Meta learning via Linear Representation

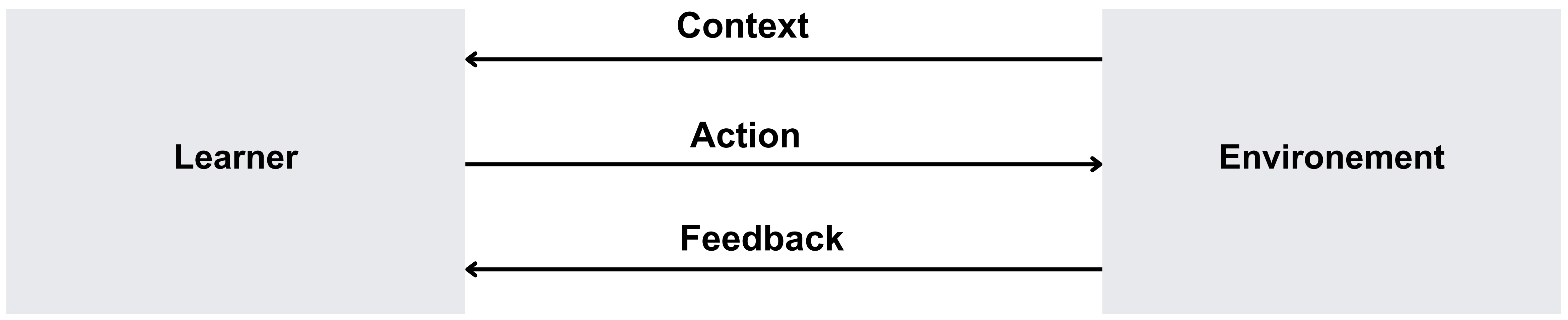
Yessin Moakher, Paul Le Van Kiem, David Kerriou

Outline

1. Problem Formulation: Meta-Learning for Contextual Linear Bandits
2. Representation Estimation Methods
3. Experimental Results

