(Sequential) Importance Sampling Bandits

Iñigo Urteaga and Chris H. Wiggins
{inigo.urteaga, chris.wiggins}@columbia.edu

Department of Applied Physics and Applied Mathematics
Data Science Institute
Columbia University
New York City, NY 10027

August 8, 2018

Abstract

The multi-armed bandit (MAB) problem is a sequential allocation task where the goal is to learn a policy that maximizes long term payoff, where only the reward of the executed action is observed; i.e., sequential optimal decisions are made, while simultaneously learning how the world operates. In the stochastic setting, the reward for each action is generated from an unknown distribution. To decide the next optimal action to take, one must compute sufficient statistics of this unknown reward distribution, e.g., upper-confidence bounds (UCB), or expectations in Thompson sampling. Closed-form expressions for these statistics of interest are analytically intractable except for simple cases. We here propose to leverage Monte Carlo estimation and, in particular, the flexibility of (sequential) importance sampling (IS) to allow for accurate estimation of the statistics of interest within the MAB problem. IS methods estimate posterior densities or expectations in probabilistic models that are analytically intractable. We first show how IS can be combined with state-of-the-art MAB algorithms (Thompson sampling and Bayes-UCB) for classic (Bernoulli and contextual linear-Gaussian) bandit problems. Furthermore, we leverage the power of sequential IS to extend the applicability of these algorithms beyond the classic settings, and tackle additional useful cases. Specifically, we study the dynamic linear-Gaussian bandit, and both the static and dynamic logistic cases too. The flexibility of (sequential) importance sampling is shown to be fundamental for obtaining efficient estimates of the key sufficient statistics in these challenging scenarios.

1 Introduction

The multi-armed bandit (MAB) problem considers the strategy one must devise when playing a row of slot machines: i.e., which arms to play to maximize returns. This analogy extends to a wide range of interesting real-world challenges that require online learning while simultaneously maximizing some notion of reward: e.g., a doctor must prescribe one of several medicines to a patient; a manager must allocate resources to one of several competing projects; or an e-commerce service must decide which of several ads to display. This setting is more formally referred to as the theory of sequential decision processes, a particular study area within machine learning known as reinforcement learning (46).

Interest in sequential decision processes has recently intensified in both academic and industrial communities, although its foundations in statistics can be traced back to the first decades of the past century, with important contributions by Thompson (48) and later Robbins (41). Very recently, the publication of works by Chapelle and Li (10) and others in industry have shown the field's impact in digital advertising and products. At the same time, an academic renaissance of the study of the multi-armed bandit problem from both a practical (31) and a theoretical (1, 36, 44) perspective has flourished.

Over the years, several algorithms have been proposed to overcome the exploration-exploitation tradeoff in the MAB problem: some based on heuristics (5), some based on optimal strategies with geometrically discounted future rewards (22), and others based on upper confidence bounds (28, 29). Bayesian counterparts of UCB-type algorithms have also been recently proposed (26). A key contribution to this revival period was the observation that one of the oldest heuristics, i.e., Thompson sampling (47, 48), has been empirically and theoretically proven to perform competitively (2, 3, 10, 27, 42, 43, 45).

Bayesian modeling of the MAB problem facilitates not only generative and interpretable modeling, but sequential and batch processing algorithm development as well. Two prime examples of the Bayesian approach to the MAB problem are Thompson sampling as in (44) and Bayes-UCB in (26). However, the application of these are limited by the complexity of the assumed reward functions, since one must both sample from the distributions modeled and/or calculate their expected rewards. This is cumbersome except in the case of simple models, e.g., those within the exponential family of distributions (27).

We here introduce sampling methods, which extend the applicability of Bayesian MAB algorithms by permitting more complex models: those for which sampling may be performed even if analytic computation of summary statistics is infeasible. This approach complements the variational approach (7), recently proposed for both general reinforcement learning problems (8, 33), and posterior sampling-based algorithms as well (30, 49). Variational inference provides a very general method for approximating generative models, but does not provide optimality guarantees.

We focus on importance sampling (IS) methods, which are a general technique for estimating properties of a distribution, using only samples generated from a different distribution. These methods are used to estimate posterior densities or expectations in problems with probabilistic models that are too complex to treat analytically. Furthermore, they are the foundation of sequential Monte Carlo (SMC) methods (4, 18, 21), which have been successful in many applications of science and engineering (14, 24, 40, 50). These methods require that one can evaluate the likelihood of the observations up to a proportionality constant, and furthermore, they provide tight convergence guarantees under general assumptions (11, 15).

Our contribution is unique to the MAB problem in that we provide a SIS-based MAB method that (i) approximates the posterior densities of interest via random measures; (ii) requires knowledge of the reward function only up to a proportionality constant; and (iii) is applicable to time-varying parameter models, i.e., dynamic bandits.

Our work extends existing MAB policy algorithms beyond their original settings by leveraging the advances in SMC methods from the approximate inference community. The goal is to provide a flexible framework for solving a rich class of MAB problems, such as dynamic bandits, for which we are not aware of other general alternatives.

We formally introduce the MAB problem and SIS methods in Section 2, before providing the description of the proposed SIS based MAB framework in Section 3. We evaluate its performance for Thompson sampling and Bayes-UCB based policies in Section 4, and conclude with promising research directions suggested by these results in Section 5.

2 Problem Statement

2.1 Multi-armed bandits

We consider the problem of maximizing the rewards resulting from sequentially chosen actions $a \in \{1, \ldots, A\}$ (named *arms* in the bandit literature). The reward function is stochastic, parameterized by the intrinsic properties of each arm (i.e., parameters $\theta \in \Theta$) and potentially depends on a context x, e.g., $x \in \mathbb{R}^d$.

At each round t, the reward y_t is observed only for one chosen arm a_t (one of A possible arms) and is independently and identically drawn from its distribution: $y_t \sim p_{a_t}(y|x_t, \theta_{a,t})$. We allow for time-varying context and parameters (note the subscript $_t$ in both), although for static bandits, parameters are constant (i.e., $\theta_{a,t} = \theta_a$, $\forall t$). This same problem formulation includes non-contextual bandits, which may be described by fixing the context to a constant value $x_t = x$.

In the MAB problem, the next arm to play is chosen based upon the history observed, which contains the set of given contexts, played arms, and observed rewards up to time t, denoted as $\mathcal{H}_{1:t} = \{y_{1:t}, a_{1:t}, x_{1:t}\}$, with $y_{1:t} \equiv (y_1, \cdots, y_t)$, $a_{1:t} \equiv (a_1, \cdots, a_t)$, and $x_{1:t} \equiv (x_1, \cdots, x_t)$.

The goal of a bandit algorithm is to maximize its cumulative reward, or alternatively minimize its cumulative regret – the loss incurred due to not knowing the best arm a_t^* at each time t. Due to the stochastic nature of the bandit, regret is expressed via the expected reward $\mu_{a,t} = \mathbb{E}_a\{y|x_t, \theta_{a,t}\}$. The cumulative *expected* regret in a time horizon T (not necessarily known a priori) is

$$R_T = \mathbb{E}\left\{\sum_{t=0}^{T} \mu_{a^*,t} - \mu_{a,t}\right\} , \tag{1}$$

where for each time instant t, $\mu_{a^*,t}$ denotes the true expected reward of the optimal arm, $\mu_{a,t}$ the expected reward of the played arm, and the outer expectation is over the arm selection choices made by the algorithm; i.e., the MAB policy.

When the parameters of the arms are known, one can readily determine the optimal selection policy $a_t^* = \text{argmax}_a\, \mu_{a,t}$. However, when there is a lack of knowledge about the model, one needs to learn the properties of the environment (i.e., the parameters of the reward distribution), as one interacts with the world (i.e., decides which action to take next). Hence, one must take into account the uncertainty on the unknown (and possibly dynamic) parameters of the world.

In a Bayesian approach to the MAB problem, prior knowledge on the model and parameters is incorporated into the algorithm. As one interacts with the environment, a Bayesian algorithm updates the parameter posterior, capturing the full state of knowledge via

$$p(\theta_t|\mathcal{H}_{1:t}) \propto p_{a_t}(y_t|x_t, \theta_t) p(\theta_t|\mathcal{H}_{1:t-1}) , \tag{2}$$

where $p_{a_t}(y_t|x_t, \theta_t)$ is the likelihood of the observed reward y_t after playing arm a_t at time t, and θ_t refers to the union of all per-arm parameters at time t, i.e., $\theta_t \equiv \{\theta_{a=1,t}, \cdots, \theta_{a=A,t}\}$. This posterior is the key component for the MAB problem, both for algorithms based on posterior sampling and those based on confidence intervals.

For the former (e.g., Thompson sampling), one uses $p(\theta_t|\mathcal{H}_{1:t})$ to compute the probability of an arm being optimal. To that end, the conditional probability of each arm being optimal given some context and the set of parameters, $p(a = a^*|x_t, \theta_t)$, is marginalized with respect

to the updated posterior

$$p_{a,t} = \int p(a = a^*|x_t, \theta_t) p(\theta_t|\mathcal{H}_{1:t}) \mathrm{d}\theta = \int I(\mu_{a,t}) p(\theta_t|\mathcal{H}_{1:t}) \mathrm{d}\theta \ , \\ \text{where } I(\mu_{a,t}) = \begin{cases} 1, \ \mu_{a,t} = \max_{a'} \{\mu_{a',t}\} \ , \\ 0, \ \text{otherwise} \ . \end{cases} \tag{3}$$

For the latter (e.g., Bayes-UCB), $p(\theta_t|\mathcal{H}_{1:t})$ is critical to determine the distribution of the expected rewards

$$p(\mu_{a,t}) = \int p(\mu_{a,t}|\theta_t) p(\theta_t|\mathcal{H}_{1:t}) \mathrm{d}\theta \ , \tag{4}$$

required for computation of the expected reward quantile value of interest $q_{a,t}(\alpha_t)$, i.e.,

$$\Pr\left[\mu_{a,t} > q_{a,t}(\alpha_t)\right] = \alpha_t \ . \tag{5}$$

Note that we allow the possible case wherein α_t depends on time, as in (26).

Analytical expressions for the parameter posteriors $p(\theta_t|\mathcal{H}_{1:t})$ are available only for few reward functions (e.g., Bernoulli and linear contextual Gaussian), but not for many other useful cases, such as logistic rewards. Furthermore, computation of the key summary statistics in Eqns. (3) and (5) can be challenging for many distributions. These issues become even more imperative when dealing with dynamic parameters, i.e., in environments that evolve over time. To overcome these issues, we propose to leverage (sequential) importance sampling.

2.2 Sequential Importance Sampling

Monte Carlo (MC) methods are a family of numerical techniques based on repeated random sampling, which have been shown to be flexible enough for both numerical integration and drawing samples from probability distributions of interest.

Importance sampling (IS) is a MC technique for estimating properties of a distribution when obtaining samples from the distribution is difficult. The basic idea of IS is to draw, from an alternative distribution, samples which are subsequently weighted to guarantee estimation accuracy (and often reduced variance). These methods are used both to approximate posterior densities, and to compute expectations in probabilistic models, i.e.,

$$\bar{f} = \int f(\varphi) p(\varphi) \mathrm{d}\varphi \ , \tag{6}$$

when these are too complex to treat analytically.

In short, IS relies on a proposal distribution $\pi(\cdot)$, from which one draws M samples $\varphi^{(m)} \sim \pi(\varphi), \ m = 1, \cdots, M$, and a set of weights

$$\widetilde{w}^{(m)} = \frac{p(\varphi^{(m)})}{\pi(\varphi^{(m)})} \ , \quad \text{with} \quad w^{(m)} = \frac{\widetilde{w}^{(m)}}{\sum_{m=1}^{M} \widetilde{w}^{(m)}} \ . \tag{7}$$

If the support of $\pi(\cdot)$ includes the support of the distribution of interest $p(\cdot)$, one computes the IS estimator of a test function based on the normalized weights $w^{(m)}$,

$$\bar{f}_M = \sum_{m=1}^{M} w^{(m)} f\left(\varphi^{(m)}\right) \ , \tag{8}$$

with convergence guarantees under weak assumptions

$$\bar{f}_M \xrightarrow[M\to\infty]{a.s} \bar{f} \,. \tag{9}$$

Note that IS can also be interpreted as a sampling method where the true posterior distribution is approximated by a random measure

$$p(\varphi) \approx p_M(\varphi) = \sum_{m=1}^{M} w^{(m)} \delta\left(\varphi^{(m)} - \varphi\right) \,, \tag{10}$$

which leads to estimates that are nothing but the test function integrated with respect to the empirical measure

$$\bar{f}_M = \int f(\varphi) p_M(\varphi) \mathrm{d}\varphi = \sum_{m=1}^{M} f\left(\varphi^{(m)}\right) w^{(m)} \,. \tag{11}$$

In many practical scenarios, observations are acquired sequentially in time, and one is interested in learning about the state of the world as data are collected. In these circumstances, one needs to infer all the unknown quantities in an online fashion. Furthermore, it is likely that the underlying parameters evolve over time. If the dynamics are modeled with known linearity and Gaussianity assumptions, then one can analytically obtain the posterior distributions of interest in closed form: i.e., the celebrated Kalman (25) filter (KF).

On the contrary, practical scenarios often require more lax assumptions: nonlinear functions, uncertainty on parameters, non-Gaussian distributions, etc. For these scenarios, sequential importance sampling (SIS), also known as sequential Monte Carlo or particle filtering, has been shown to be of great flexibility and value. These are simulation-based methods that provide a convenient solution to computing online approximations to posterior distributions.

In sequential importance sampling, one considers a proposal distribution that factorizes over time

$$\pi(\varphi_{0:t}) = \pi(\varphi_t|\varphi_{1:t-1})\pi(\varphi_{1:t-1}) = \prod_{\tau=1}^{t} \pi(\varphi_\tau|\varphi_{1:\tau-1})\pi(\varphi_0) \,, \tag{12}$$

which helps in matching the model dynamics $p(\varphi_t|\varphi_{1:t-1})$ to allow for recursive evaluation of the importance weights

$$w_t^{(m)} \propto \frac{p(\varphi_t|\varphi_{1:t-1})}{\pi(\varphi_t|\varphi_{1:t-1})} w_{t-1}^{(m)} \,. \tag{13}$$

One problem with SIS following the above weight update scheme is that, as time evolves, the distribution of the importance weights becomes more and more skewed, resulting in few (or just one) non-zero weights.

To overcome this degeneracy, an additional selection step, known as resampling (32), is added. In its most basic setting, one replaces the weighted empirical distribution with an equally weighted random measure at every time instant, where the number of offspring for each sample is proportional to its weight. This is known as the Sequential Importance Resampling (SIR) method (23), which we rely on for our proposed framework in Section 3. We acknowledge that any of the numerous methodological improvements within the SMC literature (such as alternative resampling mechanisms (32, 37)) are readily applicable to our proposed methods and likely to have a positive impact on performance.

3 Proposed framework

In this paper, we leverage (sequential) importance sampling to compute the posteriors and sufficient statistics of interest for the MAB problem. Specifically, we consider the SIR method for (dynamic) bandits, where the world (might) evolve over time, i.e., $\theta_t \sim p(\theta_t|\theta_{1:t-1})$, and rewards are sequentially observed for the played arms: $y_t \sim p_{a_t}(y|x_t, \theta_t)$.

Mathematically, the MAB with per-arm stochastic reward functions with dynamic parameters is modeled as

$$\begin{cases} \theta_{a,t} \sim p(\theta_{a,t}|\theta_{a,1:t-1}) \ , \\ y_t \sim p_{a_t}(y|x_t, \theta_{a,t}) \ , \\ \mu_{a,t} = \mathbb{E}_{a_t}\{y|x_t, \theta_{a,t}\} \ . \end{cases} \tag{14}$$

In order to compute sufficient statistics of the rewards of each arm over time, sequential updates of their parameter posteriors as in Eqn. (2) are required. To that end, we implement the SIR method (23): the proposal distribution follows the assumed parameter dynamics, i.e., $\pi(\theta_{a,t}) = p(\theta_{a,t}|\theta_{a,1:t-1})$; weights are updated based on the likelihood of observed rewards, i.e., $p_a(y_t|x_t, \theta_{a,t})$; and the random measure is resampled at every time instant.

In the proposed SIR-based MAB framework (see full details in Algorithm 1), the fundamental operation is to sequentially update the random measure $p_M(\theta_{a,t})$ that approximates the true posterior $p(\theta_{a,t}|\mathcal{H}_{1:t})$, as it allows for computation of any statistic a MAB policy might require. Thus, the proposed framework is generic for MAB algorithms that compute test functions of per-arm parametric reward distributions. As a result, we are able to extend the applicability of existing policy algorithms beyond their original assumptions, from static to time-evolving bandits.

Algorithm 1 presents SIR for the general MAB problem, with specific instructions for Thompson sampling and Bayes-UCB policies. It is described in terms of generic likelihood and transition distributions $p_a(y|x_t, \theta_t)$ and $p(\theta_{a,t}|\theta_{a,1:t-1})$, respectively. For Algorithm 1 to be implemented in practice, the likelihood function must be computable up to a proportionality constant, and one needs to be able to draw samples from the transition density, which will depend on case-by-case assumed model dynamics (see Appendix A for details of both in several bandit settings).

The SIR method approximates each per-arm parameter posterior separately. That is, the dimensionality of the estimation problem depends on the size of the per-arm parameters, and not on the number of arms of the bandit. Therefore, there will be no particle degeneracy due to increased number of arms.

We note that particle stream degeneracy is an important issue when the smoothing distribution is of interest, i.e., $p(\theta_{a,1:t}|\mathcal{H}_{1:t})$. However, in the bandit setting, one cares about the posterior density of the parameters at each time instant, i.e., the filtering density $p(\theta_{a,t}|\mathcal{H}_{1:t})$, for which there are strong theoretical SMC convergence guarantees (see details in (15) and (11)). We reiterate that the resampling and propagation steps in Algorithm 1 are necessary to attain accurate and non-degenerate sequential approximations to the true posterior.

In the following, we describe how SIR can be used for both posterior sampling-based and UCB-type policies; i.e., which are the specific instructions to execute in steps 5 and 7 within Algorithm 1 for Thompson sampling and Bayes-UCB.

