Identifiable latent bandits: Combining observational data and exploration for personalized healthcare

Ahmet Zahid Balcıoğlu, Emil Carlsson, Fredrik D. Johansson
Computer Science and Engineering Department
Chalmers University of Technology, Sweden

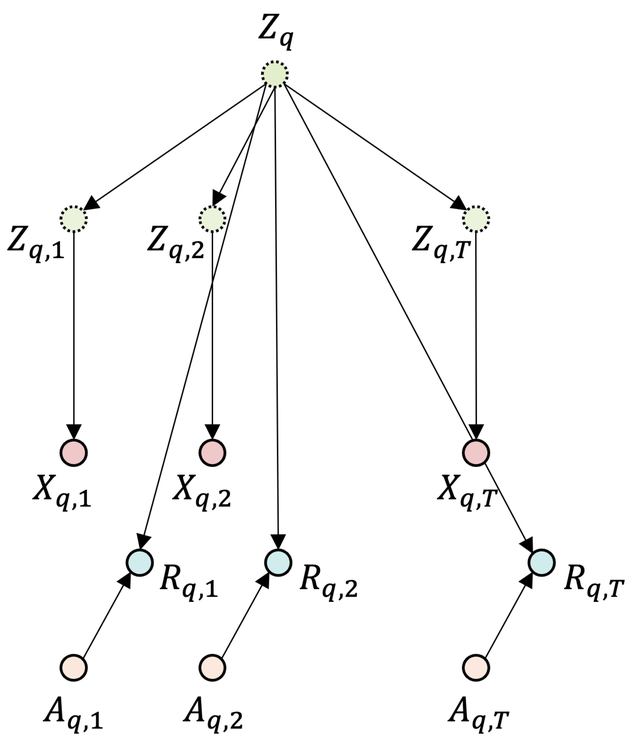
Overview

We turn a **multi-armed bandit** problem into a **linear bandit** problem by using observational data to learn a **latent variable model (LVM)** linear in the reward. We show that such a model enables better personalized treatment exploration in healthcare scenarios with multiple alternative treatments.

Main Contributions:

- Prior work has shown that exploiting the latent structure is more sample efficient for exploration compared to Thompson sampling [1]. We propose a learning algorithm for a latent bandit with a **continuous vector-valued latent state** which is recovered using an **identifiable nonlinear LVM.**
- We introduce mean-contrastive learning, a generalization of identifiability of time-contrastive learning [2].
- We propose two latent bandit algorithms that exploit the latent variable model in the regret minimization setting.
- We show in synthetic data that our algorithms are more **sample-efficient than MAB**, both when a perfect model is used and when the model has been learned from observational data.

Model



$$
\begin{aligned}
&Z_q = U \\
&Z_{q,t} = Z_q + \eta_t \quad \text{for} \quad \eta_t \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}) \\
&X_t = g(Z_q) = g(Z_q + \eta_t) \\
&R = \theta_A^T Z_q + \epsilon_A
\end{aligned} \tag{1}
$$

Assumptions:

(a) Each instance q is generated according to the structural equations (1) where each source variable $Z_{q,t}$ is stationary with respect to patient $q \in [Q]$.

(b) U follows a non-parametric product distribution p_u.

(c) The nonlinear transformation g is smooth and invertible.

Learning an LVM and θ_A using contrastive learning

- We observe patient history $\mathcal{D} = \{(X_t, A_t, R_t)_1^{T_1}, \ldots, (X_t, A_t, R_t)_Q^{T_Q}\}$
- Use observational data X_t, and patient indicators q to train the LVM using contrastive learning.
- Using theorem 1, we have guarantees that the hidden layers of our LVM recovers the true latent state Z.
- Estimate θ_A parameters for the reward model using patient history $\mathcal{D}$ and the estimated Z.

Inference and Results

Theorem 1: Identifiability of Structural Equations

(a) We observe data which is generated by independent sources according to the stuctural equations (1).

(b) The dimension of the LVM ($\mathbf{h}$) is equal to the dimension of the data $\mathbf{x}$.

(c) The patient means $M = [\mathbf{z}_1, ..., \mathbf{z}_Q]^\top \in \mathbb{R}^{n \times Q}$ have rank n; patients are distinct.