Contextual linear bandits

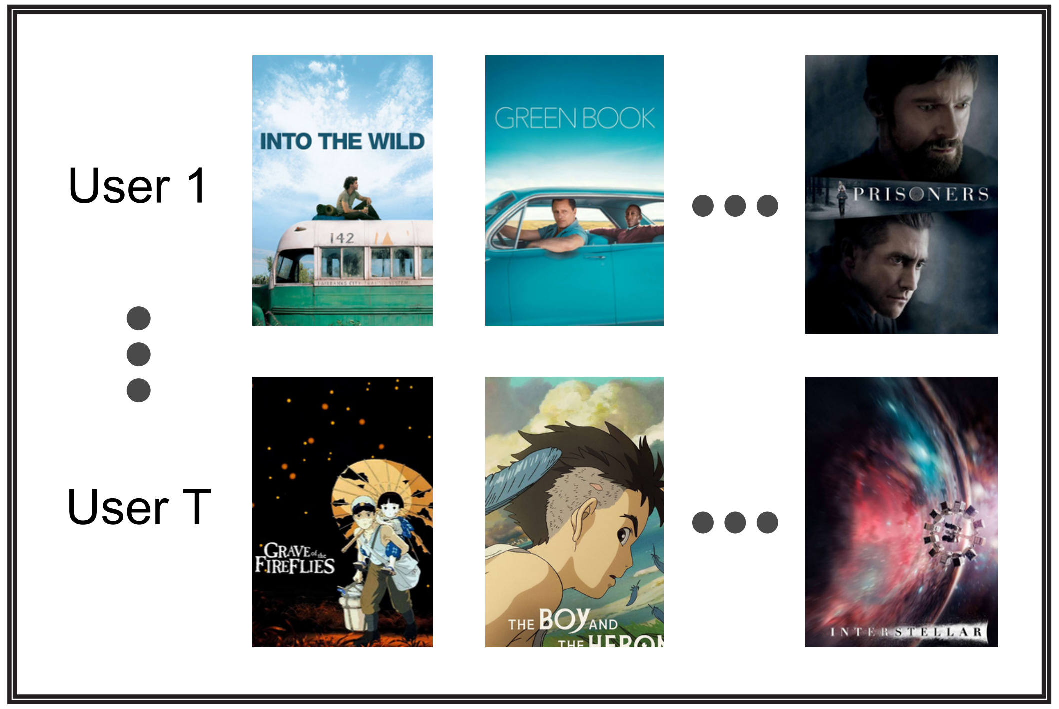
Interaction with user t on iteration n



$$\underbrace{y_{t,n}}_{\text{feedback}} = \underbrace{x_{t,n}^\top}_{\text{action given the context}} \underbrace{w_t^*}_{\text{unknown}} + \underbrace{\eta_{t,n}}_{\text{noise}}$$

Meta learning

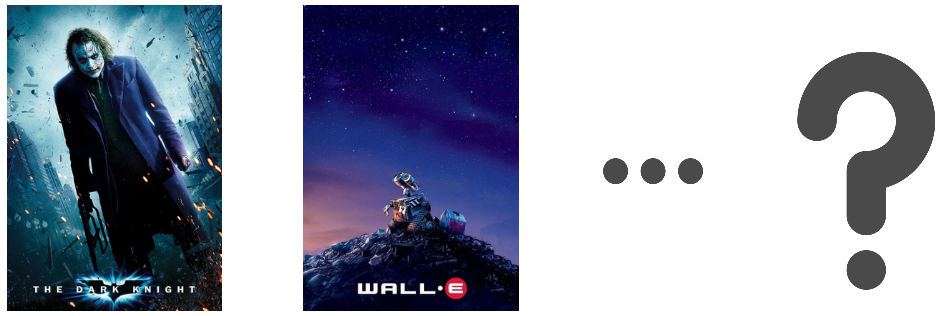
Movies recommender system



Dataset:

$$\{(x_{t,n}, y_{t,n})\}_{n=1,t=1}^{N,T}$$

New user :