Algorithm 1 SIR for MAB

Require: A, $p(\theta_a)$, $p(\theta_{a,t}|\theta_{a,1:t-1})$, $p_a(y|x,\theta)$, M (for UCB we also require α_t)

Draw initial samples from the parameter prior

$$\overline{\theta}_{a,0} \sim p(\theta_a), \forall a \in A\ , \quad \text{and} \quad w_{a,0}^{(m)} = \frac{1}{M}\ .$$

for $t = 0, \cdots, T-1$ **do**

  Receive context x_{t+1}

  for $a = 1, \cdots, A$ **do**

    Estimate sufficient statistics for the MAB policy, given updated $\{w_t^{(m)}\}$ and $\{\theta_{a,1:t}^{(m)}\}$

    Thompson sampling:

      Draw a sample $s \sim \text{Cat}\left(w_t^{(m)}\right)$,

      Propagate the sample parameter $\theta_{a,t+1}^{(s)} \sim p(\theta_{a,t+1}|\theta_{a,1:t}^{(s)})$,

      Set $\mu_{a,t+1} = \mathbb{E}\left\{y|x_{t+1}, \theta_{a,t+1}^{(s)}\right\}$.

    Bayes-UCB:

      Draw samples $m \sim \text{Cat}\left(w_t^{(m)}\right)$, $m = 1, \cdots, M$,

      Propagate parameters $\theta_{a,t+1}^{(m)} \sim p(\theta_{a,t+1}|\theta_{a,1:t}^{(m)})$,

      Set $\mu_{a,t+1}^{(m)} = \mathbb{E}\left\{y|x_{t+1}, \theta_{a,t+1}^{(m)}\right\}$,

      Estimate quantile $q_{a,t+1}(\alpha_{t+1})$ as in Eqn. (20).

  end for

  Decide next action a_{t+1} to play

    Thompson sampling:

      $a_{t+1} = \text{argmax}_a\, \mu_{a,t+1}$

    Bayes-UCB:

      $a_{t+1} = \text{argmax}_a\, q_{a,t+1}(\alpha_{t+1})$

  Observe reward y_{t+1} for played arm

  Update posterior following SIR steps

    Resample $m = 1, \cdots, M$ parameters $\overline{\theta}_{a,1:t}^{(m)}$,

      where m is drawn with replacement according to the importance weights $w_t^{(m)}$.

    Propagate resampled parameters by drawing from the transition density

$$\theta_{a,t+1}^{(m)} \sim p(\theta_{a,t+1}|\overline{\theta}_{a,1:t}^{(m)})\ ,\ m = 1, \cdots, M. \tag{15}$$

    Weight samples based on likelihood of y_{t+1}

$$\widetilde{w}_{t+1}^{(m)} \propto p(y_{t+1}|x_{t+1}, \theta_{a,t+1}^{(m)})\ ,\ m = 1, \cdots, M. \tag{16}$$

    Normalize weights

$$w_{t+1}^{(m)} = \frac{\widetilde{w}_{t+1}^{(m)}}{\sum_{m=1}^{M} \widetilde{w}_{t+1}^{(m)}}\ ,\ m = 1, \cdots, M. \tag{17}$$

end for

3.1 SIR-based MAB policies

3.1.1 SIR-based Thompson Sampling

Thompson sampling is a probability matching algorithm that randomly selects an action to play according to the probability of it being optimal. Thompson sampling has been empirically proven to perform satisfactorily and to enjoy provable optimality properties, both for problems with and without context (2, 3, 27, 42, 43). It requires computation of the optimal probability as in Eqn. (3), which is in general analytically intractable. Alternatively, Thompson sampling operates by drawing a sample parameter $\theta_t^{(s)}$ from its updated posterior $p(\theta_t|\mathcal{H}_{1:t})$, and picking the optimal arm for such sample, i.e.,

$$a_t^* = \underset{a}{\operatorname{argmax}}\, \mu_{a,t}^{(s)}, \text{ where } \mu_{a,t}^{(s)} = \mathbb{E}_a\{y_t|x_t, \theta_{a,t}^{(s)}\} \,. \tag{18}$$

As pointed out already, the posterior distribution $p(\theta_t|\mathcal{H}_{1:t})$ is for many cases of applied interest either analytically intractable or hard to sample from. We propose to use the SIR-based random measure $p_M(\theta_t)$ instead, as it provides an accurate approximation to the true with high probability.

3.1.2 SIR-based Bayes-UCB

Bayes-UCB (26) is a Bayesian approach to UCB type algorithms, where Bayesian quantiles are used as proxies for upper confidence bounds. Kaufmann (26) has proven the asymptotic optimality of Bayes-UCB's finite-time regret bound for the Bernoulli case, and argues that it provides an unifying framework for several variants of the UCB algorithm for parametric MABs. However, its application is limited to reward models where the quantile functions are analytically tractable.

We propose instead to compute the quantile function of interest by means of the SIR approximation to the parameter posterior, where one can evaluate the expected reward at each round t based on the available posterior samples, i.e., $\mu_{a,t}^{(m)} = \mathbb{E}_a\{y_t|x_t, \theta_{a,t}^{(m)}\}$.

The quantile value

$$\Pr[\mu_{a,t} > q_{a,t}(\alpha_t)] = \alpha_t \tag{19}$$

is then computed by

$$q_{a,t}(\alpha_t) := \max\{\mu| \sum_{m|\mu_{a,t}^m > \mu} w^m \geq \alpha_t\} \,. \tag{20}$$

3.2 SIR for MAB models

We describe here how the proposed SIR-based algorithm operates in terms of likelihood $p_a(y|x_t, \theta_t)$, and transition $p(\theta_{a,t}|\theta_{a,1:t-1})$ distributions, both for static and dynamic bandits. Full details on computing specific MAB reward distributions $p_a(y|x_t, \theta_t)$ of interest are provided in Appendix A .

3.2.1 SIR for static bandits

These are bandits where there are no time-varying parameters: i.e., $\theta_t = \theta, \forall t$. For these settings, SIR-based parameter propagation becomes troublesome (35), and to mitigate such issues, several alternatives have been proposed in the SMC community: artificial parameter evolution (23), kernel smoothing (35), and density assisted techniques (19).

We resort to density assisted importance sampling[1], where one approximates the posterior of the unknown parameters with a density of choice. Density assisted importance sampling is a well studied SMC approach that extends random-walking and kernel-based alternatives (19, 23, 34), with its asymptotic correctness guaranteed for the static parameter case.

Specifically, we approximate the posterior of the unknown parameters θ, given the current state of knowledge, with a Gaussian distribution $p(\theta_a|\mathcal{H}_{1:t}) \approx \mathcal{N}\left(\theta_a|\hat{\theta}_{a,t}, \hat{C}_{\theta_{a,t}}\right)$. The sufficient statistics are computed based on samples and weights of the random measure $p_M(\theta_{a,t}) = \sum_{m=1}^{M} w_t^{(m)} \delta\left(\theta_{a,t} - \theta_{a,t}^{(m)}\right)$ available at each time instant:

$$\begin{aligned} \hat{\theta}_{a,t} &= \sum_{i=1}^{M} w_{a,t}^{(m)} \theta_{a,t}^{(m)} \ , \\ \hat{C}_{\theta_{a,t}} &= \sum_{i=1}^{M} w_{a,t}^{(m)} (\theta_{a,t}^{(m)} - \hat{\theta}_{a,t})(\theta_{a,t}^{(m)} - \hat{\theta}_{a,t})^{\top} \ . \end{aligned} \tag{21}$$

For static bandits, when propagating parameters in Algorithm 1, one draws from $p(\theta_{a,t+1}|\theta_{a,1:t}) = p(\theta_{a,t}|\mathcal{H}_{1:t}) \approx \mathcal{N}\left(\theta_a|\hat{\theta}_{a,t}, \hat{C}_{\theta_{a,t}}\right)$.

3.3 SIR for dynamic bandits

A widely applicable model for time-evolving bandit parameters is the general linear model. That is, the parameters of each arm $\theta_a \in \mathbb{R}^d$ follow dynamics of the form

$$\theta_{a,t} = L_a \theta_{a,t-1} + \epsilon_a \ , \qquad \epsilon_a \sim \mathcal{N}\left(\epsilon_a|0, C_a\right) \ , \tag{22}$$

where $L_a \in \mathbb{R}^{d\times d}$ and $C_a \in \mathbb{R}^{d\times d}$. With known parameters, the transition distribution is Gaussian, i.e., $\theta_{a,t} \sim \mathcal{N}\left(\theta_{a,t}|L_a\theta_{a,t-1}, C_a\right)$; while for the unknown parameter case, the marginalized transition density [2] is a multivariate-t, i.e., $\theta_{a,t} \sim \mathcal{T}\left(\theta_{a,t}|\nu_{a,t}, m_{a,t}, R_{a,t}\right)$ with sufficient statistics as in Eqns. 23-24 below, where each equation holds separately for each arm a (the subscript has been suppressed for clarity of presentation, and subscript 0 indicates assumed prior parameters):

$$\begin{cases} \nu_t = \nu_0 + t - d \ , \\ m_t = L_{t-1}\theta_{t-1} \ , \\ R_t = \frac{V_{t-1}}{\nu_t\left(1-\theta_{t-1}^{\top}(UU^{\top})^{-1}\theta_{t-1}\right)} \ , \end{cases} \tag{23}$$

and

$$\begin{cases} \Theta_{t_0:t_1} = [\theta_{t_0}\theta_{t_0+1}\cdots\theta_{t_1-1}\theta_{a,t_1}] \in \mathbb{R}^{d\times(t_1-t_0)} \ , \\ B_{t-1} = \left(\Theta_{0:t-2}\Theta_{0:t-2}^{\top} + B_0^{-1}\right)^{-1} \ , \\ L_{t-1} = \left(\Theta_{1:t-1}\Theta_{0:t-2}^{\top} + A_0 B_0^{-1}\right) B_{t-1} \ , \\ V_{t-1} = \left(\Theta_{1:t-1} - L_{t-1}\Theta_{0:t-2}\right)\left(\Theta_{1:t-1} - L_{t-1}\Theta_{0:t-2}\right)^{\top} \\ \qquad\qquad + \left(L_{t-1} - L_0\right) B_0^{-1} \left(L_{t-1} - L_0\right)^{\top} + V_0 \ , \\ UU^{\top} = \left(\theta_{t-1}\theta_{t-1}^{\top} + B_{t-1}^{-1}\right) \ . \end{cases} \tag{24}$$

[1] We acknowledge that any of the more advanced SMC techniques that mitigate the challenges of estimating constant parameters (e.g., parameter smoothing (9, 38, 39) or nested SMC methods (12, 16)) can only improve the accuracy of the proposed method.

[2] Details of the derivation are provided in Appendix A .

We marginalize the unknown parameters of the transition distributions above (i.e., we perform Rao-Blackwellization), in order to reduce the degeneracy and variance of the SMC estimates (17, 20). One estimates the predictive posterior of per-arm parameters, as a mixture of the transition densities conditioned on previous samples

$$p_M(\theta_{a,t+1}) = \sum_{m=1}^{M} w_t^{(m)} p(\theta_{a,t+1}|\theta_{a,1:t}^{(m)}) \ . \tag{25}$$

For the dynamic bandit case, these transition distributions are used when propagating parameters in Algorithm 1.

The propagation of parameter samples in the SIR algorithm is fundamental for the accuracy of the sequential approximation to the posterior, and the performance of the SIR-based MAB policy as well. The increasing uncertainty of the parameter posterior encourages exploration of arms that have not been played recently, but may have evolved into new parameter spaces with exploitable rewards distributions. That is, as the dynamics of unobserved arms result in broad SIR posteriors (increased uncertainty about parameters), MAB policies are more likely to explore such arm, reduce their posterior's uncertainty, and in turn, update the exploration-exploitation balance.

4 Evaluation

We now empirically evaluate the proposed SIR-based MAB framework. We first consider static Bernoulli and contextual linear-Gaussian bandits (i.e., Gaussian distributed bandits, whose mean reward is linear in the contextual features), which are well-studied MAB models (2, 3, 10, 26, 27, 44). Their study serves as validation of the quality of the proposed SMC approximation to posteriors, as the performance loss of the bandit is negligible: i.e. we show that the SIR-based methods work almost as good as the analytical alternative.

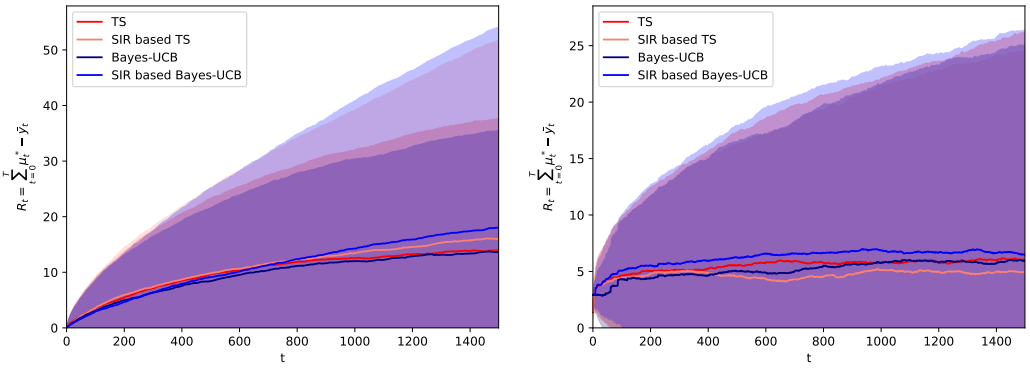
We further show the flexibility of the proposed method in more challenging scenarios, where there are no closed form posteriors available, e.g., logistic rewards and dynamic bandits. Note that with SMC, we extend the applicability of policies designed for static bandits to non-stationary cases, which are of interest in practice. To the best of our knowledge, no other approximate methods exist for the studied dynamic cases.

In all cases, the key performance metric is the cumulative regret defined in Eqn. (1), all results are averaged over at least 250 realizations, and SIR-based methods are implemented with $M = 1000$ samples. Due to space constraints, figures shown here are illustrative selected examples, although drawn conclusions are based on extensive experiments with different number of bandit arms and parameterizations (provided in Appendix B).

4.1 Static bandits

We first compare the performance of Algorithm 1 to Thompson sampling (TS) and Bayes-UCB in their original formulations: static bandits with Bernoulli and contextual linear-Gaussian reward functions. Fig. 1 shows how SIR-based algorithms perform similarly to the benchmark policies with analytical posteriors.

Note the increased uncertainty due to the MC approximation to the posterior, which finds its analytical justification in Eqn. (9), and can be empirically reduced by increasing the number M of IS samples used (see the impact of sample size M in figures included in Appendix B).



(a) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.6$

(b) Contextual linear-Gaussian bandit, $\theta_0 = (-1, -1), \theta_1 = (1, 1), \sigma^2 = 0.5$

Figure 1: Mean regret (standard deviation shown as shaded region) in example static bandits.

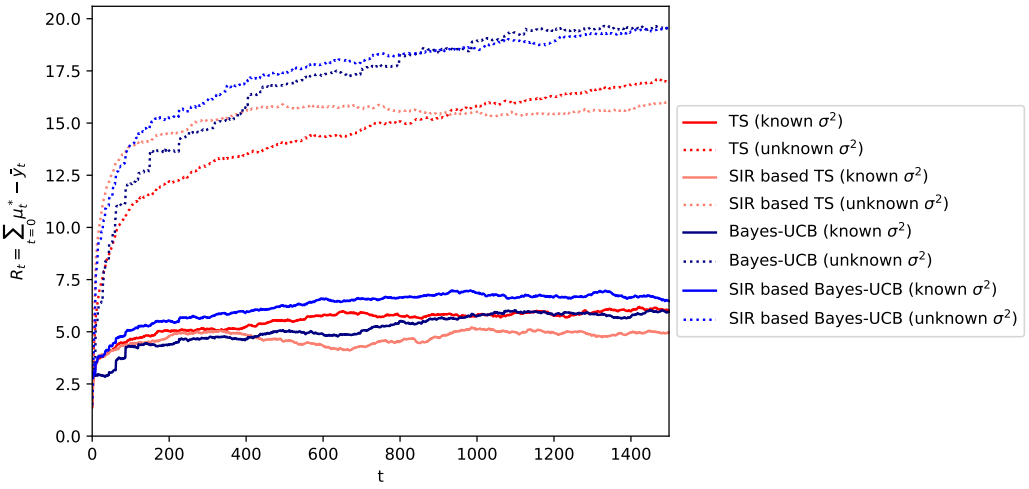


Figure 2: Mean regret for the contextual linear-Gaussian bandit $\theta_0 = (-1, -1), \theta_1 = (1, 1)$ with unknown reward variance $\sigma^2 = 0.5$ and random contexts.

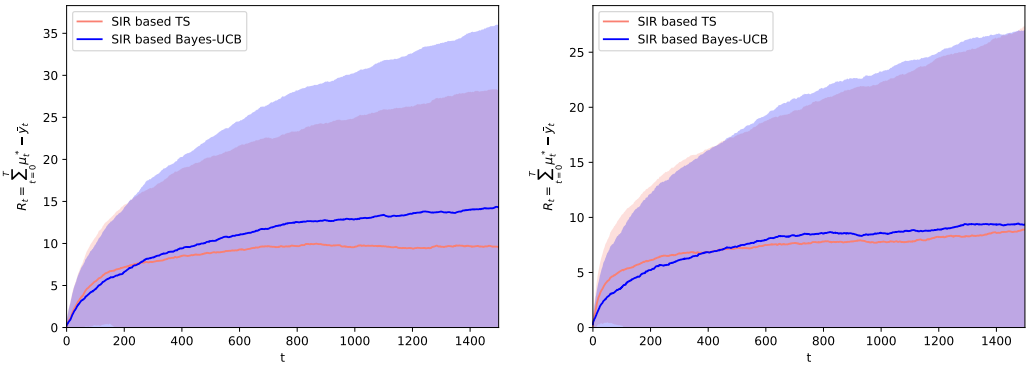
$M = 1000$ samples suffice in all our experiments for accurate estimation of parameter posteriors. Advanced and dynamic determination of SIR sample size is an active research area within the SMC community, but out of the scope of this paper.

We also evaluate a more realistic scenario for the contextual linear-Gaussian case with unknown reward variance σ^2 in Fig. 2, where SIR-based approaches are shown to be competitive as well. We reiterate that results are satisfactory across a wide range of parameterizations and bandit sizes (results for extensive evaluations are provided in Appendix B), which validate the accuracy of the proposed SIR-based method.

Overall, the random measure approximation to the posteriors of the parameters of interest is accurate enough to allow for MAB policies to find the right exploration-exploitation tradeoff.

In several applications, binary rewards are well-modeled as depending on contextual factors (10, 45). The logistic reward function is suitable for these scenarios but, due to the unavailability of Bayesian closed-form posterior updates (see subsection A.3), one needs to resort to approximations, e.g., the ad-hoc Laplace approximation proposed by Chapelle and Li (10).

Our proposed SIR-based framework is readily applicable, as one only needs to evaluate the logistic reward likelihood to compute the IS weights for the MAB policy of choice. Fig. 3 shows how quickly SIR-based Thompson sampling and Bayes-UCB achieve the right exploration-exploitation tradeoff for different logistic parameterizations.