Then, in the limit of infinite data, the outputs of the feature extractor are equal to patient mean distribution up to an invertible affine transformation. In other words, with $\hat{\mathbf{z}} := \mathbf{h}(\mathbf{x}; \boldsymbol{\varphi})$,

$$\mathbf{A}\hat{\mathbf{z}} + \mathbf{d} = g^{-1}(x) = \mathbf{z}$$

for some constant invertible matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, a constant vector $\mathbf{d} \in \mathbb{R}^d$.

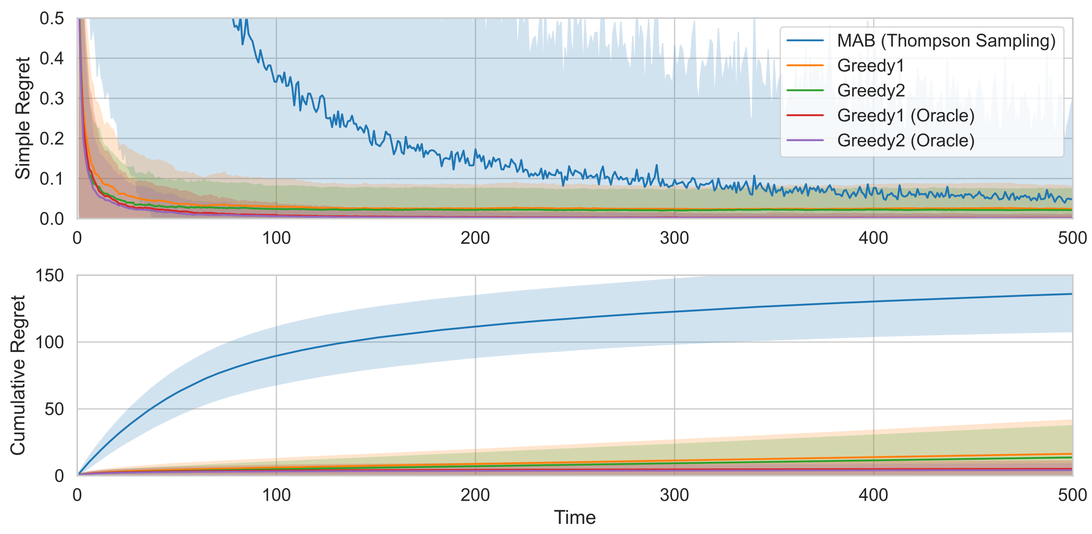
Algorithm 1: Inference Time Greedy1 and Greedy2 Algorithms

Inference Time: Infer Z
1: **for** $t = 1, \ldots, T$ **do**
2: Use the LVM to get an estimate of the latent variable $\hat{\mathbf{z}}_{q,t}$,
3: *For Greedy1*: Update the belief about the true mean $\hat{\mathbf{z}}_q = \hat{\mathbb{E}}[\hat{\mathbf{z}}_{q,t}] := \frac{1}{t} \sum_{t'=1}^{t} \mathbf{h}(\mathbf{x}_{t'})$.
4: *For Greedy2*: Update the belief about the true mean using

$$\hat{\mathbf{z}}_q = \arg\min_{\mathbf{z}} \sum_{t'=1}^{t} \left(R_{t'} - \theta_{A_{t'}}^T \mathbf{z}\right)^2 + \|\mathbf{z} - \hat{\mathbb{E}}[\hat{\mathbf{z}}_{q,t}]\|^2. \quad (2)$$

5: Choose the next action according to $a_t = \arg\max_{a \in \mathcal{A}} \theta_a^T \hat{\mathbf{z}}_q$.
6: **end for**

Empirical Results Simple and cumulative regret for bandit algorithms with the learned LVM and the true latent model (oracle).



LVM fitting results. L layers in the MLP, T_o time steps. Mean correlation coefficient (MCC) for $\hat{z}$ and average R^2 for reward estimates.

L	T_o	MCC_Z	R_R^2
2	100	0.89	0.94
2	200	0.91	0.93
2	300	0.87	0.93
4	100	0.89	0.84
4	200	0.90	0.90
4	300	0.94	0.88

References

[1] Fast Treatment Personalization with Latent Bandits in Fixed-Confidence Pure Exploration
Kinyanjui, Newton Mwai, Emil Carlsson, and Fredrik D. Johansson
Transactions on Machine Learning Research (2023)

[2] Unsupervised feature extraction by time-contrastive learning and nonlinear ICA
Aapo Hyvarinen and Hiroshi Morioka
Advances in neural information processing systems, 2016