ECE 599 - Project
Bias-Variance Analysis of Support Vector Machines for the Development of SVM-Based Ensemble Methods

An Vuong, Rong Yu

October 2, 2024

Abstract

This document contains the report for the Winter 2024 ECE599 - Statistical learning course project. It goes through the insights and derivations mentioned in the paper *Bias-Variance Analysis of Support Vector Machines for the Development of SVM-Based Ensemble Methods.*

1 Introduction

In this section, we briefly summarize the main contributions of [VD04]. We will go through the detailed derivations in the following sections.

The paper used the bias-variance decomposition proposed in [Dom00] to characterize the changes in bias and variance with respect to the hyperparameters of SVM, such as kernel types and margin cost. The authors achieved this by proposing methods to empirically estimate bias and variance using both synthesized and real-world datasets. This information was then used to guide the design of a new ensemble method for SVM-based classifiers. This new method is called Low Bias Bagged SVMs, which is a heuristic way to select base SVM classiers that have lower biases, the ensemble of these classifiers showed promising results across a wide range of tests.

In the next sections, we will provide details about the following:

1. Bias-Variance decomposition from [Dom00] and its application to 0/1 loss
2. Generalization bound of SVM
3. Generalization bound of Kernelized SVM

2 Definitions & Notations

We introduce our notations in this section. We deliberately modified some notations compared to the papers, to make them more similar to what was taught in class. We consider the following setting:

1. Training data points $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \mathcal{C} =$ set of all class labels
2. $(\mathbf{x}, y) \overset{\text{iid}}{\sim} P_{\mathbf{X},Y}(\mathbf{x}, y)$
3. Training set $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$
4. Denote $\mathcal{L}$ as some learning algorithm. Training on dataset D, this will produce a classier $f_D = \mathcal{L}(D)$. Then $\hat{y} = f_D(\mathbf{x})$ is the model prediction.
5. We use the 0/1 loss $l(y, \hat{y}) = I(y \neq \hat{y})$
6. The expected loss of $\mathcal{L}$ is then given by:

$$\begin{aligned}\mathbb{E}[\mathcal{L}] &= \mathbb{E}_D\left[\mathbb{E}_{Y|\mathbf{X}}[l(y, \hat{y})]\right] && (1)\\ &= \mathbb{E}_D\left[\mathbb{E}_{Y|\mathbf{X}}[l(y, f_D(\mathbf{x}))]\right] && (2)\end{aligned}$$

   In other words, this is simply our empirical risk restricted to function class $\mathcal{F}$ where f belongs. $\mathcal{L}$ is some algorithm that will pick the best $f \in \mathcal{F}$ according to some metrics.

Here we also define some needed quantities:

1. The optimal prediction (Bayes' prediction) restricted to class $\mathcal{F}$:

$$y^*(\mathbf{x}) = \arg\min_{\hat{y}\in\mathcal{F}} \mathbb{E}_{Y|\mathbf{X}}[l(y,\hat{y})] \tag{3}$$
$$= \arg\min_{\hat{y}\in\mathcal{F}} \mathbb{E}_{Y|\mathbf{X}}[l(y,f_D(\mathbf{x}))] \tag{4}$$
$$= \arg\min_{\hat{y}\in\mathcal{F}} \mathbb{E}_{Y|\mathbf{X}}[I(y \neq f_D(\mathbf{x}))] \tag{5}$$

   This is the mode of $P_{\mathbf{X}|Y}(\mathbf{x}|y)$.

2. The main prediction of $f_D(.)$ is:

$$\hat{y}_m = \arg\min_{y'\in\mathcal{C}} \mathbb{E}_D[l(\hat{y},y')] \tag{6}$$
$$= \arg\min_{y'\in\mathcal{C}} \mathbb{E}_D[l(f_D(\mathbf{x}),y')] \tag{7}$$
$$= \arg\min_{y'\in\mathcal{C}} \mathbb{E}_D[I(f_D(\mathbf{x}) \neq y')] \tag{8}$$

   For 0/1 loss, this is the class predicted most often by $f_D(.)$, i.e. the "mode" of classifier f_D.

