Project: Efficient Thompson Sampling for Contextual Logistic Bandits

Tuan Nguyen

Motivation

- Bandits
  - Good general framework to study learner's decision making strategies under uncertainty
  - A branch of Reinforcement Learning (RL)
    - In Bandits, context in each round is independent from previous rounds
  - Practical applications: recommendation systems
- Thompson Sampling for Bandits
  - Related to Bayesian inference and posterior sampling

Contextual Bandits Setting

- A game consists of T rounds, T is really large or infinity
- At each round t, the system randomly assigns K items, each item has observable context/features $x_{t,k} \in \mathbb{R}^d$
  - x_k is an observable random variable
- Learner chooses an item k and observes a stochastic reward

$$y_t \sim p(y|x_{t,k}, \theta^*)$$

  - $p(.)$ is the underlying probabilistic model with a **fixed** parameter θ^*
  - y_t is a observable random variable
- Learner's objective: maximizing the final cumulative rewards or minimize the cumulative regrets

$$\max \left[\mathrm{Reward}_T = \sum_t^T y_t \right]$$

$$\min \left[\mathrm{Regret}_T = \sum_t^T E[y_t|x_{t,k^*}, \theta^*] - E[y_t|x_{t,\hat{k}}, \theta^*] \right]$$

An Approach to Solve Contextual Bandits

- While playing, learner collects all previous observations

$$\mathcal{D}_{t-1} = \{a_1, x_1, y_1, \ldots, a_{t-1}, x_{t-1}, y_{t-1}\}$$

- Based on $\mathcal{D}_{t-1}$, learner tries to approximate $\tilde{\theta}_t \approx \theta^*$
- Then, based on $\hat{\theta}_t$, learner select the potentially optimal item

$$\hat{k} = \operatorname*{argmin}_k E[y | x_{t,k}, \tilde{\theta}_t]$$

- Intuition: As learner collects more observations, it maybe makes better estimations of $\tilde{\theta}_t$, selects better item, and gets better rewards

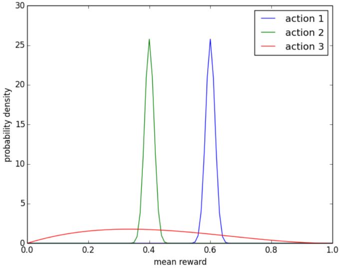
Exploration versus Exploitation $\hat{\theta}_t$

- MLE of $\hat{\theta}_t$

$$\hat{\theta}_t = \operatorname*{argmax}_{\theta} p(\mathcal{D}_{t-1}|\theta)$$

- However, using MLE is greedy in exploitation, and potentially stuck with suboptimal items
  - Learner would exploit a few items that have features fit $\hat{\theta}_t$
  - In turn, $\hat{\theta}_t$ is updated to the rewards of these few items
  - Repeat this process forever

Exploration versus Exploitation $\hat{\theta}_t$



- Simplified case for intuition:
  - Item or action k maps to θ_k, its features is simplified to one hot
  - Reward distribution for item k is simplified to posterior distribution $p(\theta_k|\mathcal{D}_{t-1})$
  - Exploiting the item that have maximal MLE $\hat{\theta}_k$ is not good
    - ★ At this round, observed reward only minimally affects posterior of $\hat{\theta}_k$
    - ★ Thus, learner chooses the item forever and ignores other potentially better items

source: Russo et al., 2020

ϵ-Greedy Strategy

- Strategy
  - With probability $1 - \epsilon$, exploit optimal item according to MLE $\hat{\theta}$. Otherwise, select other items uniformly
- Note: learner will explore forever even after certain about optimal item.

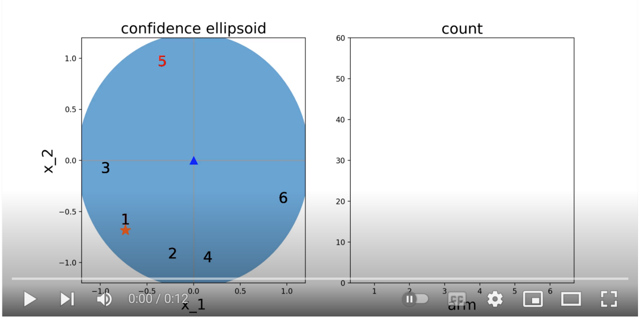
Upper Confidence Bound (UCB) Strategy

- Strategy: optimistism in the face of uncertainty
  - Learner selects the potentially optimal item

$$\hat{k} = \underset{k}{\operatorname{argmin}} \max_{\theta} E[y|x_{t,k}, \theta] \quad \text{s.t. } \|\theta - \hat{\theta}_t\|_{M_t} \leq h(t)$$

