

An introduction to the analysis of support vector machines

Ingo Steinwart
LAUR 06-6249

September 13, 2006

Informal Description of the Learning Goal

- ▶ X space of input samples
 Y space of labels, usually $Y \subset \mathbb{R}$.
- ▶ Already observed samples

$$T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$$

- ▶ **Goal:**
With the help of T find a function $f : X \rightarrow \mathbb{R}$ which predicts label y for new, unseen x .
- ▶ **Question:**
How do we assess the quality of f ?

Illustration: Binary Classification

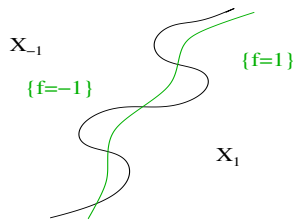
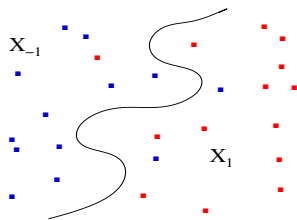
► **Problem:**

The set X is divided into two *unknown* classes X_{-1} and X_1 .

► **Goal:**

Find approximately the classes X_{-1} and X_1 .

► **Illustration:**



Left: Negative (blue) and positive (red) samples.

Right: Behaviour of a decision function (green) $f : X \rightarrow Y$.

Machine Learning: Applications

Some Applications

- ▶ Handwritten character recognition
- ▶ Diagnostics in engineering and medical science
- ▶ Bioinformatics
- ▶ Internet search engines

Formal Definition of Statistical Learning

► Basic Assumptions:

- P is an *unknown* probability measure on $X \times Y$.
- $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ sampled from P^n .
- $L : Y \times \mathbb{R} \rightarrow [0, \infty]$ *loss function* that measures cost $L(y, t)$ of predicting y by t .
- Future (x, y) will also be sampled from P .

► Goal:

Find a function $f_T : X \rightarrow \mathbb{R}$ with small *risk*

$$\mathcal{R}_{L,P}(f_T) := \int_{X \times Y} L(y, f_T(x)) dP(x, y) .$$

► Interpretation:

Average future cost of predicting by f_T should be small.

Questions in Statistical Learning I

► **Bayes risk:**

$$\mathcal{R}_{L,P}^* := \inf \{ \mathcal{R}_{L,P}(f) \mid f : X \rightarrow \mathbb{R} \text{ measurable} \} .$$

► **Learning method:**

Assigns to every training set T a predictor $f_T : X \rightarrow \mathbb{R}$.

► **Consistency:**

A learning method is called *universally consistent* if

$$\mathcal{R}_{L,P}(f_T) \rightarrow \mathcal{R}_{L,P}^* \quad \text{in probability} \quad (1)$$

for $n \rightarrow \infty$ and every probability measure P on $X \times Y$.

► **Good news:**

Many learning methods are universally consistent.

First result: Stone (1977), AoS

Questions in Statistical Learning II

► **Rates:**

Does there exist a learning method and a convergence rate $a_n \searrow 0$ such that

$$\mathbb{E}_{T \sim P^n} \mathcal{R}_{L,P}(f_T) - \mathcal{R}_{L,P}^* \leq C_P a_n, \quad n \geq 1,$$

for every probability measure P on $X \times Y$.

► **Bad news:** (Devroye, 1982, IEEE TPAMI)

No! (if $|Y| \geq 2$, $|X| = \infty$, and L “non-trivial”)

► **Good news:**

Yes, if one makes some “mild?!” assumptions on P .

Too many results in this direction to mention them.

Questions in Statistical Learning III

▶ **The “ideal” learning algorithm:**

- ▶ Is universally consistent.
- ▶ Has “optimal” learning rates on certain classes of distributions.
- ▶ Runs efficiently on a computer.
- ▶ Has a good record on real-world problems.