3. Utilizing the optimal prediction, the noise is defined as:

$$N(\mathbf{x}) = \mathbb{E}_{Y|\mathbf{X}}[l(y,y^*)] \tag{9}$$

   This is the amount of error inherent to the data, and cannot be reduced by any learning algorithms.

From these, we can define bias and variance as follows:

1. Bias, whether the "mode" of f_D matches the optimal prediction as point $\mathbf{x}$:

$$B(\mathbf{x}) = l(y^*,\hat{y}_m) \tag{10}$$
$$= I(y^* \neq \hat{y}_m) \tag{11}$$

2. Variance, how different the prediction is across different samples of traing dataset:

$$V(\mathbf{x}) = \mathbb{E}_D[l(\hat{y}_m,\hat{y})] \tag{12}$$
$$= \mathbb{E}_D[I(\hat{y}_m \neq f_D(\mathbf{x}))] \tag{13}$$

   Note that $B(\mathbf{x})$ and $V(\mathbf{x})$ are point-wise measures.

3 Effects of bias and variance

From the definitions of bias and variance, there are two important observations:

1. When bias is zero (unbiased), i.e. the "mode" of the classifier $f_D(.)$ agrees with the Bayes' prediction, more variance and noise will increase the error, since now we have $\hat{y} \neq \hat{y}_m = y^*$.

2. In the other direction, when bias is non-zero (biased), more variance and noise can actually reduce the error, since it can push $\hat{y}$ close to y^*. Note that for 0/1 loss, biased means $B(\mathbf{x}) = 1$.

These observations are summarized in Figure 1. From this analysis, it would be helpful if we can decompose the expected loss (risk) $\mathbb{E}_D\left[\mathbb{E}_{Y|\mathbf{X}}[l(y,\hat{y})]\right]$ in terms of $B(\mathbf{x})$ and $V(\mathbf{x})$. [Dom00] provided this decomposition and its details are presented in the following section.

4 Domingo's Bias-Variance Decomposition

In [Dom00], the author proposed the following decomposition, which holds for certain loss functions l:

$$\mathbb{E}[\mathcal{L}] = \mathbb{E}_D\left[\mathbb{E}_{Y|\mathbf{X}}[l(y,f_D(\mathbf{x}))]\right] \tag{14}$$
$$= c_1 N(\mathbf{x}) + B(\mathbf{x}) + c_2 V(\mathbf{x}) \tag{15}$$

Here we show the proof for the 0/1 loss in binary classification problems, (15) is valid with $c_1 = P_D(\hat{y} = y^*) - 1$, $c_2 = 1$ if $\hat{y}_m = y^*$, and $c_2 = -1$ otherwise. $P_D(\hat{y} = y^*)$ denotes the probability over the training set D that $\hat{y}$ produced by $\mathcal{L}$ coincides with the Bayes' prediction y^*.

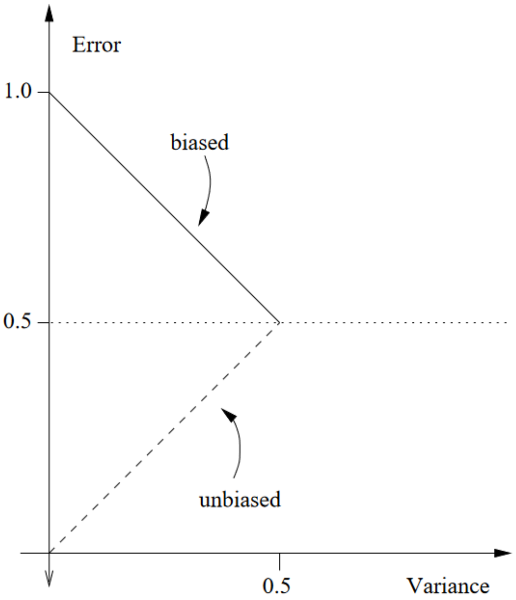

Figure 1: Effects of biased and unbiased variance on the error. The unbiased variance increases, while the biased variance decreases the error. [VD04]

Proof.
We begin by showing:

$$\mathbb{E}_{Y|\mathbf{X}}[l(y,\hat{y})] = l(y^*,\hat{y}) + c_0\mathbb{E}_{Y|\mathbf{X}}[l(y^*,y)] \tag{16}$$

with $c_0 = 1$ if $\hat{y} = y^*$ and $c_0 = -1$ if $\hat{y} \neq y^*$.