  - The region $\|\theta - \hat{\theta}_t\|_{M_t} \leq h(t)$ is a confidence ellipsoid around the estimated $\hat{\theta}$
  - Learner optimistically select the combination of the item and θ on the confidence bound that has maximal reward expectation
  - As round passes, the ellipsoid will shrink smaller, learner explores lesser and eventually stops
- Notes:
  - Theoretically proven to have sublinear regret bounds and strong empirical performance (Auer, 2003; Filippi et al., 2010; Faury et al., 2020; Jun et al., 2021; Faury et al., 2022)
    - ★ Regret bound $\tilde{O}(d\sqrt{T} + \kappa)$
  - Largest number of research in bandits

Upper Confidence Bound (UCB) Strategy



source: from Prof. Kwang-Sung Jun

Thompson Sampling

- Strategy: Bayesian inference and posterior sampling
  - Learner approximates posterior $p(\theta|\mathcal{D}_{t-1})$ and samples $\tilde{\theta}_t$ from it

$$p(\theta|\mathcal{D}_{t-1}) \propto p(\theta)p(\mathcal{D}_{t-1}|\theta)$$

  - As round passes, the variance of the posterior will shrink, learner explores lesser and eventually stops
- Notes
  - Empirically perform better than UCB (Chapelle and Li, 2011; Li et al., 2010; Dumitrascu et al., 2018)
  - Theoretically proven to have similar regret bounds as UCB (Agrawal and Goyal, 2013; Russo and Van Roy, 2016; Abeille and Lazaric, 2017; Dong et al., 2019)

Probabilistic Models in Contextual Bandits

- Mixture of Gaussians (Urteaga and Wiggins, 2021)

$$p(y|\boldsymbol{x}, \boldsymbol{w}_i, \sigma_i, \pi_i) = \sum_i \pi_i \mathcal{N}(y|\boldsymbol{x}^\top \boldsymbol{w}_i, \sigma_i^2)$$

- Generalized linear models (GLM) (Filippi et al., 2010)

$$p(y|\boldsymbol{x}, \boldsymbol{w}, \sigma) = \exp\left[\frac{y\psi(\boldsymbol{x}^\top \boldsymbol{w}) - A(\psi(\boldsymbol{x}^\top \boldsymbol{w}))}{\sigma^2} + c(y, \sigma^2)\right]$$

where $\psi(.)$ is a link function

  - Linear regression, Gaussian reward, identity link function (Agrawal and Goyal, 2013; Abeille and Lazaric, 2017)

$$p(y|\boldsymbol{x}, \boldsymbol{w}, \sigma) = \mathcal{N}(y|\boldsymbol{x}^\top \boldsymbol{w}, \sigma)$$

Contextual Logistic Bandits

- Logistic regression, Bernoulli reward, sigmoid link function (Dong et al., 2019)

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{w}) = \mathrm{Ber}(y_i|sigm(\boldsymbol{x}_i^\top \boldsymbol{w}))$$
$$p(\mathcal{D}_{t-1}|w) = \prod_i^{t-1} \frac{(e^{\boldsymbol{x}_i^\top \boldsymbol{w}})^{y_i}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{w}}}$$

- More challenging compared to linear regression
  - Nonlinearity, discrete rewards
  - No closed-form MLE, need to use numerical optimization methods
  - Challenging to approximate posterior and to do posterior sampling
    - Laplace approximation (Chapelle and Li, 2011)
    - Polya-Gamma Gibbs sampling (Dumitrascu et al., 2018; Polson et al., 2013)

Recent Work: Polya-Gamma Thompson Sampling (PG-TS) (Dumitrascu et al., 2018)

- Intuition
  - Reframe the discrete rewards as functions of latent variables with PG distributions over a continous space
  - With the PG latent variable, the logistic likelihood becomes mixture of Gaussians with PG mixing distributions

Recent Work: Polya-Gamma Thompson Sampling (PG-TS)

- PG augmentation scheme

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^\infty e^{-\omega\psi^2/2} p(\omega) d\omega$$

where $\kappa = a - b/2$, ω $\mathrm{PG}(b, 0)$

- The logistic likelihood becomes mixture of Gaussians with PG mixing distributions.

