Stability-Based PAC-Bayes Analysis for Multi-view Learning Algorithms

Shiliang Sun[*,a], Mengran Yu[a], John Shawe-Taylor[b], Liang Mao[a]

[a] *School of Computer Science and Technology, East China Normal University,*
3663 North Zhongshan Road, Shanghai 200062, P.R. China
[b] *Department of Computer Science, University College London,*
Gower Street, London WC1E 6BT, United Kingdom

Abstract

Multi-view learning exploits structural constraints among multiple views to effectively learn from data. Although it has made great methodological achievements in recent years, the current generalization theory is still insufficient to prove the merit of multi-view learning. This paper blends stability into multi-view PAC-Bayes analysis to explore the generalization performance and effectiveness of multi-view learning algorithms. We propose a novel view-consistency regularization to produce an informative prior that helps to obtain a stability-based multi-view bound. Furthermore, we derive an upper bound on the stability coefficient that is involved in the PAC-Bayes bound of multi-view regularization algorithms for the purpose of computation, taking the multi-view support vector machine as an example. Experiments provide strong evidence on the advantageous generalization bounds of multi-view learning over single-view learning. We also explore strengths and weaknesses of the proposed stability-based bound compared with previous

[*] Corresponding author. Tel.: +86-21-62233507; fax: +86-21-62232584.
Email address: `slsun@cs.ecnu.edu.cn` (Shiliang Sun)

non-stability multi-view bounds experimentally.

Key words: Multi-view Learning, PAC-Bayes Analysis, Stability, Generalization

1. Introduction

Data in the real world may come from multiple views. For example, in the web page classification task, a web page is described by two views where one is the content in the web page itself and the other is the anchor text of any web page linking to it. In the medical diagnosis field, diagnostic decisions are usually taken by various results of examinations (3D labeled magnetic resonance image, ultrasonic image, etc). Multi-view learning was proposed to learn from this type of data and leverage peculiarities owned by these views [1]. Intuitively, multi-view learning performs better than single-view learning since single-view learning neglects the consistency between different views by simply concatenating them together into a single long view.

In recent years, multi-view learning has made great progress in both practice and algorithms. On one hand, there are numerous practical applications related to multi-view learning, such as health [2], biology [3], multimedia [4] and ecology [5]. On the other hand, a large number of multi-view learning algorithms in various settings have been constructed, including multi-view transfer learning [6], multi-task multi-view learning [7], multi-view semi-supervised learning [8], multiple kernel learning [9], etc. However, theoretical analysis of multi-view learning is still insufficient. For instance, there is a lack of explicit theoretical evidence that multi-view learning is often superior to single-view learning.

In statistical learning theory, PAC-Bayes bounds provide state-of-the-art predictions of the generalization performance compared with bounds employing Rademacher complexity and VC dimension [10]. PAC-Bayes theory gives an upper bound on the true generalization error by the empirical error and the Kullback-Leibler (KL) divergence between prior and posterior distributions of a learning algorithm. It has great flexibility because it allows selecting prior distributions of learners that are not necessarily correct [11, 12, 13]. There are two ways to investigate the PAC-Bayes theory. Data-dependent priors usually split the data into two parts, one of which is used to learn a meaningful prior because the prior cannot depend on the data used in the empirical risk term of the PAC-Bayes bound [14, 15, 16]. Distribution-dependent priors could be well-designed for directly computing the KL term [17, 18, 19].

PAC-Bayes analysis has recently surged in popularity, but most studies concentrate on the generalization performance of single-view learning algorithms. The latest work derived relatively tighter PAC-Bayes bounds from different aspects. Dziugaite and Roy [20] reduced the PAC-Bayes bound from the perspective of differential privacy with the strategy of data-dependent priors, while Rivasplata et al. [21] shrank the bound from the stability of learning algorithms with the strategy of distribution-dependent priors. A learning algorithm is stable if slightly changed training sets yield similar solutions [22]. Previous experiments demonstrated the feasibility of blending stability into PAC-Bayes analysis. The more restrictive a stability criterion is, the tighter the generalization bound will be [23, 24].

The first PAC-Bayes analysis of multi-view learning algorithms was conducted by Sun et al. [25], who obtained multiple bounds by considering different prior distributions. Experiments on the multi-view support vector machine (MvSVM)

[26, 27] demonstrated the feasibility and rationality of adopting PAC-Bayes theory to analyze the generalization performance of multi-view algorithms. However, multi-view bounds were lower than single-view bounds only on two of six datasets, which could not provide strong support that multi-view learning is usually theoretically advantageous over single-view learning.

This paper adopts the latest development of shrinking PAC-Bayes bounds with stability to analyze generalization bounds of multi-view learning algorithms. The highlights of this paper are summarized as follows.

- We propose a novel view-consistency regularization which produces an informative prior to deduce an effective stability-based PAC-Bayes bound for general multi-view learning algorithms.

- For computing specific bounds of multi-view algorithms in practice, we upper-bound the stability coefficient where the MvSVM is taken as an example.

- Experimental results demonstrate the superiority of multi-view learning over single-view learning in terms of classification errors and stability-based PAC-Bayes bounds on nine datasets.

- We also illustrate advantages and disadvantages of the stability-based multi-view bound compared with previous non-stability multi-view bounds proposed by Sun et al. [25] from two aspects of tightness and ability to support model selection.

To the best of our knowledge, this is the first exploration to analyze the generalization performance of multi-view learning algorithms by blending stability into the PAC-Bayes framework.

The rest of the paper is organized as follows. First, we introduce the multi-view learning problem and the related PAC-Bayes bounds as preliminaries in Section 2. Section 3 derives our main stability-based multi-view PAC-Bayes bound with a new view-consistency regularization and also delivers a stability-based multi-view PAC-Bayes bound without the view-consistency regularization for performance comparisons. The upper bound on the stability coefficient of a multi-view regularization algorithm, taking the MvSVM as an example, is derived in Section 4 for computation purposes. Experimental results are reported in Section 5 and finally conclusions are presented in Section 6.

2. Preliminary Work

We first introduce the multi-view learning problem, and then present relavant PAC-Bayes bounds to pave the way for the proposed stability-based multi-view bound.

2.1. Multi-view Learning

Consider a multi-view learning algorithm which maps a multi-view dataset $\mathcal{X}$ to a function f where $\mathcal{X} = \left\{ \left(\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, y^{(i)} \right) \right\}_{i=1,2,\ldots,m}$ with $\mathbf{x}_v^{(i)} \in \mathcal{R}^{n_v}$ and $v = 1, 2$. For binary classification, the label $y \in \{-1, 1\}$. Multi-view learning aims to learn one function from each view and jointly optimize the two functions to promote the generalization performance.

Consistency and diversity are two fundamental properties for designing multi-view algorithms [28], where the consistency principle intends to maximize the agreement between two views and the diversity principle tries to increase the disagreement of examples misclassified by the two classifiers. Diversity-based algorithms mostly rely on ensemble learning. In the current work, consistency attracts more attention. The representatives of consistency-based algorithms contain co-training styles [29, 30], margin-consistency styles [31, 32] and co-regularization styles [33, 34]. Co-training algorithms mean that multiple learners are trained alternately and they can predict labels for unlabeled data for each other. Margin-consistency algorithms restrict the margin variables from multiple views to be consistent. Co-regularization algorithms regard the disagreement of two views as a regularization term of the objective function.

In this paper, we focus on the theoretical analysis of co-regularization algorithms. A commonly used co-regularization term is the sum of pairwise distances between the outputs from the two views evaluated at the training examples. Specialized to the MvSVM algorithm which is employed in the experiments, the regularization term can be represented as

$$\sum_{i=1}^{m} \left(\mathbf{u}_1^\top \mathbf{x}_1^{(i)} - \mathbf{u}_2^\top \mathbf{x}_2^{(i)} \right)^2, \tag{1}$$

where $\mathbf{u}_1$ and $\mathbf{u}_2$ are weight vectors of view 1 and view 2, respectively. The inputs $\mathbf{x}_1$ and $\mathbf{x}_2$ can be transformed into another feature space by some kernel function.

2.2. Basic Single-view PAC-Bayes Bounds

Consider a single-view learning algorithm, which learns a function f on a training sample $\mathcal{X}$ that includes m instances. Suppose $\mathcal{D}$ is the true distribution of an example $(\mathbf{x}, y)$. For binary classification, the output label $y \in \{-1, 1\}$ and the classifier uses the 0-1 loss function. Assume that Q is the posterior distribution and P is the prior distribution of the classifier.

The true error $e_{\mathcal{D}}$ and empirical error $e_{\mathcal{X}}$ of the classifier are defined as

$$e_{\mathcal{D}} = \Pr_{(\mathbf{x},y)\sim\mathcal{D}}(f(\mathbf{x}) \neq y), \tag{2}$$

$$e_{\mathcal{X}} = \Pr_{(\mathbf{x},y)\sim\mathcal{X}}(f(\mathbf{x}) \neq y) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{I}\left(f\left(\mathbf{x}^{(i)}\right) \neq y^{(i)}\right), \tag{3}$$

where $\mathbf{I}(\cdot)$ is an indicator function. Define the average true error as

$$E_{Q,\mathcal{D}} = \mathbb{E}_{f\sim Q} e_{\mathcal{D}}, \tag{4}$$

and the average empirical error as

$$E_{Q,\mathcal{X}} = \mathbb{E}_{f\sim Q} e_{\mathcal{X}}, \tag{5}$$

in terms of the posterior distribution of the classifier. The following Theorem 1 provides the basic PAC-Bayes bound on $E_{Q,\mathcal{D}}$ for the current binary classification.

Theorem 1 (Langford [13] [Theorem 5.1])**.** *For any data distribution $\mathcal{D}$, for any prior distribution P of a classifier f, for any $\delta \in (0, 1]$:*

$$\Pr_{\mathcal{X}\sim\mathcal{D}^m}\left(\forall Q : \mathit{KL}_{+}(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}}) \leq \frac{\mathit{KL}(Q||P) + \ln(\frac{m+1}{\delta})}{m}\right) \geq 1 - \delta. \tag{6}$$

The KL divergence $\mathrm{KL}(Q||P) = \mathbb{E}_{f\sim Q} \ln \frac{Q}{P}$ measures the difference between the prior distribution P and the posterior distribution Q of the classifier. $\mathrm{KL}_{+}(q||p)$ is defined by $q \ln \frac{q}{p} + (1 - q) \ln \frac{1-q}{1-p}$ for $p > q$ and 0 otherwise.

For an SVM classifier represented by $f_{\mathbf{u}}(\mathbf{x}) = \mathrm{sign}(\mathbf{u}^{\mathsf{T}}\phi(\mathbf{x}))$ where $\phi(\mathbf{x})$ is a projection from the original space to the feature space by some kernel function, the corresponding PAC-Bayes bound is the following where prior and posterior distributions of the classifier are reduced to prior and posterior distributions of its weight vector $\mathbf{u}$.

Corollary 1 (Langford [13] [Corollary 5.4])**.** *Consider the prior is* $P(\mathbf{u}) = \mathcal{N}(0, \mathbf{I})$ *and the posterior is* $Q(\mathbf{u}) = \mathcal{N}(\mu\boldsymbol{\omega}, \mathbf{I})$ *where* μ *and* $\boldsymbol{\omega}$ *separately indicate the norm and the direction of the mean vector. For any data distribution* $\mathcal{D}$*, for any* $\delta \in (0, 1]$*:*

$$\Pr_{\mathcal{X}\sim\mathcal{D}^m}\left(\forall\boldsymbol{\omega}, \mu : KL_+(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}}) \leq \frac{\frac{\mu^2}{2} + \ln(\frac{m+1}{\delta})}{m}\right) \geq 1 - \delta. \tag{7}$$

To bound the average true error $E_{Q,\mathcal{D}}$, it is necessary to calculate the average empirical error $E_{Q,\mathcal{X}}$. Equation (8) shows how to compute it under the posterior distribution $Q(\mathbf{u}) = \mathcal{N}(\mu\boldsymbol{\omega}, \mathbf{I})$ with $||\boldsymbol{\omega}|| = 1$,

$$E_{Q,\mathcal{X}} = \mathbb{E}_{\mathcal{X}}\left[F\left(\mu\frac{y\boldsymbol{\omega}^{\mathsf{T}}\phi(\mathbf{x})}{||\phi(\mathbf{x})||}\right)\right], \tag{8}$$

where $F(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$. The generalization error of the original SVM classifier is upper-bounded by at most twice the average true error on the Gibbs classifier [35].

2.3. Stability-Based Single-view PAC-Bayes Bounds

Define $\mathbf{z} = (\mathbf{x}, y)$ for shortening the notation. For a single-view learning algorithm A, it learns a function f on the training set $\mathcal{X}$ and learns another function $f^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}$ on the changed training set where an example $\mathbf{z}^{(i)}$ is substituted by another example $\mathbf{z}^{(j)}$. The stability coefficient of A is defined as

$$\beta_m = \sup_{i\in[m]}\left|\left|f^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}} - f\right|\right|_\infty, \tag{9}$$

where m is the number of instances. Obviously, the smaller the stability coefficient is, the more stable the learning algorithm is. To blend the stability into the PAC-Bayes analysis for getting a tighter bound, Rivasplata et al. [21] proposed a new Gaussian assumption for the prior distribution of the classifier and gave the stability-based single-view PAC-Bayes bound as follows.

Theorem 2 (Rivasplata et al. [21] [Theorem 2])**.** *Let* A *be a single-view learning algorithm, whose stability coefficient is* β_m*. Consider the prior is* $P = \mathcal{N}(\mathbb{E}[f], \sigma^2\mathbf{I})$ *and the posterior is* $Q = \mathcal{N}(f, \sigma^2\mathbf{I})$*. For any data distribution* $\mathcal{D}$*, for any variance* σ^2*, for any* $\delta \in (0, 1]$*, with probability at least* $1 - \delta$ *the following holds:*

$$KL_+(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}}) \leq \frac{\frac{1}{2\sigma^2} m\beta_m^2 \left(1 + \sqrt{\frac{1}{2}\ln\left(\frac{2}{\delta}\right)}\right)^2 + \ln\left(\frac{m+1}{\frac{\delta}{2}}\right)}{m}. \tag{10}$$

For a specific learning algorithm, it is necessary to calculate the upper bound on the stability coefficient β_m. Taking the SVM as an example, the optimal function is

$$\mathbf{SVM}_\lambda(\mathbf{u}) = \arg\min_{\mathbf{u}} \left(\frac{1}{m}\sum_{i=1}^{m} l\left(f_{\mathbf{u}}\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) + \lambda||\mathbf{u}||^2\right), \tag{11}$$

where l is the hinge loss which is convex. In this context, stability of the SVM reduces to stability of its learned weight vector $\mathbf{u}$ [21]. Then the upper bound on its stability coefficient [22, Theorem 22] is

$$\beta_m \leq \frac{\mathcal{K}^2}{2m\lambda}, \tag{12}$$

where $\mathcal{K}^2$ is the upper bound of the taken kernel function. Hence, the stability-based PAC-Bayes bound of an SVM classifier is given below.

Corollary 2 (Rivasplata et al. [21] [Corollary 3])**.** *Let* $\mathbf{u}$ *be weight vector of an SVM algorithm and* β_m *be the stability coefficient of* $\mathbf{u}$*. Consider the prior distribution is* $P = \mathcal{N}(\mathbb{E}[\mathbf{u}], \sigma^2\mathbf{I})$ *and the posterior distribution is* $Q = \mathcal{N}(\mathbf{u}, \sigma^2\mathbf{I})$*. For any data distribution* $\mathcal{D}$*, for any variance* σ^2*, with probability at least* $1 - \delta$ *the following holds:*

$$KL_{+}(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}}) \leq \frac{\frac{\mathcal{K}^4}{8m\sigma^2\lambda^2}\left(1 + \sqrt{\frac{1}{2}\ln\left(\frac{2}{\delta}\right)}\right)^2 + \ln\left(\frac{m+1}{\frac{\delta}{2}}\right)}{m}. \tag{13}$$

In general, the above stability-based bound is competitive with the non-stability bound when Q in Corollary 1 equals $\mathcal{N}(\mathbf{u}, \sigma^2\mathbf{I})$. Furthermore, it can achieve tighter performance when the hyper-parameters λ takes larger values.

2.4. Non-stability Multi-view Generalization Bounds

Previous theoretical research of multi-view learning mostly focused on co-regularization algorithms. Rademacher complexity was adopted to analyze the generalization performance of the SVM-2K [36], which was extended to the semi-supervised learning setting by Szedmak and Shawe-Taylor [37]. The corresponding bound relies on the empirical estimate for Rademacher complexity and takes expectation under the data generating distribution. That means, this approach implicitly depends on the data generating distribution to define the function class, while PAC-Bayes framework explicitly defines the prior which benefits to encoding complex prior knowledge in terms of the data generating distribution. Sridharan and Kakade [38] provided an analysis of multi-view learning in an information theoretic framework and also gave performance bounds of co-regularization algorithms and SVM-2K with respect to Rademacher complexity. Their bounds compares to the

Bayes optimal predictor and relies on unlabeled data. Different from these works, we provide generalization bounds of multi-view regularization algorithms from the PAC-Bayes framework.

