Chapter 4 Support Vector Machines

Ruochen Xu

Carnegie Mellon University

ruochenx@cmu.edu

February 3, 2015
SVMs separable case

Notions

- **Training Data** $z_i = (x_i, y_i) \sim D$, $S = z_1, z_2, ..., z_m \sim D^m$, with $y_i = f(x_i)$
- Determining a hypothesis $h \in H$, a binary classifier, with small generalization error: $R_D(h) = \Pr_{x \sim D}[h(x) \neq f(x)]$
- **Linear Classifiers** $H = \{x \mapsto \text{sign}(w \cdot x + b) : w \in \mathbb{R}^N, b \in \mathbb{R}\}$
- The solution returned by the SVM algorithm is the hyperplane with the maximum *margin*, or distance to the closest points, and is thus known as the *maximum-margin hyperplane*.

![Diagram showing linear classifers and margins.](image)
SVMs separable case

- Distance of $\mathbf{x}_0 \in \mathbb{R}^N$ to a hyperplane is $\frac{|\mathbf{w} \cdot \mathbf{x}_0 + b|}{||\mathbf{w}||}$
- Canonical Hyperplane: scale $\mathbf{w}$ and $b$ appropriately such that $\min_{(\mathbf{x},y) \in S} |\mathbf{w} \cdot \mathbf{x} + b| = 1$
- For a canonical hyperplane, the margin $\rho$ is given by:

$$\rho = \min_{(\mathbf{x},y) \in S} \frac{|\mathbf{w} \cdot \mathbf{x} + b|}{||\mathbf{w}||} = \frac{1}{||\mathbf{w}||}$$

Primal optimization problem

$$\min_{\mathbf{w}, b} \frac{1}{2} ||\mathbf{w}||^2 \quad (1)$$
subject to: $y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \forall i \in [1, m]$

- Objective function is strictly convex, so the optimization problem admits a unique solution. The optimization problem is in fact a specific instance of quadratic programming (QP), a family of problems extensively studied in optimization.
The Lagrange variables $\mathbf{\alpha} = (\alpha_1, \ldots, \alpha_m^T) \in \mathbb{R}_m^+$

The Lagrangian of primal optimization problem

$$L(\mathbf{w}, b, \mathbf{\alpha}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1] \quad (2)$$

The KKT conditions are obtained by setting the gradient of the Lagrangian with respect to the primal variables $\mathbf{w}$ and $b$ to zero and by writing the complementarity conditions:

$$\nabla_\mathbf{w} L = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \quad (3)$$

$$\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0 \quad (4)$$

$$\forall i, \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0 \Rightarrow \alpha_i = 0 \lor y_i (\mathbf{w} \cdot \mathbf{x}_i + b) = 1 \quad (5)$$
Dual optimization problem

To derive the dual form of the constrained optimization problem (1), we plug in (3) and apply (4) on (2).

\[ L = \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i x_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{m} \alpha_i y_i b + \sum_{i=1}^{m} \alpha_i \]  

(6)

\[ L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \]  

(7)

Dual optimization problem for separable case

\[ \max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \]  

(8)

subject to \( \alpha_i \geq 0 \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, \forall i \in [1, m] \)
The hypothesis returned by SVMs:

$$h(\mathbf{x}) = \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b\right)$$  \hspace{1cm} (9)$$

For any support vector $\mathbf{x}_i$, $\mathbf{w} \cdot \mathbf{x}_i + b = y_i$, and thus $b$ can be obtained via

$$b = y_i - \sum_{j=1}^{m} \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{x}_i)$$ \hspace{1cm} (10)$$

The dual optimization problem (8) and the expressions (9) and (10) reveal an important property of SVMs: the hypothesis solution depends only on inner products between vectors and not directly on the vectors themselves.
Leave-one-out analysis

- A learning algorithm $A$
- A fix sample $S$ of size $m$
- The leave-one-out error of $A$ on a sample $S$ is

$$
\hat{R}_{LOO}(A) = \frac{1}{m} \sum_{i=1}^{m} 1_{h_{S-\{x_i\}}(x_i) \neq y_i}
$$

(11)

**Lemma 4.1**

The average leave-one-out error for samples of size $m \geq 2$ is an unbiased estimate of the average generalization error for samples of size $m - 1$:

$$
E_{S \sim D^m} [\hat{R}_{LOO}(A)] = E_{S' \sim D^{m-1}} [R(h_{S'})]
$$

(12)

where $D$ denotes the distribution according to which points are drawn.
Theorem 4.1

