Supplemental Materials: Fast and Robust Multi-View Multi-Task Learning via Structured Sparsity

Proofs for Section: The AGILE method

Proof of Lemma 1

Proof. According to the definition of $\mathbf{U}_t$ and $\mathcal{W}_t$, we have

$$\mathbf{U}_t\mathcal{W}_t = \sum_{v=1}^{V} \mathbf{U}_t^v \mathbf{w}_t^v \mathbf{e}_v^\top = \sum_{v=1}^{V} \|\mathbf{U}_t^v \mathbf{w}_t^v\|_2 \frac{\mathbf{U}_t^v \mathbf{w}_t^v}{\|\mathbf{U}_t^v \mathbf{w}_t^v\|_2} \mathbf{e}_v^\top, \tag{1}$$

where $\mathbf{e}_v$ is the vector of canonical basis with its v-th element equal to 1 and 0 otherwise.

- Predictions from multiple views are orthogonal. In this case, $\{\frac{\mathbf{U}_t^v \mathbf{w}_t^v}{\|\mathbf{U}_t^v \mathbf{w}_t^v\|_2}\}_{v=1}^V$ are orthonormal to each other. Recall that the trace norm actually equal to the ℓ_1-norm on the singular values. According to (1) and singular value decomposition, we have

$$\|\mathbf{U}_t\mathcal{W}_t\|_* = \sum_{v=1}^{V} |\, \|\mathbf{U}_t^v \mathbf{w}_t^v\|_2 \,| = \|\mathbf{U}_t\mathcal{W}_t\|_{1,2} = \|\mathbf{U}_t\mathbf{w}_t\|_{G_1}. \tag{2}$$

  According to Cauchy–Schwarz inequality, the following inequality holds,

$$\sum_{v=1}^{V} \|\mathbf{U}_t^v \mathbf{w}_t^v\|_2 \leq \sum_{v=1}^{V} \|\mathbf{U}_t^v\|_{op} \|\mathbf{w}_t^v\|_2. \tag{3}$$

  where $\|\cdot\|_{op}$ denotes the matrix operator norm, such that given an arbitrary matrix $\mathbf{A}$, $\|\mathbf{A}\|_{op}$ equals to the largest singular value of $\mathbf{A}$,

- Predictions from multiple views are consistent, i.e., $\mathbf{U}_t^v \mathbf{w}_t^v = \mathbf{U}_t^{v'} \mathbf{w}_t^{v'}$, $v \neq v'$. Thus, we have

$$\mathbf{U}_t\mathcal{W}_t = \mathbf{U}_t^1 \mathbf{w}_t^1 \mathbf{1}_V^\top = \sqrt{V} \left\|\mathbf{U}_t^1 \mathbf{w}_t^1\right\|_2 \frac{\mathbf{U}_t^1 \mathbf{w}_t^1}{\|\mathbf{U}_t^1 \mathbf{w}_t^1\|_2} \frac{\mathbf{1}_V^\top}{\sqrt{V}}. \tag{4}$$

Then the following equality holds,

$$\|\mathbf{U}_t\mathcal{W}_t\|_* = |\sqrt{V} \left\|\mathbf{U}_t^1 \mathbf{w}_t^1\right\|_2| = \sqrt{\sum_{v=1}^{V} \|\mathbf{U}_t^v \mathbf{w}_t^v\|_2^2} = \|\mathbf{U}_t\mathcal{W}_t\|_F = \|\mathbf{U}_t\mathbf{w}_t\|_2. \tag{5}$$

Again, according to Cauchy–Schwarz inequality, we have

$$\|\mathbf{U}_t\mathbf{w}_t\|_2 \leq \|\mathbf{U}_t\|_{op} \|\mathbf{w}_t\|_2. \tag{6}$$

□

Proof of Lemma 2

To verify Lemma 2, we need to prove the correctness of the following Lemma.

