Improving the Generalization Performance of Multi-class SVM via Angular Regularization

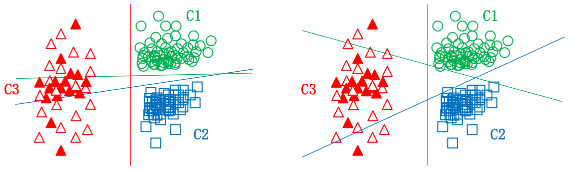
Jianxin Li[1], **Haoyi Zhou**[1], **Pengtao Xie**[2,3], **Yingchun Zhang**[1]
[1] School of Computer Science and Engineering, Beihang University
[2] Machine Learning Department, Carnegie Mellon University
[3] Petuum Inc, USA
{lijx, zhouhy, zhangyc}@act.buaa.edu.cn, pengtaox@cs.cmu.edu

Abstract

In multi-class support vector machine (MSVM) for classification, one core issue is to regularize the coefficient vectors to reduce overfitting. Various regularizers have been proposed such as ℓ_2, ℓ_1, and trace norm. In this paper, we introduce a new type of regularization approach – angular regularization, that encourages the coefficient vectors to have larger angles such that class regions can be widen to flexibly accommodate unseen samples. We propose a novel angular regularizer based on the singular values of the coefficient matrix, where the uniformity of singular values reduces the correlation among different classes and drives the angles between coefficient vectors to increase. In generalization error analysis, we show that decreasing this regularizer effectively reduces generalization error bound. On various datasets, we demonstrate the efficacy of the regularizer in reducing overfitting.

Table 1: Classification Accuracy of MSVM-ℓ_2

Dataset	Yeast	Usps	YaleB
Dim. of Features	8	256	1024
Train Accuracy (%)	57.23	97.31	97.70
Test Accuracy (%)	52.00	90.48	93.25



(a) Unregularized MSVM (b) Angular regularized MSVM

Figure 1: (a) Without angular regularization, the small angle between the green and blue lines narrows the decision region C3, which is consequently unable to accommodate unseen samples; (b) with angular regularization, the angle between the two lines is enlarged, which subsequently widens C3, that is more flexible to accommodate unseen data.

1 Introduction

Multi-class classification, which classifies a data sample $\mathbf{x}$ into one of $K > 2$ classes, is a fundamental task in machine learning. Among the various methods developed to accomplish this task, multi-class support vector machine (MSVM) [Weston and Watkins, 1998; Crammer and Singer, 2001; Shalev-Shwartz *et al.*, 2011; Kumar *et al.*, 2012] is a premier model that possesses both empirical efficacy and sound theoretical properties. At the core of MSVM is the principle of *maximum margin learning* [Crammer and Singer, 2001], encouraging the margin – which in the multi-class case is defined as the difference between the classification score $\mathbf{w}_y^\top \mathbf{x}$ corresponding to the true label y and the largest score of other classes $\max_{y' \neq y} \mathbf{w}_{y'}^\top \mathbf{x}$ – to be large, such that different classes can be well discriminated.

In training multi-class classification models, one typical problem is overfitting where the model performs well on training data but poorly on testing data. To improve the generalization performance on unseen samples, a number of regularization methods have been studied, such as ℓ_2-norm [Vapnik and Vapnik, 1998] encouraging large margin, ℓ_1-norm [Bradley and Mangasarian, 1998; Wang and Shen, 2012] promoting sparsity and trace norm [Amit *et al.*, 2007] encouraging low-rankness. While their effectiveness have been widely demonstrated, there is still much room for improvement. As an example, Table 1 shows the classification accuracy of ℓ_2-regularized MSVM on several datasets, where the gap between training and testing accuracy are still substantial.

In this paper, we study a new type of regularizer that encourages the coefficient vectors (equivalently, the hyperplanes parameterized by them) in MSVM to have large angles, for the purpose to control overfitting. Fig. 1 illustrates the idea. Without regularization (Fig. 1(a)), the two hyperplanes denoted by the green and blue lines share a small angle, generating a narrow decision region C3 which is supposed to accommodate data samples of class 3 (denoted by triangles). While C3 successfully accommodates most training samples (denoted by filled triangles), many test samples (denoted by void triangles) fall out of C3 and get misclassified due to the narrowness of this region, incurring severe overfitting. To alleviate this problem, an angular regularizer can be applied to enlarge the angle between these two hyperplanes, which consequently widens C3, enabling it to successfully accommodate unseen samples. Previously, angle-based regularizers have been studied in other contexts, such as ensemble learning [Guo, 2016] and latent variable modeling [Xie *et al.*, 2015a; Xie, 2015]. To our best knowledge, our work represents the first one introducing angular regularization into multi-class SVM. While the existing angular regularizers have shown promising results, they bear certain limitations, for instance, not amenable for theoretical analysis or optimization. We define a new angular regularizer based on the singular values of the coefficient matrix. These singular values are encouraged to approach uniformity, which reduces the correlation among different classes. Decreasing this regularizer can effectively enlarge the angles between coefficient vectors. In analysis, it is proved that decreasing the regularizer can reduce generalization error. In experiments on various datasets, we demonstrate the efficacy of this regularizer in preventing overfitting.

The major contributions of this paper are:

- We introduce the angular regularization into MSVM to improve its generalization performance.
- We propose a new angular regularizer that is amenable for optimization and theoretical analysis.
- Theoretically, we analyze how the angular regularizer affects the generalization performance.
- Empirically, we evaluate the effectiveness of the regularizer in control overfitting on various datasets.

The rest of this paper is organized as follows. Section 2 introduce the angular regularizer, which is applied to MSVM. Section 3 and 4 present the theoretical analysis and experimental results. Section 5 reviews related works and Section 6 concludes the paper.

2 Methods

In this section, we develop a new angular regularizer and apply it to regularize MSVM.

