Learning Personalized Attribute Preference via Multitask AUC Optimization

Zhiyong Yang Qianqian Xu Xiaochun Cao, Qingming Huang

Presented By Zhiyong Yang

https://joshuaas.github.io

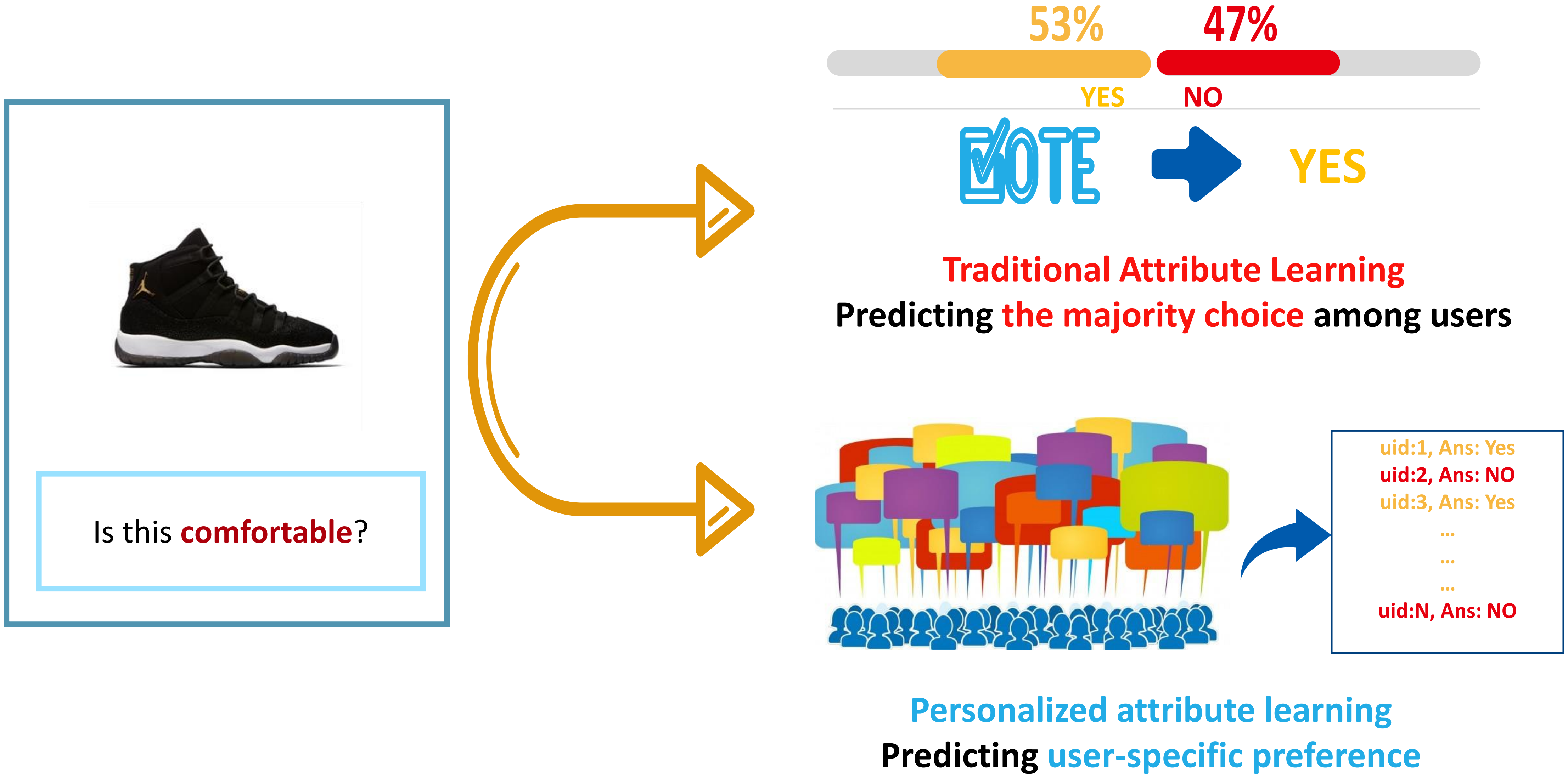
Outline

- Introduction
- Related Work
- Methodology
- Theoretical Analyses
- Experiments

Outline

- Introduction
- Related Work
- Methodology
- Theoretical Analyses
- Experiments

Attribute Learning
Consensus vs. Personalization



Personalized Attribute Learning
The problems

Problem B: How to guarantee that a positive labeled instance has a higher rank than negatively labeled instances ?