Let $h_S$ be the hypothesis returned by SVMs for a sample $S$, and let $N_{SV}(S)$ be the number of support vectors that define $h_S$. Then,

$$E_{S \sim D^m} [R(h_S)] \leq E_{S \sim D^{m+1}} \left[ \frac{N_{SV}(S)}{m + 1} \right]$$

(13)

Note here we assume $S$ be a linearly separable sample of $m + 1$

Proof If $x$ is not a support vector for $h_S$, removing it does not change the SVM solution. Thus, $h_{S-\{x\}} = h_S$ and $h_{S-\{x\}}$ correctly classifies $x$. By contraposition, if $h_{S-\{x\}}$ misclassifies $x$, $x$ must be a support vector, which implies

$$\hat{R}_{LOO}(SVM) \leq \frac{N_{SV}(S)}{m + 1}$$

$$E_{S \sim D^m} [R(h_S)] = E_{S \sim D^{m+1}} [\hat{R}_{LOO}(SVM)] \leq E_{S \sim D^{m+1}} \left[ \frac{N_{SV}(S)}{m + 1} \right]$$
SVMs non-separable case

- Problem: data often not linearly separable in practice. For any hyperplane, there exists \( x_i \) such that \( y_i[w \cdot x_i + b] \not\geq 1 \)
- Idea: relax constraints using *slack variables* \( y_i[w \cdot x_i + b] \geq 1 - \xi_i \)

\[ w \cdot x + b = 0 \]
\[ w \cdot x + b = +1 \]
\[ w \cdot x + b = -1 \]

- Support vectors: points along the margin or outliers.
- Soft margin: \( \rho = \frac{1}{\|w\|} \)
SVMs non-separable case

Primal optimization problem

\[
\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i^\rho
\]

subject to: \( y_i(w \cdot x_i + b) \geq 1 - \xi_i \land \xi_i \geq 0, \forall i \in [1, m] \)

- Lagrangian: for all \( w \in R^N, b \in R, \alpha_i \geq 0, \beta \geq 0 \)

\[
L(w, b, \xi, \alpha, \beta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i [y_i(w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=0}^{m} \beta_i \xi_i
\]
KKT conditions:

\[
\nabla_w L = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i \\
\n\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0 \\
\n\nabla_{\xi_i} L = C - \alpha_i \beta_i = 0 \Rightarrow \alpha_i + \beta_i = C \\
\forall i, \alpha_i [y_i (w \cdot x_i + b) - 1 + \xi_i] = 0 \Rightarrow \alpha_i = 0 \lor y_i (w \cdot x_i + b) = 1 - \xi_i \\
\forall i, \beta_i \xi_i = 0 \Rightarrow \beta_i = 0 \lor \xi_i = 0 
\]
SVMs non-separable case

Dual Optimization Problem for non-separable case

\[
\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (16)
\]

subject to \(0 \leq \alpha_i \leq C \land \sum_{i=1}^{m} \alpha_i y_i = 0, \forall i \in [1, m]\)

Solution:

\[
h(x) = sgn(w \cdot x + b) = sgn\left(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b\right) \quad (17)
\]

\[
b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j \cdot x_i) \quad (18)
\]

for any \(x_i\) with \(0 < \alpha_i < C\)
SVMs non-separable case

Dual Optimization Problem for non-separable case

\[
\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)
\]

subject to \(0 \leq \alpha_i \leq C \land \sum_{i=1}^{m} \alpha_i y_i = 0, \forall i \in [1, m]\)

Solution:

\[
h(x) = \text{sgn}(w \cdot x + b) = \text{sgn}(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b)
\]

\[
b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j \cdot x_i)
\]

for any \(x_i\) with \(0 < \alpha_i < C\)
Talagrand’s lemma

Let $\Phi : \mathbb{R} \to \mathbb{R}$ be an $l$-Lipschitz. Then, for any hypothesis set $H$ of real-valued functions, the following inequality holds:

$$\hat{\mathcal{R}}(\Phi \circ H) \leq l \hat{\mathcal{R}}(H) \quad (22)$$

**Proof** First we fix a sample $S = (x_1, \ldots, x_m)$, then, by definition,

$$\hat{\mathcal{R}}_S(\Phi \circ H) = \frac{1}{m} \mathbb{E} \left[ \sup_{h \in H} \sum_{i=1}^{m} \sigma_i(\Phi \circ h)(x_i) \right]$$

