On the Optimization of Margin Distribution

Meng-Zhang Qian\textsuperscript{1}, Zheng Ai\textsuperscript{1}, Teng Zhang\textsuperscript{2} and Wei Gao\textsuperscript{1}

\textsuperscript{1}National Key Laboratory for Novel Software Technology, Nanjing University, China
\textsuperscript{2}School of Computer Science and Technology, Huazhong University of Science and Technology, China
\{qianmz, aiz, gaow\}@lamda.nju.edu.cn, tengzhang@hust.edu.cn

Abstract

Margin has played an important role on the design and analysis of learning algorithms during the past years, mostly working with the maximization of the minimum margin. Recent years have witnessed the increasing empirical studies on the optimization of margin distribution according to different statistics such as medium margin, average margin, margin variance, etc., whereas there is a relative paucity of theoretical understanding.

In this work, we take one step on this direction by providing a new generalization error bound, which is heavily relevant to margin distribution by incorporating ingredients such as average margin and semi-variance, a new margin statistics for the characterization of margin distribution. Inspired by the theoretical findings, we propose the MsvM\textsubscript{AV}, an efficient approach to achieve better performance by optimizing margin distribution in terms of its empirical average margin and semi-variance. We finally conduct extensive experiments to show the superiority of the proposed MsvM\textsubscript{AV} approach.

1 Introduction

Margin has played an important role on the design of learning algorithms from the pioneer work [Vapnik, 1982], which proposed the famous Support Vector Machines (SVMs) by maximizing the minimum margin, i.e. the smallest distance from the instances to the classification boundary. Boser \textit{et al.} [1992] introduced the kernel technique for SVMs to relax the linear separation. Large margin has been one of the most important principles on the design of learning algorithms in the history of machine learning [Cortes and Vapnik, 1995; Schapire \textit{et al.}, 1998; Rossset \textit{et al.}, 2003; Shivashwamy and Jebara, 2010; Ji \textit{et al.}, 2021], even for recent deep learning [Sokolić \textit{et al.}, 2017; Weinstein \textit{et al.}, 2020].

Various margin-based bounds have been presented to study the generalization performance of learning algorithms. Bartlett and Shawe-Taylor [1999] possibly presented the first generalization margin bounds based on VC dimension and fat-shattering dimension. Bartlett and Mendelson [2002] introduced the famous margin bounds based on Rademacher complexity, a data-dependent and finite-sample complexity measure. Kabán and Durrant [2020] took advantage of geometric structure to provide margin bounds for compressive learning. Grönlund \textit{et al.} [2020] presented the near-tight margin generalization bound for SVMs. Margin has also been an ingredient to analyze the generalization performance for other algorithms such as boosting [Schapire \textit{et al.}, 1998; Breiman, 1999; Gao and Zhou, 2013], and deep learning [Bartlett \textit{et al.}, 2017; Wei and Ma, 2020].

Margin distribution has been considered as an important ingredient on the design and analysis of learning algorithms, and the basic idea is to optimize some margin statistics, relevant to the whole margin distribution rather than single margin. Garg and Roth [2003] introduced the model complexity measure to optimize margin distribution. Pelckmans \textit{et al.} [2007] optimized margin distribution via average margin, while Aiolli \textit{et al.} [2008] tried to maximize the minimum margin and average margin. Zhang and Zhou [2014] proposed the large margin distribution machine by considering average margin and margin variance simultaneously, which motivates the design of a series learning algorithms on the optimization of margin distribution [Cheng \textit{et al.}, 2016; Rastogi \textit{et al.}, 2020]. For deep learning, Jiang \textit{et al.} [2019] introduced some margin distribution statistics, such as total variation, median quartile, etc., to analyze the generalization of neural networks. There is a relative paucity of theoretical understanding on how to correlate margin distribution with the generalization of learning algorithms.

This work tries to fill the gap between theoretical and empirical studies on the optimization of margin distribution, and the main contributions can be summarized as follows:

- We present a new generalization error bound, which is heavily relevant to margin distribution by incorporating factors such as average margin and semi-variance. Here, semi-variance is a new statistics, counting the average of squared distances between average margin and the instances’ margin, that is smaller than average margin.

- Motivated from our theoretical result, we develop the MsvM\textsubscript{AV} approach, which tries to achieve better generalization performance by optimizing margin distribution in terms of empirical average margin and semi-variance. We find the closed-form solution in optimization, and improve its efficiency via Sherman-Morrison formula.
• We conduct extensive empirical studies to validate the effectiveness of the MsM approach in comparisons with the state-of-the-art algorithms on large-margin or margin distribution optimization.