$$L_i(\boldsymbol{w}|\omega_i, x_i, y_i) = \frac{(e^{\boldsymbol{x}_i^\top \boldsymbol{w}})^{y_i}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{w}}} \propto e^{\kappa_i \boldsymbol{x}_i^\top \boldsymbol{w}} \int_0^\infty e^{-\omega_i (\boldsymbol{x}_i^\top \boldsymbol{w})^2/2} p(\omega_i) d\omega_i$$

$$p(\boldsymbol{w}|\omega_i, \mathcal{D}_{t-1}) = p(\boldsymbol{w}) \prod_i^{t-1} L_i(\boldsymbol{w}|\omega_i, x_i, y_i)$$

Recent Work: Polya-Gamma Thompson Sampling (PG-TS)

- Thus, $\boldsymbol{w}$ can be draw from a Gaussian distribution, parameterized by PG augmentation ω_i

$$(\omega_i|\boldsymbol{w}) \sim \mathrm{PG}(1, \boldsymbol{x}_i^\top \boldsymbol{w})$$
$$(\boldsymbol{w}|\omega_i, \mathcal{D}_{t-1}) \sim N(m_\omega, V_\omega)$$

where $V_\omega = (X^\top \Omega X + V_0^{-1})^{-1}$, $m_\omega = V_\omega(X^\top \kappa + V_0^{-1} m_0)$

- Benefits
  - PG distribution can be easily sampled with high acceptance rate

Recent Work: Polya-Gamma Thompson Sampling (PG-TS)

Algorithm 3 PG-TS: Pólya-Gamma augmented Thompson Sampling

Input: $\mathbf{b}$, $\mathbf{B}$, M, $\mathcal{D} = \emptyset$, $\boldsymbol{\theta}_0 \sim MVN(\mathbf{b}, \mathbf{B})$
for $t = 1, 2, \ldots$ **do**
  Receive contexts $\mathbf{x}_{t,a} \in \mathbb{R}^d$
  $\boldsymbol{\theta}_t^{(0)} \leftarrow \boldsymbol{\theta}_{t-1}$
  for $m = 1$ **to** M **do**
    for $i = 1$ **to** $t-1$ **do**
      $\omega_i | \boldsymbol{\theta}_t^{(m-1)} \sim PG(1, \mathbf{x}_{i,a_i}^\top \boldsymbol{\theta}_t^{(m-1)})$
    $\boldsymbol{\Omega}_{t-1} = diag(\omega_1, \omega_2, \ldots \omega_{t-1})$
    $\boldsymbol{\kappa}_{t-1} = \left[r_1 - \frac{1}{2}, ..., r_{t-1} - \frac{1}{2}\right]^\top$
    $\mathbf{V}_\omega \leftarrow (\mathbf{X}_{t-1}^\top \boldsymbol{\Omega}_{t-1} \mathbf{X}_{t-1} + \mathbf{B}^{-1})^{-1}$
    $\mathbf{m}_\omega \leftarrow \mathbf{V}_\omega (\mathbf{X}^\top \boldsymbol{\kappa}_{t-1} + \mathbf{B}^{-1}\mathbf{b})$
    $\boldsymbol{\theta}_t^{(m)} | \boldsymbol{r}_{t-1}, \boldsymbol{\omega} \sim MVN(\mathbf{m}_\omega, \mathbf{V}_\omega)$
  $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t^{(M)}$
  Select arm $a_t \leftarrow argmax_a \mu(\mathbf{x}_{t,a}^\top \boldsymbol{\theta}_t)$
  Observe reward $r_t \in \{0, 1\}$
  $\mathcal{D} = \mathcal{D} \cup \{\mathbf{x}_{t,a_t}, a_t, r_t\}$

Extending the Polya-Gamma Thompson Sampling

- Extending the posterior sampling
  - Applying Hamiltonian MC
  - Applying Stein's Variational Inference
- Extending PG-TS from Bernoulli to Categorical rewards

Coding and Evaluation

- Coding
  - https://github.com/iosband/ts_tutorial (Russo et al., 2020)
  - https://github.com/iurteaga/bandits (Urteaga and Wiggins, 2021)
- Evaluation
  - Measure the quality of samples generated from posterior samplers?
  - Empirical performance on reward and regret over time horizon
  - Theoretical analysis on regret bounds (not presented in the PG-TS paper)**