

# **Bounding Generalization for Support Vector Machines**

Steve Kroon

28 March 2003

## Contents

1. Bias-variance tradeoff
2. Regression SVMs
3. Risk and PAC bounds
4. Hoeffding's theorem
5. Covers and covering numbers
6. Deriving PAC bounds
7. Bounding covering numbers - generic
8. Bounding covering numbers - SVMs
9. Generalization bounds for SVMs
10. Example
11. General comments and conclusion

## 1 Bias-variance tradeoff

- Regression context
- $Y = f(X) + \Theta$ , with  $\Theta \sim (0, \theta^2)$  independent of  $X$ .

$$\begin{aligned}\text{EPE}(\hat{f}) &= \mathbb{E} \left[ (Y - \hat{f}(X))^2 \right] \\ &= \mathbb{E} \left[ (Y - f(X))^2 \right] + \mathbb{E} \left[ (f(X) - \mathbb{E}_X \hat{f}(X))^2 \right] + \mathbb{E} \left[ (\mathbb{E}_X \hat{f}(X) - \hat{f}(X))^2 \right] \\ &= \theta^2 + \text{Bias}^2(\hat{f}(X)) + \text{Var}(\hat{f}(X))\end{aligned}$$

- Linear model, i.e. hypothesis space

$$\mathcal{H} = \left\{ f_\beta : X \rightarrow \beta^T \begin{bmatrix} 1 \\ X \end{bmatrix} \right\}$$

- **Example 1** Least squares estimator  
unbiased
- **Example 2** Ridge regression estimator  
small increase in bias vs potentially large decrease in variance

## 2 Regression SVMs

- *Shrinkage methods* handle the bias-variance tradeoff:

$$\hat{f} = \arg \min_{g \in \mathcal{F}} \sum_{i=1}^N L(y_i - g(x_i)) + \lambda P(g)$$

- Loss function  $L$ , Penalty function  $P$ , Regularization constant  $\lambda$
- Linear model:  $\hat{f}(x) = \hat{\beta}^T x = \sum_{i=1}^N \hat{\alpha}_i x_i^T x$ , where

- Ridge regression:

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^N (y_i - \beta^T x_i)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

- Lasso:

$$\hat{f} = \arg \min_{\beta} \sum_{i=1}^N (y_i - \beta^T x_i)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

- The linear SVM:

$$\hat{f} = \arg \min_{\beta} \sum_{i=1}^N |y_i - \beta^T x_i|_{\varepsilon} + \lambda \sum_{j=1}^p \beta_j^2$$

- $\varepsilon$ -insensitive loss function

$$|p|_{\varepsilon} = \begin{cases} 0 & \text{if } |p| < \varepsilon \\ |p| - \varepsilon & \text{otherwise} \end{cases}$$

- Assumes error of less than  $\varepsilon$  is adequate

- Find  $\hat{\alpha}_i$  by solving quadratic programming problem
- Typically few  $\hat{\alpha}_i \neq 0$ . Corresponding  $x_i$  are SVs
- i.e.  $\hat{\beta}$  has a *sparse* representation in the  $x_i$
- SVM variants — different loss and penalty functions
- **Example** Ridge regression - different loss function
- Non-linear SVMs using *kernel trick*. Replace

$$\hat{f}(x) = \sum_{i=1}^N \hat{\alpha}_i x_i^T x$$

by

$$\hat{f}(x) = \sum_{i=1}^N \hat{\alpha}_i K(x_i, x)$$

- *Kernel*  $K(x, y)$  is dot product in high-dimensional space

$$K(x, y) = \Phi(x)^T \Phi(y)$$

- $\hat{\alpha}_i$ 's trained using  $K$
- **Example** Polynomial kernel

$$K(x, y) = (x^T y + 1)^d$$

results in a polynomial regression function

### 3 Risk and PAC bounds

- A function  $g$  errs when loss is non-zero, i.e.

$$L(Y - g(X)) > 0$$

- Expected risk: probability of error of  $g$

$$R(g) = P(L(Y - g(X)) > 0) = \mathbb{E}[\mathbb{I}(L(Y - g(X)) > 0)]$$

- $\mathbb{I}$  is an indicator function

$$\mathbb{I}(p) = \begin{cases} 0 & \text{if } p \text{ is false} \\ 1 & \text{if } p \text{ is true} \end{cases}$$

- $R(\hat{f})$  depends on joint distribution of  $X$  and  $Y$
- PAC bound: probably approximately correct
- We need, with probability  $1 - \delta$ ,

$$P(|Y - \hat{f}(X)| > \varepsilon) \leq \sigma$$

- For  $\varepsilon$ -insensitive loss, this means with probability  $1 - \delta$  (probably),  $R(\hat{f})$  is small (approximately correct)

