Overview of Thompson sampling in RL and BCI applications

Yang Li, Yuliang Xu

On Thompson Sampling

Introducing (Refreshing) on Thompson Sampling

- Bernoulli Multi-armed Bandit: K actions, Action $k \in \{1, \ldots, K\}$ produces a reward that follows Bernoulli(θ_k). $(\theta_1, \ldots, \theta_K)$ are unknown, but are fixed over time.
- Thompson Sampling for Bernoulli Multi-armed Bandit:
  - **Initialize**: Assume for each $k = 1, \ldots, K$, the prior distribution of θ_k is Beta(α_k, β_k).
  - **Sample**: At each time point, sample $\hat{\theta}_k$ from Beta(α_k, β_k), as an estimate of θ_k, $k = 1, \ldots, K$.
  - **Choose action**: Apply action $a_t = \text{argmax}_k \hat{\theta}_k$, and get a reward of r_t.
  - **Update**: Update the distribution of θ_k with the posterior distribution, i.e.

$$(\alpha_k, \beta_k) \leftarrow (\alpha_k, \beta_k) + I(k = a_t) \cdot (r_t, 1 - r_t)$$

  - **Repeat**: Repeat the procedure.

Introducing (Refreshing) on Thompson Sampling

- General idea of Thompson Sampling:
  - **Initialize**: Assume the prior distribution p for some parameters.
  - **Update**: Compute the posterior distribution of the parameters using some statistics. Update p with the posterior distribution.
  - **Sample**: Sample the parameters from the distribution p.
  - **Choose action**: Compute the optimal policy from the model with sampled parameters.
  - **Repeat**: Repeat the procedure.

Regret Analysis

- Regret Bound for Thompson Sampling for Bernoulli Multi-armed Bandit. (Agrawal and Goyal, 2012) [1]
- Without loss of generality,assume $\theta_1 = max_i \theta_i$. Let $\Delta_i = \theta_1 - \theta_i$, the Thompson Sampling for N-armed Bernoulli Bandit has expected regret

$$E(R(T)) \leq O\left(\left(\sum_{a=2}^{N} \frac{1}{\Delta_a^2}\right)^2 \ln(T)\right)$$

in time T.

Regret Analysis

- Regret Bound for Thompson Sampling for Bernoulli Multi-armed Bandit. (Kaufmann et al., 2012) [4]
- Without loss of generality,assume $\theta_1 = max_i \theta_i$. Let $\Delta_i = \theta_1 - \theta_i$, For every $\epsilon > 0$, there exists a problem-dependent constant $C(\epsilon, \theta_1, \dots, \theta_N)$, such that

$$E(R(T)) \leq (1+\epsilon) \sum_{a=2}^{N} \frac{\Delta_a(\ln(T) + \ln\ln(T))}{K(\theta_a, \theta_1)} + C(\epsilon, \theta_1, \dots, \theta_N),$$

where $K(p,q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$.

Regret Analysis

- Problem Independent Regret Bound for Thompson Sampling for Bernoulli Stochastic Bandit. (Agrawal and Goyal, 2013) [2]
- The Thompson Sampling for N-armed Bernoulli Bandit has expected regret

$$E(R(T)) \leq O\left(\sqrt{NT\ln(T)}\right)$$

in time T.

Regret Analysis

- Regret Bound for Thompson Sampling for 1-Dimensional Exponential Family Bandit. (Korda et al., 2013) [5]
- 1-Dimensional Exponential Family Bandit: The outcome follows an one-dimensional exponential family $p(x|\theta) = A(x)\exp(T(x)\theta - F(\theta))$.
- Jeffreys prior: $\pi_J(\theta) \propto \sqrt{|F''(\theta)|}$
- The Thompson Sampling for 1-Dimensional Exponential Family Bandit with Jeffreys prior follows

$$\lim_{T\to\infty} \frac{E(R(T))}{\ln(T)} = \sum_{a=1}^{K} \frac{\mu(\theta_{a^*}) - \mu(\theta_a)}{K(\theta_a, \theta_{a^*})},$$

where $K(\theta, \theta^{'})^{=KL(p_\theta, p_{\theta'})}$ is the Kullback-Leibler divergence.

Regret Analysis

- Regret Bound for Thompson Sampling for a more general setting (Russo and Van Roy, 2014) [8]
- Model Setting:
  - Set of Actions $\mathcal{A}$.
  - At time point t, the agent can only select the action from a subset of the action set possibly random $\mathcal{A}_t \subset \mathcal{A}$.
  - After getting the action set, the agent select an action $A_t \in \mathcal{A}_t$, based on the history $H_t := (\mathcal{A}_1, A_1, R_1, \dots, \mathcal{A}_{t-1}, A_{t-1}, R_{t-1}, \mathcal{A}_t)$, and distribution $\pi_t(H_t)$.
  - After selecting the action A_t, the agent get a reward R_t, and $E(R_t|H_t, \theta, A_t) = f_\theta(A_t)$

Regret Analysis

- Regret Bound for Thompson Sampling for a more general setting (Russo and Van Roy, 2014)
- An example regarding the random action set.
  The contextual MAB model:
  - An exogenous Markov process X_t taking values in a set $\mathcal{X}$ influences rewards.
  - The expected reward at time t is given by $f_\theta(a, X_t)$.
  - We can define $\mathcal{A}' := \{(x, a) : x \in \mathcal{A}, a \in \mathcal{A}(x)\}$, and $\mathcal{A}'_t = \{(X_t, a) : a \in \mathcal{A}(X_t)\}$.