Sun et al. [25] adopted the PAC-Bayes framework to evaluate the generalization performance of the MvSVM algorithm. They considered two prior distributions of weight parameters $\mathbf{u}_1$ and $\mathbf{u}_2$ from two views and multiplied them with a view-consistent function in order to obtain the prior distribution of the concatenated weight $\mathbf{u}$. The posterior was isotropic Gaussian distribution $\mathcal{N}(\mu\boldsymbol{\omega}, \mathbf{I})$ where μ and $\boldsymbol{\omega}$ are the norm and the direction of the mean vector, respectively. Four bounds were proposed with the same view-consistent function

$$V(\mathbf{u}_1, \mathbf{u}_2) = \exp\left\{-\frac{1}{2\sigma_2^2}\mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)}\left[\mathbf{x}_1^\mathsf{T}\mathbf{u}_1 - \mathbf{x}_2^\mathsf{T}\mathbf{u}_2\right]^2\right\}. \tag{14}$$

The view-consistent function made the prior distribution more concerned with weights which produced agreement of classifiers from two views. Choosing priors of isotropic Gaussians centered at origin $\mathcal{N}(\mathbf{0}, \mathbf{I})$, they obtained Theorem 3 and Theorem 4. Theorem 5 and Theorem 6 were provided by considering priors of isotropic Gaussians centered at the expected outputs of the algorithm $\mathcal{N}(\eta\boldsymbol{\omega}_p, \mathbf{I})$ where $\boldsymbol{\omega}_p = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}]$. The difference between Theorem 3 and Theorem 4 (Theorem 5 and Theorem 6) was whether they involved the dimensionality n, which was caused by different inequalities. The n-independent bounds could be utilized when the dimensionality of the feature space using the kernel trick went to infinity. Their experimental results indicated that there was much space and possibility for further developments of multi-view PAC-Bayes analysis. Based on that, our paper proposes a novel view-consistent function and blends the uniform stability of algorithms to promote theoretical analysis.

Theorem 3 (Sun et al. [25] [Theorem 5])**.** *For any data distribution $\mathcal{D}$, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, the following formula holds:*

$$\forall \boldsymbol{\omega}, \mu : \mathit{KL}_{+}(E_{Q,\mathcal{X}} || E_{Q,\mathcal{D}}) \leq \frac{-\frac{n}{2} \ln \left[h_m - \left(\sqrt[n]{\left(\frac{R}{\sigma_2}\right)^2 + 1} - 1 \right) \sqrt{\frac{1}{2m} \ln\left(\frac{3}{\delta}\right)} \right]_{+}}{m} + \frac{\frac{r_m}{2\sigma_2^2} + \frac{R^2}{2\sigma_2^2}\left(1+\mu^2\right) \sqrt{\frac{1}{2m} \ln\left(\frac{3}{\delta}\right)} + \frac{\mu^2}{2} + \ln\left(\frac{m+1}{\frac{\delta}{3}}\right)}{m}, \tag{15}$$

where $h_m = \frac{1}{m}\sum_{i=1}^{m} \left| \mathbf{I} + \frac{\tilde{\mathbf{x}}^{(i)}\tilde{\mathbf{x}}^{(i)\mathsf{T}}}{\sigma_2^2} \right|^{\frac{1}{n}}, r_m = \frac{1}{m}\sum_{i=1}^{m} \left[\tilde{\mathbf{x}}^{(i)\mathsf{T}}\tilde{\mathbf{x}}^{(i)} + \mu^2 \left(\boldsymbol{\omega}^{\mathsf{T}}\tilde{\mathbf{x}}^{(i)}\right)^2 \right]$, $R = \mathit{sup}_{\tilde{\mathbf{x}}} ||\tilde{\mathbf{x}}||$ *and* $\tilde{\mathbf{x}} = \left[\mathbf{x}_1^{\mathsf{T}}, -\mathbf{x}_2^{\mathsf{T}}\right]^{\mathsf{T}}$.

Theorem 4 (Sun et al. [25] [Theorem 6])**.** *For any data distribution $\mathcal{D}$, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, the following formula holds:*

$$\begin{aligned} &\forall \boldsymbol{\omega}, \mu : \mathit{KL}_{+}(E_{Q,\mathcal{X}} || E_{Q,\mathcal{D}}) \\ &\leq \frac{\frac{h_m}{2} + \frac{1}{2}\left(\frac{(1+\mu^2)R^2}{\sigma_2^2} + \ln\left(1 + \frac{R^2}{\sigma_2^2}\right) \right) \sqrt{\frac{1}{2m} \ln\frac{2}{\delta}} + \frac{\mu^2}{2} + \ln\left(\frac{m+1}{\frac{\delta}{2}}\right)}{m}, \end{aligned} \tag{16}$$

where $h_m = \frac{1}{m}\sum_{i=1}^{m} \left(\frac{1}{\sigma_2^2} \left[\tilde{\mathbf{x}}^{(i)\mathsf{T}}\tilde{\mathbf{x}}^{(i)} + \mu^2 \left(\boldsymbol{\omega}^{\mathsf{T}}\tilde{\mathbf{x}}^{(i)}\right)^2 \right] - \ln \left| \mathbf{I} + \frac{\tilde{\mathbf{x}}^{(i)}\tilde{\mathbf{x}}^{(i)\mathsf{T}}}{\sigma_2^2} \right| \right)$.

Theorem 5 (Sun et al. [25] [Theorem 7])**.** *For any data distribution $\mathcal{D}$, for any $\delta \in (0, 1]$, for any $\boldsymbol{\omega}$, μ, and η, with probability at least $1 - \delta$, the following formula* *holds:*

$$
\begin{aligned}
&\mathit{KL}_+(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}})\\
&\leq \frac{-\frac{n}{2}\ln\left[h_m-\left(\sqrt[n]{(R/\sigma_2)^2+1}-1\right)\sqrt{\frac{1}{2m}\ln\frac{4}{\delta}}\quad\right]_+ + \frac{r_m}{2\sigma_2^2}+\ln\left(\frac{m+1}{\frac{\delta}{4}}\right)}{m}\\
&+\frac{\frac{1}{2}\left(\frac{\eta R}{\sqrt{m}}\left(2+\sqrt{2\ln\frac{4}{\delta}}\right)+||\eta\boldsymbol{\omega}_p-\mu\boldsymbol{\omega}||+\mu\right)^2+\frac{R^2+\mu^2R^2+4\eta\mu\sigma_2^2R}{2\sigma_2^2}\sqrt{\frac{1}{2m}\ln\frac{4}{\delta}}}{m},
\end{aligned}
\tag{17}
$$

where $\boldsymbol{\omega}_p=\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}]$, $h_m=\frac{1}{m}\sum_{i=1}^m\left|\mathbf{I}+\frac{\tilde{\mathbf{x}}^{(i)}\tilde{\mathbf{x}}^{(i)\mathsf{T}}}{\sigma_2^2}\right|^{\frac{1}{n}}$ *and*
$r_m=\frac{1}{m}\sum_{i=1}^m\left[\tilde{\mathbf{x}}^{(i)\mathsf{T}}\tilde{\mathbf{x}}^{(i)}-2\eta\mu\sigma_2^2y^{(i)}\left(\boldsymbol{\omega}^\mathsf{T}\mathbf{x}^{(i)}\right)+\mu^2\left(\boldsymbol{\omega}^\mathsf{T}\tilde{\mathbf{x}}^{(i)}\right)^2\right]$.

Theorem 6 (Sun et al. [25] [Theorem 8])**.** *For any data distribution $\mathcal{D}$, for any $\delta\in(0,1]$, for any $\boldsymbol{\omega}$, μ, and η, with probability at least $1-\delta$, the following formula holds:*

$$
\begin{aligned}
\mathit{KL}_+(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}})\leq\;&\frac{\frac{1}{2}\left(\frac{\eta R}{\sqrt{m}}\left(2+\sqrt{2\ln\frac{3}{\delta}}\right)+||\eta\boldsymbol{\omega}_p-\mu\boldsymbol{\omega}||+\mu\right)^2+\frac{h_m}{2}}{m}\\
&+\frac{\frac{R^2+4\eta\mu\sigma_2^2R+\mu^2R^2+\sigma_2^2\ln\left(1+\frac{R^2}{\sigma_2^2}\right)}{2\sigma_2^2}\sqrt{\frac{1}{2m}\ln\frac{3}{\delta}}+\frac{\mu^2}{2}+\ln\left(\frac{m+1}{\frac{\delta}{3}}\right)}{m},
\end{aligned}
\tag{18}
$$

where $h_m=\frac{1}{m}\sum_{i=1}^m\left(\frac{1}{\sigma_2^2}\left[\tilde{\mathbf{x}}^{(i)\mathsf{T}}\tilde{\mathbf{x}}^{(i)}-2\eta\mu\sigma_2^2y^{(i)}\left(\boldsymbol{\omega}^\mathsf{T}\mathbf{x}^{(i)}\right)+\mu^2\left(\boldsymbol{\omega}^\mathsf{T}\tilde{\mathbf{x}}^{(i)}\right)^2\right]\right)$
$+\frac{1}{m}\sum_{i=1}^m\left(-\ln\left|\mathbf{I}+\frac{\tilde{\mathbf{x}}^{(i)}\tilde{\mathbf{x}}^{(i)\mathsf{T}}}{\sigma_2^2}\right|\right)$.

Another relevant work was done by Goyal et al. [28]. They considered a two-level hierarchy for distributions of multiple views. Specific prior and learned posterior distributions were utilized for specific views. Hyper-prior and learned hyper-posterior distributions were considered over the views. Based on this strategy, they derived both probabilistic and expected risk bounds theoretically which exhibited a notion of diversity, and designed two multi-view algorithms based on the two-level PAC-Bayes strategy. Compared with their bounds, our probabilistic PAC-Bayes bound is derived by the consistency-dependent prior distribution which combines the priors from each view. As another difference, our bound is proposed to analyze the generalization performance of multi-view algorithms, rather than design algorithms and optimize the model parameters.

3. Stability-Based PAC-Bayes Bounds for Multi-view Algorithms

We firstly describe the stability-based multi-view PAC-Bayes bound with a novel view-consistency regularization which makes a constraint between the outputs of the two views. Then we also derive a stability-based multi-view PAC-Bayes bound without the view-consistency constraint for performance comparisons.

We consider the linear classifiers of the form $f(\mathbf{x}) = \text{sign}\left(\mathbf{u}^\mathsf{T}\mathbf{x}\right)$ where $\mathbf{u} = \left[\mathbf{u}_1^\mathsf{T}, \mathbf{u}_2^\mathsf{T}\right]^\mathsf{T}$ is the concatenated weight vector from two views and $\mathbf{x} = \left[\mathbf{x}_1^\mathsf{T}, \mathbf{x}_2^\mathsf{T}\right]^\mathsf{T}$ is the concatenated feature vector. Note that the feature vector can be transformed into another feature space by some kernel function $\phi(\mathbf{x})$. The average weight vector is $\mathbb{E}[\mathbf{u}] = \left[\mathbb{E}[\mathbf{u}_1]^\mathsf{T}, \mathbb{E}[\mathbf{u}_2]^\mathsf{T}\right]^\mathsf{T}$ where $\mathbb{E}[\mathbf{u}]$ is the expected output of the multi-view learning algorithm. We also define $\tilde{\mathbf{x}} = \left[\mathbf{x}_1^\mathsf{T}, -\mathbf{x}_2^\mathsf{T}\right]^\mathsf{T}$ for marking conveniently.

3.1. The Bound with the View-Consistency Regularization

Consider that the prior distributions of weights from the two views are

$$P_1(\mathbf{u}_1) = \mathcal{N}(\mathbb{E}[\mathbf{u}_1], \sigma_1^2\mathbf{I}), \tag{19}$$

$$P_2(\mathbf{u}_2) = \mathcal{N}(\mathbb{E}[\mathbf{u}_2], \sigma_1^2 \mathbf{I}). \tag{20}$$

The prior distribution of the weights on the concatenated space is defined as

$$P(\mathbf{u}) \propto P_1(\mathbf{u}_1) P_2(\mathbf{u}_2) V(\mathbf{u}_1, \mathbf{u}_2), \tag{21}$$

where $V(\mathbf{u}_1, \mathbf{u}_2)$ is a view-consistency regularization function.

The previous view-consistency regularization function used in Sun et al. [25] was defined in Equation (14), which made the prior place large probability mass on weight vectors where the random classifiers from two views agree well on all examples averagely. To produce a stability-based prior distribution, we define our novel view-consistency regularization function as

$$V(\mathbf{u}_1, \mathbf{u}_2) = \exp\left\{ -\frac{1}{2\sigma_2^2} \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)} \left[\mathbf{x}_1^\top (\mathbf{u}_1 - \mathbb{E}[\mathbf{u}_1]) - \mathbf{x}_2^\top (\mathbf{u}_2 - \mathbb{E}[\mathbf{u}_2]) \right]^2 \right\}. \tag{22}$$

It indicates that the prior distribution focuses more on the weights which result in less fluctuations of agreements between the random classifiers from two views on all instances averagely.

According to Equation (21), we firstly obtain the prior distribution of the weights as follows (the detailed derivation is given in Appendix A).

$$P(\mathbf{u}) = \exp\left\{ -\frac{1}{2} (\mathbf{u} - \mathbb{E}[\mathbf{u}])^\top \left(\frac{\mathbf{I}}{\sigma_1^2} + \frac{\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top\right]}{\sigma_2^2} \right) (\mathbf{u} - \mathbb{E}[\mathbf{u}]) \right\}. \tag{23}$$

Hence, $P(\mathbf{u}) = \mathcal{N}(\mathbb{E}[\mathbf{u}], \Sigma)$ with $\Sigma = \left(\frac{\mathbf{I}}{\sigma_1^2} + \frac{\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top\right]}{\sigma_2^2} \right)^{-1}$. Suppose that the posterior distribution is $Q(\mathbf{u}) = \mathcal{N}(\mathbf{u}, \mathbf{I})$. With the following formula of the KL divergence between two Gaussian distributions

$$\begin{aligned} &\mathrm{KL}(\mathcal{N}(\mathbf{u}_0, \Sigma_0) || \mathcal{N}(\mathbf{u}_1, \Sigma_1)) \\ &= \frac{1}{2} \left(\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|} \right) + \mathrm{tr}\left(\Sigma_1^{-1} \Sigma_0 \right) + (\mathbf{u}_1 - \mathbf{u}_0)^\top \Sigma_1^{-1} (\mathbf{u}_1 - \mathbf{u}_0) - n \right), \end{aligned} \tag{24}$$

where n is the dimensionality of the feature space, we obtain the KL divergence between the prior distribution $P(\mathbf{u})$ and the posterior distribution $Q(\mathbf{u})$ of a multi-view learning algorithm as follows (the derivation is provided in Appendix B).

$$\begin{aligned} \mathrm{KL}(Q(\mathbf{u})||P(\mathbf{u})) = & \frac{1}{2}\left[-\ln\left(\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\mathbb{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}]}{\sigma_2^2}\right|\right)+\frac{1}{\sigma_2^2}\mathbb{E}\left[\tilde{\mathbf{x}}^\mathsf{T}\tilde{\mathbf{x}}\right]\right] \\ & +\frac{1}{2}\left[+(\mathbb{E}[\mathbf{u}]-\mathbf{u})^\mathsf{T}\mathbb{E}\left[\frac{\mathbf{I}}{\sigma_1^2}+\frac{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}}{\sigma_2^2}\right](\mathbb{E}[\mathbf{u}]-\mathbf{u})+\frac{n}{\sigma_1^2}-n\right]. \end{aligned} \tag{25}$$

Since our defined prior distribution involves the input distribution via the view-consistent function, Equation (25) contains expectation over the input distribution which is unable to be computed directly. Hence, we upper-bound the KL term to provide the stability-based multi-view PAC-Bayes bound that considers the view-consistency regularization. The detailed derivation is given in Appendix C and the corresponding bound is as follows.

Theorem 7. *Let* A *be a multi-view learning algorithm, whose stability coefficient is* β_m. *Consider that the prior distribution is given in (23) and the posterior distribution is* $Q(\mathbf{u}) = \mathcal{N}(\mathbf{u}, \mathbf{I})$. *For any data distribution* $\mathcal{D}$, *for any* $\delta \in (0, 1]$, *with probability at least* $1-\delta$, *the following inequality holds:*

$$\begin{aligned} KL_+(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}}) \leq & \frac{-\frac{n}{2}\ln\left[h_m-\left(\sqrt[n]{\left(\frac{R}{\sigma_2}\right)^2+\frac{1}{\sigma_1^2}}-\frac{1}{\sigma_1^2}\right)\sqrt{\frac{1}{2m}\ln\left(\frac{3}{\delta}\right)}\right]_+ + R^2}{m} \\ & + \frac{\left(\frac{1}{\sigma_1^2}+\frac{R^2}{\sigma_2^2}\right)m\beta_m^2\left(1+\sqrt{\frac{1}{2}\ln\frac{3}{\delta}}\right)^2+\frac{n}{\sigma_1^2}-n+\ln\left(\frac{m+1}{\frac{\delta}{3}}\right)}{m}, \end{aligned} \tag{26}$$

where $h_m = \frac{1}{m}\sum_{i=1}^m \left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\tilde{\mathbf{x}}^{(i)}\tilde{\mathbf{x}}^{(i)\mathsf{T}}}{\sigma_2^2}\right|^{\frac{1}{n}}$, $(\cdot)_+ = max(0,\cdot)$, $R = sup_{\tilde{\mathbf{x}}}||\tilde{\mathbf{x}}||$ *and* n *is the dimensionality of the feature space.*

It is clear that the stability coefficient β_m controls the tightness of the PAC-Bayes bound. The smaller the upper bound on the β_m is, the tighter the multi-view PAC-Bayes bound will be. Hence, a more stable algorithm will likely lead to a tighter PAC-Bayes bound and achieve better generalization performance.