Find me a pair of sporty shoes

Positive Instances

Negative Instances

More Sporty

Less Sporty

Problem A： How to model the correlation of the user behaviors?

Contributions

- We propose a novel model for personalized attribution learning:
  - **For problem A :** Regarding the annotations prediction for each user as a specific task, we proposed **a three-level decomposition.**
  - **For problem B :** AUC optimization based framework along with **an efficient evaluation method.**
- For theoretical analysis, we have the following contributions:
  - 1) A novel closed-form solution
  - 2) A novel evaluation method for AUC loss and gradients, which yields ***20x speed-up*** at most.
  - 3) The convergence analysis and novel generation error bound

Outline

- Introduction
- **Related Work**
- Methodology
- Theoretical Analyses
- Experiments

Related Work

- Attribute Learning
  - 1) User adaption based method (Kovashka and Grauman 2013)
  - 2) Word shade discovery based method (Kovashka and Grauman 2013)
  - Require ***pre-training on a large pool*** or ***initialization with specific method*** ; ***Neglect*** the merit of AUC; ***No*** theoretical guarantee.
- Multi-task Learning
  - The most relevant works are grouping and clustering based multi-task learning algorithms
  - We extends these methods with ***AUC optimization***, ***new closed-form solution*** and ***new generalization error bounds***

Outline

- Introduction
- Related Work
- **Methodology**
- Theoretical Analyses
- Experiments

Roadmap

For Problem A → Task Parameter Modeling → Three-level Decomposition

For Problem B → AUC optimization → Novel Efficient Evaluation Method

Optimization → Proximal Gradient Descent → Novel closed-form Solution

Notations and Problem Definition

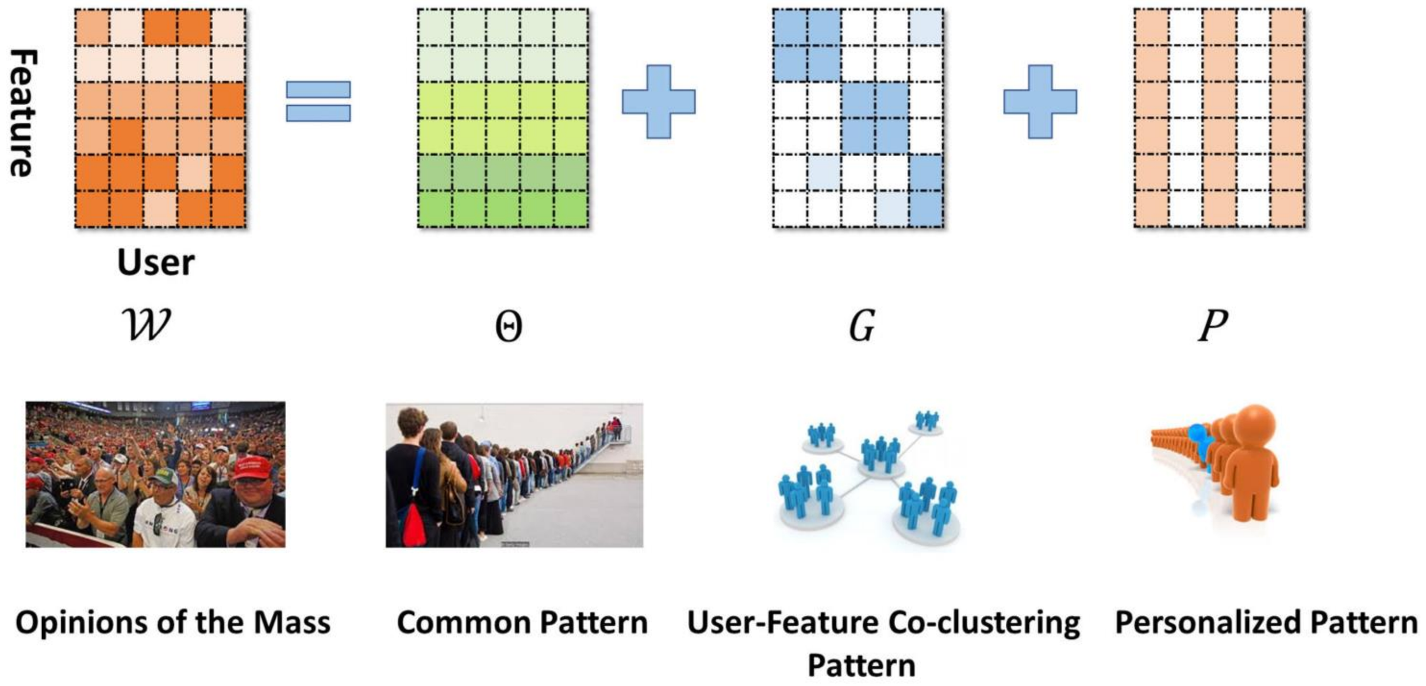
- #User: U; For a given user i, #positively/negatively labeled instances are denotes as $n_{+,i}$/$n_{-,i}$; $\mathcal{S}_{+,i} = \{k \mid y_k^{(i)} = 1\}$ and $\mathcal{S}_{-,i} = \{k \mid y_k^{(i)} = -1\}$.