Regret Analysis

- Regret Bound for Thompson Sampling for a more general setting (Russo and Van Roy, 2014)
- The Regret is defined by

$$R(T, \pi, \theta) = \sum_{t=1}^{T} \mathrm{E}(\max_{a \in \mathcal{A}_t} f_\theta(a) - f_\theta(A_t) \mid \theta).$$

- The Bayesian Regret is defined by

$$BR(T, \pi) = \mathrm{E}_\theta(R(T, \pi, \theta))$$

with respect to the prior distribution over θ

Regret Analysis

- Regret Bound for Thompson Sampling for a more general setting (Russo and Van Roy, 2014)
- Bandit with finite actions:
- **Theorem 1.** Let π^{TS} be the policy generated from Thompson Sampling. If $\mathcal{A} = K < \infty$, and $R_t \in [0,1]$, we have

$$BR(T, \pi^{\mathrm{TS}}) \leq 2\min\{K, T\} + 4\sqrt{KT(2 + 6\log(T))} = O(\sqrt{|\mathcal{A}|T\log(T)})$$

Regret Analysis

- Regret Bound for Thompson Sampling for a more general setting (Russo and Van Roy, 2014)
- Linear Bandit: Reward function are parameterized by a vector $\theta \in \Theta \subset \mathbb{R}^d$, and there is a known feature mapping $\phi : \mathcal{A} \to \mathbb{R}^d$, such that $f_\theta(a) = \phi(a)^T \theta$
- **Theorem 2.** If Θ and $\phi(a)$ are bounded, $R_t - f_\theta(A_t)$ conditioned on (H_t, A_t, θ) is sub-Gaussian, then

$$BR(T, \pi^{\mathrm{TS}}) = O(d\sqrt{T}\log(T)).$$

Regret Analysis

- Regret Bound for Thompson Sampling for a more general setting (Russo and Van Roy, 2014)
- **Theorem 3.** If $\mathcal{A}$ is finite, $(f_\theta(a) : a \in \mathcal{A})$ follows a multivariate Gaussian distribution with marginal variances bounded by 1, $R_t - f_\theta(A_t)$ is independent of (H_t, θ, A_t), and $\{R_t - f_\theta(A_t) | t \in \mathbb{N}\}$ is an iid sequence of zero mean Gaussian random variables with variance σ^2, then

$$BR(T, \pi^{\mathrm{PS}}) \leq 1 + 2\sqrt{T\gamma_T \ln(1+\sigma^{-2})^{-1} \ln\left(\frac{(T^2+1)|\mathcal{A}|}{\sqrt{2\pi}}\right)},$$

where γ_T is the maximum possible information gain, defined as the difference between the entropy of prior and posterior.

On Brain-Computer-Interface

Introducing EEG-BCI speller system.

An Electroencephalogram Brain-computer Interface (EEG-BCI) Spelling System is a device that enables people to 'type' in words without using the physical keyboard. x‘

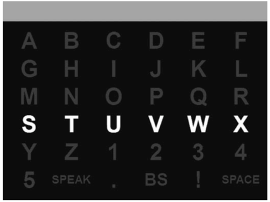


Figure: The panel of a EEG-BCI speller system

Current methods in BCI-speller system

- **Classical design** Random ordering and exhaustive row and column flashes: flashing every row and every column each loop, and repeating for 4-10 loops to decide a letter[9]. That means we need 48-120 flashes to decide the letter!
- **Current methods**
  - Most previous literature focusing on the classification of P300 signal with target letter, instead of an online learning setting. Barely any method used TS.

...continued

- **Current methods**
  - POMPD [7]: Independently select row and column, use flashes as both state and action, applicable for selecting one letter, but not efficient for typing word. Reward is almost greedy.
  - Hierarchy model of variable-sized flash groups based on a language model [6]: Easily susceptible to error propagation, not friendly to users with weaker cognitive ability.
  - Adaptive optimization [3]: Greedy approach for stimulus selection, letter-by-letter approach.

Our proposed method

Goal Determine the target letter (word) with minimum number of flashes.

