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
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**

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

Problem formulation

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

Optimal MAB policy

\[
a_t^* = \arg\max_{a' \in \mathcal{A}} \mu_{t,a'}(x_t, \theta^*), \quad \text{where } \mu_{t,a}(x_t, \theta^*) = \mathbb{E} \{ Y|a, x_t, \theta^* \}
\]
Multi-armed bandit

Problem formulation

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

Optimal MAB policy

\[
a_t^* = \arg\max_{a' \in A} \mu_{t,a'}(x_t, \theta^*), \quad \text{where} \quad \mu_{t,a}(x_t, \theta^*) = \mathbb{E}\{Y|a, x_t, \theta^*\}
\]

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 \( \mathcal{H}_{1:t} = \{x_{1:t}, a_{1:t}, y_{1:t}\} \)

\[
x_{1:t} \equiv (x_1, \ldots, x_t), \quad a_{1:t} \equiv (a_1, \ldots, a_t), \quad y_{1:t} \equiv (y_1, a_1, \ldots, y_t, a_t)
\]
Bayesian MAB algorithms

**Upper-confidence bounds**

\[ a_t = \arg\max_{a' \in A} q_{t,a'}(\alpha_t) \]

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

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

Computed by integrating out unknown parameters

\[ p(\mu_{t,a}) = \int p(\mu_{t,a}|x_t, \theta_t)p(\theta_t|H_{1:t-1})d\theta_t \]
Bayesian MAB algorithms

Thompson sampling

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

Computed via

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

with (sampled) approximation

\[ a_t = \arg \max_{a' \in \mathcal{A}} \mu_{t,a'}(x_t, \theta_t^{(s)}) \text{, with } \theta_t^{(s)} \sim p(\theta_t | \mathcal{H}_{1:t-1}) \]
Challenge in Bayesian MAB algorithms

No analytical solution

\[ p(\theta_t|\mathcal{H}_{1:t}) \propto p_a(y_t|x_t, \theta_t)p(\theta_t|\theta_{t-1})p(\theta_{t-1}|\mathcal{H}_{1:t-1}) \]

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

\[
w^{(m)}_t \propto \frac{p(\varphi_t|\varphi_{1:t-1})}{\pi(\varphi_t|\varphi_{1:t-1})} w^{(m)}_{t-1}
\]

3. Resample the random measure over time

\[
\overline{\varphi}^{(m)}_t = \varphi^{(m')}_t
\]

with \(m'\) drawn with replacement according to importance weights

\[
w^{(m')}_t \sim \text{Cat} \left( w^{(m)}_t \right)
\]
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

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

and

\[ w_{t,a}^{(m_{t,a})} \propto p_{a_t} \left( y_{t} | x_{t}, \theta^{(m_{t,a})}_{t,a} \right) \]
Sequential Monte Carlo for latent MAB parameters

Sequentially updated parameter posterior approximation

**Sequential Importance Resampling**

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

where

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

and

\[
w^{(m_{t,a})}_{t,a} \propto p_{a_{t}} \left( y_{t}|x_{t}, \theta^{(m_{t,a})}_{t,a} \right)
\]

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

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

\[
a_{t+1} = \argmax_{a' \in A} \mu_{t+1,a'} \left( x_{t+1}, \theta_{t+1,a'}^{(s)} \right)
\]
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

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

$$a_{t+1} = \arg\max_{a' \in A} \mu_{t+1,a'} \left(x_{t+1}, \theta_{t+1,a'}^{(s)}\right)$$

Bayes-UCB

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

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

$$a_{t+1} = \arg\max_{a' \in A} q_{t+1,a'}(\alpha_{t+1})$$
SMC-based framework for dynamic models

General linear dynamics

\[
\theta_{t,a} = L_a \theta_{t-1,a} + \epsilon_a , \quad \epsilon_a \sim \mathcal{N}(\epsilon_a|0, \Sigma_a) ,
\]

results in transition densities

\[
\theta_{t,a} \sim \begin{cases} 
\mathcal{N}(\theta_{t,a}|L_a \theta_{t-1,a}, \Sigma_a) & \text{with known parameters} \\
\mathcal{T}(\theta_{t,a}|\nu_{t,a}, m_{t,a}, R_{t,a}) & \text{with unknown parameters}
\end{cases}
\]
SMC-based framework for complex 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

\[ p_a(Y|x, \theta) = \mathcal{N} \left( Y | x^\top w_a, \sigma_a^2 \right) = e^{-\frac{(y-x^\top w_a)^2}{2\sigma_a^2}} \frac{1}{\sqrt{2\pi\sigma_a^2}} \]
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

\[
R_t = \sum_{t=0}^{T} \mu^* - \bar{y}_t
\]

- SIR-TS (known dynamics)
- SIR-TS (unknown dynamics)
- SIR-BUCB (known dynamics)
- SIR-BUCB (unknown dynamics)
SMC-based framework in real MABs

Yahoo News Recommendation data

\[ \hat{p}(a_t = a^* | H_{1:t}) \]

- A=0
- A=1
- A=2
- A=3
- A=4
- A=5
- A=6
- A=7
- A=8
- A=9
- A=10
- A=11
- A=12
- A=13
- A=14
- A=15
- A=16
- A=17
- A=18
- A=19
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
Conclusion and next steps

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?