Sequential Monte Carlo Bandits:

A flexible framework for complex and dynamic bandits

Iñigo Urteaga and Chris H. Wiggins

Applied Physics and Applied Mathematics
Data Science Institute

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

September 25, 2019

Multi-armed bandit

Practical challenges

- Reward generating process might change in practice

  Dynamic time-varying models

Multi-armed bandit

Practical challenges

- Reward generating process might change in practice
  Dynamic time-varying models
- Reward specific algorithms
  A flexible framework for complex models

Multi-armed bandit

Practical challenges

- Reward generating process might change in practice
  Dynamic time-varying models
- Reward specific algorithms
  A flexible framework for complex models
- Can't compute parameter posterior and/or their sufficient statistics
  Approximate inference

Multi-armed bandit

Practical challenges

- Reward generating process might change in practice
  Dynamic time-varying models
- Reward specific algorithms
  A flexible framework for complex models
- Can't compute parameter posterior and/or their sufficient statistics
  Approximate inference

Our proposed approach

Sequential Monte Carlo for Bayesian MAB algorithms

Multi-armed bandit

Problem formulation

$$\begin{cases} \theta_t^* \sim p(\theta_t^*|\theta_{t-1}^*) & \text{In-time transition density} \\ y_t \sim p_{a_t}(Y|x_t, \theta_t^*) & \text{Context-dependent parametric reward model} \end{cases}$$

Multi-armed bandit

Problem formulation

$$\begin{cases} \theta_t^* \sim p(\theta_t^*|\theta_{t-1}^*) & \text{In-time transition density} \\ y_t \sim p_{a_t}(Y|x_t, \theta_t^*) & \text{Context-dependent parametric reward model} \end{cases}$$

Optimal MAB policy

$$a_t^* = \underset{a' \in \mathcal{A}}{\operatorname{argmax}}\, \mu_{t,a'}(x_t, \theta^*), \text{ where } \mu_{t,a}(x_t, \theta^*) = \mathbb{E}\left\{Y|a, x_t, \theta^*\right\}$$

Multi-armed bandit

Problem formulation

$$\begin{cases} \theta_t^* \sim p(\theta_t^*|\theta_{t-1}^*) & \text{In-time transition density} \\ y_t \sim p_{a_t}(Y|x_t, \theta_t^*) & \text{Context-dependent parametric reward model} \end{cases}$$

Optimal MAB policy

$$a_t^* = \underset{a' \in \mathcal{A}}{\operatorname{argmax}}\, \mu_{t,a'}(x_t, \theta^*), \text{ where } \mu_{t,a}(x_t, \theta^*) = \mathbb{E}\left\{Y|a, x_t, \theta^*\right\}$$

Compute parameter posterior

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

as we observe history $\mathcal{H}_{1:t} = \{x_{1:t}, a_{1:t}, y_{1:t}\}$

$$x_{1:t} \equiv (x_1, \cdots, x_t),\ a_{1:t} \equiv (a_1, \cdots, a_t),\ y_{1:t} \equiv (y_{1,a_1}, \cdots, y_{t,a_t})$$

Bayesian MAB algorithms

Upper-confidence bounds

$$a_t = \underset{a' \in \mathcal{A}}{\operatorname{argmax}} \, q_{t,a'}(\alpha_t)$$

Quantile value of interest $q_{t,a}(\alpha_t)$, i.e.,

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

Computed by integrating out unknown parameters

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

Bayesian MAB algorithms

Thompson sampling

$$a_t \sim \mathbb{P}\left(a = a_t^* | x_t, \mathcal{H}_{1:t-1}\right)$$

Computed via

$$\mathbb{P}\left(a = a_t^* | x_t, \mathcal{H}_{1:t-1}\right) = \int \mathbb{1}\left[a = \underset{a' \in \mathcal{A}}{\operatorname{argmax}}\, \mu_{t,a'}(x_t, \theta_t)\right] p(\theta_t | \mathcal{H}_{1:t-1}) \mathrm{d}\theta_t$$

with (sampled) approximation

$$a_t = \underset{a' \in \mathcal{A}}{\operatorname{argmax}}\, \mu_{t,a'}\left(x_t, \theta_t^{(s)}\right), \text{ with } \theta_t^{(s)} \sim p(\theta_t | \mathcal{H}_{1:t-1})$$