The rest of this paper is organized as follows. Section 2 introduces some preliminaries. Section 3 presents theoretical analysis. Section 4 proposes the MsM approach. Section 5 conducts extensive empirical studies, and Section 6 concludes with future work.

2 Preliminaries
Let \( \mathcal{X} \subseteq \mathbb{R}^d \) and \( \mathcal{Y} = \{+1, -1\} \) denote the instance and label space, respectively. Suppose that \( \mathcal{D} \) is an underlying (unknown) distribution over the product space \( \mathcal{X} \times \mathcal{Y} \). Let \( S_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \) be a training sample with each element drawn independently and identically (i.i.d.) from distribution \( \mathcal{D} \). We use \( \Pr_D[\cdot] \) and \( E_D[\cdot] \) to refer to the probability and expectation according to distribution \( \mathcal{D} \), respectively.

Let \( \mathcal{H} = \{h: \mathcal{X} \to \{-1, +1\}\} \) be a function space. We define the classification error (or generalization risk) with respect to function \( h \in \mathcal{H} \) and distribution \( \mathcal{D} \) as

\[
\mathcal{E}(h) = \Pr_D[\mathrm{sgn}[h(x)] \neq y] = E_D[\mathbb{I}[yh(x) \leq 0]],
\]

where the sign function \( \mathrm{sgn}[\cdot] \) returns \(+1\), \(0\) and \(-1\) if the argument is positive, zero and negative, respectively, and the indicator function \( \mathbb{I}[\cdot] \) returns \(1\) when the argument is true, and \(0\) otherwise.

Given an example \((x, y)\), the margin of \( h \in \mathcal{H} \) is defined as \( yh(x) \), which can be viewed as a measure of the confidence of the classification. We further define the average margin of \( h \in \mathcal{H} \) over distribution \( \mathcal{D} \) as

\[
\theta_h = E_{(x,y) \sim \mathcal{D}}[yh(x)].
\]

We also introduce the empirical Rademacher complexity [Bartlett and Mendelson, 2002] to measure the complexity of function space \( \mathcal{H} \) as follows:

\[
\hat{\mathcal{R}}_n(\mathcal{H}) = E_{\sigma_1, \sigma_2, \ldots, \sigma_n}[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i)],
\]

where each \( \sigma_i \) is a Rademacher variable with \( \Pr[\sigma_i = +1] = \Pr[\sigma_i = -1] = 1/2 \) for \( i \in [n] \).

We finally introduce some notations used in this work. Write \([d] = \{1, 2, \ldots, d\}\) for integer \( d > 0 \), and \( \langle w, x \rangle \) represents the inner product of \( w \) and \( x \). Let \( I_d \) be the identity matrix of size \( d \times d \), and denote by \( \top \) the transpose of vectors or matrices. For positive \( f(n) \) and \( g(n) \), we write \( f(n) = O(g(n)) \) if \( g(n)/f(n) \to c \) for constant \( c < +\infty \).

3 Theoretical Analysis
We begin with the squared margin loss as follows:

**Definition 1.** For \( \theta > 0 \), we define the squared margin loss \( \ell_\theta \) with respect to function \( h \in \mathcal{H} \) as

\[
\ell_\theta(h, (x, y)) = \left[(1 - yh(x))/\theta \right]_+^2,
\]

where \((a)_+ = \max(0, a)\).

This is a simple extension from the traditional margin loss [Bartlett and Mendelson, 2002], while we consider the squared loss and unbounded constraint for the negative \( yh(x) \). The margin parameter \( \theta \) is generally irrelevant to learned function \( h \) and data distribution in most previous theoretical and algorithmic studies.

In this work, we select margin parameter \( \theta \) as the average margin when \( \theta_h > 0 \), to correlate generalization performance with margin distribution, that is,

\[
\theta = \theta_h = E_D[yh(x)],
\]

which is dependent on data distribution and learned function. Given training sample \( S_n \), we try to learn a function \( h \) by minimizing the squared margin loss as follows:

\[
\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\theta_h} \left[1 - \frac{y_i h(x_i)}{\theta_h}\right]_+\right)^2.
\]

For simplicity, we further introduce the notion of margin semi-variance [Markowitz, 1952] as follows:

**Definition 2.** Given function \( h \in \mathcal{H} \) and training sample \( S_n \), we define the margin semi-variance as

\[
SV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\theta_h} \left[1 - \frac{y_i h(x_i)}{\theta_h}\right]_+\right)^2,
\]

where \( \theta_h \) denotes the average margin defined by Eqn. (1).