2.1 Multi-class Support Vector Machine (MSVM)

MSVM [Crammer and Singer, 2001; Weston and Watkins, 1998; Guermeur and Monfrini, 2011] is a well-established method for K-class classification. Given a set of training samples $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, each represented with a feature vector $\mathbf{x} \in \mathbb{R}^D$ and a class label $y \in \mathcal{Y} = \{1, \ldots, K\}$, MSVM aims to learn a coefficient vector $\mathbf{w}$ for each class to separate the K classes apart. At the core of MSVM is the large margin principle [Crammer and Singer, 2001]: given a sample $(\mathbf{x}, y)$, the classification score $\mathbf{w}_y^\top \mathbf{x}$ of the true label y is preferred to be larger than the scores of other classes with a certain margin (typically set to 1), namely

$$\mathbf{w}_y^\top \mathbf{x} - \mathbf{w}_r^\top \mathbf{x} \geq 1, \tag{1}$$

where $r = \mathrm{argmax}_{k \in \mathcal{Y}, k \neq y} \mathbf{w}_k^\top \mathbf{x}$. Let $\mathbf{W} \in \mathbb{R}^{D \times K}$ denote the coefficient matrix where the k-th column vector belongs to class k, MSVM solves the following optimization problem

$$\min_{\mathbf{W}} \frac{1}{m} \sum_{i=1}^{m} \max(0,\ 1 + \mathbf{w}_{r_i}^\top \mathbf{x}_i - \mathbf{w}_{y_i}^\top \mathbf{x}_i) + \frac{\lambda}{2} \|\mathbf{W}\|_2^2, \tag{2}$$

where the first term is a hinge loss encouraging large margin and the second term is an ℓ_2 regularizer.

2.2 A New Angular Regularizer

We propose to use an angular regularization to improve the generalization performance of MSVM. Previous angular regularizers [Yu *et al.*, 2011; Bao *et al.*, 2013; Xie *et al.*, 2015a] either are not amenable for optimization or do not facilitate theoretical analysis. In this section, we aim to define a new angular regularizer that overcomes these limitations.

Given K coefficient vectors $\{\mathbf{w}_k\}_{k=1}^K$, we first measure the pairwise angles between them:

$$\theta_{ij} = \arccos(\tfrac{|\mathbf{w}_i^\top \mathbf{w}_j|}{\|\mathbf{w}_i\|_2 \|\mathbf{w}_j\|_2}). \tag{3}$$

Similar to [Xie *et al.*, 2015a], we place an absolute value over $\mathbf{w}_i^\top \mathbf{w}_j$ to make θ_{ij} insensitive to vector flip (from $\mathbf{w}$ to $-\mathbf{w}$). Then a straightforward angular regularizer $\mathcal{R}(\mathbf{W})$ can be defined as the negative of the minimum angle

$$\mathcal{R}(\mathbf{W}) = -\mathrm{min}_{i \neq j} \theta_{ij} \quad . \tag{4}$$

Minimizing $\mathcal{R}(\mathbf{W})$ encourages these coefficient vectors to have larger angles. As will be discussed later, this definition facilitates theoretical analysis. However, it is not amenable for optimization since it is non-smooth. To address this issue, we use a smooth function $\widehat{\mathcal{R}}(\mathbf{W})$ to approximate $\mathcal{R}(\mathbf{W})$ and perform optimization over $\widehat{\mathcal{R}}(\mathbf{W})$ instead.

A Decorrelation Regularizer To define $\widehat{\mathcal{R}}(\mathbf{W})$, we begin with interpreting the angles from a statistical perspective. For the ease of presentation, we impose a constraint over $\mathbf{W}$:

$$\mathbf{W}^\top \mathbf{1} = 0 \tag{5}$$

where $\mathbf{1} \in \mathbb{R}^D$ is a vector whose elements are all 1. This constraint ensures the row vectors of $\mathbf{W}$ sum to 0. We will remove this constraint later on. Given two classes i and j, together with their coefficient vectors $\mathbf{w}_i$ and $\mathbf{w}_j$, we can treat i and j as two random variables and $\{w_{in}\}_{n=1}^D$, $\{w_{jn}\}_{n=1}^D$ as samples drawn from them. The empirical correlation of these two random variables are defined as

$$\rho_{ij} = \frac{\sum_{n=1}^D (w_{in} - \bar{w}_i)(w_{jn} - \bar{w}_j)}{\sqrt{\sum_{n=1}^D (w_{in} - \bar{w}_i)^2 \sum_{n=1}^D (w_{jn} - \bar{w}_j)^2}}, \tag{6}$$

where $\bar{w}_i = \frac{1}{n} \sum_{n=1}^D w_{in}$ and $\bar{w}_j = \frac{1}{n} \sum_{n=1}^D w_{jn}$. According to the constraint in Eq.(5), we have $\bar{w}_i = 0$ and $\bar{w}_j = 0$, then $\rho_{ij} = \cos(\theta_{ij})$, which indicates that to encourage the angle to be larger, we can equivalently suppress the correlation.

Based on this idea, we define a regularizer $\widehat{\mathcal{R}}(\mathbf{W})$ that encourages small correlation. First we compute the $K \times K$ Gram matrix $\mathbf{G} = \mathbf{W}^\top \mathbf{W}$, where $G_{ij} = \mathbf{w}_i^\top \mathbf{w}_j$. Suppose $\mathbf{W}$ is a column full-rank matrix and $K < D$, then $\mathbf{G}$ is a full-rank matrix with rank K. Let $\mathbf{G} = \sum_{i=1}^K \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$ be the eigen-decomposition where λ_i is an eigenvalue and $\mathbf{u}_i$ is the associated eigenvector. To understand the geometry meanings of these eigenvalues and eigenvectors, we can take a row view (Fig. 2) of $\mathbf{W}$. Different from the column view where each column vector is a D-dimensional coefficient vector, in the row view, each row vector can be treated as a sample of a K-dimensional random vector by regarding the classes as random variables. According to Eq.(5), these samples have zero mean.

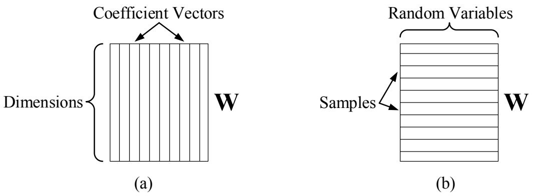
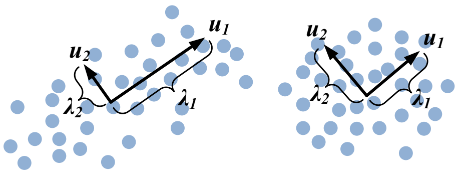


Figure 2: (a) Column view of $\mathbf{W}$; (b) Row view of $\mathbf{W}$.



