JMLR: Workshop and Conference Proceedings 45:407–422, 2015 ACML 2015

# One-Pass Multi-View Learning

**Yue Zhu** ZHUY@LAMDA.NJU.EDU.CN
**Wei Gao** GAOW@LAMDA.NJU.EDU.CN
**Zhi-Hua Zhou** ZHOUZH@LAMDA.NJU.EDU.CN
*National Key Laboratory for Novel Software Technology, Nanjing University*
*Collaborative Innovation Center of Novel Software Technology and Industrialization*
*Nanjing 210023, China*

**Editor:** Geoffrey Holmes and Tie-Yan Liu

## Abstract

Multi-view learning has been an important learning paradigm where data come from multiple channels or appear in multiple modalities. Many approaches have been developed in this field, and have achieved better performance than single-view ones. Those approaches, however, always work on small-size datasets with low dimensionality, owing to their high computational cost. In recent years, it has been witnessed that many applications involve large-scale multi-view data, e.g., hundreds of hours of video (including visual, audio and text views) is uploaded to YouTube every minute, bringing a big challenge to previous multi-view algorithms. This work concentrates on the large-scale multi-view learning for classification and proposes the One-Pass Multi-View (OPMV) framework which goes through the training data only once without storing the entire training examples. This approach jointly optimizes the composite objective functions with consistency linear constraints for different views. We verify, both theoretically and empirically, the effectiveness of the proposed algorithm.

**Keywords:** Classification; multi-view learning; one-pass; variable linear equality constraints.

## 1. Introduction

Nowadays, a tremendous quantity of data is continuously generated from various views. For example, hundreds of hours of video is uploaded to YouTube every minute, which appears in multiple modalities or views, namely visual, audio and text views; a large number of bilingual news are reported every day, with the description in each language as a view; numerous academic papers are published with text content and citation links, which can also be regarded as multiple views.

Multi-view learning has been an important learning paradigm to handle such tasks with different views (represented as different feature vectors), and many approaches (Blum and Mitchell, 1998; Guo, 2013; Guo and Xiao, 2012; Li et al., 2014; White et al., 2012) have been proposed. For example, in web-page classification, co-training (Blum and Mitchell, 1998) achieves a better performance than single-view approaches by combinative label propagation over the text-content view and the page-link view (Wang and Zhou, 2010); in cross language text categorization, SCMV (Guo and Xiao, 2012) outperforms single-view approaches, by taking advantage of the common latent subspace on different language view.

The exploitation of the cross view latent relationship always accompanies with high computational cost, and most previous approaches work on small-size datasets of low dimensionality, e.g., training examples are usually fewer than 5,000 and the number of features is no more than 1,000. However, in recent years, it has been witnessed that many real applications involve large scale multi-view data, video, webpages, bilingual news, to name but a few, where a large volume of data comes in a short time period, making it infeasible to store the entire dataset in memory before an optimization procedure is applied. Therefore, it is a challenge for previous approaches to tackle the tasks of the large-scale multi-view data. To the best of our knowledge, this is the first work to explore such large-scale multi-view learning problems.

In this paper, we propose the one-pass multi-view framework for large-scale multi-view classifications. We address this problem by jointly optimizing the composite objective functions for different views, where the consistency constraints are expressed with linear equalities for different views. This framework can be viewed as a generalization of online ADMM optimization, and the main difference is that the traditional online ADMM considers the invariable linear constraints, whereas we have to consider the variable constraints according to the pass of training data, so as to keep the classifiers' consistency. We present a regret bound under such setting. Besides, we conduct extensive empirical studies on 27 datasets to show the effectiveness and efficiency of our approach. Details of the experiments are exhibited in Section 6.

## 2. Related Work

Multi-view learning has been an important learning paradigm during the past decade. Blum and Mitchell (Blum and Mitchell, 1998) introduced the famous co-training, whose basic idea is to train a classifier on each view in an alternative manner. Theoretical analysis shows that co-training succeeds if two sufficient and redundant views are conditionally independent to each other. Then this condition was relaxed by $\epsilon$-expansion (Balcan et al., 2005). Further, Wang and Zhou (Wang and Zhou, 2010) presented a sufficient and necessary condition for co-training. Besides, many variants have been developed for multi-view learning (Nigam and Ghani, 2000; Wang and Zhou, 2007, 2010).

Another popular paradigm for multi-view learning is to capture the cross-view relationship by a common subspace. The basic assumption is that different views of the identical example should be close to each other after mapping into a common latent subspace (Chaudhuri et al., 2009; Guo, 2013; Guo and Xiao, 2012; White et al., 2012; Li et al., 2014). Therefore, the cross-view relationships are explored to deal with various multi-view tasks, and improve the performance in practice. Xu et al. (Xu et al., 2013) gave an extensive review on multi-view learning. It is noteworthy that most previous multi-view approaches work only on small-size datasets with low dimensionality, which makes it difficult to handle large-scale and high-dimensional multi-view tasks.

Online learning has been an efficient strategy to build large-scale learning systems, whose study could be traced back to the 1950's of Perceptron algorithm (Rosenblatt, 1958). It has attracted much attention during the past years (Cesa-Bianchi and Lugosi, 2006; Hazan et al., 2007). Many first-order optimization approaches have been complemented in an online style. For example, the online composite objective mirror descent (COMID) (Duchi

et al., 2010) can be viewed as an extension of mirror descent (Beck and Teboulle, 2003), while online regularized dual averaging (RDA) (Xiao, 2010) is generalized from the dual averaging technique (Nesterov, 2009). All those methods work on the single view, and cannot be directly applied to the multi-view learning exploring the cross view relationship.

Another relevant approach is the Alternating Direction Method of Multipliers (ADMM), first introduced by Gabay and Mercier (Gabay and Mercier, 1976). In practice, ADMM shows many excellent properties such as easy applicability (Boyd et al., 2011; Yogatama and Smith, 2014), convenient distributed implementation (Boyd et al., 2011; Zhang and Kwok, 2014), good performance, etc. Wang and Banerjee (Wang and Banerjee, 2012) presented the first online ADMM, and some variants have been presented in (Ouyang et al., 2013; Suzuki, 2013; Zhong and Kwok, 2014). All those approaches focus on the optimization under invariable linear constraints.

## 3. Preliminaries

In multi-view learning, each instance $\mathbf{x}$ is described with several different disjoint spaces of features. Without loss of generality, we focus on a two-view setting for the sake of simplicity in this work. Specifically, let $\mathcal{X} = \mathcal{X}^1 \times \mathcal{X}^2$ be the instance space where $\mathcal{X}^1$ and $\mathcal{X}^2$ are two view spaces, and $\mathcal{Y} = \{+1, -1\}$ denotes the label space. Suppose that $\mathcal{D}$ is an unknown (underlying) distribution over $\mathcal{X} \times \mathcal{Y}$, and what we observe is a training sample $S_n = \{(\mathbf{x}_1^1, \mathbf{x}_1^2; y_1), (\mathbf{x}_2^1, \mathbf{x}_2^2; y_2), \ldots, (\mathbf{x}_n^1, \mathbf{x}_n^2; y_n)\}$ where each example is drawn independently and identically (i.i.d.) from the distribution $\mathcal{D}$.

Let $\mathcal{H}^1$ and $\mathcal{H}^2$ be the function space for each view, respectively. For notational simplicity, we denote by $[n] = \{1, 2, \ldots, n\}$ for integer $n > 0$. In this work, all vectors are assumed to be in a finite dimensional inner product space under the inner product $\langle \cdot, \cdot \rangle$. For two vectors $\mathbf{u}$ and $\mathbf{v}$ of identical size, let $\mathbf{u} \otimes \mathbf{v}$ denote their outer product matrix, and let the symbol $^\top$ denote the transpose operation on vectors and matrices.