Lemma 1. *For any real matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, the following inequality holds;*

$$\|\mathbf{A}\|_F \leq \|\mathbf{A}\|_* \leq \|\mathbf{A}\|_{2,1}. \tag{7}$$

Proof. For any real matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, we have the following definitions for matrix norms

$$\|\mathbf{A}\|_F = \sqrt{\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top)}, \quad \|\mathbf{A}\|_* = \mathrm{Tr}(\sqrt{\mathbf{A}\mathbf{A}^\top}), \tag{8}$$

and

$$\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{n} \|\mathbf{a}_{i\cdot}\|_2 = \mathrm{Tr}(\sqrt{\mathrm{Diag}(\mathbf{A}\mathbf{A}^\top)}), \tag{9}$$

where $\text{Diag}(\mathbf{A}\mathbf{A}^\top) \in \mathbb{R}^{n\times n}$ denotes a diagonal matrix with same diagonal elements of $\mathbf{A}\mathbf{A}^\top \in \mathbb{R}^{n\times n}$. Let the SVD of $\mathbf{A}$ be $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times r}$, $\mathbf{V} \in \mathbb{R}^{m\times r}$ and $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ with the i-th diagonal $\mathbf{\Sigma}_{ii} = \sigma_i$, we have

$$\|\mathbf{A}\|_F = \sum_{i=1}^{r} \sigma_i^2, \quad \|\mathbf{A}\|_* = \sum_{i=1}^{r} \sigma_i, \tag{10}$$

indicating that $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_*$ equal to ℓ_2-norm and ℓ_1-norm of singular values of $\mathbf{A}$, respectively. Since for an arbitrary vector $\mathbf{v}$, the inequality $\|\mathbf{v}\|_2 \le \|\mathbf{v}\|_1$ is always satisfied, we then have $\|\mathbf{A}\|_F \le \|\mathbf{A}\|_*$.

Let $\mathbf{M} = \sqrt{\mathbf{A}\mathbf{A}^\top}$ and $\mathbf{m}_{i\cdot}$ be the i-th row of $\mathbf{M}$, we have

$$\left(\sqrt{\text{Diag}(\mathbf{A}\mathbf{A}^\top)}\right)_{ii} = \|\mathbf{m}_{i\cdot}\|_2 = \sqrt{m_{ii}^2 + \sum_{j\neq i} m_{ij}^2} \ge m_{ii} = \left(\sqrt{\mathbf{A}\mathbf{A}^\top}\right)_{ii}, \ i \in [n], \tag{11}$$

where m_{ii} is the i-th diagonal element of $\mathbf{M}$. Hence we have $\|\mathbf{A}\|_* \le \|\mathbf{A}\|_{2,1}$. □

According to Lemma 1 and the fact $\left\|\mathbf{A}^\top\right\|_{2,1} = \|\mathbf{A}\|_{1,2}$, we have the following inequality

$$\|\mathbf{U}_t\mathcal{W}_t\|_F \le \|\mathbf{U}_t\mathcal{W}_t\|_* \le \|\mathbf{U}_t\mathcal{W}_t\|_{1,2}. \tag{12}$$

Given the facts that $\|\mathbf{U}_t\mathcal{W}_t\|_{1,2} = \|\mathbf{U}_t\mathbf{w}_t\|_{G_1}$ in (2) and $\|\mathbf{U}_t\mathcal{W}_t\|_F = \|\mathbf{U}_t\mathbf{w}_t\|_2$ in (5), the above inequality becomes

$$\|\mathbf{U}_t\mathbf{w}_t\|_2 \le \|\mathbf{U}_t\mathcal{W}_t\|_* \le \|\mathbf{U}_t\mathbf{w}_t\|_{G_1}. \tag{13}$$