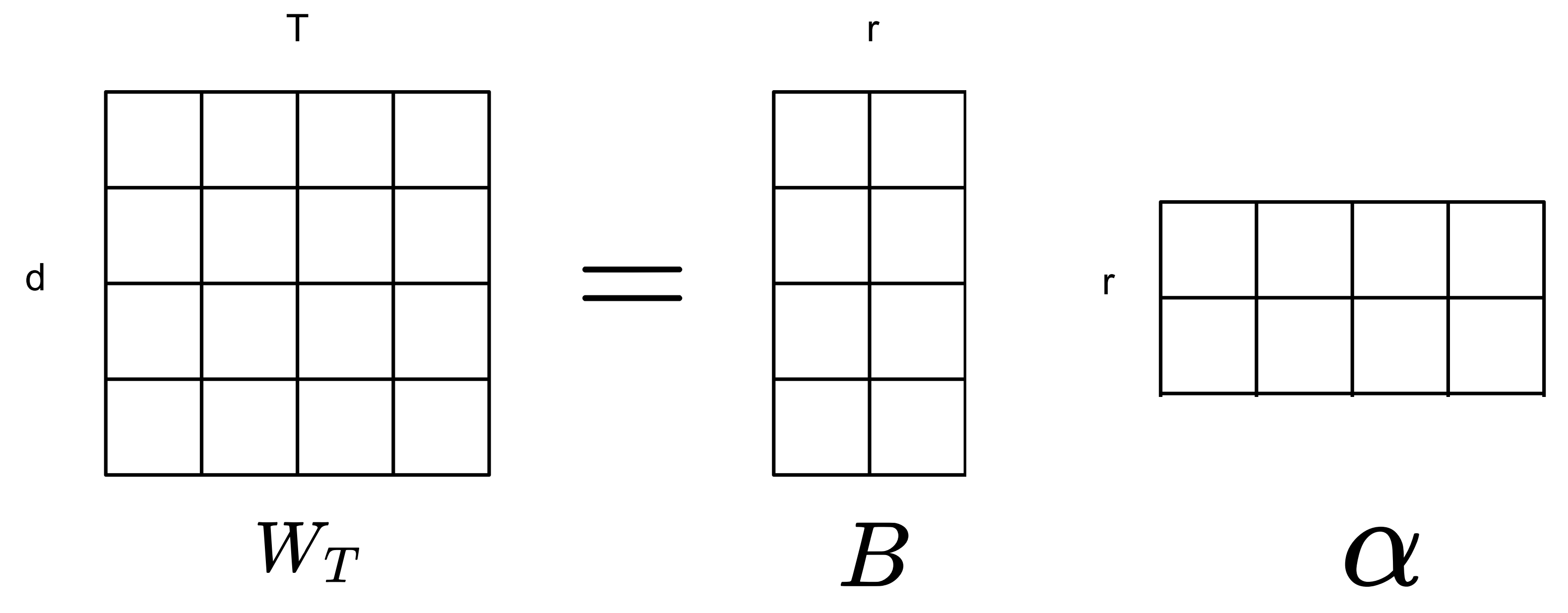


few labels

Meta learning using representation learning

Low rank assumption:

$$\mathbf{W_T} = [w_1, \ldots, w_T]$$



- Learn the unknown low-dimensional representation B shared across all tasks.
- Learn the task-specific vector for the new task (using, for example, simple linear regression in our case).

Meta learning using representation learning

Greedy policy algorithm :

$$y_{T+1,n} = \langle x_{T+1,n}, B\alpha_{T+1} \rangle + \eta_{T+1,n}, n \in \{1, \ldots, N\}$$

After estimating B, the new task can be learned through the following algorithm:

- First step, n = 1 : Select a random action and observe the feedback.
- For n=2 to N do :
  - $\hat{\alpha}_{T+1,n} \in \arg\min \sum_{i=1}^{n-1} \left(y_{T+1,i} - \langle x_{T+1,i}, \hat{B}\alpha \rangle \right)^2$
  - $x_{T+1,n} \in \arg\max_{x \in D_{T+1,n}} \langle x, \hat{B}\hat{\alpha}_{T+1,n} \rangle$
  - Observe feedback

Estimating B

First Method:

Trace norm regularization [L.Cella, K.Lounici, G.Pacreau and M.Pontil, ICML'23]

- We have a regression problem under the assumption of a low-rank matrix. The problem is defined as follows:

$$min_{W\in\mathcal{C},\mathrm{rank}(W)\ll\min(d,T)} \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{t,i}-\langle\mathbf{x}_{t,i},W_t\rangle\right)^2,$$

- We convexify this problem by using the trace norm:

$$\min_{W\in\mathcal{C}} \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\underbrace{\left(y_{t,i}-\langle\mathbf{x}_{t,i},W_t\rangle\right)^2}_{\text{regression}}+\underbrace{\lambda\|W\|_*}_{\text{regularization}},$$

- **Use SVD of Ŵ to get B.**

Trace norm regularization

Numerical simulation

[S.Ji and J.Ye, ICML'09]

- Computing the sub-gradient of the trace norm, making proximal algorithms computationally expensive. We use an algorithm based on the fact that:

$$D_\lambda(A) = \arg\min_W \frac{1}{2}\|W - A\|_F^2 + \lambda\|W\|_*,$$

where $D_\lambda(A) = U\Sigma_\lambda V^T$, and Σ_λ diagonal matrix having $(\Sigma_\lambda)_{ii} = \max\{0, \Sigma_{ii} - \lambda\}$

- to optimize the following problem:

$$\min_W f(W) + \lambda\|W\|_*,$$

- by linearizing the function f using its quadratic approximation.

$$P_{tk}(W, W_{k-1}) = f(W_{k-1}) + \langle W - W_{k-1}, \nabla f(W_{k-1})\rangle + \frac{t_k}{2}\|W - W_{k-1}\|_F^2$$

Trace norm regularization

Regression matrix error

Theorem 1:

The estimation error on W is given by

$$\frac{1}{T}\|\widehat{W} - W_T\|_F^2 \leq c(\frac{rT\lambda^2}{n} + C^2\Xi_n(u))$$

$$\text{where } \Xi_n(u) = \frac{u}{nT} + \frac{r(d+T)(d+u+\log(T))}{n^2T}$$

- Optimality condition + assumption on the distribution

Estimating B

Second Method:

