Multi-task Linear Bandits

Marta Soare, Ouais Alsharif, Alessandro Lazaric, Joelle Pineau

Motivation

- **Recommendation System:** users with similar features have similar preferences over different items.
- **Personalized Healthcare:** patients with similar symptoms and medical history have the similar reactions to treatments.
- **Games:** an agent might reduce the exploratory steps needed to discover an environment, by using the knowledge acquired on previous similar environments.

Problem Setting

Sequential multi-task linear bandit: The learner faces a sequence of (unknown) linear bandit tasks $(\theta_1, \theta_2, \ldots, \theta_j, \ldots)$.

Linear bandit task:

- Set of arms $\mathcal{X} \subseteq \mathbb{R}^d$, $|\mathcal{X}| = K$ and $||x||_2 \le L$, $\forall x \in \mathcal{X}$.
- Reward model for task j:

$$r_j(x) = x^\top \theta_j + \eta$$

where $\theta_j \in \mathbb{R}^d$ is **unknown** and η is an R sub-Gaussian noise.

- Cumulative regret wrt the optimal arm $x_j^* = \arg\max_{x \in \mathcal{X}} x^\top \theta_j$

$$R_{n_j} = \sum_{s=1}^{n_j} (x_j^* - x_s)^\top \theta_j$$

Task Similarity: there exists $\varepsilon > 0$ such that for any pair of tasks (θ, θ') we have $||\theta - \theta'||_2 \le \varepsilon$.

Better performance on a particular task can be achieved by leveraging information from different but similar tasks.

Tools

Single-task ordinary least-squares (OLS) estimate: for any task j, the estimate of θ_j after n rewards is

$$A_{j,n} = \sum_{t=1}^{n} x_t x_t^\top, \quad b_{j,n} = \sum_{t=1}^{n} x_t r_t, \quad \boxed{\widehat{\theta}_{j,n} = A_{j,n}^{-1} b_{j,n}}$$

Single-task regularized least-squares estimate: for any task j, the estimate of θ_j after n_j rewards is

$$A_{j,n}^{\lambda} = \Big(\sum_{t=1}^{n} x_t x_t^\top + \lambda I\Big), \quad b_{j,n} = \sum_{t=1}^{n} x_t r_t, \quad \boxed{\widehat{\theta}_{j,t}^{\lambda} = (A_{j,n}^{\lambda})^{-1} b_{j,n}}$$

Single-task prediction error Thm.2 in [1] with probability $1-\delta$

$$\boxed{|x^\top \theta_j - x^\top \widehat{\theta}_{j,n}^{\lambda}| \le ||x||_{(A_{j,n}^{\lambda})^{-1}} \left(R\sqrt{d \log\left(\frac{1 + nL^2/\lambda}{\delta}\right)} + \lambda^{1/2} ||\theta_j|| \right) = B_{j,n}(x).}$$

Multi-Task Estimates

Multi-task estimates: use all the past samples to construct an estimate for the current task

$$\widetilde{A}_{m+1,t} = \sum_{j=1}^{m} A_j + A_{m+1,t}; \quad \widetilde{b}_{m+1,t} = \sum_{j=1}^{m} b_j + b_{m+1,t}$$

$$\boxed{\widetilde{\theta}_{m+1,t} = \widetilde{A}_{m+1,t}^{-1} \widetilde{b}_{m+1,t}}$$

Average target task:

$$\mathbb{E}[\widetilde{\theta}_{m+1,t}] = \widetilde{\theta}_{m+1,t}^{*}$$

Multi-task Regularized estimates:

$$\widetilde{A}_{m+1,t}^{\lambda} = \widetilde{A}_{m+1,t} + \lambda I, \quad \boxed{\widetilde{\theta}_{m+1,t}^{\lambda} = (\widetilde{A}_{m+1,t}^{\lambda})^{-1} \cdot \widetilde{b}_{m+1,t}}$$

Use at the same time single-task and multi-task estimates to construct upper confidence bounds $B(x), \widetilde{B}(x)$.