If $\hat{y} = y^*$, $l(y^*,\hat{y}) = I(y^* \neq \hat{y}) = 0$ and $\mathbb{E}_{Y|\mathbf{X}}[l(y^*,y)] = \mathbb{E}_{Y|\mathbf{X}}[l(\hat{y},y)]$ thus $c_0 = 1$ is trivially true.

If $\hat{y} \neq y^*$, recall that this is binary classification, in this case if $y \neq y^*$, this implies $\hat{y} = y$ and vice versa. Hence $P_{Y|\mathbf{X}}(y = \hat{y}) = P_{Y|\mathbf{X}}(y \neq y^*)$. This gives:

$$\begin{aligned}
\mathbb{E}_{Y|\mathbf{X}}[l(y,\hat{y})] &= \mathbb{E}_{Y|\mathbf{X}}[I(y \neq \hat{y})] = P_{Y|\mathbf{X}}(y \neq \hat{y}) && (17)\\
&= 1 - P_{Y|\mathbf{X}}(y = \hat{y}) && (18)\\
&= 1 - P_{Y|\mathbf{X}}(y \neq y^*) && (19)\\
&= 1 - \mathbb{E}_{Y|\mathbf{X}}[l(y,y^*)] && (20)\\
&= l(y^*,\hat{y}) - \mathbb{E}_{Y|\mathbf{X}}[l(y,y^*)] \quad (\text{note that } y^* \neq \hat{y} \to l(y^*,\hat{y}) = 1) && (21)
\end{aligned}$$

Equation (21) implies $c_0 = -1$ when $\hat{y} \neq y^*$, thus complete the proof for (16).

In the same manner, we can show that:

$$\mathbb{E}_D[l(y^*,\hat{y})] = l(y^*,\hat{y}_m) + c_2\mathbb{E}_D[l(\hat{y}_m,\hat{y})] \tag{22}$$

with $c_2 = 1$ if $\hat{y}_m = y^*$ and $c_2 = -1$ if $\hat{y}_m \neq y^*$.

Apply equation (16) to (14), we have:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}] &= \mathbb{E}_D\left[\mathbb{E}_{Y|\mathbf{X}}[l(y,f_D(\mathbf{x}))]\right] = \mathbb{E}_D\left[\mathbb{E}_{Y|\mathbf{X}}[l(y,\hat{y})]\right] && (23)\\
&= \mathbb{E}_D\left[l(y^*,\hat{y}) + c_0\mathbb{E}_{Y|\mathbf{X}}[l(y^*,y)]\right] && (24)\\
&= \mathbb{E}_D[l(y^*,\hat{y})] + \mathbb{E}_D[c_0]\mathbb{E}_{Y|\mathbf{X}}[l(y^*,y)] \quad (l(y^*,y) \text{ is independent of } D) && (25)
\end{aligned}$$

We have:

$$\begin{aligned}
\mathbb{E}_D[c_0] &= P_D(\hat{y} = y^*) \times 1 + P_D(\hat{y} \neq y^*) \times (-1) && (26)\\
&= 2P_D(\hat{y} = y^*) - 1 = c_1 && (27)
\end{aligned}$$

This completes the proof for (15). □

5 Generalization bound of SVM

5.1 Hard-margin SVM

For binary classification, SVM classifier is a function class of the form $f(\mathbf{x}) = \text{sgn}(\mathbf{w}^T\mathbf{x} + b)$. $f(.)$ will assign either $-$ or $+$ for a data point $\mathbf{x}$, where $\mathbf{w}, \mathbf{x} \in \mathbb{R}^d$, $b \in \mathbb{R}$. If the data is to be classified correctly, the hyperplane produced by SVM should ensure that:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) > 0,\ \forall(\mathbf{x}_i, y_i) \in D \tag{28}$$

where $D = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ is the training set. This can be formulated as a convex optimization problem:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}||\mathbf{w}||^2 \tag{29}$$

$$\text{s.t.}\ \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1,\ \forall i = 1, \ldots, n \tag{30}$$