The margin semi-variance essentially counts the average of squared deviation between average margin and the margins \( y_i h(x_i) \), which are smaller than average margin. This yields an equivalent expression for Eqn. (2) as follows:

\[
\min_{h \in \mathcal{H}} \frac{1}{\theta_h} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\theta_h} \left[1 - \frac{y_i h(x_i)}{\theta_h}\right]_+\right)^2\right),
\]

that is, optimizing the squared margin loss with parameter \( \theta = \theta_h \) is equivalent to minimizing margin semi-variance and maximizing average margin simultaneously.

For most real applications, we could learn some relatively-good functions from sufficient training data. Motivated from the notion of weak learner in boosting [Freund and Schapire, 1996], we formally define the set of relatively-good functions for function space \( \mathcal{H} \) as follows:

\[
\mathcal{H}_\nu = \{h \in \mathcal{H}, \theta_h \geq \nu\} \text{ for some small constant } \nu > 0.
\]

Essentially, a relatively-good function is similar to a weak learner, which achieves slightly better performance than the randomly-guessed classifier.

We now present the main theoretical result as follows:

**Theorem 1.** For small constant \( \nu > 0 \), let \( \mathcal{H} \) be a function space with relative-good set \( \mathcal{H}_\nu \). For any \( \delta \in (0, 1) \) and for every \( h \in \mathcal{H}_\nu \), the following holds with probability at least \( 1 - \delta \) over the training sample \( S_n \):

\[
\mathcal{E}(h) \leq \frac{SV(h)}{\theta_h^2} + O \left(\frac{\hat{\mathcal{R}}_n(\mathcal{H}_\nu)}{\theta_h^2} + \sqrt{\frac{1}{2n} \ln \frac{4n}{\delta}}\right),
\]

with empirical Rademacher complexity \( \hat{\mathcal{R}}_n(\mathcal{H}_\nu) \leq \hat{\mathcal{R}}_n(\mathcal{H}) \).
This theorem presents a new generalization error bound, which is heavily relevant to margin distribution by incorporating factors such as average margin and semi-variance. This could shed some new insights on the design of algorithms on the optimization of margin distribution as shown in Section 4.

The proof follows the empirical Rademacher complexity [Bartlett and Mendelson, 2002], while the challenge lies in the distribution-dependent average margin \( \theta_h \). We solve it by constructing a sequence of intervals for average margin \( \theta_h \), and the detailed proof is presented in [Qian et al., 2022].

It remains to study the empirical Rademacher complexity in Theorem 1, and we focus on linear and kernel functions. For instance space \( \mathcal{X} = \{ x \in \mathbb{R}^d : \| x \| \leq r \} \) and linear function space \( \mathcal{H} = \{ h(x) = w \top x : \| w \| \leq \Lambda \} \), let \( \mathcal{H}_i \) denote the set of relatively-good classifiers. We upper bound the empirical Rademacher complexity as

\[
\hat{R}_{S_n}(\mathcal{H}_i) \leq \hat{R}_{S_n}(\mathcal{H}) \leq r\Lambda / \sqrt{n}
\]

from the work of [Shalev-Shwartz and Ben-David, 2014]. For kernel function \( \kappa(\cdot, \cdot) \), we have the kernel function space

\[
\mathcal{H} = \{ h(x) = \sum_{i=1}^n a_i \kappa(x_i, x) : \sum_{i,j=1}^n a_ia_j \kappa(x_i, x_j) \leq \Lambda^2 \}.
\]

We could upper bound the empirical Rademacher complexity for kernel functions, from [Bartlett and Mendelson, 2002],

\[
\hat{R}_{S_n}(\mathcal{H}_i) \leq \hat{R}_{S_n}(\mathcal{H}) \leq 2\Lambda^2 / n \left( \sum_{i=1}^n \kappa(x_i, x_i) \right)^{1/2}.
\]

It is also noteworthy that the average margin \( \theta_i \) is unknown on the design of new algorithms because of the unknown distribution \( \mathcal{D} \), and we resort to the empirical average margin from training sample \( S_n \) in practice.

4 The MsvMAv Approach

Motivated from Theorem 1, this section develops the MsvMAv approach on the optimization of margin distribution, and we focus on linear and kernel functions.