1. Action space $\mathscr{A}$: a set of row or column flashes a_i.
   $|\mathscr{A}| = C(n,r)$, $n =$ total number of flashes, $r =$ number of flashes chosen at each iteration.
2. State $\mathscr{X}$: [Bandit problem] Only one state. [MDP] Each letter is viewed as one state, π is determined by the linguistic probability.
3. Reward $r_{ij} = R(x_j, a_i)$: a summary statistic of the P300 time series signal when the set of flashes a_i is shown to the human, given the true letter is x_j. (Example).

Illustrating P300 signal and motivation for modifying TS

Constraint Psychological Refractory Period (PRP) effect: EEG signal cannot discern between two consecutive target events.

- Suppose our target letter is T, i.e. $a_i = (4, 8)$ contains the target letter, if we number rows from $1, ..., 6$, and number columns from $7, ..., 12$.
- P300 signal for flash a_i is recorded from the moment a_i start to flash, till 300 ms after.
- In the following example, P300 spike falls into the time interval for flashes $(7, 4, 8, 9)$, although only $(4, 8)$ contains the true target letter T.

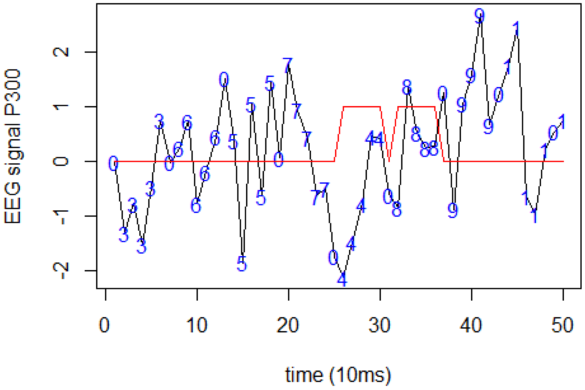


Figure: Illustrating P300 signal

Modified Thompson Sampling Algorithm

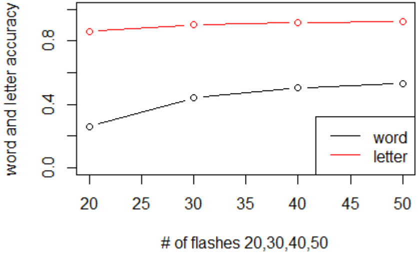
Motivation for modifying TS
Instead of selecting $a_i = \text{argmax}_{x_i} \mathbb{E}[r_i(\theta_i; \cdot)|X = x_i, A = a_i]$, we select a group of actions/flashes at a time. $G^{(t)} = \text{argmax}_{G \subset \mathcal{A}} \frac{1}{|G|} \sum_{i \in G} |\theta_i|$.
In the simple Bernoulli setting, define $\theta_k = \text{Prob}\left(a_k \text{contains the target letter}\right)$.
Advantage

- Selecting a group avoids the overlapping PRP effect, since a group of actions will be assigned the same reward.

...Continued

- Efficiently identify the target letter by searching for the row and column that contains the target letter, requiring less flashes.
- TS effectively explores the action space, which could be more robust under random reward. Friendly to users with weak cognitive ability.

Initial simulation study



- The initial simulation of the simple Bernoulli setting, with random reward function with normal noise.
- Only independent bandit TS is considered.
- The target word is `THOMPSON`. Each letter is flashed the same number of times.
- Already better than the current design that takes 48-120 flashes to identify a letter.

Improvement

1. Consider a MDP problem with the transition matrix determined by a linguistic model.
2. Design stopping rules to assign number of flashes adaptively across letters.
3. Try different prior settings.

Shipra Agrawal and Navin Goyal.
Analysis of thompson sampling for the multi-armed bandit problem.
In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Shipra Agrawal and Navin Goyal.
Further optimal regret bounds for thompson sampling.
In *Artificial intelligence and statistics*, pages 99–107. PMLR, 2013.

Dmitry Kalika, L. Collins, C. Throckmorton, and B. Mainsah.
Adaptive stimulus selection in erp-based brain-computer interfaces by maximizing expected discrimination gain.
2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 1405–1410, 2017.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos.
Thompson sampling: An asymptotically optimal finite-time analysis.
In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos.
Thompson sampling for 1-dimensional exponential family bandits.
arXiv preprint arXiv:1307.3400, 2013.

Rui Ma, Navid Aghasadeghi, Julian Jarzebowski, Timothy Bretl, and Todd P Coleman.
A stochastic control approach to optimally designing hierarchical flash sets in p300 communication prostheses.
IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society, 20(1):102—112, January 2012.

Jaeyoung Park and Kee-Eung Kim.
A pomdp approach to optimizing p300 speller bci paradigm.
IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society, 20:584–94, 04 2012.

Daniel Russo and Benjamin Van Roy.
Learning to optimize via posterior sampling.

Mathematics of Operations Research, 39(4):1221–1243, 2014.

David E. Thompson, Kirsten L. Gruis, and Jane E. Huggins.
A plug-and-play brain-computer interface to operate commercial assistive technology.
Disability and Rehabilitation: Assistive Technology, 9(2):144–150, 2014.

Thank you!