Implementation and convergence analysis for Section: Optimization algorithm

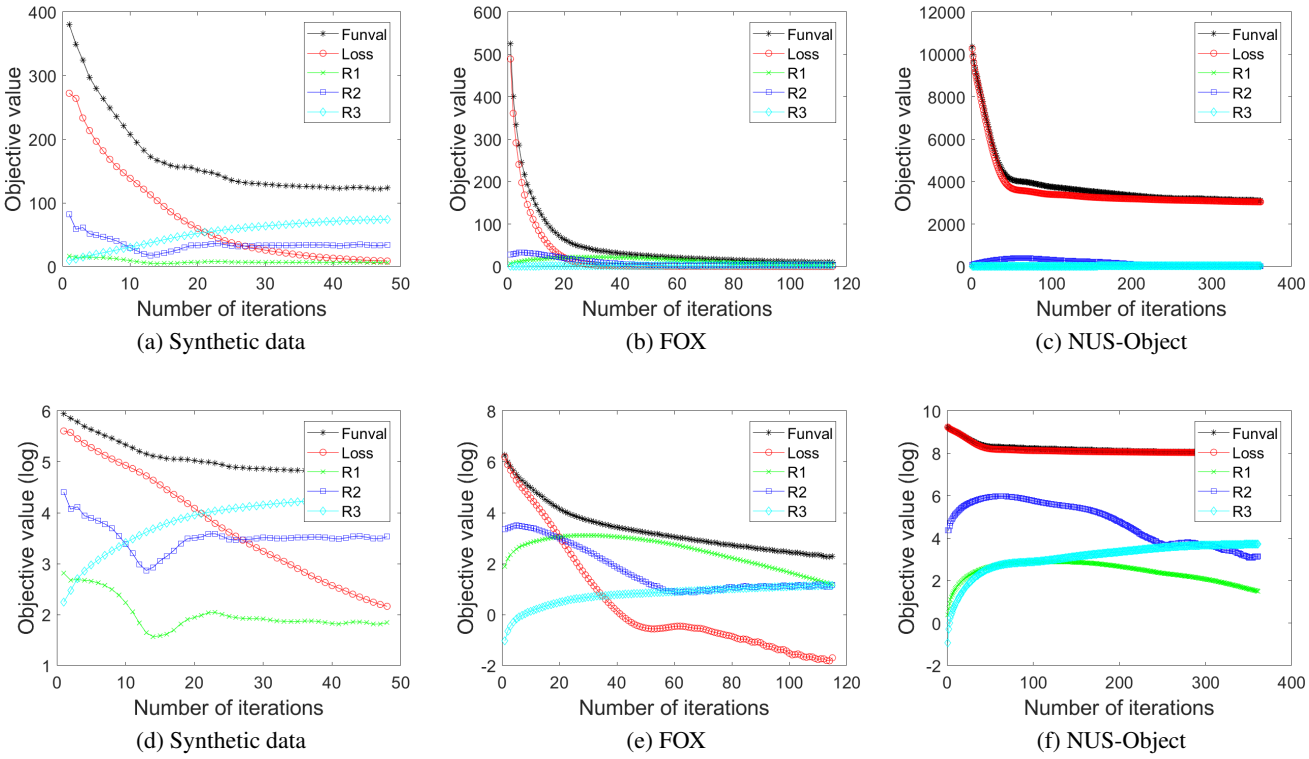
In Algorithm 1, we provide the optimization algorithm of AGILE discussed in Sec.4 of the main paper. Note that, we apply fast-ADMM [Goldstein *et al.*, 2014] and Accelerated Proximal Method (APM) [Nesterov, 2013] in Algorithm 1 to accelerate the optimization algorithm, leading to a quadratic convergence rate. In Algorithm 1, prox-$\ell_{2,1}$ and prox-ℓ_{G_1} denote the proximal operator for $\ell_{2,1}$-norm regularized problem and ℓ_{G_1}-norm regularized problem, respectively.

Algorithm 1 AGILE: Optimization algorithm

Input: $\{\{\mathbf{X}_t^v\}_{v=1}^V\}_{t=1}^T$, $\{\{\mathbf{U}_t^v\}_{v=1}^V\}_{t=1}^T$, $\{\mathbf{y}_t\}_{t=1}^T$, α, β, γ.
Output: $\mathbf{\Theta} = \mathbf{W} + \mathbf{H}$.