4.1 Linear Functions

For linear space \( \mathcal{H} = \{ h_w(x) = (w, x) : \| w \| = 1 \} \) and training sample \( S_n \), we have the empirical average margin

\[
\hat{\theta}_w = \frac{1}{n} \sum_{i=1}^n y_i(w, x_i).
\]

For simplicity, we omit a bias term on the design of algorithm, and we will augment \( w \) and instance \( x \) with bias term \( b \) and \( 1 \) in experiments, respectively, as shown in Section 5. Our optimization problem can be formally written as

\[
\min_{\| w \| = 1} \left\{ \frac{SV(w)}{\hat{\theta}_w} \right\},
\]

where empirical average margin \( \hat{\theta}_w > 0 \) and empirical margin semi-variance \( SV(w) = \sum_{i=1}^n (\hat{\theta}_w - y_i(w, x_i))^2 / n \).

Obviously, this is a non-convex optimization problem, and we would optimize the empirical margin semi-variance and average margin alternatively.

Initialize the linear function \( w_0 \) by optimizing empirical average margin, that is,

\[
w_0 = \arg \max_{\| w \| = 1} \frac{1}{n} \sum_{i=1}^n y_i(w, x_i) = \frac{1}{n} \sum_{i=1}^n \| \sum_{i=1}^n y_i x_i \|_2,
\]

where we solve \( w_0 \) from its dual problem using Lagrangian function, and the details are given by Qian et al. [2022].

Optimization of Empirical Margin Semi-variance

In the \( k \)-th iteration \( (k \geq 1) \) with previous classifier \( w_{k-1} \), we first calculate the empirical average margin \( \hat{\theta}_{w_{k-1}} \) as

\[
\hat{\theta}_{w_{k-1}} = \frac{1}{n} \sum_{i=1}^n y_i(w_{k-1}, x_i).
\]

We then introduce the minimization of empirical margin semi-variance as follows:

\[
\min_{w} \left\{ \frac{1}{n} \sum_{i=1}^n (\hat{\theta}_{w_{k-1}} - y_i(w, x_i))^2 + \beta_k \| w - w_{k-1} \|^2_2 \right\},
\]

where \( \beta_k \) is a proximal regularization parameter. We now introduce the following index set, to present a closed-form solution for the above minimization,

\[
A_k = \{ i : y_i(w_{k-1}, x_i) < \hat{\theta}_{w_{k-1}} \text{ for } i \in [n] \},
\]

i.e., the index set of instance with margins below the empirical average margin \( \hat{\theta}_{w_{k-1}} \). We can rewrite the minimization of empirical margin semi-variance as

\[
\min_{w} \left\{ \sum_{i \in A_k} (\hat{\theta}_{w_{k-1}} - y_i(w, x_i))^2 + \beta_k \| w - w_{k-1} \|^2_2 \right\}.
\]

Denote by \( w'_k \) the minimizer of the above problem, and we obtain the closed-form solution for \( w'_k \) as follows

\[
\left( I_d + \sum_{i \in A_k} x_ix_i^\top / n\beta_k \right)^{-1} \left( \hat{\theta}_{w_{k-1}} \sum_{i \in A_k} y_i x_i + w_{k-1} \right).
\]

One problem is to calculate the inverse in Eqn. (6), which takes \( O(d^3) \) computational costs (\( d \) is dimensionality). This remains one challenge to deal with high-dimensional tasks.

We now present an efficient method to calculate the inverse in Eqn. (6). For simplicity, we denote by

\[
M_k = \left( I_d + \sum_{i \in A_k} x_ix_i^\top / n\beta_k \right)^{-1} \text{ for } k = 1, 2, \ldots ,
\]

and it is easy to derive the following recursive relation:

\[
M_k^{-1} = M_{k-1}^{-1} - \sum_{i \in A_k \setminus A_{k-1}} x_ix_i^\top / n\beta + \sum_{i \in A_k \setminus A_{k-1}} x_ix_i^\top / n\beta,
\]

with \( M_0 = I_d \).

We calculate \( M_k \) efficiently from \( M_{k-1} \) and Sherman-Morrison formula [Sherman and Morrison, 1950]. In other
Algorithm 1 The MsvMAV Approach

Input: Training sample $S_n$, iteration number $T$, and proximal parameters $\alpha_k$ and $\beta_k$