(a) Logistic bandit with $\theta_{0,i} = -0.5, \theta_{1,i} = 0.5, \forall i$. (b) Logistic bandit with $\theta_{0,i} = -1.0, \theta_{1,i} = 1.0, \forall i$.

Figure 3: Mean regret (standard deviation shown as shaded region) in logistic bandits with random contexts.

These results indicate that the impact of only observing rewards of the played arms is minimal for the proposed SIR-based method. The parameter posterior uncertainty associated with the SIR-based estimation is automatically accounted for by both algorithms, as they explore rarely-played arms if the uncertainty is high.

However, we do observe a slight performance deterioration of Bayes-UCB, which we argue is related to the quantile value used ($\alpha_t \propto 1/t$). It was analytically justified by Kaufmann (26) for Bernoulli rewards, but might not be optimal for other reward functions and, more importantly, for the SIR-based approximation to parameter posteriors.

On the contrary, Thompson sampling is more flexible, automatically adjusting to the uncertainty of the posterior approximation, and thus, attaining reduced regret.

4.2 Dynamic bandits

Full potential of the proposed SIR-based algorithm is harnessed when facing the most interesting and challenging bandits: those with time-evolving parameters.

We consider the linear case (as formulated in Section 3.3), because it allows us to (i) validate the SIR-based approximation to the optimal posterior (i.e., the KF for the linear and Gaussian case); and (ii) show its flexibility and robustness to more realistic and challenging MAB models (with unknown parameters, nonlinear functions, and non-Gaussian distributions).

We have evaluated different parameterizations of the model as in Eqn. (22) (all provided in Appendix B), but we here focus on a two-armed contextual dynamic bandit with parameters

$$\begin{aligned}
\begin{pmatrix}\theta_{0,0,t}\\ \theta_{0,1,t}\end{pmatrix} &= \begin{pmatrix}0.9 & -0.1\\ -0.1 & 0.9\end{pmatrix}\begin{pmatrix}\theta_{0,0,t-1}\\ \theta_{0,1,t-1}\end{pmatrix} + \epsilon_0 \ , \qquad && \epsilon_0 \sim \mathcal{N}\left(\epsilon|0, 0.1\cdot \mathrm{I}\right) \ , \\
\begin{pmatrix}\theta_{1,0,t}\\ \theta_{1,1,t}\end{pmatrix} &= \begin{pmatrix}0.9 & 0.1\\ 0.1 & 0.9\end{pmatrix}\begin{pmatrix}\theta_{1,0,t-1}\\ \theta_{1,1,t-1}\end{pmatrix} + \epsilon_1 \ , \qquad && \epsilon_1 \sim \mathcal{N}\left(\epsilon|0, 0.1\cdot \mathrm{I}\right) \ .
\end{aligned} \tag{26}$$

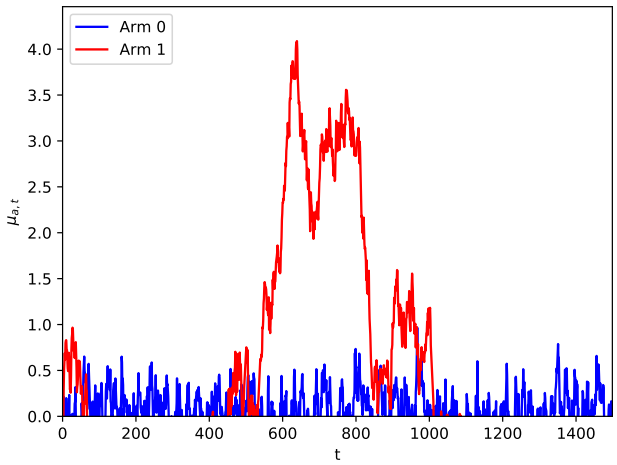


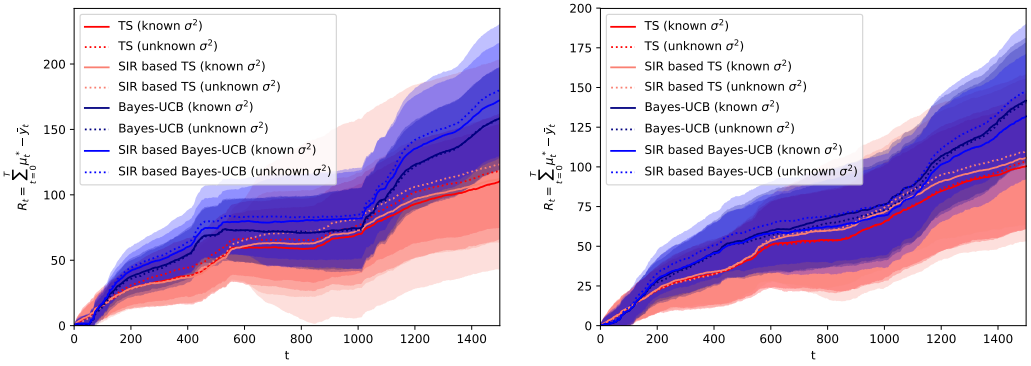
Figure 4: Expected per-arm rewards over time for a contextual linear dynamic bandit. Note how the optimal arm switches around $t = \{50, 500, 1000\}$.

Fig. 4 illustrates the time-evolution of the expected rewards for a realization of Eqn. (26). We consider this setting of special interest because the induced expected rewards change over time and so, the decision on the optimal arm swaps accordingly. We evaluate the proposed SIR-based methods for bandits with dynamics as in Eqn. (26), and both contextual linear-Gaussian and logistic rewards.

We show in Fig. 5 that the regret of SIR-based methods, for the contextual linear-Gaussian case with known parameters, is equivalent to the optimal case (i.e., the KF). Furthermore, even for the scenarios where the reward variance σ^2 is unknown, and thus the Gaussianity assumption needs to be dropped (instead modeling bandit rewards via Student-t distributions), SIR-based methods perform comparably well.

Finally, we evaluate in Fig. 6 the most challenging contextual linear-Gaussian bandit case, where none of the parameters of the model (A, C, σ^2) are known; i.e., one must sequentially learn the underlying dynamics, in order to make informed online decisions. Due to the flexibility of SIS in approximating the key parameter posteriors, SIR-based Thompson sampling and Bayes-UCB are both able to accurately learn their evolution and thus, reach the exploitation-exploration balance.

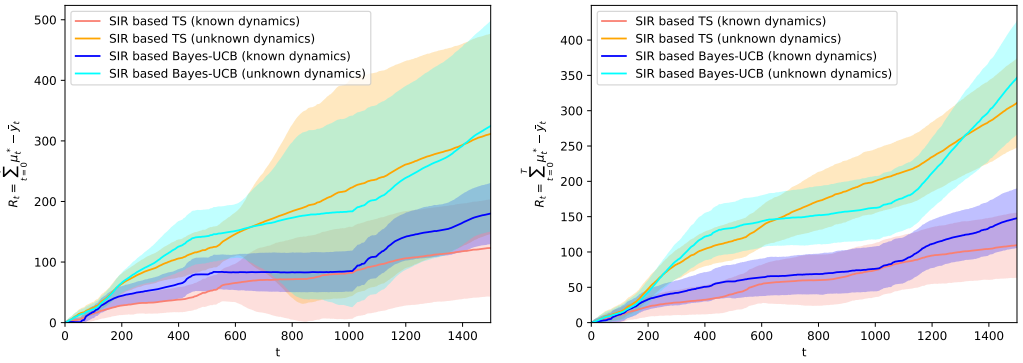
We observe noticeable increases in regret when the dynamics of the parameters (as in Fig. 4) swap the optimal arm. Note that these changes in parameter dynamics impact Bayes-UCB more profoundly as time evolves. This effect is also observed for linearly dynamic bandits with logistic reward functions (see Fig. 7 for logistic rewards with both static and random contexts).



(a) Static context.

(b) Uniformly random context.

Figure 5: Mean regret (standard deviation shown as shaded region) in contextual linear-Gaussian bandits with known dynamics. Notice the regret bumps when optimal arms swap at $t = \{500, 1000\}$, and how our proposed SIR-based methods adjust.



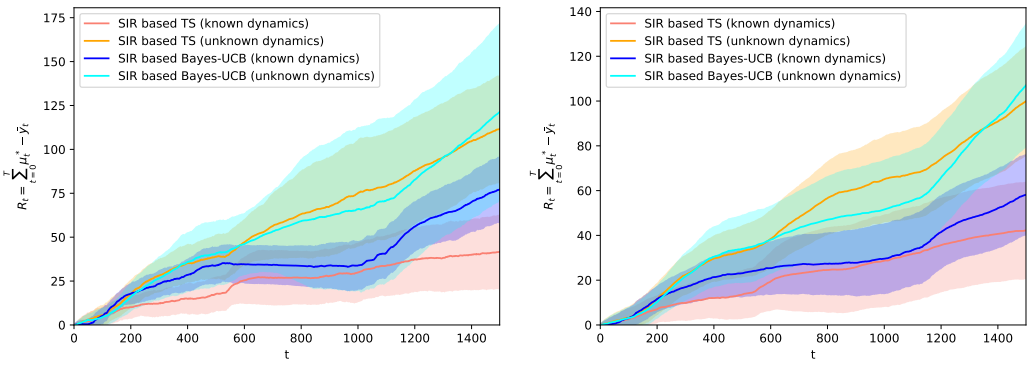
(a) Static context.

(b) Uniformly random context.

Figure 6: Mean regret (standard deviation shown as shaded region) in contextual linear-Gaussian bandits with unknown dynamics. Notice the regret bumps when optimal arms swap at $t = \{500, 1000\}$, and how our proposed SIR-based methods adjust.

We argue that the shrinking quantile value $\alpha_t \propto 1/t$ explains this behavior. It was originally proposed in (26) based on confidence bounds of static reward models, which tend to shrink with more observations of the bandit. However, the uncertainty of the evolving parameter posteriors (due to the dynamics of unobserved arms) might result in broader distributions and thus, the inadequacy of the shrinking α_t. More generally, the need to determine appropriate quantile values α_t for each model is a drawback for Bayes-UCB type algorithms.

On the contrary, Thompson sampling does not require any parameter tweaking. It relies on samples from the posterior, which SIR is able to approximate accurately enough for it to operate successfully, even in the most challenging MAB scenarios as in Figs. 6 and 7.



(a) Static context.

(b) Uniformly random context.

Figure 7: Mean regret (standard deviation shown as shaded region) in contextual linear logistic dynamic bandits. Notice the regret bumps when optimal arms swap at $t = \{500, 1000\}$, and how our proposed SIR-based methods adjust.

Due to the Bayesian approach to MAB, our proposed SIR-based Thompson sampling not only estimates the evolving parameters θ_t, but also their corresponding uncertainty. As such, both when a given arm's dynamics are unclear, or when an arm is not sampled for a while, the uncertainty of its estimated posterior grows. As a result, Thompson sampling is more likely to explore that arm again, in order to achieve the right exploration-exploitation balance, as shown in our results.

4.3 Bandits for personalized news article recommendation

Finally, we consider the application of the proposed SIR-based methods for recommendation of personalized news articles, in a similar fashion as done by Chapelle and Li (10). Online content recommendation represents an important example of reinforcement learning, as it requires efficient balancing of the exploration and exploitation tradeoff.

We use a dataset[3] that contains a fraction of user click logs for news articles displayed in the Featured Tab of the Today Module on Yahoo! Front Page during the first ten days in May 2009. The articles to be displayed were originally chosen uniformly at random from a hand-picked pool of high-quality articles. As such, the candidate pool was originally dynamic. However, we picked a subset of 20 articles shown in May 08th and collected all logged user interactions, for a total of 500354.

The goal is to choose the most interesting article to users, or in bandit terms, to maximize the total number of clicks on the recommended articles, i.e., the average click-through rate (CTR). In the dataset, each user is associated with six features: a bias term and 5 features that correspond to the membership features constructed via the conjoint analysis with a bilinear model described in (13).

We treat each article as an arm ($A = 20$), and the logged reward is whether the article is clicked or not by the user ($y_t = \{1, 0\}$). We pose the problem as a MAB with logistic rewards, so that we can account for the user features ($x_t \in \mathbb{R}^6$).

[3] Available at R6A - Yahoo! Front Page Today Module User Click Log Dataset.

One may further hypothesize that the news recommendation system should evolve over time, as the relevance of news might change during the course of the day. As a matter of fact, our proposed framework readily accommodates these assumptions.

We consider both static and dynamic bandits with logistic rewards, and implement the proposed SIR-based Thompson sampling, due to its flexibility and the lack of parameter tuning required. Summary CTR results are provided in Table 1. Observe the flexibility of the dynamic bandit, which is able to pick up the dynamic popularity of certain articles over time (see Fig. 8).

		CTR	Normalized CTR
Model	Logistic rewards, static arms	0.0670 +/- 0.0088	1.6149 +/- 0.2313
	Logistic rewards, time-evolving arms	0.0655 +/- 0.0082	1.5765 +/- 0.2042

Table 1: CTR results for SIR-based bandits on the news article recommendation data. The normalized CTR is with respect to a random baseline.

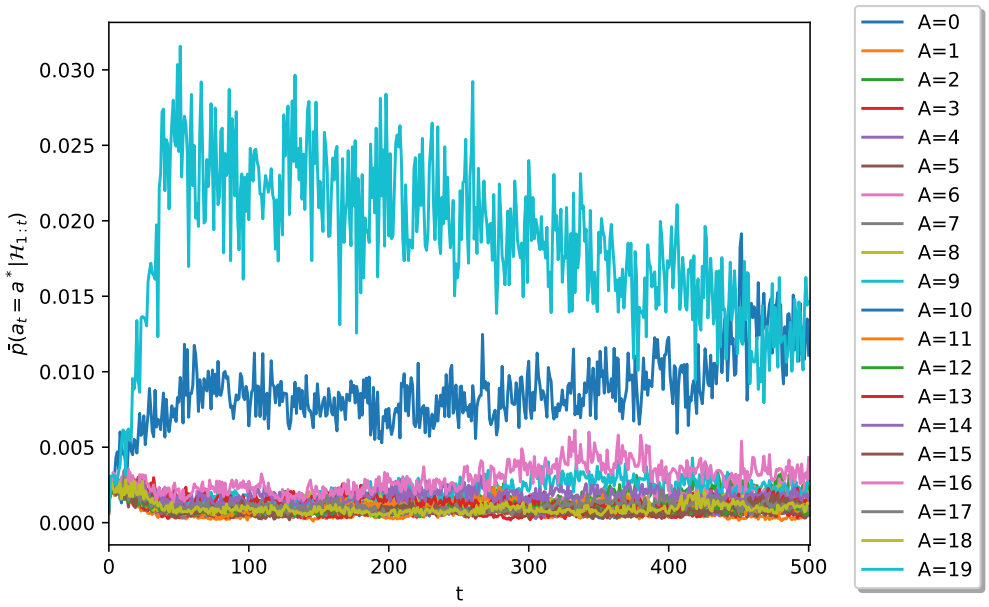


Figure 8: Empirical probability of SIR-based contextual dynamic logistic Thompson sampling policy picking each arm over time. Notice how the algorithm captures the changing popularity of articles over time.

5 Conclusion

We have presented a (sequential) importance sampling-based framework for the MAB problem, where we combine sequential Monte Carlo inference with state-of-the-art Bayesian MAB policies. The proposed algorithmic setting allows for interpretable modeling of complex reward functions and time-evolving bandits. The methods sequentially learn the dynamics of the bandit from online data, and are able to find the exploration-exploitation balance.

In summary, we extend the applicability of Bayesian MAB policies (Thompson sampling and Bayes-UCB in particular) by accommodating complex models of the world with SIR-based inference of the unknowns. Empirical results show good cumulative regret performance of the proposed framework in simulated challenging models (e.g., contextual logistic dynamic bandits), and practical scenarios (personalized news article recommendation) where complex models of data are required.

5.1 Software and Data

The implementation of the proposed method is available in this public repository. It contains all the software required for replication of the findings of this study.

Acknowledgments

This research was supported in part by NSF grant SCH-1344668.

References

[1] S. Agrawal and N. Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. *CoRR*, abs/1111.1797, 2011.

[2] S. Agrawal and N. Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. *CoRR*, abs/1209.3352, 2012.

[3] S. Agrawal and N. Goyal. Further Optimal Regret Bounds for Thompson Sampling. *CoRR*, abs/1209.3353, 2012.

[4] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *Signal Processing, IEEE Transactions on*, 50(2):174–188, 2 2002. ISSN 1053-587X.

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2-3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352.

[6] J. M. Bernardo and A. F. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 2009. ISBN 9780470317716. doi: 10.1002/9780470316870.

[7] C. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag New York, 2006.

[8] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight Uncertainty in Neural Networks. In *Proceedings of the 32Nd International Conference on International*

Conference on Machine Learning - Volume 37, ICML'15, pages 1613–1622. JMLR.org, 2015.

[9] C. M. Carvalho, M. S. Johannes, H. F. Lopes, and N. G. Polson. Particle Learning and Smoothing. *Statist. Sci.*, 25(1):88–106, 02 2010.

[10] O. Chapelle and L. Li. An Empirical Evaluation of Thompson Sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2249–2257. Curran Associates, Inc., 2011.

[11] N. Chopin. Central Limit Theorem for Sequential Monte Carlo Methods and Its Application to Bayesian Inference. *The Annals of Statistics*, 32(6):2385–2411, 2004. ISSN 00905364.

[12] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos. SMC$\hat{2}$: an efficient algorithm for sequential analysis of state-space models. *ArXiv e-prints*, Jan. 2011.

[13] W. Chu, S.-T. Park, T. Beaupre, N. Motgi, A. Phadke, S. Chakraborty, and J. Zachariah. A Case Study of Behavior-driven Conjoint Analysis on Yahoo!: Front Page Today Module. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1097–1104, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557138.

[14] D. Creal. A Survey of Sequential Monte Carlo Methods for Economics and Finance. *Econometric Reviews*, 31(3):245–296, 2012.

[15] D. Crisan and A. Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3):736–746, Mar 2002. ISSN 1053-587X. doi: 10.1109/78.984773.

[16] D. Crisan and J. Míguez. Nested particle filters for online parameter estimation in discrete-time state-space Markov models. *ArXiv e-prints*, Aug 2013.

[17] P. M. Djurić and M. F. Bugallo. *Particle Filtering*, chapter 5, pages 271–331. Wiley-Blackwell, 2010. ISBN 9780470575758. doi: 10.1002/9780470575758.ch5.

[18] P. M. Djurić, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Míguez. Particle Filtering. *IEEE Signal Processing Magazine*, 20(5):19–38, 9 2003.

[19] P. M. Djurić, M. F. Bugallo, and J. Míguez. Density assisted particle filters for state and parameter estimation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, volume 2, pages ii – 701–704, 5 2004. doi: 10.1109/ICASSP.2004.1326354.