Given the training sample $S_n$, the goal for multi-view learning is to select two functions $h^1 \in \mathcal{H}^1$ and $h^2 \in \mathcal{H}^2$ so as to minimize the empirical 0/1 loss as follows:

$$\min_{h^1 \in \mathcal{H}^1, h^2 \in \mathcal{H}^2} \sum_{i=1}^{n} I[h^1(\mathbf{x}_i^1) \neq y_i] + I[h^2(\mathbf{x}_i^2) \neq y_i]$$

under the constraints

$$h^1(\mathbf{x}_i^1) = h^2(\mathbf{x}_i^2) \quad \text{for} \quad i \in [n].$$

Here $I[\cdot]$ denotes the indicator function which returns 1 if the argument is true; and 0 otherwise. The constraints $h^1(\mathbf{x}_i^1) = h^2(\mathbf{x}_i^2)$ are also called *consistency constraints*.

Note that the above formulation is a general formulation for multi-view classification, in which the indicator function (i.e., 0/1 loss) is non-convex and discontinuous, thus the direct optimization often leads to NP-hard problems. In practice, we consider some surrogate losses $\ell$ (such as hinge loss, exponential loss) that can be optimized efficiently. For simplicity, we study the linear function space, i.e.,

$$\mathcal{H}^1 = \{\mathbf{w}^1 \colon \|\mathbf{w}^1\| \leq B\} \quad \text{and} \quad \mathcal{H}^2 = \{\mathbf{w}^2 \colon \|\mathbf{w}^2\| \leq B\},$$

and our approach can be generalized to non-linear classifiers. Therefore, the optimization problem can be rewritten as

$$\begin{aligned} \min_{\mathbf{w}^1\in\mathcal{H}^1,\mathbf{w}^2\in\mathcal{H}^2} & \sum_{i=1}^{n} \ell(\langle \mathbf{w}^1, \mathbf{x}_i^1\rangle, y_i) + \ell(\langle \mathbf{w}^2, \mathbf{x}_i^2\rangle, y_i) \\ \text{s.t.} \quad & \text{sign}(\langle \mathbf{w}^1, \mathbf{x}_i^1\rangle) = \text{sign}(\langle \mathbf{w}^2, \mathbf{x}_i^2\rangle) \quad \text{for} \quad i \in [n], \end{aligned} \tag{1}$$

where $\text{sign}(\langle \mathbf{w}^1, \mathbf{x}_i^1\rangle)$ and $\text{sign}(\langle \mathbf{w}^2, \mathbf{x}_i^2\rangle)$ denote the predicted labels in different views, respectively.

## 4. One-Pass Multi-view Framework

Due to the sign(.) function in the consistency constraints

$$\text{sign}(\langle \mathbf{w}^1, \mathbf{x}_i^1\rangle) = \text{sign}(\langle \mathbf{w}^2, \mathbf{x}_i^2\rangle) \text{ for } i \in [n],$$

it is difficult to design an efficient algorithm to optimize the formulation given by Eqn. (1). For computation, we take the consistency constraints as

$$\langle \mathbf{w}^1, \mathbf{x}_i^1\rangle = \langle \mathbf{w}^2, \mathbf{x}_i^2\rangle \text{ for } i \in [n].$$

Thus, we have

$$\min_{\mathbf{w}^1\in\mathcal{H}^1,\mathbf{w}^2\in\mathcal{H}^2} \sum_{i=1}^{n} \ell(\langle \mathbf{w}^1, \mathbf{x}_i^1\rangle, y_i) + \ell(\langle \mathbf{w}^2, \mathbf{x}_i^2\rangle, y_i) \;\; : \;\; \langle \mathbf{w}^1, \mathbf{x}_i^1\rangle = \langle \mathbf{w}^2, \mathbf{x}_i^2\rangle \quad \text{for} \quad i \in [n]. \tag{2}$$

To efficiently deal with the large scale multi-view tasks, we develop an online learning algorithm to optimize Eqn. (2), which goes through the dataset only once. Let $(\mathbf{x}_t^1, \mathbf{x}_t^2; y_t) \in S_n$ denote the labeled example passes in the iteration $t \in [T]$ , where $T$ denotes the total number of iterations, and it will be equal to the number of training examples in our experiments. The optimization task for the iteration $t$ can be cast as follows:

$$\min_{\mathbf{w}^1\in\mathcal{H}^1,\mathbf{w}^2\in\mathcal{H}^2} \phi_t(\mathbf{w}^1) + \psi_t(\mathbf{w}^2) \;\; : \;\; \langle \mathbf{w}^1, \mathbf{x}_t^1\rangle = \langle \mathbf{w}^2, \mathbf{x}_t^2\rangle \tag{3}$$

where $\phi_t(\mathbf{w}^1) = \ell(\langle \mathbf{w}^1, \mathbf{x}_t^1\rangle, y_t) + \lambda\Omega(\mathbf{w}^1)$, $\psi_t(\mathbf{w}^2) = \ell(\langle \mathbf{w}^2, \mathbf{x}_t^2\rangle, y_t) + \lambda\Omega(\mathbf{w}^2)$, $\lambda > 0$ and $\Omega$ is a regularization.

Notice that the formulation given by Eqn. (3) is a composite objective function with linear equality constraints. It is similar to but different from the online ADMM formulation. The main difference is that the linear coefficients in the constraints are fixed all the time in the online ADMM, whereas those in our consistency constraints vary as training data pass one by one; therefore, our formulation given by Eqn. (3) can be viewed as a generization of the online ADMM framework.

Specifically in this work, we consider the $L_2$ norm regularization, i.e., $\Omega(\mathbf{w}^1) = \|\mathbf{w}^1\|_2^2$ and $\Omega(\mathbf{w}^2) = \|\mathbf{w}^2\|_2^2$. In our experiments, we select hinge loss as our surrogate loss $\ell$. The augmented Lagrangian for Eqn. (3) is given by

$$L_t(\mathbf{w}^1, \mathbf{w}^2, u_t) = \phi_t(\mathbf{w}^1) + \psi_t(\mathbf{w}^2) + u_t(\langle \mathbf{w}^1, \mathbf{x}_t^1\rangle - \langle \mathbf{w}^2, \mathbf{x}_t^2\rangle) + \frac{\rho}{2}\big(\langle \mathbf{w}^1, \mathbf{x}_t^1\rangle - \langle \mathbf{w}^2, \mathbf{x}_t^2\rangle\big)^2$$

where $\mathbf{w}^1$ and $\mathbf{w}^2$ are the primal variables, $u_t$ is the dual variable, $\rho$ ($\rho > 0$) is the penalty parameter. By introducing $\alpha_t = u_t/\rho$, we have [1]

$$L_t(\mathbf{w}^1, \mathbf{w}^2, \alpha_t) = \phi_t(\mathbf{w}^1) + \psi_t(\mathbf{w}^2) - \alpha_t^2 + \frac{\rho}{2}(\langle \mathbf{w}^1, \mathbf{x}_t^1\rangle - \langle \mathbf{w}^2, \mathbf{x}_t^2\rangle + \alpha_t)^2.$$

For computational simplicity, we linearize $\phi_t(\mathbf{w}^1)$ as

$$\phi_t(\mathbf{w}^1) = \phi_t(\mathbf{w}_t^1) + \langle \nabla\phi_t(\mathbf{w}_t^1), \mathbf{w}^1 - \mathbf{w}_t^1\rangle,$$

where $\nabla\phi_t(\mathbf{w}_t^1)$ denotes the gradient of $\phi_t(\mathbf{w}^1)$ at $\mathbf{w}^1 = \mathbf{w}_t^1$. We update $\mathbf{w}_{t+1}^1$ in each iteration $t$ by

$$\mathbf{w}_{t+1}^1 \leftarrow \arg\min_{\mathbf{w}^1} \langle \nabla\phi_t(\mathbf{w}_t^1), \mathbf{w}^1\rangle + \frac{1}{\eta} B_\Psi(\mathbf{w}^1, \mathbf{w}_t^1) + \frac{\rho}{2}\left(\langle \mathbf{w}^1, \mathbf{x}_t^1\rangle - \langle \mathbf{w}_t^2, \mathbf{x}_t^2\rangle + \alpha_t\right)^2, \tag{4}$$

where Bregman divergence $B_\Psi$ is introduced to control the distance between $\mathbf{w}_t^1$ and $\mathbf{w}_{t+1}^1$, and $\eta$ is the learning rate. Throughout this work, we consider the Euclidean distance, i.e., $B_\Psi(\mathbf{w}^1, \mathbf{w}_t^1) = \frac{1}{2}\|\mathbf{w}^1 - \mathbf{w}_t^1\|_2^2$.