Using Lagrange multipliers, the dual problem is:

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j \tag{31}$$

$$\text{s.t.}\ \ \alpha_i \geq 0 \tag{32}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{33}$$

In the dual problem, the classification rule is $\hat{y} = \text{sgn}(\sum_{i=1}^{n} y_i\alpha_i\mathbf{x}^T\mathbf{x}_i + b)$. The interesting thing in this formulation is the appearance of the inner product $\mathbf{x}_i^T\mathbf{x}_j$. Instead of the usual dot product, we can choose any valid inner products here, this makes SVM much more flexible and can even deal with non-linear problems.

In the hard-margin formulation, the underlying loss is still 0/1 loss. So the generalization bound follows exactly what we have in class. With probability $1 - \delta$, for all hard-margin linear SVM classifiers $\hat{y}$ in $\mathbb{R}^d$:

$$R[\hat{y}] \leq \hat{R}[\hat{y}] + \sqrt{\frac{2d_{vc}\log(\frac{en}{d_{vc}})}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \tag{34}$$

$$= \hat{R}[\hat{y}] + \sqrt{\frac{2(d+1)\log(\frac{en}{d+1})}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \tag{35}$$

Next, we extend this bound to the soft-margin case.

5.2 Soft-margin SVM

Since the hard-margin SVM formulation will yield no solution for the non-separable case. To combat this, we simply modify the problem to allow misclassification error within a margin:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i}^{n} \xi_i \tag{36}$$

$$\text{s.t.}\ \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i,\ \forall i = 1, \ldots, n \tag{37}$$

$$\xi_i \geq 0 \tag{38}$$

The dual problem is almost exactly the same as before, just with one more constraints on α_i:

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j \tag{39}$$

$$\text{s.t.}\ \ C > \alpha_i \geq 0 \tag{40}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{41}$$

In the soft-margin formulation, the loss is no longer 0/1. Instead, it is now the hinge loss:

$$l(y, \hat{y}) = \max(0, 1 - y\hat{y}) \tag{42}$$

$$\hat{y} = \mathbf{w}^T\mathbf{x} + b \tag{43}$$

Here, we derive the generalization bound for the hinge loss. Using the same notations and steps as in the lectures. The only difference will be in Step 1 (McDiarmid's inequality) and Step 3 (we are using hinge loss instead of 0/1 loss).

For **Step 1**, assuming the loss function is ranging in an interval no more than B, following the lecture notes we get to:

$$\Phi(z) - \Phi(z') \leq \sup_{g\in\mathcal{G}} \left\{ \frac{1}{n}\big(g(z'_i) - g(z_i)\big) \right\} \tag{44}$$

$$\leq \frac{B}{n} \tag{45}$$

Now apply McDiarmid's inequality we have:

$$P(\Phi(z) \geq \mathbb{E}[\Phi(z)] + \epsilon) \leq e^{-\frac{2\epsilon^2}{n \cdot \frac{B^2}{n^2}}} \tag{46}$$

$$= e^{-\frac{2n\epsilon^2}{B^2}} = \delta \tag{47}$$

This means $\epsilon = B\sqrt{\frac{\log(\frac{1}{\delta})}{2n}}$. Thus, we have with probability $1 - \delta$:

$$\Phi(z) \leq \mathbb{E}[\Phi(z)] + B\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \quad \square \tag{48}$$

Step 2 follows as is, which gives:

$$\mathbb{E}[\Phi(z)] \leq 2\mathcal{R}_{\mathcal{G}}(n) \tag{49}$$

For **Step 3**, we first prove the following Lemma:

Lemma 1. For any L-Lipschitz function $g \in \mathcal{G}$ acting on predictors $f \in \mathcal{F}$, we have:

$$\mathcal{R}_{\mathcal{G}}(n) \leq L\mathcal{R}_{\mathcal{F}}(n) \tag{50}$$

This means for any Lipschitz loss functions, we can bound the Radamacher complexity of class of loss functions by the Radamacher of the predictor class.