[20] A. Doucet, N. de Freitas, K. P. Murphy, and S. J. Russell. Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, UAI '00, pages 176–183, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-709-9.

[21] A. Doucet, N. D. Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[22] J. C. Gittins. Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979. ISSN 00359246.

[23] N. J. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEEE Proceedings*, 140(2):107 –113, 4 1993. ISSN 0956-375X.

[24] E. L. Ionides, C. Bretó, and A. A. King. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443, 2006.

[25] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[26] E. Kaufmann, O. Cappe, and A. Garivier. On Bayesian Upper Confidence Bounds for Bandit Problems. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 592–600, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[27] N. Korda, E. Kaufmann, and R. Munos. Thompson Sampling for 1-Dimensional Exponential Family Bandits. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1448–1456. Curran Associates, Inc., 2013.

[28] T. L. Lai. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. *The Annals of Statistics*, 15(3):1091–1114, 1987. ISSN 00905364.

[29] T. L. Lai and H. Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1):4–22, mar 1985. ISSN 0196-8858. doi: 10.1016/0196-8858(85)90002-8.

[30] S. Lamprier, T. Gisselbrecht, and P. Gallinari. Variational Thompson Sampling for Relational Recurrent Bandits. In M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 405–421, Cham, 2017. Springer International Publishing. ISBN 978-3-319-71246-8.

[31] L. Li, W. Chu, J. Langford, and R. E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. *CoRR*, abs/1003.0146, 2010.

[32] T. Li, M. Bolić, and P. M. Djurić. Resampling Methods for Particle Filtering: Classification, implementation, and strategies. *Signal Processing Magazine, IEEE*, 32(3):70–86, 5 2015. ISSN 1053-5888.

[33] Z. C. Lipton, X. Li, J. Gao, L. Li, F. Ahmed, and L. Deng. Efficient Dialogue Policy Learning with BBQ-Networks. *ArXiv e-prints*, Aug. 2016.

[34] J. Liu and M. West. *Combined Parameter and State Estimation in Simulation-Based Filtering*, chapter 10, pages 197–223. Springer New York, New York, NY, 2001. ISBN 978-1-4757-3437-9. doi: 10.1007/978-1-4757-3437-9_10.

[35] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer, 2001.

[36] O.-A. Maillard, R. Munos, and G. Stoltz. Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences. In *Conference On Learning Theory*, 2011.

[37] L. Martino, V. Elvira, and F. Louzada. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386 – 401, 2017. ISSN 0165-1684.

[38] J. Olsson and J. Westerborn. Efficient particle-based online smoothing in general hidden Markov models: the PaRIS algorithm. *ArXiv e-prints*, Dec 2014.

[39] J. Olsson, O. Cappé, R. Douc, and E. Moulines. Sequential Monte Carlo smoothing with application to parameter estimation in non-linear state space models. *ArXiv Mathematics e-prints*, Sep 2006.

[40] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, 2004. ISBN 9781580538510.

[41] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, (58):527–535, 1952.

[42] D. Russo and B. V. Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[43] D. Russo and B. V. Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.

[44] S. L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010. ISSN 1526-4025. doi: 10.1002/asmb.874.

[45] S. L. Scott. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31:37–49, 2015. Special issue on actual impact and future perspectives on stochastic modelling in business and industry.

[46] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press: Cambridge, MA, 1998.

[47] W. R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444.

[48] W. R. Thompson. On the Theory of Apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935. ISSN 00029327, 10806377.

[49] I. Urteaga and C. Wiggins. Variational inference for the multi-armed contextual bandit. In A. Storkey and F. Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 698–706, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

[50] P. J. van Leeuwen. Particle Filtering in Geophysical Systems. *Monthly Weather Review*, 12(137):4089–4114., 2009.

A MAB models

We now describe the key distributions for some MAB models of interest.

A.1 Bernoulli rewards

The Bernoulli distribution is well suited for applications with binary returns (i.e., success or failure of an action) that don't depend on a context. The rewards $y \in \{0, 1\}$ of each arm are modeled as independent draws from a Bernoulli distribution with success probabilities θ_a, i.e.,

$$p_a(y|\theta) = \text{Ber}\left(y|\theta_a\right) = \theta_a^y (1-\theta_a)^{(1-y)} \ . \tag{27}$$

For this reward distribution, the parameter conjugate prior distribution is the Beta distribution

$$p(\theta_a|\alpha_{a,0}, \beta_{a,0}) = \text{Beta}\left(\theta_a|\alpha_{a,0}, \beta_{a,0}\right) = \frac{\Gamma\left(\alpha_0 + \beta_0\right)}{\Gamma\left(\alpha_0\right)\Gamma\left(\beta_0\right)} \theta_a^{\alpha_0 - 1} (1-\theta_0)^{\beta_0 - 1} \ . \tag{28}$$

After observing actions $a_{1:t}$ and rewards $y_{1:t}$, the parameter posterior follows an updated Beta distribution

$$p(\theta_a|a_{1:t}, y_{1:t}, \alpha_{a,0}, \beta_{a,0}) = p(\theta_a|\alpha_{a,t}, \beta_{a,t}) = \text{Beta}\left(\theta_a|\alpha_{a,t}, \beta_{a,t}\right) \ , \tag{29}$$

with sequential updates

$$\begin{cases} \alpha_{a,t} = \alpha_{a,t-1} + y_t \cdot \mathbb{1}[a_t = a] \ , \\ \beta_{a,t} = \beta_{a,t-1} + (1-y_t) \cdot \mathbb{1}[a_t = a] \ , \end{cases} \tag{30}$$

or, alternatively, batch updates

$$\begin{cases} \alpha_{a,t} = \alpha_{a,0} + \sum_{t|a_t=a} y_t \ , \\ \beta_{a,t} = \beta_{a,0} + \sum_{t|a_t=a} (1-y_t) \ . \end{cases} \tag{31}$$

The expected reward for each arm follows

$$p(\mu_a|\theta_a) = p(\theta_a|a_{1:t}, y_{1:t}) = \text{Beta}\left(\theta_a|\alpha_{a,t}, \beta_{a,t}\right) \ , \tag{32}$$

and the quantile function is based on the Beta distribution

$$q_{a,t+1}(\alpha_{t+1}) = Q\left(1-\alpha_{t+1}, \text{Beta}\left(\theta_a|\alpha_{a,t}, \beta_{a,t}\right)\right) \ . \tag{33}$$

A.2 Contextual linear-Gaussian rewards

For bandits with continuous rewards, the Gaussian distribution is often applicable, where contextual dependencies can also be included. The contextual linear-Gaussian reward model is suited for these scenarios, where the expected reward of each arm is modeled as a linear combination of a d-dimensional context vector $x \in \mathbb{R}^d$ and the idiosyncratic parameters of the arm $w_a \in \mathbb{R}^d$, i.e.,

$$p_a(y|x, \theta) = \mathcal{N}\left(y|x^\top w_a, \sigma_a^2\right) = \frac{e^{-\frac{(y-x^\top w_a)^2}{2\sigma_a^2}}}{\sqrt{2\pi\sigma_a^2}} \ . \tag{34}$$

We denote as $\theta \equiv \{w, \sigma\}$ the set of all the parameters.

For this reward distribution, the parameter conjugate prior distribution is the Normal Inverse Gamma distribution

$$\begin{aligned} p(w_a, \sigma_a^2 | u_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}) &= \text{NIG}\left(w_a, \sigma_a^2 | u_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}\right) \\ &= \mathcal{N}\left(w_a | u_{a,0}, \sigma_a^2 V_{a,0}\right) \cdot \Gamma^{-1} \sigma_a^2 | \alpha_{a,0}, \beta_{a,0} \\ &= \frac{e^{-\frac{1}{2}(w_a - u_{a,0})^\top (\sigma_a^2 V_{a,0})^{-1}(w_a - u_{a,0})}}{(2\pi)^{1/2} \sigma_a \left|V_{a,0}\right|^{-1/2}} \cdot \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} (\sigma_a^2)^{-\alpha_0 - 1} e^{-\frac{\beta_0}{(\sigma_a^2)}} . \end{aligned} \tag{35}$$

After observing actions $a_{1:t}$ and rewards $y_{1:t}$, the parameter posterior follows an updated NIG distribution

$$\begin{aligned} p(w_a, \sigma_a^2 | a_{1:t}, y_{1:t}, u_{a,0}, V_{a,0}, \alpha_{a,0}, \beta_{a,0}) &= p\left(w_a, \sigma_a^2 | u_{a,t}, V_{a,t}, \alpha_{a,t}, \beta_{a,t}\right) \\ &= \text{NIG}\left(w_a, \sigma_a^2 | u_{a,t}, V_{a,t}, \alpha_{a,t}, \beta_{a,t}\right) , \end{aligned} \tag{36}$$

with sequential updates

$$\begin{cases} V_{a,t}^{-1} = V_{a,t-1}^{-1} + x_t x_t^\top \cdot \mathbb{1}[a_t = a] , \\ u_{a,t} = V_{a,t}\left(V_{a,t-1}^{-1} u_{a,t-1} + x_t y_t \cdot \mathbb{1}[a_t = a]\right) , \\ \alpha_{a,t} = \alpha_{a,t-1} + \frac{\mathbb{1}[a_t = a]}{2} , \\ \beta_{a,t} = \beta_{a,t-1} + \frac{\mathbb{1}[a_t = a](y_{t_a} - x_t^\top u_{a,t-1})^2}{2\left(1 + x_t^\top V_{a,t-1} x_t\right)} , \end{cases} \tag{37}$$

or, alternatively, batch updates

$$\begin{cases} V_{a,t}^{-1} = V_{a,0}^{-1} + x_{1:t|t_a} x_{1:t|t_a}^\top , \\ u_{a,t} = V_{a,t}\left(V_{a,0}^{-1} u_{a,0} + x_{1:t|t_a} y_{1:t|t_a}\right) , \\ \alpha_{a,t} = \alpha_{a,0} + \frac{|t_a|}{2} , \\ \beta_{a,t} = \beta_{a,0} + \frac{\left(y_{1:t|t_a}^\top y_{1:t|t_a} + u_{a,0}^\top V_{a,0}^{-1} u_{a,0} - u_{a,t}^\top V_{a,t}^{-1} u_{a,t}\right)}{2} , \end{cases} \tag{38}$$

where $t_a = \{t | a_t = a\}$ indicates the set of time instances when arm a is played.

The expected reward for each arm follows

$$p(\mu_a | x, \sigma_a^2, u_{a,t}, V_{a,t}) = \mathcal{N}\left(\mu_a | x^\top u_{a,t},\ \sigma_a^2 \cdot x^\top V_{a,t} x\right) . \tag{39}$$

and the quantile function is based on this Gaussian distribution

$$q_{a,t+1}(\alpha_{t+1}) = Q\left(1 - \alpha_{t+1}, \mathcal{N}\left(\mu_a | x^\top u_{a,t},\ \sigma_a^2 \cdot x^\top V_{a,t} x\right)\right) . \tag{40}$$

For the more realistic scenario where the reward variance σ_a^2 is unknown, we can marginalize it and obtain

$$\begin{aligned} p(\mu_a | x, u_{a,t}, \sigma_a^2, V_{a,t}) &= \mathcal{T}\left(\mu_a | 2\alpha_{a,t}, x^\top u_{a,t},\ \frac{\beta_{a,t}}{\alpha_{a,t}} \cdot x^\top V_{a,t} x\right) \\ &= \frac{\Gamma\left(\frac{2\alpha_{a,t}+1}{2}\right)}{\Gamma\left(\frac{2\alpha_{a,t}}{2}\right) \sqrt{\pi 2\alpha_{a,t} \frac{\beta_{a,t}}{\alpha_{a,t}} x^\top V_{a,t} x}} \cdot \left(1 + \frac{1}{(2\alpha_{a,t})} \left(\frac{(\mu_a - x^\top u_{a,t})^2}{\frac{\beta_{a,t}}{\alpha_{a,t}} \cdot x^\top V_{a,t} x}\right)\right)^{-\frac{2\alpha_{a,t}+1}{2}} . \end{aligned} \tag{41}$$

The quantile function for this case is based on the Student's-t distribution

$$q_{a,t+1}(\alpha_{t+1}) = Q\left(1-\alpha_{t+1}, \mathcal{T}\left(\mu_a | 2\alpha_{a,t}, x^\top u_{a,t}, \tfrac{\beta_{a,t}}{\alpha_{a,t}} \cdot x^\top V_{a,t} x\right)\right) . \tag{42}$$

Note that one can use the above results for bandits with no context, by replacing $x = I$ and obtaining $\mu_a = u_{a,t}$.

A.3 Contextual linear logistic rewards

The logistic function is applicable for problems where returns are binary (i.e., success or failure of an action), but depend on a d-dimensional context vector $x \in \mathbb{R}^d$ and idiosyncratic parameters of each arm θ_a. The contextual linear logistic reward model follows

$$p_a(y|x,\theta) = \frac{e^{y\cdot(x^\top \theta_a)}}{1+e^{(x^\top \theta_a)}} . \tag{43}$$

For this reward distribution, the posterior of the parameters can not be computed in closed form and, neither, the quantile function of the expected rewards $\mu_a = y \cdot (x^\top \theta_a)$.

A.4 Linearly dynamic bandits

Let us consider a general linear model for the dynamics of the parameters of each arm $\theta_a \in \mathbb{R}^d$:

$$\theta_{a,t} = L_a \theta_{a,t-1} + \epsilon_a , \qquad \epsilon_a \sim \mathcal{N}\left(\epsilon_a | 0, C_a\right) , \tag{44}$$

where $L_a \in \mathbb{R}^{d\times d}$ and $C_a \in \mathbb{R}^{d\times d}$. One can immediately determine that, for linearly dynamic bandits with known parameters, the parameters follow

$$\theta_{a,t} \sim \mathcal{N}\left(\theta_{a,t} | L_a \theta_{a,t-1}, C_a\right) . \tag{45}$$

However, it is unrealistic to assume that the parameters are known in practice. We thus marginalize them out by means of the following conjugate priors for the matrix A and covariance matrix C (we drop the per arm subscript a for clarity)

$$\begin{aligned} p(A, C | L_0, B_0, \nu_0, V_0) &= \mathrm{NIW}\left(A, C | L_0, B_0, \nu_0, V_0\right) = p(A|L_0, B_0, C) p(C|\nu_0, V_0) \\ &= \mathcal{MN}\left(A | L_0, C, B_0\right) \mathrm{IW}\left(C|\nu_0, V_0\right) , \end{aligned} \tag{46}$$

where the matrix variate Gaussian distribution follows

$$\mathcal{MN}\left(A|L_0, C, B_0\right) = \frac{e^{-\frac{1}{2}\mathrm{tr}\left\{B_0^{-1}(A-L_0)^\top C^{-1}(A-L_0)\right\}}}{(2\pi)^{(d\cdot d)/2}\left|B_0\right|^{d/2}\left|C\right|^{d/2}} , \tag{47}$$

and the Inverse Wishart

$$\mathrm{IW}\left(C|\nu_0, V_0\right) = \frac{\left|C\right|^{-\frac{\nu_0+d+1}{2}} e^{-\frac{1}{2}\mathrm{tr}\left\{C^{-1}V_0\right\}}}{2^{\frac{\nu_0 \cdot d}{2}}\left|V_0\right|^{-\frac{\nu_0}{2}}\Gamma\left(\frac{\nu_0}{2}\right)} . \tag{48}$$

We integrate out the unknown parameters A and C to derive the predictive density, i.e., the distribution of θ_t, given all the past data $\theta_{1:t}$. One can show that the resulting distribution is a multivariate t-distribution

$$f(\theta_t|\theta_{1:t-1}) = \mathcal{T}\left(\theta_t|\nu_t, m_t, R_t\right) \propto \left|1 + \frac{1}{\nu_t}\left(\theta_t - m_t\right) R_t^{-1}\left(\theta_t - m_t\right)^\top\right|^{-\frac{\nu_t+d}{2}} , \tag{49}$$

where ν_t denotes degrees of freedom, $m_t \in \mathbb{R}^d$ is the location parameter, and $R_t \in \mathbb{R}^{d\times d}$ represents the scale matrix (6). These follow

$$\begin{cases} \nu_t = \nu_0 + t - d \ , \\ m_t = L_{t-1}\theta_{t-1} \ , \\ R_t = \frac{V_{t-1}}{\nu_t\left(1-\theta_{t-1}^\top (UU^\top)^{-1}\theta_{t-1}\right)} \ , \end{cases} \tag{50}$$

where the sufficient statistics of the parameters are

$$\begin{cases} B_{t-1} = \left(\Theta_{0:t-2}\Theta_{0:t-2}^\top + B_0^{-1}\right)^{-1} \ , \\ L_{t-1} = \left(\Theta_{1:t-1}\Theta_{0:t-2}^\top + A_0 B_0^{-1}\right) B_{t-1} \ , \\ V_{t-1} = \left(\Theta_{1:t-1} - L_{t-1}\Theta_{0:t-2}\right)\left(\Theta_{1:t-1} - L_{t-1}\Theta_{0:t-2}\right)^\top \\ \qquad\qquad + \left(L_{t-1} - L_0\right) B_0^{-1} \left(L_{t-1} - L_0\right)^\top + V_0 \ , \\ UU^\top = \left(\theta_{t-1}\theta_{t-1}^\top + B_{t-1}^{-1}\right) \ , \end{cases} \tag{51}$$

and we have defined the stacked parameter matrix

$$\Theta_{t_0:t_1} = [\theta_{t_0}\theta_{t_0+1}\cdots\theta_{t_1-1}\theta_{a,t_1}] \in \mathbb{R}^{d\times(t_1-t_0)} \ . \tag{52}$$

All in all, for linear dynamic bandits with unknown parameters, the per-arm parameters follow

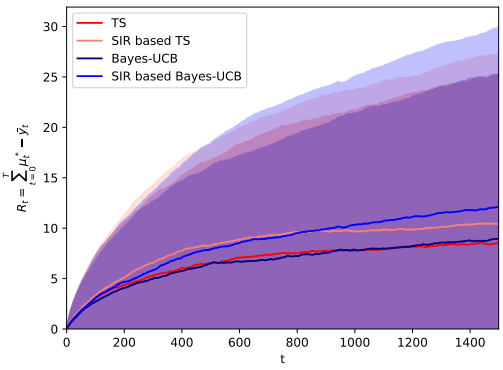
$$\theta_{a,t} \sim \mathcal{T}\left(\theta_{a,t}|\nu_{a,t}, m_{a,t}, R_{a,t}\right) \ . \tag{53}$$

B Evaluation

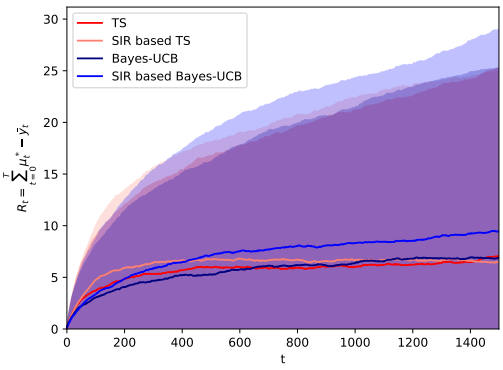
In the following pages, we provide results for other parameterizations of the evaluated bandits.