Challenge in Bayesian MAB algorithms

No analytical solution

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

in complex and dynamic MAB models

Challenge in Bayesian MAB algorithms

No analytical solution

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

in complex and dynamic MAB models

Approximate solution

with sequential Monte Carlo (SMC) methods

Sequential Monte Carlo

(Sequential) Importance Sampling

1. A proposal distribution that factorizes over time

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

2. Recursive evaluation of the importance weights

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

3. Resample the random measure over time

$$\overline{\varphi}_t^{(m)} = \varphi_t^{(m')}$$

with m' drawn with replacement according to importance weights

$$w_t^{(m')} \sim \mathrm{Cat}\left(w_t^{(m)}\right)$$

Sequential Monte Carlo for latent MAB parameters

Sequentially updated parameter posterior approximation

Sequential Importance Resampling

$$p(\theta_{t,a}|\mathcal{H}_{1:t}) \approx p_M(\theta_{t,a}|\mathcal{H}_{1:t}) = \sum_{m_{t,a}=1}^{M} w_{t,a}^{(m_{t,a})}\delta\left(\theta_{a,t} - \theta_{a,t}^{(m_{t,a})}\right)$$

where

$$\theta_{t,a}^{(m_{t,a})} \sim p(\theta_{t,a}|\overline{\theta}_{t-1,a}^{(m_{t,a})}) \;\; \forall a \in \mathcal{A}$$

and

$$w_{t,a_t}^{(m_{t,a_t})} \propto p_{a_t}\left(y_t|x_t, \theta_{t,a_t}^{(m_{t,a_t})}\right)$$

Sequential Monte Carlo for latent MAB parameters

Sequentially updated parameter posterior approximation

Sequential Importance Resampling

$$p(\theta_{t,a}|\mathcal{H}_{1:t}) \approx p_M(\theta_{t,a}|\mathcal{H}_{1:t}) = \sum_{m_{t,a}=1}^{M} w_{t,a}^{(m_{t,a})} \delta\left(\theta_{a,t} - \theta_{a,t}^{(m_{t,a})}\right)$$

where

$$\theta_{t,a}^{(m_{t,a})} \sim p(\theta_{t,a}|\overline{\theta}_{t-1,a}^{(m_{t,a})}) \;\; \forall a \in \mathcal{A}$$

and

$$w_{t,a_t}^{(m_{t,a_t})} \propto p_{a_t}\left(y_t|x_t, \theta_{t,a_t}^{(m_{t,a_t})}\right)$$

Approximation

with convergence guarantees!

SMC-based framework

Use SMC posterior $p_M(\theta_{t,a}|\mathcal{H}_{1:t})$

To estimate sufficient statistics of the MAB policy

SMC-based framework

Use SMC posterior $p_M(\theta_{t,a}|\mathcal{H}_{1:t})$

To estimate sufficient statistics of the MAB policy

Thompson sampling

$$\theta_{t+1,a}^{(s)} \sim p\left(\theta_{t+1,a}|\theta_{t,a}^{(s)}\right), \text{ with } s \sim \mathrm{Cat}\left(w_{t,a}^{(m_{t,a})}\right)$$
$$a_{t+1} = \operatorname{argmax}_{a' \in \mathcal{A}} \mu_{t+1,a'}\left(x_{t+1}, \theta_{t+1,a'}^{(s)}\right)$$

SMC-based framework

Use SMC posterior $p_M(\theta_{t,a}|\mathcal{H}_{1:t})$

To estimate sufficient statistics of the MAB policy

Thompson sampling

$$\theta_{t+1,a}^{(s)} \sim p\left(\theta_{t+1,a}|\theta_{t,a}^{(s)}\right), \text{ with } s \sim \text{Cat}\left(w_{t,a}^{(m_{t,a})}\right)$$

$$a_{t+1} = \operatorname{argmax}_{a' \in \mathcal{A}} \mu_{t+1,a'}\left(x_{t+1}, \theta_{t+1,a'}^{(s)}\right)$$

Bayes-UCB

$$\theta_{t+1,a}^{(m'_a)} \sim p\left(\theta_{t+1,a}|\theta_{t,a}^{(m'_a)}\right), \text{ with } m'_a \sim \text{Cat}\left(w_{t,a}^{(m_{t,a})}\right)$$

Compute $q_{t+1,a}(\alpha_{t+1}) := \max\{\mu \mid \sum_{m|\mu_{t+1,a}^m > \mu} w_{t,a}^m \geq \alpha_{t+1}\}$

$$a_{t+1} = \operatorname{argmax}_{a' \in \mathcal{A}} q_{t+1,a'}(\alpha_{t+1})$$

SMC-based framework for dynamic models

General linear dynamics

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

results in transition densities

$$\theta_{t,a} \sim \begin{cases} \mathcal{N}\left(\theta_{t,a} | L_a \theta_{t-1,a}, \Sigma_a\right) & \text{with known parameters} \\ \mathcal{T}\left(\theta_{t,a} | \nu_{t,a}, m_{t,a}, R_{t,a}\right) & \text{with unknown parameters} \end{cases}$$