- Data Set: $\mathcal{S} = \left\{(\boldsymbol{X}^{(1)}, \boldsymbol{y}^{(1)}), \cdots, (\boldsymbol{X}^{(U)}, \boldsymbol{y}^{(U)})\right\}$

- Our goal : Learn an attribute ranker for each user $\boldsymbol{f}^{(i)}(\boldsymbol{x}) = \boldsymbol{W}^{(i)\top}\boldsymbol{x}$.

A Hierarchical Decomposition of the Task Parameters

- θ: the common factor ***shared among users***(tasks)
- G: expected to be ***block-diagonal***
- P: expected to be ***column-wise sparse*** (only outlier users have non-zero values)



The AUC Loss

- Population AUC loss: The probability to mis-rank positive, negative instance pairs *(unknown distribution, non-differentiable)*

- Sample based AUC loss : The frequency on the sample *(non-differentiable)*

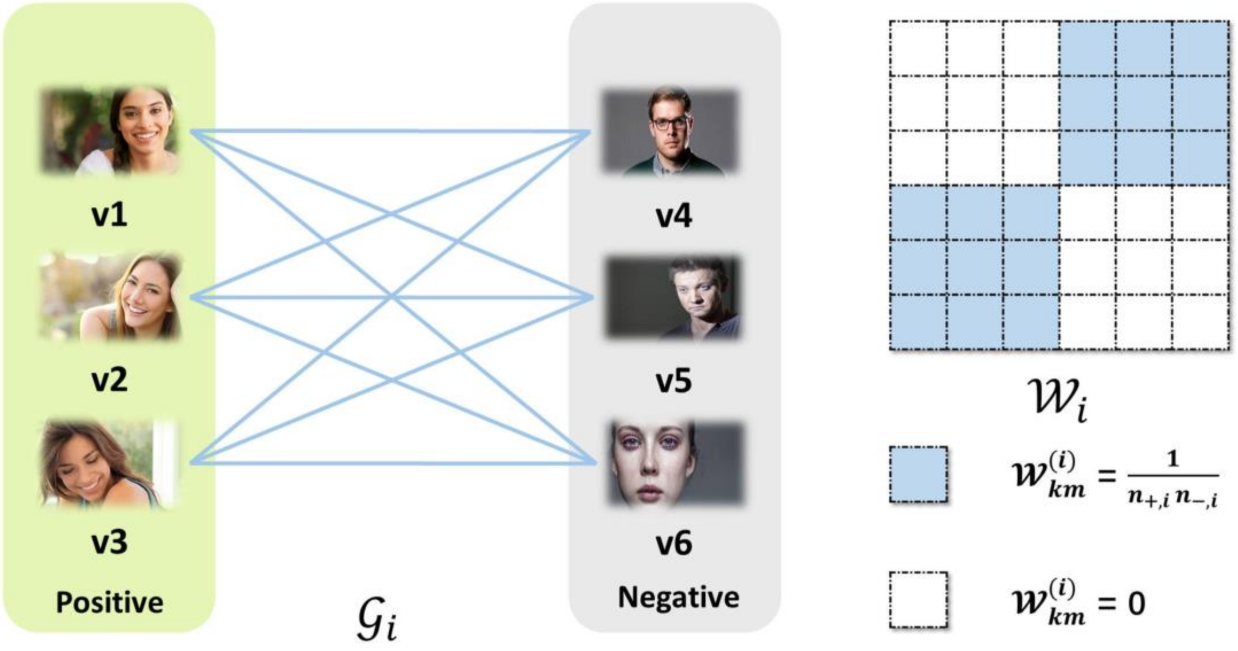
$$\ell_{AUC}^{(i)} = \sum_{x_p \in \mathcal{S}_{+,i}} \sum_{x_q \in \mathcal{S}_{-,i}} \frac{I(\boldsymbol{x}_p, \boldsymbol{x}_q)}{n_{+,i} n_{-,i}}$$