(a) $\lambda_1 > \lambda_2$ (b) $\lambda_1 \approx \lambda_2$

Figure 3: The geometry meanings of eigenvalue λ_i.

Under this view, $\mathbf{G}$ can be thought of as an empirical covariance matrix of the random vector. Consider a point cloud formed by the D samples in the K-dimensional space, as is well known in Principle Component Analysis [Jolliffe, 2002], an eigenvector $\mathbf{u}_i$ of $\mathbf{G}$ represents a principal direction of the point cloud and the associated eigenvalue λ_i tells the variability of points along that direction. As shown in Fig. 3, the larger λ_i is, the more spread out the points along the direction $\mathbf{u}_i$. The level of disparity among eigenvalues indicates the level of correlations among the K dimensions. The more different the eigenvalues are, the higher the correlation is. In Fig. 3(a), the two eigenvalues have a large discrepancy and the two dimensions are strongly correlated. In contrast, in Fig. 3(b) where λ_1 and λ_2 are close, the correlation between the two dimensions is minimal. To reduce correlation, we encourage the eigenvalues to be uniform. We first transform the eigenvalues $\{\lambda_i\}_{i=1}^K$ into a probability distribution $p(X)$ through normalization

$$p(X=i) = \frac{\lambda_i}{\sum_{j=1}^K \lambda_j}, \tag{7}$$

then encourage $p(X)$ to be close to a uniform distribution $q(X=i) = \frac{1}{K}$, where the "closeness" is measured using Kullback-Leibler divergence

$$\begin{array}{rcl} KL(q\|p) & = & \sum_{i=1}^K \frac{1}{K} \log \frac{1/K}{\lambda_i / \sum_{j=1}^K \lambda_j} \\ & = & \log \sum_{j=1}^K \lambda_j - \frac{1}{K} \sum_{i=1}^K \log \lambda_i - \log K \end{array} . \tag{8}$$

Note that the i-th eigenvalue λ_i of $\mathbf{W}^\top\mathbf{W}$ is the square of the i-th singular value of $\mathbf{W}$. Hence, $KL(q\|p)$ is essentially defined over the singular values of $\mathbf{W}$. Using the two facts

$$\mathrm{tr}(\mathbf{G}) = \sum_{j=1}^K \lambda_j, \quad \det(\mathbf{G}) = \prod_{i=1}^K \lambda_i$$

we have

$$KL(q\|p) = \log \mathrm{tr}(\mathbf{W}^\top\mathbf{W}) - \frac{1}{K} \log \det(\mathbf{W}^\top\mathbf{W}) - \log K.$$

Dropping the constant, we define a decorrelation regularizer

$$\widehat{\mathcal{R}}(\mathbf{W}) = \log \mathrm{tr}(\mathbf{W}^\top\mathbf{W}) - \frac{1}{K} \log \det(\mathbf{W}^\top\mathbf{W}). \tag{9}$$

subject to $\mathbf{W}^\top\mathbf{1} = 0$. From now on, we drop this constraint, which does not affect theoretical results given in next section. $\widehat{\mathcal{R}}(\mathbf{W})$ is a smooth function that is much easier to optimize than $\mathcal{R}(\mathbf{W})$. Next, we show that $\widehat{\mathcal{R}}(\mathbf{W})$ can achieve similar effect as $\mathcal{R}(\mathbf{W})$: enlarging angles between vectors in $\mathbf{W}$.

Connection between $\widehat{\mathcal{R}}(\mathbf{W})$ and $\mathcal{R}(\mathbf{W})$ The following theorem states that $\widehat{\mathcal{R}}(\mathbf{W})$ is closely aligned with $\mathcal{R}(\mathbf{W})$.

Theorem 1. *Let $\nabla\widehat{\mathcal{R}}$ be the gradient of $\widehat{\mathcal{R}}(\mathbf{W})$ w.r.t $\mathbf{W}$. $\exists \kappa > 0$, such that $\forall\, \eta \in (0, \kappa)$, $\mathcal{R}(\mathbf{W} - \eta\nabla\widehat{\mathcal{R}}) \leq \mathcal{R}(\mathbf{W})$.*

Proof Due to space limit, we present the proof sketch of Theorem 1, and the following lemma is needed.

Lemma 1. *(Extension of Lemma 2 in [Xie* et al.*, 2015a]) Let the weight vector $\mathbf{w}_k$ of hyperplane k be decomposed into $\mathbf{w}_k = \mathbf{x}_k + l_k\mathbf{e}_k$, where $\mathbf{x}_k = \sum_{j=1, j\neq k}^K \alpha_j \mathbf{w}_j$ lies in the subspace L spanned by $\{\mathbf{w}_1, \dots, \mathbf{w}_K\} \backslash \{\mathbf{w}_k\}$, $\mathbf{e}_k$ is in the orthogonal complement of L, $\|\mathbf{e}_k\| = 1$, $\mathbf{e}_k \cdot \mathbf{w}_k > 0$, l_k is a positive scalar. Then the gradient of $\widehat{\mathcal{R}}(\mathbf{W})$ w.r.t $\mathbf{w}_k$ is $p_k\mathbf{x}_k + q_k\mathbf{e}_k$, where p_k is a positive scalar .*