$$= \frac{1}{m} \mathbb{E} \left[ \sup_{h \in H} \sum_{i=1}^{m} \sigma_i(\Phi \circ h)(x_i) \right]$$

where $u_{m-1}(h) = \sum_{i=1}^{m-1} \sigma_i(\Phi \circ h)(x_i)$. By definition of the supremum, for any $\epsilon > 0$, there exist $h_1, h_2 \in H$ such that

$$u_{m-1}(h_1) + (\Phi \circ h_1)(x_m) \geq (1 - \epsilon) \left[ \sup_{h \in H} u_{m-1}(h) + (\Phi \circ h)(x_m) \right]$$

and

$$u_{m-1}(h_2) - (\Phi \circ h_2)(x_m) \geq (1 - \epsilon) \left[ \sup_{h \in H} u_{m-1}(h) - (\Phi \circ h)(x_m) \right].$$
Talagrand's lemma cont.

Thus, for any $\epsilon > 0$, by definition of $E_{\sigma_m}$,

$$(1 - \epsilon) E_{\sigma_m} \left[ \sup_{h \in H} u_{m-1}(h) + \sigma_m(\Phi \circ h)(x_m) \right]$$

$$= (1 - \epsilon) \left[ \frac{1}{2} \sup_{h \in H} u_{m-1}(h) + (\Phi \circ h)(x_m) + \frac{1}{2} \sup_{h \in H} u_{m-1}(h) - (\Phi \circ h)(x_m) \right]$$

$$\leq \frac{1}{2} [u_{m-1}(h_1) + (\Phi \circ h_1)(x_m)] + \frac{1}{2} [u_{m-1}(h_2) - (\Phi \circ h_2)(x_m)].$$

Let $s = \text{sgn}(h_1(x_m) - h_2(x_m))$. Then, the previous inequality implies

$$(1 - \epsilon) E_{\sigma_m} \left[ \sup_{h \in H} u_{m-1}(h) + \sigma_m(\Phi \circ h)(x_m) \right]$$

$$\leq \frac{1}{2} [u_{m-1}(h_1) + u_{m-1}(h_2) + sl(h_1(x_m) - h_2(x_m))] \quad \text{(Lipschitz property)}$$

$$= \frac{1}{2} [u_{m-1}(h_1) + slh_1(x_m)] + \frac{1}{2} [u_{m-1}(h_2) - slh_2(x_m)] \quad \text{(rearranging)}$$

$$\leq \frac{1}{2} \sup_{h \in H} [u_{m-1}(h) + slh(x_m)] + \frac{1}{2} \sup_{h \in H} [u_{m-1}(h) - slh(x_m)] \quad \text{(definition of sup)}$$

$$= E_{\sigma_m} \left[ \sup_{h \in H} u_{m-1}(h) + \sigma_m lh(x_m) \right]. \quad \text{(definition of } E_{\sigma_m})$$
Since the inequality holds for all \( \varepsilon > 0 \), we have

\[
\mathbb{E} \left[ \sup_{h \in H} u_{m-1}(h) + \sigma_m (\Phi \circ h)(x_m) \right] \leq \mathbb{E} \left[ \sup_{h \in H} u_{m-1}(h) + \sigma_m l h(x_m) \right].
\]

Proceeding in the same way for all other \( \sigma_i \)'s \( i \neq m \) proves the lemma. \( \blacksquare \)

### Margin loss function

\[\rho > 0, \; L_\rho(y, y') = \Phi_\rho(yy')\]

\[
\Phi_\rho(x) = \begin{cases} 
0 & \text{if } \rho \leq x \\
1 - \frac{x}{\rho} & \text{if } 0 \leq x \leq \rho \\
1 & \text{if } x \leq 0
\end{cases}
\]
Talagrand’s lemma cont.

Since the inequality holds for all $\epsilon > 0$, we have

$$
E_{\sigma_m} \left[ \sup_{h \in H} u_{m-1}(h) + \sigma_m(\Phi \circ h)(x_m) \right] \leq E_{\sigma_m} \left[ \sup_{h \in H} u_{m-1}(h) + \sigma_m l h(x_m) \right].
$$

Proceeding in the same way for all other $\sigma_i$s ($i \neq m$) proves the lemma. ■

**Margin loss function**

$\rho > 0$, $L_\rho(y, y') = \Phi_\rho(yy')$

$$
\Phi_\rho(x) = \begin{cases} 
0 & \text{if } \rho \leq x \\
1 - \frac{x}{\rho} & \text{if } 0 \leq x \leq \rho \\
1 & \text{if } x \leq 0 
\end{cases}
$$
Empirical margin loss