Output: $w$

1: Initialize $M_0 = I_d$, $A_0 = \emptyset$ and $w_0$ by Eqn. (3)
2: for $k = 1, 2, \cdots, T$ do
3: Compute empirical average margin $\hat{\theta}_{w_{k-1}}$ by Eqn. (4)
4: Solve the index set $A_k$ by Eqn. (5)
5: Compute $M_k = (I_d + \sum_{i \in A_k} x_i x_i^\top / n \beta_k)^{-1}$ by Eqns. (7) and (8)
6: Compute the minimizer $w'_k$ for empirical margin semi-variance by Eqn. (9)
7: Solve the empirical average margin maximizer $w_k$ by Eqn. (10), and normalize $w_k = w_k / \|w_k\|_2$
8: end for
9: return $w = w_T$

words, we initialize $M' = M_{k-1}$, and make the following updates iteratively, based on Sherman-Morrison formula,

$$M' = M' - \frac{M' x_i x_i^\top M'}{x_i^\top M' x_i - n \beta_k} \quad \text{for } i \in A_{k-1} \setminus A_k,$$  \hspace{1cm} (7)

$$M' = M' - \frac{M' x_i x_i^\top M'}{x_i^\top M' x_i + n \beta_k} \quad \text{for } i \in A_k \setminus A_{k-1}.$$  \hspace{1cm} (8)

We then obtain $M_k = M'$, and the minimizer of empirical margin semi-variance is given by

$$w'_k = M' \left( \frac{\hat{\theta}_{w_{k-1}}}{n \beta_k} \sum_{i \in A_k} y_i x_i + w_{k-1} \right).$$  \hspace{1cm} (9)

Optimization of Empirical Average Margin

We now study the maximization of empirical average margin, which can be formalized as:

$$w_k = \arg \min_w \left\{ - \frac{1}{n} \sum_{i=1}^n y_i \langle w, x_i \rangle + \alpha_k \|w - w'_k\|_2^2 \right\},$$

where $\alpha_k$ is a proximal regularization parameter. It is easy to obtain the closed-form solution as follows:

$$w_k = w'_k + \frac{1}{2 \alpha_k n} \sum_{i=1}^n y_i x_i.$$  \hspace{1cm} (10)

We obtain $w_k = w_k / \|w_k\|$ in the $k$-th iteration. Algorithm 1 presents a detailed description of our MsvMAV approach.

4.2 Kernelization

This section focuses on kernel mapping $\phi : \mathcal{X} \rightarrow \mathbb{H}$ for Hilbert space $\mathbb{H}$, we consider $h(x) = \langle w, \phi(x) \rangle$ with $w \in \mathbb{H}$ and $\phi(x) \in \mathbb{H}$. The optimization problem is given by

$$\min_w \left\{ \frac{\tilde{S}V(w)}{\hat{\theta}_w} \right\},$$

where average margin $\hat{\theta}_w = \sum_{i=1}^n y_i \langle w, \phi(x_i) \rangle / n > 0$, and margin semi-variance

$$\tilde{S}V(w) = \frac{1}{n} \sum_{i=1}^n \left[ (\hat{\theta}_w - y_i \langle w, \phi(x_i) \rangle)^+ \right]^2.$$
where $\beta_k$ is a proximal regularization parameter. We introduce the index set $A_k = \{i \in [n] : \langle y_i, K(a) \rangle < \theta_{a_k} \}$, and obtain the empirical margin semi-variance minimizer

$$a'_k = M_k \left( (K + I_n)a_{k-1} + \frac{1}{n} \sum_{i \in A_k} y_i K_i \right).$$

where we use the Sherman-Morrison formula to calculate

$$M_k = \left( \sum_{i \in A_k} \frac{K_i K_i^\top}{n \beta} + K + I_n \right)^{-1}.$$

We finally maximize the empirical average margin based on the following optimization problem:

$$\min_{a_k} \left\{ -\frac{1}{n} \sum_{i=1}^{n} y_i \langle K_i, a_k \rangle + \alpha_k (a_k - a'_k)^\top K (a_k - a'_k) \right\},$$

where $\alpha_k$ is a proximal regularization parameter, and it is easy to get the closed-form solution as follows:

$$a_k = a'_k + \frac{y_1, y_2, \ldots, y_n}{2 \alpha_k n}.$$

We get the final $a_k = a_k/\|a_k\|_K$ in the $k$-th iteration.

## 5 Empirical Study

In this section, we present extensive empirical studies to verify the effectiveness of our proposed MsVMAv approach. We consider 30 datasets, including 20 regular and 10 large-scale datasets. The number of instances varies from 208 to 88588 while the feature dimensionality ranges from 2 to 1836, covering a wide range of properties. The statistics for all datasets can be found in [Qian et al., 2022].

We compare our proposed MsVMAv approach with state-of-the-art algorithms on large-margin and margin distribution optimization: 1) SVM [Boser et al., 1992], 2) SVR [Drucker et al., 1997] with binary targets, 3) LSSVM [Suykens et al., 2002], 4) MAMC [Pelckmans et al., 2007], 5) ODM [Zhang and Zhou, 2019], 6) FMM [Ji et al., 2021]. The details of compared algorithms can be found in [Qian et al., 2022].