▶ **So far:**

- ▶ It has not been found.
- ▶ But research is getting closer to it

Reproducing Kernel Hilbert Spaces I

- ▶ $k : X \times X \rightarrow \mathbb{R}$ is a **kernel**

: \Leftrightarrow there exist a Hilbert space H and a map $\Phi : X \rightarrow H$ with

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle \quad \text{for all } x, x' \in X.$$

\Leftrightarrow all $(k(x_i, x_j))_{i,j=1}^n$ are symmetric and positive semi-definite.

- ▶ **RKHS** of k : the “smallest” such H consisting of functions.
 - ▶ “Construction”: Take the “completion” of

$$\left\{ \sum_{i=1}^n \alpha_i k(x_i, \cdot) : n \in \mathbb{N}, \alpha_1, \dots, \alpha_n \in \mathbb{R}, x_1, \dots, x_n \in X \right\}$$

equipped with the dot product

$$\left\langle \sum_{i=1}^n \alpha_i k(x_i, \cdot), \sum_{j=1}^m \beta_j k(\hat{x}_j, \cdot) \right\rangle := \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i, \hat{x}_j).$$

- ▶ **Feature map:** $\Phi : x \mapsto k(x, \cdot)$.

Reproducing Kernel Hilbert Spaces II

► **Polynomial Kernels:**

For $a \geq 0$ and $m \in \mathbb{N}$ let

$$k(x, x') := (\langle x, x' \rangle + a)^m, \quad x, x' \in \mathbb{R}^d .$$

► **Gaussian RBF kernels:**

For $\sigma > 0$ let

$$k_\sigma(x, x') := \exp(-\sigma^2 \|x - x'\|_2^2), \quad x, x' \in \mathbb{R}^d .$$

The parameter $1/\sigma$ is called *width*.

► **Denseness of Gaussian RKHSs:**

The RKHS H_σ of k_σ is dense in $L_p(\mu)$ for all $p \in [1, \infty)$ and all probability measures μ on \mathbb{R}^d .

Support Vector Machines I

- ▶ **Support vector machines (SVMs)** solve the problem

$$f_{T,\lambda} = \arg \min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)) , \quad (2)$$

- ▶ H is a RKHS,
- ▶ $T = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ is a training set,
- ▶ $\lambda > 0$ is a *free* regularization parameter,
- ▶ $L : Y \times \mathbb{R} \rightarrow [0, \infty)$ is a *convex* loss, e.g.
 - ▶ *hinge loss*: $L(y, t) := \max\{0, 1 - yt\}$
 - ▶ *least squares loss*: $L(y, t) := (y - t)^2$.
- ▶ **Representer Theorem:**
The *unique* solution is of the form $f_{T,\lambda} = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$.
Minimization actually takes place over $\{\alpha_1, \dots, \alpha_n\}$.

Support Vector Machines II

Questions:

- ▶ Universally consistent?
- ▶ Learning rates?
- ▶ Efficient algorithms?
- ▶ Performance on real world problems?
- ▶ Additional properties?

Notations

► Empirical risk

$$\mathcal{R}_{L,T}(f) := \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i))$$

► Regularized risks

$$\mathcal{R}_{L,T,\lambda}(f) := \lambda \|f\|_H^2 + \mathcal{R}_{L,T}(f)$$

$$\mathcal{R}_{L,P,\lambda}(f) := \lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f)$$

► SVM solutions

$$f_{T,\lambda} := \arg \min_{f \in H} \mathcal{R}_{L,T,\lambda}(f)$$

$$f_{P,\lambda} := \arg \min_{f \in H} \mathcal{R}_{L,P,\lambda}(f)$$

Basic Assumptions and Consequences

- ▶ **Boundedness:** $L(y, 0) \leq 1$ for all $y \in Y$.