Initialize $\mathbf{W}$, $\mathbf{H}$, $\mathbf{P}$ and $\mathbf{Q}$, and set $a_0, a_1 \coloneqq 1$, $k \coloneqq 1$.
repeat
  $\hat{\mathbf{W}}^{(k)} \leftarrow \mathbf{W}^{(k)} + \frac{a_{k-1}-1}{a_k}\left(\mathbf{W}^{(k)} - \mathbf{W}^{(k-1)}\right)$.
  $\hat{\mathbf{H}}^{(k)} \leftarrow \mathbf{H}^{(k)} + \frac{a_{k-1}-1}{a_k}\left(\mathbf{H}^{(k)} - \mathbf{H}^{(k-1)}\right)$.
  $\widehat{\mathbf{P}}^{(k)} \leftarrow \mathbf{P}^{(k)} + \frac{a_{k-1}-1}{a_k}\left(\mathbf{P}^{(k)} - \mathbf{P}^{(k-1)}\right)$.
  $\widehat{\mathbf{Q}}^{(k)} \leftarrow \mathbf{Q}^{(k)} + \frac{a_{k-1}-1}{a_k}\left(\mathbf{Q}^{(k)} - \mathbf{Q}^{(k-1)}\right)$.
  Determine the learning rate η by line search.
  $\mathbf{W}^{(k+1)} \leftarrow \text{prox-}\ell_{2,1}\left(\hat{\mathbf{W}}^{(k)} - \eta\nabla_w f(\hat{\mathbf{W}}^{(k)}), \alpha\eta\right)$.
  $\mathbf{H}^{(k+1)} \leftarrow \text{prox-}\ell_{G_1}\left(\hat{\mathbf{H}}^{(k)} - \eta\nabla_h f(\hat{\mathbf{H}}^{(k)}), \gamma\eta\right)$.
  $[\mathbf{M}^{(k)}, \mathbf{\Sigma}^{(k)}, \mathbf{N}^{(k)}] \leftarrow \text{SVD}\left(\mathbf{U}\mathcal{W}^{(k)} - \widehat{\mathbf{Q}}^{(k)}\right)$.
  $\mathbf{P}^{(k+1)} \leftarrow \mathbf{M}^{(k)}\tilde{\mathbf{\Sigma}}^{(k)}\mathbf{N}^{(k)\top}$, where $\tilde{\mathbf{\Sigma}}_{ii}^{(k)} \leftarrow \max\{0, \mathbf{\Sigma}_{ii}^{(k)} - \frac{\beta}{2\rho}\}, \forall i$.
  $\mathbf{Q}^{(k+1)} \leftarrow \widehat{\mathbf{Q}}^{(k)} + \mathbf{P}^{(k+1)} - \mathbf{U}\mathcal{W}^{(k)}$.
  $a_{k+1} \leftarrow \frac{1+\sqrt{4a_k^2+1}}{2}$.
  $k \leftarrow k+1$.
until *Convergence*

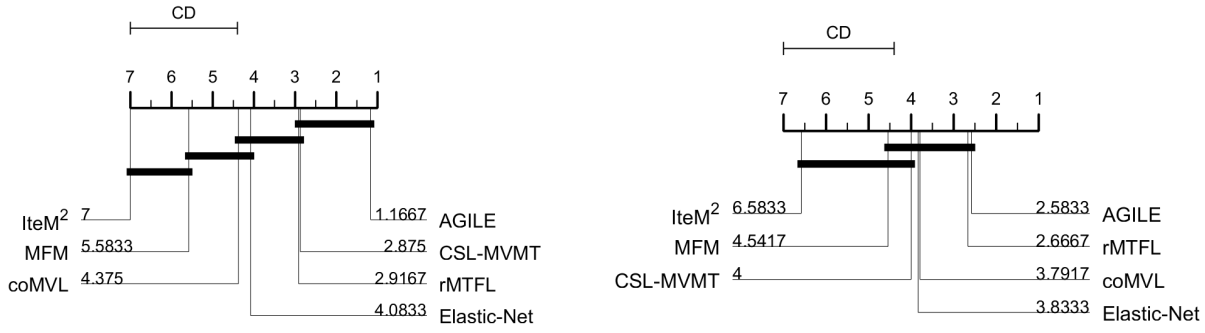
To evaluate the convergence ability of Algorithm 1, we conduct experiment on one synthetic dataset and two real-world datasets, FOX and NUS-Object. In this experiment, we randomly select 30%, 30%, 20% and 20% of total samples as labeled training set, unlabeled training set, validation set and testing set, respectively, and set the parameters of AGILE as $\alpha = \beta = \gamma = 1$. We terminate Algorithm 1 once the relative change of its objective is below 10^{-5}. Figure 1 shows the convergence curves of the objective function value by Algorithm 1. Figure 1 shows that the objective function value converges within 400 iterations, demonstrating the efficiency of the proposed algorithm.



(a) Synthetic data (b) FOX (c) NUS-Object

(d) Synthetic data (e) FOX (f) NUS-Object

Figure 1: Convergence analysis of Algorithm 1 ($\alpha = \beta = \gamma = 1$) on one synthetic dataset and two real-world datasets. The algorithm converged at the 48th, 115th and 361st iteration on the synthetic data, FOX, and NUS-Object, respectively. **The 1st row** shows the original objective value, while **the 2nd row** shows the objective value in the logarithmic scale. In each sub-figure, Funval and Loss denote the objective value and value of loss function, respectively, and R1, R2 and R3 denote the values of three regularization terms.