Some of the proof techniques are borrowed from [Xie *et al.*, 2015a]. Let $s(\mathbf{W}) = \cos(-\mathcal{R}(\mathbf{W})) = \cos(\theta_{i*j*})$. If $s(\mathbf{W}) = 0$, then $s(\mathbf{W} - \eta\nabla\widehat{\mathcal{R}}) - s(\mathbf{W}) = |(p_{i*} + p_{j*})s(\mathbf{W}) + o(\eta)|[1 + f(\eta) + o(\eta)] = o(\eta)$, such that $\lim_{\eta\to 0} \frac{s(\mathbf{W}-\eta\nabla\widehat{\mathcal{R}})-s(\mathbf{W})}{\eta} = 0$. Otherwise, if $s(\mathbf{W}) \neq 0$, then $\frac{s(\mathbf{W}-\eta\nabla\widehat{\mathcal{R}})}{s(\mathbf{W})} = 1 + \psi_{i*j*}\eta + o(\eta)$, where $\psi_{i*j*} = -[\frac{p_{i*}l_{i*}^2 - q_{i*}l_{i*}}{\|\mathbf{w}_{i*}\|_2} + \frac{p_{j*}l_{j*}^2 - q_{j*}l_{j*}}{\|\mathbf{w}_{j*}\|_2}]$. We have $\frac{p_{i*}l_{i*}^2 - q_{i*}l_{i*}}{\|\mathbf{w}_{i*}\|_2} = \frac{l_{i*}}{K(\mathbf{e}_{i*}\cdot\mathbf{w}_{i*})\|\mathbf{w}_{i*}\|^2} > 0$ and $\frac{p_{j*}l_{j*}^2 - q_{j*}l_{j*}}{\|\mathbf{w}_{j*}\|_2} > 0$, so $\psi_{i*j*} < 0$. Thus, $\lim_{\eta\to 0} \frac{s(\mathbf{W}-\eta\nabla\widehat{\mathcal{R}})-s(\mathbf{W})}{\eta} = s(\mathbf{W}) \cdot \lim_{\eta\to 0} \frac{\frac{s(\mathbf{W}-\eta\nabla\widehat{\mathcal{R}})}{s(\mathbf{W})} - 1}{\eta} = s(\mathbf{W})\psi_{i*j*} < 0$. Overall, $\exists\kappa > 0$, such that $\forall\eta \in (0, \kappa)$, we have $\frac{s(\mathbf{W}-\eta\nabla\widehat{\mathcal{R}})-s(\mathbf{W})}{\eta} \leq 0$, i.e. $\mathcal{R}(\mathbf{W} - \eta\nabla\widehat{\mathcal{R}}) \leq \mathcal{R}(\mathbf{W})$. □

From Theorem 1, we can see the negative gradient of $\widehat{\mathcal{R}}(\mathbf{W})$ is a descent direction of $\mathcal{R}(\mathbf{W})$. Hence these two regularizers have closely-aligned shape and the same local optimal points. $\widehat{\mathcal{R}}(\mathbf{W})$ can be utilized as a close approximation of $\mathcal{R}(\mathbf{W})$. Decreasing $\widehat{\mathcal{R}}(\mathbf{W})$ effectively decreases $\mathcal{R}(\mathbf{W})$, which enlarges the angles.

MSVM with Angular Regularization Given $\widehat{\mathcal{R}}(\mathbf{W})$, we can utilize it to enlarge the angles between coefficient vectors in MSVM, for the sake of reducing overfitting. An angle-regularized MSVM (AR-MSVM) problem can be defined as:

$$\min_{\mathbf{W}} \frac{1}{m} \sum_{i=1}^m \max(0,\, 1 + \mathbf{w}_{r_i}^\top \mathbf{x}_i - \mathbf{w}_{y_i}^\top \mathbf{x}_i) + \frac{\lambda}{2}\|\mathbf{W}\|_2^2 + \frac{\beta}{2}\widehat{\mathcal{R}}(\mathbf{W}), \tag{10}$$

where β is the regularization parameter. We use a stochastic sub-gradient method [Shalev-Shwartz *et al.*, 2011; Wang *et al.*, 2010] to solve the AR-MSVM problem. The gradient of

$\widehat{\mathcal{R}}(\mathbf{W})$ is

$$\nabla\widehat{\mathcal{R}} = \frac{2\mathbf{W}}{\mathrm{tr}(\mathbf{W}^\top\mathbf{W})} - \frac{2}{K}\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}. \qquad (11)$$

The subgradient of the hinge loss $L = \max(0,\ 1 + \mathbf{w}_{r_i}^\top\mathbf{x}_i - \mathbf{w}_{y_i}^\top\mathbf{x}_i)$ is a zero matrix if $\mathbf{w}_{y_i}^\top - \mathbf{w}_{r_i}^\top\mathbf{x}_i > 1$. Otherwise, the k-th column vector of the sub-gradient matrix is $\mathbf{x}_i$ if $k \neq y_i$ and is $-\mathbf{x}_i$ if $k = y_i$.

3 Analysis

In this section, we analyze how the angular regularizer $\mathcal{R}(\mathbf{W})$ affects the generalization performance of MSVM. We begin with a recap of basic terminologies. Let $L(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{D}}[l(\mathbf{x},y)]$ be the true risk of a hypothesis h, where $\mathbb{D}$ represents the true distribution and $l(\mathbf{x},y)$ is the loss function. Let $\widehat{L}(h) = \frac{1}{m}\sum_{i=1}^m l(\mathbf{x}_i, y_i)$ denote the empirical risk. Let $h^* \in \mathrm{argmin}_{h\in H} L(h)$ and $\hat{h} \in \mathrm{argmin}_{h\in H} \widehat{L}(h)$ be the true and empirical risk minimizer respectively. The generalization error is defined as $L(\hat{h}) - L(h^*)$ which measures how well the hypothesis $\hat{h}$ learned on the training data generalizes on the unseen data.

In the analysis of AR-MSVM, the hypothesis set is defined as $H = \{h | h(\mathbf{x}, y) = \mathbf{w}_y^\top\mathbf{x},\ \mathbf{x} \in \mathbb{R}^D,\ y \in \mathcal{Y} = \{1, \ldots, K\}\}$. Let H^1 be the best-case hypothesis set where different classes can be perfectly separated: $\forall(\mathbf{x}, y), r \in \mathcal{Y}\backslash\{y\}, \mathbf{w}_y^\top\mathbf{x} \geq \mathbf{w}_r^\top\mathbf{x}$. For each hypothesis $h \in H$, its multi-class margin for the input-output pair $(\mathbf{x}, y)$ is $\rho_h(\mathbf{x}, y) = h(\mathbf{x}, y) - \max_{k\neq y} h(\mathbf{x}, k)$. And the loss set is defined as $\mathcal{A} = \{l | l(\mathbf{x}, y) = \Phi_p(\rho_h(\mathbf{x}, y))\}$, where $\Phi_p(x) = \max(0, 1 - x/p)$ is a p-margin hinge loss. In addition, we assume $\|\mathbf{x}\|_2 \leq C_1$ and $\|\mathbf{w}\|_2 \leq C_2$.