MultiTask-LinUCB

Input: budgets $\{n_j\}_j$, arms $\mathcal{X} \subset \mathbb{R}^d$, regularizer λ
$j = 1$
$A = \lambda I_d, \widetilde{b} = b = 0_d, \widetilde{A}_j = \lambda I_d, \hat{\theta}_j = A^{-1} b$
for $t = 1, \ldots, n_j$ **do**
  Choose: $x_t = \arg\max_{x \in \mathcal{X}} (x_t^\top \hat{\theta}_j + B_{j,t}(x))$
  Observe reward: $r_t = x_t^\top \theta_j + \eta_t$
  Update A, b and the estimate $\hat{\theta}_j = A^{-1} b$
end for
for $j = 2, \ldots, m+1$ **do**
  $\widetilde{A}_j = \widetilde{A}_j + A - \lambda I_d, \ \widetilde{b} = \widetilde{b} + b, \ \widetilde{\theta}_j = \widetilde{A}_j^{-1} \widetilde{b}$
  $A = \lambda I_d, b = 0_d, \hat{\theta}_j = A^{-1} b$
  for $t = 1, \ldots, n_j$ **do**
    $x_t = \arg\max_{x \in \mathcal{X}} \min \left[x^\top \hat{\theta}_j + B_{j,t}(x); \ x^\top \widetilde{\theta}_j + \widetilde{B}_{j,t}(x) \right]$
    Observe reward: $r_t = x_t^\top \theta_j + \eta_t$
    Update: $A, b, \hat{\theta}_j, B_{j,t}, \widetilde{A}_j, \widetilde{b}, \widetilde{\theta}_j, \widetilde{B}_{j,t}$
  end for
end for

Main result

Theorem 1 *Let $\widetilde{\theta}_{m+1,t}^{\lambda}$ be the multi-task regularized least-squares estimate. Then, for any $\delta \ge 0$, for any $t \ge 1$, with probability greater than $1-\delta$ it holds that:*

$$\boxed{|x^\top (\widetilde{\theta}_{m+1,t}^{\lambda} - \theta_{m+1})| \le ||x||_{(\widetilde{A}_{m+1,t}^{\lambda})^{-1}} \left(R\sqrt{d \log\left(\frac{\det\left(\widetilde{A}_{m+1,t}^{\lambda}\right)^{1/2} \det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2} S \right) + x^\top \varepsilon = \widetilde{B}_{m+1,t}(x).}$$

Estimation error

The estimation error of the MT least-squares estimate is upper-bounded by

$$|x^\top (\widetilde{\theta}_{m+1,t}^{\lambda} - \widetilde{\theta}_{m+1,t}^{*})| \le$$

$$||x||_{(\widetilde{A}_{m+1}^{\lambda})^{-1}} \left(R\sqrt{2 \log\left(\frac{\det(\widetilde{A}_{m+1}^{\lambda})^{1/2} \det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2} S \right)$$