To solve the minimization problem in Eqn. (4), $\mathbf{w}_{t+1}^1$ can be calculated as

$$\mathbf{w}_{t+1}^1 \leftarrow \left(\frac{1}{\eta}\mathbb{I} + \rho\mathbf{x}_t^1 \otimes \mathbf{x}_t^1\right)^{-1} \times \left(\frac{1}{\eta}\mathbf{w}_t^1 - \nabla\phi_t(\mathbf{w}_t^1) + \rho(\langle \mathbf{w}_t^2, \mathbf{x}_t^2\rangle - \alpha_t)\mathbf{x}_t^1\right), \tag{5}$$

where $\mathbb{I}$ is an identical matrix of size $d^1 \times d^1$ and $d^1$ is the dimensionality of $\mathcal{X}^1$. This update involves an inverse operation of a $d^1 \times d^1$ matrix, which will take high computational cost and memory space when the dimensionality goes large.

Incorporating the Sherman–Morrison formula (Sherman and Morrison, 1950) into the previous analysis, we update $\mathbf{w}_{t+1}^1$ as [2]

$$\mathbf{w}_{t+1}^1 \leftarrow \eta\mathbf{v}_t^1 - \beta_t^1\mathbf{w}_t^1, \tag{6}$$

where

$$\mathbf{v}_t^1 = -\nabla\phi_t(\mathbf{w}_t^1) + \frac{1}{\eta}\mathbf{w}_t^1 + \rho(\langle \mathbf{w}_t^2, \mathbf{x}_t^2\rangle - \alpha_t)\mathbf{x}_t^1 \tag{7}$$

$$\beta_t^1 = \frac{\rho\eta^2\langle \mathbf{x}_t^1, \mathbf{v}_t^1\rangle}{1 + \rho\eta\langle \mathbf{x}_t^1, \mathbf{x}_t^1\rangle}. \tag{8}$$

From Eqn. (6), it is noteworthy that we do not need to calculate and store the $d^1 \times d^1$ matrix $(\mathbb{I}/\eta + \rho\mathbf{x}_t^1 \otimes \mathbf{x}_t^1)^{-1}$ in Eqn. (5), thus our algorithm can be directly applied to high-dimensional datasets.

In a similar manner, we can update $\mathbf{w}_{t+1}^2$ as

$$\mathbf{w}_{t+1}^2 \leftarrow \eta\mathbf{v}_t^2 - \beta_t^2\mathbf{w}_t^2, \tag{9}$$

1. This is the scaled augmented Lagrangian.
2. In the Eqn. (6), $\mathbf{v}_t^{1(2)}$ and $\beta_t^{1(2)}$ are middle results for the update, where $\mathbf{v}_t^{1(2)}$ is a vector and $\beta_t^{1(2)}$ is a number. Besides, the parameter $\lambda$ is in $\phi$ ($\psi$) in $\mathbf{v}_t^{1(2)}$.

**Algorithm 1** The One-Pass Multi-View (OPMV) Approach

**Input**: The regularization parameter $\lambda > 0$, the penalty parameter $\rho > 0$, the learning rate $\eta > 0$, and the training sample.
**Output**: Two classifiers $\mathbf{w}^1$ and $\mathbf{w}^2$.
**Initialize**: $\mathbf{w}_0^1 = \mathbf{0}$, $\mathbf{w}_0^2 = \mathbf{0}$ and $\alpha_0 = 0$.
**Process**:

1: **for** $t = 0, 1, \cdots, T-1$ **do**
2: Receive a training example $(\mathbf{x}_t^1, \mathbf{x}_t^2, y_t)$;
3: Update the first classifier $\mathbf{w}_{t+1}^1$ by Eqns. (6)-(8);
4: Update the second classifier $\mathbf{w}_{t+1}^2$ by Eqns. (9)-(11);
5: Update the dual variable, by Eqn. (12).
6: **end for**
7: $\mathbf{w}^1 = \mathbf{w}_T^1$ and $\mathbf{w}^2 = \mathbf{w}_T^2$.

with

$$\mathbf{v}_t^2 = -\nabla\psi_t(\mathbf{w}_t^2) + \frac{1}{\eta}\mathbf{w}_t^2 + \rho(\langle \mathbf{w}_{t+1}^1, \mathbf{x}_t^1\rangle + \alpha_t)\mathbf{x}_t^2 \tag{10}$$

$$\beta_t^2 = \frac{\rho\eta^2\langle \mathbf{x}_t^2, \mathbf{v}_t^2\rangle}{1 + \rho\eta\langle \mathbf{x}_t^2, \mathbf{x}_t^2\rangle}. \tag{11}$$

Finally, let the gradient of the scaled augmented Lagrange w.r.t. $\alpha_t$ equal to 0, we derive the update rule for $\alpha_{t+1}$ as

$$\alpha_{t+1} \leftarrow \alpha_t + \langle \mathbf{w}_{t+1}^1, \mathbf{x}_t^1\rangle - \langle \mathbf{w}_{t+1}^2, \mathbf{x}_t^2\rangle. \tag{12}$$

Algorithm 1 highlights the key steps of the proposed One-Pass Multi-View (*OPMV*) algorithm. We initialize $\alpha_0 = 0$, $\mathbf{w}_0^1 = \mathbf{0}$ and $\mathbf{w}_0^2 = \mathbf{0}$, where the sizes of $\mathbf{w}_0^1$ and $\mathbf{w}_0^2$ are $d^1$ (dimensionality of $\mathcal{X}_1$) and $d^2$ (dimensionality of $\mathcal{X}_2$), respectively. We update $\mathbf{w}_{t+1}^1$, $\mathbf{w}_{t+1}^2$ and $\alpha_{t+1}$ in iteration and finally output $\mathbf{w}_T^1$ and $\mathbf{w}_T^2$. In the test stage, we predict the label for a new example $(\mathbf{x}_{new}^1, \mathbf{x}_{new}^2)$ as $y = \text{sign}(\langle \mathbf{w}_T^1, \mathbf{x}_{new}^1\rangle + \langle \mathbf{w}_T^2, \mathbf{x}_{new}^2\rangle)$.

