Metric Entropy in Learning Theory and Small Deviations

Thomas Kühn

Universität Leipzig, Germany

NSF/CBMS Regional Conference in Mathematical Sciences

"Small Deviation Probabilities: Theory and Applications"

University of Alabama in Huntsville, 4-8 June 2012

Outline of the talk

1. Introduction to Learning Theory - some examples
2. A formal model of learning
3. Error analysis and entropy numbers
4. Covering numbers of Gaussian RKHs and small deviations of smooth Gaussian processes

1. A short introduction to Learning Theory

1. A short introduction to Learning Theory

- Learning Theory
  - goal: to approximate an unknown function (or some features of a function) from data samples, possibly perturbed by noise
  - Learning Theory relies on
    - statistics (draw information from random samples)
    - approximation theory
    - functional analysis

1. A short introduction to Learning Theory

- Learning Theory
  – goal: to approximate an unknown function (or some features of a function) from data samples, possibly perturbed by noise
  – Learning Theory relies on
    – statistics (draw information from random samples)
    – approximation theory
    – functional analysis

- Literature
  – Monograph by Felipe Cucker and Ding-Xuan Zhou
  "Learning Theory. An Approximation Theory Viewpoint"
  Cambridge University Press 2007

  – Survey Article by Felipe Cucker and Steve Smale
  "On the mathematical foundations of learning"
  Bull. Amer. Math. Soc. 39 (2002), 1–49.

- Example 1. Linear regression

- Example 1. Linear regression

– Let m data points $(x_1, y_1), ..., (x_m, y_m) \in \mathbb{R}^2$ be given.

- Example 1. Linear regression

– Let m data points $(x_1, y_1), ..., (x_m, y_m) \in \mathbb{R}^2$ be given.

– We seek the line $y = ax + b$ that fits best to the data.

- Example 1. Linear regression

– Let m data points $(x_1, y_1), ..., (x_m, y_m) \in \mathbb{R}^2$ be given.

– We seek the line $y = ax + b$ that fits best to the data.

– This line should minimize the quadratic error

$$Q(a,b) = \sum_{i=1}^{m} (y_i - ax_i - b)^2$$

- Example 1. Linear regression

– Let m data points $(x_1, y_1), ..., (x_m, y_m) \in \mathbb{R}^2$ be given.

– We seek the line $y = ax + b$ that fits best to the data.

– This line should minimize the quadratic error

$$Q(a, b) = \sum_{i=1}^{m} (y_i - ax_i - b)^2$$

- Example 2. Curve fitting

- Example 1. Linear regression

– Let m data points $(x_1, y_1), ..., (x_m, y_m) \in \mathbb{R}^2$ be given.

– We seek the line $y = ax + b$ that fits best to the data.

– This line should minimize the quadratic error

$$Q(a, b) = \sum_{i=1}^{m} (y_i - ax_i - b)^2$$

- Example 2. Curve fitting

– Assume a physical law is goverened by an unknown function $f : \mathbb{R} \to \mathbb{R}$, which can be parametrized by N parameters, e.g.

$f = f_a = \sum_{j=0}^{d} a_j x^j =$ polynomial of degree $\leq d$, i.e. $N = d + 1$

- Example 1. Linear regression

– Let m data points $(x_1, y_1), \dots, (x_m, y_m) \in \mathbb{R}^2$ be given.

– We seek the line $y = ax + b$ that fits best to the data.

– This line should minimize the quadratic error

$$Q(a,b) = \sum_{i=1}^{m} (y_i - ax_i - b)^2$$

- Example 2. Curve fitting

– Assume a physical law is goverened by an unknown function $f : \mathbb{R} \to \mathbb{R}$, which can be parametrized by N parameters, e.g.

$f = f_a = \sum_{j=0}^{d} a_j x^j =$ polynomial of degree $\leq d$, i.e. $N = d + 1$

– exact measurements $\curvearrowright y_i = f(x_i)$

– noisy data $\curvearrowright y_i = f(x_i) + \varepsilon$, where ε is a r.v. of mean 0

– We seek the coefficient vector $a = (a_0, a_1, ..., a_d) \in \mathbb{R}$ such that

$$\sum_{i=1}^{m} (f_a(x_j) - y_i)^2 \to \min .$$

– We seek the coefficient vector $a = (a_0, a_1, ..., a_d) \in \mathbb{R}$ such that

$$\sum_{i=1}^{m} (f_a(x_j) - y_i)^2 \to \min .$$

– Noisy data. Let ε_x, $x \in \mathbb{R}$, be a family of random variables with $\mathbb{E}\varepsilon_x = f(x)$. Then the y_i are drawn randomly from ε_{x_i}.

- We seek the coefficient vector $a = (a_0, a_1, ..., a_d) \in \mathbb{R}$ such that

$$\sum_{i=1}^{m} (f_a(x_j) - y_i)^2 \to \min .$$

- Noisy data. Let ε_x, $x \in \mathbb{R}$, be a family of random variables with $\mathbb{E}\varepsilon_x = f(x)$. Then the y_i are drawn randomly from ε_{x_i}.
- Sometimes: x_i chosen randomly from a probability ρ_X on $\mathbb{R}$.
- More general starting point:
  measure ρ on $\mathbb{R} \times \mathbb{R}$ capturing both ρ_X and ε_x.

