Cutting to the Chase with Warm-Start Contextual Bandits

Bastian Oetomo, R. Malinga Perera, Renata Borovica-Gajic, Benjamin I. P. Rubinstein

School of Computing and Information Systems
The University of Melbourne

{*b.oetomo, malinga.perera*}*@student.unimelb.edu.au*,
{*renata.borovica, brubinstein*}*@unimelb.edu.au*

Motivation

Multi-armed bandits have gained more popularity: news, movie recommendation, crowd sourcing, self-driving databases.

Cold-start problem: Multi-armed bandits suffer relatively poor early round performance due to exploration.

We would like to use some data before the bandit deployment if they exist.

Contextual Bandit

The stochastic contextual multi-armed bandit (MAB) has a setting as follows. In round t, the MAB:

1. observes k possible actions (*arms*) with each arm i having *context vectors* $\boldsymbol{x}_t(i) \in \mathbb{R}^d$;
2. selects or *pulls* an arm $i_t \in [k]$;
3. observes random reward $R_{i_t}(t)$ for the pulled arm i_t which depends on the context $\boldsymbol{x}_t(i)$.

The goal is to minimise the *cumulative regret* up to round T:

$$Reg(T) = \sum_{t=1}^{T} \left(\mathbb{E}[R_{i_t}(t) \mid \boldsymbol{x}_t(i_t)] - \mathbb{E}[R_{i_t^\star}(t) \mid \boldsymbol{x}_t(i_t^\star)] \right) ,$$

where $i_t^\star$ is an optimal arm to pull at round t.

Linear Thompson Sampling [1, 2, 3]

Assume that the expected reward and context have a linear relationship

$$r_t(i_t) = \boldsymbol{\theta}_\star^T \boldsymbol{x}_t(i_t) + \epsilon_t(i_t)$$

for some unknown parameter vector $\boldsymbol{\theta}_\star \in \mathbb{R}^d$

Ridge regression is invoked: $\hat{\boldsymbol{\theta}}_t = \boldsymbol{V}_t^{-1} \sum_{s=1}^{t-1} \boldsymbol{x}_s(i_s) r_s(i_s)$, where $\boldsymbol{V}_t = \lambda \boldsymbol{I} + \sum_{s=1}^{t-1} \boldsymbol{x}_s(i_s) \boldsymbol{x}_s^T(i_s)$.

Adding exploration term yields $\tilde{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_t + \beta_t \boldsymbol{V}_t^{-1/2} \boldsymbol{\eta}_t$, where $\boldsymbol{\eta}_t \sim \mathcal{D}^{TS}$.

Choose an arm i that maximises $\tilde{\boldsymbol{\theta}}_t^T \boldsymbol{x}_t(i)$.

If $\mathcal{D}^{TS}$ follows some concentration and anti-concentration properties, β_t can be chosen in such a way that LinTS is Hannan consistent.

Warm-Starting LinTS

In some cases, some related data exist, which might help us to reduce exploration.

Let $\hat{\boldsymbol{\mu}}$ be our guess for the weight of the first phase dataset with covariance matrix $\boldsymbol{\Sigma}_\mu$. Let $\alpha > 0$ measures the similarity of the two datasets.

Rewriting $\boldsymbol{\theta}_\star = \hat{\boldsymbol{\mu}} + \boldsymbol{\delta}_\star$, then $y_t(i_t) = r_t(i_t) - \hat{\boldsymbol{\mu}}^T \boldsymbol{x}_t(i_t) = \boldsymbol{\delta}_\star^T \boldsymbol{x}_t(i_t) + \epsilon_t(i_t)$.

The initial prior is $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, (\boldsymbol{\Sigma}_\mu + \alpha^{-1} \boldsymbol{I}_d))$ and the posterior is $\mathcal{N}(\hat{\boldsymbol{\delta}}_t, R^2 \boldsymbol{V}_t^{-1})$, where

$$\hat{\boldsymbol{\delta}}_t = \boldsymbol{V}_t^{-1} \sum_{s=1}^{t-1} \boldsymbol{x}_s(i_s) y_s(i_s), \qquad \boldsymbol{V}_t = R^2 \boldsymbol{V}_1 + \sum_{s=1}^{t-1} \boldsymbol{x}_s(i_s) \boldsymbol{x}_s^T(i_s)$$

Warm-Starting LinTS

Algorithm 1 Warm Start Linear Thompson Sampler

Input: $\hat{\boldsymbol{\mu}}, \alpha, \boldsymbol{\Sigma}_\mu, \delta, T, R$

Initialize $\hat{\boldsymbol{\delta}}_1 \leftarrow \mathbf{0}$, $\boldsymbol{V}_1 \leftarrow R^2(\boldsymbol{\Sigma}_\mu + \alpha^{-1}\boldsymbol{I}_d)^{-1}$,

$\delta' \leftarrow \frac{\delta}{4T}, \boldsymbol{b}_1 \leftarrow \mathbf{0}$

for $t = 1, \ldots, T$ **do**

  Sample $\boldsymbol{\eta}_t \sim \mathcal{D}^{TS}$

  $\tilde{\boldsymbol{\theta}}_t \leftarrow \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\delta}}_t + \beta_t(\delta')\boldsymbol{V}_t^{-1/2}\eta_t$

  $i_t \leftarrow s \in \arg\max_{i \in [k]} \tilde{\boldsymbol{\theta}}_t^T \boldsymbol{x}_t(i)$

  Pull arm i_t and observe reward $r_t(i_t) = R_{i_t}(t) | \boldsymbol{x}_t(i_t)$

  $y_t(i_t) \leftarrow r_t(i_t) - \hat{\boldsymbol{\mu}}^T \boldsymbol{x}_t(i_t)$

  $\boldsymbol{V}_{t+1} \leftarrow \boldsymbol{V}_t + \boldsymbol{x}_t(i_t)\boldsymbol{x}_t^T(i_t)$

  $\boldsymbol{b}_{t+1} \leftarrow \boldsymbol{b}_t + y_t(i_t)\boldsymbol{x}_t(i_t)$

  $\hat{\boldsymbol{\delta}}_{t+1} \leftarrow \boldsymbol{V}_{t+1}^{-1}\boldsymbol{b}_{t+1}$

end for

Extension to ϵ-Greedy and LinUCB

The core idea of our method to warm-start is to set up the initial parameters properly.