- Sample based surrogate loss: using the square function $s(t) = (1-t)^2$ to replace the mis-ranking indicator (*available and differentiable*)

$$\ell_i(\boldsymbol{f}^{(i)}, \boldsymbol{y}^{(i)}) = \sum_{x_p \in \mathcal{S}_{+,i}} \sum_{x_q \in \mathcal{S}_{-,i}} \frac{s\left(\boldsymbol{f}^{(i)}(\boldsymbol{x}_p) - \boldsymbol{f}^{(i)}(\boldsymbol{x}_q)\right)}{n_{+,i} n_{-,i}}$$

Efficient Evaluation

- ***AUC-graph for each user***: a bipartite graph with edges cross different labels.
  - **Vertexes** : The image objects to be labeled.
  - **Edges**: between every pair of positive and negative instances with weight $\mathcal{W}_{km}^{(i)} = \frac{1}{n_{+,i}n_{-,i}}$.
- The AUC loss is a quadratic form of the graph Laplacian $\boldsymbol{L}^{(i)}$



Proposition 1. *For any $\boldsymbol{A} \in \mathbb{R}^{n\times a}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times b}$,where a and b are positive integers. $\boldsymbol{A}^\top \boldsymbol{L}^{(i)}\boldsymbol{B}$, and $\boldsymbol{A}^\top \boldsymbol{L}^{(i)}$ could be finished within $\mathcal{O}\big(n_i(a+b+ab)\big) = \mathcal{O}(abn_i)$ and $\mathcal{O}(an_i)$, respectively.*

The overall loss

$$(P^*)\ \min_{\boldsymbol{\theta},\boldsymbol{G},\boldsymbol{P}} \underbrace{\sum_{i} \sum_{\boldsymbol{x}_p \in \mathcal{S}_{+,i}} \sum_{\boldsymbol{x}_q \in \mathcal{S}_{-,i}} \frac{s\Big({\boldsymbol{W}^{(i)}}^\top (\boldsymbol{x}_p - \boldsymbol{x}_q)\Big)}{n_{+,i} n_{-,i}}}_{\mathcal{L}(\boldsymbol{W})}$$

$$+ \lambda_1 \underbrace{\|\boldsymbol{\theta}\|_2^2}_{\mathcal{R}_1(\boldsymbol{\theta})} + \lambda_2 \underbrace{\sum_{\kappa+1}^{min\{d,U\}} \sigma_i^2(\boldsymbol{G})}_{\mathcal{R}_2(\boldsymbol{G})} + \lambda_3 \underbrace{\|\boldsymbol{P}\|_{1,2}}_{\mathcal{R}_3(\boldsymbol{P})}$$

$$s.t \quad \boldsymbol{W}^{(i)} = \boldsymbol{\theta} + \boldsymbol{G}^{(i)} + \boldsymbol{P}^{(i)}$$

Optimization
Proximal Gradient Descent

For each iteration step k, giving a reference point $\boldsymbol{W}^{ref_k} = (\boldsymbol{\theta}^{ref_k}, \boldsymbol{G}^{ref_k}, \boldsymbol{P}^{ref_k})$, then the proximal gradient method updates the variables as :

$$\boldsymbol{\theta}^k := \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{2} \left\| \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}^k \right\|_2^2 + \frac{\lambda_1}{\rho_k} \|\boldsymbol{\theta}\|_2^2 \tag{2}$$

$$\boxed{\boldsymbol{G}^k := \underset{\boldsymbol{G}}{\operatorname{argmin}} \frac{1}{2} \left\| \boldsymbol{G} - \tilde{\boldsymbol{G}}^k \right\|_F^2 + \frac{\lambda_2}{\rho_k} \sum_{\kappa+1}^{min\{d,U\}} \sigma_i^2(\boldsymbol{G})} \tag{3}$$