- Example 3. Pattern recognition

- Example 3. Pattern recognition

X – matrices with entries in the interval $[0, 1]$
Interpretation: Each entry represents a pixel in a gray scale of a digitized photograph of a handwritten letter A,B,...,Z.

- Example 3. Pattern recognition

X – matrices with entries in the interval $[0, 1]$
Interpretation: Each entry represents a pixel in a gray scale of a digitized photograph of a handwritten letter A,B,...,Z.

$Y = \left\{y = (\lambda_j) \in \mathbb{R}^{26} : \lambda_j \geq 0, \sum \lambda_j = 1\right\}$ with the interpretation
$\lambda_1 = Prob(x = A), \lambda_2 = Prob(x = B), ..., \lambda_{26} = Prob(x = Z)$

- Example 3. Pattern recognition

X – matrices with entries in the interval $[0,1]$
Interpretation: Each entry represents a pixel in a gray scale of a digitized photograph of a handwritten letter A,B,...,Z.

$Y = \{y = (\lambda_j) \in \mathbb{R}^{26} : \lambda_j \geq 0, \sum \lambda_j = 1\}$ with the interpretation
$\lambda_1 = Prob(x = A), \lambda_2 = Prob(x = B), ..., \lambda_{26} = Prob(x = Z)$

– Goal: to learn the ideal function $f : X \to Y$ which associates to a handwritten letter the vector y of probabilities.
– "Learning" f means to find a good approximation to f within a given class.

- Example 3. Pattern recognition

X – matrices with entries in the interval $[0,1]$
Interpretation: Each entry represents a pixel in a gray scale of a digitized photograph of a handwritten letter A,B,...,Z.

$Y = \left\{ y = (\lambda_j) \in \mathbb{R}^{26} : \lambda_j \geq 0, \sum \lambda_j = 1 \right\}$ with the interpretation
$\lambda_1 = Prob(x = A), \lambda_2 = Prob(x = B), ..., \lambda_{26} = Prob(x = Z)$

– Goal: to learn the ideal function $f : X \to Y$ which associates to a handwritten letter the vector y of probabilities.
– "Learning" f means to find a good approximation to f within a given class.

– The approximation to f is constructed from a set of samples of handwritten letters x, each with a label y.
– Samples (x_i, y_i) are drawn randomly from a probability ρ on $X \times Y$.
– In practice, ρ is concentrated around pairs (x, e_j).
– The function f to be learned is the regression function f_ρ.

Roughly speaking: $f_\rho(x) =$ average of the y-values of $\{x\} \times Y$

- 2. A formal model of learning

- 2. A formal model of learning

  X – a compact metric space (e.g. a domain in $\mathbb{R}^n$)
  $Y = \mathbb{R}^k$ – for simplicity $k = 1$
  ρ – a probability measure on $Z := X \times Y$

- 2. A formal model of learning

X – a compact metric space (e.g. a domain in $\mathbb{R}^n$)
$Y = \mathbb{R}^k$ – for simplicity $k = 1$
ρ – a probability measure on $Z := X \times Y$

(Least squares) error of $f : X \to Y$

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) = \int_Z (f(x) - y)^2 \, d\rho$$

- 2. A formal model of learning

  X – a compact metric space (e.g. a domain in $\mathbb{R}^n$)
  $Y = \mathbb{R}^k$ – for simplicity $k = 1$
  ρ – a probability measure on $Z := X \times Y$

  (Least squares) error of $f : X \to Y$

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) = \int_Z (f(x) - y)^2 \, d\rho$$

– The local error for input x and output y is $(f(x) - y)^2$.

$\curvearrowright$ $\mathcal{E}(f) =$ average over $X \times Y$ of the local errors

– Problem. Which f minimizes the error $\mathcal{E}(f)$?

- 2. A formal model of learning

  X – a compact metric space (e.g. a domain in $\mathbb{R}^n$)
  $Y = \mathbb{R}^k$ – for simplicity $k = 1$
  ρ – a probability measure on $Z := X \times Y$

  (Least squares) error of $f : X \to Y$

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) = \int_Z (f(x) - y)^2 \, d\rho$$

– The local error for input x and output y is $(f(x) - y)^2$.

$\curvearrowright$ $\mathcal{E}(f)$ = average over $X \times Y$ of the local errors

– Problem. Which f minimizes the error $\mathcal{E}(f)$?