The generalization error bound of AR-MSVM is given in the following theorem.

Theorem 2. *Fix $p > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following multi-class classification genralization bound holds for all $h \in H \rightsquigarrow H^1$:*

$$L(\hat{h}) - L(h^*) \leq \frac{8K^2C_1C_2}{p\sqrt{m}} + \frac{2}{\sqrt{m}}$$
$$+ \left(\frac{1}{p}\sqrt{(1 - \tfrac{1}{K})(\cos(-\mathcal{R}(\mathbf{W})) + \tfrac{K+1}{K-1})C_1C_2} + 1\right)\sqrt{\frac{2\log(2/\delta)}{m}}.$$

With the objective function in Eq.(10) decreasing, the *maximum margin* principle encourages H approaching H^1. As can be seen, the generalization error bound is an increasing function of the angular regularizer $\mathcal{R}(\mathbf{W})$. Hence decreasing $\mathcal{R}(\mathbf{W})$ can reduce the error bound and improve the generalization performance.

3.1 Proof Sketch

Inspired by [Xie *et al.*, 2015b], we bound $L(\hat{h}) - L(h^*)$ using Rademacher complexity $R_m(\mathcal{A}) = \mathbb{E}[\sup_{l\in\mathcal{A}} \frac{1}{m}\sum_{i=1}^m \sigma_i \cdot l(\mathbf{x}_i, y_i)]$, where σ_i is uniform over $\{-1, 1\}$.

Lemma 2. *[Percy, 2015] Fix $p > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$*

$$L(\hat{h}) - L(h^*) \leq 4R_m(\mathcal{A}) + B\sqrt{\frac{2\log(2/\delta)}{m}} \qquad (12)$$

for $B \geq \sup_{\mathbf{x},y,l} |l(\mathbf{x}, y)|$.

Table 2: Statistics of Datasets

Dataset	#Classes	#Train	#Test	#Features
YaleB	38	1500	914	1024
ImageNet-50	50	30K	10K	128
Covtype	7	100K	40K	54
Shuttle	7	30450	14500	9
New-thyroid	3	108	107	5
Yeast	10	1134	350	8
Dermatology	6	323	35	33
Page-Blocks	5	4924	548	10
Wine-Quality-Red	6	1439	160	11
Zoo	7	89	12	16

Then we bound $R_m(\mathcal{A})$. Let $\widetilde{H} = \{z = (\mathbf{x}, y) \mapsto \rho_h(\mathbf{x}, y)\}$ be an auxiliary hypothesis set, we have

Lemma 3. *For a fixed $p > 0$, we have*

$$R_m(\mathcal{A}) \leq \frac{1}{2}R_m^{\|}(l \circ \widetilde{H}) \leq \frac{8K^2C_1C_2}{p\sqrt{m}} + \frac{2}{\sqrt{m}}. \qquad (13)$$

The $R_m^{\|}(\mathcal{A})$ denotes $R_m(\mathcal{A})$'s absolute form [Bartlett and Mendelson, 2002]. The second inequality is from the following proposition and the Lipschitz condition of p-margin hinge loss.

Proposition 1. *(Extension of Theorem 1 in [Cortes* et al.*, 2013]) For any fixed $y \in \mathcal{Y}$, $H_\mathcal{X}$ only takes $\mathbf{x}$ into consideration. From $\widetilde{H}$, the empirical Rademacher complexity has*

$$R_m^{\|}(\widetilde{H}) \leq K^2 R_m^{\|}(H_\mathcal{X}). \qquad (14)$$

Next we bound B with the same Lipschitz condition and analyze $\rho_h(\mathbf{x}, y)$'s distribution w.r.t H. The key step is making use of $\mathcal{R}(\mathbf{W})$ to upper bound $\|\mathbf{W}\|_{op}$ (operator norm).

Lemma 4. *Fix $p > 0$. Then, for any hypothesis set H, $|l(\mathbf{x}, y)|$ can be bounded as $|l(\mathbf{x}, y)| \leq \frac{2}{p}C_1C_2 + 1$.*

When $H \rightsquigarrow H^1$, a tighter bound is acquired

$$|l(\mathbf{x}, y)| \leq \frac{1}{p}\sqrt{(1 - \tfrac{1}{K})(\cos(-\mathcal{R}(\mathbf{W})) + \tfrac{K+1}{K-1})C_1C_2} + 1. \quad (15)$$

Putting the pieces together[1], we can obtain Theorem 2.

4 Experiments

In this section, we present experimental results.

4.1 Experimental Setup

We evaluated our method on ten datasets. Table 2 summaries their statistics. The regularization parameters λ and β are tuned in the range $[2^{-20}, 2^{-19}, \ldots, 2^{20}]$ via 5-fold cross validations. In the stochastic sub-gradient descent algorithm, the mini-batch size and number of epochs are set to 20 and 50 respectively. The learning rate is set according to ADADELTA [Zeiler, 2012] in all methods. All the experiments are conducted over 10 random train/test splits and the results are averaged over the 10 runs. We compare AR-MSVM with the following baselines: (1) C-SVC [Chang and Lin, 2011]: performing multi-class classification using an one-against-one strategy. (2) AMM [Wang *et al.*, 2011]: a multi-class classification method based on adaptive multi-hyperplanes; (3) MSVM-ℓ_1 [Shalev-Shwartz and Tewari, 2011]: ℓ_1-regularized MSVM; (4) CS-MSVM:

[1] For more details, please refer to the complementary document at *http://act.buaa.edu.cn/lijx/pubs/ijcai17-2077-supply.pdf*.