For each dataset, we scale all features into the interval $[0, 1]$, and augment each instance $x$ with constant 1 for the bias of linear model. The empirical average margin $\bar{\theta}_w$ may be smaller than zero in experiments, when the proximal regularization parameter $\beta_w$ is set too small. In such case, we take the opposite model $-w$ so as to keep the positiveness of empirical average margin.

For our MsVMAv approach, parameters $\alpha_k$ and $\beta_k$ are set
to be constant and selected by 5-fold cross validation from \( \{2^{-10}, 2^{-8}, \ldots, 2^{10}\} \), and the width of Gaussian kernel is chosen from \( \{2^{-10}/d, 2^{-8}/d, \ldots, 2^{10}/d\} \). We select the maximum iteration number \( T = 100 \) as a stopping criteria for MsvM\( \text{Av} \). For SVM, SVR, LSSVM and ODM, we set regularization parameter \( C \in \{2^{-10}, 2^{-8}, \ldots, 2^{10}\} \) by 5-fold cross validation again, and the others are set according to their respective references, also shown in [Qian et al., 2022].

We first compare the margin distributions of our proposed MsvM\( \text{Av} \) approach with other algorithms. Figure 1 illustrates the cumulative margin distributions of different algorithms on four datasets, and similar trends can be observed on other datasets. As can be seen, the curves of our MsvM\( \text{Av} \) approach generally lie on the rightmost side, which shows the margin distributions of MsvM\( \text{Av} \) are generally better than that of SVM, ODM and FMM.

We further analyze the generalization performance of our proposed MsvM\( \text{Av} \) approach with other compared algorithms. All experiments are conducted by repeating 30 times of random partitions of datasets with 80% and 20% of data for training and testing, respectively. The test accuracies are obtained by averaging over 30 times. Tables 1 and 2 show the empirical results of our MsvM\( \text{Av} \) and other algorithms with linear and Gaussian kernel functions, respectively.

From Tables 1 and 2, our proposed MsvM\( \text{Av} \) approach takes significantly better performance than other algorithms for linear and kernel functions, since win/tie/loss counts show that our approach wins for most datasets, and rarely loses. One intuitive explanation is that our MsvM\( \text{Av} \) approach achieves better margin distribution by maximizing the empirical average margin and minimizing empirical margin semi-variance, as shown in Figure 1. SVM, SVR and FMM maximize the minimum margin, which ignores the margin distribution. LSSVM and MAMC essentially maximize average margin only, which fails to learn from other margin statistics. ODM takes the average margin and margin variance into consideration, but the process of margin variance minimization could constrain some large margins.

This section omits partial empirical results due to the page limit, including the empirical curves of margin distributions, as well as running time comparisons for our MsvM\( \text{Av} \) and other compared algorithms. Relevant results can be found in our full work [Qian et al., 2022].

### 6 Conclusion

Large margin has been one of the most important principles on the design of algorithms in machine learning, and recent empirical studies show new insights on the optimization of margin distribution yet without theoretical supports. This work takes one step on this direction by providing a new generalization error bound, which is heavily relevant to margin distribution by incorporating factors such as average margin and semi-variance. Based on the theoretical results, we develop the MsvM\( \text{Av} \) approach for margin distribution optimization, and extensive experiments verify its superiority. An interesting future work is to exploit more effective statistics to characterize the whole margin distribution.

### Acknowledgements

The authors want to thank the reviewers for helpful comments and suggestions. This work is partially supported by the NSFC (61921006, 61876078, 62006088). Wei Gao is the corresponding author.