  ρ_X = marginal probability of ρ on X, i.e.

$$\rho_X(A) = \rho(\{(x, y) \in X \times Y : x \in A\}) \quad , \quad A \subset X.$$

  $\rho(\,.\,|x)$ = conditional probability w.r.to x on Y

By Fubini we have for ρ-integrable $g(x,y)$

$$\int_{X\times Y} g(x,y)\,d\rho = \int_X \left(\int_Y g(x,y)\,d\rho(y|x) \right) d\rho_X(x).$$

By Fubini we have for ρ-integrable $g(x, y)$

$$\int_{X\times Y} g(x,y)\,d\rho = \int_X \left(\int_Y g(x,y)\,d\rho(y|x)\right) d\rho_X(x).$$

The regression function $f_\rho : X \to Y$ of f is defined as

$$f_\rho(x) = \int_Y y\,d\rho(y|x)\,.$$

General assumption. f_ρ is bounded

By Fubini we have for ρ-integrable $g(x,y)$

$$\int_{X\times Y} g(x,y)\,d\rho = \int_X \left(\int_Y g(x,y)\,d\rho(y|x)\right) d\rho_X(x).$$

The regression function $f_\rho : X \to Y$ of f is defined as

$$f_\rho(x) = \int_Y y\,d\rho(y|x)\,.$$

General assumption. f_ρ is bounded

Further notation.

$$\sigma^2(x) := \int_Y (y - f(x))^2\,d\rho(y|x)$$

$$\sigma_\rho^2 := \int_X \sigma^2(x)\,d\rho_X = \mathcal{E}(f\rho)\,.$$

σ_ρ measures how well conditioned ρ is.

Proposition. For every $f : X \to Y$ we have

$$\mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 \, d\rho_X + \sigma_\rho^2 \, .$$

Proposition. For every $f : X \to Y$ we have

$$\mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 \, d\rho_X + \sigma_\rho^2 \, .$$

Proof. By definition of f_ρ we have $\int_Y (f_\rho(x) - y) \, d\rho(y|x) = 0$ for all $x \in X$, whence

$$\begin{aligned}
\mathcal{E}(f) &= \int_Z (f(x) - f_\rho(x) + f_\rho(x) - y)^2 \, d\rho \\
&= \int_X (f(x) - f_\rho(x))^2 \, d\rho_X(x) + \underbrace{\int_{X\times Y} (f_\rho(x) - y)^2 \, d\rho}_{=\sigma_\rho^2} \\
&\quad + 2 \cdot \int_X (f(x) - f_\rho(x)) \underbrace{\int_Y (f_\rho(x) - y) \, d\rho(y|x)}_{=0} \, d\rho_X(x) \, .
\end{aligned}$$

Proposition. For every $f : X \to Y$ we have

$$\mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 \, d\rho_X + \sigma_\rho^2 \,.$$

Proof. By definition of f_ρ we have $\int_Y (f_\rho(x) - y) \, d\rho(y|x) = 0$ for all $x \in X$, whence

$$\begin{aligned}
\mathcal{E}(f) &= \int_Z (f(x) - f_\rho(x) + f_\rho(x) - y)^2 \, d\rho \\
&= \int_X (f(x) - f_\rho(x))^2 \, d\rho_X(x) + \underbrace{\int_{X \times Y} (f_\rho(x) - y)^2 \, d\rho}_{=\sigma_\rho^2} \\
&\quad + 2 \cdot \int_X (f(x) - f_\rho(x)) \underbrace{\int_Y (f_\rho(x) - y) \, d\rho(y|x)}_{=0} \, d\rho_X(x) \,.
\end{aligned}$$

$\curvearrowright$ $\mathcal{E}(f) \geq \sigma_\rho^2$ This lower bound for the error depends only on ρ.

Sampling.

– Draw m pairs (x_i, y_i) independently according to ρ.

$\curvearrowright$ sample $\quad \mathbf{z} \in Z^m \quad , \quad \mathbf{z} = ((x_1, y_1), ..., (x_m, y_m))$

Sampling.

- Draw m pairs (x_i, y_i) independently according to ρ.
  $\curvearrowright$ sample $\quad \mathbf{z} \in Z^m \quad , \quad \mathbf{z} = ((x_1, y_1), ..., (x_m, y_m))$
- Empirical mean of a r.v. ξ w.r.to the sample $\mathbf{z} \in Z^m$

$$\mathbb{E}_{\mathbf{z}}\, \xi := \frac{1}{m} \sum_{i=1}^{m} \xi(z_i)^2$$

- Empirical error of f w.r.to the sample $\mathbf{z} \in Z^m$

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

Sampling.

– Draw m pairs (x_i, y_i) independently according to ρ.

$\curvearrowright$ sample $\quad \mathbf{z} \in Z^m \quad , \quad \mathbf{z} = ((x_1, y_1), ..., (x_m, y_m))$

– Empirical mean of a r.v. ξ w.r.to the sample $\mathbf{z} \in Z^m$

$$\mathbb{E}_{\mathbf{z}}\, \xi := \frac{1}{m} \sum_{i=1}^{m} \xi(z_i)^2$$

– Empirical error of f w.r.to the sample $\mathbf{z} \in Z^m$

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

– For $f : X \to Y$ define the function

$$f_Y : X \times Y \to Y \quad \text{by} \quad f_Y(x, y) := f(x) - y\,.$$

$\curvearrowright \quad \mathcal{E}(f) = \mathbb{E}\, f_Y^2 \quad$ and $\quad \mathcal{E}_{\mathbf{z}}(f) = \mathbb{E}_{\mathbf{z}} f_Y^2$

Hypothesis spaces and target functions.