Method of moments [N.Tripuraneni, C.Jin, and M.Jordan, ICML'21]

- We calculate the moment of order 2 :

$$\Sigma = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i^2 x_i x_i^T\right) = 2\Gamma + (1 + \mathrm{tr}(\Gamma))\,\mathrm{I_d},$$

$$\text{where } \Gamma = \frac{1}{n}\sum_{i=1}^{n} B\alpha_i\alpha_i^T B^T.$$

- We can retrieve the space spanned by the columns of B by applying PCA on Σ with dimension r, the rank of B.

Estimating B

Why does it work ?

$$\Sigma = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i^2 x_i x_i^T\right) = 2\Gamma + (1 + \mathrm{tr}(\Gamma))\,\mathrm{I_d},$$

$$\text{where } \Gamma = \frac{1}{n}\sum_{i=1}^{n} B\alpha_i\alpha_i^T B^T.$$

- This works because the space spanned by the r-th first eigenvectors of Γ is equal to the space spanned by the columns of B
- We must know r

Contribution :

- We showed that rg(Γ) = rg(B)
- Why not try to estimate r by estimating the rank of Γ ?

Method of moments

Error on Σ :

Theorem 2:

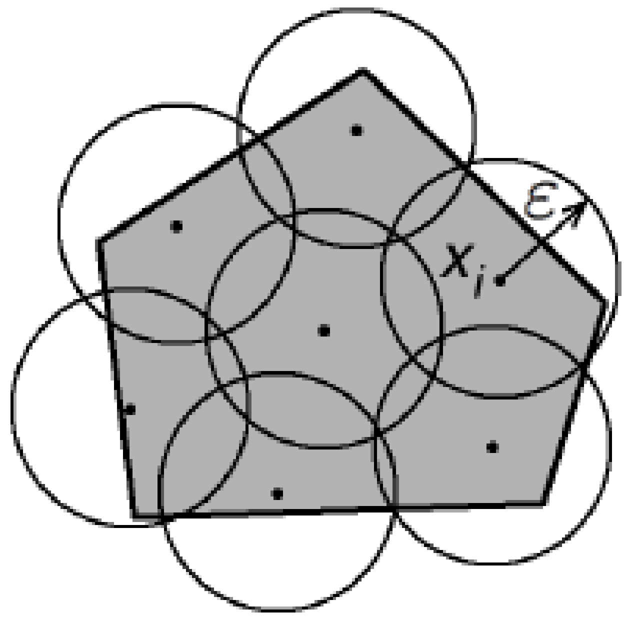
The estimation error on Σ is given by

$$\|\Sigma_n - \Sigma\| = \tilde{O}_{\mathbb{P}}((1 \vee \max_{1\leq i\leq n} \|w_i\| \vee \max_{1\leq i\leq n} \|w_i\|^6)(\sqrt{\frac{d}{n}} \vee \frac{d}{n}))$$

Method of moments

Proof sketch : Epsilon-Net

Recouvrement d’un ensemble par des boules de taille ε



Norme d’opérateur d’une matrice symétrique avec un ε-Net N de la sphère unité

$$\|A\| \leq \frac{1}{1-2\epsilon} \sup_{x \in N} |\langle Ax, x \rangle|$$

Il existe un recouvrement de la sphère unité par un ε-Net de taille finie

Method of moments

Proof sketch :

$$
\begin{aligned}
\Sigma_n - \Sigma \quad &= \frac{1}{n}\sum_{i=1}^{n} \eta_i^2 x_i x_i^T - \mathbb{E}(\eta_i^2 x_i x_i^T) \\
&+ \frac{1}{n}\sum_{i=1}^{n} 2\eta_i x_i^T w_i x_i x_i^T - \mathbb{E}(2\eta_i x_i^T w_i x_i x_i^T) \\
&+ \frac{1}{n}\sum_{i=1}^{n} x_i^T w_i w_i^T x_i x_i x_i^T - \mathbb{E}(x_i^T w_i w_i^T x_i x_i x_i^T)
\end{aligned}
$$

- Eta-conditioning
- Hanson-Wright

- Eta-conditioning
- Projection on w
- Hanson-Wright
- Bernstein

Error on the estimation of B

- Both methods 1 and 2 rely on estimating a matrix and subsequently using it to estimate its singular vectors.
- **But what guarantees the “convergence” of the singular vectors ?**

Without loss of generality, using the symmetrization trick, we examine the "convergence" of the eigenspaces.