Consequence: $\|f_{P,\lambda}\|_H \leq \lambda^{-1/2}$.

$$\lambda \|f_{P,\lambda}\|_H^2 \leq \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) \leq \mathcal{R}_{L,P,\lambda}(0) \leq 1$$

- ▶ **Lipschitz continuity:** $|L(y, t) - L(y, t')| \leq |t - t'|$.

Consequence I: $\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}(g) \leq \|f - g\|_\infty$.

Consequence II: $L(y, t) \leq 1 + |t|$.

- ▶ **Compactness:** X compact and k continuous with $\|k\|_\infty \leq 1$.

Consequence I: H is compactly embedded in $C(X)$.

Proof uses Arzelà-Ascoli

Consequence II: $\|f\|_\infty \leq \|f\|_H$.

$$|f(x)| = |\langle f, k(x, \cdot) \rangle| \leq \|f\|_H \cdot \|k(x, \cdot)\|_H = \|f\|_H \cdot \sqrt{k(x, x)}$$

The Basic Decomposition I

Goal: Estimate

$$P^n \left(T \in (X \times Y)^n : \mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) \leq \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) + \varepsilon \right)$$

To this end we observe

$$\begin{aligned} & \mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) - \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) \\ \leq & \mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) - \mathcal{R}_{L,T,\lambda}(f_{T,\lambda}) & = \mathcal{R}_{L,P}(f_{T,\lambda}) - \mathcal{R}_{L,T}(f_{T,\lambda}) \\ & + \mathcal{R}_{L,T,\lambda}(f_{T,\lambda}) - \mathcal{R}_{L,T,\lambda}(f_{P,\lambda}) & \leq 0 \\ & + \mathcal{R}_{L,T,\lambda}(f_{P,\lambda}) - \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) & = \mathcal{R}_{L,T}(f_{P,\lambda}) - \mathcal{R}_{L,P}(f_{P,\lambda}) \\ \leq & 2 \sup_{f \in \lambda^{-1/2} B_H} |\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f)| \end{aligned}$$

The Basic Decomposition I

Goal: Estimate

$$P^n \left(T \in (X \times Y)^n : \mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) \leq \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) + \varepsilon \right)$$

So far

$$\mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) - \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) \leq 2 \sup_{f \in \lambda^{-1/2} B_H} |\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f)|$$

Let \mathcal{G}_δ be a δ -net of $\lambda^{-1/2} B_H$ with respect to $\|\cdot\|_\infty$. Then:

$$\sup_{f \in \lambda^{-1/2} B_H} |\mathcal{R}_{L,T}(f) - \mathcal{R}_{L,P}(f)| \leq \sup_{g \in \mathcal{G}_\delta} |\mathcal{R}_{L,T}(g) - \mathcal{R}_{L,P}(g)| + 2\delta$$

The Basic Decomposition I

Goal: Estimate

$$P^n \left(T \in (X \times Y)^n : \mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) \leq \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) + \varepsilon \right)$$

So far

$$\mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) - \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) \leq 2 \sup_{g \in \mathcal{G}_\delta} |\mathcal{R}_{L,T}(g) - \mathcal{R}_{L,P}(g)| + 4\delta.$$

This implies

$$\begin{aligned} & P^n \left(T \in (X \times Y)^n : \mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) \leq \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) + 2\varepsilon + 4\delta \right) \\ & \geq P^n \left(T \in (X \times Y)^n : \sup_{g \in \mathcal{G}_\delta} |\mathcal{R}_{L,T}(g) - \mathcal{R}_{L,P}(g)| \leq \varepsilon \right) \\ & \geq 1 - \sum_{g \in \mathcal{G}_\delta} P^n \left(T \in (X \times Y)^n : |\mathcal{R}_{L,T}(g) - \mathcal{R}_{L,P}(g)| > \varepsilon \right) \end{aligned}$$