No existing closed-form solutions

$$\boldsymbol{P}^k := \underset{\boldsymbol{P}}{\operatorname{argmin}} \frac{1}{2} \left\| \boldsymbol{P} - \tilde{\boldsymbol{P}}^k \right\|_F^2 + \frac{\lambda_3}{\rho_k} \|\boldsymbol{P}\|_{1,2} \tag{4}$$

where: $\tilde{\boldsymbol{\theta}}^k = \boldsymbol{\theta}^{ref_k} - \frac{1}{\rho_k} \nabla_\theta \mathcal{L}(\boldsymbol{W}^{ref_k})$, $\tilde{\boldsymbol{G}}^k = \boldsymbol{G}^{ref_k} - \frac{1}{\rho_k} \nabla_{\boldsymbol{G}} \mathcal{L}(\boldsymbol{W}^{ref_k})$, $\tilde{\boldsymbol{P}}^k = \boldsymbol{P}^{ref_k} - \frac{1}{\rho_k} \nabla_P \mathcal{L}(\boldsymbol{W}^{ref_k})$

Optimization
A novel closed-form solution

Proposition 2. *An Optimal Solution of (3) is:*

$$\boldsymbol{G}^* = \boldsymbol{U}\mathcal{T}_{\kappa,\frac{\lambda_3}{\rho_k}}(\boldsymbol{\Sigma})\boldsymbol{V}^\top \tag{6}$$

where $\boldsymbol{U\Sigma V}^\top$ *is a SVD decomposition of* $\tilde{\boldsymbol{G}}^k$, $\mathcal{T}_{\kappa,c}$ *maps* $\boldsymbol{\Sigma} = diag(\sigma_1, \cdots, \cdots, \sigma_{min\{d,U\}})$ *to a diagonal matrix having the same size with* $\mathcal{T}_{\kappa,c}(\boldsymbol{\Sigma})_{ii} = (\frac{1}{2c+1})^{I[i>\kappa]}\sigma_i$.

Outline

- Introduction
- Related Work
- Methodology
- **Theoretical Analyses**
- Experiments

Lipschitz Continuity of the Gradients

- Lipschitz Continuity of the Gradients is a well-known assumption for the proximal gradient methods.
- In Theorem 1, we show that this assumption is satisfied by our method.
- The results show that the Lipschitz depends on:
  - the ***number of users,***
  - ***the numerical stability of the inputs***
  - ***the degree to which the labels are imbalanced***

Theorem 1 (Lipschitz Continuous Gradient)**.** *Suppose that the data is bounded in the sense that:*

$$\forall i,\ \|\boldsymbol{X}^{(i)}\|_2 = \sigma_{X_i} < \infty,\ n_{+,i} \geq 1,\ n_{-,i} \geq 1.$$

Given two arbitrary distinct parameters $\boldsymbol{W}, \boldsymbol{W}'$ *:*

$$\|\nabla\mathcal{L}(vec(\boldsymbol{W})) - \nabla\mathcal{L}(vec(\boldsymbol{W}'))\| \leq \gamma\Delta\boldsymbol{W}$$

where: $\gamma = 3U\sqrt{(2U+1)}\max_i\left\{\frac{n_i\sigma^2_{X_i}}{n_{+,i}n_{-,i}}\right\}$, $vec(\boldsymbol{W}) = [\boldsymbol{\theta}, vec(\boldsymbol{G}), vec(\boldsymbol{P})]$, $\Delta\boldsymbol{W} = \|vec(\boldsymbol{W}) - vec(\boldsymbol{W}')\|$.

Convergence Analysis

Lemma 1. *The function* $\sum_{i=\kappa+1}^{min\{d,U\}} \sigma_i^2(\mathbf{G})$ *is continuous with respect to* $\mathbf{G}$.