3.2. The Bound without the View-Consistency Regularization

To illustrate the performance of our proposed view-consistency regularization function, we now neglect the view-consistency constraint and directly adapt the single-view stability-based PAC-Bayes bound to the multi-view learning setting. It means that the outputs among different views are independent, which may cause large differences when data in different views reveal different extents of information.

Without the view-consistency regularization between two views, the prior distribution of weights on the concatenated space is simply proportional to the product of prior distributions of weights from two views. Consider that $P(\mathbf{u}_1) = \mathcal{N}(\mathbb{E}[\mathbf{u}_1], \sigma^2\mathbf{I})$ and $P(\mathbf{u}_2) = \mathcal{N}(\mathbb{E}[\mathbf{u}_2], \sigma^2\mathbf{I})$. Then the prior distribution of a learner is given by

$$P(\mathbf{u}) \propto P(\mathbf{u}_1)P(\mathbf{u}_2) = \exp\left\{-\frac{1}{2}(\mathbf{u} - \mathbb{E}[\mathbf{u}])^\mathsf{T}\frac{1}{\sigma^2}(\mathbf{u} - \mathbb{E}[\mathbf{u}])\right\}. \tag{27}$$

That is, $P(\mathbf{u}) = \mathcal{N}(\mathbb{E}[\mathbf{u}], \sigma^2\mathbf{I})$. Given that $Q(\mathbf{u}) = \mathcal{N}(\mathbf{u}, \sigma^2\mathbf{I})$, we obtain the KL divergence between the two distributions according to Equation (24) as follows

$$\begin{aligned} \mathrm{KL}(Q(\mathbf{u})||P(\mathbf{u})) &= \frac{1}{2}\left(\mathrm{tr}\left(\frac{\mathbf{I}}{\sigma^2}\cdot\sigma^2\mathbf{I}\right) + (\mathbb{E}[\mathbf{u}] - \mathbf{u})^\mathsf{T}\frac{1}{\sigma^2}(\mathbb{E}[\mathbf{u}] - \mathbf{u}) - n\right) \\ &= \frac{1}{2\sigma^2}(\mathbb{E}[\mathbf{u}] - \mathbf{u})^2. \end{aligned} \tag{28}$$

We give the stability-based multi-view bound without the view-consistency regularization below (the detailed derivation is given in Appendix D).

Theorem 8. *Let* A *be a multi-view learning algorithm, whose stability coefficient is* β_m. *Consider that the prior distribution is given in (27) and the posterior distribution is* $Q(\mathbf{u}) = \mathcal{N}(\mathbf{u}, \sigma^2\mathbf{I})$. *For any data distribution* $\mathcal{D}$, *for any variance* σ^2, *for any* $\delta \in (0, 1]$, *with probability at least* $1 - \delta$, *the following inequality holds:*

$$KL_+(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}}) \leq \frac{\frac{1}{2\sigma^2}m\beta_m^2\left(1+\sqrt{\frac{1}{2}\ln\left(\frac{2}{\delta}\right)}\right)^2 + \ln\left(\frac{m+1}{\frac{\delta}{2}}\right)}{m}. \tag{29}$$

Lacking the view-consistency constraint, the priors of weights from different views are independent of each other which produce the same prior distribution as that in the single-view learning setting. In addition, the above formula again shows that if an algorithm is more stable, its generalization performance will be better.

4. Upper-Bounding the Stability Coefficient of Multi-view Algorithms

To compute the specific PAC-Bayes bound of a multi-view learning algorithm, it is necessary to derive the upper bound on its stability coefficient β_m. The multi-view regularization algorithm minimizes the following objective to learn an optimal weight vector $\mathbf{u}_*$ on the training set $\mathcal{X}$

$$\mathcal{J}(\mathbf{u}) = \frac{1}{2m}\sum_{i=1}^{m} l\left(f_{\mathbf{u}}\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) + \lambda_1 N(\mathbf{u}) + \lambda_2 M(\mathbf{u}), \tag{30}$$

where l is the loss function, N and M are convex functions from $\mathcal{U}$ to $\mathcal{R} \cup \{-\infty, +\infty\}$, and λ_1 and λ_2 are regularization coefficients. Consider the loss function is σ-admissible, whose definition is provided in Definition 1.

Definition 1. *A loss function l is σ-admissible if possible predictions y_1, y_2 satisfy the following inequality:*

$$|l(y_1, y) - l(y_2, y)| \leq \sigma |y_1 - y_2| . \tag{31}$$

When $\mathcal{J}$ is everywhere differentiable, the Bregman divergence associated to $\mathcal{J}$ of $\mathbf{u}$ to $\mathbf{v}$ is defined by $\forall \mathbf{v}, \mathbf{u} \in \mathcal{U}, d_{\mathcal{J}}(\mathbf{v}, \mathbf{u}) = \mathcal{J}(\mathbf{v}) - \mathcal{J}(\mathbf{u}) - \langle \mathbf{v} - \mathbf{u}, \triangledown \mathcal{J}(\mathbf{u}) \rangle$ where $\triangledown \mathcal{J}(\mathbf{u})$ is a subgradient of $\mathcal{J}$ in $\mathbf{u}$. We give Lemma 1 to prepare for upper-bounding the stability coefficient. The proof is given in Appendix E.

Lemma 1. *Let l be α-admissible and $\mathbf{u}_*^{\mathbf{z}^{(i)} \rightarrow \mathbf{z}^{(j)}}$ denote a new optimal weight vector on the changed training set where the example $\mathbf{z}^{(i)}$ is substituted by another example $\mathbf{z}^{(j)}$. When l, N, and M are differentiable, we have*

$$\begin{aligned}
&\lambda_1 d_N \left(\mathbf{u}_*, \mathbf{u}_*^{\mathbf{z}^{(i)} \rightarrow \mathbf{z}^{(j)}} \right) + \lambda_2 d_M \left(\mathbf{u}_*, \mathbf{u}_*^{\mathbf{z}^{(i)} \rightarrow \mathbf{z}^{(j)}} \right) + \lambda_1 d_N \left(\mathbf{u}_*^{\mathbf{z}^{(i)} \rightarrow \mathbf{z}^{(j)}}, \mathbf{u}_* \right) \\
&+ \lambda_2 d_M \left(\mathbf{u}_*^{\mathbf{z}^{(i)} \rightarrow \mathbf{z}^{(j)}}, \mathbf{u}_* \right) \leq \frac{\alpha}{2m} \left(\left| \Delta \mathbf{u}_*^{\mathsf{T}} \mathbf{x}^{(i)} \right| + \left| \Delta \mathbf{u}_*^{\mathsf{T}} \mathbf{x}^{(j)} \right| \right),
\end{aligned}$$

where $\Delta \mathbf{u}_* = \mathbf{u}_*^{\mathbf{z}^{(i)} \rightarrow \mathbf{z}^{(j)}} - \mathbf{u}_*$, $d_N \left(\mathbf{u}_*, \mathbf{u}_*^{\mathbf{z}^{(i)} \rightarrow \mathbf{z}^{(j)}} \right)$ and $d_M \left(\mathbf{u}_*, \mathbf{u}_*^{\mathbf{z}^{(i)} \rightarrow \mathbf{z}^{(j)}} \right)$ are the Bregman divergences associated to N and M of $\mathbf{u}_*$ to $\mathbf{u}_*^{\mathbf{z}^{(i)} \rightarrow \mathbf{z}^{(j)}}$, respectively.

We specialize Lemma 1 to the MvSVM, whose objective function is

$$\begin{aligned}
&\mathrm{MvSVM}_{\lambda_1, \lambda_2}(\mathbf{u}) = \underset{\mathbf{u}_1, \mathbf{u}_2}{\arg\min} \frac{1}{2m} \sum_{i=1}^{m} l \left(f_{\mathbf{u}} \left(\mathbf{x}^{(i)} \right), y^{(i)} \right) \\
&+ \underset{\mathbf{u}_1, \mathbf{u}_2}{\arg\min} \left(\lambda_1 \left(||\mathbf{u}_1||^2 + ||\mathbf{u}_2||^2 \right) + \lambda_2 \sum_{i=1}^{m} \left(\mathbf{u}_1^{\mathsf{T}} \mathbf{x}_1^{(i)} - \mathbf{u}_2^{\mathsf{T}} \mathbf{x}_2^{(i)} \right)^2 \right),
\end{aligned} \tag{32}$$

where l is the hinge loss function and $\mathbf{u} = [\mathbf{u}_1^{\mathsf{T}}, \mathbf{u}_2^{\mathsf{T}}]^{\mathsf{T}}$. In this context, the upper bound on the stability coefficient of the MvSVM algorithm is given by Theorem 9.

Theorem 9. *Consider an MvSVM algorithm with a stability coefficient β_m. Let $\mathcal{K}^2$ be the upper bound of some kernel function. We obtain the upper bound on the stability coefficient*

$$\beta_m \leq \frac{\mathcal{K}^2}{2m\left(\lambda_1 + \lambda_2 R^2\right)}, \tag{33}$$

where $R = \sup_{\tilde{\mathbf{x}}} ||\tilde{\mathbf{x}}||$ and $\tilde{\mathbf{x}} = [\mathbf{x}_1^\mathsf{T}, -\mathbf{x}_2^\mathsf{T}]^\mathsf{T}$.

Proof. Comparing the target function of the MvSVM algorithm with the general form of multi-view regularization algorithms, we get $N(\mathbf{u}) = ||\mathbf{u}||^2$, $M(\mathbf{u}) = (\mathbf{u}^\mathsf{T}\tilde{\mathbf{x}})^2$ where $\tilde{\mathbf{x}} = [\mathbf{x}_1^\mathsf{T}, -\mathbf{x}_2^\mathsf{T}]^\mathsf{T}$. The Bregman divergences associated to N and M of $\mathbf{u}_*$ to $\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}$ are given as follows

$$d_N\left(\mathbf{u}_*, \mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}\right) = \left|\left|\mathbf{u}_* - \mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}\right|\right|^2 = ||\Delta\mathbf{u}_*||^2.$$

$$\begin{aligned} d_M\left(\mathbf{u}_*, \mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}\right) &= \left(\mathbf{u}_*^\mathsf{T}\tilde{\mathbf{x}} - \mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}\mathsf{T}}\tilde{\mathbf{x}}\right)^2 = \left(\Delta\mathbf{u}_*^\mathsf{T}\tilde{\mathbf{x}}\right)^2 \\ &= \left|\left|\Delta\mathbf{u}_*^\mathsf{T}\tilde{\mathbf{x}}\right|\right|^2 \leq \left|\left|\Delta\mathbf{u}_*^\mathsf{T}\right|\right|^2 ||\tilde{\mathbf{x}}||^2 \leq \left|\left|\Delta\mathbf{u}_*^\mathsf{T}\right|\right|^2 R^2 = ||\Delta\mathbf{u}_*||^2 R^2, \end{aligned}$$

where $R = \sup_{\tilde{\mathbf{x}}} ||\tilde{\mathbf{x}}||$. Applying Lemma 1, we obtain

$$2\lambda_1||\Delta\mathbf{u}_*||^2 + 2\lambda_2||\Delta\mathbf{u}_*||^2R^2 \leq \frac{\alpha}{2m}\left(\left|\Delta\mathbf{u}_*^\mathsf{T}\mathbf{x}^{(i)}\right| + \left|\Delta\mathbf{u}_*^\mathsf{T}\mathbf{x}^{(j)}\right|\right).$$

The reproducing property of a reproducing kernel Hilbert space (RKHS) is written as $\forall\mathbf{x}, \mathbf{u}^\mathsf{T}\mathbf{x} = \langle\mathbf{u}, k(\mathbf{x}, \cdot)\rangle$. With Cauchy-Schwarz's inequality, we have $\forall\mathbf{x}, |\mathbf{u}^\mathsf{T}\mathbf{x}| \leq ||\mathbf{u}||\sqrt{k(\mathbf{x}, \mathbf{x})}$. It is direct to get that

$$\left|\Delta\mathbf{u}_*^\mathsf{T}\mathbf{x}^{(i)}\right| \leq ||\Delta\mathbf{u}_*||\sqrt{k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(i)}\right)} \leq \mathcal{K}||\Delta\mathbf{u}_*||,$$

where we denote $\sqrt{k(\mathbf{x}, \mathbf{x})} \leq \mathcal{K}$. In the same way, $\left|\Delta\mathbf{u}_*^\mathsf{T}\mathbf{x}^{(j)}\right| \leq \mathcal{K}||\Delta\mathbf{u}_*||$. Hence,

$$2\lambda_1||\Delta\mathbf{u}_*||^2 + 2\lambda_2||\Delta\mathbf{u}_*||^2R^2 \leq \frac{\alpha}{2m}\cdot 2\mathcal{K}||\Delta\mathbf{u}_*|| \leq \frac{\alpha\mathcal{K}}{m}||\Delta\mathbf{u}_*||.$$

Then we get

$$||\Delta\mathbf{u}_*|| \leq \frac{\alpha\mathcal{K}}{2m(\lambda_1+\lambda_2R^2)},$$

and thus

$$\left|\Delta\mathbf{u}_*^\mathsf{T}\mathbf{x}^{(i)}\right| \leq \frac{\alpha\mathcal{K}^2}{2m(\lambda_1+\lambda_2R^2)}.$$

Since the loss function is the hinge loss which is 1-admissible, that is, $\alpha = 1$. Recalling the definition of β_m in Section 2.3, we give the upper bound on the stability coefficient

$$\beta_m = \sup_{i\in[m],\forall\mathbf{x}} \left|\left|\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}\mathsf{T}}\mathbf{x} - \mathbf{u}_*^\mathsf{T}\mathbf{x}\right|\right|_\infty \leq \sup_{i\in[m]}\left|\Delta\mathbf{u}_*^\mathsf{T}\mathbf{x}^{(i)}\right| \leq \frac{\mathcal{K}^2}{2m(\lambda_1+\lambda_2R^2)}.$$

□

Consequently, a direct application of Theorem 7 together with Theorem 9 gives the view-consistency stability-based PAC-Bayes bound for the MvSVM algorithm as below.

Corollary 3. *Consider an MvSVM algorithm with a stability coefficient β_m. Suppose that the prior distribution is given in (23) and the posterior distribution is $Q(\mathbf{u}) = \mathcal{N}(\mathbf{u},\mathbf{I})$. For any data distribution $\mathcal{D}$, for any $\delta \in (0,1]$, with probability at least $1-\delta$, the following inequality holds:*

$$\boldsymbol{KL}_+(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}}) \leq \frac{-\frac{n}{2}\ln\left[h_m - \left(\sqrt[n]{\left(\frac{R}{\sigma_2}\right)^2+\frac{1}{\sigma_1^2}} - \frac{1}{\sigma_1^2}\right)\sqrt{\frac{1}{2m}\ln\left(\frac{3}{\delta}\right)}\right]_+ + R^2}{m}$$
$$+\frac{\left(\frac{1}{\sigma_1^2}+\frac{R^2}{\sigma_2^2}\right)\frac{\mathcal{K}^4}{4m(\lambda_1+\lambda_2R^2)^2}\left(1+\sqrt{\frac{1}{2}\ln\frac{3}{\delta}}\right)^2 + \frac{n}{\sigma_1^2} - n + \ln\left(\frac{m+1}{\frac{\delta}{3}}\right)}{m}, \tag{34}$$

where $h_m = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{\mathbf{I}}{\sigma_1^2} + \frac{\tilde{\mathbf{x}}^{(i)}\tilde{\mathbf{x}}^{(i)\top}}{\sigma_2^2}\right|^{\frac{1}{n}}$, $(\cdot)_+ = max(0,\cdot)$, $R = sup_{\tilde{\mathbf{x}}}||\tilde{\mathbf{x}}||$ *and* n *is the dimensionality of the feature space.*

Compared with the four non-stability PAC-Bayes bounds of the MvSVM algorithm which are introduced in Theorems 3, 4, 5, and 6, our stability-based bound has a term involving the hyper-parameters λ_1 and λ_2 of the algorithm. This term decreases as the values of λ_1 and λ_2 increase, which control the strength of penalization on the norms of classifiers from two views and the disagreement between the classifiers, respectively. It indicates that the stability-based bound is sensitive to the hyper-parameters of the algorithm.