Table 3: Classification results (%) on six multi-class datasets

Dataset	Metric	C-SVC	AMM	MSVM-ℓ_1	CS-MSVM	IC-MSVM	Div-MSVM	MSVM-Struct	MSVM-ℓ_2	AR-MSVM
YaleB	*Acc*	55.31 ±0.1	83.32 ±2.5	85.12 ±4.4	94.13 ±1.6	89.38 ±0.9	94.25 ±0.4	91.50 ±0.0	93.90 ±0.5	**94.55** ±0.8
	F	54.74 ±1.0	86.88 ±1.8	87.11 ±2.6	93.71 ±1.3	90.54 ±1.4	93.87 ±0.3	91.39 ±0.0	93.51 ±0.5	**94.38** ±0.7
ImageNet-50	*Acc*	91.59 ±0.4	92.27 ±0.1	91.23 ±0.0	92.86 ±0.1	92.96 ±0.1	92.93 ±0.1	89.36 ±0.0	92.32 ±0.1	**93.12** ±0.1
	F	91.70 ±0.4	92.31 ±0.1	91.38 ±0.0	92.90 ±0.1	92.98 ±0.1	92.97 ±0.1	89.65 ±0.0	92.35 ±0.0	**93.17** ±0.1
Covtype	*Acc*	71.35 ±0.5	70.21 ±0.3	56.12 ±0.4	70.29 ±1.1	70.82 ±0.2	70.70 ±0.1	65.84 ±0.1	69.45 ±0.2	**71.75** ±0.1
	F	50.97 ±3.1	49.84 ±2.4	37.03 ±3.5	50.67 ±2.1	50.82 ±0.1	50.53 ±0.5	36.80 ±0.1	45.64 ±1.8	**51.31** ±0.9
Shuttle	*Acc*	**98.86** ±0.3	93.81 ±0.3	79.41 ±0.4	96.02 ±0.0	96.22 ±0.1	94.74 ±0.2	66.57 ±0.0	92.83 ±0.1	97.08 ±0.3
	F	52.94 ±2.1	54.12 ±0.3	24.90 ±0.0	57.20 ±0.1	57.57 ±0.1	55.61 ±0.2	33.66 ±0.0	53.55 ±0.2	**58.53** ±0.3
New-thyroid	*Acc*	76.92 ±0.9	86.77 ±1.8	86.15 ±0.0	78.46 ±10.9	78.69 ±3.3	91.54 ±0.8	89.23 ±0.0	90.87 ±0.0	**92.15** ±0.5
	F	58.73 ±1.0	83.66 ±2.4	87.50 ±0.0	75.32 ±1.4	78.46 ±2.1	90.08 ±1.1	87.50 ±0.0	89.24 ±0.0	**90.91** ±0.7
Yeast	*Acc*	**59.71** ±1.5	48.63 ±5.3	53.71 ±2.4	53.61 ±6.2	54.14 ±0.2	53.14 ±2.5	49.93 ±0.1	51.86 ±3.8	54.80 ±1.1
	F	**61.81** ±2.0	44.87 ±5.1	48.63 ±4.8	52.80 ±4.1	52.07 ±2.0	52.64 ±3.7	37.09 ±0.1	49.45 ±3.1	54.65 ±4.3

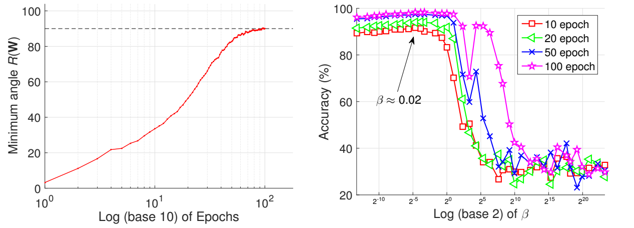
Table 4: Classification accuracy (%) on four imbalance datasets

Dataset	Metric	Global-CS	CS	Static-SMT	SL-SMT	AdaB.NC	MSVM-ℓ_2	AR-MSVM	Δ
Dermatology	*Acc*	95.78	95.44	95.60	94.30	**97.08**	84.29 ± 8.959	94.57 ± 1.622	+10.28
Page-Blocks	*Acc*	91.67	89.32	69.04	89.34	88.29	88.93 ± 2.237	**92.27** ± 1.019	+3.34
Wine-Quality-Red	*Acc*	39.33	37.82	30.74	37.93	37.22	49.94 ± 5.064	**52.94** ± 1.793	+3.00
Zoo	*Acc*	95.02	93.02	95.35	93.02	95.02	93.28 ± 2.500	**95.38** ± 1.510	+2.10

MSVM regularized by a cosine similarity regularizer [Yu *et al.*, 2011]; (5) IC-MSVM: MSVM regularized by an incoherence regularizer [Bao *et al.*, 2013]; (6) Div-MSVM: MSVM regularized by a diversity-promoting regularizer [Xie *et al.*, 2015a]; (7) MSVM-Struct [Tsochantaridis *et al.*, 2004]: ℓ_2-regularized MSVM solved with a cutting plane algorithm; (8) MSVM-ℓ_2 [Wang *et al.*, 2010]: ℓ_2-regularized MSVM solved with a stochastic gradient descent algorithm; Two metrics are used for evaluation: accuracy (Acc) and F-measure (F).

4.2 Results

Table 3 shows the classification results on six datasets, from which we observe that: (1) AR-MSVM greatly improves upon vanilla MSVM methods, including MSVM-ℓ_2 and MSVM-Struct. Comparing the last two columns, we can see AR-MSVM outperforms MSVM-ℓ_2 (which works better than MSVM-Struct) on all the datasets. This demonstrates that the angular regularizer can effectively improve the generalization performance on test data. (2) AR-MSVM achieves better performance than CS-MSVM, IC-MSVM and Div-MSVM. These three methods use previously proposed regularizers to encourage large angles as well. But the incoherence (IC) regularizer and diversity-promoting (Div) regularizer are non-smooth, bringing in difficulty for optimization, which we conjecture is the major reason being inferior to the angular regularizer proposed in this paper. The cosine similarity (CS) regularizer can encourage the sum of angles to increase, but cannot guarantee each individual angle is increased. The AR regularizer we proposed strictly increases the minimal angle, hence ensuring all angles are increased. (3) AR-MSVM outperforms MSVM-ℓ_1, demonstrating that angular regularization is more effective in control overfitting than ℓ_1 regularization. (4) AR-MSVM outperforms two non-MSVM methods: C-SVC and AMM, demonstrating its competitive performance in solving multi-class classification problems. (5) AR-MSVM achieves better improvement on F-measure than on accuracy. Accuracy is biased toward classes with larger number of samples. F-measure aims to reduce this bias and is more suitable to measure the performance on small-sample classes. This indicates AR-MSVM's ability to better classify small-sample classes.