SMC-based framework for complex models

Complex reward models

Likelihood function known up to proportionality constant

$$w_{t,a}^{(m_{t,a})} \propto p_a(Y|x,\theta)$$

SMC-based framework for complex models

Complex reward models

Likelihood function known up to proportionality constant

$$w_{t,a}^{(m_{t,a})} \propto p_a(Y|x,\theta)$$

Contextual Gaussian

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

SMC-based framework for complex models

Complex reward models

Likelihood function known up to proportionality constant

$$w_{t,a}^{(m_{t,a})} \propto p_a(Y|x,\theta)$$

Contextual Gaussian

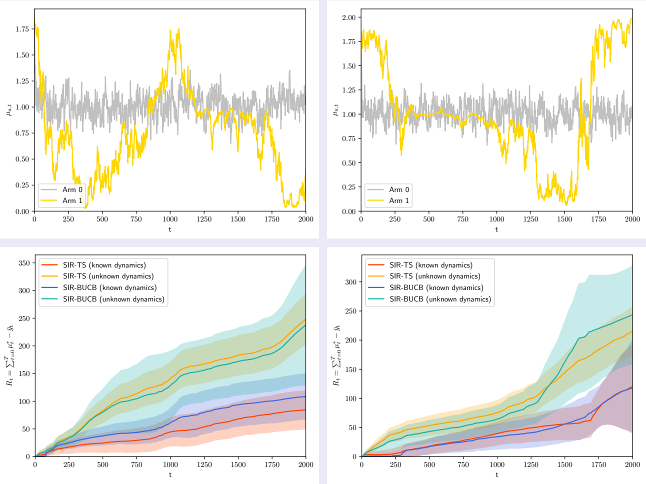
$$p_a(Y|x,\theta) = \mathcal{N}\left(Y|x^\top w_a, \sigma_a^2\right) = \frac{e^{-\frac{(y-x^\top w_a)^2}{2\sigma_a^2}}}{\sqrt{2\pi\sigma_a^2}}$$

Categorical-softmax rewards

$$p_a(Y=c|x,\theta_a) = \frac{e^{(x^\top \theta_{a,c})}}{\sum_{c'=1}^{C} e^{(x^\top \theta_{a,c'})}}$$

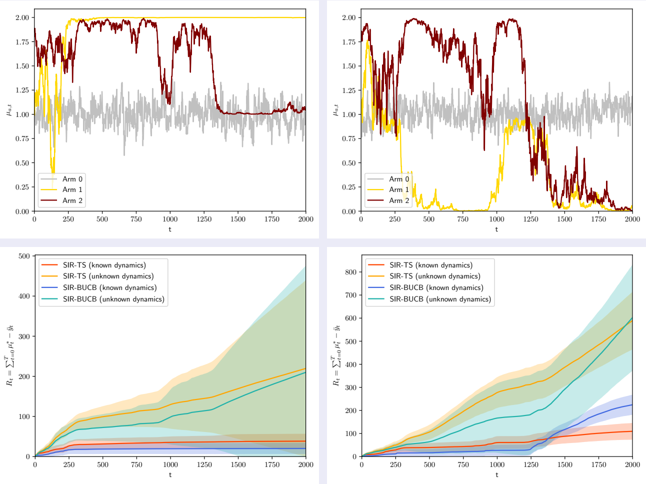
SMC-based framework in simulated MABs

Two-armed contextual 3-categorical bandit



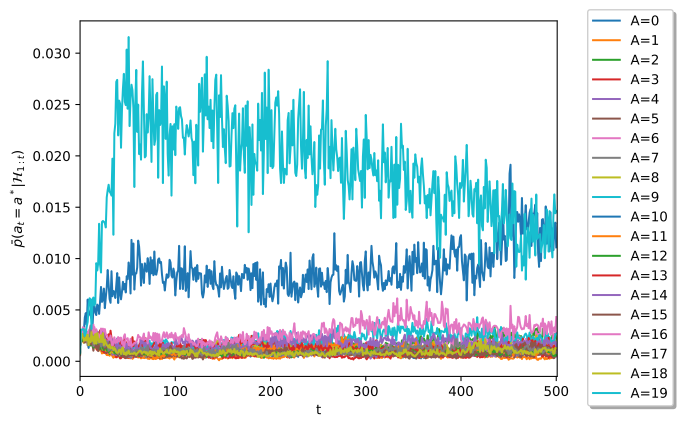
SMC-based framework in simulated MABs

Three-armed contextual 3-categorical bandit



SMC-based framework in real MABs

Yahoo News Recommendation data



Contribution

SMC-based MAB method

- Approximates parameter posteriors with random measures
- Reward function known only up to a proportionality constant
- Time-varying parameter models that we can sample from

Contribution

SMC-based MAB method

- Approximates parameter posteriors with random measures
- Reward function known only up to a proportionality constant
- Time-varying parameter models that we can sample from

A flexible MAB framework

For solving a rich class of MAB problems,
such as dynamic and nonlinear bandits

Open questions

Regret bounds

SMC posterior convergence, but...

Open questions

Regret bounds

SMC posterior convergence, but...

Dynamics of the MAB problem

Optimal arm changes

Open questions

Regret bounds

SMC posterior convergence, but...

Dynamics of the MAB problem

Optimal arm changes

Dimensionality of the MAB problem

Dependency on number of arms

Thanks

Questions?