focusing on the exploration problems provide theoretical guarantees on regret bounds, however, the bounds are often tied with very strong underlying assumptions. In [8], for instance, the authors provide a regret bound that holds with probability $1-\delta$ and is of the order $\mathcal{O}\Big(\sqrt{Td\ln^3(KT\ln(T)/\delta}\Big)$ for d dimensional context vector with T rounds and K actions. The underlying assumption for obtaining such regret bounds is that the payoff (or reward) is a linear function of the context features. Similar assumptions underlie the work in [4], where the authors present Thompson sampling for contextual bandits with linear payoffs.
For the second category, different base learners for contextual bandits have been investigated. In [5], the authors employ neural networks to model the value of rewards given the contexts. Subsequently deep learning for contextual bandits have also been explored in [31, 25, 16]. In [13], the authors use random forest as a base learner for contextual bandits. The proposed learner is optimal up to a logarithmic factor where the computational cost of the algorithm is linear with respect to the time horizon. In [12], the authors use decision tree learners for contextual bandits, and then propose Thompson sampling for such non-parametric learners.
Our work pertains to the second category, where we use TM as a base learner for contextual bandits. Our TM learner supports incremental training with streaming data. In contrast to popular baseline learners such as artificial neural networks, both the learned arm-context model and the process of learning are easy to follow and explain. The interpretability is attained using propositional functions of context features used by TM for arm selection.

B Parameters

The *Noisy XOR* dataset was already binarized and for *MNIST*, we binarize the features having value ≥ 75 to be 1 and 0 otherwise. For other datasets, the maximum number of bits for binarization per feature is given by Table 3.

Table 3: Maximum number of bits per feature

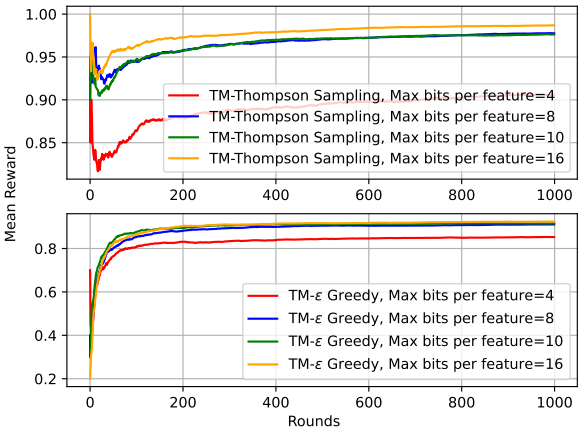
DATASETS	CONTEXT DIM	MAX BITS PER FEATURE
IRIS	4	4
BREAST CANCER (DIAGNOSTIC)	30	10
ADULT	15	10
STATLOG SHUTTLE	9	10
COVERTYPE	54	10
SIMULATED ARTICLE	4	10
MOVIE LENS	10	8

Table 4: TM learner configuration

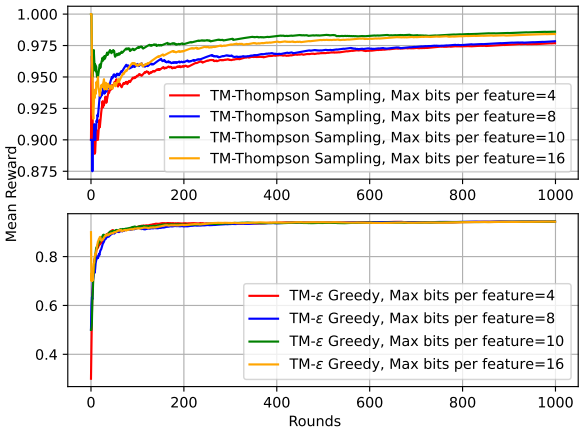
DATASETS	#CLAUSES	T	S	STATE BITS
IRIS	1200	1000	8.0	10
BREAST CANCER (DIAGNOSTIC)	650	300	5.0	10
NOISY XOR	1000	700	5.0	8
ADULT	1200	800	5.0	8
STATLOG SHUTTLE	1200	800	5.0	8
COVERTYPE	1200	800	5.0	8
SIMULATED ARTICLE	2000	1500	5.0	10
MOVIE LENS	4000	3000	8.0	8
MNIST	5000	4000	5.0	8

Table 4 shows the configuration of the contextual bandit learners using TM. The same TM configurations are used for both TM with $\varepsilon-$greedy arm selection and TM with Thompson sampling.