- $$A = U\Sigma V^T, \quad \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} u_k \\ \pm v_k \end{bmatrix} = \pm\sigma_k \begin{bmatrix} u_k \\ \pm v_k \end{bmatrix}$$
- $$\| \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \| = \|A\|$$

Error on the estimation of B

Theorem 3:

$$\text{If } \|\Sigma - \Sigma_n\| < \frac{\min(\lambda_r - \lambda_{r+1}, \lambda_{s-1} - \lambda_s)}{4}$$

$$\text{then } \|P_{\lambda_j} - \widehat{P_{\lambda_j}}\| \leq \frac{4 \cdot \|\Sigma - \Sigma_n\|}{\min(\lambda_r - \lambda_{r+1}, \lambda_{s-1} - \lambda_s)}$$

- Use the notion of the resolvent of a matrix to derive an analytical expression for the projection matrix of the eigenspaces.

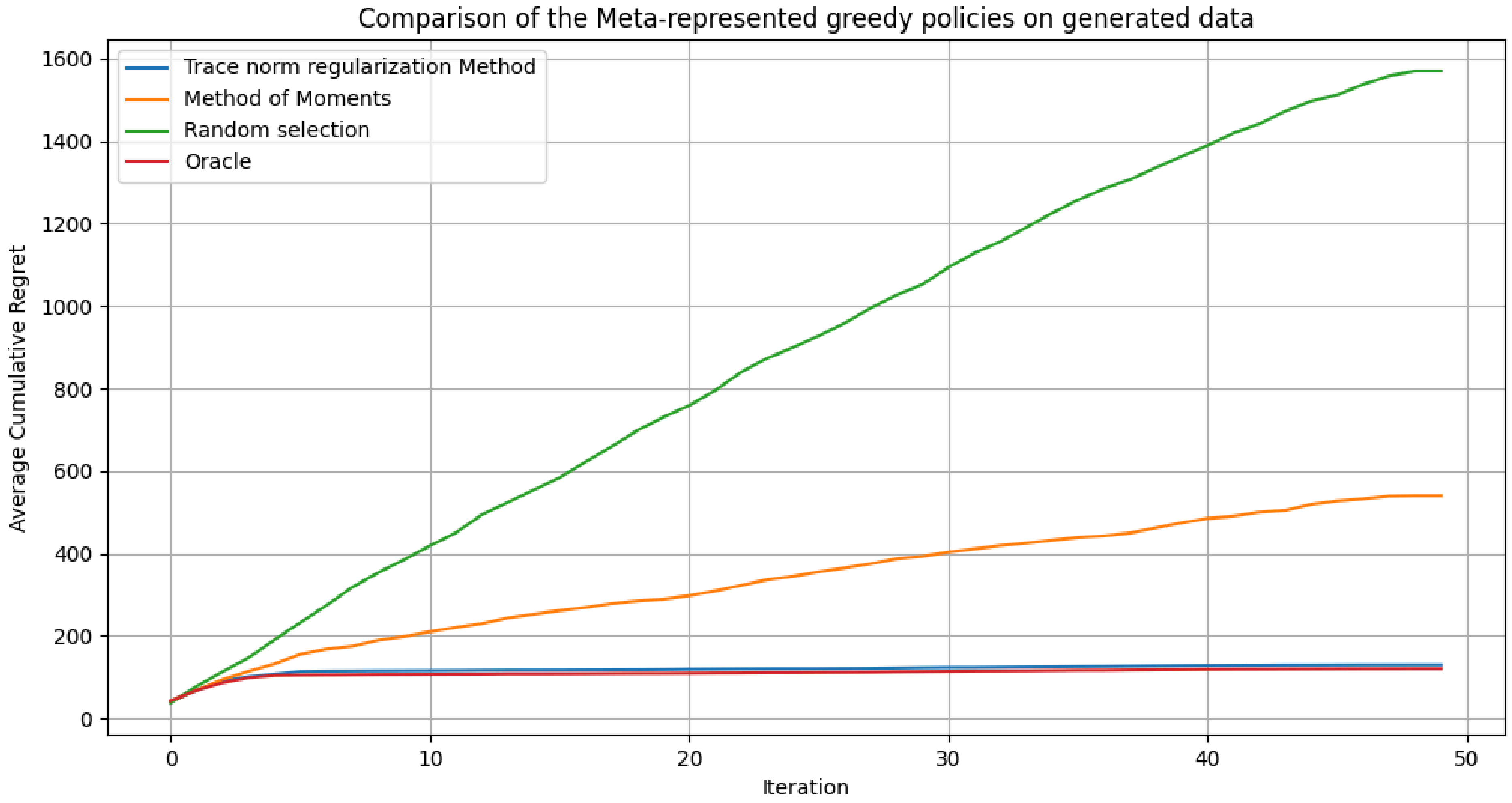
$$R_A(z) = (A - zI)^{-1} \qquad\qquad P_{\lambda_i} = \frac{1}{2\pi i}\oint_{\Gamma_i} R_A(\eta)\, d\eta$$