Hoeffding's Inequality

So far:

$$\begin{aligned} & P^n \left(T \in (X \times Y)^n : \mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) \leq \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) + 2\varepsilon + 4\delta \right) \\ & \geq 1 - \sum_{g \in \mathcal{G}_\delta} P^n \left(T \in (X \times Y)^n : |\mathcal{R}_{L,T}(g) - \mathcal{R}_{L,P}(g)| > \varepsilon \right) \end{aligned}$$

Hoeffding's inequality: For all $h : Z \rightarrow [0, B]$ we have

$$P^n \left((z_1, \dots, z_n) \in Z^n : \left| \frac{1}{n} \sum_{i=1}^n h(z_i) - \mathbb{E}_P h \right| > \varepsilon \right) \leq 2e^{-\frac{2\varepsilon^2 n}{B^2}}$$

In our case $h(x, y) := L(y, g(x)) \leq 1 + \|g\|_\infty \leq 1 + \lambda^{-1/2}$ leads to

$$P^n \left(T \in (X \times Y)^n : |\mathcal{R}_{L,T}(g) - \mathcal{R}_{L,P}(g)| > \varepsilon \right) \leq 2e^{-\frac{2\varepsilon^2 \lambda n}{(1+\sqrt{\lambda})^2}}$$

An Oracle Inequality

So far:

$$P^n \left(\mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) \leq \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) + 2\varepsilon + 4\delta \right) \leq 1 - 2|\mathcal{G}_\delta| e^{-\frac{2\varepsilon^2 \lambda n}{(1+\sqrt{\lambda})^2}}$$

Assumption: there are constants $a \geq 1$ and $p > 0$ such that

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \leq a\varepsilon^{-2p}, \quad \varepsilon > 0.$$

Then we have $\ln |\mathcal{G}_\delta| \leq a\lambda^{-1}\delta^{-2p}$ and hence with probability not less than $1 - e^{-\tau}$ we have

$$\mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) < \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) + 8 \left(\frac{a}{\lambda^{1+p}n} \right)^{\frac{1}{2+2p}} + 4\sqrt{\frac{\tau}{\lambda n}}$$

An Oracle Inequality: Consequences

Oracle Inequality:

With probability not less than $1 - e^{-\tau}$ we have

$$\mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) \leq \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) + 8 \left(\frac{a}{\lambda^{1+p}n} \right)^{\frac{1}{2+2p}} + 4 \sqrt{\frac{\tau}{\lambda n}}$$

Example: H fixed Gaussian RKHS

- ▶ We have $\lim_{\lambda \rightarrow 0} \mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) = \mathcal{R}_{L,P}^*$.
- ▶ We can choose p to be arbitrarily close to 0.

Consequence I:

If $\lambda_n \rightarrow 0$ and $\lambda_n^{1+p} n \rightarrow \infty$ for some $p > 0$ then the SVM is universally consistent.

Consequence II:

The oracle inequality can be used to select λ and σ in a data-dependent fashion.

Oracle too Conservative: Reason I

Observation:

We often have constants $c > 0$ and $\alpha \in (0, 1]$ such that

$$\mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) - \mathcal{R}_{L,P}^* \leq c\lambda^\alpha, \quad \lambda > 0.$$

By the oracle inequality we then have with high probability that

$$\lambda \|f_{T,\lambda}\|_H^2 \leq \mathcal{R}_{L,P,\lambda}(f_{T,\lambda}) - \mathcal{R}_{L,P}^* \leq c\lambda^\alpha + 8 \left(\frac{a}{\lambda^{1+p}n} \right)^{\frac{1}{2+2p}} + 4 \sqrt{\frac{\tau}{\lambda n}}.$$

⇒ For “typical” values of λ and n our initial bound $\lambda \|f_{T,\lambda}\|_H^2 \leq 1$ is too loose!

Oracle too Conservative: Reason II

Observation:

- ▶ For P satisfying $\mathcal{R}_{L,P}(f) = 0$ for a suitable $f \in H$ rates up to the order of n^{-1} are possible.
- ▶ The oracle inequality cannot produce rates faster than $n^{-1/2}$.