## 5. Theoretical Study

It is necessary to introduce another lemma as follows:

**Lemma 1** *Let $f(\mathbf{w})$ be a convex function. For any scalar $r > 0$ and any vector $\mathbf{u}$, let*

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} f(\mathbf{w}) + r\|\mathbf{w} - \mathbf{u}\|_2^2.$$

*For any subgradient $\mathbf{g} \in \partial f(\mathbf{w}^*)$, we have*

$$\langle \mathbf{g}, \mathbf{w}^* - \mathbf{w}\rangle \leq r(\|\mathbf{w} - \mathbf{u}\|_2^2 - \|\mathbf{w} - \mathbf{w}^*\|_2^2 - \|\mathbf{u} - \mathbf{w}^*\|_2^2).$$

**Proof** For convex function $f(\mathbf{w})$, the optimal solution of $\mathbf{w}^*$ gives

$$\langle \mathbf{g} + 2r(\mathbf{w}^* - \mathbf{u}), \mathbf{w} - \mathbf{w}^* \rangle \geq 0.$$

Combining with

$$2\langle \mathbf{w}^* - \mathbf{u}, \mathbf{w}^* - \mathbf{w} \rangle = \|\mathbf{w} - \mathbf{w}^*\|_2^2 + \|\mathbf{u} - \mathbf{w}^*\|_2^2 - \|\mathbf{w} - \mathbf{u}\|_2^2,$$

we complete the proof. ■

We assume classifiers and subgradients are all bounded for each iteration $t \in [T]$, i.e.,

**Assumption 1** $\|\mathbf{x}_t^1\| \leq B_0$ and $\|\mathbf{x}_t^2\| \leq B_0$,

**Assumption 2** $\|\mathbf{w}_t^i\| \leq B_1$ and $\|\mathbf{w}_*^i\| \leq B_1$ with $i \in \{1, 2\}$,

**Assumption 3** $\|\nabla \phi_t(\mathbf{w}_t^1)\| \leq B_2$ and $\|\nabla \psi_t(\mathbf{w}_t^2)\| \leq B_2$.

We now give our main result for regret bounds as follows:

**Theorem 2** *Let the sequences* $\{\mathbf{w}_t^1, \mathbf{w}_t^2, \alpha_t\}$ *be generated by Algorithm 1. Then, we have*

$$\sum_{t=1}^{T} \phi_t(\mathbf{w}_t^1) + \psi_t(\mathbf{w}_t^2) - \min_{(\mathbf{w}_*^1, \mathbf{w}_*^2) \in \mathcal{R}} \sum_{t=1}^{T} \phi_t(\mathbf{w}_*^1) + \psi_t(\mathbf{w}_*^2) \leq (B_1 + B_2 + 4B_0^2 B_1^2) T^{1/2} + 4B_0 B_1^2 / T,$$

*by setting* $\rho = T^{-3/2}$ *and* $\eta = T^{-1/2}$*, under Assumptions 1-3. Here*

$$\mathcal{R} = \{(\mathbf{w}_*^1, \mathbf{w}_*^2) \colon \mathit{sign}(\langle \mathbf{w}_*^1, \mathbf{x}_t^1 \rangle) = \mathit{sign}(\langle \mathbf{w}_*^2, \mathbf{x}_t^2 \rangle) \text{ for } t \in [T]\}.$$

**Proof** Since $\mathbf{w}_{t+1}^1$ is an optimal solution of Eqn. (4), it holds that, from Lemma 2

$$\begin{aligned} &\big\langle \nabla \phi_t(\mathbf{w}_t^1) + \rho(\langle \mathbf{w}_{t+1}^1, \mathbf{x}_t^1 \rangle - \langle \mathbf{w}_t^2, \mathbf{x}_t^2 \rangle + \alpha_t)\mathbf{x}_t^1, \mathbf{w}_{t+1}^1 - \mathbf{w}_*^1 \big\rangle \\ &\qquad \leq \left( \|\mathbf{w}_*^1 - \mathbf{w}_t^1\|_2^2 - \|\mathbf{w}_*^1 - \mathbf{w}_{t+1}^1\|_2^2 - \|\mathbf{w}_{t+1}^1 - \mathbf{w}_t^1\|_2^2 \right) / 2\eta. \end{aligned}$$

For a convex function $\phi_t$, we have

$$\phi_t(\mathbf{w}_t^1) - \phi_t(\mathbf{w}_*^1) \leq \langle \nabla \phi_t(\mathbf{w}_t^1), \mathbf{w}_t^1 - \mathbf{w}_*^1 \rangle.$$

Combining with the previous two inequalities, we have

$$\begin{aligned} &\rho(\langle \mathbf{w}_{t+1}^1, \mathbf{x}_t^1 \rangle - \langle \mathbf{w}_t^2, \mathbf{x}_t^2 \rangle + \alpha_t)\langle \mathbf{x}_t^1, \mathbf{w}_{t+1}^1 - \mathbf{w}_*^1 \rangle + \phi_t(\mathbf{w}_t^1) - \phi_t(\mathbf{w}_*^1) \\ &\qquad \leq \langle \nabla \phi_t(\mathbf{w}_t^1), \mathbf{w}_t^1 - \mathbf{w}_{t+1}^1 \rangle + \left( \|\mathbf{w}_*^1 - \mathbf{w}_t^1\|_2^2 - \|\mathbf{w}_*^1 - \mathbf{w}_{t+1}^1\|_2^2 - \|\mathbf{w}_{t+1}^1 - \mathbf{w}_t^1\|_2^2 \right) / 2\eta. \end{aligned}$$

From Young's inequality, we have

$$\langle \nabla \phi_t(\mathbf{w}_t^1), \mathbf{w}_t^1 - \mathbf{w}_{t+1}^1 \rangle \leq \eta \|\nabla \phi_t(\mathbf{w}_t^1)\|_2^2 / 2 + \|\mathbf{w}_t^1 - \mathbf{w}_{t+1}^1\|_2^2 / 2\eta.$$

This derives that

$$\rho(\langle \mathbf{w}_{t+1}^1, \mathbf{x}_t^1\rangle - \langle \mathbf{w}_t^2, \mathbf{x}_t^2\rangle + \alpha_t)\langle \mathbf{x}_t^1, \mathbf{w}_{t+1}^1 - \mathbf{w}_*^1\rangle + \phi_t(\mathbf{w}_t^1) - \phi_t(\mathbf{w}_*^1) \\ \leq \eta\|\nabla\phi_t(\mathbf{w}_t^1)\|_2^2/2 + \left(\|\mathbf{w}_*^1 - \mathbf{w}_t^1\|_2^2 - \|\mathbf{w}_*^1 - \mathbf{w}_{t+1}^1\|_2^2\right)/2\eta. \quad (13)$$

In a similar manner, we can prove that

$$\rho(\langle \mathbf{w}_{t+1}^2, \mathbf{x}_t^2\rangle - \langle \mathbf{w}_{t+1}^1, \mathbf{x}_t^1\rangle - \alpha_t)\langle \mathbf{x}_t^2, \mathbf{w}_{t+1}^2 - \mathbf{w}_*^2\rangle + \psi_t(\mathbf{w}_t^2) - \psi_t(\mathbf{w}_*^2) \\ \leq \eta\|\nabla\psi_t(\mathbf{w}_t^2)\|_2^2/2 + \left(\|\mathbf{w}_*^2 - \mathbf{w}_t^2\|_2^2 - \|\mathbf{w}_*^2 - \mathbf{w}_{t+1}^2\|_2^2\right)/2\eta. \quad (14)$$