Let $C(X)$ be the Banach space of continuous functions on X, eqquipped with the sup- norm $\quad \|f\|_\infty = \sup_{x \in X} |f(x)|$

Hypothesis spaces and target functions.

Let $C(X)$ be the Banach space of continuous functions on X, eqquipped with the sup- norm $\|f\|_\infty = \sup_{x\in X} |f(x)|$

Hypothesis space – a (typically compact) subset $\mathcal{H}$ of $C(X)$

Target function – any function $f_{\mathcal{H}} \in \mathcal{H}$ that minimizes the error $\mathcal{E}(f)$ over $f \in \mathcal{H}$, i.e. any optimizer of

$$\min_{f\in\mathcal{H}} \int_Z (f(x) - y)^2 \, d\rho \, .$$

Hypothesis spaces and target functions.

Let $C(X)$ be the Banach space of continuous functions on X, eqquipped with the sup- norm $\|f\|_\infty = \sup_{x \in X} |f(x)|$

Hypothesis space – a (typically compact) subset $\mathcal{H}$ of $C(X)$

Target function – any function $f_{\mathcal{H}} \in \mathcal{H}$ that minimizes the error $\mathcal{E}(f)$ over $f \in \mathcal{H}$, i.e. any optimizer of

$$\min_{f \in \mathcal{H}} \int_Z (f(x) - y)^2 \, d\rho \, .$$

We have $\mathcal{E}(f) = \int_X (f - f_\rho)^2 + \sigma_\rho^2 \quad \curvearrowright \quad f_{\mathcal{H}}$ is also an optimizer of

$$\min_{f \in \mathcal{H}} \int_X (f - f_\rho)^2 \, d\rho_X \, .$$

Defect function of $f : X \to Y$ w.r.to a sample $\mathbf{z} \in Z^m$

$$L_{\mathbf{z}}(f) = L_{\rho,\mathbf{z}}(f) = \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)$$

Defect function of $f : X \to Y$ w.r.to a sample $\mathbf{z} \in Z^m$

$$L_{\mathbf{z}}(f) = L_{\rho,\mathbf{z}}(f) = \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)$$

Definition. We say, $f : X \to Y$ is M-bounded, if for some subset $U \subset Z$ with $\rho(U) = 1$ and all $(x, y) \in U$,

$$|f(x) - y| \le M\,.$$

Defect function of $f : X \to Y$ w.r.to a sample $\mathbf{z} \in Z^m$

$$L_{\mathbf{z}}(f) = L_{\rho,\mathbf{z}}(f) = \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)$$

Definition. We say, $f : X \to Y$ is M-bounded, if for some subset $U \subset Z$ with $\rho(U) = 1$ and all $(x, y) \in U$,

$$|f(x) - y| \leq M \,.$$

Proposition. For any two M-bounded functions and all $\mathbf{z} \in U^m$,

$$|L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)| \leq 4M \cdot \|f_1 - f_2\|_\infty \,.$$

Defect function of $f : X \to Y$ w.r.to a sample $\mathbf{z} \in Z^m$

$$L_{\mathbf{z}}(f) = L_{\rho,\mathbf{z}}(f) = \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)$$

Definition. We say, $f : X \to Y$ is M-bounded, if for some subset $U \subset Z$ with $\rho(U) = 1$ and all $(x, y) \in U$,

$$|f(x) - y| \leq M\,.$$

Proposition. For any two M-bounded functions and all $\mathbf{z} \in U^m$,

$$|L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)| \leq 4M \cdot \|f_1 - f_2\|_\infty\,.$$

Proof. From $\quad (f_1(x) - y)^2 - (f_2(x) - y)^2$
$$= \Big((f_1(x) - y) - (f_2(x) - y)\Big)(f_1(x) - f_2(x) \quad \text{we get}$$

$$|\mathcal{E}(f_1) - \mathcal{E}(f_2)| \leq \int_Z \Big(\underbrace{|f_1(x) - y|}_{\leq M} + \underbrace{|f_2(x) - y|}_{\leq M} \Big) \underbrace{|f_1(x) - f_2(x)|}_{\leq \|f_1 - f_2\|_\infty} \, d\rho$$

$$|\mathcal{E}(f_1) - \mathcal{E}(f_2)| \leq \int_Z \Big(\underbrace{|f_1(x) - y|}_{\leq M} + \underbrace{|f_2(x) - y|}_{\leq M} \Big) \underbrace{|f_1(x) - f_2(x)|}_{\leq \|f_1 - f_2\|_\infty} \, d\rho$$