Remarks:

- The regularization term on G is a ***non-convex*** and ***non-smooth spectral*** function.
- It is hard to analysis the loss function based on convex definition of the sub-gradients.
- By virtue of Lemma 1, the generalized sub-gradient for lower semi-continuous functions (defined in Rockafellar and Wets 2009) is then well-defined for the loss function

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media,2009

Convergence Analysis

Theorem 2. *Assume that the initial solutions $\boldsymbol{\theta}^0, \boldsymbol{G}^0, \boldsymbol{P}^0$ are bounded, the following properties hold :*

Sufficient descent

- *1) The sequence $\{\mathcal{F}(\boldsymbol{\theta}^k, \boldsymbol{G}^k, \boldsymbol{P}^k)\}$ is non-increasing in the sense that :* $\forall k, \exists C_{k+1} > 0$

$$\boxed{\mathcal{F}(\boldsymbol{\theta}^{k+1}, \boldsymbol{G}^{k+1}, \boldsymbol{P}^{k+1}) \leq \mathcal{F}(\boldsymbol{\theta}^k, \boldsymbol{G}^k, \boldsymbol{P}^k) - C_{k+1}(\|\Delta(\boldsymbol{\theta}^k)\|_2^2 + \|\Delta(\boldsymbol{G}^k)\|_F^2 + \|\Delta(\boldsymbol{P}^k)\|_F^2)}$$

- *2)* $\lim_{k\to\infty} \boldsymbol{\theta}^k - \boldsymbol{\theta}^{k+1} = 0,\ \lim_{k\to\infty} \boldsymbol{G}^k - \boldsymbol{G}^{k+1} = 0,\ \lim_{k\to\infty} \boldsymbol{P}^k - \boldsymbol{P}^{k+1} = 0.$
- *3) The parameter sequences $\{\boldsymbol{\theta}^k\}_k$, $\{\boldsymbol{G}^k\}_k$, $\{\boldsymbol{P}^k\}_k$ are bounded*
- *4) Every limit point of $\{\boldsymbol{\theta}^k, \boldsymbol{G}^k, \boldsymbol{P}^k\}_k$ is a critical point of the problem.*
- *5)* $\forall\, T \geq 1, \exists\, C_T > 0$ *:*

sublinear convergence rate

$$\boxed{\min_{0\leq k<T} \left(\|\Delta(\boldsymbol{\theta}^k)\|_2^2\right) \leq \frac{C_T}{T},\ \min_{0\leq k<T} \left(\|\Delta(\boldsymbol{G}^k)\|_F^2\right) \leq \frac{C_T}{T},\ \min_{0\leq k<T} \left(\|\Delta(\boldsymbol{P}^k)\|_F^2\right) \leq \frac{C_T}{T}.}$$

Generalization Error Bound

- Hypothesis Space:

$$\Theta = \left\{(\boldsymbol{\theta}, \boldsymbol{G}, \boldsymbol{P}) : \sqrt{\mathcal{R}_1(\boldsymbol{\theta})} \leq \psi_1, \mathcal{R}_2(\boldsymbol{G}) \leq \psi_2, \|\boldsymbol{G}\|_2 \leq \sigma_{max} < \infty, \mathcal{R}_3(\boldsymbol{P}) \leq \psi_3\right\}$$

- We have the following Bound：

Population loss

Surrogate loss On the sample

Theorem 3. *Assume that* $\exists \Delta_\chi > 0$*, all the instances are sampled such that,* $\|x\| \leq \Delta_\chi$ *.Define* $C = (\psi_1 + \sqrt{\psi_2 + \kappa \cdot \sigma_{max}^2} + \psi_3)\,\zeta$ *as* $\zeta = \Delta_\chi C$*, we have, for all* $\delta \in (0,1)$*, for all* $(\boldsymbol{\theta}, \boldsymbol{G}, \boldsymbol{P}) \in \Theta$ *:*

$$\boxed{\mathbb{E}_{\mathcal{D}}\Big(\sum_i \ell_{AUC}^{(i)}\Big)} \leq \boxed{\mathcal{L}(\boldsymbol{W})} + \boxed{\sum_{i=1}^{U} \frac{B_1}{\sqrt{(n_i\chi_i(1-\chi_i))}} + B_2\sqrt{\frac{ln(\frac{2}{\delta})}{\sum_{i=1}^{U} n_i\chi_i(1-\chi_i)}}}$$