The following expressions list the convergence rates of these bounds as functions of the number of examples m:

$$O\left(\left(\ln\sqrt{1/m} + \sqrt{1/m} + \ln m\right)/m\right), \quad (Theorem\ 3) \tag{35}$$

$$O\left(\left(\sqrt{1/m} + \ln m\right)/m\right), \quad (Theorem\ 4) \tag{36}$$

$$O\left(\left(\ln\sqrt{1/m} + 1/m + \sqrt{1/m} + \ln m\right)/m\right), \quad (Theorem\ 5) \tag{37}$$

$$O\left(\left(1/m + \sqrt{1/m} + \ln m\right)/m\right), \quad (Theorem\ 6) \tag{38}$$

$$O\left(\left(\ln\sqrt{1/m} + 1/m + \ln m\right)/m\right). \quad (Corollary\ 3) \tag{39}$$

Although convergence rates of all these bounds are dominated by the $\frac{\ln m}{m}$ term, our stability-based bound improves the rate in the other terms. Specifically, we have the inequality when $m > 1$: $\left(\ln\sqrt{1/m} + 1/m\right)/m < \left(\ln\sqrt{1/m} + \sqrt{1/m}\right)/m < \sqrt{1/m}/m$, which means that the stability-based bound shrinks faster than the bounds delivered by Theorems 3 and 4. The bounds presented by Theorems 5 and 6 drop slower than the stability-based bound since the inequality holds when $m > 1$: $\left(\ln\sqrt{1/m} + 1/m\right)/m < \left(\ln\sqrt{1/m} + 1/m + \sqrt{1/m}\right)/m < \left(1/m + \sqrt{1/m}\right)/m$. We present quantitative comparisons of these bounds in the experiments.

For experimental comparisons, we give the stability-based multi-view bound without the view-consistency regularization as follow by combining Theorem 8 and Theorem 9.

Corollary 4. *Consider an MvSVM algorithm with a stability coefficient β_m. Suppose that the prior distribution is given in (27) and the posterior distribution is $Q(\mathbf{u}) = \mathcal{N}(\mathbf{u}, \sigma^2\mathbf{I})$. For any data distribution $\mathcal{D}$, for any $\delta \in (0,1]$, with probability at least $1-\delta$, the following inequality holds:*

$$KL_{+}(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}}) \leq \frac{\frac{\mathcal{K}^4}{8\sigma^2 m(\lambda_1+\lambda_2 R^2)^2}\left(1+\sqrt{\frac{1}{2}\ln\left(\frac{2}{\delta}\right)}\right)^2 + \ln\left(\frac{m+1}{\frac{\delta}{2}}\right)}{m}. \tag{40}$$

The stability-based bound without the view-consistency constraint seems similar to the single-view stability-based bound presented by Corollary 2, expect for the first term on the right hand side. This term comes from the upper bound on the stability coefficient of the MvSVM, which is different from that of the SVM.

4.1. Remarks on the Regularization Parameters λ_1 and λ_2

To theoretically illustrate benefits of multi-view learning and stability, we investigate the influence of the regularization hyper-parameters λ_1 and λ_2 on different bounds, including the single-view non-stability bound (Corollary 1), the single-view stability-based bound (Corollary 2), the multi-view non-stability bound (Theorem 3), and the multi-view stability-based bound (Corollary 3). To this end, we apply the multi-view bounds to the multi-view regularization problem (32) and the single-view bounds to the same problem but without the co-regularization term.

Since the non-stability bounds do not depend on λ_1 and λ_2 explicitly, we first use the following relationship, due to the equivalence between regularization and constraint for convex optimization problems,

$$||\mathbf{u}_1||^2 + ||\mathbf{u}_2||^2 \leq \frac{k_1}{\lambda_1}, \tag{41}$$

$$\sum_{i=1}^{m} \left(\mathbf{u}_1^\mathsf{T} \mathbf{x}_1^{(i)} - \mathbf{u}_2^\mathsf{T} \mathbf{x}_2^{(i)} \right)^2 \leq \frac{k_2}{\lambda_2}, \tag{42}$$

where k_1, $k_2 > 0$ are some constants, to bound the corresponding terms in the single-view non-stability bound and the multi-view non-stability bound. We have

$$\forall \boldsymbol{\omega}, \mu : \mathrm{KL}_+(E_{Q,\mathcal{X}} || E_{Q,\mathcal{D}}) \leq \frac{\frac{k_1}{2\lambda_1} + \ln(\frac{m+1}{\delta})}{m}, \tag{43}$$

and

$$\forall \boldsymbol{\omega}, \mu : \mathrm{KL}_+(E_{Q,\mathcal{X}} || E_{Q,\mathcal{D}}) \leq \frac{-\frac{n}{2} \ln \left[h_m - \left(\sqrt[n]{\left(\frac{R}{\sigma_2}\right)^2 + 1} - 1 \right) \sqrt{\frac{1}{2m} \ln\left(\frac{3}{\delta}\right)} \right]_+}{m} + \frac{\frac{R^2}{2\sigma_2^2} + \frac{k_2}{2m\sigma_2^2\lambda_2} + \frac{R^2}{2\sigma_2^2}\left(1 + \frac{k_1}{\lambda_1}\right)\sqrt{\frac{1}{2m}\ln\left(\frac{3}{\delta}\right)} + \frac{k_1}{2\lambda_1} + \ln\left(\frac{m+1}{\frac{\delta}{3}}\right)}{m}, \tag{44}$$

where $h_m = \frac{1}{m}\sum_{i=1}^{m}\left|\mathbf{I} + \frac{\tilde{\mathbf{x}}^{(i)}\tilde{\mathbf{x}}^{(i)\top}}{\sigma_2^2}\right|^{\frac{1}{n}}$, $R = \sup_{\tilde{\mathbf{x}}}||\tilde{\mathbf{x}}||$ and $\tilde{\mathbf{x}} = \left[\mathbf{x}_1^\top, -\mathbf{x}_2^\top\right]^\top$.

Comparison of multi-view stability-based bound with single-view bounds On one hand, note that the single-view non-stability bound and the single-view stability-based bound depend on λ_1 in the forms of $\frac{1}{\lambda_1}$ and $\frac{1}{\lambda_1^2}$, respectively, which suggests that the stability-based bound is more sensitive to the change in the strength of regularization than the non-stability one. Our multi-view stability-based bound relies on λ_1 and λ_2 in the form of $\frac{1}{(\lambda_1+\lambda_2 R^2)^2}$, where the additional view-consistency constraint could directly empower the L_2-regularization to decrease the bound. This provides an explanation on how view-consistency regularization in multi-view learning algorithms can lead to better generalization.

Comparison of multi-view bounds On the other hand, the multi-view non-stability bound depends on λ_1 and λ_2 in a different way, which is in the forms of $\frac{1}{\lambda_1}$ and $\frac{1}{\lambda_2}$ separately. Although increasing the strength of view-consistency regularization could decrease the bound, it is unclear whether the multi-view non-stability bound could improve over the corresponding single-view non-stability one by comparing Equation (44) and Equation (43). However, our multi-view stability-based bound is controlled by λ_1 and λ_2 in a coupled way due to stability, which could indicate the potential superiority over the single-view stability-based bound by comparing $\frac{1}{(\lambda_1+\lambda_2 R^2)^2}$ in Corollary 3 and $\frac{1}{\lambda_1^2}$ in Corollary 2. This demonstrates the advantage of stability on explaining why the multi-view stability-based bound may outperform the single-view stability-based bound. Furthermore, the multi-view stability-based bound decreases faster than the multi-view non-stability bound when enhancing the strength of regularization, since the multi-view stability-based bound is influenced by the hyper-parameters λ_1 and λ_2 in a higher order. This provides another benefit of stability on tightening multi-view bounds.

5. Experiments

We train the SVM and MvSVM on nine datasets to compare multi-view learning algorithms with single-view learning algorithms and evaluate the performance of the proposed stability-based multi-view bound. We first introduce datasets and experimental configurations, and then report the experimental results.

5.1. Datasets

We describe the characteristics of all nine datasets in detail. Table 1 lists the number of instances, the number of positive instances, the number of negative instances, the feature dimensionality of view 1, the feature dimensionality of view 2, the features of view 1, and the features of view 2.

Ads The Ads dataset is used to classify web images into advertisements and non-advertisements [39]. The two views represent the image itself (terms in the image's caption, URL and alt text) and other objects in the web page (terms in the page and destination URLs), respectively.

Course The Course dataset is used for the web page classification problem, which judges whether a web page is a course web page or not [40]. A web page is described by two views where one is the content in the web page itself and the other is the anchor text of any web page linking to it.

Hand The Hand dataset is taken from the UCI machine learning repository [41] to distinguish images of ten handwritten digits. For simplicity, we recognize two classes where one is (1, 2, 3) and the other is (4, 5, 6). Each example is represented by Fourier coefficients and Karhunen-Loéve coefficients, which serve as two views, respectively.

Syn We treat two randomly generated direction vectors as two different views and sample 2000 points, half of which belong to the positive class. If the inner product between the direction vector and the feature vector is positive, the point is considered as the positive, otherwise as the negative class. Finally, we add Gaussian white noise to form the Syn dataset.

Cora The Cora dataset classifies the fields to which scientific publications belong [42]. There are seven categories. We take the field of the most publications as the positive class and the rest fields are set to be the negative. Each instance is expressed by words used in the publication and citation links between other publications and itself, respectively.

Wis The Wis dataset is a subset of the Course dataset, used to distinguish whether a web page is a student web page or not. The two views of each web page are words in the page and words in the links referring to it.

Attack The Attack dataset is used for classifying types of violent attacks. The original dataset contains six categories and we take the types of the two most violent attacks as the positive and the negative class. One view shows the information of the violent attack and the other view expresses the relations between the current violent attack and other violent attacks that occurred in the same place and were held by the same organization.

Ionos The Ionos dataset is extracted from the UCI machine learning repository for classifying the types of structure in the ionosphere [41]. We randomly divide

Table 1: Descriptions of nine datasets

datasets	#samples	#pos	#neg	#view 1	#view 2	view 1	view 2
Ads	3279	459	2820	587	967	images	objects except images
Course	1051	230	821	500	87	web contents	anchor texts
Hand	1200	600	600	76	64	Fourier coefficients	Karhunen-Loéve coefficients
Syn	2000	1000	1000	50	50	one direction vector	the other direction vector
Cora	2708	818	1890	1433	2708	words	citation links
Wis	265	122	143	1703	265	web contents	anchor texts
Attack	1060	498	562	106	1060	violent attacks	relationships
Ionos	351	225	126	7	17	randomly split	randomly split
Breast	699	458	241	4	5	randomly split	randomly split

the features into two views by a ratio of 2:8.

Breast The Breast dataset is also from the UCI machine learning repository and is used to decide whether one has breast cancer or not [41]. We split the features into two views averagely.

5.2. Configurations

Each dataset is randomly split into a training set and a test set with a proportion of 8:2. All the experiments are performed 10 times. We train the SVM classifier and the MvSVM classifier using the linear kernel with data of different views. In the training process, we adopt the standard forms of the SVM algorithm and the MvSVM algorithm, which are

$$\mathrm{SVM}_C(\mathbf{u}) = \arg\min_{\mathbf{u}} \left(C \sum_{i=1}^{m} l\left(f_{\mathbf{u}}\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) + \frac{1}{2}||\mathbf{u}||^2 \right), \tag{45}$$

$$
\begin{aligned}
\mathrm{MvSVM}_{C_1,C_2}(\mathbf{u}) &= \operatorname*{arg\,min}_{\mathbf{u}_1,\mathbf{u}_2} C_1 \sum_{i=1}^{m} l\left(f_{\mathbf{u}}\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) \\
&+ \operatorname*{arg\,min}_{\mathbf{u}_1,\mathbf{u}_2} \left(\frac{1}{2}\left(||\mathbf{u}_1||^2 + ||\mathbf{u}_2||^2\right) + C_2 \sum_{i=1}^{m} \left(\mathbf{u}_1^\top \mathbf{x}_1^{(i)} - \mathbf{u}_2^\top \mathbf{x}_2^{(i)}\right)^2 \right).
\end{aligned}
\tag{46}
$$

The model parameters C in the SVM and C_1, C_2 in the MvSVM are selected by three-fold cross-validation, whose ranges are $\{2^{-8}, 2^{-7}, \cdots, 1, 2, 4\} \times C_0$ and $\{2^{-4}, 2^{-3}, \cdots, 1, 2, 4\} \times C_0$, respectively, where C_0 is the reciprocal of the upper bound of the feature space transformed by the kernel function.

To compute the upper bounds on the stability coefficients of the SVM algorithm and the MvSVM algorithm which are separately presented in Equation (12) and in Equation (33), we illustrate the conversions between the standard forms and the λ-forms. The λ-forms of the SVM algorithm and the MvSVM algorithm are presented as Equation (11) and Equation (32), respectively. Comparing the standard forms with λ-forms of the SVM and the MvSVM, we could get

$$
\lambda = \frac{1}{2mC}, \tag{47}
$$

and

$$
\lambda_1 = \frac{1}{4mC_1}, \quad \lambda_2 = \frac{C_2}{2mC_1}. \tag{48}
$$

According to the forementioned theoretical analysis, the stability-based multi-view bound will tend to decrease when the hyper-parameters λ_1 and λ_2 take larger values. Therefore, smaller values of C_1 and larger values of C_2 might lead to a tighter bound.

In addition, all PAC-Bayes bounds are computed with a confidence of $\delta = 0.05$. Let σ_2 be 100 to calculate the multi-view PAC-Bayes bounds. For view-consistency multi-view stability-based bounds, σ_1 is equal to 100. For multi-view stability-based bounds without the view-consistency constraint and single-view stability-based bounds, we fix σ to be 1.

5.3. Results

To illustrate the superiority of multi-view learning algorithms over single-view learning algorithms, we compare classification errors and the corresponding stability-based PAC-Bayes bounds. We are interested in the strengths and weaknesses of various multi-view bounds, especially in their tightness and their ability to support model selection.

5.3.1. Multi-view Algorithms vs. Single-view Algorithms

We illustrate the superior performance of multi-view learning over single-view learning in terms of classification errors and stability-based PAC-Bayes bounds.

Classification Errors The averages and standard deviations of classification errors of different learning algorithms are shown in Table 2. To illustrate the significance of these empirical errors, it also includes the confidence intervals when the confidence is 95%. SVM-v1 and SVM-v2 are the SVM classifiers trained on view 1 and view 2, respectively. By concatenating view 1 and view 2 into a long single view, we obtain results of SVM-v3 with the SVM algorithm. The MvSVM-v3 classifier is trained on view 1 and view 2. By randomly dividing view 1 (2) into two parts, we obtain the MvSVM-v1 (MvSVM-v2) classifier.

From the table, the MvSVM algorithm achieves the highest performance on seven of the nine datasets. It demonstrates that simply concatenating two views into a single long view usually cannot capture the structural information among

Table 2: Averages (%), standard deviations (%) and confidence intervals (%) of classification errors. Each dataset contains two rows where the first row shows averages and standard deviations, and the second row is confidence intervals at the 95% confidence. The bold numbers mean the best performance.

Error	SVM-v1	SVM-v2	SVM-v3	MvSVM-v1	MvSVM-v2	MvSVM-v3
Hand	5.25±1.69	3.58±1.01	**1.33±0.38**	6.21±1.32	5.17±0.95	1.92±0.93
	[4.20, 6.30]	[2.96, 4.20]	[1.09, 1.57]	[5.39, 7.03]	[4.58, 5.76]	[1.35, 2.49]
Cora	12.03±1.76	18.60±1.11	**10.44±1.02**	13.39±1.54	18.89±0.65	11.68±1.06
	[10.94, 13.12]	[17.81, 19.29]	[9.81, 11.07]	[11.86, 14.93]	[18.49, 19.29]	[11.02, 12.34]
Attack	7.78±2.11	43.07±3.14	8.07±2.67	**7.52±0.82**	39.46±1.18	10.39±0.98
	[6.47, 9.09]	[41.13, 45.01]	[6.41, 9.73]	[7.01, 8.03]	[38.73, 40.19]	[9.78, 11.00]
Ionos	15.29±5.39	18.43±6.44	15.00±5.31	**13.29±3.02**	18.14±3.69	13.71±3.88
	[11.95, 18.63]	[14.44, 22.42]	[11.71, 18.29]	[11.42, 15.16]	[15.85, 20.43]	[11.30, 16.12]
Ads	4.48±0.70	3.66±0.55	3.19±0.47	5.08±0.62	4.16±0.45	**2.85±0.65**
	[4.05, 4.91]	[3.32, 4.00]	[2.90, 3.48]	[4.70, 5.46]	[3.88, 4.44]	[2.45, 3.25]
Course	9.33±1.08	6.38±1.27	5.24±0.98	17.14±1.37	5.43±1.44	**4.62±0.90**
	[8.66, 10.00]	[5.59, 7.17]	[4.36, 5.85]	[16.29, 17.99]	[4.54, 6.32]	[4.06, 5.18]
Syn	13.45±1.60	17.08±0.90	8.12±1.21	39.97±1.93	38.90±2.62	**7.30±0.75**
	[12.46, 14.44]	[16.52, 17.64]	[7.37, 8.87]	[38.77, 41.17]	[37.27, 40.53]	[6.83, 7.77]
Wis	13.21±12.00	33.21±6.55	10.38±2.85	14.15±9.30	38.49±4.46	**10.19±4.81**
	[5.77, 20.65]	[29.15, 37.27]	[8.62, 12.14]	[8.39, 19.91]	[35.27, 41.26]	[7.21, 13.17]
Breast	5.36±1.23	5.79±1.76	7.57±1.82	6.14±1.76	8.50±4.40	**4.29±1.78**
	[4.60, 6.12]	[4.70, 6.88]	[6.44, 8.70]	[5.50, 7.23]	[5.82, 11.22]	[3.19, 5.39]

multiple views well, while multi-view learning provides the merit by constraining outputs of multiple views. It also reflects the advantage of multi-view learning over single-view learning in terms of empirical errors. Furthermore, for the Ionos and Breast (artificial constructed multi-view datasets), they all get the smallest classification errors with multi-view learning algorithms. This illustrates that multi-view learning may improve the performance of single-view learning on a single-view dataset in some cases.