A similar argument applies for $\mathbf{z} \in U^m$,

$$|\mathcal{E}_{\mathbf{z}}(f_1) - \mathcal{E}_{\mathbf{z}}(f_2)| \leq 2M \cdot \|f_1 - f_2\|_\infty$$

and by triangle inequality the proof is finished.

$$|\mathcal{E}(f_1) - \mathcal{E}(f_2)| \leq \int_Z \Big(\underbrace{|f_1(x) - y|}_{\leq M} + \underbrace{|f_2(x) - y|}_{\leq M} \Big) \underbrace{|f_1(x) - f_2(x)|}_{\leq \|f_1 - f_2\|_\infty} \, d\rho$$

A similar argument applies for $\mathbf{z} \in U^m$,

$$|\mathcal{E}_{\mathbf{z}}(f_1) - \mathcal{E}_{\mathbf{z}}(f_2)| \leq 2M \cdot \|f_1 - f_2\|_\infty$$

and by triangle inequality the proof is finished.

Consequences.

1. The error functions $\mathcal{E}, \mathcal{E}_{\mathbf{z}} : \mathcal{H} \to \mathbb{R}$ are continuous.
2. If $\mathcal{H}$ is a compact subset of $C(X)$ such that all $f \in \mathcal{H}$ are M-bounded, then the (not necessarily unique) minimizers $f_{\mathcal{H}}$ and $f_{\mathbf{z}}$ exist.
   If $\mathcal{H}$ is convex and compact, then $f_{\mathcal{H}}$ is unique.)

3. Error analysis and metric entropy

3. Error analysis and metric entropy

Recall $\quad \mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 \, d\rho_X + \sigma_\rho^2.$

3. Error analysis and metric entropy

Recall $\quad \mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 \, d\rho_X + \sigma_\rho^2.$
Taking $f = f_{\mathbf{z}}$ and using similar arguments as in the proof gives

$$\mathcal{E}(f_{\mathbf{z}}) = \underbrace{\int_X (f_{\mathbf{z}}(x) - f_{\mathcal{H}}(x))^2 \, d\rho_X}_{= \text{ sample error } \mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}})}$$

$$+ \underbrace{\int_X (f_{\mathcal{H}}(x) - f_\rho(x))^2 \, d\rho_X + \sigma_\rho^2}_{= \text{ approximation error } \mathcal{E}(f_{\mathcal{H}})}$$

3. Error analysis and metric entropy

Recall $\quad \mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 \, d\rho_X + \sigma_\rho^2.$
Taking $f = f_{\mathbf{z}}$ and using similar arguments as in the proof gives

$$\mathcal{E}(f_{\mathbf{z}}) = \underbrace{\int_X (f_{\mathbf{z}}(x) - f_{\mathcal{H}}(x))^2 \, d\rho_X}_{= \text{ sample error } \mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}})}$$

$$+ \underbrace{\int_X (f_{\mathcal{H}}(x) - f_\rho(x))^2 \, d\rho_X + \sigma_\rho^2}_{= \text{ approximation error } \mathcal{E}(f_{\mathcal{H}})}$$

– The sample error depends on ρ only through the sample $\mathbf{z} \in Z^m$
$\curvearrowright$ bounds will hold only with a certain confidence

– The approximation error depends heavily on ρ through $f\rho$.
$\curvearrowright$ bounds will depend on parameters measuring the behaviour of f_ρ

3. Error analysis and metric entropy

Recall $\mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 \, d\rho_X + \sigma_\rho^2$.
Taking $f = f_{\mathbf{z}}$ and using similar arguments as in the proof gives

$$\mathcal{E}(f_{\mathbf{z}}) = \underbrace{\int_X (f_{\mathbf{z}}(x) - f_{\mathcal{H}}(x))^2 \, d\rho_X}_{= \text{ sample error } \mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}})} + \underbrace{\int_X (f_{\mathcal{H}}(x) - f_\rho(x))^2 \, d\rho_X + \sigma_\rho^2}_{= \text{ approximation error } \mathcal{E}(f_{\mathcal{H}})}$$

– The sample error depends on ρ only through the sample $\mathbf{z} \in Z^m$
$\curvearrowright$ bounds will hold only with a certain confidence

– The approximation error depends heavily on ρ through $f\rho$.
$\curvearrowright$ bounds will depend on parameters measuring the behaviour of f_ρ

Goal. Show that under appropriate assumptions on ρ and $\mathcal{H}$,
$\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}})$ becomes arbitrarily small with high probability as $m \to \infty$.

Now SMALL DEVIATIONS come into play!

Now SMALL DEVIATIONS come into play!

Hoeffding's inequality. Let ξ be a random variable on a probability space Z with mean $\mathbb{E}\xi = \mu$ and $|\xi(z) - \mu| \leq M$ for almost all $z \in Z$. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu \geq \varepsilon\right) \leq \exp\left(-\frac{m\varepsilon^2}{2M^2}\right) .$$

Here $\mathbb{P}$ means the probability of all $z = (z_i) \in Z^m$ satisfying the respective inequality.

Now SMALL DEVIATIONS come into play!