Proof.

$$\mathcal{R}_{\mathcal{G}}(n) = \mathbb{E}_Z\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\Big[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} g(z_i)\sigma_i\Big]\right] \tag{51}$$

$$= \mathbb{E}_Z\left[\frac{1}{2}\mathbb{E}_{\sigma_1,\ldots,\sigma_{n-1}}\Big[\sup_{g\in\mathcal{G}}\big\{\frac{1}{n}\sum_{i=1}^{n} g(z_i)\sigma_i + \frac{1}{n}g(z_n)\big\} + \sup_{g\in\mathcal{G}}\big\{\frac{1}{n}\sum_{i=1}^{n} g(z_i)\sigma_i - \frac{1}{n}g(z_n)\big\}\Big]\right] \tag{52}$$

$$\text{(change the expectation and spell out the values for } \sigma_n) \tag{53}$$

$$= \mathbb{E}_Z\left[\frac{1}{2}\mathbb{E}_{\sigma_1,\ldots,\sigma_{n-1}}\Big[\sup_{g,g'\in\mathcal{G}}\big\{\frac{1}{n}\sum_{i=1}^{n} g(z_i)\sigma_i + \frac{1}{n}g(z_n) + \frac{1}{n}\sum_{i=1}^{n} g'(z_i)\sigma_i - \frac{1}{n}g'(z_n)\big\}\Big]\right] \tag{54}$$

$$= \mathbb{E}_Z\left[\frac{1}{2}\mathbb{E}_{\sigma_1,\ldots,\sigma_{n-1}}\Big[\sup_{g,g'\in\mathcal{G}}\big\{\frac{1}{n}\sum_{i=1}^{n} g(z_i)\sigma_i + \frac{1}{n}\sum_{i=1}^{n} g'(z_i)\sigma_i + \underbrace{\frac{1}{n}g(z_n) - \frac{1}{n}g'(z_n)}_{\text{(apply Lipschitz)}}\big\}\Big]\right] \tag{55}$$

$$\le \mathbb{E}_Z\left[\frac{1}{2}\mathbb{E}_{\sigma_1,\ldots,\sigma_{n-1}}\Big[\sup_{g,g'\in\mathcal{G},f\in\mathcal{F}}\big\{\frac{1}{n}\sum_{i=1}^{n}\big(g(z_i) + g'(z_i)\big)\sigma_i + \frac{1}{n}L|f(x_n) - f'(x_n)|\big\}\Big]\right] \tag{56}$$

$$= \mathbb{E}_Z\left[\frac{1}{2}\mathbb{E}_{\sigma_1,\ldots,\sigma_{n-1}}\Big[\sup_{g,g'\in\mathcal{G},f\in\mathcal{F}}\big\{\frac{1}{n}\sum_{i=1}^{n}\big(g(z_i) + g'(z_i)\big)\sigma_i + \frac{1}{n}L(f(x_n) - f'(x_n))\big\}\Big]\right] \tag{57}$$

$$\text{(due to the supremum)} \tag{58}$$

$$= \mathbb{E}_Z\left[\frac{1}{2}\mathbb{E}_{\sigma_1,\ldots,\sigma_{n-1}}\Big[\sup_{g\in\mathcal{G},f\in\mathcal{F}}\big\{\frac{1}{n}\sum_{i=1}^{n} g(z_i)\sigma_i + \frac{1}{n}Lf(x_n)\big\} + \sup_{g\in\mathcal{G},f\in\mathcal{F}}\big\{\frac{1}{n}\sum_{i=1}^{n} g(z_i)\sigma_i - \frac{1}{n}Lf(x_n)\big\}\Big]\right] \tag{59}$$

$$= \mathbb{E}_Z\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_{n-1}}\Big[\sup_{g\in\mathcal{G},f\in\mathcal{F}}\big\{\frac{1}{n}\sum_{i=1}^{n-1} g(z_i)\sigma_i + \frac{1}{n}\sigma_n Lf(x_n)\big\}\Big]\right] \tag{60}$$

$$= \ldots \text{(keep repeating steps (52)-(60) to remove } g(z_i)) \tag{61}$$

$$\le \mathbb{E}_Z\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i Lf(x_i)\Big]\right] \tag{62}$$

$$= L\mathcal{R}_{\mathcal{F}}(n) \quad \square \tag{63}$$

Since hinge loss is 1-Lipschitz, we can then replace Step 3 from class lecture with Lemma 1 (with L=1) and continue as is. By the end of Step 3 we have:

$$\Phi(z) = \sup_{g\in\mathcal{G}}\Big\{\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z)\Big\} \le 2\mathcal{R}_{\mathcal{F}}(n) + B\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \tag{64}$$