## 4 Hoeffding's theorem

**Theorem (Wassily Hoeffding)** Let  $\tau_1, \dots, \tau_m$  be a sequence of independent, identically distributed bounded random variables, with  $\tau_i$  falling into the interval  $[a_i, a_i + b_i]$  with probability one. Denote their average by  $S_m$ . Then, for any  $\epsilon \geq 0$ ,

$$\Pr \{E(S_m) - S_m \geq \epsilon\} \leq \exp\left(-\frac{2m^2\epsilon^2}{\sum_{i=1}^m b_i^2}\right).$$

- Empirical risk: proportion of errors on a sample  $s$

$$R_s(g) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(L(y_i - g(x_i)) > 0)$$

- Let  $\tau_i = \mathbb{I}(L(y_i - g(x_i)) > 0)$
- Then  $R_s(g) = S_m$  and  $R(g) = E(S_m)$ , so

$$\Pr(R(g) - R_s(g) \geq \rho) = \Pr(R(g) \geq R_s(g) + \rho) \leq \exp(-2m\rho^2)$$

- Finite hypothesis space  $\mathcal{H}$ :

$$\Pr(\sup_{h \in \mathcal{H}} (R(h) - R_s(h)) \geq \rho) \leq |\mathcal{H}| \exp(-2m\rho^2)$$

## 5 Covers and covering numbers

- Hypothesis space is usually not finite, e.g. linear model
- Workaround: find a finite set  $P$  which approximates  $\mathcal{H}$  well
- $P$  is a  $\gamma$ -cover of  $\mathcal{H}$  with respect to points  $\mathbf{a} = \{a_1, a_2, \dots, a_n\}$  if for any element  $h \in \mathcal{H}$  there is a  $p(h) \in P$  so that  $|h(a_i) - p(h)(a_i)| \leq \gamma$  for all  $i$
- **Example**  $P = \{\frac{1}{4}, \frac{3}{4}\}$  is a  $\frac{1}{2}$ -cover of  $\mathcal{H} = \{c : 0 < c < 1\}$  with respect to  $\mathbb{R}$
- Minimal size of such a  $\gamma$ -cover is the  $\gamma$ -covering number of  $\mathcal{H}$  with respect to the points  $\mathbf{a}$ ,  $\mathcal{N}(\gamma, \mathcal{H}, \mathbf{a})$
- Expected margin risk:

$$R^\gamma(g) = \mathbb{E}[\mathbb{I}(|y - g(x)| > \varepsilon - \gamma)]$$

- Empirical margin risk:

$$R_s^\gamma(g) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(|y_i - g(x_i)| > \varepsilon - \gamma)$$

- These risks increase with  $\gamma$

## 6 Deriving PAC bounds

- 2 important results:

$$R(h) \leq R^{\frac{\gamma}{2}}(p(h))$$

$$R_s^\gamma(h) \geq R_s^{\frac{\gamma}{2}}(p(h))$$

- Replace original risks with  $\frac{\gamma}{2}$ -margin risks, and apply Hoeffding:

$$\Pr(\sup_{h \in \mathcal{H}} (R(h) - R_s^\gamma(h)) \geq \rho) \leq \mathcal{N}\left(\frac{\gamma}{2}, \mathcal{H}, \text{dom}(X)\right) \exp(-2m\rho^2)$$

- Only if covering number is finite
- Next workaround (double sample): For  $N > \frac{2}{\rho^2}$

$$\Pr(\sup_{h \in \mathcal{H}} (R^{\frac{\gamma}{2}}(h) - R_{s_1}^{\frac{\gamma}{2}}(h)) \geq \rho) \leq 2\Pr(\sup_{h \in \mathcal{H}} (R_{s_2}^{\frac{\gamma}{2}}(h) - R_{s_1}^{\frac{\gamma}{2}}(h)) \geq \frac{\rho}{2})$$

- Using this approach yields (no longer using Hoeffding), for example

$$\begin{aligned} \Pr(\sup_{h \in \mathcal{H}} (R(h) - R_s^\gamma(h)) \geq \rho) &\leq 6N \sup_{s_1, s_2} \left\{ \mathcal{N}\left(\frac{\gamma}{3}, \mathcal{H}, (s_1, s_2)\right) \right\} \exp\left(-\frac{N\rho^2}{9}\right) \\ &= 6N \mathcal{N}\left(\frac{\gamma}{3}, \mathcal{H}, 2N\right) \exp\left(-\frac{N\rho^2}{9}\right) \end{aligned}$$