(a) AUC (b) Accuracy

Figure 2: CD diagrams (0.05 significance level) of seven comparing methods in two evaluation metrics. The performance of two methods is regarded as significantly different if their average ranks differ by at least the Critical Difference (CD).

Data preparation for Section: Experiments

To compare the performances of multi-task multi-view learning methods, we conduct experiments on the following four real-world dual-heterogeneous datasets.

- **FOX**: The datasetrefers to the FOX web news classification [Qian and Zhai, 2014] with 4 categories (*tasks*): health, sports, science/technology and travel, where each article (*instance*) has 2 *views*: text view and image view. Titles, abstracts and text body contents in one article (*instance*) are extracted as the text view, while the associated images are processed as the image view, consisting of seven groups of color features. The text features are stored in ℓ_2-normalized TFIDF vector representation, and the image features concatenate 7 groups of color features.
- **Mirflickr**: The Mirflickr datasetcollects 25,000 Flickr images from Mar. 2007 to Jun. 2008 for image annotation [Huiskes and Lew, 2008]. Here, we treat the 15 relevant class labels as the ground truth categories (*tasks*) of images, such as clouds, sea, flower, dog, car and people. Each image (*instance*) is represented by two types (*views*) of features: image edge histogram and image homogeneous texture.
- **NUS-Object**: This datasetis an object image dataset extracted from the NUS-WIDE dataset [Chua *et al.*, July 8 10 2009] for web image annotation and retrieval. Images (*instances*) are annotated by 31 class labels (*tasks*), like book, car, computer, flower, horse, train, plane, etc, and each image is represented by 5 types (*views*) of low-level features, including 64-D color histogram, 144-D color correlation, 73-D edge direction histogram, 128-D wavelet texture and 225-D block-wise color moments.
- **NUS-Scene**: This datasetis a scene image dataset of the NUS-WIDE dataset [Chua *et al.*, July 8 10 2009] for web image annotation and retrieval. Images (*instances*) can be associated with 33 class labels (*tasks*), such as beach, building, airport, forest, moon, sky and road. Similar with NUS-Object, each image in the NUS-Scene dataset is represented by 5 types (*views*) of low-level features: 64-D color histogram, 144-D color correlation, 73-D edge direction histogram, 128-D wavelet texture and 225-D block-wise color moments.

As a preprocessing for the above datasets, we filter out the text features represented by ℓ_2-normalized TFIDF vector representation, whose frequency is less than 1%. In addition, we eliminate the features with constant values, and remove the samples with no associated categorization. Moreover, we discard the tasks with a relatively small number of positive instances, for example, in the Mirflickr dataset, 7 tasks are discarded as the percentage of positive instances is less than 5%, leading to a MVMTL problem with 8 tasks.

Statistical test for Section: Experiments

To perform statistical test on experimental results in Table 2 of the main paper, we apply Nemenyi test [Demšar, 2006], which allows to statistically evaluate the performance between every two methods. In Nemenyi test, the performance of two methods is regarded as significantly different if their average ranks differ by at least the critical difference (CD). Fig. 2 shows the CD diagrams for four evaluation metrics at 0.05 significance level. In each subfigure, the CD is given above the axis, where the averaged rank is marked. In Fig. 2, algorithms which are not significantly different are connected by a thick line. In terms of AUC, AGILE achieved statistically comparable performances with CSL-MVMT and rMTFL, and statistically superior performances than Elastic-Net, coMVL, MFM and IteM^2. In Accuracy, AGILE ranked 1st among seven comparing methods, and statistically outperformed IteM^2. On average, the robust multi-task learning method, rMTFL, performed second best. This is probably because the real-world datasets indeed exhibit tasks/views outliers, and two robust methods, AGILE and rMTFL, can successfully cope with this setting.

References

[Chua *et al.*, July 8 10 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009.

[Demšar, 2006] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[Goldstein *et al.*, 2014] Tom Goldstein, Brendan O'Donoghue, Simon Setzer, and Richard Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.

[Huiskes and Lew, 2008] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[Nesterov, 2013] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[Qian and Zhai, 2014] Mingjie Qian and Chengxiang Zhai. Unsupervised feature selection for multi-view clustering on text-image web news data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1963–1966. ACM, 2014.