Summing over $t = 0, 1, \ldots, T-1$, incorporating with Eqns. (13) and (14), we will get

$$\begin{aligned}
&\sum_{t=0}^{T-1} \phi_t(\mathbf{w}_t^1) + \psi_t(\mathbf{w}_t^2) - \phi_t(\mathbf{w}_*^1) - \psi_t(\mathbf{w}_*^2) \\
&\quad \leq \frac{1}{2\eta}\left(\|\mathbf{w}_T^2\| + \|\mathbf{w}_T^1\|\right) + \frac{\eta}{2}\sum_{t=0}^{T-1}\left(\|\nabla\phi_t(\mathbf{w}_t^1)\|_2^2 + \|\nabla\psi_t(\mathbf{w}_t^2)\|_2^2\right) \\
&\quad + \rho\sum_{t=0}^{T-1}\langle \mathbf{w}_t^2 - \mathbf{w}_{t+1}^2\rangle\langle \mathbf{x}_t^1, \mathbf{w}_{t+1}^1 - \mathbf{w}_*^1\rangle + \rho\sum_{t=0}^{T-1}\alpha_{t+1}(\langle \mathbf{x}_t^2, \mathbf{w}_{t+1}^2 - \mathbf{w}_*^2\rangle - \langle \mathbf{x}_t^1, \mathbf{w}_{t+1}^1 - \mathbf{w}_*^1\rangle).
\end{aligned}$$

From Eqn. (12), we have

$$\alpha_t = \sum_{i=1}^{t-1}\langle \mathbf{w}_{i+1}^1, \mathbf{x}_i^1\rangle - \langle \mathbf{w}_{i+1}^2, \mathbf{x}_i^2\rangle,$$

which yields that $|\alpha_t| \leq 2(t-1)B_0B_1$. Therefore, we have

$$\sum_{t=0}^{T-1} \phi_t(\mathbf{w}_t^1) + \psi_t(\mathbf{w}_t^2) - \phi_t(\mathbf{w}_*^1) - \psi_t(\mathbf{w}_*^2) \leq B_1/\eta + \eta T B_2 + \rho(4TB_0B_1^2 + 4T^2B_0^2B_1^2).$$

By setting $\rho = T^{-3/2}$ and $\eta = T^{-1/2}$ we complete the proof as desired. ■

## 6. Experiments

### 6.1. Datasets

We conduct our experiments on 27 real datasets in multi-view learning, including Cora (McCallum et al., 2000), IMDB (Bisson and Grimal, 2012), News Group (Hussain et al., 2010) and Reuter (Amini et al., 2009). Those datasets have been well-investigated by previous researchers. The multi-class datasets Cora, IMDB and Reuter have been transformed into binary ones by randomly partitioning classes into two groups, where each group contains similar number of examples. The detail of the datasets are summarized in Table 1.

Table 1: Detail description of datasets: let $n$ be the number of examples, and $d^1$ and $d^2$ denote the dimensionality of the first and second view, respectively.

| Dataset | $n$ | $d^1$ | $d^2$ | Dataset | $n$ | $d^1$ | $d^2$ | Dataset | $n$ | $d^1$ | $d^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Rt.EN-FR | 18,758 | 21,531 | 24,892 | Rt.GR-FR | 29,953 | 34,279 | 24,892 | Rt.SP-GR | 12,342 | 11,547 | 34,262 |
| Rt.EN-GR | 18,758 | 21,531 | 34,215 | Rt.GR-IT | 29,953 | 34,279 | 15,505 | Rt.SP-IT | 12,342 | 11,547 | 15,500 |
| Rt.EN-IT | 18,758 | 21,531 | 15,506 | Rt.GR-SP | 29,953 | 34,279 | 11,547 | Cora | 2,708 | 2,708 | 1,433 |
| Rt.EN-SP | 18,758 | 21,531 | 11,547 | Rt.IT-EN | 24,039 | 15,506 | 21,517 | IMDB | 617 | 1,878 | 1,398 |
| Rt.FR-EN | 26,648 | 24,893 | 21,531 | Rt.IT-FR | 24,039 | 15,506 | 24,892 | NG.M2 | 500 | 2,000 | 2,000 |
| Rt.FR-GR | 26,648 | 24,893 | 34,287 | Rt.IT-GR | 24,039 | 15,506 | 34,278 | NG.M5 | 500 | 2,000 | 2,000 |
| Rt.FR-IT | 26,648 | 24,893 | 15,503 | Rt.IT-SP | 24,039 | 15,506 | 11,547 | NG.M10 | 500 | 2,000 | 2,000 |
| Rt.FR-SP | 26,648 | 24,893 | 11,547 | Rt.SP-EN | 12,342 | 11,547 | 21,530 | NG.NG1 | 400 | 2,000 | 2,000 |
| Rt.GR-EN | 29,953 | 34,279 | 21,531 | Rt.SP-FR | 12,342 | 11,547 | 24,892 | NG.NG2 | 1,000 | 2,000 | 2,000 |

### 6.2. Compared Approaches

We compare our *OPMV* approach with one single-view approach, four state-of-the-art (batch) multi-view approaches and the batch version of the proposed optimization. It is noteworthy that all compared methods require to store the entire training data in memory and scan the training data several times. In contrast, our approaches scan the data only once without storing of training data. The details of the compared approaches are listed as follows:

- *SV*: We concatenate two views so as to form a new single view, and then apply SVM for classification.
- *CCAMV*: We firstly use CCA to extract the latent common subspace representation and then perform SVM in the common space for classification.
- *CSLMV*: We firstly perform the convex subspace learning approach (White et al., 2012) to find a common space between two views, then SVM is applied in the common space for classification.
- *NMFMV*: We apply non-negative matrix factorization (NMF) based approach (Li et al., 2014) to extract the latent representation and learn the linear classifier by SVM.
- *SCMV*: We use the *SCMV* approach proposed in (Guo and Xiao, 2012) for comparison, which simultaneously learn the subspace projection and the classifier.
- *ADMMMV*: We implement the batch version optimization for Eqn. (2) by using ADMM.

### 6.3. Experiment Setting

All experiments are performed with Matlab 7 on a node of computational cluster with 12 cores (2.53GHz each). 5-fold cross-validation is executed on training sets to decide the learning rate $\eta \in 2^{[-8:8]}$ and the regularization parameters $\lambda \in 1e[-16:0]$. The penalty parameter $\rho$ is pre-defined as 1 for our *OPMV* approach. For compared approaches, the dimensionality of common subspace $k \in \{5, 10, 20, 40, 80\}$ is also tuned by 5-fold cross validation on the training sets. The performances of all approaches are evaluated by average accuracy over 10 independent runs.

Table 2: Comparison of test accuracies (mean ± std.) on datasets Cora, IMDB and News Group and Reuter. 'N/A' means that no result returns after 8 hours. •(◦) indicates that *OPMV* is significantly better(worse) than the compared method (paired t-tests at 95% significance level).