Main ingredients of such an approach:

- ▶ Consider functions of the form

$$g_{f,r} := \frac{\mathbb{E}L \circ f - L \circ f}{r + \mathbb{E}L \circ f}, \quad r > 0, f \in H.$$

- ▶ Use a **variance bound** $\mathbb{E}(L \circ f)^2 \leq B\mathbb{E}L \circ f$.
- ▶ Apply **Bernstein's inequality** instead of Hoeffding's inequality.

But: This approach still uses $\|\cdot\|_\infty$ -covering numbers.

How to use more suitable covering numbers

Roadmap:

- ▶ Use a stronger concentration inequality which can directly deal with suprema.
- ▶ Use covering numbers to bound expectations and *not* concentrations.

Main ingredient of this approach: Talagrand's inequality

Talagrand's inequality

Talagrand's inequality

$\mathcal{G} \subset \mathcal{L}_\infty(Z)$ separable such that $\mathbb{E}_\mu g = 0$, $\mathbb{E}_\mu g^2 \leq \sigma^2$ and $\|g\|_\infty \leq B$ for all $g \in \mathcal{G}$. Define $h : Z^n \rightarrow \mathbb{R}$ by

$$h(z_1, \dots, z_n) := \sup_{g \in \mathcal{G}} \left| \sum_{j=1}^n g(z_j) \right|, \quad z = (z_1, \dots, z_n) \in Z.$$

Then for all $\tau > 0$ and $\gamma > 0$ we have

$$P^n \left(\left\{ z \in Z^n : h(z) \geq (1+\gamma)\mathbb{E}_{P^n} h + \sqrt{2\tau n\sigma^2} + \left(\frac{2}{3} + \frac{1}{\gamma}\right)\tau B \right\} \right) \leq e^{-\tau}.$$

Interpretation: Generalization of Bernstein's inequality to suprema.

Applying Talagrand's inequality I

Notations:

- ▶ $Z := X \times Y$ and $z := (x, y)$.
- ▶ $f_{L,P}^* : X \rightarrow \mathbb{R}$ function satisfying $\mathcal{R}_{L,P}(f_{L,P}^*) = \mathcal{R}_{L,P}^*$.
- ▶ $g_f(x, y) := L(y, f(x)) - L(y, f_{L,P}^*(x))$
- ▶ For $r > 0$ define $h_r : Z^n \rightarrow \mathbb{R}$ by

$$h_r(z_1, \dots, z_n) := \sup_{\substack{f \in H \\ \lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^* \leq r}} \left| \sum_{j=1}^n \frac{\mathbb{E} g_f - g_f(z_j)}{\mathbb{E} g_f + r} \right|.$$

Talagrand's inequality yields

$$P^n \left(\left\{ z \in Z^n : h_r(z) \geq (1+\gamma) \mathbb{E}_{P^n} h_r + \sqrt{2\tau n \sigma^2} + \left(\frac{2}{3} + \frac{1}{\gamma} \right) \tau B \right\} \right) \leq e^{-\tau}.$$

Applying Talagrand's inequality II

Major difficulty: Find an upper bound of

$$\mathbb{E}_{P^n} h_r = \mathbb{E}_{(z_1, \dots, z_n) \sim P^n} \sup_{\substack{f \in H \\ \lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^* \leq r}} \left| \sum_{j=1}^n \frac{\mathbb{E} g_f - g_f(z_j)}{\mathbb{E} g_f + r} \right|.$$

Main steps:

- ▶ “Symmetrization” \Rightarrow omits $\mathbb{E} g_f$ in the numerator
adds outer \mathbb{E} over Rademacher variables
- ▶ “Peeling” \Rightarrow omit $\mathbb{E} g_f + r$ in the denominator.
- ▶ “Dudley’s entropy integral” bounds the rest by entropy numbers.