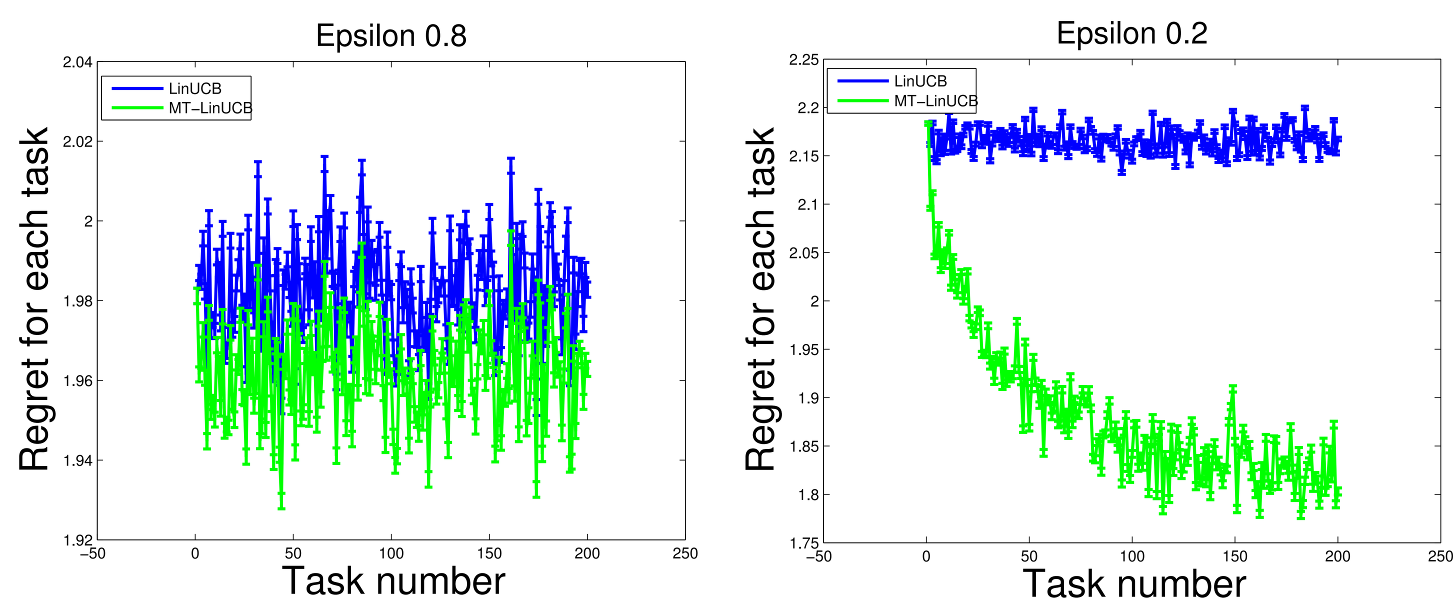
Multi-Task Bias

Under the **task similarity assumption** the approximation error of the multi-task estimates is

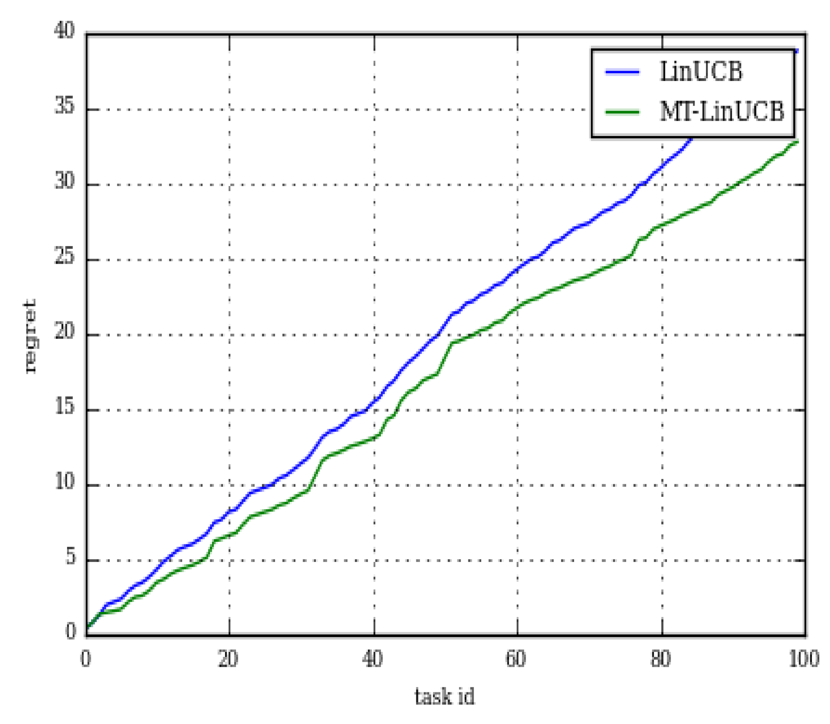
$$||\widetilde{\theta}_{m+1,t}^{*} - \theta_{m+1}|| \le \varepsilon, \text{ w.p. } 1-\delta.$$

Experiments

Synthetic experiments



Boston Housing dataset



Setting

- 200 tasks with $\theta_1, \ldots, \theta_{200} \in \mathbb{R}^2$, randomly generated
- $\max_\theta ||\theta||_2 = 1.1 \cdot \sqrt{2}$, $||\varepsilon||_2$ ranges in $\{0.8, 0.6, 0.4, 0.2\} \cdot \sqrt{2}$
- 100 samples for each task

Results

- MT-LinUCB is never worse than LinUCB.
- As the difference between tasks reduces, the advantage of MT-LinUCB becomes more evident.
- For $\varepsilon = 0.2\sqrt{2}$, the regret of MT-LinUCB decreases with every additional task, while the regret for LinUCB remains constant over time.

- Data points are scaled to have a norm of 1.
- Twenty different clusters, each cluster center is θ_m^* for task m.
- For each task, we drew 10 sets of 3 arms each.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances Neural Information Processing Systems 24*, 2011.

Acknowledgements

Université de Lille 1 Sciences et Technologies
CompLACS
Google
ANR