holds with probability at least $1-\delta$*, where* $B_1 = 8\sqrt{2}C\Delta_\chi(1+\zeta)$*,* $B_2 = 10\sqrt{2}(1+\zeta)\zeta$*,* $\chi_i = \frac{n_{+,i}}{n_i}$*. The distribution* $\mathcal{D} = \otimes_{i=1}^{U}(\mathcal{D}_{+,i} \otimes \mathcal{D}_{-,i})$*, where for user* i*,* $\mathcal{D}_{+,i}$*,* $\mathcal{D}_{-,i}$ *are conditional distributions for positive and negative instances, respectively.*

$$\boxed{\mathcal{O}\Big(\sum_{i=1}^{U} \frac{1}{\sqrt{(n_i\chi_i(1-\chi_i))}}\Big)}$$

Outline

- Introduction & motivation
- Related Work
- Methodology
- Theoretical Analyses
- Experiments

Competitors

- Standard Lasso
- Robust Multi-Task Learning (RMTL)
- Robust Multi-Task Feature Learning (rMTFL)
- Joint Feature Learning (JFL)
- The Clustered Multi-Task Learning Method (CMTL)
- The task-feature coclusters based multi-task method (COMT)
- Reduced Rank Multi-stage multi-task learning (RAMU)

Simulated Dataset

- We generate a simulated dataset with 100 users and ***500,000*** annotations.
- Table 1 shows the average performance comparison, where our algorithm significantly outperforms the second-best.
- From Table 2, we see that the proposed AUC evaluation method yields an ***20x speed-up at most***

Table 1: AUC Comparison on Simulation Dataset

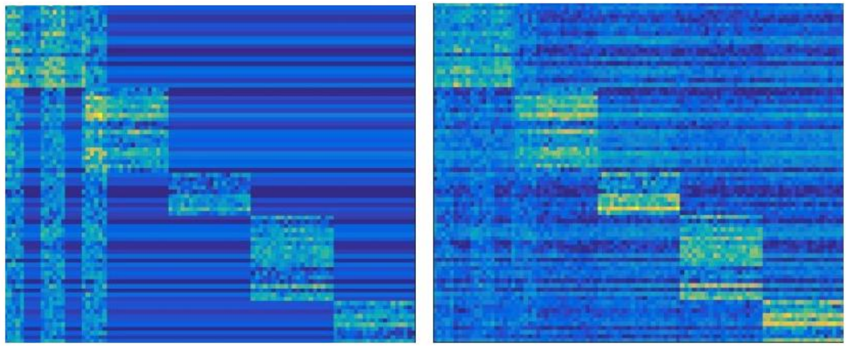
Alg	RMTL	rMTFL	LASSO	JFL
mean	83.48	83.45	83.57	83.49
Alg	CMTL	COMT	RAMU	Ours
mean	83.47	83.44	83.50	**99.65**

Table 2: Running Time Comparison (seconds): Original s-tands for the original AUC evaluation, wheres ours stands for our acceleration scheme.

ratio	20%	40%	60%	80%	100%
Orginal	18.57	74.22	151.86	268.55	nan
Ours	**3.06**	**5.50**	**8.65**	**12.46**	**15.82**

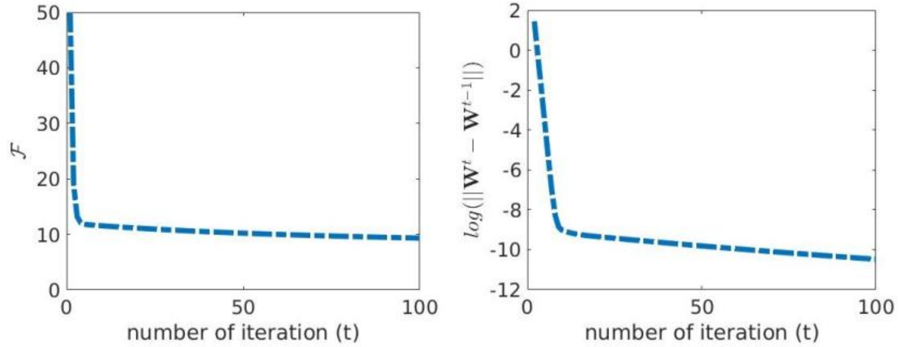
Simulated Dataset

- Figure 3 shows that our algorithm could roughly recover the structure of the true parameters .
- In Figure 4, we see that the convergence behavior coincides with the theoretical results.