The Final Oracle Inequality

Assumptions and notations:

- ▶ $Y \subset [-1, 1]$.
- ▶ $L(y, \cdot) : \mathbb{R} \rightarrow [0, \infty)$ has a minimum in $[-1, 1]$.
- ▶ $\check{t} := \max\{-1, \min\{1, t\}\}$.
- ▶ Variance bound:
 $\exists V \geq 1$ and $\vartheta \in [0, 1] \forall f \in H$:

$$\mathbb{E}_P(L \circ \check{f} - L \circ f_{L,P}^*)^2 \leq V \cdot (\mathbb{E}_P(L \circ \check{f} - L \circ f_{L,P}^*))^{\vartheta}$$

- ▶ Entropy assumption:
 $\exists a \geq 1$ and $p \in (0, 1)$:

$$\sup_{D \in \mathcal{X}^n} e_m(B_H, L_2(D)) \leq a m^{-\frac{1}{2p}}, \quad m, n \geq 1.$$

The Final Oracle Inequality

The improved oracle inequality (slightly simplified)

- ▶ Variance bound:

$$\mathbb{E}_P(L \circ \check{f} - L \circ f_{L,P}^*)^2 \leq V \cdot (\mathbb{E}_P(L \circ \check{f} - L \circ f_{L,P}^*))^{\vartheta}$$

- ▶ Entropy assumption:

$$\sup_{D \in X^n} e_m(B_H, L_2(D)) \leq a m^{-\frac{1}{2p}}, \quad m, n \geq 1.$$

Then with probability not less than $1 - e^{-\tau}$ we have

$$\mathcal{R}_{L,P}(\check{f}_{T,\lambda}) - \mathcal{R}_{L,P}^* \leq K \cdot \left(\mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) - \mathcal{R}_{L,P}^* + \tau \left(\frac{a^{2p}}{\lambda p n} \right)^{\frac{1}{2-p-\vartheta(1-p)}} \right)$$

Discussion I

- ▶ New oracle inequality gives the best known (and sometimes optimal) rates for a-priori chosen λ_n .
- ▶ Using the oracle inequality in model selection to find optimal λ in a data-dependent way recovers these rates.
No knowledge on distribution is required!
- ▶ In general, optimality for interesting distributions is unknown.
- ▶ Actual oracle inequality holds for a variety of penalized empirical risk minimizers.

Discussion II

Example: Using Gaussian RKHSs with different widths.

- ▶ Entropy assumption $\sup_{D \in X^n} e_m(B_{H_\sigma}, L_2(D)) \leq a_\sigma m^{-\frac{1}{2p}}$ implies oracle inequality

$$\mathcal{R}_{L,P}(\check{f}_{T,\lambda}) - \mathcal{R}_{L,P}^* \leq K \cdot \left(\mathcal{R}_{L,P,\lambda}(f_{P,\lambda}) - \mathcal{R}_{L,P}^* + \tau \left(\frac{a_\sigma^{2p}}{\lambda p n} \right)^{\frac{1}{2-p-\vartheta(1-p)}} \right)$$

- ▶ If σ changes with n so does a_σ .
⇒ substantial influence on RHS of oracle inequality.
- ▶ **Question:** What is the behaviour of a_σ ?

Discussion III

Question: What is the behaviour of a_σ in

$$\sup_{D \in X^n} e_m(B_{H_\sigma}, L_2(D)) \leq a_\sigma m^{-\frac{1}{2p}}$$

- ▶ “Off-the-shelf-estimate” using e.g. $H_\sigma \subset C^s$ does not yield suitable estimate for a_σ .
- ▶ Zhou, JOC, 2002, proved an exponential bound in m but his constant behaves like σ^{2d+1}
- ▶ Only known result which balances p and σ in a suitable fashion is (roughly) of the form $a_\sigma = c\sigma^{(1-p)d}$.