Step 4 and **Step 5** follow as is from the lectures, here we know the VC-dimension of a hyperplane in $\mathbb{R}^d$ is $d+1$, utilizing this we have:

$$\mathcal{R}_{\mathcal{F}}(n) \le \sqrt{\frac{2\log(\Delta_{\mathcal{F}}(n))}{n}} \quad \text{(Massart's Lemma)} \tag{65}$$

$$\Delta_{\mathcal{F}}(n) \le \Big(\frac{en}{d_{vc}}\Big)^{d_{vc}} \quad \text{(Bound on growth function for } n \ge d_{vc}) \tag{66}$$

$$= \Big(\frac{en}{d+1}\Big)^{d+1} \tag{67}$$

By plugging these results back into (64), we have:

$$\Phi(z) = \sup_{g\in\mathcal{G}}\Big\{\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z)\Big\} \tag{68}$$

$$\le 2\sqrt{\frac{2(d+1)\log\left(\frac{en}{d+1}\right)}{n}} + B\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \tag{69}$$

Finally, we have with probability $1-\delta$, for all soft-margin linear SVM classifiers $\hat{y}$ in $\mathbb{R}^d$:

$$R[\hat{y}] \leq \hat{R}[\hat{y}] + 2\sqrt{\frac{2(d+1)\log\left(\frac{en}{d+1}\right)}{n}} + B\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \quad \square \tag{70}$$

Note that this bound also works for the hard-margin case, which makes sense because hinge loss upper bound the 0/1 loss. But if we want to apply for Kernel SVM, whose VC dimension might be infinite, this bound will become trivial. Hence, we try to develop a different bound below.

Alternatively: We can do a bit better by realizing that (36) can be re-written as an unconstrained optimization problem:

$$\min_{\mathbf{w},b} \quad \lambda||\mathbf{w}||^2 + \sum_i^n \max\{0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\} \tag{71}$$

where $\lambda = \frac{1}{2nC}$. We can then move the norm into the constraint and obtain the equivalent minimization of the hinge loss:

$$\min_{\mathbf{w},b} \quad \sum_i^n \max\{0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\} \tag{72}$$

$$\text{s.t. } ||\mathbf{w}||^2 \leq R^2 \tag{73}$$

Note that we have absorbed both C and λ into R to simplify the derivation.

The problem now becomes minimizing the hinge loss restricted to a norm ball in $\mathbb{R}^d$ with radius R. Based on this observation, we can think of our function class as $\mathcal{F} = \{f : \mathbf{x} \to \langle \mathbf{w}, \mathbf{x}\rangle + b \mid ||\mathbf{w}||^2 \leq R^2\}$, where we switched to $\langle \mathbf{w}, \mathbf{x}\rangle$ to denote the general inner product. Without loss of generality, we can ignore b (or absorb it into $\mathbf{w}$) and get $\mathcal{F} = \{f : \mathbf{x} \to \langle \mathbf{w}, \mathbf{x}\rangle \mid ||\mathbf{w}||^2 \leq R\}$. From this, we can bound $\mathcal{R}_{\mathcal{F}}(n)$ differently:

$$\mathcal{R}_{\mathcal{F}}(n) = \mathbb{E}_Z\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(\mathbf{x}_i)\right]\right] \tag{74}$$

$$= \mathbb{E}_Z\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{\mathbf{w}:||\mathbf{w}||^2\leq R} \frac{1}{n}\sum_{i=1}^n \sigma_i \langle \mathbf{w}, \mathbf{x}_i\rangle\right]\right] \tag{75}$$