Given a sample $S = (x_1, ..., x_m)$ and a hypothesis $h$, the empirical margin loss is defined by

$$\hat{R}_\rho(h) = \frac{1}{m} \sum_{i=1}^{m} \Phi_\rho(y_i h(x_i))$$  \hspace{1cm} (23)$$

$$\hat{R}_\rho(h) \leq \frac{1}{m} \sum_{i=1}^{m} 1_{y_i h(x_i) \leq \rho}$$  \hspace{1cm} (24)$$
Margin Theory

Theorem 4.4 Margin bound for binary classification

Let $H$ be a set of real-valued functions. Fix $\rho > 0$, then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all $h \in H$:

\[
R(h) \leq \hat{R}_\rho(h) + \frac{2}{\rho} \mathcal{R}_m(H) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{25}
\]

\[
R(h) \leq \hat{R}_\rho(h) + \frac{2}{\rho} \hat{\mathcal{R}}_S(H) + 3 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \tag{26}
\]
Proof Let $\tilde{H} = \{ z = (x, y) \mapsto yh(x) : h \in H \}$. Consider the family of functions taking values in $[0, 1]$:

$$\tilde{\mathcal{H}} = \{ \Phi_\rho \circ f : f \in \tilde{H} \}.$$ 

By theorem 3.1, with probability at least $1 - \delta$, for all $g \in \tilde{\mathcal{H}}$,

$$\mathbb{E}[g(z)] \leq \frac{1}{m} \sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(\tilde{\mathcal{H}}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}},$$

and thus, for all $h \in H$,

$$\mathbb{E}[\Phi_\rho(yh(x))] \leq \hat{R}_\rho(h) + 2\mathcal{R}_m(\Phi_\rho \circ \tilde{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$
Since $1_{u \leq 0} \leq \Phi_\rho(u)$ for all $u \in \mathbb{R}$, we have $R(h) = \mathbb{E}[1_{y_h(x) \leq 0}] \leq \mathbb{E}[\Phi_\rho(y_h(x))]$, thus

$$R(h) \leq \tilde{R}_\rho(h) + 2\mathcal{R}_m(\Phi_\rho \circ \tilde{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$ \hspace{1cm} (4.40)

$\mathcal{R}_m$ is invariant to a constant shift, therefore we have

$$\mathcal{R}_m(\Phi_\rho \circ \tilde{H}) = \mathcal{R}_m((\Phi_\rho - 1) \circ \tilde{H}).$$

Since $(\Phi_\rho - 1)(0) = 0$ and since $(\Phi_\rho - 1)$ is $1/\rho$-Lipschitz as with $\Phi_\rho$, by lemma 4.2, we have $\mathcal{R}_m(\Phi_\rho \circ \tilde{H}) \leq \frac{1}{\rho} \mathcal{R}_m(\tilde{H})$ and $\mathcal{R}_m(\tilde{H})$ can be rewritten as follows:

$$\mathcal{R}_m(\tilde{H}) = \frac{1}{m} \mathbb{E}_{S,\sigma} \left[ \sup_{h \in H} \sum_{i=1}^{m} \sigma_i y_i h(x_i) \right] = \frac{1}{m} \mathbb{E}_{S,\sigma} \left[ \sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \right] = \mathcal{R}_m(H).$$

This proves (4.40). The second inequality, (4.41), can be derived in the same way by using the second inequality of theorem 3.1, (3.4), instead of (3.3). ⊢
Theorem 4.5

Let $H$ be a set of real-valued functions. Then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all $h \in H$ and $\rho \in (0, 1)$:

$$R(h) \leq \hat{R}_\rho(h) + \frac{4}{\rho} \mathcal{R}_m(H) + \sqrt{\log \log_2 \frac{2}{\rho}} \sqrt{\frac{m}{\rho}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$  \hspace{0.5cm} (27)

$$R(h) \leq \hat{R}_\rho(h) + \frac{4}{\rho} \hat{\mathcal{R}}(H) + \sqrt{\log \log_2 \frac{2}{\rho}} \sqrt{\frac{m}{\rho}} + 3 \sqrt{\frac{\log \frac{4}{\delta}}{2m}}$$  \hspace{0.5cm} (28)
Theorem 4.5 cont.