Stability-Based PAC-Bayes Bounds Table 3 presents the stability-based PAC-Bayes bounds of multi-view random classifiers and single-view random classifiers. SPB represents the stability-based single-view bound which is introduced in Corollary 2. MPB and MPBc correspond to directly adapted and view-consistency multi-view bounds, respectively. The MPBc bounds are always tighter than the SPB bounds on all the nine datasets, which demonstrates that multi-view learning algorithms are superior to single-view learning algorithms in terms of generalization performance. In addition, the MPB bounds are clearly worse than the MPBc bounds. This demonstrates that the newly proposed view-consistency function indeed contributes to tightening the multi-view bounds.

5.3.2. Stability-Based Multi-view Bounds vs. Non-stability Multi-view Bounds

We explore the strengths and weaknesses of various multi-view bounds from the aspects of tightness and the ability to support model selection. Since the MPB bound is not our focus, it is not compared with other non-stability multi-view bounds. MPBod and MPBoi are separately dimension dependent and independent non-stability multi-view bounds where the priors of isotropic Gaussians centered

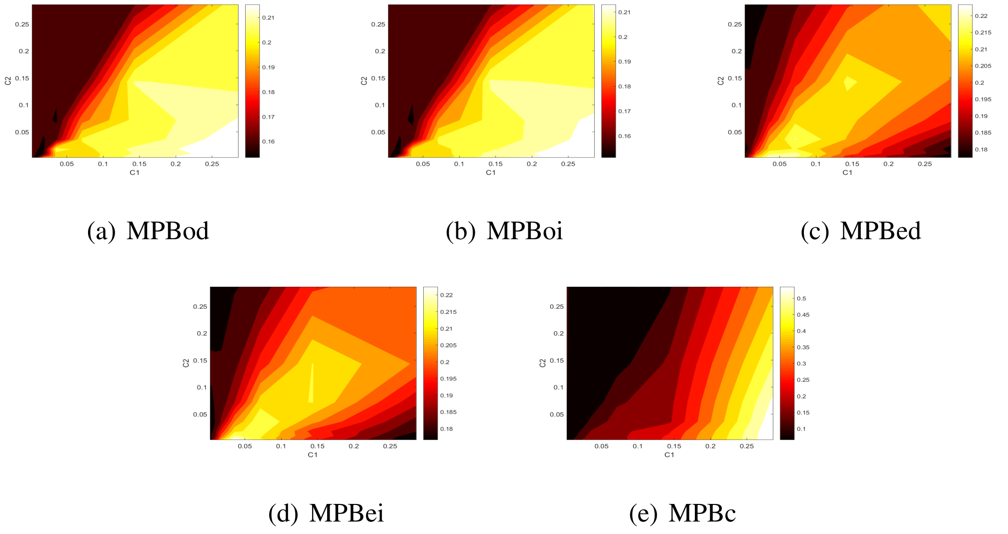
Table 3: Averages (%) and standard deviations (%) of multi-view and single-view stability-based PAC-Bayes bounds over expected Gibbs errors. SPB-$*$ are the single-view bounds over the SVM-$*$ Gibbs classifiers. MPB-$*$ and MPBc-$*$ correspond to directly adapted and view-consistency multi-view bounds over the MvSVM-$*$ Gibbs classifiers, respectively. The tightest bounds are shown in bold.

Bound	Ads	Course	Hand	Syn	Cora	Wis	Attack	Ionos	Breast
SPB-v1	41.16	41.30	50.34	55.52	52.07	64.29	52.89	55.37	59.47
	(±0.00)	(±0.00)	(±0.30)	(±0.00)	(±0.00)	(±0.00)	(±0.14)	(±0.17)	(±0.00)
MPB-v1	37.36	35.36	57.77	58.13	54.59	70.71	60.63	55.23	68.17
	(±0.10)	(±0.00)	(±0.00)	(±0.19)	(±0.00)	(±0.00)	(±0.00)	(±0.00)	(±0.00)
MPBc-v1	**22.28**	**22.78**	**21.43**	**46.56**	**28.99**	**40.64**	**30.36**	**31.85**	**47.66**
	(±0.11)	(±0.15)	(±0.27)	(±0.20)	(±0.82)	(±0.17)	(±0.31)	(±0.56)	(±0.12)
SPB-v2	42.62	39.29	37.84	55.09	46.22	64.24	57.04	60.00	59.45
	(±0.17)	(±0.00)	(±0.26)	(±0.00)	(±0.00)	(±0.00)	(±0.00)	(±0.18)	(±0.00)
MPB-v2	41.43	34.48	56.41	58.07	43.47	65.12	62.17	62.88	68.52
	(±0.11)	(±0.00)	(±0.20)	(±0.00)	(±0.00)	(±0.00)	(±0.00)	(±0.00)	(±0.00)
MPBc-v2	**25.47**	**20.85**	**16.16**	**45.69**	**26.84**	**36.40**	**40.32**	**36.07**	**47.03**
	(±0.00)	(±0.28)	(±0.29)	(±0.12)	(±0.17)	(±0.64)	(±0.36)	(±0.49)	(±0.31)
SPB-v3	46.46	41.98	45.54	55.26	52.29	64.29	55.93	59.68	59.44
	(±0.22)	(±0.00)	(±0.36)	(±0.00)	(±0.00)	(±0.00)	(±0.00)	(±0.00)	(±0.00)
MPB-v3	45.66	35.94	53.09	58.91	56.10	71.39	64.30	64.59	68.87
	(±0.21)	(±0.00)	(±0.43)	(±0.00)	(±0.12)	(±0.00)	(±0.00)	(±0.11)	(±0.00)
MPBc-v3	**26.75**	**16.00**	**12.35**	**36.79**	**31.67**	**45.03**	**34.11**	**35.46**	**46.92**
	(±0.12)	(±0.31)	(±0.32)	(±0.00)	(±0.17)	(±0.15)	(±0.16)	(±0.42)	(±0.16)

at origin $\mathcal{N}(\mathbf{0}, \mathbf{I})$, which are introduced in Theorem 3 and Theorem 4. MPBed and MPBei are separately dimension dependent and independent non-stability multi-view bounds where the priors of isotropic Gaussians centered at expected outputs of the algorithms $\mathcal{N}(\eta \boldsymbol{\omega}_p, \mathbf{I})$ where $\boldsymbol{\omega}_p = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}]$, which are introduced in Theorem 5 and Theorem 6. MPBc is our proposed stability-based bound.

Tightness To explicitly illustrate the tightness of various multi-view bounds, we present the differences between these bounds and their corresponding expected Gibbs errors estimated on test data over different hyper-parameters C_1, C_2 on Ionos, Hand, Syn, and Breast datasets which are separately shown in Figure 1, Figure 2, Figure 3, and Figure 4. All PAC-Bayes bounds upper-bound the MvSVM-v3 algorithm. There is no doubt that the discrepancies between the MPBod and MPBoi bounds are subtle since they are derived in the same way though with different inequalities. This is also true for the MPBed and MPBei bounds.

From these figures, the obvious difference among the non-stability bounds and the stability-based bound is that the MPBc bound is sensitive to the value of C_1 and it becomes loose for large values of C_1. Meanwhile, the MPBc bound is slightly sensitive to C_2 where the bound becomes tighter for larger values of C_2. This relationship becomes more obvious when C_1 is small, which is demonstrated by Figure 1, Figure 3, and Figure 4. This is in line with our expectations. The dependency of the non-stability multi-view bounds on C_1 and C_2 is irregular since it varies in different ways on different datasets. For example, the MPBod and the MPBoi bounds achieve better performance when C_1 and C_2 take smaller values on the Hand dataset, while for the Syn dataset, they become tighter when C_1 and C_2 are set to larger values. The MPBed and MPBei bounds perform better with larger



(a) MPBod

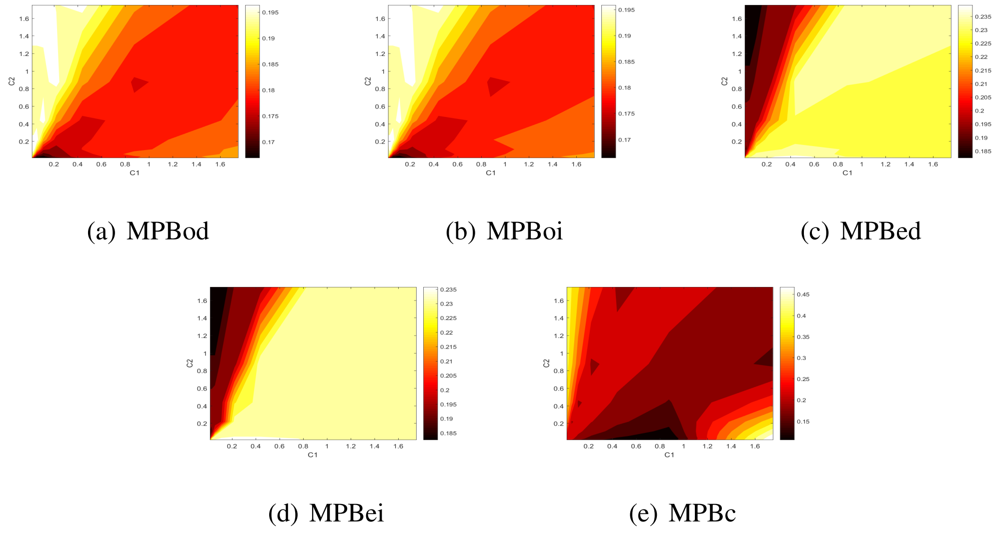
(b) MPBoi

(c) MPBed

(d) MPBei

(e) MPBc

Figure 1: Tightness of multi-view bounds on Ionos dataset shown as the difference between these bounds and the corresponding expected Gibbs errors. Smaller values are preferred.



(a) MPBod

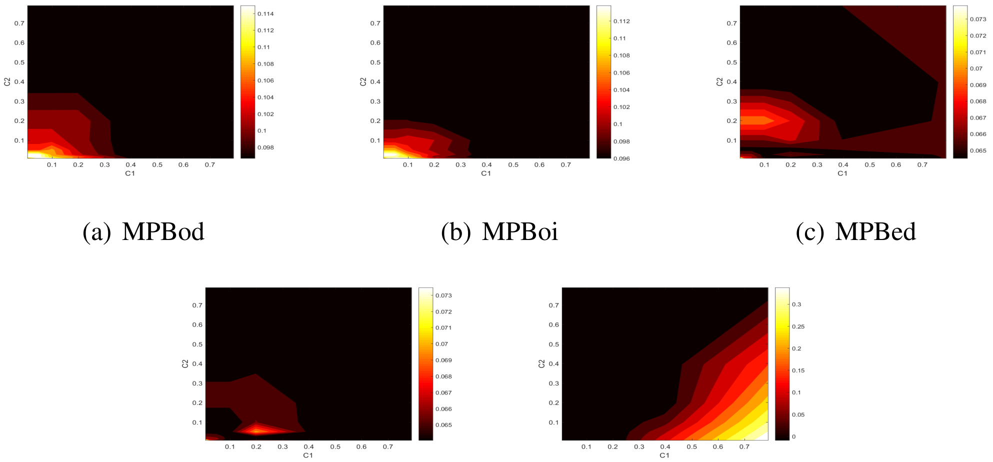
(b) MPBoi

(c) MPBed

(d) MPBei

(e) MPBc

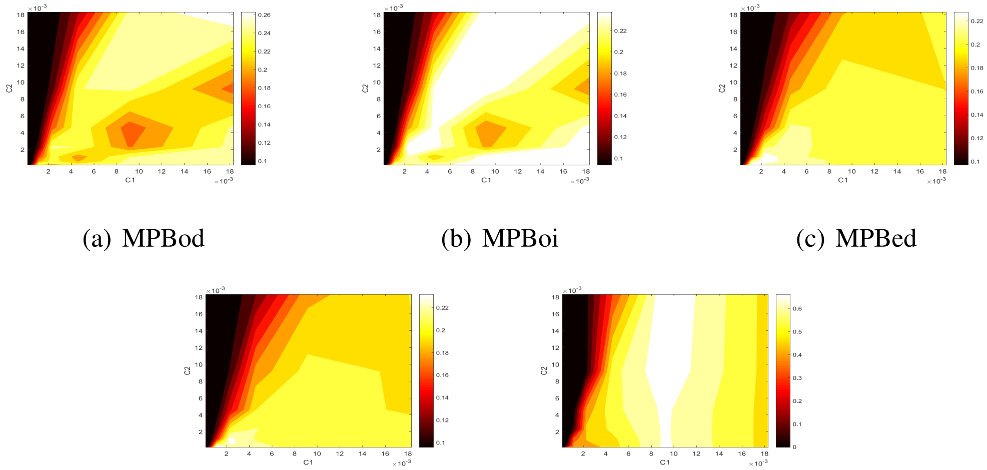
Figure 2: Tightness of multi-view bounds on Hand dataset shown as the difference between these bounds and the corresponding expected Gibbs errors. Smaller values are preferred.



(a) MPBod (b) MPBoi (c) MPBed

(d) MPBei (e) MPBc

Figure 3: Tightness of multi-view bounds on Syn dataset shown as the difference between these bounds and the corresponding expected Gibbs errors. Smaller values are preferred.



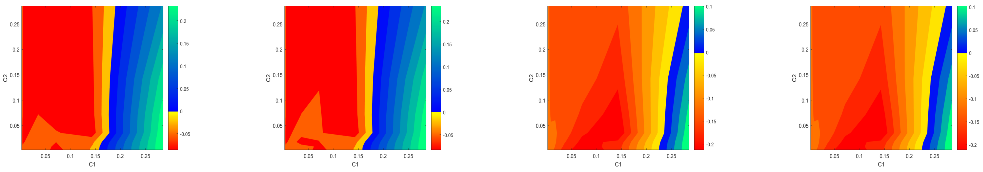
(a) MPBod (b) MPBoi (c) MPBed

(d) MPBei (e) MPBc

Figure 4: Tightness of multi-view bounds on Breast dataset shown as the difference between these bounds and the corresponding expected Gibbs errors. Smaller values are preferred.

C_1 and smaller C_2 on the Ionos dataset, while they become looser on the Breast dataset in the same condition.

We also present advantages and disadvantages of our MBPc bound over the previous non-stability bounds on Ionos, Hand, Syn, and Breast datasets which are shown in Figure 5, Figure 6, Figure 7, and Figure 8, respectively. The stability-based MPBc bound obviously performs better than other non-stability bounds when C_1 takes smaller values on the four datasets. While C_1 is set to larger values, the MPBod and MPBoi bounds achieve comparatively better performance on the four datasets. Figure 5 and Figure 7 also illustrate the better performance of the MPBc bound over other bounds with larger values of C_2. In a nutshell, our proposed MPBc bound is superior to previous non-stability bounds when the hyper-parameter C_1 is within the smaller range or C_2 is in the larger range, that is, when the algorithm is strongly regularized or the outputs of two view tend to agree.



(a) MPBc vs. MPBod

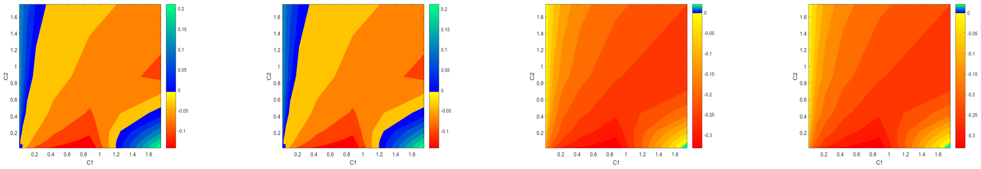
(b) MPBc vs. MPBoi

(c) MPBc vs. MPBed

(d) MPBc vs. MPBei

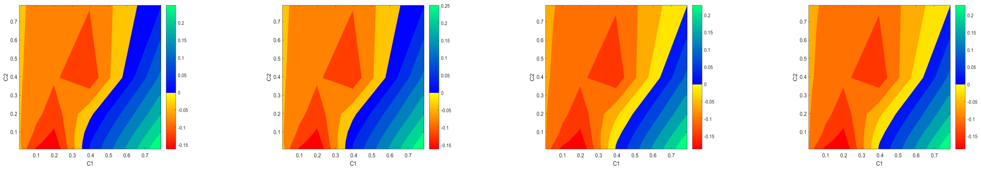
Figure 5: Differences among the MPBc bound and the MPBod, MPBoi, MPBed, MPBei bounds over expected Gibbs errors on Ionos dataset. The MPBc bound is preferred when differences are negative and it performs worse than other bounds when differences are positive.