(a) (b)

Figure 4: (a) AR enlarges the minimal angle; (b) Sensitivity to the regularization parameter.

To verify the ability of AR in enlarging angles, we plot how the minimal angle changes as the algorithm proceeds. Fig. 4(a) shows the curve on the New-thyroid dataset with $\beta \approx 0.02$. It can be seen that the minimal angle consistently increases, which demonstrates that AR is able to drive the angles to become large.

4.3 Parameter Sensitivity

We study the sensitivity of AR-MSVM to the regularization parameter β. Fig. 4(b) shows how accuracy varies as β increases on the New-thyroid dataset. Initially, increasing β improves accuracy, because a larger β results in larger angles that help prevent overfitting. However, further increasing β causes the accuracy to drop. That is because, if β is too large, the regularizer dominates the hinge loss and impairs the large-margin learning.

4.4 Imbalanced Classification

Inspired by the study in [Xie *et al.*, 2015a] that large-angle regularizer can effectively capture infrequent patterns when the frequency of patterns is imbalanced, we are interested in how the AR helps with imbalanced classification [Wang and Yao, 2012] where the number of samples of different classes

Table 5: *Acc* (%) on each class in the Page-Blocks dataset

Class	1	2	3	4	5	sum
#Train	4421	296	25	79	103	4924
#Test	492	33	3	8	12	548
MSVM-ℓ_2	96.75	21.21	66.67	12.50	16.67	89.05
AR-MSVM	95.93	54.55	66.67	50.00	75.00	92.15

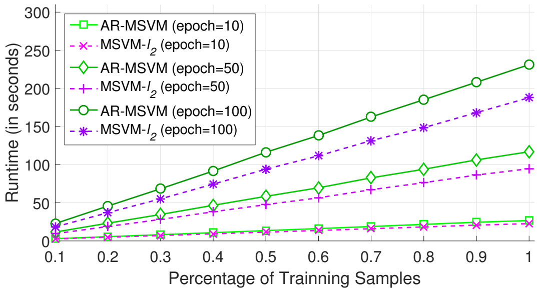


Figure 5: Runtime on the Covtype dataset.

is imbalanced. In Table 4, we compare with state-of-the-art methods dealing with class imbalance, including Global-CS, CS, Static-SMT, SL-SMT and AdaB.NC (summarized in [FernáNdez *et al.*, 2013]). As can be seen, on three datasets, our method achieves the best accuracy, demonstraing its ability to handle imbalanced classification. To further verify this, we measure the per-class accuracy on the Page-Blocks dataset in Table 5. On small classes (2-5), AR-MSVM performs much better than MSVM. On large class (1), they are comparable. This indicates that AR-MSVM is able to improve the performance on small classes without sacrificing that on large classes.

4.5 Computational Time

We evaluate how much extra time is incurred after adding the angular regularizer to MSVM. The algorithms were implemented in MATLAB and the experiments were run on a Linux machine with a 2.00GHz Xeon CPU and 256G memory. A comparison of runtime is shown in Fig. 5. As can be seen, the runtime of AR-MSVM is close to MSVM. The extra time cost is moderate.

5 Related Work

5.1 Multi-class Classification

Many works [Lee *et al.*, 2004; Guermeur and Monfrini, 2011] have been proposed for multi-class classification. A detailed comparison is provided in [Wang and Xue, 2014]. Wang *et al.* [2011] proposed an Adaptive Multi-hyperplane Machine (AMM) algorithm that consists of a set of hyperplanes (weights), each assigned to one of the multiple classes, and predicts based on the associated class of the weight that provides the largest prediction. Lapin *et al.* [2015] proposed a top-k multiclass SVM that directly optimizes the top-k performance in image classification, to deal with class ambiguity. This method generalizes multiclass SVM based on a tight convex upper bound of the top-k error.

To avoid overfitting, several regularizers have been proposed. The ℓ_2-norm [Vapnik and Vapnik, 1998] penalizes the magnitude of coefficients while encouraging large margin. The ℓ_1-norm [Bradley and Mangasarian, 1998; Wang and Shen, 2012], which promotes sparsity, can effectively control overfitting in high dimensional feature spaces. Amit *et al.* [2007] uses trace norm to encourage low-rankness on the coefficient matrix, for the sake of capturing common structure shared by different classes. Guo [2016] proposed an angle-based regularization method in the ensemble learning of multiple binary SVMs.

5.2 Angular Regularizers

Several regularizers have been proposed to control the angles (equivalently, the cosine similarity) between vectors, in the context of ensemble learning, latent variable modeling and deep learning. Yu *et al.* [2011] computes the pairwise cosine similarity s_{ij} between K vectors and defines the regularizer as $\sum_{1\leq i<j\leq K}(1-s_{ij})$. Similarly, Bao *et al.* [2013] aggregates the cosine similarity scores into a regularizer $-\log(\frac{1}{K(K-1)}\sum_{1\leq i<j\leq K}\beta|s_{ij}|)^{\frac{1}{\beta}}$ where $\beta > 0$. In [Xie *et al.*, 2015a], the regularizer is defined as mean of $\arccos(|s_{ij}|)$ minus the variance of $\arccos(|s_{ij}|)$. The variance term is utilized to encourage the vectors to evenly spread out to different directions. These regularizers have two limitations. First, they are not amenable for theoretical analysis. Second, some of them are non-smooth functions that present great challenges for optimization.