$$= \mathbb{E}_Z\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{\mathbf{w}:||\mathbf{w}||^2\leq R} \frac{1}{n}\left\langle \mathbf{w}, \sum_{i=1}^n \sigma_i \mathbf{x}_i\right\rangle\right]\right] \tag{76}$$

$$\leq \mathbb{E}_Z\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{\mathbf{w}:||\mathbf{w}||^2\leq R} \frac{1}{n}||\mathbf{w}||\Big|\Big|\sum_{i=1}^n \sigma_i \mathbf{x}_i\Big|\Big|\right]\right] \tag{77}$$

$$\text{(Cauchy-Schwarz)} \tag{78}$$

$$\leq \mathbb{E}_Z\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\frac{R}{n}\Big|\Big|\sum_{i=1}^n \sigma_i \mathbf{x}_i\Big|\Big|\right]\right] \tag{79}$$

$$= \mathbb{E}_Z\left[\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\frac{R}{n}\left(\Big|\Big|\sum_{i=1}^n \sigma_i \mathbf{x}_i\Big|\Big|^2\right)^{1/2}\right]\right] \tag{80}$$

$$\leq \mathbb{E}_Z\left[\left(\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\frac{R}{n}\Big|\Big|\sum_{i=1}^n \sigma_i \mathbf{x}_i\Big|\Big|^2\right]\right)^{1/2}\right] \quad \text{(convexity of norm)} \tag{81}$$

$$= \frac{R}{n}\mathbb{E}_Z\left[\left(\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sum_{i,j} \sigma_i\sigma_j\langle \mathbf{x}_i, \mathbf{x}_j\rangle\right]\right)^{1/2}\right] \tag{82}$$

$$= \frac{R}{n}\mathbb{E}_Z\left[\left(\sum_i^n \mathbb{E}[\sigma_i^2]||\mathbf{x}_i||^2 + n(n-1)\sum_{i,j,i\neq j}^n \underbrace{\mathbb{E}[\sigma_i]\mathbb{E}[\sigma_j]}_{=0 \text{ due to independence}} \langle \mathbf{x}_i, \mathbf{x}_j\rangle\right)^{1/2}\right] \tag{83}$$

$$= \frac{R}{n}\mathbb{E}_Z\left[\left(\sum_i^n \mathbb{E}[\sigma_i^2]||\mathbf{x}_i||^2\right)^{1/2}\right] = \frac{R}{n}\mathbb{E}_Z\left[\left(\sum_i^n ||\mathbf{x}_i||^2\right)^{1/2}\right] \tag{84}$$

$$\leq \frac{R}{n}\sqrt{n}\max_i ||\mathbf{x}_i|| = \frac{R}{\sqrt{n}}\max_i ||\mathbf{x}_i|| \quad \square \tag{85}$$

To take care of the constant B, we use the following simple argument. Since $||\mathbf{w}|| \leq R$, the maximal loss happens when $\langle \mathbf{w}, \mathbf{x} \rangle$ and $\langle \mathbf{w}, \mathbf{x}' \rangle$ are opposite in sign. In this case, the difference in the hinge losses is:

$$\begin{aligned} B &= \sup \Big\{ \max\{0, 1 - y(\mathbf{w}^T \mathbf{x} + b)\} - \max\{0, 1 - y(\mathbf{w}^T \mathbf{x}' + b)\} \Big\} && (86) \\ &\leq 2 + 2||\mathbf{w}|| \max_i ||\mathbf{x}_i|| \quad \text{(Cauchy-Schwarz)} && (87) \\ &\leq 2(1 + R \max_i ||\mathbf{x}_i||) && (88) \end{aligned}$$

Plugging (85), (88) into (64) we obtain a bound that is independent of the VC-dimension, instead it depends on the maximum norm of training data, as well as the radius of the norm ball.

With probability $1 - \delta$, for all soft-margin linear SVM classifiers $\hat{y}$ in $\mathbb{R}^d$:

$$R[\hat{y}] \leq \hat{R}[\hat{y}] + \frac{2R}{\sqrt{n}} \max_i ||\mathbf{x}_i|| + 2(1 + R \max_i ||\mathbf{x}_i||) \sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \quad \square \tag{89}$$

This bound is better in the sense that the generalization error actually vanishes as $n \to \infty$ regardless of the VC dimension. The dependence on the maximum norm of training points also makes sense, since for data points with large norm (i.e., lies deep in one halfplane), the misclassification risk will be huge for hinge loss.