- $\mathcal{N}\left(\frac{\gamma}{3}, \mathcal{H}, 2N\right)$  is the *Growth function*
- Bounds above generate PAC bounds, e.g. for all  $h$ , with probability  $1 - \delta$ ,

$$R(h) < R_s^\gamma(h) + \sqrt{\frac{1}{2N} \left( \ln \mathcal{N}\left(\frac{\gamma}{2}, \mathcal{H}, \text{dom}(X)\right) - \ln \delta \right)}$$

- Holds only for a pre-specified  $\gamma$
- Generalize: for all  $h$ , with probability  $1 - \delta$ , simultaneously for all  $\gamma \in (0, \varepsilon]$ :

$$R(h) < R_s^\gamma(h) + \sqrt{\frac{1}{2N} \left( \ln \mathcal{N}\left(\frac{\gamma}{4}, \mathcal{H}, \text{dom}(X)\right) - \ln \frac{\delta\gamma}{2\varepsilon} \right)}$$

- Distribution-free

## 7 Bounding covering numbers - generic

- Need to evaluate the Growth function
- Generally done using *scale-sensitive dimensions*
- Fat-shattering function is an example - measures complexity of the hypothesis space
- Then bound Growth function in terms of the fat-shattering function at scale related to margin

## 8 Bounding covering numbers - SVMs

- Newer approach for SVMs uses functional analysis technique of entropy numbers
- Uses eigenvalues of  $\mathbf{X}^T \mathbf{X}$
- Finite input space (double samples also)
- Consider the restricted model space

$$\mathcal{H}(c) = \{f_\beta : X \rightarrow \beta^T X, \|\beta\| \leq c\}$$

- For  $\gamma > \Upsilon(n)$ ,  $\mathcal{N}(\gamma, \mathcal{H}(c), \text{dom}(X)) < \lceil c \rceil 2^{n-1}$
- Allowing an intercept

$$\mathcal{H}_B(c) = \{f_\beta : X \rightarrow \beta^T X + b, \|\beta\| \leq c, |b| < B\}$$

- For  $\gamma > \Upsilon(n)$ ,

$$\mathcal{N}(\gamma, \mathcal{H}_B(c), \text{dom}(X)) < \left\lceil \frac{2B}{\gamma} \right\rceil \lceil c \rceil 2^{n-1}$$