Model Selection It is worth comparing correlations between the bounds and their corresponding expected Gibbs errors estimated on test data in order to illus-



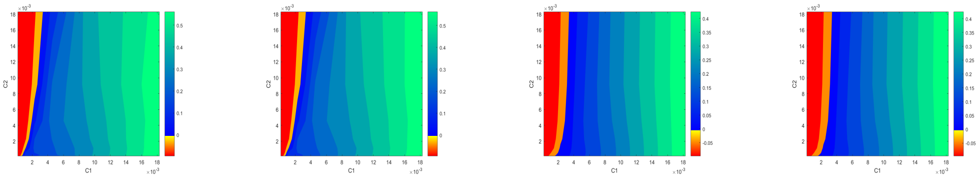
(a) MPBc vs. MPBod (b) MPBc vs. MPBoi (c) MPBc vs. MPBed (d) MPBc vs. MPBei

Figure 6: Differences among the MPBc bound and the MPBod, MPBoi, MPBed, MPBei bounds over expected Gibbs errors on Hand dataset. The MPBc bound is preferred when differences are negative and it performs worse than other bounds when differences are positive.



(a) MPBc vs. MPBod (b) MPBc vs. MPBoi (c) MPBc vs. MPBed (d) MPBc vs. MPBei

Figure 7: Differences among the MPBc bound and the MPBod, MPBoi, MPBed, MPBei bounds over expected Gibbs errors on Syn dataset. The MPBc bound is preferred when differences are negative and it performs worse than other bounds when differences are positive.

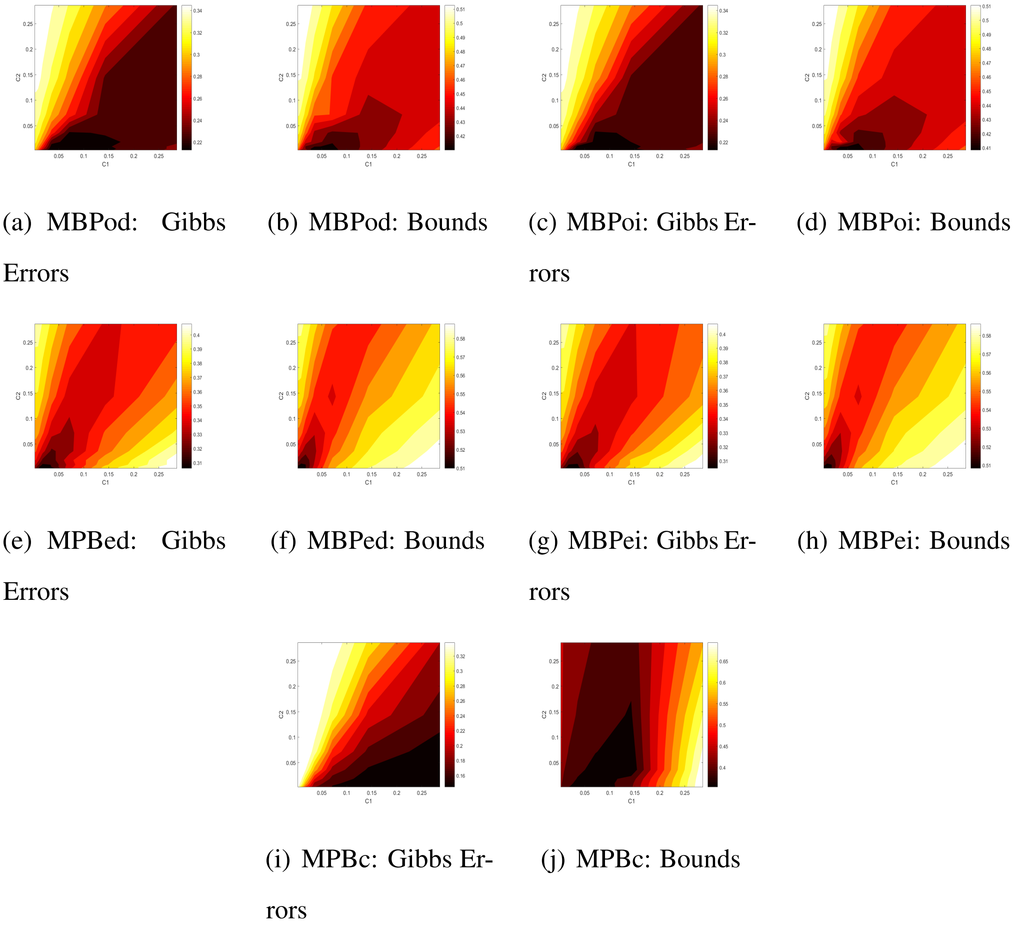


(a) MPBc vs. MPBod (b) MPBc vs. MPBoi (c) MPBc vs. MPBed (d) MPBc vs. MPBei

Figure 8: Differences among the MPBc bound and the MPBod, MPBoi, MPBed, MPBei bounds over expected Gibbs errors on Breast dataset. The MPBc bound is preferred when differences are negative and it performs worse than other bounds when differences are positive.

trate their ability for model selection. Figure 9, Figure 10, Figure 11, and Figure 12 show the correlations over different hyper-parameters C_1, C_2 on Ionos, Hand, Syn, and Breast datasets, respectively. All the PAC-Bayes bounds are calculated on the MvSVM-v3 algorithm. The MPBod and MPBoi bounds (MPBed and MPBei bounds) perform similarly again.

From these figures, the behaviors of the non-stability bounds follow more closely the behaviors of the corresponding Gibbs errors surfaces, while the MPBc bounds tend to perform consistently with their Gibbs errors for small values of C_1 or for large values of C_2. However, our MPBc bounds are able to select better values for C_1 and C_2 which lead to smaller Gibbs errors. For example, on the Ionos dataset, the MPBc bound identifies classifiers with smaller Gibbs errors around 0.16 compared with the Gibbs errors of 0.22 corresponding to the MPBod and MPBoi bounds and the Gibbs errors of 0.31 corresponding to the MPBed and MPBei bounds.



(a) MBPod: Gibbs Errors

(b) MBPod: Bounds

(c) MBPoi: Gibbs Errors

(d) MBPoi: Bounds

(e) MPBed: Gibbs Errors

(f) MBPed: Bounds

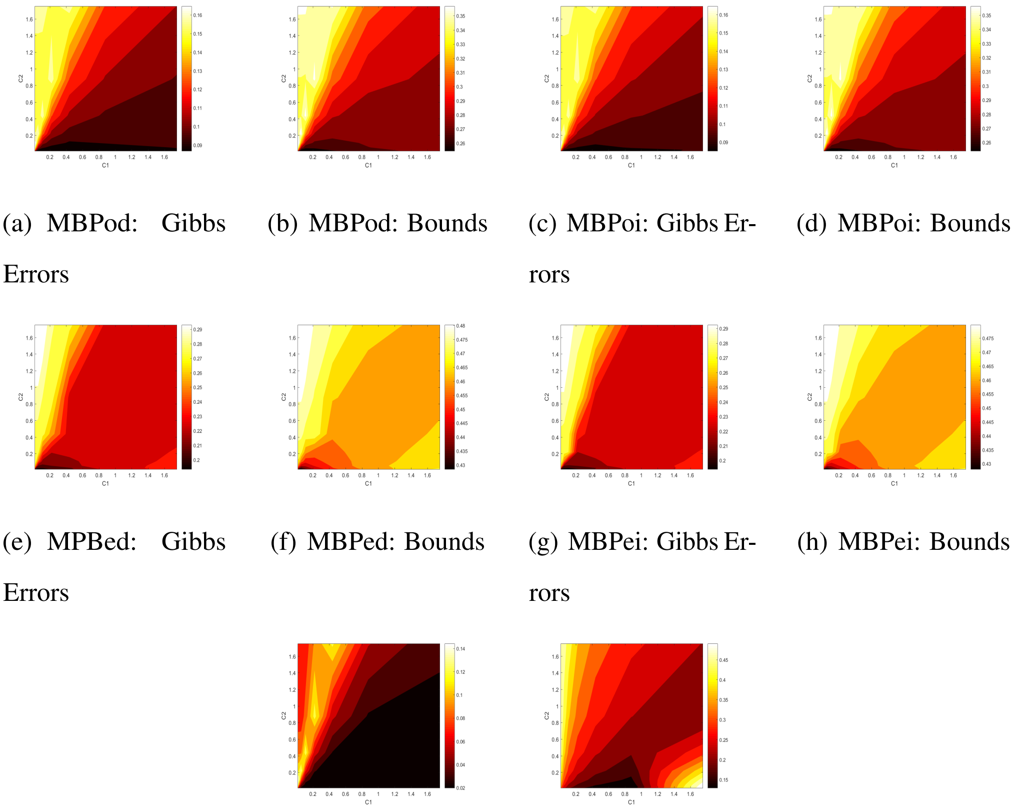
(g) MBPei: Gibbs Errors

(h) MBPei: Bounds

(i) MPBc: Gibbs Errors

(j) MPBc: Bounds

Figure 9: Ability to support model selection of various multi-view bounds on Ionos dataset shown as the bounds and the corresponding expected Gibbs errors in pairs.



(a) MBPod: Gibbs Errors

(b) MBPod: Bounds

(c) MBPoi: Gibbs Errors

(d) MBPoi: Bounds

(e) MPBed: Gibbs Errors

(f) MBPed: Bounds

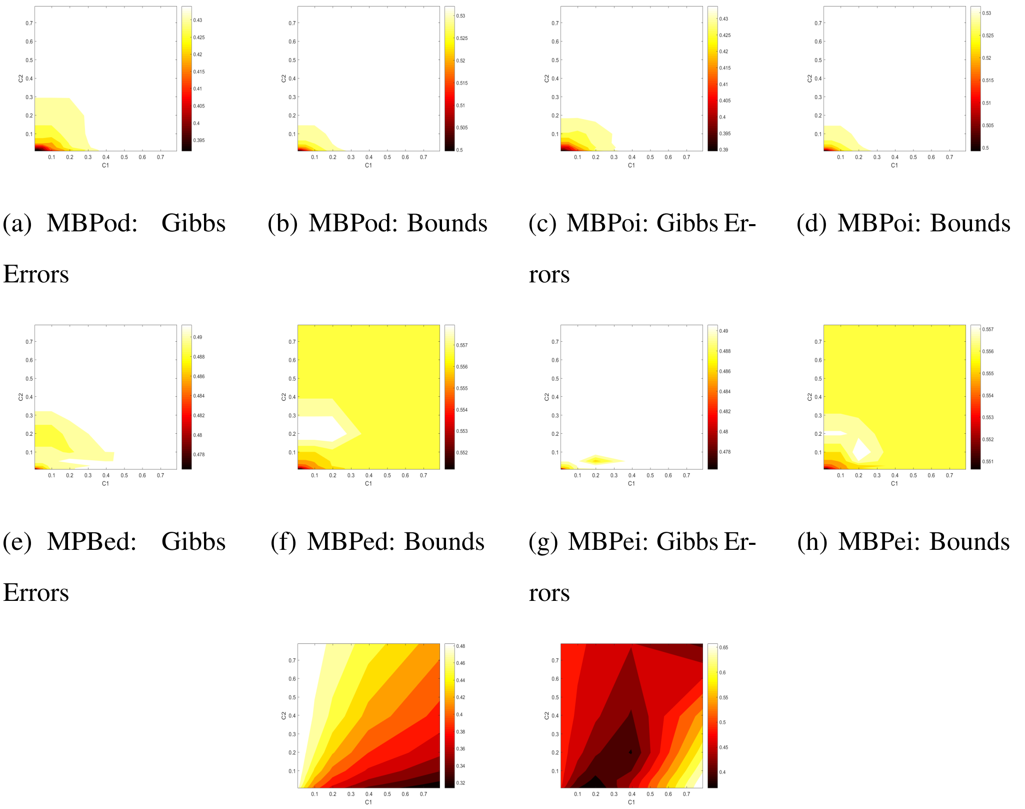
(g) MBPei: Gibbs Errors

(h) MBPei: Bounds

(i) MPBc: Gibbs Errors

(j) MPBc: Bounds

Figure 10: Ability to support model selection of various multi-view bounds on Hand dataset shown as the bounds and the corresponding expected Gibbs errors in pairs.



(a) MBPod: Gibbs Errors

(b) MBPod: Bounds

(c) MBPoi: Gibbs Errors

(d) MBPoi: Bounds

(e) MPBed: Gibbs Errors

(f) MBPed: Bounds

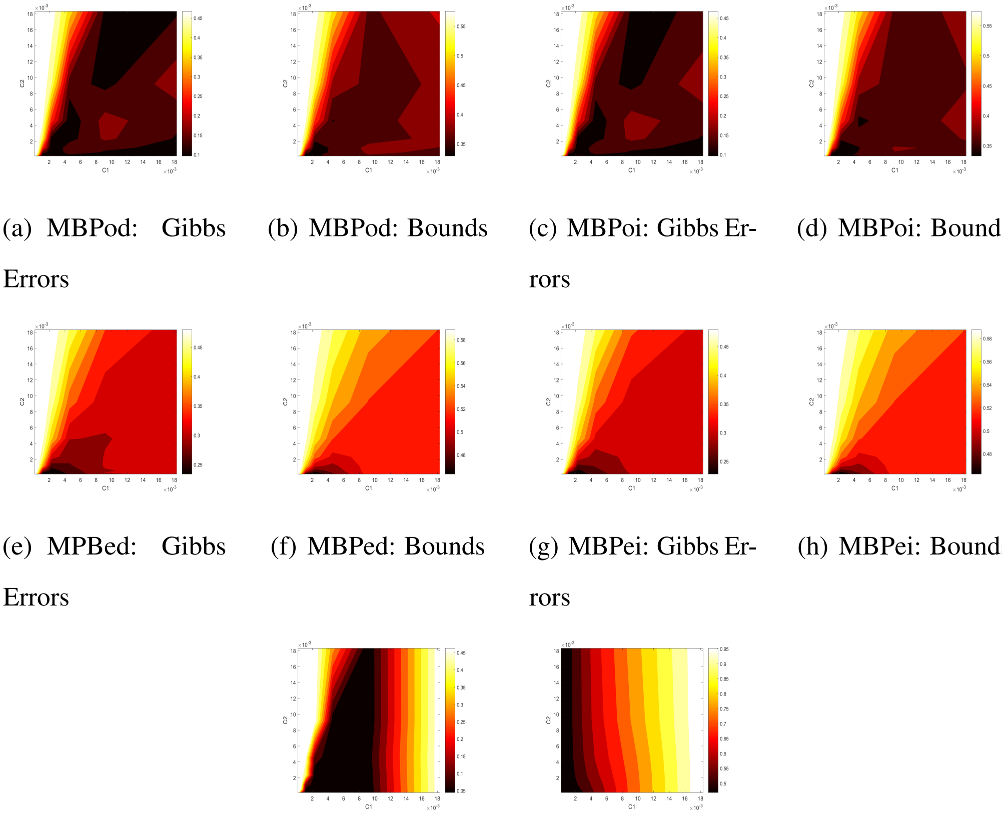
(g) MBPei: Gibbs Errors

(h) MBPei: Bounds

(i) MPBc: Gibbs Errors

(j) MPBc: Bounds

Figure 11: Ability to support model selection of various multi-view bounds on Syn dataset shown as the bounds and the corresponding expected Gibbs errors in pairs.



(a) MBPod: Gibbs Errors

(b) MBPod: Bounds

(c) MBPoi: Gibbs Errors

(d) MBPoi: Bounds

(e) MPBed: Gibbs Errors

(f) MBPed: Bounds

(g) MBPei: Gibbs Errors

(h) MBPei: Bounds

(i) MPBc: Gibbs Errors

(j) MPBc: Bounds

Figure 12: Ability to support model selection of various multi-view bounds on Breast dataset shown as the bounds and the corresponding expected Gibbs errors in pairs.

6. Conclusion

In this paper, we have proposed a stability-based multi-view PAC-Bayes bound with a novel view-consistency regularization. To employ the stability-based bound in experiments, we have derived an upper bound on the stability coefficient of the MvSVM algorithm. Experimental results have provided strong evidence to demonstrate the superiority of multi-view learning over single-view learning and have validated the advantages and disadvantages of the proposed stability-based bound over the previous non-stability multi-view bounds. Furthermore, the designed view-consistency regularization function has been shown to be beneficial to tightening the PAC-Bayes bound. In the future, it would be interesting to exploit the view-consistency regularization and the proposed bound to motivate new multi-view algorithms.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Project 62076096, Shanghai Municipal Project 20511100900, and Shanghai Knowledge Service Platform Project (No. ZF1213). The authors also acknowledge support of the UK Defence Science and Technology Laboratory (DSTL) and Engineering and Physical Research Council (EPSRC) under grant number EP/R013616/1. This is part of the collaboration between US DOD, UK MOD and UK EPSRC under the Multidisciplinary University Research Initiative.