B.1 Static bandits

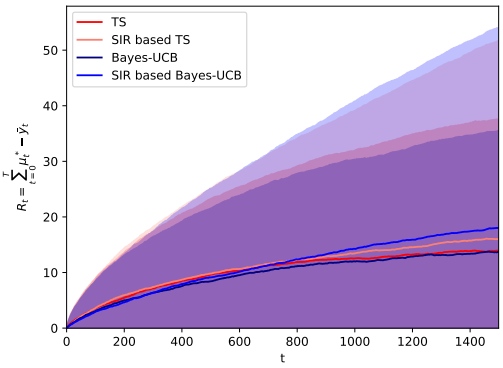
B.2 Bernoulli bandits, A=2



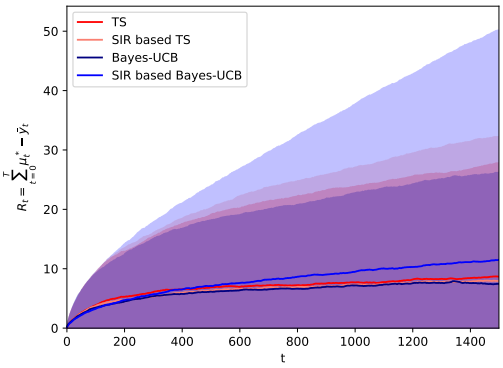
(a) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.2$.



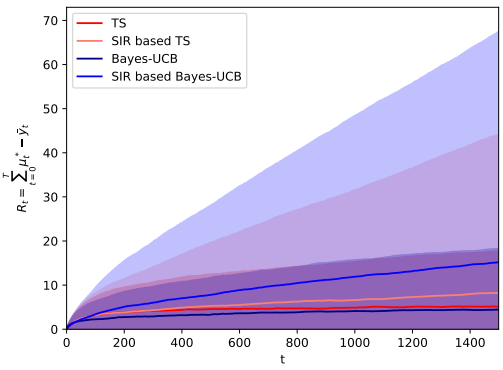
(b) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.3$.



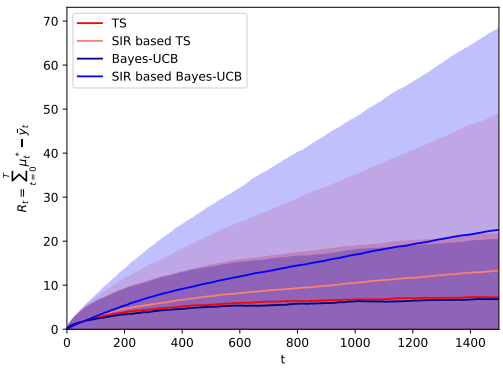
(c) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.6$.



(d) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.7$.

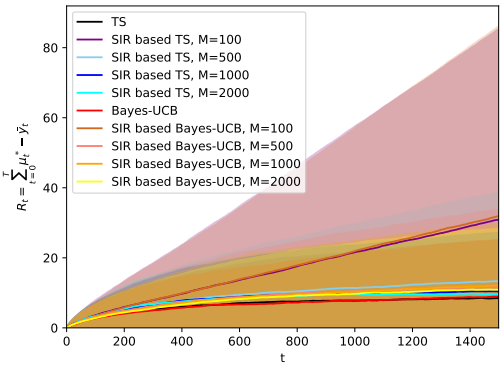


(e) Bernoulli bandit, $\theta_0 = 0.7, \theta_1 = 0.9$.

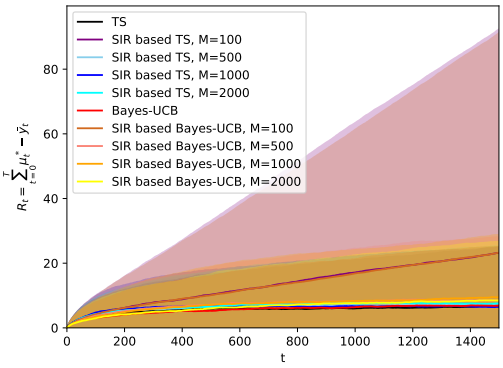


(f) Bernoulli bandit, $\theta_0 = 0.8, \theta_1 = 0.9$.

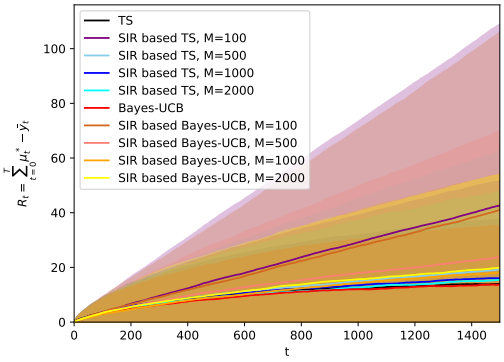
Figure 9: Mean regret (standard deviation shown as shaded region) in static two-armed Bernoulli bandits.



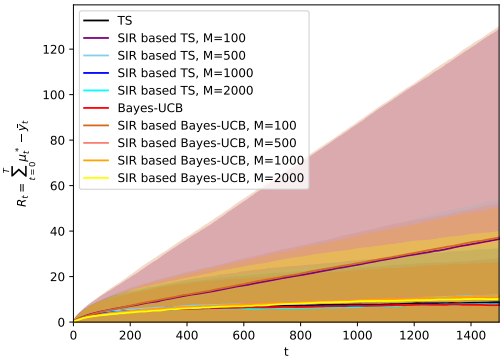
(a) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.2$.



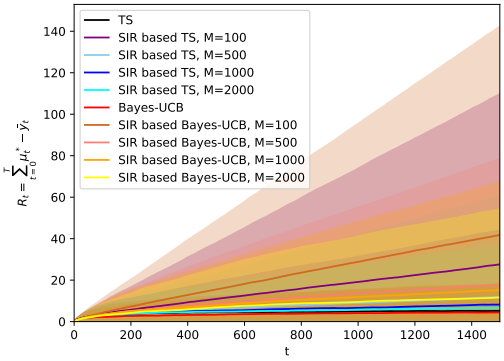
(b) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.3$.



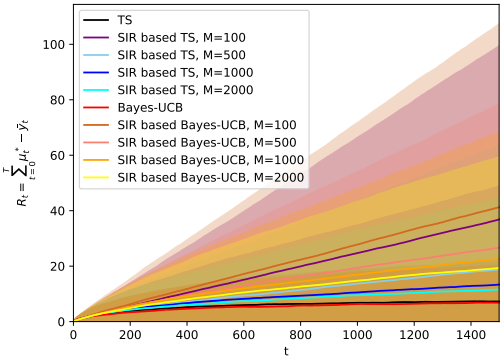
(c) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.6$.



(d) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.7$.



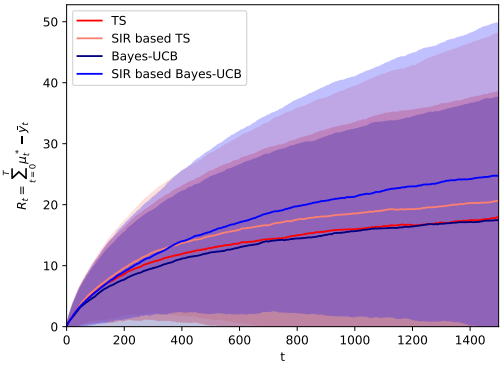
(e) Bernoulli bandit, $\theta_0 = 0.7, \theta_1 = 0.9$.



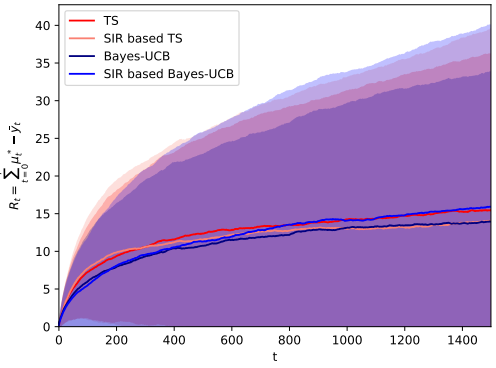
(f) Bernoulli bandit, $\theta_0 = 0.8, \theta_1 = 0.9$.

Figure 10: Mean regret (standard deviation shown as shaded region) in static two-armed Bernoulli bandits.

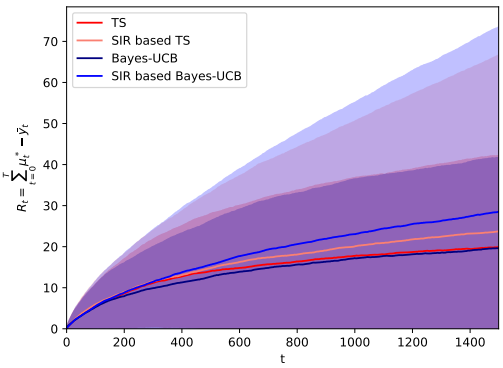
B.3 Bernoulli bandits, A=3



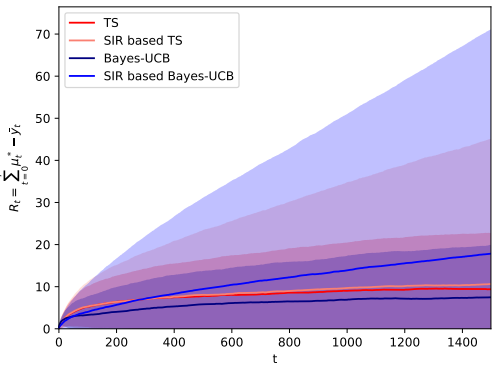
(a) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.2, \theta_2 = 0.3$.



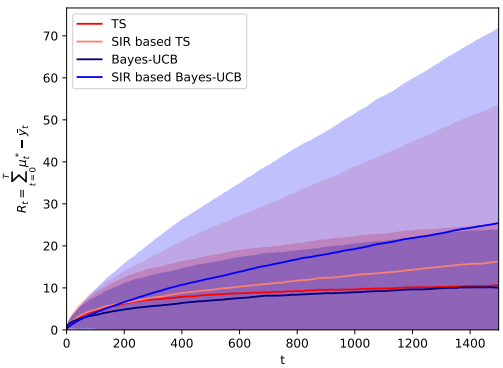
(b) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.3, \theta_2 = 0.5$.



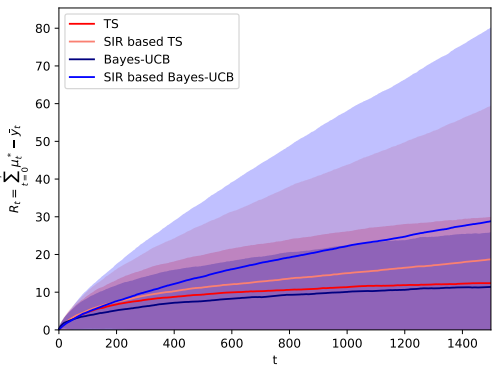
(c) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.6, \theta_2 = 0.7$.



(d) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.7, \theta_2 = 0.9$.

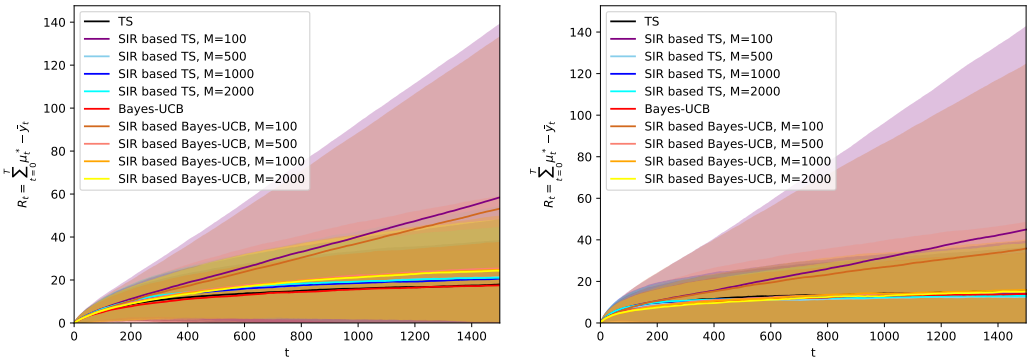


(e) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.8, \theta_2 = 0.9$.

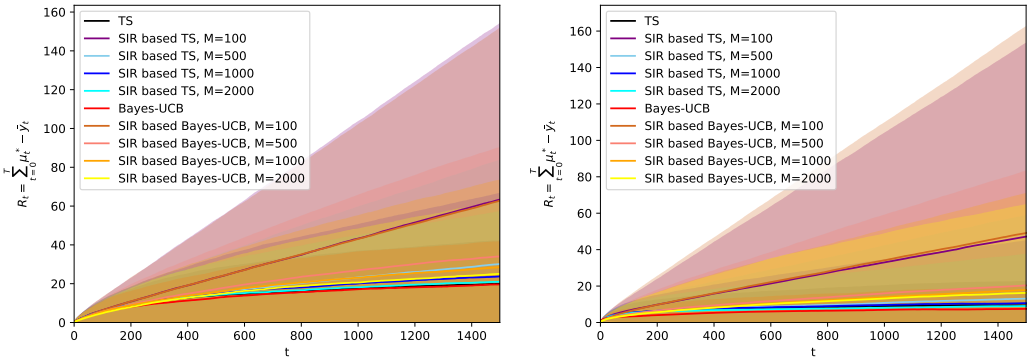


(f) Bernoulli bandit, $\theta_0 = 0.7, \theta_1 = 0.8, \theta_2 = 0.9$.

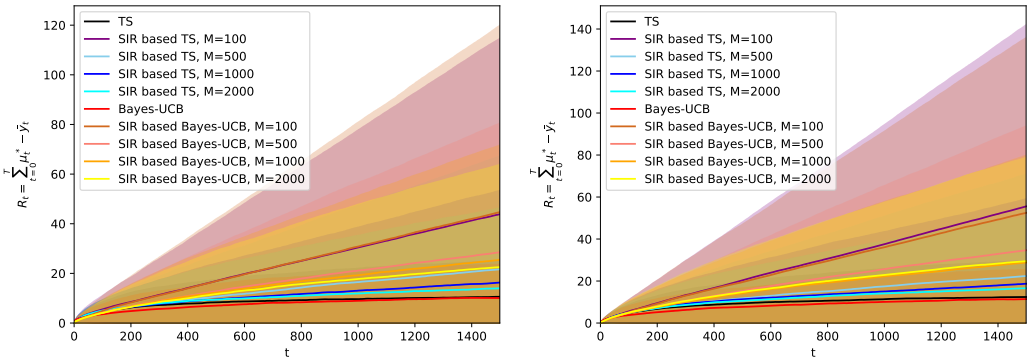
Figure 11: Mean regret (standard deviation shown as shaded region) in static three-armed Bernoulli bandits.



(a) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.2, \theta_2 = 0.3$. (b) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.3, \theta_2 = 0.5$.



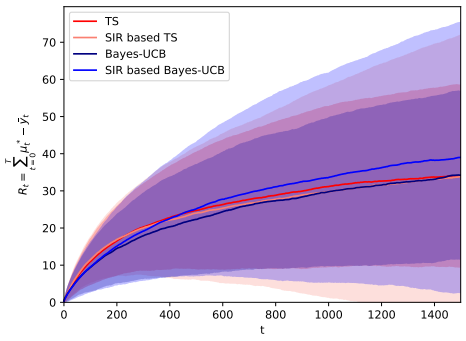
(c) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.6, \theta_2 = 0.7$. (d) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.7, \theta_2 = 0.9$.



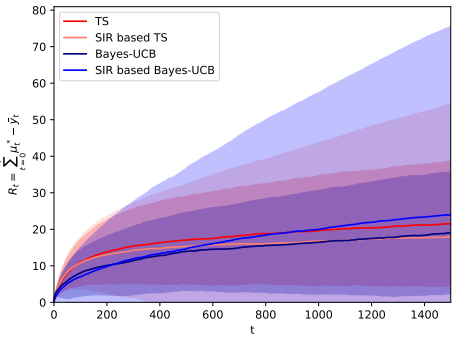
(e) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.8, \theta_2 = 0.9$. (f) Bernoulli bandit, $\theta_0 = 0.7, \theta_1 = 0.8, \theta_2 = 0.9$.

Figure 12: Mean regret (standard deviation shown as shaded region) in static three-armed Bernoulli bandits.

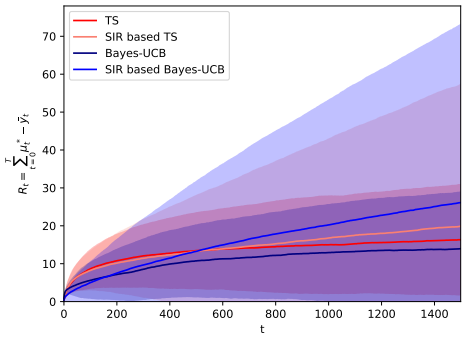
B.4 Bernoulli bandits, A=5



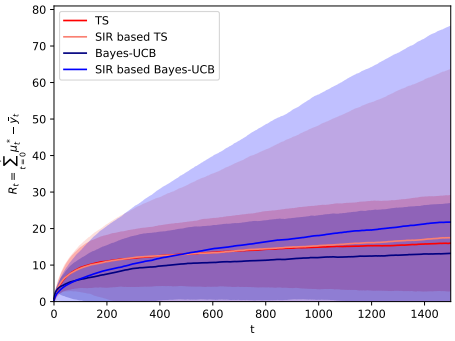
(a) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.2, \theta_2 = 0.3, \theta_3 = 0.4, \theta_4 = 0.5$.



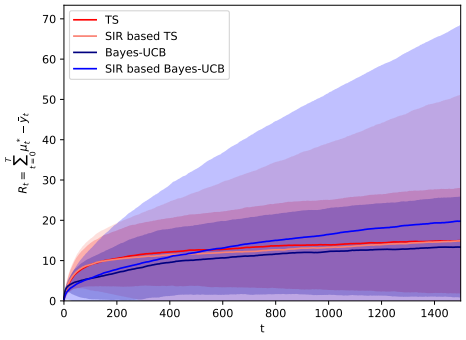
(b) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.3, \theta_2 = 0.5, \theta_3 = 0.6, \theta_4 = 0.8$.