Hoeffding's inequality. Let ξ be a random variable on a probability space Z with mean $\mathbb{E}\xi = \mu$ and $|\xi(z) - \mu| \leq M$ for almost all $z \in Z$. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu \geq \varepsilon\right) \leq \exp\left(-\frac{m\varepsilon^2}{2M^2}\right) .$$

Here $\mathbb{P}$ means the probability of all $z = (z_i) \in Z^m$ satisfying the respective inequality.

Proof. follows from Markov's inequality, Taylor's expansion of the exponential function, and convexity of $\exp(cx)$.

Now SMALL DEVIATIONS come into play!

Hoeffding's inequality. Let ξ be a random variable on a probability space Z with mean $\mathbb{E}\xi = \mu$ and $|\xi(z) - \mu| \leq M$ for almost all $z \in Z$. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu \geq \varepsilon\right) \leq \exp\left(-\frac{m\varepsilon^2}{2M^2}\right).$$

Here $\mathbb{P}$ means the probability of all $z = (z_i) \in Z^m$ satisfying the respective inequality.

Proof. follows from Markov's inequality, Taylor's expansion of the exponential function, and convexity of $\exp(cx)$.

Let $f : X \to Y$ be M-bounded , i.e. $|f(x) - y| \leq M$ almost surely.

For the random variable $\xi = (f(x) - y)^2$ on $Z = X \times Y$ we have

$$\mathbb{E}\xi = 0 \quad \text{and} \quad |\xi| \leq M^2 .$$

We can apply Hoeffding's inequality to ξ, this gives – for any fixed f – a bound on the defect function $L_{\mathbf{z}} = \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)$, where $\mathbf{z} \in Z^m$.

We can apply Hoeffding's inequality to ξ, this gives – for any fixed f – a bound on the defect function $L_{\mathbf{z}} = \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)$, where $\mathbf{z} \in Z^m$.

$$\mathbb{P}\Big(L_{\mathbf{z}}(f) \geq \varepsilon\Big) \leq \exp\left(-\frac{m\varepsilon^2}{2M^4}\right)$$

We can apply Hoeffding's inequality to ξ, this gives – for any fixed f – a bound on the defect function $L_{\mathbf{z}} = \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)$, where $\mathbf{z} \in Z^m$.

$$\mathbb{P}\Big(L_{\mathbf{z}}(f) \geq \varepsilon\Big) \leq \exp\left(-\frac{m\varepsilon^2}{2M^4}\right)$$

At this point METRIC ENTROPY shows up, in the form of covering numbers.

We can apply Hoeffding's inequality to ξ, this gives – for any fixed f – a bound on the defect function $L_{\mathbf{z}} = \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)$, where $\mathbf{z} \in Z^m$.

$$\mathbb{P}\Big(L_{\mathbf{z}}(f) \geq \varepsilon\Big) \leq \exp\left(-\frac{m\varepsilon^2}{2M^4}\right)$$

At this point METRIC ENTROPY shows up, in the form of covering numbers.

Assume that $\mathcal{H} = B_1 \cup \ldots \cup B_\ell$ and consider the events

$$A = \left\{\mathbf{z} \in Z^m : \sup_{f\in\mathcal{H}} L_{\mathbf{z}}(f) \geq \varepsilon\right\} \quad , \quad A_j = \left\{\mathbf{z} \in Z^m : \sup_{f\in B_j} L_{\mathbf{z}}(f) \geq \varepsilon\right\}$$

We can apply Hoeffding's inequality to ξ, this gives – for any fixed f – a bound on the defect function $L_{\mathbf{z}} = \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)$, where $\mathbf{z} \in Z^m$.

$$\mathbb{P}\Big(L_{\mathbf{z}}(f) \geq \varepsilon\Big) \leq \exp\left(-\frac{m\varepsilon^2}{2M^4}\right)$$

At this point METRIC ENTROPY shows up, in the form of covering numbers.

Assume that $\mathcal{H} = B_1 \cup \ldots \cup B_\ell$ and consider the events

$$A = \left\{\mathbf{z} \in Z^m : \sup_{f \in \mathcal{H}} L_{\mathbf{z}}(f) \geq \varepsilon\right\} \quad , \quad A_j = \left\{\mathbf{z} \in Z^m : \sup_{f \in B_j} L_{\mathbf{z}}(f) \geq \varepsilon\right\}$$

Then $A = \bigcup_{j=1}^{\ell} A_j$, whence $\mathbb{P}(A) \leq \sum_{j=1}^{\ell} \mathbb{P}(A_j)$, i.e.

$$\mathbb{P}\Big(\sup_{f \in \mathcal{H}} L_{\mathbf{z}}(f) \geq \varepsilon\Big) \leq \sum_{\ell=1}^{\ell} \mathbb{P}\Big(\sup_{f \in B_j} L_{\mathbf{z}}(f) \geq \varepsilon\Big)$$

Let now $\ell = \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right)$ and choose $f_1, ..., f_\ell$ such that the balls B_j with centers f_j and radius $\frac{\varepsilon}{8M}$ cover $\mathcal{H}$.