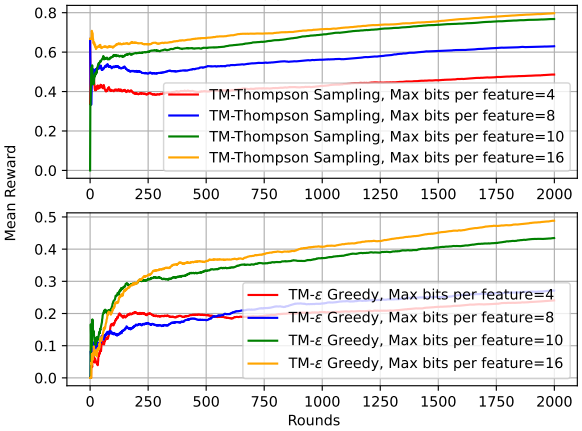
C Empirical analysis of sub-optimal binarization in TM



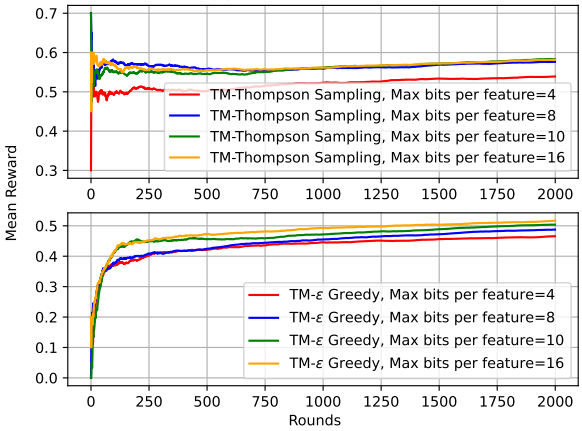
(a) Iris



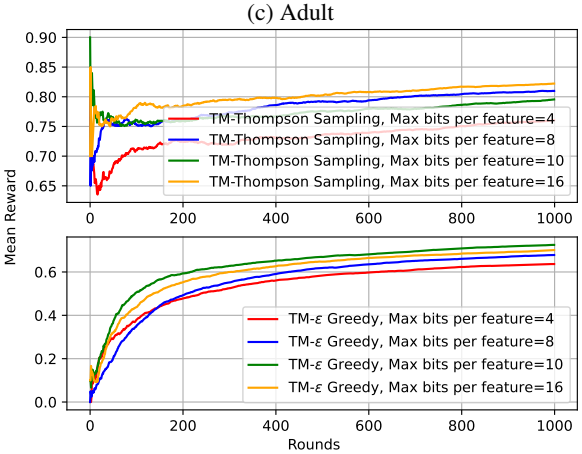
(b) Breast Cancer



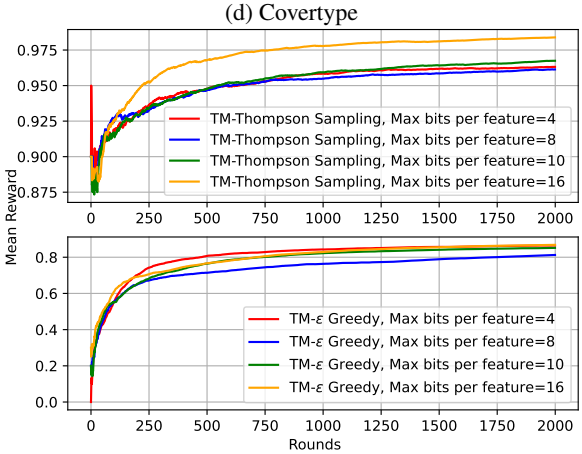
(c) Adult



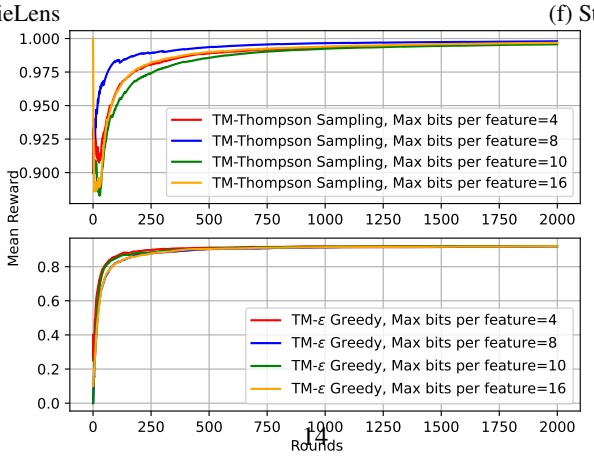
(d) Covertype



(e) MovieLens



(f) Statlog



(g) Simulated Article

Figure 4: Empirical analysis with sub-optimal binarization