6 Generalization bound of Kernelized SVM

Note: we try to extend to Kernel SVM, but we are not sure if the result is correct here.

In kernelized SVM, we replace the inner product in (39) by a valid kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \tag{90}$$

where $\phi(.)$ is a mapping in feature space. Usually explicitly calculating $\phi(\mathbf{x})$ is expensive, but calculating $K(\mathbf{x}_i, \mathbf{x}_j)$ is not, hence the popularity of kernel method. When using the kernel method for soft-margin SVM, we can imagine that our function class $\mathcal{F}$ is now:

$$\mathcal{F} = \{f : \mathbf{x} \to \langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b \mid ||\mathbf{w}||^2 \leq R^2\} \tag{91}$$

Then by following the same steps as before, we get to:

$$\begin{aligned} \mathcal{R}_{\mathcal{F}}(n) &\leq \frac{R}{n} \mathbb{E}_Z \left[\Big(\sum_i^n K(\mathbf{x}_i, \mathbf{x}_i) \Big)^{1/2} \right] && (92) \\ &= \frac{R}{n} \mathbb{E}_Z \left[\Big(\sum_i^n \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_i) \rangle \Big)^{1/2} \right] && (93) \\ &= \frac{R}{n} \mathbb{E}_Z \left[\sqrt{\text{trace}(\mathbf{G})} \right] && (94) \\ &(\mathbf{G} \text{ is the Gram matrix, } \mathbf{G}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j). && (95) \\ &= \frac{R}{n} \sqrt{\text{trace}(\mathbf{G})} && (96) \\ &\leq \frac{R}{n} \sqrt{n \sigma_{\max}(\mathbf{G})} = R \sqrt{\frac{\sigma_{\max}(\mathbf{G})}{n}} && (97) \\ &(\sigma_{\max}(\mathbf{G}) \text{ is the largest eigenvalue of } \mathbf{G}) && (98) \end{aligned}$$

We can also perform a similar analysis for B:

$$\begin{aligned} B &= \sup \Big\{ \max\{0, 1 - y(\langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b)\} - \max\{0, 1 - y(\langle \mathbf{w}, \phi(\mathbf{x}') \rangle + b)\} \Big\} && (99) \\ &\leq 2 + 2||\mathbf{w}|| \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \quad \text{(Cauchy-Schwarz)} && (100) \\ &\leq 2(1 + R\sigma_{\max}(\mathbf{G})) && (101) \end{aligned}$$

Thus, we have the following bound. With probability $1 - \delta$, for all soft-margin kernelized SVM classifiers $\hat{y}$ in $\mathbb{R}^d$ with Kernel Gram matrix $\mathbf{G}$:

$$R[\hat{y}] \leq \hat{R}[\hat{y}] + 2R \sqrt{\frac{\sigma_{\max}(\mathbf{G})}{n}} + 2(1 + R\sigma_{\max}(\mathbf{G})) \sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \quad \square \tag{102}$$

Instead of depending on the data point with maximum norm, the bound now depends on $\sigma_{\max}(\mathbf{G})$, which represents the dominant component of the Gram matrix, i.e., the dominant feature in the latent space $\phi(\mathbf{x})$.

7 Conclusion

In this short paper, we summarized the main idea of [VD04], which was built upon the Bias-Variance technique introduced in [Dom00]. We then attempted to derive the generalization bounds of SVM models, for hard-margin, soft-margin, and kernelized versions. The derivations followed closely the 5 steps introduced in class. Parts of the proof were motivated by the course CSC588 - Machine Learning Theory, but they did not explicitly dervive a bound for SVM, so we are not 100% sure if our results are correct.

References

[Dom00] Pedro Domingos. A unified bias-variance decomposition. In *Proceedings of 17th international conference on machine learning*, pages 231–238. Morgan Kaufmann Stanford, 2000.

[VD04] Giorgio Valentini and Thomas G Dietterich. Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. *Journal of Machine Learning Research*, 5(Jul):725–775, 2004.