- Bounds hold for hypothesis space, not only SVMs

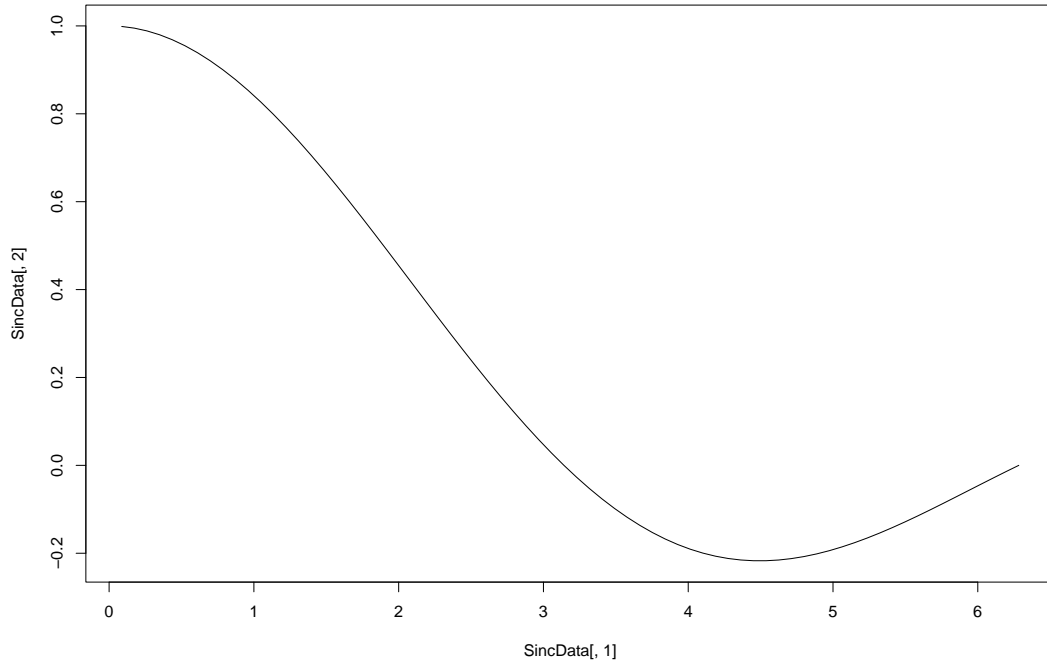
## 9 A generalization bound for SVMs

- Can apply to all linear functions of  $X$
- Substitute bounds on covering numbers in original bounds
- Can generalize by using kernels, then uses Gram matrix  $G_{ij} = K(x_i, x_j)$
- Thus for all  $h \in \mathcal{H}_B(c)$ , with prob.  $1 - \delta$ , simultaneously for  $\gamma \in (8\Upsilon(n), \varepsilon]$ :

$$R(h) < R_s^\gamma(h) + \sqrt{\frac{1}{2N} \left( n \ln 2 + \ln \frac{\lceil \frac{8B}{\gamma} \rceil \lceil c \rceil \varepsilon}{\delta \gamma} \right)}$$