ϵ-**Greedy:** With the choice of $\hat{\boldsymbol{\theta}}_t$ and $\boldsymbol{V}_t$ as before, explore with probability ϵ by choosing an arm uniformly at random, or exploit with probability $1-\epsilon$ by choosing an arm i that maximises $\hat{\boldsymbol{\theta}}_t^T \boldsymbol{x}_t(i)$.

LinUCB[4]**:** We have $\boldsymbol{\theta}^T \boldsymbol{x} \sim \mathcal{N}((\hat{\boldsymbol{\mu}} + \boldsymbol{V}_t^{-1}\boldsymbol{b}_t)^T \boldsymbol{x}, R^2 \boldsymbol{x}^T \boldsymbol{V}_t^{-1} \boldsymbol{x})$. Hence, we choose an arm which maximises $(\hat{\boldsymbol{\mu}} + \boldsymbol{V}_t^{-1}\boldsymbol{b}_t)^T \boldsymbol{x} + \rho R \sqrt{\boldsymbol{x}^T \boldsymbol{V}_t^{-1} \boldsymbol{x}}$.

In both cases, the update procedure remains the same: subtract $\hat{\boldsymbol{\mu}}^T \boldsymbol{x}_t(i_t)$ from the original reward and fit ridge regression according to the equations for $\hat{\boldsymbol{\theta}}_t$ and $\boldsymbol{V}_t$ as before.

It Performs Better in Early Rounds

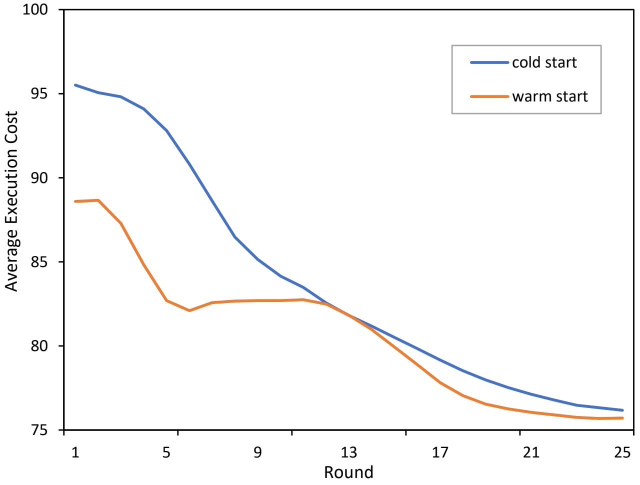

Figure: Index Selection Performance in TPC-H database

It Performs As Good As Baseline

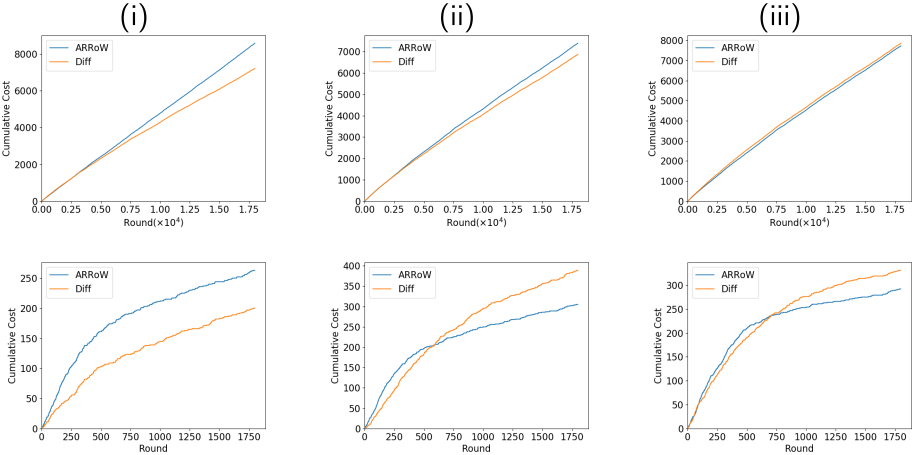

Figure: Comparison between our algorithm and [5] for datasets Letters (top) and Numbers (bottom) with learners (i) ϵ-greedy, (ii) LinUCB and (iii) TS.

Conclusion and Future Work

Setting up the initial parameters provides a way to warm-start linear bandit.

Motivated by Linear Thompson Sampling, the result was extended into ϵ-Greedy and LinUCB.

Warm-starting the bandit improves the performance in the early rounds.

Our method is as good as baseline while provides flexibility on how to choose the initial guess.

References

W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3–4, pp. 285–294, 1933.

S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *International Conference on Machine Learning*, pp. 127–135, 2013.

M. Abeille, A. Lazaric, *et al.*, "Linear Thompson sampling revisited," *Electronic Journal of Statistics*, vol. 11, no. 2, pp. 5165–5197, 2017.

L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*, pp. 661–670, ACM, 2010.

C. Zhang, A. Agarwal, H. D. Iii, J. Langford, and S. Negahban, "Warm-starting contextual bandits: Robustly combining supervised and bandit feedback," in *Proceedings of the 36th International Conference on Machine Learning*, pp. 7335–7344, 2019.