Theoretical Foundations for Transfer Learning

Inspired by recent work developing theoretical foundations for transfer learning Tripuraneni et al. [2020], we take an additional step to study the provable benefits of transfer learning (via) representation learning for sequential decision making problems. The principal challenge in sequential decision-making systems is to balance exploration of the environment with exploitation of high-reward actions. In a multi-task setting where agents (or bandits) can transfer information between tasks, this paradigm also becomes relevant at a higher level: agents can act collaboratively to learn (mutually beneficial) shared structure or act selfishly to only optimize their own reward. To explore these tradeoff, we study a set of P linear bandit models which are learned concurrently for T steps with underlying parameters coupled to span the same low-dimensional subspace.

More formally, we consider the contextual linear bandit setting where at each round t, the pth bandit learner receives a context $\mathcal{X}_{t,p} \subset \mathbb{R}^d$ and selects an action $\mathbf{x}_{t,p} \in \mathbb{R}^d$ with $\mathbf{x}_{t,p} \in \mathcal{X}_{t,p}$. Given an (unobserved) linear feature matrix $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_r) \in \mathbb{R}^{d \times r}$ with orthonormal columns and r-dimensional task parameter $\boldsymbol{\alpha}_p \in \mathbb{R}^r$ the learner receives a noisily, linearly generated reward:

$$r_{t,p} = \mathbf{x}_{t,p}^\top \mathbf{B} \boldsymbol{\alpha}_p + \xi_{t,p}, \tag{1}$$

where $\xi_{t,p}$ is an i.i.d. noise process. In this framework, the mean rewards vectors $\boldsymbol{\theta}_p = \mathbf{B}\boldsymbol{\alpha}_p$ across bandit instances are coupled since they lie in a common low-dimensionsal subspace (the column space of $\mathbf{B}$). We make the following standard assumptions on the noise, covariates, and parameters of the underlying problems:

- The noise variables $\xi_{t,m}$ are R-subgaussian for all $p \in [P]$ and $t \in [T]$.
- The covariates $\mathbf{x}_{t,p}$ satisfy the bound $\|\mathbf{x}_{t,p}\| \leq L$ for all for all $p \in [P]$ and $t \in [T]$.
- All task parameters satisfy the bound $\|\boldsymbol{\alpha}_p\|_2 \leq S$ for all $p \in [P]$.

In order to state an informal version of a result we consider a situation with completely parallel, collaborative agents. Here each bandit instance symmetrically and simultaneously plays for a total of T rounds in coordination. The goal is for each agent to minimize its regret $\mathcal{R}(T) = \sum_{t=1}^{T} r_{t,p}^\star - r_{t,p}$ where $r_{t,p}^\star$ denotes the best reward available at time t, and instance p. The overall algorithm functions as a two-stage strategy. At a high level, there is a master meta-algorithm which employs an explore-then-commit strategy at the meta-level over all the bandit instances. That is, it commands each bandit learner to sacrifice $E = \lceil \epsilon_p T \rceil$ rounds for $p \in [P]$ of random exploration, in service of learning the shared subspace $\hat{\mathbf{B}}$ (this is done via a modification of the the subspace recovery algorithm from [Tripuraneni et al., 2020]). After this time period each bandit instance can use a softly projected (onto $\hat{\mathbf{B}}$) LinUCB-style algorithm (inspired by the algorithm in [Valko et al., 2014]) to accrue rewards acting in isolation. In order to implement this strategy at first pass we ensure each agent explores equally so $\epsilon_p = \epsilon$. Then, we obtain

Theorem 1 (Informal)**.** *Assume the matrix of task parameters* $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p]$ *satisfies* $\tilde{\kappa}(\mathbf{A}) \leq O(1)$*. Then there is a two-stage parallel meta-ETC strategy with per-agent regret,*

$$\mathcal{R}(T) \leq \tilde{O}(r\sqrt{T}) + \tilde{O}\left((\frac{d^3 r^2 T^2}{P})^{1/3} \right)$$

with probability at least $1 - \gamma$*.*

We now make several comments on the result:

- The aforementioned results shows two-phase behavior for the regret. The first term represents the (optimal) regret achievable by oracle knowledge of the shared subspace. The second term represents the price of the learning of the shared subspace (and bias due to using a stochastic estimate of it in the second stage).

- In the limit $P \to \infty$ a nearly perfect speed-up w.r.t to dimension is obtained: the regret of each instance scales as $\tilde{O}(r\sqrt{T})$ as opposed to the $\tilde{O}(d\sqrt{T})$ regret achieved by acting in isolation. In this limit, the optimal regret can be attained regardless of all other problem parameters.

These results invite two directions for future consideration. First are these results optimal? That is are there better schemes for (meta)-coordination and exploitation? This would require proving matching upper and lower bounds. Additionally, a second exciting direction for consideration is to ask if these results can be extended to the setting of full reinforcement learning which moves beyond that of contextual bandits.

References

Nilesh Tripuraneni, Chi Jin, and Michael I Jordan. Provable meta-learning of linear representations. *arXiv preprint arXiv:2002.11684*, 2020.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomávs Kocák. Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*, pages 46–54, 2014.