let $U \subset Z$ be a subset of full measure such that

$$\sup_{f \in \mathcal{H}} |f(x) - y| \leq M \quad \text{for all } z \in U\,.$$

Let now $\ell = \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right)$ and choose $f_1, ..., f_\ell$ such that the balls B_j with centers f_j and radius $\frac{\varepsilon}{8M}$ cover $\mathcal{H}$.

let $U \subset Z$ be a subset of full measure such that

$$\sup_{f \in \mathcal{H}} |f(x) - y| \leq M \quad \text{for all } z \in U\,.$$

For all $f \in B_j$ and all $z \in U$ we have

$$|L_{\mathbf{z}}(f) - L_{\mathbf{z}}(f_j)| \leq 4M \cdot \|f - f_j\|_\infty \leq 4M \cdot \frac{\varepsilon}{8M} \leq \frac{\varepsilon}{2}\,.$$

Let now $\ell = \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right)$ and choose $f_1, \ldots, f_\ell$ such that the balls B_j with centers f_j and radius $\frac{\varepsilon}{8M}$ cover $\mathcal{H}$.

let $U \subset Z$ be a subset of full measure such that

$$\sup_{f \in \mathcal{H}} |f(x) - y| \leq M \quad \text{for all } z \in U\,.$$

For all $f \in B_j$ and all $z \in U$ we have

$$|L_{\mathbf{z}}(f) - L_{\mathbf{z}}(f_j)| \leq 4M \cdot \|f - f_j\|_\infty \leq 4M \cdot \frac{\varepsilon}{8M} \leq \frac{\varepsilon}{2}\,.$$

Triangle inequality gives: $\sup_{f \in B_j} L_{\mathbf{z}}(f) \geq \varepsilon \Longrightarrow L_{\mathbf{z}}(f_j) \geq \frac{\varepsilon}{2}$

and consequently we obtain from Hoeffding's inequality,

$$\mathbb{P}\Big(\sup_{f \in B_j} L_{\mathbf{z}}(f) \geq \varepsilon\Big) \leq \mathbb{P}\Big(L_{\mathbf{z}}(f_j) \geq \frac{\varepsilon}{2}\Big) \leq \exp\left(-\frac{m\varepsilon^2}{8M^4}\right).$$

Putting everything together we get the following uniform bound for the defect.

Theorem. Let $\mathcal{H}$ be a compact M-bounded subset of $C(X)$. Then, for all $\varepsilon > 0$ and all $m \in N$,

$$\mathbb{P}_{z \in Z^m}\Big(\sup_{f \in \mathcal{H}} L_{\mathbf{z}}(f) \leq \varepsilon\Big) \geq 1 - \mathcal{N}\Big(\mathcal{H}, \frac{\varepsilon}{8M}\Big) \exp\Big(-\frac{m\varepsilon^2}{8M^4}\Big).$$

Putting everything together we get the following uniform bound for the defect.

Theorem. Let $\mathcal{H}$ be a compact M-bounded subset of $C(X)$. Then, for all $\varepsilon > 0$ and all $m \in N$,

$$\mathbb{P}_{z\in Z^m}\Big(\sup_{f\in\mathcal{H}} L_{\mathbf{z}}(f) \leq \varepsilon\Big) \geq 1 - \mathcal{N}\Big(\mathcal{H}, \frac{\varepsilon}{8M}\Big) \exp\Big(- \frac{m\varepsilon^2}{8M^4}\Big).$$

The same technique gives similar bounds for the sample error.

$$\mathbb{P}_{z\in Z^m}\Big(\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \leq \varepsilon\Big) \geq 1 - \Big[\mathcal{N}\Big(\mathcal{H}, \frac{\varepsilon}{16M}\Big) + 1\Big] \exp\Big(- \frac{m\varepsilon^2}{32M^4}\Big).$$

4. Covering numbers of Gaussian RKHSs and small deviations of Gaussian random fields

4. Covering numbers of Gaussian RKHSs and small deviations of Gaussian random fields

- The positive definite Gaussian kernel

$$K(x,y) = \exp(-\sigma^2\|x-y\|_2^2) \quad , \quad x,y \in [0,1]^d \quad , \quad \sigma > 0,$$

generates a RKHS $H_\sigma([0,1]^d)$ which is compactly embedded in $C([0,1]^d)$. In particular, the unit ball in H_σ often serves as hypothesis space $\mathcal{H}$ in learning theory. As shown before, covering numbers are of central importance in the error analyis.

4. Covering numbers of Gaussian RKHSs and small deviations of Gaussian random fields

- The positive definite Gaussian kernel

$$K(x,y) = \exp(-\sigma^2 \|x-y\|_2^2) \quad , \quad x, y \in [0,1]^d \quad , \quad \sigma > 0,$$

generates a RKHS $H_\sigma([0,1]^d)$ which is compactly embedded in $C([0,1]^d)$. In particular, the unit ball in H_σ often serves as hypothesis space $\mathcal{H}$ in learning theory. As shown before, covering numbers are of central importance in the error analyis.