A. Derivation of the Prior Distribution

Consider that the prior distributions of weight vectors from the two views are $P_1(\mathbf{u}_1) = \mathcal{N}(\mathbb{E}[\mathbf{u}_1], \sigma_1^2\mathbf{I})$ and $P_2(\mathbf{u}_2) = \mathcal{N}(\mathbb{E}[\mathbf{u}_2], \sigma_1^2\mathbf{I})$. We derive the prior distribution of the weight vector on the concatenated space as follows

$$\begin{aligned}
&P(\mathbf{u}) \propto P_1(\mathbf{u}_1)P_2(\mathbf{u}_2)V(\mathbf{u}_1, \mathbf{u}_2)\\
&\propto \mathcal{N}\left(\mathbb{E}[\mathbf{u}_1], \sigma_1^2\mathbf{I}\right) \cdot \mathcal{N}\left(\mathbb{E}[\mathbf{u}_2], \sigma_1^2\mathbf{I}\right) \cdot V\left(\mathbf{u}_1, \mathbf{u}_2\right)\\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{u}_1 - \mathbb{E}[\mathbf{u}_1])^\mathsf{T}\frac{1}{\sigma_1^2}(\mathbf{u}_1 - \mathbb{E}[\mathbf{u}_1])\right\}\exp\left\{-\frac{1}{2}(\mathbf{u}_2 - \mathbb{E}[\mathbf{u}_2])^\mathsf{T}\frac{1}{\sigma_1^2}(\mathbf{u}_2 - \mathbb{E}[\mathbf{u}_2])\right\}\\
&\quad \exp\left\{-\frac{1}{2\sigma_2^2}\mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)}\left[\mathbf{x}_1^\mathsf{T}(\mathbf{u}_1 - \mathbb{E}[\mathbf{u}_1]) - \mathbf{x}_2^\mathsf{T}(\mathbf{u}_2 - \mathbb{E}[\mathbf{u}_2])\right]^2\right\}\\
&= \exp\left\{-\frac{1}{2}(\mathbf{u} - \mathbb{E}[\mathbf{u}])^\mathsf{T}\frac{1}{\sigma_1^2}(\mathbf{u} - \mathbb{E}[\mathbf{u}])\right\}\exp\left\{-\frac{1}{2\sigma_2^2}\mathbb{E}_{\tilde{\mathbf{x}}}\left[\tilde{\mathbf{x}}^\mathsf{T}(\mathbf{u} - \mathbb{E}[\mathbf{u}])\right]^2\right\}\\
&= \exp\left\{-\frac{1}{2}(\mathbf{u} - \mathbb{E}[\mathbf{u}])^\mathsf{T}\frac{1}{\sigma_1^2}(\mathbf{u} - \mathbb{E}[\mathbf{u}])\right\}\exp\left\{-\frac{1}{2\sigma_2^2}\mathbb{E}_{\tilde{\mathbf{x}}}\left[(\mathbf{u} - \mathbb{E}[\mathbf{u}])^\mathsf{T}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}(\mathbf{u} - \mathbb{E}[\mathbf{u}])\right]\right\}\\
&= \exp\left\{-\frac{1}{2}(\mathbf{u} - \mathbb{E}[\mathbf{u}])^\mathsf{T}\frac{1}{\sigma_1^2}(\mathbf{u} - \mathbb{E}[\mathbf{u}])\right\}\exp\left\{-\frac{1}{2\sigma_2^2}(\mathbf{u} - \mathbb{E}[\mathbf{u}])^\mathsf{T}\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}\right](\mathbf{u} - \mathbb{E}[\mathbf{u}])\right\}\\
&= \exp\left\{-\frac{1}{2}(\mathbf{u} - \mathbb{E}[\mathbf{u}])^\mathsf{T}\left(\frac{\mathbf{I}}{\sigma_1^2} + \frac{\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}\right]}{\sigma_2^2}\right)(\mathbf{u} - \mathbb{E}[\mathbf{u}])\right\}.
\end{aligned} \tag{49}$$

B. Derivation of the KL-Divergence Between the Prior and Posterior Distributions

Consider a given prior distribution $P(\mathbf{u}) = \mathcal{N}(\mathbb{E}[\mathbf{u}], \Sigma)$ with $\Sigma = \left(\frac{\mathbf{I}}{\sigma_1^2} + \frac{\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}\right]}{\sigma_2^2}\right)^{-1}$ and a posterior distribution $Q(\mathbf{u}) = \mathcal{N}(\mathbf{u}, \mathbf{I})$, the KL divergence between them is derived as follows

$$\begin{aligned}
&\mathrm{KL}(Q(\mathbf{u})||P(\mathbf{u})) \\
&= \frac{1}{2}\left[-\ln\left(\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}\right]}{\sigma_2^2}\right|\right)+\mathrm{tr}\left(\frac{\mathbf{I}}{\sigma_1^2}+\frac{\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}\right]}{\sigma_2^2}\right)\right] \\
&\quad+\frac{1}{2}\left[(\mathbb{E}[\mathbf{u}]-\mathbf{u})^{\mathsf{T}}\left(\frac{\mathbf{I}}{\sigma_1^2}+\frac{\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}\right]}{\sigma_2^2}\right)(\mathbb{E}[\mathbf{u}]-\mathbf{u})-n\right] \\
&= \frac{1}{2}\left[-\ln\left(\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}\right]}{\sigma_2^2}\right|\right)+\mathrm{tr}\left(\frac{\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}\right]}{\sigma_2^2}\right)\right] \\
&\quad+\frac{1}{2}\left[\frac{1}{\sigma_1^2}[\mathbb{E}[\mathbf{u}]-\mathbf{u}]^2+(\mathbb{E}[\mathbf{u}]-\mathbf{u})^{\mathsf{T}}\left(\frac{\mathbb{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}]}{\sigma_2^2}\right)(\mathbb{E}[\mathbf{u}]-\mathbf{u})+\frac{n}{\sigma_1^2}-n\right] \\
&= \frac{1}{2}\left[-\ln\left(\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}\right]}{\sigma_2^2}\right|\right)+\frac{1}{\sigma_2^2}\mathbb{E}\left[\tilde{\mathbf{x}}^{\mathsf{T}}\tilde{\mathbf{x}}\right]+\frac{1}{\sigma_1^2}[\mathbb{E}[\mathbf{u}]-\mathbf{u}]^2\right] \\
&\quad+\frac{1}{2}\left[\frac{1}{\sigma_2^2}\mathbb{E}\left[\left(\tilde{\mathbf{x}}^{\mathsf{T}}(\mathbb{E}[\mathbf{u}]-\mathbf{u})\right)^2\right]+\frac{n}{\sigma_1^2}-n\right] \\
&= \frac{1}{2}\left[-\ln\left(\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\mathbb{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}]}{\sigma_2^2}\right|\right)+\frac{1}{\sigma_2^2}\mathbb{E}\left[\tilde{\mathbf{x}}^{\mathsf{T}}\tilde{\mathbf{x}}+\left(\tilde{\mathbf{x}}^{\mathsf{T}}(\mathbb{E}[\mathbf{u}]-\mathbf{u})\right)^2\right]\right] \\
&\quad+\frac{1}{2}\left[\frac{1}{\sigma_1^2}[\mathbb{E}[\mathbf{u}]-\mathbf{u}]^2+\frac{n}{\sigma_1^2}-n\right] \\
&= \frac{1}{2}\left[-\ln\left(\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\mathbb{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}]}{\sigma_2^2}\right|\right)+\frac{1}{\sigma_2^2}\mathbb{E}\left[\tilde{\mathbf{x}}^{\mathsf{T}}\tilde{\mathbf{x}}+(\mathbb{E}[\mathbf{u}]-\mathbf{u})^{\mathsf{T}}\left(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}\right)(\mathbb{E}[\mathbf{u}]-\mathbf{u}))\right]\right] \\
&\quad+\frac{1}{2}\left[\frac{1}{\sigma_1^2}(\mathbb{E}[\mathbf{u}]-\mathbf{u})^{\mathsf{T}}(\mathbb{E}[\mathbf{u}]-\mathbf{u})+\frac{n}{\sigma_1^2}-n\right] \\
&= \frac{1}{2}\left[-\ln\left(\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\mathbb{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}]}{\sigma_2^2}\right|\right)+\frac{1}{\sigma_2^2}\mathbb{E}\left[\tilde{\mathbf{x}}^{\mathsf{T}}\tilde{\mathbf{x}}\right]\right] \\
&\quad+\frac{1}{2}\left[(\mathbb{E}[\mathbf{u}]-\mathbf{u})^{\mathsf{T}}\mathbb{E}\left[\frac{\mathbf{I}}{\sigma_1^2}+\frac{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}}{\sigma_2^2}\right](\mathbb{E}[\mathbf{u}]-\mathbf{u})+\frac{n}{\sigma_1^2}-n\right].
\end{aligned} \tag{50}$$

C. Derivation of Theorem 7

We upper-bound the KL divergence between the prior and the posterior distributions of a multi-view learning algorithm which is presented in Equation (25).

With Jensen's inequality we have

$$-\ln\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\mathbb{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top]}{\sigma_2^2}\right| = -n\ln\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\mathbb{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top]}{\sigma_2^2}\right|^{\frac{1}{n}} \leq -n\ln\mathbb{E}\left[\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top}{\sigma_2^2}\right|^{\frac{1}{n}}\right]. \tag{51}$$

Define $h_m = h\left(\tilde{\mathbf{x}}^{(1)},\cdots,\tilde{\mathbf{x}}^{(i)},\cdots,\tilde{\mathbf{x}}^{(m)}\right) = \frac{1}{m}\sum_{i=1}^m\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\tilde{\mathbf{x}}^{(i)}\tilde{\mathbf{x}}^{(i)\top}}{\sigma_2^2}\right|^{\frac{1}{n}}$ where m is the number of examples. Since the rank of matrix $\frac{\tilde{\mathbf{x}}^{(i)}\tilde{\mathbf{x}}^{(i)\top}}{\sigma_2^2}$ is 1 with the nonzero eigenvalue being $\frac{||\tilde{\mathbf{x}}^{(i)}||^2}{\sigma_2^2}$ and the determinant of a positive semi-definite matrix is equal to the product of its eigenvalues, the following inequality holds

$$\begin{aligned}
&\sup_{\tilde{\mathbf{x}}^{(1)},\cdots,\tilde{\mathbf{x}}^{(m)},\bar{\mathbf{x}}^{(i)}}\left|h\left(\tilde{\mathbf{x}}^{(1)},\cdots,\tilde{\mathbf{x}}^{(i)},\cdots,\tilde{\mathbf{x}}^{(m)}\right)-h\left(\tilde{\mathbf{x}}^{(1)},\cdots,\bar{\mathbf{x}}^{(i)},\cdots,\tilde{\mathbf{x}}^{(m)}\right)\right|\\
&=\frac{1}{m}\left|\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\tilde{\mathbf{x}}^{(i)}\tilde{\mathbf{x}}^{(i)\top}}{\sigma_2^2}\right|^{\frac{1}{n}}-\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\bar{\mathbf{x}}^{(i)}\bar{\mathbf{x}}^{(i)\top}}{\sigma_2^2}\right|^{\frac{1}{n}}\right| \qquad (52)\\
&\leq\frac{1}{m}\left(\sqrt[n]{(R/\sigma_2)^2+\frac{1}{\sigma_1^2}}-\frac{1}{\sigma_1^2}\right),
\end{aligned}$$

where $R=\sup_{\tilde{\mathbf{x}}}||\tilde{\mathbf{x}}||$. By McDiarmid's inequality, we have for all $\epsilon>0$,

$$P\left\{\mathbb{E}\left[\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top}{\sigma_2^2}\right|^{\frac{1}{n}}\right]\geq h_m-\epsilon\right\}\geq 1-\exp\left(\frac{-2m\epsilon^2}{\left(\sqrt[n]{(R/\sigma_2)^2+\frac{1}{\sigma_1^2}}-\frac{1}{\sigma_1^2}\right)^2}\right). \tag{53}$$

Let the right hand side be $1-\frac{\delta}{3}$, then $\epsilon=\left(\sqrt[n]{(R/\sigma_2)^2+\frac{1}{\sigma_1^2}}-\frac{1}{\sigma_1^2}\right)\sqrt{\frac{1}{2m}\ln\frac{3}{\delta}}$. Hence, we have with probability at least $1-\frac{\delta}{3}$,

$$\mathbb{E}\left[\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top}{\sigma_2^2}\right|^{\frac{1}{n}}\right]\geq h_m-\left(\sqrt[n]{(R/\sigma_2)^2+\frac{1}{\sigma_1^2}}-\frac{1}{\sigma_1^2}\right)\sqrt{\frac{1}{2m}\ln\frac{3}{\delta}}. \tag{54}$$

Therefore, we get the following inequality

$$\begin{aligned}
&-\ln\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\mathbb{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}]}{\sigma_2^2}\right| \le -n\ln\mathbb{E}\left[\left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}}{\sigma_2^2}\right|^{\frac{1}{n}}\right] \\
&\le -n\ln\left[h_m-\left(\sqrt[n]{(R/\sigma_2)^2+\frac{1}{\sigma_1^2}}-\frac{1}{\sigma_1^2}\right)\sqrt{\frac{1}{2m}\ln\frac{3}{\delta}}\quad\right]_+,
\end{aligned} \tag{55}$$

where $[\cdot]_+ = \max(\cdot, 0)$.

With $R = \sup_{\tilde{\mathbf{x}}} ||\tilde{\mathbf{x}}||$, we have

$$\mathbb{E}\left[\tilde{\mathbf{x}}^\mathsf{T}\tilde{\mathbf{x}}\right] \le R^2. \tag{56}$$

Since the equation $\left(\mathbf{I}+\frac{\sigma_1^2}{\sigma_2^2}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}\right)\tilde{\mathbf{x}} = \tilde{\mathbf{x}}+\frac{\sigma_1^2}{\sigma_2^2}\tilde{\mathbf{x}}||\tilde{\mathbf{x}}||^2 = \left(1+\frac{\sigma_1^2}{\sigma_2^2}||\tilde{\mathbf{x}}||^2\right)\tilde{\mathbf{x}}$ holds, $1+\frac{\sigma_1^2}{\sigma_2^2}||\tilde{\mathbf{x}}||^2$ is an eigenvalue and $\tilde{\mathbf{x}}$ is an eigenvector. For other eigenvectors $\mathbf{w}$, there is $\left(\mathbf{I}+\frac{\sigma_1^2}{\sigma_2^2}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}\right)\mathbf{w} = \mathbf{w}$ because of $\mathbf{w}^\mathsf{T}\tilde{\mathbf{x}} = 0$. It indicates that the eigenvalues are 1 or $1+\frac{\sigma_1^2}{\sigma_2^2}||\tilde{\mathbf{x}}||^2$. Then the third term of the KL divergence can be upper-bounded

$$\begin{aligned}
&(\mathbb{E}[\mathbf{u}]-\mathbf{u})^\mathsf{T}\mathbb{E}\left[\frac{\mathbf{I}}{\sigma_1^2}+\frac{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}}{\sigma_2^2}\right](\mathbb{E}[\mathbf{u}]-\mathbf{u}) \\
&= \mathbb{E}\left[\frac{1}{\sigma_1^2}(\mathbb{E}[\mathbf{u}]-\mathbf{u})^\mathsf{T}\left(\mathbf{I}+\frac{\sigma_1^2}{\sigma_2^2}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}\right)(\mathbb{E}[\mathbf{u}]-\mathbf{u})\right] \\
&\le \frac{1}{\sigma_1^2}\mathbb{E}\left[\left(1+\frac{\sigma_1^2}{\sigma_2^2}||\tilde{\mathbf{x}}||^2\right)||\mathbb{E}[\mathbf{u}]-\mathbf{u}||^2\right] \\
&\le \left(\frac{1}{\sigma_1^2}+\frac{R^2}{\sigma_2^2}\right)||\mathbb{E}[\mathbf{u}]-\mathbf{u}||^2
\end{aligned} \tag{57}$$

According to the following inequality (see Corollary 8 of Rivasplata et al. [21]),

$$||\mathbb{E}[\mathbf{u}]-\mathbf{u}||^2 \le m\beta_m^2\left(1+\sqrt{\frac{1}{2}\ln\frac{3}{\delta}}\right)^2, \tag{58}$$

we have

$$(\mathbb{E}[\mathbf{u}]-\mathbf{u})^\mathsf{T}\mathbb{E}\left[\frac{\mathbf{I}}{\sigma_1^2}+\frac{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\mathsf{T}}{\sigma_2^2}\right](\mathbb{E}[\mathbf{u}]-\mathbf{u}) \le \left(\frac{1}{\sigma_1^2}+\frac{R^2}{\sigma_2^2}\right)m\beta_m^2\left(1+\sqrt{\frac{1}{2}\ln\frac{3}{\delta}}\right)^2. \tag{59}$$

With the basic PAC-Bayes theorem, we have with probability at least $1-\frac{\delta}{3}$,

$$P_{\mathcal{X}\sim\mathcal{D}^m}\left\{\mathrm{KL}_+(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}}) \leq \frac{\mathrm{KL}(Q||P)+\ln\left(\frac{m+1}{\frac{\delta}{3}}\right)}{m}\right\} \geq 1-\frac{\delta}{3}. \tag{60}$$

Combining the inequalities (55), (56), (59), (60), we therefore reach the stability-based multi-view PAC-Bayes bound as follows

$$\mathrm{KL}_+(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}}) \leq \frac{-\frac{n}{2}\ln\left[h_m-\left(\sqrt[n]{\left(\frac{R}{\sigma_2}\right)^2+\frac{1}{\sigma_1^2}}-\frac{1}{\sigma_1^2}\right)\sqrt{\frac{1}{2m}\ln\left(\frac{3}{\delta}\right)}\right]_+ + R^2}{m} \\ + \frac{\left(\frac{1}{\sigma_1^2}+\frac{R^2}{\sigma_2^2}\right)m\beta_m^2\left(1+\sqrt{\frac{1}{2}\ln\frac{3}{\delta}}\right)^2+\frac{n}{\sigma_1^2}-n+\ln\left(\frac{m+1}{\frac{\delta}{3}}\right)}{m}. \tag{61}$$

where $h_m = \frac{1}{m}\sum_{i=1}^m \left|\frac{\mathbf{I}}{\sigma_1^2}+\frac{\tilde{\mathbf{x}}^{(i)}\tilde{\mathbf{x}}^{(i)\top}}{\sigma_2^2}\right|^{\frac{1}{n}}$ and n is the dimensionality of the feature space.