(a) Ground-Truth (b) Learned Parameter

Figure 3: The Potential of our proposed method to Recover the Expected Structure of the Parameters

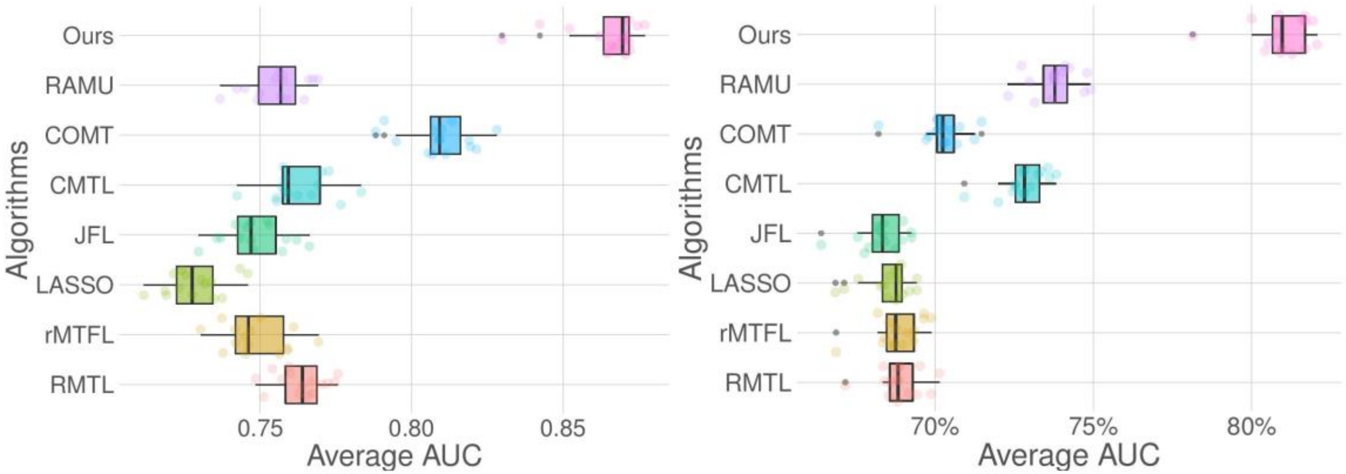


(a) Loss Convergence (b) Paramter Convergence

Figure 4: The Convergence Behavior On Simulation Dataset: a)shows the loss convergence, whereas b) exhibits the convergence property in terms of the parameters.

Real World Datasets

- **Shoes Dataset** contains *14,658* online shopping images and *90,000* personalized annotations on *7* shoes attributes .
- **Sun Attribute Dataset** contains roughly 14,000 scene images, 64,900 personalized annotations on 5 attributes
- The performance comparison results show the superiority of our algorithm.



(a) Shoes Dataset (b) Sun Attribute Dataset

Alg	Attibutes											
	Shoes							Sun				
	BR	CM	FA	FM	OP	ON	PT	CL	MO	OP	RU	SO
RMTL	79.31	84.99	66.90	85.08	75.67	67.22	75.14	69.36	62.71	75.28	67.91	69.23
rMTFL	70.90	83.78	67.27	85.91	73.71	65.21	77.11	69.27	62.15	75.80	68.16	68.76
LASSO	68.46	80.48	65.90	84.01	71.47	64.60	75.08	67.64	61.83	75.39	68.57	69.13
JFL	72.00	83.10	67.26	85.93	73.02	65.39	77.09	68.63	61.94	75.00	67.17	68.78
CMTL	74.54	85.16	68.21	85.32	75.06	68.17	77.62	72.55	66.61	79.78	72.34	72.82
COMT	84.24	88.68	69.66	89.19	80.93	72.99	80.62	70.69	63.72	76.93	69.43	70.44
RAMU	78.33	84.58	65.78	84.68	75.25	66.72	73.50	72.95	69.25	79.81	74.39	72.50
Ours	**92.95**	**90.92**	**73.24**	**92.65**	**87.95**	**81.07**	**86.22**	**79.31**	**78.19**	**86.50**	**81.88**	**78.98**

THANKS!