---

| Dataset  | MsvM\( \text{Av} \) | SVM | SVR | LSSVM | ODM | MAMC | FMM |
|----------|------------------|-----|-----|-------|-----|------|-----|
| advertise | .9837 ± .0014    | .9820 ± .0023 | .9835 ± .0019 | .9848 ± .0016 | .9838 ± .0016 | .9547 ± .0026 | .9810 ± .0028 |
| australian | .8580 ± .0078    | .8225 ± .0087 | .8210 ± .0080 | .8357 ± .0111 | .8329 ± .0088 | .8302 ± .0122 | .8237 ± .0060 |
| bibtex   | .7554 ± .0036    | .7506 ± .0050 | .7508 ± .0053 | .7505 ± .0056 | .7557 ± .0045 | .6714 ± .0397 | .7509 ± .0050 |
| biodeg   | .8834 ± .0081    | .8687 ± .0115 | .8580 ± .0089 | .8712 ± .0093 | .8845 ± .0095 | .8559 ± .0098 | .8915 ± .0096 |
| breastw | .9783 ± .0013    | .9710 ± .0013 | .9703 ± .0050 | .9713 ± .0018 | .9710 ± .0013 | .9714 ± .0022 | .9710 ± .0013 |
| diabetes | .7576 ± .0074    | .7574 ± .0094 | .7502 ± .0090 | .7411 ± .0112 | .7504 ± .0082 | .6779 ± .0197 | .7385 ± .0105 |
| emotions | .7986 ± .0122    | .7899 ± .0156 | .7756 ± .0164 | .8115 ± .0124 | .8101 ± .0147 | .7952 ± .0120 | .8078 ± .0126 |
| german   | .7465 ± .0099    | .7443 ± .0096 | .7198 ± .0174 | .7353 ± .0143 | .7285 ± .0156 | .7198 ± .0154 | .7277 ± .0127 |
| halloffame | .9677 ± .0025    | .9625 ± .0020 | .9578 ± .0033 | .9523 ± .0060 | .9617 ± .0025 | .9510 ± .0028 | .9590 ± .0034 |
| hill-valley | .6826 ± .0184    | .6668 ± .0177 | .6482 ± .0135 | .7116 ± .0678 | .5950 ± .0360 | .5402 ± .0320 | .7886 ± .0299 |
| k1c      | .8739 ± .0042    | .8701 ± .0057 | .8689 ± .0045 | .8703 ± .0044 | .8716 ± .0055 | .8764 ± .0027 | .8706 ± .0038 |
| parkinson | .9573 ± .0223    | .9282 ± .0203 | .9393 ± .0193 | .9393 ± .0224 | .9385 ± .0157 | .9214 ± .0174 | .9402 ± .0167 |
| pbcseq   | .7350 ± .0143    | .7214 ± .0152 | .7317 ± .0151 | .7312 ± .0165 | .7269 ± .0221 | .7076 ± .0149 | .7238 ± .0184 |
| sleepdata | .7407 ± .0129    | .7192 ± .0105 | .7211 ± .0061 | .7037 ± .0083 | .7050 ± .0083 | .7055 ± .0056 | .7182 ± .0094 |
| students | .8977 ± .0119    | .8920 ± .0079 | .8665 ± .0098 | .8898 ± .0072 | .8805 ± .0142 | .6543 ± .0185 | .8993 ± .0062 |
| titanic  | .7825 ± .0048    | .7823 ± .0052 | .7823 ± .0052 | .7822 ± .0052 | .7767 ± .0075 | .7873 ± .0052 | .7823 ± .0052 |
| tokyo1   | .9406 ± .0037    | .9241 ± .0050 | .9257 ± .0060 | .9337 ± .0054 | .9253 ± .0053 | .9229 ± .0039 | .9248 ± .0050 |
| vehicle  | .9793 ± .0083    | .9795 ± .0090 | .9856 ± .0073 | .9899 ± .0070 | .9722 ± .0088 | .9625 ± .0105 | .9805 ± .0093 |
| vertebr  | .8280 ± .0240    | .7898 ± .0098 | .7957 ± .0183 | .8108 ± .0176 | .8000 ± .0122 | .7769 ± .0126 | .7962 ± .0153 |
| wdbc     | .9819 ± .0081    | .9810 ± .0056 | .9772 ± .0066 | .9842 ± .0035 | .9795 ± .0052 | .9526 ± .0089 | .9526 ± .0089 |

### Table 2: Comparisons of the test accuracies (mean±std.) on 20 datasets. We use Gaussian kernel for all algorithms. */ indicates that our MsvM\( \text{Av} \) approach is significantly better/worse than the corresponding algorithms (pairwise t-tests at 95% significance level).
References

[Aiolli et al., 2008] F. Aiolli, G. Da San Martino, and A. Sperduti. A kernel method for the optimization of the margin distribution. In ICANN, pages 305–314, 2008.

[Bartlett and Mendelson, 2002] P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. JMLR, 3:463–482, 2002.

[Bartlett and Shawe-Taylor, 1999] P. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. Advances in Kernel Methods: Support Vector Learning, pages 43–54, 1999.

[Bartlett et al., 2017] P. Bartlett, D. J. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. In NIPS 30, pages 6240–6249, 2017.

[Boser et al., 1992] B. E. Boser, I. M. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In COLT, pages 144–152, 1992.