- Kühn (J. Complexity 2011) The covering numbers $\mathcal{N}(\varepsilon)$ of the unit ball of $H_\sigma([0,1]^d)$, considered as a compact subset of $C([0,1]^d)$, behave asymptotically like

$$\log \mathcal{N}(\varepsilon) \sim \frac{\left(\log \frac{1}{\varepsilon}\right)^{d+1}}{\left(\log \log \frac{1}{\varepsilon}\right)^{d}} \qquad \textit{as} \quad \varepsilon \to 0\,.$$

The same is true, if we consider the unit ball as a subset of $L_p([0,1]^d)$, $2 \le p < \infty$.

- Remarks.

  1. This improves earlier results of Ding-Xuan Zhou 2002/2003.

  He showed $$(\log \tfrac{1}{\varepsilon})^{\frac{d}{2}} \preceq \mathcal{N}(\varepsilon) \preceq (\log \tfrac{1}{\varepsilon})^{d+1}$$

  and conjectured that the correct bound is $(\log \frac{1}{\varepsilon})^{\frac{d}{2}+1}$.

  2. Our proof uses an explicit description of an ONB in Gaussian RKHSs, due to Steinwart/Hush/Scovel 2006.

- Remarks.

  1. This improves earlier results of Ding-Xuan Zhou 2002/2003.

  He showed $(\log \frac{1}{\varepsilon})^{\frac{d}{2}} \preceq \mathcal{N}(\varepsilon) \preceq (\log \frac{1}{\varepsilon})^{d+1}$

  and conjectured that the correct bound is $(\log \frac{1}{\varepsilon})^{\frac{d}{2}+1}$.

  2. Our proof uses an explicit description of an ONB in Gaussian RKHSs, due to Steinwart/Hush/Scovel 2006.

- Application to smooth Gaussian processes.

  Let $X = X(t), t \in T$, be a centered Gaussian process with values in a Banach space E (mostly $E = L_2$ or C or L_∞). There is a close connection between small deviation probabilities of X

  $$\mathbb{P}\left(\|X\|_E \le \varepsilon\right) \quad , \quad \varepsilon > 0$$

  and entropy numbers of operators $S : H \to E$ with

  $$\mathbb{E}e^{i\langle X,a\rangle} = e^{-\|S'a\|^2/2} \quad , \quad a \in E' .$$

  (This relation between X and S can also be expressed by the covariance structure of X.) Details of the small deviation – entropy connection have been explained in the talks of Wenbo.

- Example. Let $\sigma > 0$ and $d \in \mathbb{N}$. Consider the centered Gaussian process $X_{\sigma,d} = (X_{\sigma,d}(t))$, $t \in [0,1]^d$ with covariance structure

$$\mathbb{E}\, X_{\sigma,d}(t) X_{\sigma,d}(s) = \exp\left(-\sigma^2 \|t-s\|_2^2\right) \quad , \quad t, s \in [0,1]^d .$$

- Example. Let $\sigma > 0$ and $d \in \mathbb{N}$. Consider the centered Gaussian process $X_{\sigma,d} = (X_{\sigma,d}(t))$, $t \in [0,1]^d$ with covariance structure

$$\mathbb{E}\, X_{\sigma,d}(t) X_{\sigma,d}(s) = \exp\left(-\sigma^2 \|t-s\|_2^2\right) \quad , \quad t, s \in [0,1]^d .$$

- Kühn (J. Complexity 2011)
  The small deviation probabilities under the sup-norm satisfy

$$-\log \mathbb{P}\left(\sup_{t\in[0,1]^d} |X_{\sigma,d}(t)| \le \varepsilon \right) \sim \frac{\left(\log \frac{1}{\varepsilon}\right)^{d+1}}{\left(\log\log \frac{1}{\varepsilon}\right)^{d}} .$$

  The same estimates hold for all L_p-norms with $2 \le p < \infty$.

- Example. Let $\sigma > 0$ and $d \in \mathbb{N}$. Consider the centered Gaussian process $X_{\sigma,d} = (X_{\sigma,d}(t))$, $t \in [0,1]^d$ with covariance structure

$$\mathbb{E}\, X_{\sigma,d}(t) X_{\sigma,d}(s) = \exp\left(-\sigma^2 \|t-s\|_2^2\right) \quad , \quad t,s \in [0,1]^d \,.$$

- Kühn (J. Complexity 2011)
  The small deviation probabilities under the $\sup$-norm satisfy

$$-\log \mathbb{P}\left(\sup_{t \in [0,1]^d} |X_{\sigma,d}(t)| \le \varepsilon \right) \sim \frac{\left(\log \frac{1}{\varepsilon}\right)^{d+1}}{\left(\log\log \frac{1}{\varepsilon}\right)^{d}} \,.$$

  The same estimates hold for all L_p-norms with $2 \le p < \infty$.

THANK YOU FOR YOUR ATTENTION!