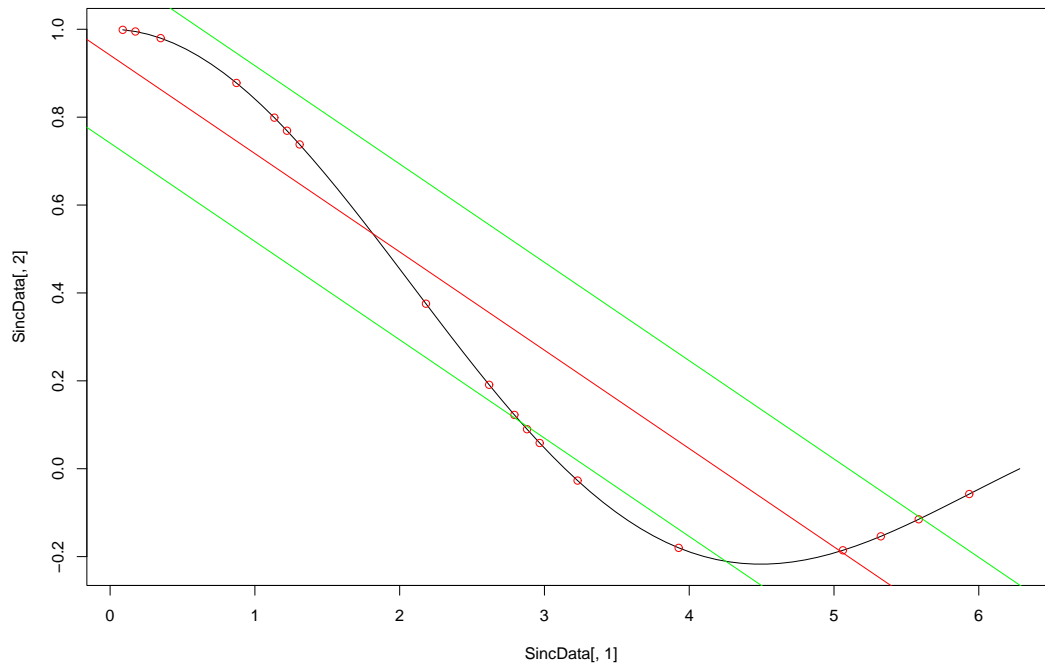
## 10 Example

$$\text{sinc}(x) = \frac{\sin(x)}{x}$$

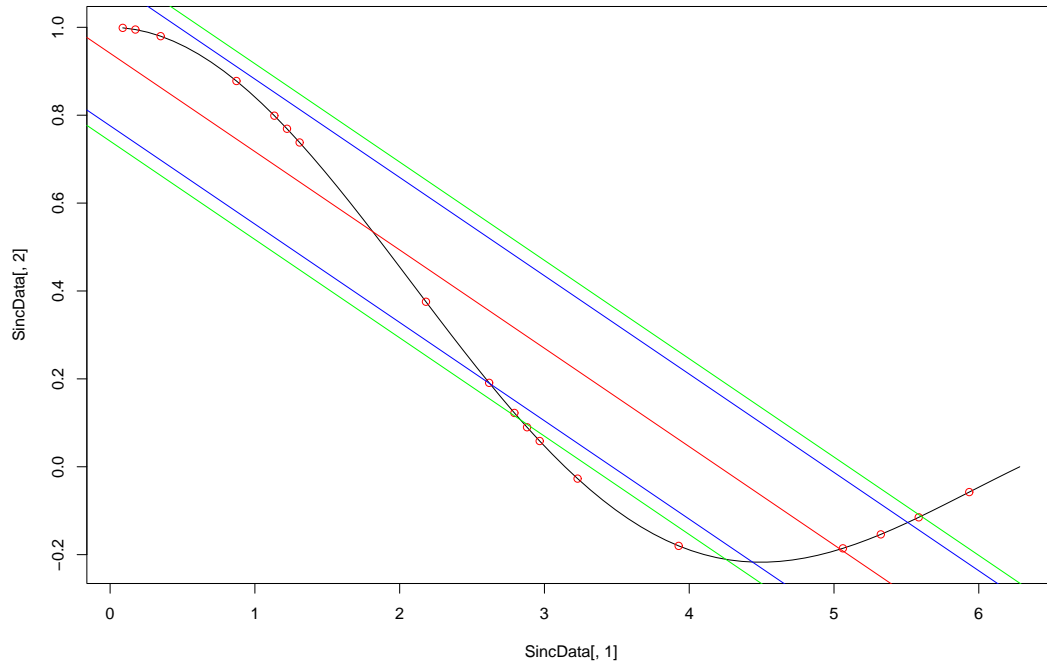


- Input space 72 points:  $\frac{k\pi}{36}$
- Training set 18 points:  $N = 18$
- $B = 1, c = \frac{1}{\pi^2}$
- Accuracy of  $\varepsilon = 0.2$  adequate
- 90% certainty, so  $\delta = 0.1$
- We use  $n = 3$ :  $\Upsilon(3) < 0.0002$

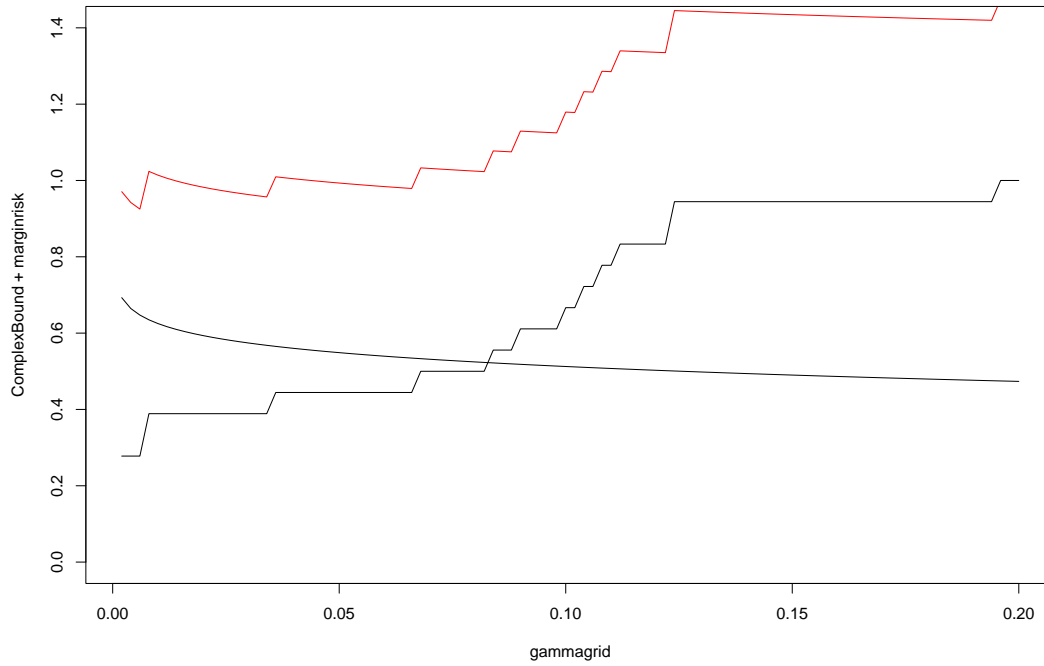
- Least squares
- $Y = -0.224X + 0.941$



- Calculate margin risk for allowable margins



$$R(h) < R_s^\gamma(h) + \sqrt{\frac{1}{36} \left[ 3 \ln 2 + \ln \frac{\lceil \frac{8}{\gamma} \rceil (0.2)}{0.1\gamma} \right]}$$



- e.g. From  $\gamma = 0.006$ ,  $R(h) < 0.925$
- Works for other estimators too

## 11 General comments and conclusion

- Bounds very loose for 2 reasons:
  - Use of loose inequalities
  - Distribution-free
- For finite  $\text{dom}(X)$ ,  $\mathcal{H}_B(c)$  too big
- Model selection
- SVM bounds limited by  $\Upsilon(n)$
- Have a nice weekend