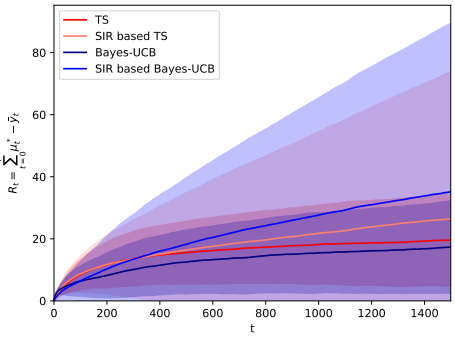
(c) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.3, \theta_2 = 0.5, \theta_3 = 0.8, \theta_4 = 0.9$.



(d) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.5, \theta_2 = 0.6, \theta_3 = 0.7, \theta_4 = 0.9$.

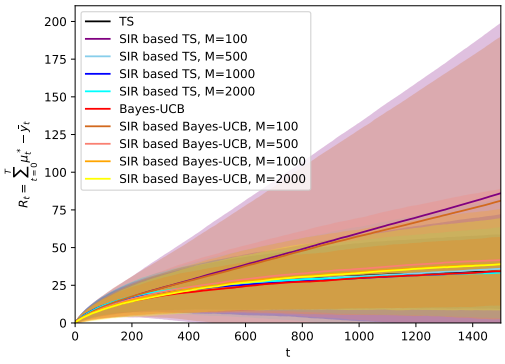


(e) Bernoulli bandit, $\theta_0 = 0.3, \theta_1 = 0.4, \theta_2 = 0.5, \theta_3 = 0.7, \theta_4 = 0.9$.

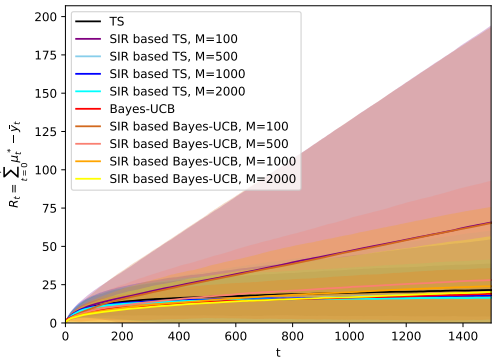


(f) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.6, \theta_2 = 0.7, \theta_3 = 0.8, \theta_4 = 0.9$.

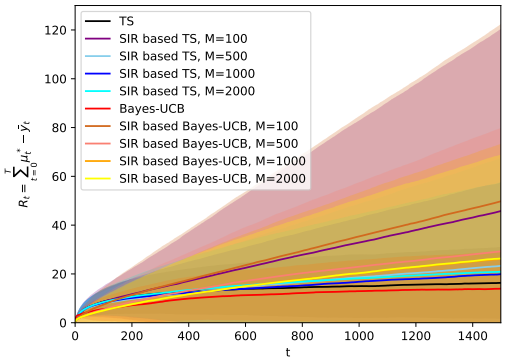
Figure 13: Mean regret (standard deviation shown as shaded region) in static five-armed Bernoulli bandits.



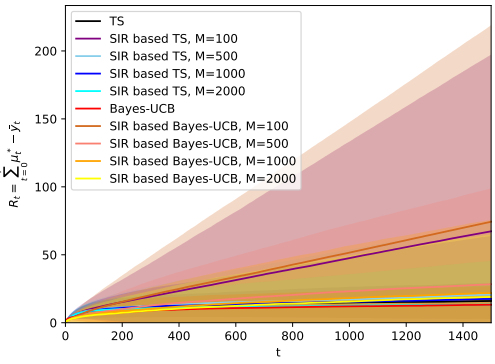
(a) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.2$, $\theta_2 = 0.3, \theta_3 = 0.4, \theta_4 = 0.5$.



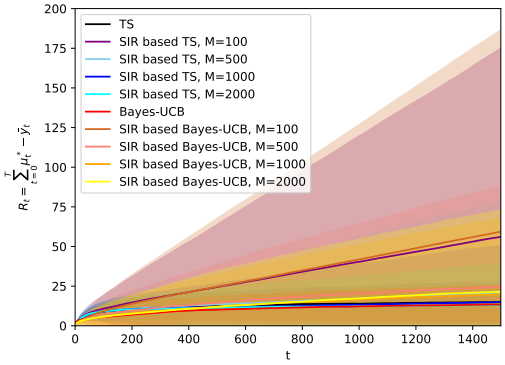
(b) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.3$, $\theta_2 = 0.5, \theta_3 = 0.6, \theta_4 = 0.8$.



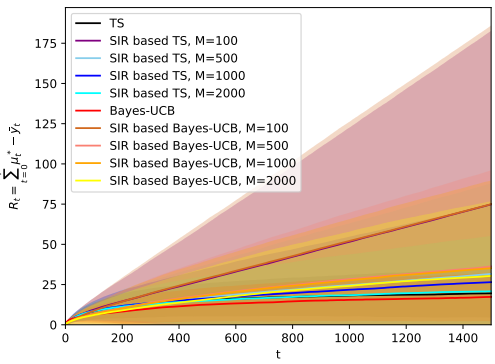
(c) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.3$, $\theta_2 = 0.5, \theta_3 = 0.8, \theta_4 = 0.9$.



(d) Bernoulli bandit, $\theta_0 = 0.1, \theta_1 = 0.5$, $\theta_2 = 0.6, \theta_3 = 0.7, \theta_4 = 0.9$.



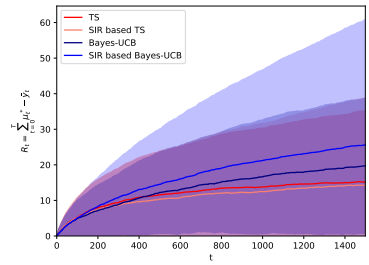
(e) Bernoulli bandit, $\theta_0 = 0.3, \theta_1 = 0.4$, $\theta_2 = 0.5, \theta_3 = 0.7, \theta_4 = 0.9$.



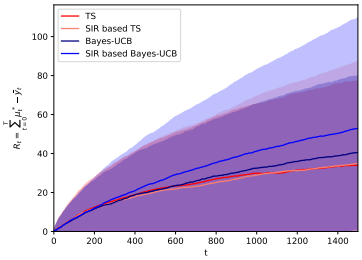
(f) Bernoulli bandit, $\theta_0 = 0.5, \theta_1 = 0.6$, $\theta_2 = 0.7, \theta_3 = 0.8, \theta_4 = 0.9$.

Figure 14: Mean regret (standard deviation shown as shaded region) in static five-armed Bernoulli bandits.

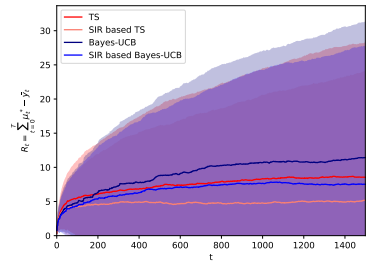
B.5 Contextual Linear Gaussian bandits, A=2



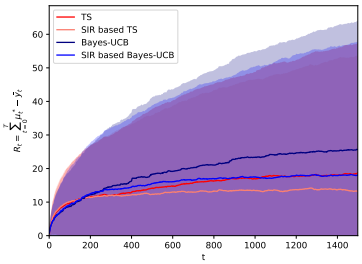
(a) Contextual Gaussian bandit, $\theta_0 = (-0.1, -0.1), \theta_1 = (0.1, 0.1), \sigma^2 = 0.5$.



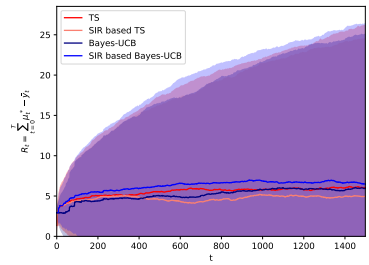
(b) Contextual Gaussian bandit, $\theta_0 = (-0.1, -0.1), \theta_1 = (0.1, 0.1), \sigma^2 = 1.0$.



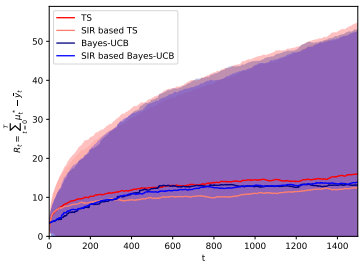
(c) Contextual Gaussian bandit, $\theta_0 = (-0.5, -0.5), \theta_1 = (0.5, 0.5), \sigma^2 = 0.5$.



(d) Contextual Gaussian bandit, $\theta_0 = (-0.5, -0.5), \theta_1 = (0.5, 0.5), \sigma^2 = 1.0$.

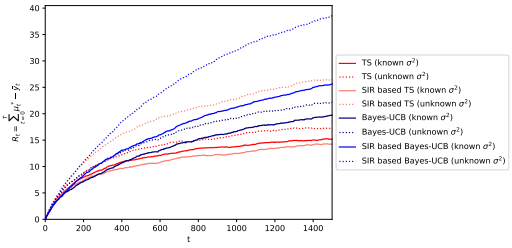


(e) Contextual Gaussian bandit, $\theta_0 = (-1, -1), \theta_1 = (1, 1), \sigma^2 = 0.5$.

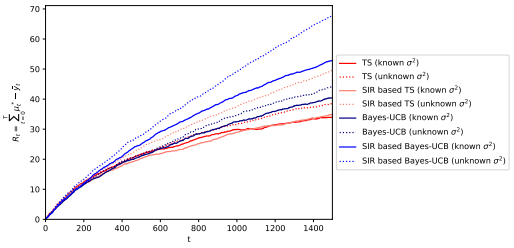


(f) Contextual Gaussian bandit, $\theta_0 = (-1, -1), \theta_1 = (1, 1), \sigma^2 = 1.0$.

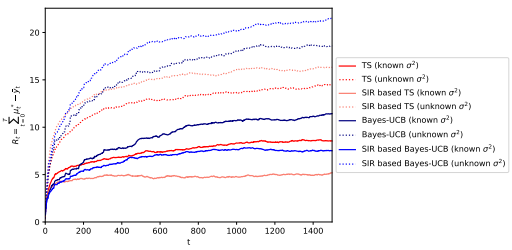
Figure 15: Mean regret (standard deviation shown as shaded region) in static two-armed contextual Gaussian bandits with reward variance.



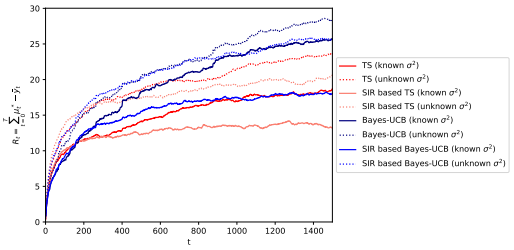
(a) Contextual Gaussian bandit, $\theta_0 = (-0.1, -0.1), \theta_1 = (0.1, 0.1), \sigma^2 = 0.5$.



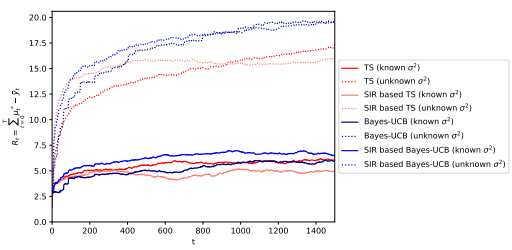
(b) Contextual Gaussian bandit, $\theta_0 = (-0.1, -0.1), \theta_1 = (0.1, 0.1), \sigma^2 = 1.0$.



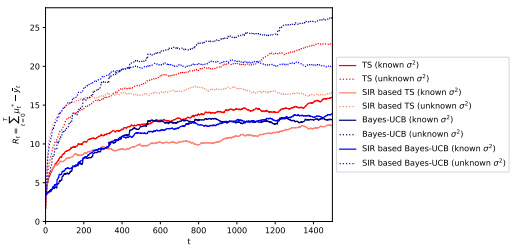
(c) Contextual Gaussian bandit, $\theta_0 = (-0.5, -0.5), \theta_1 = (0.5, 0.5), \sigma^2 = 0.5$.



(d) Contextual Gaussian bandit, $\theta_0 = (-0.5, -0.5), \theta_1 = (0.5, 0.5), \sigma^2 = 1.0$.



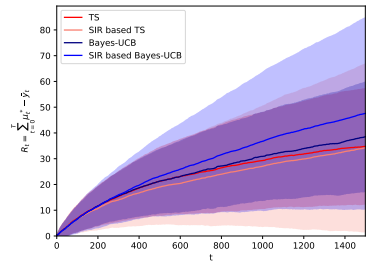
(e) Contextual Gaussian bandit, $\theta_0 = (-1, -1), \theta_1 = (1, 1), \sigma^2 = 0.5$.



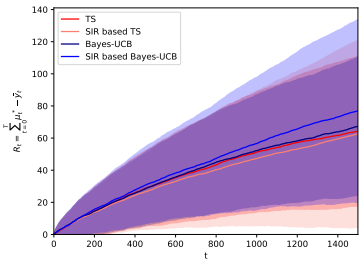
(f) Contextual Gaussian bandit, $\theta_0 = (-1, -1), \theta_1 = (1, 1), \sigma^2 = 1.0$.

Figure 16: Mean regret (standard deviation shown as shaded region) in static two-armed contextual Gaussian bandits with unknown reward variance.

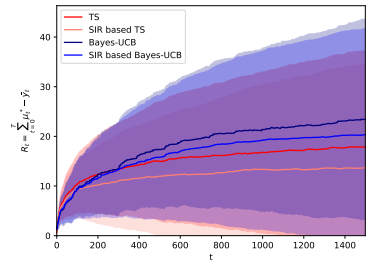
B.6 Contextual Linear Gaussian bandits, A=3



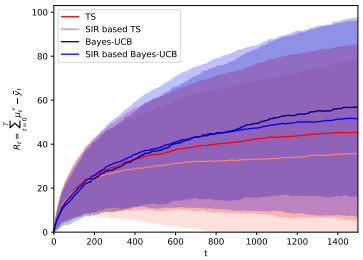
(a) Contextual Gaussian bandit, $\theta_0 = (-0.1, -0.1), \theta_1 = (0, 0), \theta_2 = (0.1, 0.1), \sigma^2 = 0.5$.



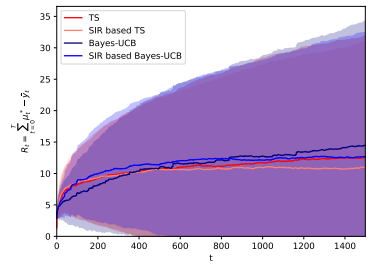
(b) Contextual Gaussian bandit, $\theta_0 = (-0.1, -0.1), \theta_1 = (0, 0), \theta_2 = (0.1, 0.1), \sigma^2 = 1.0$.



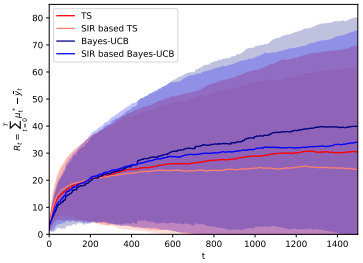
(c) Contextual Gaussian bandit, $\theta_0 = (-0.5, -0.5), \theta_1 = (0, 0), \theta_2 = (0.5, 0.5), \sigma^2 = 0.5$.



(d) Contextual Gaussian bandit, $\theta_0 = (-0.5, -0.5), \theta_1 = (0, 0), \theta_2 = (0.5, 0.5), \sigma^2 = 1.0$.

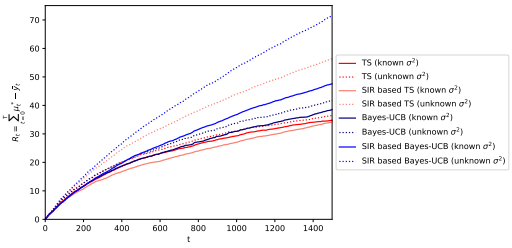


(e) Contextual Gaussian bandit, $\theta_0 = (-1, -1), \theta_1 = (0, 0), \theta_2 = (1, 1), \sigma^2 = 0.5$.

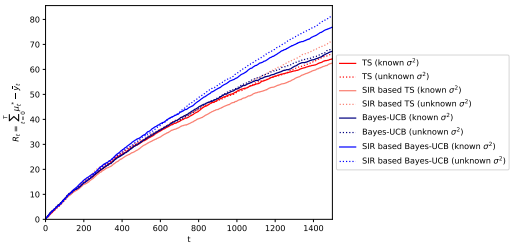


(f) Contextual Gaussian bandit, $\theta_0 = (-1, -1), \theta_1 = (0, 0), \theta_2 = (1, 1), \sigma^2 = 1.0$.

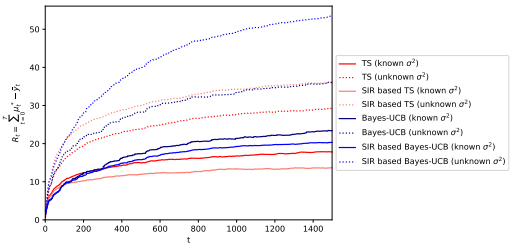
Figure 17: Mean regret (standard deviation shown as shaded region) in static three-armed contextual Gaussian bandits with reward variance.



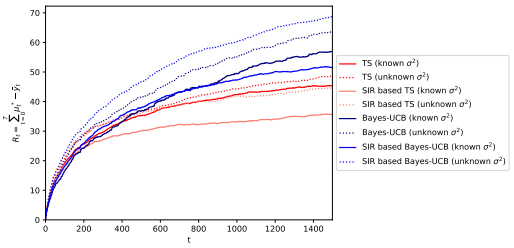
(a) Contextual Gaussian bandit, $\theta_0 = (-0.1, -0.1), \theta_1 = (0, 0), \theta_2 = (0.1, 0.1), \sigma^2 = 0.5$.



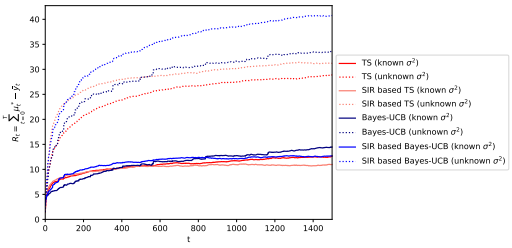
(b) Contextual Gaussian bandit, $\theta_0 = (-0.1, -0.1), \theta_1 = (0, 0), \theta_2 = (0.1, 0.1), \sigma^2 = 1.0$.



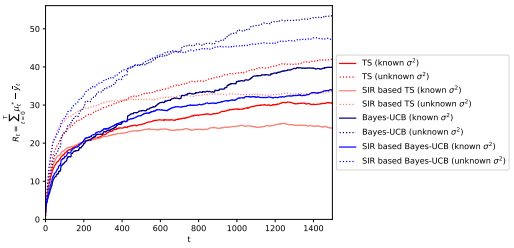
(c) Contextual Gaussian bandit, $\theta_0 = (-0.5, -0.5), \theta_1 = (0, 0), \theta_2 = (0.5, 0.5), \sigma^2 = 0.5$.