D. Derivation of Theorem 8

Suppose the prior distribution is $P(\mathbf{u}) = \mathcal{N}(\mathbb{E}[\mathbf{u}], \sigma^2\mathbf{I})$ and the posterior distribution is $Q(\mathbf{u}) = \mathcal{N}(\mathbf{u}, \sigma^2\mathbf{I})$. For the directly adapted multi-view bound, the KL divergence between the prior and the posterior distributions of a learner is given as

$$\mathrm{KL}(Q(\mathbf{u})||P(\mathbf{u})) = \frac{1}{2\sigma^2}(\mathbb{E}[\mathbf{u}]-\mathbf{u})^2. \tag{62}$$

Applying Equation (58) and Equation (60), we can get the multi-view PAC-Bayes bound without the view-consistency regularization as follows

$$P_{\mathcal{X}\sim\mathcal{D}^m}\left\{\mathrm{KL}_+(E_{Q,\mathcal{X}}||E_{Q,\mathcal{D}}) \leq \frac{\frac{1}{2\sigma^2}m\beta_m^2\left(1+\sqrt{\frac{1}{2}\ln\left(\frac{2}{\delta}\right)}\right)^2+\ln\left(\frac{m+1}{\frac{\delta}{2}}\right)}{m}\right\} \geq 1-\delta. \tag{63}$$

E. Proof of Lemma 1

For a multi-view regularization algorithm, it aims to minimize the following objective function on the training set $\mathcal{X}$,

$$\mathcal{J}(\mathbf{u}) = \frac{1}{2m}\sum_{i=1}^{m} l\left(f_{\mathbf{u}}\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) + \lambda_1 N(\mathbf{u}) + \lambda_2 M(\mathbf{u}). \tag{64}$$

If an example $\mathbf{z}^{(i)}$ is replaced by another example $\mathbf{z}^{(j)}$, the objective function could be written as

$$\mathcal{J}^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}(\mathbf{u}) = \frac{1}{2m}\sum_{t=1,\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}^{m} l\left(f_{\mathbf{u}}\left(\mathbf{x}^{(t)}\right), y^{(t)}\right) + \lambda_1 N(\mathbf{u}) + \lambda_2 M(\mathbf{u}). \tag{65}$$

The original empirical error function and the empirical error function after substituting examples are defined by

$$R(\mathbf{u}) = \frac{1}{2m}\sum_{i=1}^{m} l\left(f_{\mathbf{u}}\left(\mathbf{x}^{(i)}\right), y^{(i)}\right), \tag{66}$$

$$R^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}(\mathbf{u}) = \frac{1}{2m}\sum_{t=1,\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}^{m} l\left(f_{\mathbf{u}}\left(\mathbf{x}^{(t)}\right), y^{(t)}\right). \tag{67}$$

Define $\mathbf{u}_*$ and $\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}$ as the minimizers of Equation (64) and Equation (65), respectively. When $\mathcal{J}$ is differentiable everywhere, the Bregman divergence associated to $\mathcal{J}$ of $\mathbf{v}$ to $\mathbf{u}$ is defined by

$$\forall \mathbf{v}, \mathbf{u} \in \mathcal{U}, d_{\mathcal{J}}(\mathbf{v}, \mathbf{u}) = \mathcal{J}(\mathbf{v}) - \mathcal{J}(\mathbf{u}) - \langle \mathbf{v} - \mathbf{u}, \bigtriangledown\mathcal{J}(\mathbf{u})\rangle, \tag{68}$$

where $\bigtriangledown\mathcal{J}(\mathbf{u})$ is a subgradient of $\mathcal{J}$ in $\mathbf{u}$. When $\mathbf{u}$ is a minimizer of $\mathcal{J}(\mathbf{u})$, the following equality holds

$$\forall \mathbf{v} \in \mathcal{U}, d_{\mathcal{J}}(\mathbf{v}, \mathbf{u}_*) = \mathcal{J}(\mathbf{v}) - \mathcal{J}(\mathbf{u}_*). \tag{69}$$

Employing Equation (69), we have

$$\begin{aligned}
& d_{\mathcal{J}}\left(\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}, \mathbf{u}_*\right) + d_{\mathcal{J}^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}}\left(\mathbf{u}_*, \mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}\right) \\
& = \mathcal{J}\left(\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}\right) - \mathcal{J}(\mathbf{u}_*) + \mathcal{J}^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}(\mathbf{u}_*) - \mathcal{J}^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}\left(\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}\right) \\
& = \frac{1}{2m} l\left(\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}\top}\mathbf{x}^{(i)}, y^{(i)}\right) - \frac{1}{2m} l\left(\mathbf{u}_*^\top\mathbf{x}^{(i)}, y^{(i)}\right) \\
& \quad + \frac{1}{2m} l\left(\mathbf{u}_*^\top\mathbf{x}^{(j)}, y^{(j)}\right) - \frac{1}{2m} l\left(\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}\top}\mathbf{x}^{(j)}, y^{(j)}\right).
\end{aligned} \tag{70}$$

By the nonnegativity of divergences, we have

$$d_R\left(\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}, \mathbf{u}_*\right) + d_{R^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}}\left(\mathbf{u}_*, \mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}\right) \geq 0. \tag{71}$$

Hence, the following inequality holds

$$\begin{aligned}
& \lambda_1 d_N\left(\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}, \mathbf{u}_*\right) + \lambda_2 d_M\left(\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}, \mathbf{u}_*\right) \\
& \quad + \lambda_1 d_N\left(\mathbf{u}_*, \mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}\right) + \lambda_2 d_M\left(\mathbf{u}_*, \mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}}\right) \\
& \leq \frac{1}{2m} l\left(\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}\top}\mathbf{x}^{(i)}, y^{(i)}\right) - \frac{1}{2m} l\left(\mathbf{u}_*^\top\mathbf{x}^{(i)}, y^{(i)}\right) \\
& \quad + \frac{1}{2m} l\left(\mathbf{u}_*^\top\mathbf{x}^{(j)}, y^{(j)}\right) - \frac{1}{2m} l\left(\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}\top}\mathbf{x}^{(j)}, y^{(j)}\right) \\
& \leq \frac{1}{2m}\left(l\left(\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}\top}\mathbf{x}^{(i)}, y^{(i)}\right) - l\left(\mathbf{u}_*^\top\mathbf{x}^{(i)}, y^{(i)}\right)\right) \\
& \quad + \frac{1}{2m}\left(l\left(\mathbf{u}_*^\top\mathbf{x}^{(j)}, y^{(j)}\right) - l\left(\mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}\top}\mathbf{x}^{(j)}, y^{(j)}\right)\right) \\
& \leq \frac{\alpha}{2m}\left(\left|\Delta\mathbf{u}_*^\top\mathbf{x}^{(i)}\right| + \left|\Delta\mathbf{u}_*^\top\mathbf{x}^{(j)}\right|\right),
\end{aligned} \tag{72}$$

where the last inequality holds since the loss function l is α-admissible and $\Delta\mathbf{u}_* = \mathbf{u}_*^{\mathbf{z}^{(i)}\to\mathbf{z}^{(j)}} - \mathbf{u}_*$. Therefore, we obtain Lemma 1.

References

[1] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: Recent progress and new challenges, Information Fusion 38 (2017) 43–54.

[2] M. Fratello, G. Caiazzo, F. Trojsi, A. Russo, G. Tedeschi, R. Tagliaferri, F. Esposito, Multi-view ensemble classification of brain connectivity images for neurodegeneration type discrimination, Neuroinformatics 15 (2017) 199–213.

[3] S. Bhadra, S. Kaski, J. Rousu, Multi-view kernel completion, Machine Learning 106 (2017) 713–739.

[4] C. Ma, Y. Guo, J. Yang, W. An, Learning multi-view representation with LSTM for 3D shape recognition and retrieval, IEEE Transactions on Multimedia 21 (2019) 1169–1182.

[5] G. Goh, K. Sakloth, C. Siegel, A. Vishnu, J. Pfaendtner, Multimodal deep neural networks using both engineered and learned representations for biodegradability prediction, ArXiv Preprint arXiv:1808.04456 (2018) 1–7.

[6] P. Yang, W. Gao, Multi-view discriminant transfer learning, in: International Joint Conference on Artificial Intelligence, 2013, pp. 1848–1854.

[7] X. Zhang, X. Zhang, H. Liu, Multi-task multi-view clustering for non-negative data, in: International Joint Conference on Artificial Intelligence, 2015, pp. 4055–4061.

[8] X. Xie, S. Sun, Multi-view Laplacian twin support vector machines, Applied Intelligence 41 (2014) 1059–1068.

[9] R. Huusari, H. Kadri, C. Capponi, Multi-view metric learning in vector-

valued kernel spaces, in: International Conference on Artificial Intelligence and Statistics, 2018, pp. 415–424.

[10] P. Germain, A. Lacasse, F. Laviolette, M. Marchand, J. Roy, Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm, Journal of Machine Learning Research 16 (2015) 787–860.

[11] D. McAllester, PAC-Bayesian stochastic model selection, Machine Learning 51 (2003) 5–21.

[12] M. Seeger, PAC-Bayesian generalisation error bounds for Gaussian process classification, Journal of Machine Learning Research 3 (2002) 233–269.

[13] J. Langford, Tutorial on practical prediction theory for classification, Journal of Machine Learning Research 6 (2005) 273–306.

[14] A. Ambroladze, E. Parrado-Hernández, J. Shawe-taylor, Tighter PAC-Bayes bounds, in: Advances in Neural Information Processing Systems, 2007, pp. 9–16.

[15] P. Germain, A. Lacasse, F. Laviolette, M. Marchand, PAC-Bayesian learning of linear classifiers, in: International Conference on Machine Learning, 2009, pp. 353–360.

[16] G. K. Dziugaite, K. Hsu, W. Gharbieh, G. Arpino, D. Roy, On the role of data in pac-bayes, in: International Conference on Artificial Intelligence and Statistics, 2021, pp. 604–612.

[17] O. Catoni, PAC-Bayesian supervised classification: The thermodynamics of statistical learning, ArXiv Preprint arXiv:0712.0248 (2007) 1–175.

[18] G. Lever, F. Laviolette, J. Shawe-Taylor, Distribution-dependent PAC-Bayes priors, in: International Conference on Algorithmic Learning Theory, 2010, pp. 119–133.

[19] G. Lever, F. Laviolette, J. Shawe-Taylor, Tighter PAC-Bayes bounds through distribution-dependent priors., Theoretical Computer Science 473 (2013) 4–28.

[20] G. K. Dziugaite, D. Roy, Data-dependent PAC-Bayes priors via differential privacy, in: Advances in Neural Information Processing Systems, 2018, pp. 8430–8441.

[21] O. Rivasplata, C. Szepesvari, J. Shawe-Taylor, E. Parrado-Hernandez, S. Sun, PAC-Bayes bounds for stable algorithms with instance-dependent priors, in: Advances in Neural Information Processing Systems, 2018, pp. 9214–9224.

[22] O. Bousquet, A. Elisseeff, Stability and generalization, Journal of Machine Learning Research 2 (2002) 499–526.

[23] K. Abou-Moustafa, C. Szepesvári, An a priori exponential tail bound for k-folds cross-validation, ArXiv Preprint arXiv:1706.05801 (2017) 1–18.

[24] T. Liu, G. Lugosi, G. Neu, D. Tao, Algorithmic stability and hypothesis complexity, in: International Conference on Machine Learning, 2017, pp. 2159–2167.

[25] S. Sun, J. Shawe-Taylor, L. Mao, PAC-Bayes analysis of multi-view learning, Information Fusion 35 (2017) 117–131.

[26] V. Sindhwani, P. Niyogi, M. Belkin, A co-regularization approach to semi-supervised learning with multiple views, in: International Conference on Machine Learning Workshop on Learning with Multiple Views, 2005, pp. 74–79.

[27] V. Sindhwani, D. Rosenberg, An RKHS for multi-view learning and manifold co-regularization, in: International Conference on Machine Learning, 2008, pp. 976–983.

[28] A. Goyal, E. Morvant, P. Germain, M.-R. Amini, PAC-Bayesian analysis for a two-step hierarchical multiview learning approach, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2017, pp. 205–221.

[29] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Conference on Computational Learning Theory, 1998, pp. 92–100.

[30] S. Sun, F. Jin, Robust co-training, International Journal of Pattern Recognition and Artificial Intelligence 25 (2011) 1113–1126.

[31] S. Sun, G. Chao, Multi-view maximum entropy discrimination, in: International Joint Conference on Artificial Intelligence, 2013, pp. 1706–1712.

[32] L. Mao, S. Sun, Soft margin consistency based scalable multi-view maximum

entropy discrimination., in: International Joint Conference on Artificial Intelligence, 2016, pp. 1839–1845.

[33] J. Farquhar, D. Hardoon, H. Meng, J. S. Shawe-Taylor, S. Szedmak, Two view learning: Svm-2k, theory and practice, in: Advances in Neural Information Processing Systems, 2006, pp. 355–362.

[34] X. Xie, S. Sun, Multi-view twin support vector machines, Intelligent Data Analysis 19 (2015) 701–712.

[35] J. Langford, J. Shawe-Taylor, PAC-Bayes & margins, in: Advances in Neural Information Processing Systems, 2002, pp. 439–446.

[36] J. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, S. Szedmak, Two view learning: SVM-2K, theory and practice, in: Advances in Neural Information Processing Systems, 2006, pp. 355–362.

[37] S. Szedmak, J. Shawe-Taylor, Synthesis of maximum margin and multiview learning using unlabeled data, Neurocomputing 70 (2007) 1254–1264.

[38] K. Sridharan, S. M. Kakade, An information theoretic framework for multi-view learning, in: Conference on Learning Theory, 2008.

[39] N. Kushmerick, Learning to remove internet advertisements, in: Agents, 1999, pp. 175–181.

[40] S. Sun, J. Shawe-Taylor, Sparse semi-supervised learning using conjugate functions, Journal of Machine Learning Research 11 (2010) 2423–2455.

[41] D. Dua, C. Graff, UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml.

[42] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, T. Eliassi-Rad, Collective classification in network data, AI Magazine 29 (2008) 93–93.

About the Author—SHILIANG SUN is a professor at the School of Computer Science and Technology and the head of the Pattern Recognition and Machine Learning Research Group, East China Normal University. He received the B.E. degree in automatic control from the Department of Automatic Control, Beijing University of Aeronautics and Astronautics in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the Department of Automation and the State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing, China, in 2007. In 2004, he was entitled Microsoft Fellow. From 2009 to 2010, he was a visiting researcher at the Department of Computer Science, University College London, working within the Center for Computational Statistics and Machine Learning. From March to April 2012, he was a visiting researcher at the Department of Statistics, Rutgers University. He is a member of the PASCAL (Pattern Analysis, Statistical Modelling, and Computational Learning) network of excellence, and on the editorial boards of multiple international journals. His research interests include multi-view learning, approximate inference, Gaussian process, sequential modeling, kernel methods, and their applications.

About the Author—MENGRAN YU received the B.S. degree from East China Normal University, Shanghai, China. She is currently pursuing the M.S. degree with the School of Computer Science and Technology, East China Normal University, Shanghai, China. Her current research interests include pattern recognition and adversarial machine learning.

About the Author—JOHN SHAWE-TAYLOR is a professor at University College London (UK) where he is Director of the Centre for Computational Statistics and Machine Learning (CSML). His main research area is Statistical Learning Theory, but his contributions range from Neural Networks, to Machine Learning, to Graph Theory. John Shawe-Taylor obtained a PhD in Mathematics at Royal Holloway, University of London in 1986. He subsequently completed an MSc in the Foundations of Advanced Information Technology at Imperial College. He was promoted to Professor of Computing Science in 1996. He has published over 150 research papers. He moved to the University of Southampton in 2003 to lead the ISIS research group. He has been appointed the Director of the Centre for Computational Statistics and Machine Learning at University College, London from July 2006. He has coordinated a number of European wide projects investigating the theory and practice of Machine Learning, including the NeuroCOLT projects. He is currently the scientific coordinator of a Framework VI Network of Excellence in Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL) involving 57 partners.

About the Author—LIANG MAO is currently a postdoctoral fellow with the Pattern Recognition and Machine Learning Research Group, School of Computer Science and Technology, East China Normal University, Shanghai, China. His research interests include probabilistic models, kernel methods, statistical learning theory, and multi-view learning.