| Datasets | *OPMV* | *SV* | *CCAMV* | *CSLMV* | *NMFMV* | *SCMV* | *ADMMMV* |
|---|---|---|---|---|---|---|---|
| Cora | .902±.013 | .882±.009• | .880±.020• | .890±.013• | .901±.026 | .860±.003• | .903±.007 |
| IMDB | .602±.003 | .593±.003• | .598±.003 | .587±.010• | .607±.005 | .586±.004• | .618±.003 |
| NG.M2 | .940±.026 | .935±.017 | .880±.061• | .946±.014 | .941±.022 | .890±.040• | .945±.017 |
| NG.M5 | .933±.030 | .936±.014 | .940±.046 | .940±.035 | .911±.044 | .924±.049 | .942±.024 |
| NG.M10 | .877±.038 | .849±.039• | .862±.042• | .866±.032• | .861±.043• | .856±.037• | .871±.028 |
| NG.NG1 | .951±.030 | .943±.028• | .949±.031 | .952±.030 | .932±.026• | .920±.044• | .960±.020 |
| NG.NG2 | .921±.020 | .915±.019 | .919±.035 | .920±.020 | .920±.019• | .910±.024• | .935±.018 |
| Rt.EN-FR | .936±.003 | .926±.007• | N/A | N/A | N/A | N/A | N/A |
| Rt.EN-GR | .933±.004 | .923±.005• | N/A | N/A | N/A | N/A | N/A |
| Rt.EN-IT | .933±.004 | .924±.006• | N/A | N/A | N/A | N/A | N/A |
| Rt.EN-SP | .932±.004 | .924±.004• | N/A | N/A | N/A | N/A | N/A |
| Rt.FR-EN | .905±.004 | .891±.003• | N/A | N/A | N/A | N/A | N/A |
| Rt.FR-GR | .904±.005 | .894±.005• | N/A | N/A | N/A | N/A | N/A |
| Rt.FR-IT | .904±.004 | .891±.003• | N/A | N/A | N/A | N/A | N/A |
| Rt.FR-SP | .903±.004 | .888±.003• | N/A | N/A | N/A | N/A | N/A |
| Rt.GR-EN | .926±.004 | .899±.002• | N/A | N/A | N/A | N/A | N/A |
| Rt.GR-FR | .927±.004 | .899±.005• | N/A | N/A | N/A | N/A | N/A |
| Rt.GR-IT | .923±.004 | .903±.004• | N/A | N/A | N/A | N/A | N/A |
| Rt.GR-SP | .925±.003 | .902±.002• | N/A | N/A | N/A | N/A | N/A |
| Rt.IT-EN | .897±.003 | .877±.006• | N/A | N/A | N/A | N/A | N/A |
| Rt.IT-FR | .898±.003 | .877±.005• | N/A | N/A | N/A | N/A | N/A |
| Rt.IT-GR | .895±.004 | .878±.005• | N/A | N/A | N/A | N/A | N/A |
| Rt.IT-SP | .895±.003 | .874±.005• | N/A | N/A | N/A | N/A | N/A |
| Rt.SP-EN | .953±.004 | .922±.007• | N/A | N/A | N/A | N/A | N/A |
| Rt.SP-FR | .953±.004 | .921±.007• | N/A | N/A | N/A | N/A | N/A |
| Rt.SP-GR | .953±.005 | .925±.010• | N/A | N/A | N/A | N/A | N/A |
| Rt.SP-IT | .952±.003 | .919±.079• | N/A | N/A | N/A | N/A | N/A |

### 6.4. Experimental Results

The comparison results are summarized in Table 2 and the average running time is shown in Figure 1. Because none of batch multi-view approaches can return a result after 8 hours for large datasets Reuter, we randomly select a subset from each Reuter dataset of 3,000 examples with 400 features corresponding to 400 words of highest appearance frequency. For the sampled datasets (SmallRt for short), the comparison results and average running time are shown in Table 3 and last column in Figure 1, respectively. As can be seen, experimental results clearly verify the effectiveness and efficiency of our proposed *OPMV* approach.

First, our *OPMV* approach is superior to the single-view approach *SV* which simply concatenates two views into a single view, since the results show that our approach never loses. This also demonstrates that the exploitation of cross-view relationship, i.e., the consistency equality constraints for different view in our *OPMV* approach, is beneficial to improving the performance for multi-view learning. In addition, our *OPMV* approach is comparable to the batch *ADMMMV* algorithm which requires to store the entire training datasets.

Second, the proposed *OPMV* approach outperforms *CCAMV*, *CSLMV*, *NMFMV* and *SCMV* on small-size datasets, and those methods do not return any result on the large dataset Reuter after 8 hours. Even for sampled and dimension-reductional dataset Reuter, our *OPMV* ap-

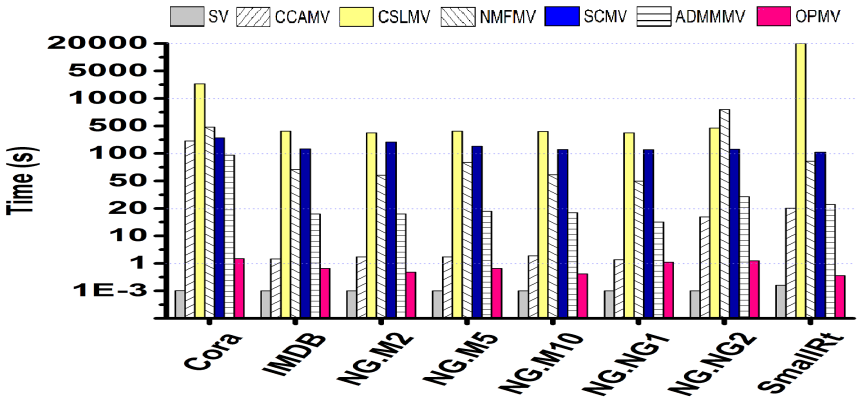


Figure 1: Time comparison on Cora, IMDB, NewsGroup and SampledReuter(SmallRt).

proach also achieves better performance as shown in Table 3. The possible reasons are that i) most previous methods learn the subspace under an un-supervised procedure; ii) non-convex optimization tends to converge to local minimum solution; and iii) previous methods are very sensitive to parameters.

Figure 1 shows that the proposed *OPMV* approach is much faster than state-of-the-art batch multi-view approaches. For example, on the dataset Cora of 2,708 examples with dimensionality of 2,708 and 1,433 for two views respectively, our *OPMV* is 50 times faster than *ADMMMV*, 110 times faster than *CCAMV*, 130 times faster than *SCMV*, 180 times faster than *NMFMV* and even 1,000 times faster than *CSLMV*. Besides, our *OPMV* takes less than 4 minutes on the dataset Reuter, which contains over 10,000 examples with at least 10,000 dimensions for each view, whereas previous multi-view approaches do not return any result even after 8 hours. Even for the sampled Reuter with size 3,000 and dimensionality 400, our *OPMV* approach is more efficient in contrast to compared batch multi-view approaches.

## 7. Convergence and Parameter Study

This section studies the parameter influence on our proposed approach. Due to the limited space, only 4 datasets are sampled for exhibition, but a similar phenomenon can be observed in the other datasets. Figure 2 reveals the iteration influence on *OPMV*, which suggests the trend of convergence with the increase of the number of iteration. Particularly, in the first 200 iterations, it converges very fast, where the curve is of a large slope.

There are three parameters in all involved in *OPMV* approaches, including regularization parameter $\lambda$, learning rate parameter $\eta$ and penalty parameter $\rho$. We study respectively the parameter influence to the learning performance as follows [3].

Figure 3 indicates the influence of $\lambda$ ranging from $1e{-}16$ to 1, whose x axis is $-\lg\lambda$ and the result is obtained with fixed $\eta(0.25)$ and $\rho(8)$. Notice that the datasets are of high dimensionality, thus both $\mathbf{w}^1$ and $\mathbf{w}^2$ are large vectors. In this circumstance, large

3. Though the penalty parameter $\rho$ is pre-defined in the experiments, we still try to make a study on the influence of this parameter.