(d) Contextual Gaussian bandit, $\theta_0 = (-0.5, -0.5), \theta_1 = (0, 0), \theta_2 = (0.5, 0.5), \sigma^2 = 1.0$.



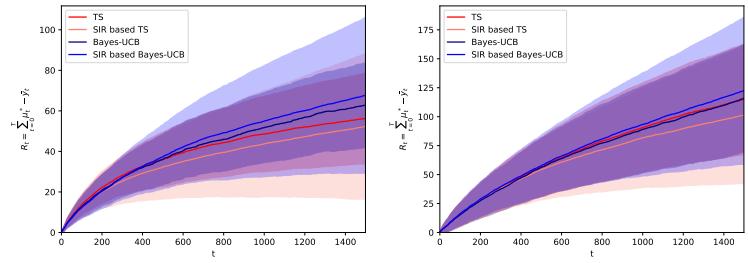
(e) Contextual Gaussian bandit, $\theta_0 = (-1, -1), \theta_1 = (0, 0), \theta_2 = (1, 1), \sigma^2 = 0.5$.



(f) Contextual Gaussian bandit, $\theta_0 = (-1, -1), \theta_1 = (0, 0), \theta_2 = (1, 1), \sigma^2 = 1.0$.

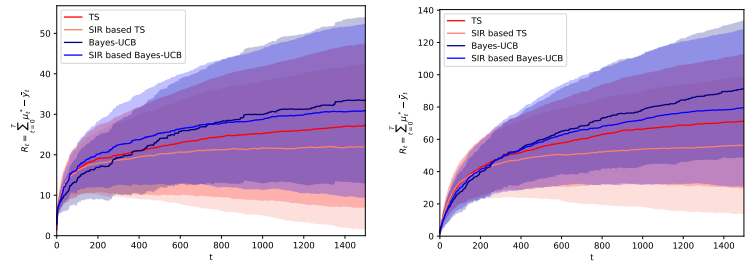
Figure 18: Mean regret (standard deviation shown as shaded region) in static three-armed contextual Gaussian bandits with unknown reward variance.

B.7 Contextual Linear Gaussian bandits, A=5



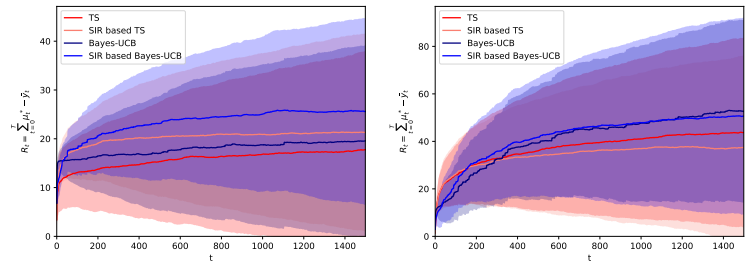
(a) Contextual Gaussian bandit, $\theta_0 = (-0.2, -0.2), \theta_1 = (-0.1, -0.1), \theta_2 = (0, 0), \theta_3 = (0.1, 0.1), \theta_4 = (0.2, 0.2), \sigma^2 = 0.5.$

(b) Contextual Gaussian bandit, $\theta_0 = (-0.2, -0.2), \theta_1 = (-0.1, -0.1), \theta_2 = (0, 0), \theta_3 = (0.1, 0.1), \theta_4 = (0.2, 0.2), \sigma^2 = 1.0.$



(c) Contextual Gaussian bandit, $\theta_0 = (-1.0, -1.0), \theta_1 = (-0.5, -0.5), \theta_2 = (0, 0), \theta_3 = (0.5, 0.5), \theta_4 = (1.0, 1.0), \sigma^2 = 0.5.$

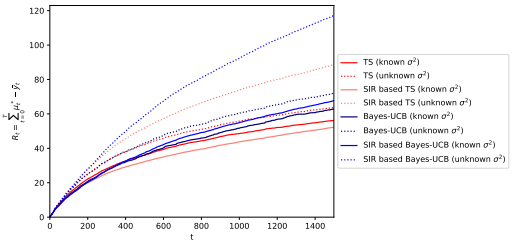
(d) Contextual Gaussian bandit, $\theta_0 = (-1.0, -1.0), \theta_1 = (-0.5, -0.5), \theta_2 = (0, 0), \theta_3 = (0.5, 0.5), \theta_4 = (1.0, 1.0), \sigma^2 = 1.0.$



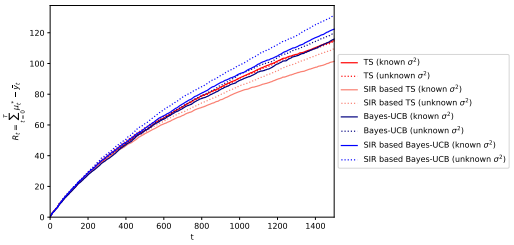
(e) Contextual Gaussian bandit, $\theta_0 = (-2, -2), \theta_1 = (-1, -1), \theta_2 = (0, 0), \theta_3 = (1, 1), \theta_4 = (2, 2), \sigma^2 = 0.5.$

(f) Contextual Gaussian bandit, $\theta_0 = (-2, -2), \theta_1 = (-1, -1), \theta_2 = (0, 0), \theta_3 = (1, 1), \theta_4 = (2, 2), \sigma^2 = 1.0.$

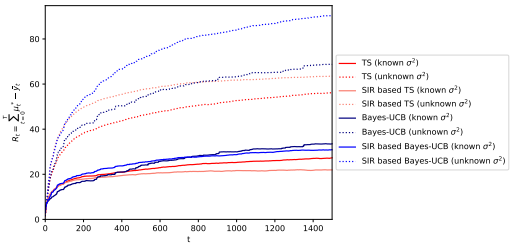
Figure 19: Mean regret (standard deviation shown as shaded region) in static five-armed contextual Gaussian bandits with reward variance.



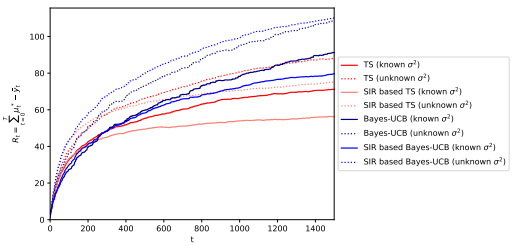
(a) Contextual Gaussian bandit,
$\theta_0 = (-0.2, -0.2), \theta_1 = (-0.1, -0.1), \theta_2 = (0, 0), \theta_3 = (0.1, 0.1), \theta_4 = (0.2, 0.2), \sigma^2 = 0.5.$



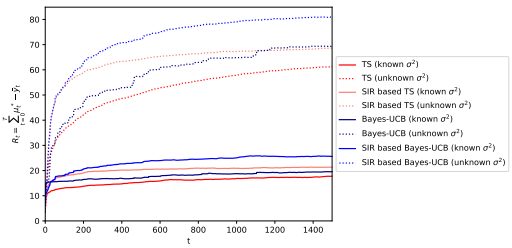
(b) Contextual Gaussian bandit,
$\theta_0 = (-0.2, -0.2), \theta_1 = (-0.1, -0.1), \theta_2 = (0, 0), \theta_3 = (0.1, 0.1), \theta_4 = (0.2, 0.2), \sigma^2 = 1.0.$



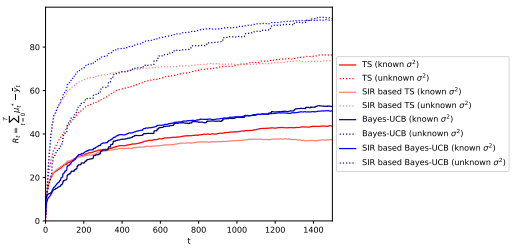
(c) Contextual Gaussian bandit,
$\theta_0 = (-1.0, -1.0), \theta_1 = (-0.5, -0.5), \theta_2 = (0, 0), \theta_3 = (0.5, 0.5), \theta_4 = (1.0, 1.0), \sigma^2 = 0.5.$



(d) Contextual Gaussian bandit,
$\theta_0 = (-1.0, -1.0), \theta_1 = (-0.5, -0.5), \theta_2 = (0, 0), \theta_3 = (0.5, 0.5), \theta_4 = (1.0, 1.0), \sigma^2 = 1.0.$



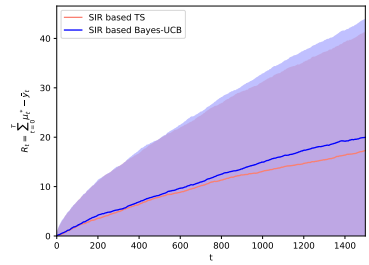
(e) Contextual Gaussian bandit,
$\theta_0 = (-2, -2), \theta_1 = (-1, -1), \theta_2 = (0, 0), \theta_3 = (1, 1), \theta_4 = (2, 2), \sigma^2 = 0.5.$



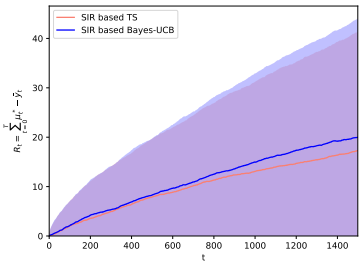
(f) Contextual Gaussian bandit,
$\theta_0 = (-2, -2), \theta_1 = (-1, -1), \theta_2 = (0, 0), \theta_3 = (1, 1), \theta_4 = (2, 2), \sigma^2 = 1.0.$

Figure 20: Mean regret (standard deviation shown as shaded region) in static five-armed contextual Gaussian bandits with unknown reward variance.

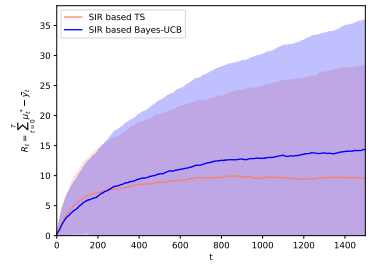
B.8 Contextual Logistic bandits, A=2



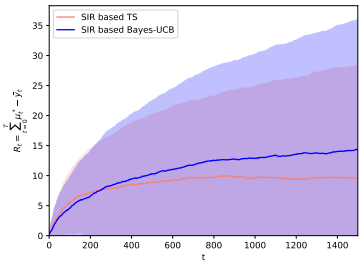
(a) Contextual Logistic bandit, $\theta_0 = (-0.1, -0.1), \theta_1 = (0.1, 0.1)$.



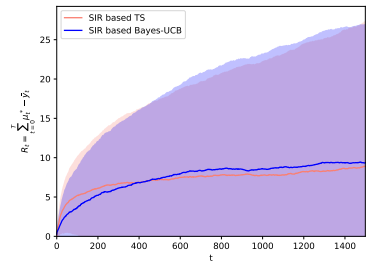
(b) Contextual Logistic bandit, $\theta_0 = (-0.1, -0.1), \theta_1 = (0.1, 0.1)$.



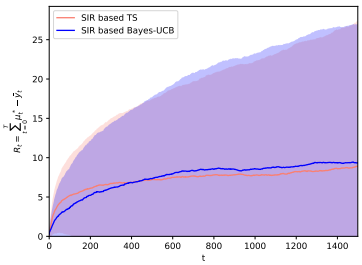
(c) Contextual Logistic bandit, $\theta_0 = (-0.5, -0.5), \theta_1 = (0.5, 0.5)$.



(d) Contextual Logistic bandit, $\theta_0 = (-0.5, -0.5), \theta_1 = (0.5, 0.5)$.



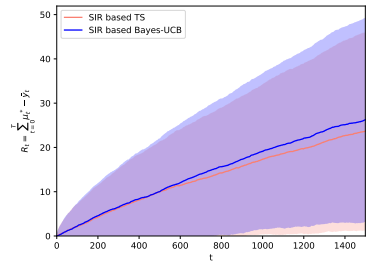
(e) Contextual Logistic bandit, $\theta_0 = (-1, -1), \theta_1 = (1, 1)$.



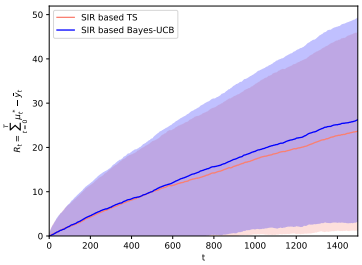
(f) Contextual Logistic bandit, $\theta_0 = (-1, -1), \theta_1 = (1, 1)$.

Figure 21: Mean regret (standard deviation shown as shaded region) in static two-armed contextual Logistic bandits.

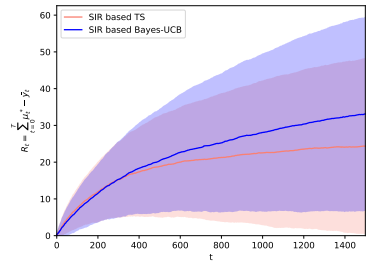
B.9 Contextual Logistic bandits, A=3



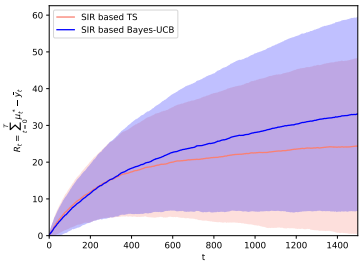
(a) Contextual Logistic bandit, $\theta_0 = (-0.1, -0.1), \theta_1 = (0, 0), \theta_2 = (0.1, 0.1)$.



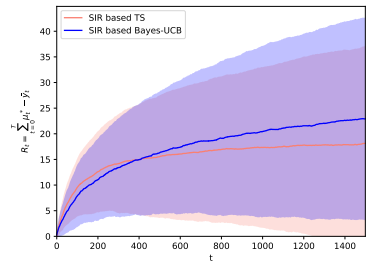
(b) Contextual Logistic bandit, $\theta_0 = (-0.1, -0.1), \theta_1 = (0, 0), \theta_2 = (0.1, 0.1)$.



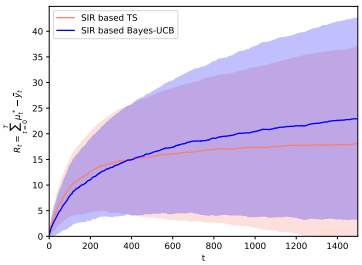
(c) Contextual Logistic bandit, $\theta_0 = (-0.5, -0.5), \theta_1 = (0, 0), \theta_2 = (0.5, 0.5)$.



(d) Contextual Logistic bandit, $\theta_0 = (-0.5, -0.5), \theta_1 = (0, 0), \theta_2 = (0.5, 0.5)$.



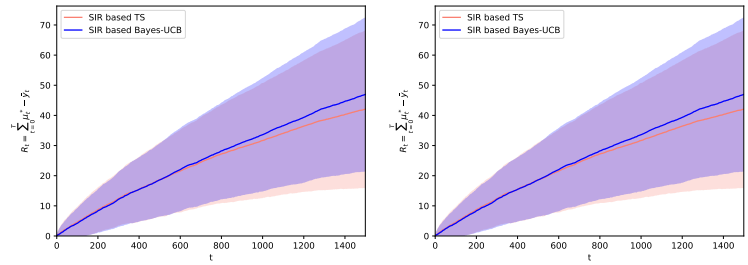
(e) Contextual Logistic bandit, $\theta_0 = (-1, -1), \theta_1 = (0, 0), \theta_2 = (1, 1)$.



(f) Contextual Logistic bandit, $\theta_0 = (-1, -1), \theta_1 = (0, 0), \theta_2 = (1, 1)$.

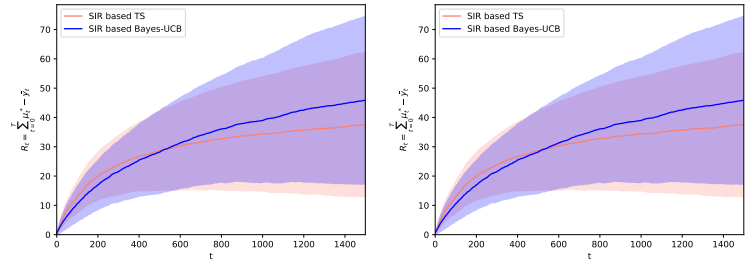
Figure 22: Mean regret (standard deviation shown as shaded region) in static three-armed contextual Logistic bandits.

B.10 Contextual Logistic bandits, A=5



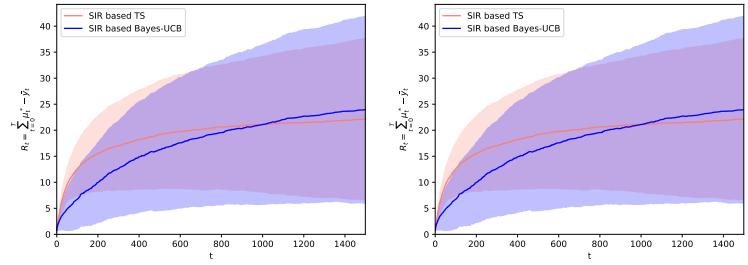
(a) Contextual Logistic bandit, $\theta_0 = (-0.2, -0.2), \theta_1 = (-0.1, -0.1), \theta_2 = (0, 0), \theta_3 = (0.1, 0.1), \theta_4 = (0.2, 0.2)$.

(b) Contextual Logistic bandit, $\theta_0 = (-0.2, -0.2), \theta_1 = (-0.1, -0.1), \theta_2 = (0, 0), \theta_3 = (0.1, 0.1), \theta_4 = (0.2, 0.2)$.



(c) Contextual Logistic bandit, $\theta_0 = (-1.0, -1.0), \theta_1 = (-0.5, -0.5), \theta_2 = (0, 0), \theta_3 = (0.5, 0.5), \theta_4 = (1.0, 1.0)$.

(d) Contextual Logistic bandit, $\theta_0 = (-1.0, -1.0), \theta_1 = (-0.5, -0.5), \theta_2 = (0, 0), \theta_3 = (0.5, 0.5), \theta_4 = (1.0, 1.0)$.



(e) Contextual Logistic bandit, $\theta_0 = (-2, -2), \theta_1 = (-1, -1), \theta_2 = (0, 0), \theta_3 = (1, 1), \theta_4 = (2, 2)$.

(f) Contextual Logistic bandit, $\theta_0 = (-2, -2), \theta_1 = (-1, -1), \theta_2 = (0, 0), \theta_3 = (1, 1), \theta_4 = (2, 2)$.

Figure 23: Mean regret (standard deviation shown as shaded region) in static five-armed contextual Logistic bandits.