- The two matrices need to be close relative to a notion of “spectral gap”.

Simulation

Synthetic data : I.i.d., gaussian random variables.

$$R(T,N) = \sum_{t=1}^{T}\sum_{n=1}^{N}(x_{t,n}^* - x_{t,n})^\top w_t^*, \qquad \text{with } x_{t,n}^* = \arg\max_{x\in D_{t,n}} x^\top w_t^*.$$



- Sub linear regret of order of

$$\sqrt{rN}\left(1 \vee \sqrt{\frac{d}{T}}\right)$$

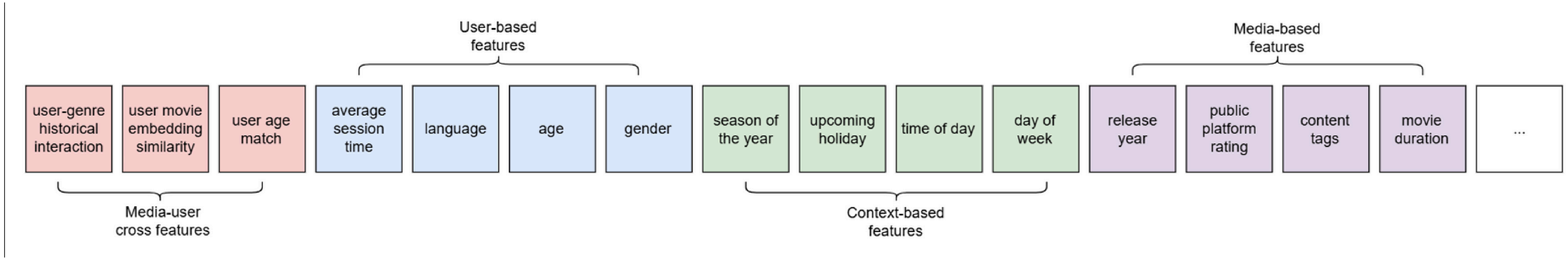
- Importance of tuning λ

Movie lens

Dataset description

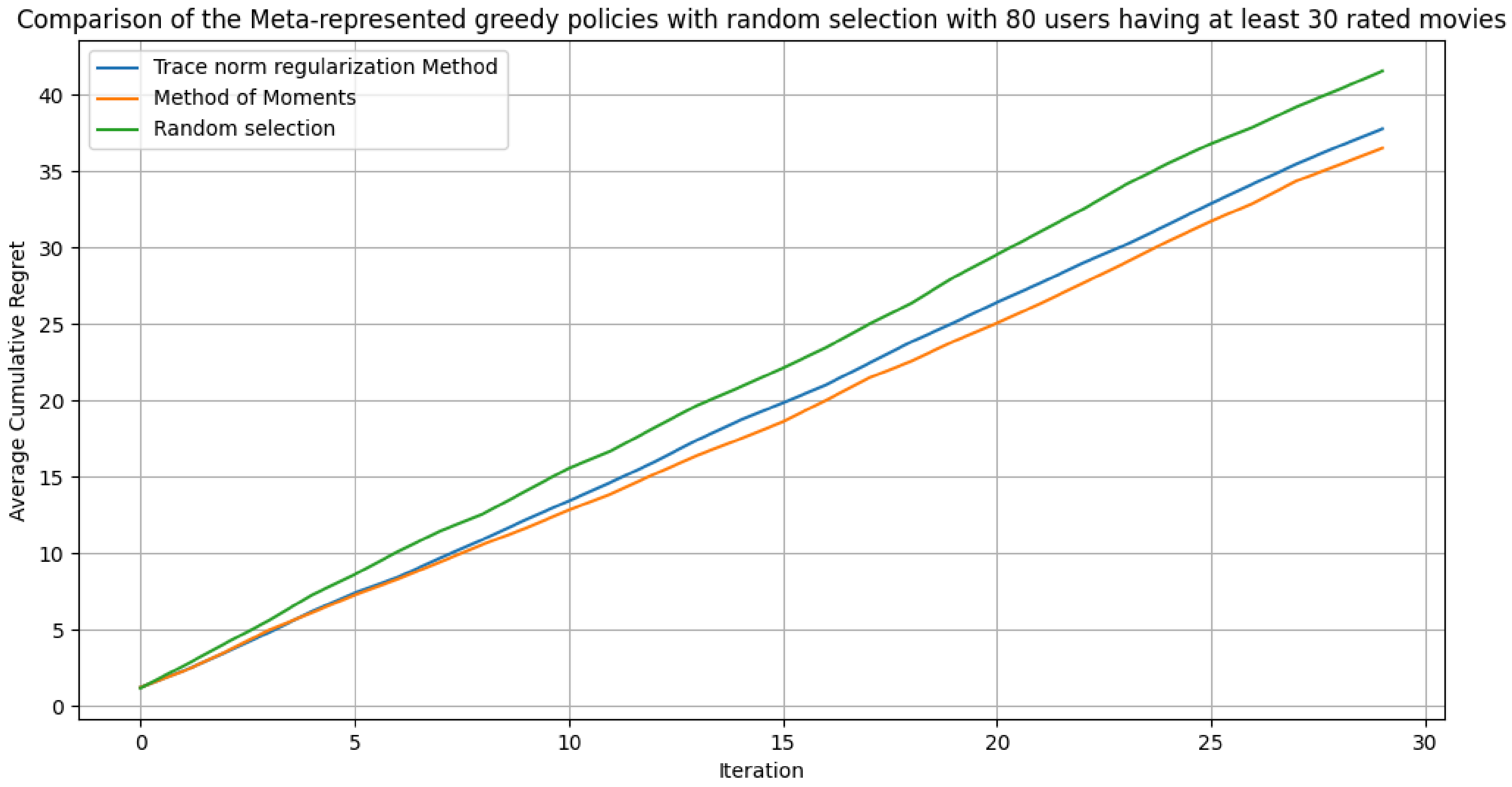
Movies lens contains 100000 ratings by 943 users on 1682 movies.
Each user has rated at least 20 movies. we have *movie_id* , *rating_date, release_date, genre, age, gender,* *occupation*.

Ideal x :



Simulation

Movie lens data :



- Theory predicts sublinear regret.

References

- Leonardo Cella, Karim Lounici, Grégoire Pacreau, and Massimiliano Pontil. Multi-task representation learning with stochastic linear bandits. In Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, 2023.
- Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm mini-mization. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML ’09, New York, NY, USA, 2009.
- Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In Marina Meila and Tong Zhang, Proceedings of the 38th International Conference on Machine Learning, 2021.