Table 3: Comparison of test accuracies (mean ± std.) on sampled datasets Reuter, which contains 3,000 examples with dimensionality of 400. 'N/A' means that no result returns after 8 hours. •(○) indicates that *OPMV* is significantly better(worse) than the compared method (paired t-tests at 95% significance level).

| Datasets | *OPMV* | *SV* | *CCAMV* | *CSLMV* | *NMFMV* | *SCMV* | *ADMMMV* |
|---|---|---|---|---|---|---|---|
| Rt.EN-FR | .876±.005 | .849±.005• | .865±.008• | N/A | .854±.010• | .841±.013• | .882±.005 |
| Rt.EN-GR | .881±.005 | .852±.008• | .849±.017• | N/A | .852±.008• | .868±.017• | .889±.006 |
| Rt.EN-IT | .872±.011 | .852±.003• | .865±.021• | N/A | .860±.005• | .867±.009• | .885±.007 |
| Rt.EN-SP | .881±.002 | .852±.004• | .874±.035 | N/A | .868±.007 | .861±.001• | .884±.003 |
| Rt.FR-EN | .842±.009 | .791±.005• | .800±.014• | N/A | .810±.010• | .796±.003• | .840±.003 |
| Rt.FR-GR | .830±.008 | .790±.005• | .795±.017• | N/A | .792±.011• | .797±.008• | .840±.004 |
| Rt.FR-IT | .836±.004 | .789±.017• | .795±.011• | N/A | .828±.003• | .794±.015• | .844±.008 |
| Rt.FR-SP | .833±.009 | .789±.002• | .801±.019• | N/A | .827±.002 | .807±.017• | .845±.005 |
| Rt.GR-EN | .882±.001 | .820±.010• | .820±.010• | N/A | .865±.004• | .863±.013• | .883±.002 |
| Rt.GR-FR | .878±.004 | .819±.005• | .800±.022• | N/A | .850±.015• | .834±.012• | .885±.004 |
| Rt.GR-IT | .880±.004 | .820±.011• | .809±.017• | N/A | .866±.004• | .856±.005• | .887±.005 |
| Rt.GR-SP | .878±.001 | .823±.003• | .810±.017• | N/A | .868±.013• | .830±.016• | .889±.001○ |
| Rt.IT-EN | .831±.004 | .791±.009• | .800±.015• | N/A | .800±.009• | .794±.010• | .839±.003○ |
| Rt.IT-FR | .830±.006 | .792±.004• | .795±.020• | N/A | .800±.009• | .795±.006• | .827±.001 |
| Rt.IT-GR | .833±.003 | .783±.006• | .790±.020• | N/A | .809±.008• | .797±.007• | .838±.004 |
| Rt.IT-SP | .830±.003 | .793±.005• | .795±.017• | N/A | .802±.009• | .800±.005• | .834±.002 |
| Rt.SP-EN | .917±.003 | .883±.011• | .887±.009• | N/A | .900±.010• | .895±.007• | .917±.005 |
| Rt.SP-FR | .915±.004 | .883±.002• | .879±.014• | N/A | .901±.016• | .897±.011• | .917±.002 |
| Rt.SP-GR | .910±.001 | .881±.012• | .891±.015• | N/A | .909±.005 | .883±.011• | .923±.002○ |
| Rt.SP-IT | .917±.004 | .880±.011• | .906±.017• | N/A | .899±.007• | .906±.015• | .923±.005 |

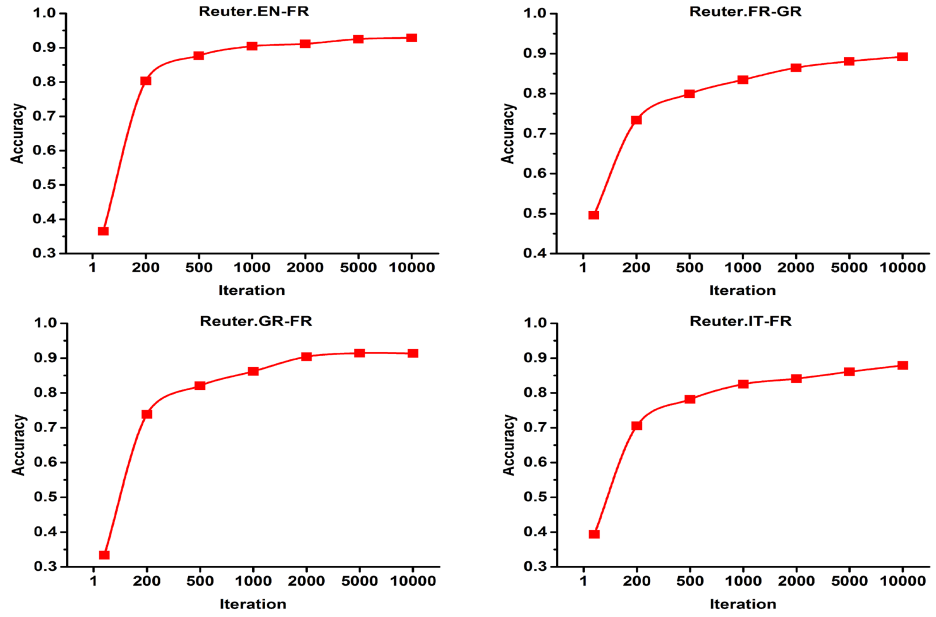


Figure 2: Influence of iterations.

regularization parameter $\lambda$ may dominate the objective function, which may lead to a worse performance. It can be observed in the Figure 3 that, at the beginning, with the decrease of $\lambda$, the accuracy goes up, and when $\lambda$ is less than $1e-4$, *OPMV* is not sensitive to it.

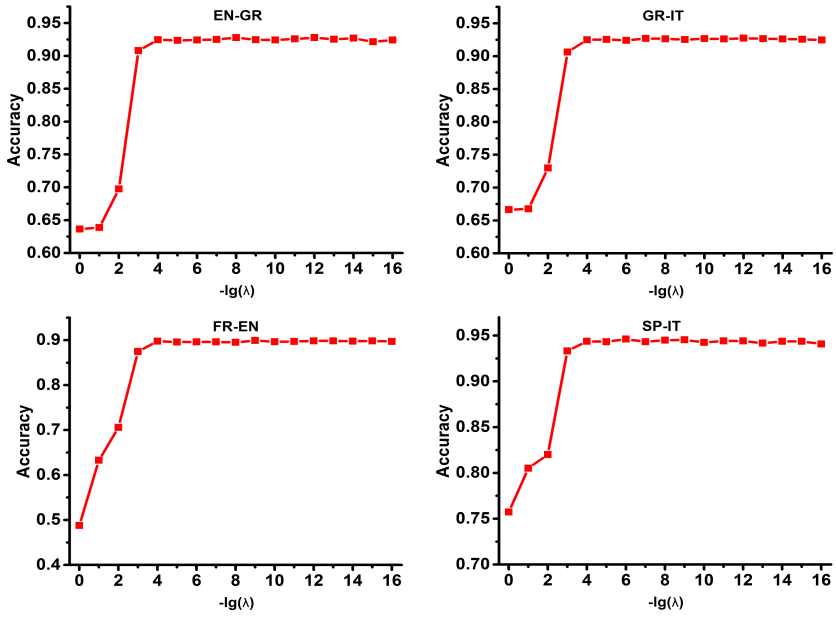


Figure 3: Influence of regularization parameter $\lambda$.

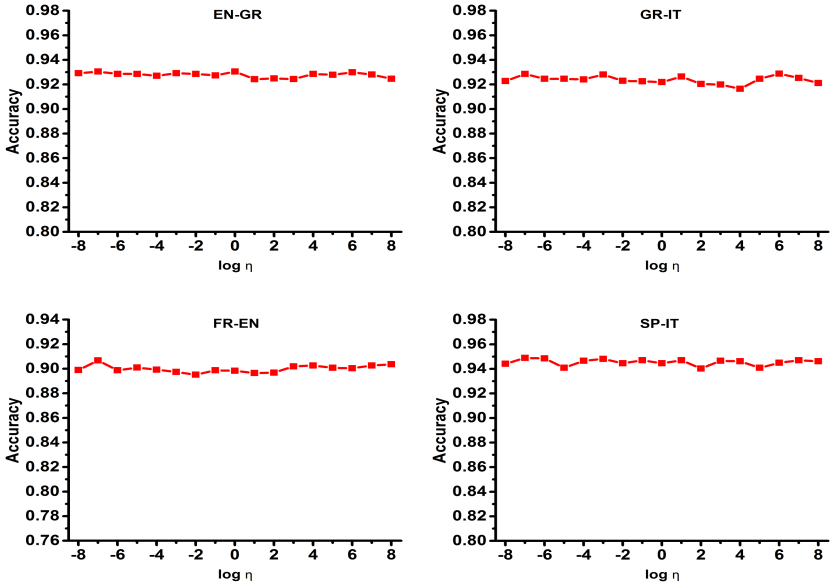


Figure 4: Influence of learning rate parameter $\eta$.