Another closely related regularizer is Determinantal Point Process (DPP) [Kulesza *et al.*, 2012]. Given K vectors $\{\mathbf{w}_i\}_{i=1}^K$, DPP defines a probability distribution $p(\{\mathbf{w}_i\}_{i=1}^K) \propto \det(\mathbf{G})$. $\mathbf{G}$ is a $K\times K$ kernel matrix where $G_{ij} = k(\mathbf{a}_i, \mathbf{a}_j)$ with $k(\cdot,\cdot)$ as a kernel function. $\det(\cdot)$ denotes the determinant of a matrix. Under a linear kernel, the DPP regularizer is analogous to $\widehat{\mathcal{R}}(\mathbf{W})$ in Eq.(9). The shortcoming of DPP is that it is sensitive to vector scaling, which can be overcome by the $\log \mathrm{tr}(\mathbf{W}^\top\mathbf{W})$ term in $\widehat{\mathcal{R}}(\mathbf{W})$.

6 Conclusion

In this paper, we introduce a new regularization method – angular regularization – into multi-class SVM, to improve its generalization performance. This regularizer encourages the coefficient vectors to have larger angles such that the decision regions can be widen to flexibly accommodate unseen data. We define a novel angular regularizer by encouraging the singular values of the coefficient matrix to approach evenness and prove its effectiveness. We analyze how this regularizer influences the generalization performance of MSVM and the results indicate that decreasing the regularizer can reduce the generalization error bound. The evaluation on a number of datasets demonstrate the capability of this regularizer in reducing overfitting.

Acknowledgements

This work is supported by China 973 Fundamental R&D Program (No.2014CB340300), NSFC program (No.61472022, 61421003), SKLSDE-2016ZX-11, and partly by the Beijing Advanced Innovation Center for Big Data and Brain Computing. The corresponding author is Jianxin Li.

References

[Amit *et al.*, 2007] Yonatan Amit, Michael Fink, Nathan Srebro, and Shimon Ullman. Uncovering shared structures in multiclass classification. In *ICML*, pages 17–24, 2007.

[Bao *et al.*, 2013] Yebo Bao, Hui Jiang, Lirong Dai, and Cong Liu. Incoherent training of deep neural networks to de-correlate bottleneck features for speech recognition. In *ICASSP*, pages 6980–6984. IEEE, 2013.

[Bartlett and Mendelson, 2002] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.

[Bradley and Mangasarian, 1998] Paul S Bradley and Olvi L Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, pages 82–90, 1998.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[Cortes *et al.*, 2013] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Multi-class classification with maximum margin multiple kernel. In *ICML*, pages 46–54, 2013.

[Crammer and Singer, 2001] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2001.

[FernáNdez *et al.*, 2013] Alberto FernáNdez, Victoria LóPez, Mikel Galar, MaríA José Del Jesus, and Francisco Herrera. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-based systems*, 42:97–110, 2013.

[Guermeur and Monfrini, 2011] Yann Guermeur and Emmanuel Monfrini. A quadratic loss multi-class svm for which a radius–margin bound applies. *Informatica*, 22(1):73–96, 2011.

[Guo, 2016] Xiaojie Guo. Exclusivity regularized machine. *arXiv:1603.08318*, 2016.

[Jolliffe, 2002] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[Kulesza *et al.*, 2012] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

[Kumar *et al.*, 2012] Abhishek Kumar, Alexandru Niculescu-Mizil, Koray Kavukcuoglu, and Hal Daume III. A binary classification framework for two-stage multiple kernel learning. In *ICML*, pages 1295–1302, 2012.

[Lapin *et al.*, 2015] Maksim Lapin, Matthias Hein, and Bernt Schiele. Top-k multiclass svm. In *NIPS*, pages 325–333, 2015.

[Lee *et al.*, 2004] Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *JASA*, 99(465):67–81, 2004.

[Percy, 2015] Liang Percy. Lecture notes of statistical learning theory. *https://web.stanford.edu/class/cs229t/notes.pdf*, 2015.

[Shalev-Shwartz and Tewari, 2011] Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l1-regularized loss minimization. *JMLR*, 12:1865–1892, 2011.

[Shalev-Shwartz *et al.*, 2011] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

[Tsochantaridis *et al.*, 2004] Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, page 104, 2004.

[Vapnik and Vapnik, 1998] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[Wang and Shen, 2012] Lifeng Wang and Xiaotong Shen. On l1-norm multiclass support vector machines. *JASA*, 102(478):583–594, 2012.

[Wang and Xue, 2014] Zhe Wang and Xiangyang Xue. Multi-class support vector machine. In *Support Vector Machines Applications*, pages 23–48. Springer, 2014.

[Wang and Yao, 2012] Shuo Wang and Xin Yao. Multiclass imbalance problems: Analysis and potential solutions. *TSMC*, 42(4):1119–1130, 2012.

[Wang *et al.*, 2010] Zhuang Wang, Koby Crammer, and Slobodan Vucetic. Multi-class pegasos on a budget. In *ICML*, pages 1143–1150, 2010.

[Wang *et al.*, 2011] Zhuang Wang, Nemanja Djuric, Koby Crammer, and Slobodan Vucetic. Trading representability for scalability: adaptive multi-hyperplane machine for nonlinear classification. In *SIGKDD*, pages 24–32. ACM, 2011.

[Weston and Watkins, 1998] Jason Weston and Chris Watkins. Multi-class support vector machines. Technical report, Citeseer, 1998.

[Xie *et al.*, 2015a] Pengtao Xie, Yuntian Deng, and Eric Xing. Diversifying restricted boltzmann machine for document modeling. In *SIGKDD*, pages 1315–1324, 2015.

[Xie *et al.*, 2015b] Pengtao Xie, Yuntian Deng, and Eric Xing. On the generalization error bounds of neural networks under diversity-inducing mutual angular regularization. *arXiv:1511.07110*, 2015.

[Xie, 2015] Pengtao Xie. Learning compact and effective distance metrics with diversity regularization. In *ECML-KDD*, pages 610–624. Springer, 2015.

[Yu *et al.*, 2011] Yang Yu, Yu-Feng Li, and Zhi-Hua Zhou. Diversity regularized machine. In *IJCAI*, volume 22, pages 1603–1608. Citeseer, 2011.

[Zeiler, 2012] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv:1212.5701*, 2012.