B.11 Dynamic bandits

Fig. 24 illustrates the time-evolution of the expected rewards for a realization of the dynamics

$$\begin{aligned}\begin{pmatrix}\theta_{0,0,t}\\ \theta_{0,1,t}\end{pmatrix} &= \begin{pmatrix}0.9 & -0.1\\ -0.1 & 0.9\end{pmatrix}\begin{pmatrix}\theta_{0,0,t-1}\\ \theta_{0,1,t-1}\end{pmatrix} + \epsilon\,,\\ \begin{pmatrix}\theta_{1,0,t}\\ \theta_{1,1,t}\end{pmatrix} &= \begin{pmatrix}0.9 & 0.1\\ 0.1 & 0.9\end{pmatrix}\begin{pmatrix}\theta_{1,0,t-1}\\ \theta_{1,1,t-1}\end{pmatrix} + \epsilon\,,\end{aligned} \tag{54}$$

and $\epsilon \sim \mathcal{N}(\epsilon|0, 0.1 \cdot \mathrm{I})$.

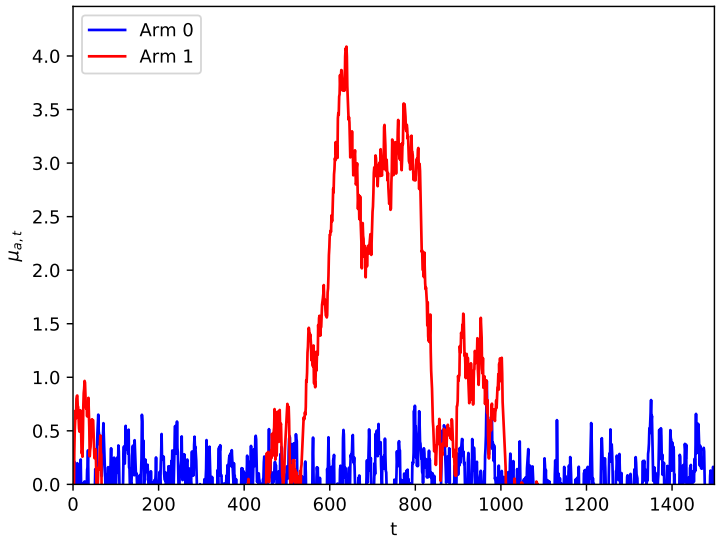


Figure 24: Expected per-arm rewards over time.

Fig. 25 illustrates the time-evolution of the expected rewards for a realization of the dynamics

$$\begin{aligned}\begin{pmatrix}\theta_{0,0,t}\\ \theta_{0,1,t}\end{pmatrix} &= \begin{pmatrix}0.5 & 0.0\\ 0.0 & 0.5\end{pmatrix}\begin{pmatrix}\theta_{0,0,t-1}\\ \theta_{0,1,t-1}\end{pmatrix} + \epsilon\,,\\ \begin{pmatrix}\theta_{1,0,t}\\ \theta_{1,1,t}\end{pmatrix} &= \begin{pmatrix}0.9 & 0.1\\ 0.1 & 0.9\end{pmatrix}\begin{pmatrix}\theta_{1,0,t-1}\\ \theta_{1,1,t-1}\end{pmatrix} + \epsilon\,,\end{aligned} \tag{55}$$

and $\epsilon \sim \mathcal{N}(\epsilon|0, 0.1 \cdot \mathrm{I})$.

Fig. 26 illustrates the time-evolution of the expected rewards for a realization of the static dynamics

$$\begin{aligned}\begin{pmatrix}\theta_{0,0,t}\\ \theta_{0,1,t}\end{pmatrix} &= \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}\theta_{0,0,t-1}\\ \theta_{0,1,t-1}\end{pmatrix} + \epsilon\,,\\ \begin{pmatrix}\theta_{1,0,t}\\ \theta_{1,1,t}\end{pmatrix} &= \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}\theta_{1,0,t-1}\\ \theta_{1,1,t-1}\end{pmatrix} + \epsilon\,,\end{aligned} \tag{56}$$

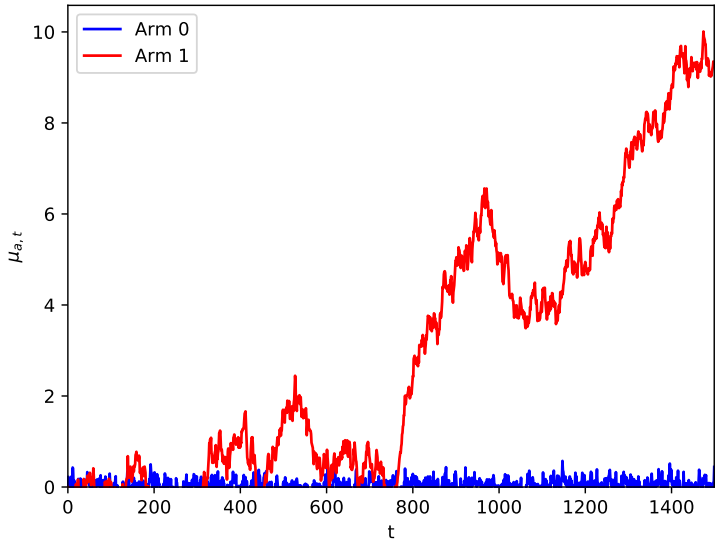


Figure 25: Expected per-arm rewards over time.

and $\epsilon \sim \mathcal{N}(\epsilon|0, 0.00001 \cdot \mathrm{I})$.

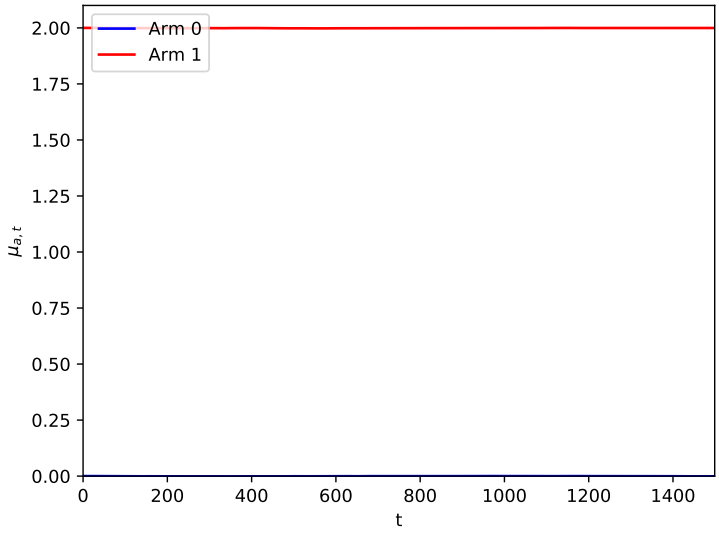
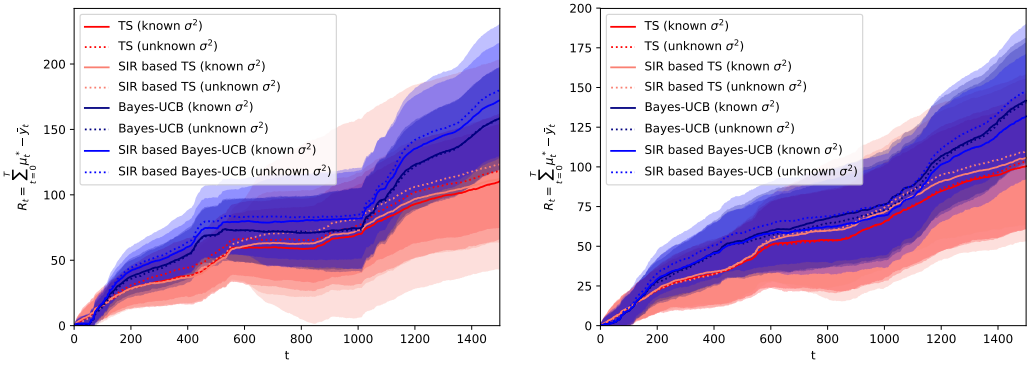
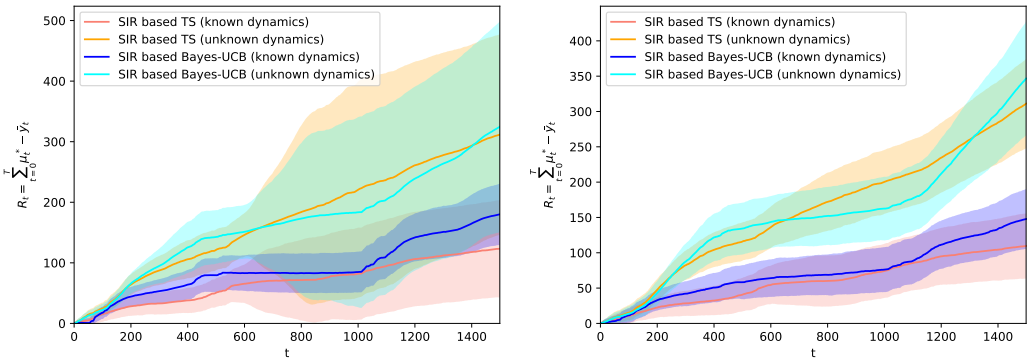


Figure 26: Expected per-arm rewards over time.



(a) Contextual logistic bandit with static context and known dynamics.

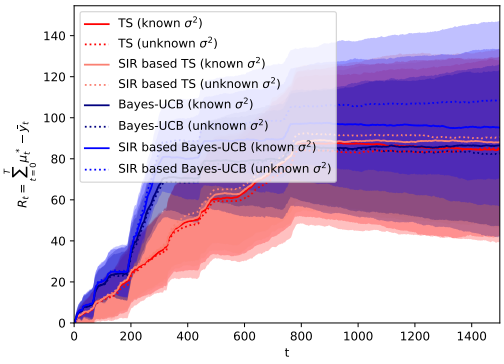
(b) Contextual logistic bandit with random context and known dynamics.



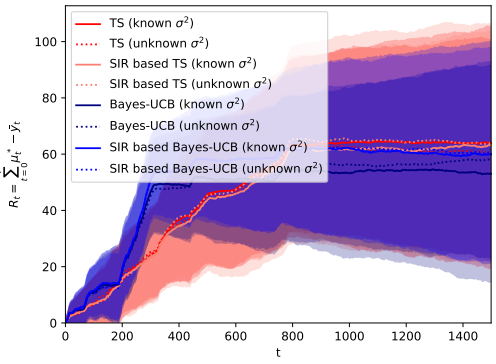
(c) Contextual logistic bandit with static context and known dynamics.

(d) Contextual logistic bandit with random context and known dynamics.

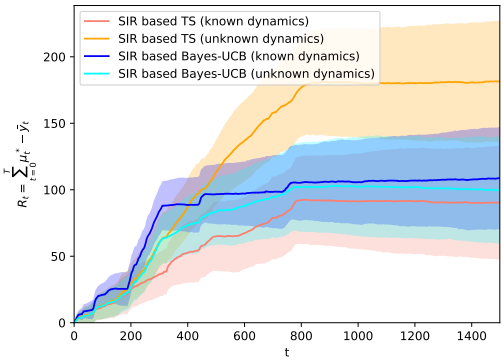
Figure 27: Mean regret (standard deviation shown as shaded region) in dynamic contextual linear-Gaussian bandits with dynamics as in Eqn. (54).



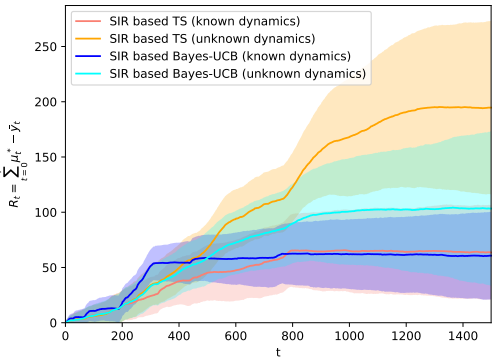
(a) Contextual logistic bandit with static context and known dynamics.



(b) Contextual logistic bandit with random context and known dynamics.

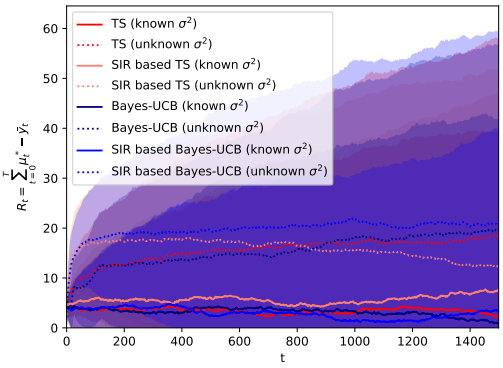


(c) Contextual logistic bandit with static context and known dynamics.

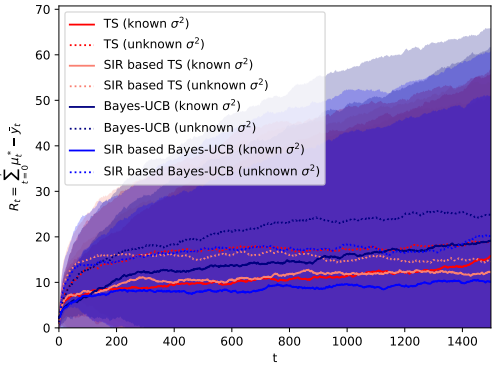


(d) Contextual logistic bandit with random context and known dynamics.

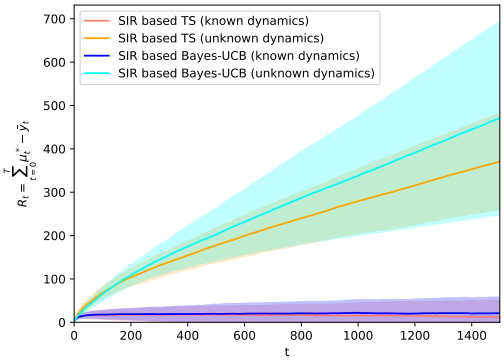
Figure 28: Mean regret (standard deviation shown as shaded region) in dynamic contextual linear-Gaussian bandits with dynamics as in Eqn. (55).



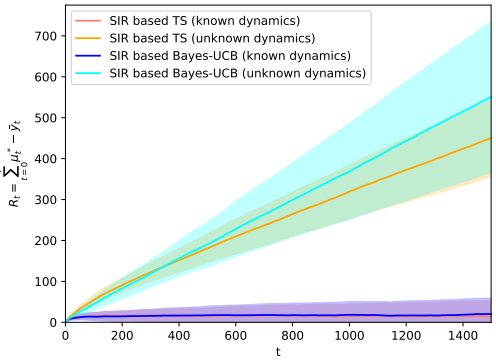
(a) Contextual logistic bandit with static context and known dynamics.



(b) Contextual logistic bandit with random context and known dynamics.

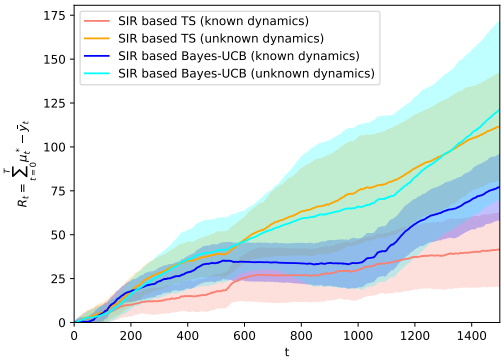


(c) Contextual logistic bandit with static context and known dynamics.

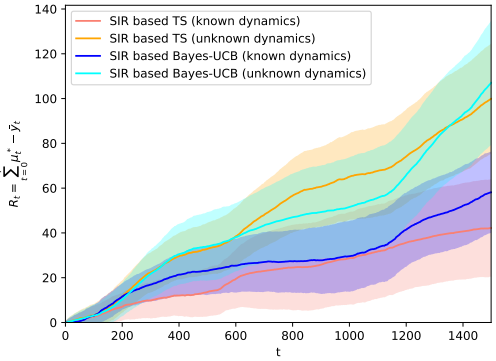


(d) Contextual logistic bandit with random context and known dynamics.

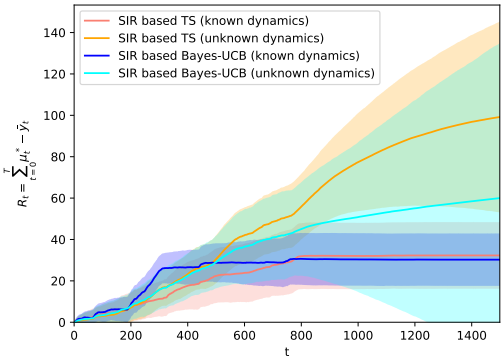
Figure 29: Mean regret (standard deviation shown as shaded region) in dynamic contextual linear-Gaussian bandits with dynamics as in Eqn. (56).



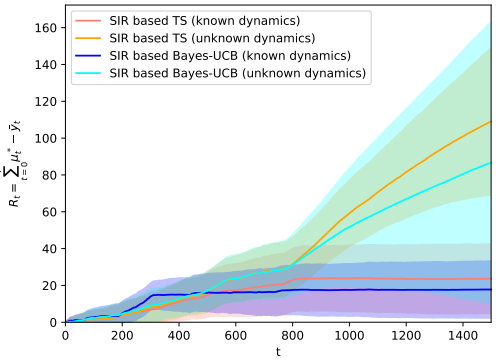
(a) Contextual logistic bandit with static context and dynamics as in Eqn. (54).



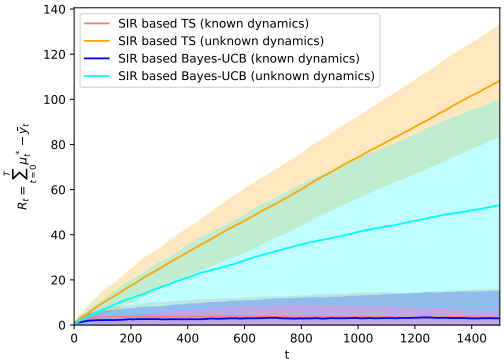
(b) Contextual logistic bandit with random context and dynamics as in Eqn. (54).



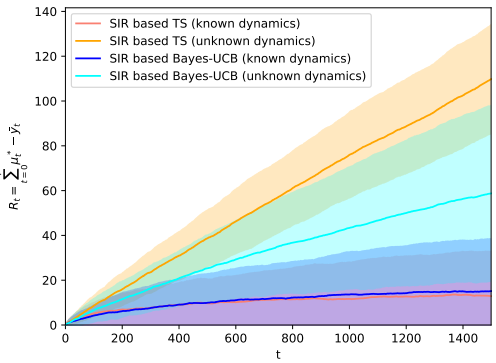
(c) Contextual logistic bandit with static context and dynamics as in Eqn. (55).



(d) Contextual logistic bandit with random context and dynamics as in Eqn. (55).



(e) Contextual logistic bandit with static context and dynamics as in Eqn. (56).



(f) Contextual logistic bandit with random context and dynamics as in Eqn. (56).

Figure 30: Mean regret (standard deviation shown as shaded region) in dynamic contextual logistic bandits.