[Breiman, 1999] L. Breiman. Prediction games and arcing algorithms. Neural Comput., 11(7):1493–1517, 1999.

[Cheng et al., 2016] F.-Y. Cheng, J. Zhang, and C.-H. Wen. Cost-sensitive large margin distribution machine for classification of imbalanced data. Pattern Recognit. Lett., 80:107–112, 2016.

[Cortes and Vapnik, 1995] C. Cortes and V. Vapnik. Support-vector networks. Mach. Learn., 20(3):273–297, 1995.

[Drucker et al., 1997] H. Drucker, J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In NIPS 9, pages 155–161, 1997.

[Freund and Schapire, 1996] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In ICML, pages 148–156, 1996.

[Gao and Zhou, 2013] W. Gao and Z.-H. Zhou. On the doubt about margin explanation of boosting. Artif. Intell., 203:1–18, 2013.

[Garg and Roth, 2003] A. Garg and D. Roth. Margin distribution and learning. In ICML, pages 210–217, 2003.

[Grønlund et al., 2020] A. Grønlund, L. Kamma, and K. G. Larsen. Near-tight margin-based generalization bounds for support vector machines. In ICML, pages 3779–3788, 2020.

[Ji et al., 2021] Z.-W. Ji, N. Srebro, and M. Telgarsky. Fast margin maximization via dual acceleration. In ICML, pages 4860–4869, 2021.

[Jiang et al., 2019] Y. Jiang, D. Krishnan, H. Mobahi, and S. Bengio. Predicting the generalization gap in deep networks with margin distributions. In ICLR, 2019.

[Kabán and Durrant, 2020] Ata Kabán and Robert J Durrant. Structure from randomness in halfspace learning with the zero-one loss. JAIR, 69:733–764, 2020.

[Markowitz, 1952] Harry Markowitz. Portfolio selection. Journal of Finance, 7(1):77–91, 1952.

[Pelckmans et al., 2007] K. Pelckmans, J. A. K. Suykens, and B. De Moor. A risk minimization principle for a class of parzen estimators. In NIPS 20, pages 1137–1144, 2007.

[Qian et al., 2022] M.-Z. Qian, Z. Ai, T. Zhang, and W. Gao. On the optimization of margin distribution. CoRR, abs/2204.14118, 2022.

[Rastogi et al., 2020] R. Rastogi, P. Anand, and S. Chandra. Large-margin distribution machine-based regression. Neural Comput. Appl., 32(8):3633–3648, 2020.

[Rosset et al., 2003] S. Rosset, J. Zhu, and T. Hastie. Margin minimizing loss functions. In NIPS 16, pages 1237–1244, 2003.

[Schapire et al., 1998] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. Ann. Stat., 26(5):1651–1686, 1998.

[Schölkopf et al., 2002] B. Schölkopf, A. J. Smola, and F. Bach. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, 2002.

[Shalev-Shwartz and Ben-David, 2014] S. Shalev-Shwartz and S. Ben-David. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, Cambridge, 2014.

[Sherman and Morrison, 1950] J. Sherman and W. J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. Ann. Math. Stat., 21(1):124–127, 1950.

[Shivaswamy and Jebara, 2010] P. K. Shivaswamy and T. Jebara. Maximum relative margin and data-dependent regularization. JMLR, 11(2), 2010.

[Sokolić et al., 2017] J. Sokolić, R. Giryes, G. Sapiro, and M. R. Rodrigues. Robust large margin deep neural networks. IEEE Trans. Signal Process., 65(16):4265–4280, 2017.

[Suykens et al., 2002] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. Least Squares Support Vector Machines. World Scientific, Singapore, 2002.

[Vapnik, 1982] V. Vapnik. Estimation of Dependences based on Empirical Data. Springer-Verlag, New York, 1982.

[Wei and Ma, 2020] C. Wei and T.-Y. Ma. Improved sample complexities for deep neural networks and robust classification via an all-layer margin. In ICLR, 2020.

[Weinstein et al., 2020] B. Weinstein, S. Fine, and Y. Hel-Or. Margin-based regularization and selective sampling in deep neural networks. CoRR, abs/2009.06011, 2020.

[Zhang and Zhou, 2014] T. Zhang and Z.-H. Zhou. Large margin distribution machine. In KDD, pages 313–322, 2014.

[Zhang and Zhou, 2019] T. Zhang and Z.-H. Zhou. Optimal margin distribution machine. IEEE Trans. Knowl. Data Eng., 32(6):1143–1156, 2019.