**Proof** Consider two sequences \((\rho_k)_{k \geq 1}\) and \((\epsilon_k)_{k \geq 1}\), with \(\epsilon_k \in (0, 1)\). By theorem 4.4, for any fixed \(k \geq 1\),

\[
\Pr \left[ R(h) - \hat{R}_{\rho_k}(h) > \frac{2}{\rho_k} \mathcal{R}_m(H) + \epsilon_k \right] \leq \exp(-2m\epsilon_k^2). \tag{4.44}
\]

Choose \(\epsilon_k = \epsilon + \sqrt{\frac{\log k}{m}}\), then, by the union bound,

\[
\Pr \left[ \exists k: R(h) - \hat{R}_{\rho_k}(h) > \frac{2}{\rho_k} \mathcal{R}_m(H) + \epsilon_k \right] \leq \sum_{k \geq 1} \exp(-2m\epsilon_k^2)
\]

\[
= \sum_{k \geq 1} \exp \left[ -2m(\epsilon + \sqrt{\frac{\log k}{m}})^2 \right]
\]

\[
\leq \sum_{k \geq 1} \exp(-2m\epsilon^2) \exp(-2 \log k)
\]

\[
= \left( \sum_{k \geq 1} \frac{1}{k^2} \right) \exp(-2m\epsilon^2)
\]

\[
= \frac{\pi^2}{6} \exp(-2m\epsilon^2) \leq 2 \exp(-2m\epsilon^2).
\]
We can choose \( \rho_k = 1/2^k \). For any \( \rho \in (0, 1) \), there exists \( k \geq 1 \) such that \( \rho \in (\rho_k, \rho_{k-1}] \), with \( \rho_0 = 1 \). For that \( k \), \( \rho \leq \rho_{k-1} = 2\rho_k \), thus \( 1/\rho_k \leq 2/\rho \) and \( \log k = \sqrt{\log \log_2(1/\rho_k)} \leq \sqrt{\log \log_2(2/\rho)} \). Furthermore, for any \( h \in H \), \( \hat{R}_{\rho_k}(h) \leq \hat{R}_\rho(h) \). Thus,

\[
\Pr \left[ \exists k : R(h) - \hat{R}_\rho(h) > \frac{4}{\rho} \mathcal{R}_m(H) + \sqrt{\frac{\log \log_2(2/\rho)}{m}} + \epsilon \right] \leq 2 \exp(-2m\epsilon^2),
\]
Theorem 4.3

Let $S \subseteq \{x : \|x\| \leq R\}$ be a sample of size $m$ and let $H = \{x \rightarrow w \cdot x : \|w\| \leq \Lambda\}$. Then, the empirical Rademacher complexity of $H$ can be bounded as follows:

$$\hat{R}_S(H) \leq \sqrt{\frac{R^2\Lambda^2}{m}}$$

(29)

\[
\begin{align*}
\hat{R}_S(H) &= \frac{1}{m} \mathbb{E}_{\sigma} \left[ \sup_{\|w\| \leq \Lambda} \sum_{i=1}^{m} \sigma_i w \cdot x_i \right] \\
&= \frac{1}{m} \mathbb{E}_{\sigma} \left[ \sup_{\|w\| \leq \Lambda} w \cdot \sum_{i=1}^{m} \sigma_i x_i \right] \\
&\leq \frac{\Lambda}{m} \mathbb{E} \left[ \left\| \sum_{i=1}^{m} \sigma_i x_i \right\| \right] \\
&\leq \frac{\Lambda}{m} \left[ \mathbb{E} \left[ \left\| \sum_{i=1}^{m} \sigma_i x_i \right\|^2 \right] \right]^{1/2} \\
&\leq \frac{\Lambda \sqrt{mR^2}}{m} \leq \sqrt{\frac{R^2\Lambda^2}{m}}.
\end{align*}
\]
Combining theorem 4.3 and theorem 4.4 gives directly the following corollary.

**Corollary 4.1**

Let $H = \{ x \rightarrow w \cdot x : \|w\| \leq \Lambda \}$ and assume that $X \subseteq \{ x : \|x\| \leq r \}$. Fix $\rho > 0$, then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$, 

$$R(h) \leq \hat{R}_\rho(h) + 2\sqrt{r^2 \Lambda^2 / \rho^2} \frac{1}{m} + \sqrt{\log \frac{1}{\delta} / 2m}$$

(30)

- generalization bound does not depend on the dimension but on the margin.
- this suggests seeking a large-margin separating hyperplane in a higher-dimensional feature space.
Mohri, Mehryar and Rostamizadeh, Afshin and Talwalkar, Ameet (2012)
Foundations of machine learning

Mehryar Mohri
The End