Figure 4, whose x axis is $\log_2 \eta$, suggests the influence of learning $\eta$ ranging from $2^{-8}$ to $2^8$, under fixed $\lambda(1e{-}12)$ and $\rho(1)$. As shown in the Figure 4, *OPMV* is not sensitive to the learning rate parameter $\eta$.

Figure 5, whose x axis is $\log_2 \rho$, indicates the influence of $\rho$ from $2^{-8}$ to $2^8$ with fixed $\lambda(1e{-}12)$ and $\rho(1)$. It can be seen that, the accuracy is slightly increasing with $\rho$ increasing

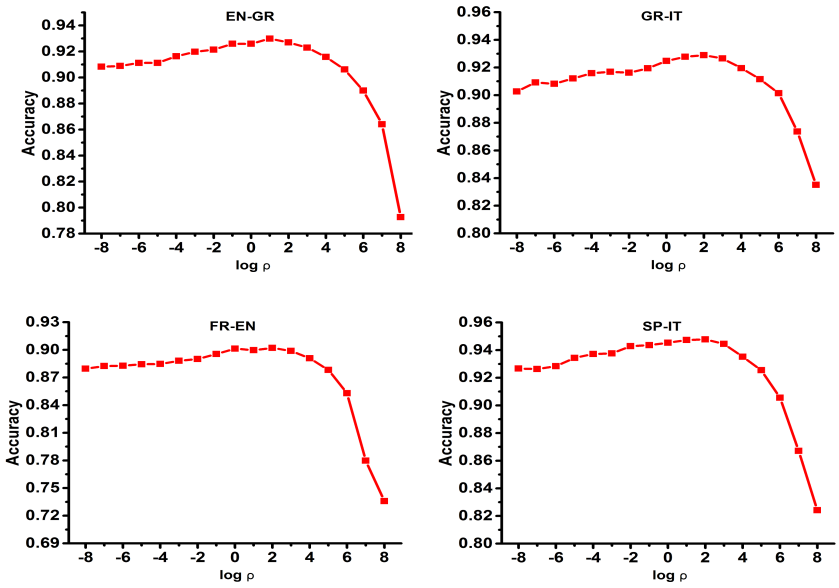


Figure 5: Influence of penalty parameter $\rho$.

from $2^{-8}$ to 4, but after that the accuracy drops. This is mainly due to the fact that larger $\rho$ will encourage the classifier outputs on different views to be more similar, but when $\rho$ goes too large, it may overfit the noisy data where classifier consistency constraint does not hold and leads to a degenerated performance.

## 8. Conclusions

Multi-view learning has been an important learning paradigm over the last decade. Many multi-view approaches have been developed, and almost all of them work on small-size datasets with low dimensionality. In this paper, we propose the one-pass multi-view approach *OPMV*, which scans each example only once without storing the entire training data. To the best of our knowledge, this is the first work to study on the large-scale multi-view learning. We address this problem by jointly optimizing composite objective functions for different views, and the consistency constraints are expressed with linear equalities for different training examples. This framework can be viewed as a generalization of ADMM optimization, and the main difference is that traditional ADMM only considers equality constraints with fixed linear coefficients, whereas our constraints vary as training data pass one by one, so as to keep classifier consistency between multiple views. Theoretically, we present a regret bound for such setting. Moreover, extensive experiments show that the proposed *OPMV* approach achieves better or comparable performance in contrast to state-of-the-art (batch) multi-view approaches, and is more efficient, e.g., hundreds of times faster than state-of-the-art batch multi-view approaches. In the future, it is interesting to study one-pass partial view problem where examples may pass with a missing view in many real applications.

## 9. Acknowledgment

This research was supported by the National Science Foundation of China (61273301, 61305067).

## References

M. Amini, N. Usunier, and C. Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. In *Advances in Neural Information Processing Systems 22*, pages 28–36. 2009.

M.-F. Balcan, A. Blum, and K. Yang. Co-training and expansion: Towards bridging theory and practice. In *Advances in Neural Information Processing Systems 18*, pages 89–96. 2005.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Gilles Bisson and Clément Grimal. Co-clustering of multi-view datasets: A parallelizable approach. In *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM'12)*, pages 828–833, Brussels, Belgium, 2012.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Conference on Computational Learning Theory (COLT' 98)*, pages 92–100, Madison, WI, 1998.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

K. Chaudhuri, S.-M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th International Conference on Machine Learning (ICML' 09)*, pages 129–136, Montreal, Canada, 2009.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT' 10)*, Haifa, Israel, 2010.

D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers and Mathematics with Applications*, 2(1):17–40, 1976.

Y.-H. Guo. Convex subspace representation learning from multi-view data. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI' 13)*, pages 387–393, Bellevue, WA, 2013.

Y.-H. Guo and M. Xiao. Cross language text classification via subspace co-regularized multi-view learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML' 12)*, pages 1615–1622, Edinburgh, Scotland, 2012.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

S. Hussain, C. Grimal, and G. Bisson. An improved co-similarity measure for document clustering. In *Proceedings of 9th International Conference on Machine Learning and Applications (ICMLA' 10)*, pages 190–197, Washinton DC, 2010.

S.-Y. Li, Y. Jiang, and Z.-H. Zhou. Partial multi-view clustering. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI' 14)*, pages 1968–1974, Quebec City, Canada, 2014.

A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming: Series A and B*, 120(1):221–259, 2009.

K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the 9th international conference on Information and Knowledge Management*, pages 86–93, Washington, D.C, 2000.

H. Ouyang, N. He, L. Tran, and A. Gray. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning (ICML' 13)*, pages 80–88, Atlanta, GA, 2013.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

J. Sherman and W. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics*, 21(1):124–127, 1950.

T. Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of the 30th International Conference on Machine Learning (ICML' 13)*, pages 392–400, Atlanta, GA, 2013.

H. Wang and A. Banerjee. Online alternating direction method. In *Proceedings of the 29th International Conference on Machine Learning (ICML' 12)*, pages 1119–1126, Edinburgh, Scotland, 2012.

W. Wang and Z.-H. Zhou. Analyzing co-training style algorithms. In *Proceedings of the 18th European Conference on Machine Learning (ECML' 07)*, pages 454–465, Warsaw, Poland, 2007.

W. Wang and Z.-H. Zhou. A new analysis of co-training. In *Proceedings of the 27th International Conference on Machine Learning (ICML' 10)*, pages 1135–1142, Haifa, Israel, 2010.

M. White, Y. Yu, X. Zhang, and D. Schuurmans. Convex multi-view subspace learning. In *Advances in Neural Information Processing Systems 25*, pages 1673–1681. 2012.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

D. Yogatama and N. Smith. Making the most of bag of words: Sentence regularization with alternating direction method of multipliers. In *Proceedings of the 31st International Conference on Machine Learning (ICML' 14)*, pages 656–664, Beijing, China, 2014.

R. Zhang and J. Kwok. Asynchronous distributed admm for consensus optimization. In *Proceedings of the 31st International Conference on Machine Learning (ICML' 14)*, pages 1701–1709, Beijing, China, 2014.

W. Zhong and J. Kwok. Fast stochastic alternating direction method of multipliers. In *Proceedings of The 31st International Conference on Machine Learning (ICML' 14)*, pages 46–54, Beijing, China, 2014.