Rademacher and Gaussian averages in Learning Theory

Olivier Bousquet
Max Planck Institute
Tübingen

Marne-la-Vallée, 25th March 2003

Outline

- Motivation
- Rademacher and Gaussian averages
- Unit balls in Banach spaces
- Examples

Motivation I

There are relationships between

- Empirical processes
- Probability in Banach spaces
- Geometry of Banach spaces
- Learning theory

Some examples

- Concentration inequalities (cf Lugosi / Massart)
  - Empirical processes
  - Combinatorial parameters (VC entropy, VC dimension)
- Combinatorial parameters (metric entropy/shattering dimension) (cf Mendelson)
- Capacity measures
- Margin/Regularization

The Learning Problem

Formalization

- $(X, Y) \sim P$ pair of random variables, values in $\mathcal{X} \times \mathcal{Y}$, P unknown joint distribution.
- Given n i.i.d. pairs (X_i, Y_i) sampled according to P, find $g : \mathcal{X} \to \mathcal{Y}$ such that $P(g(X) \neq Y)$ is small

More generally, ℓ measures the cost of errors. Minimize

$$L(g) = \mathbb{E}\left[\ell(g(X), Y)\right]$$

Motivation II

Notation: $Pf = \mathbb{E}\left[f(X,Y)\right]$, $P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i)$.

In general,

$$L(g_n) - L_n(g_n) \leq \sup_{f \in \mathcal{F}} (P - P_n) f \,.$$

For algorithms looking for small error functions, with high probability,

$$L(g_n) - L_n(g_n) \leq \sup_{f \in \mathcal{F}, Pf^2 \leq c} (P - P_n) f \,.$$

Expectations of these quantities measure the capacity of the function class.

Motivation III

Regularization algorithms (dual to large margin algorithms)

$$\min_{f \in \mathcal{F}} L_n(f) + \lambda \|f\|$$

Interesting classes have the form

$$\{f \in \mathcal{F} : \|f\| \leq B\}$$

$\Rightarrow$ Geometry of balls in Banach spaces

Rademacher Averages

For bounded functions

$$\sup_{f \in \mathcal{F}} Pf - P_n f$$

can be controlled by the random Rademacher average

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right]$$

or the random Gaussian average

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} g_i f(X_i) \right]$$

(cf Lugosi's talk)

Modulus of Continuity

With more care (and Talagrand's inequality), one can get

$$Pf - P_n f \leq K^{-1} P f^2 + cK\mathbb{E}\left[\sup_{f\in\mathcal{F}, Pf^2 \leq r} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right] + \cdots .$$

for some r.

- If Pf^2 is related to Pf then one can get a better bound from this
- What is the right value of r ?

$$r = \mathbb{E}\left[\sup_{f\in\mathcal{F}, Pf^2 \leq r} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right]$$

  Call this value r^* capacity radius ("fixed point of the modulus of continuity")

- Idea goes back to Massart (2000), Koltchinskii and Panchenko (2001). This version Bartlett, B. and Mendelson (2002)

Empirical version

$$Pf - P_n f \leq K^{-1} P f^2 + cK \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}, P_n f^2 \leq r} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \right] + \cdots .$$

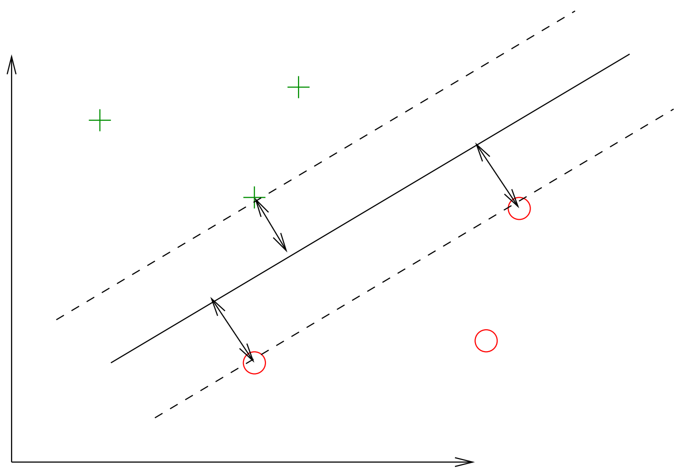
with r satisfying

$$r = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}, P_n f^2 \leq r} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \right]$$

Call this value r_n^* empirical capacity radius

Bartlett, B. and Mendelson, to appear.

Margin and Regularization



- Normalize the weight vector w such that $w.x = 1$ for closest points.
- Margin proportional to $1/\|w\|$
- Give linear penalty to errors

SVM algorithm equivalent to

$$\min_w \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i w.X_i)_+ + \lambda \|w\|^2 \ ,$$

Kernel Spaces

The SVM algorithm does this after mapping the data to a high dimensional euclidean space (cf Ben-David's talk)

Actually it solves

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ + \lambda \left\| f \right\|^2 ,$$

in a reproducing kernel Hilbert space $\mathcal{H}$ generated by $k(x, x')$ ($\mathcal{H} = \mathrm{span}\{k(x, .) : x \in \mathcal{X}\}$).

Equivalent problem

$$\min_{\|f\| \leq B} L_n(f)$$

$\Rightarrow$ estimate the capacity of balls in the RKHS $\mathcal{H}$.

Duality of Rademacher Averages

Rademacher average

$$\mathbb{E}\left[\sup_{f:\|f\|\leq 1} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i)\right]$$

Reproducing property

$$f(X_i) = \langle f, k(X_i, \cdot)\rangle_{\mathcal{H}}$$

By duality

$$\mathbb{E}\left[\sup_{f:\|f\|\leq 1} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i)\right] = \frac{1}{n}\mathbb{E}\left[\left\|\sum_{i=1}^{n} \sigma_i k(X_i, \cdot)\right\|_{\mathcal{H}}\right]$$

$\Rightarrow$ This is a general phenomenon for regularization in a Banach space

Computing Capacity Measures

Some notation

$$R_n(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \right]$$

$$\phi_n(\mathcal{F}, r) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F} : P_n f^2 \leq r} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \right]$$

$$r_n^*(\mathcal{F}) = \phi_n(\mathcal{F}, r_n^*(\mathcal{F}))$$

Convex Hull

- Motivation: Boosting type algorithms
  Choose a base class of $\{-1, 1\}$ functions, make linear combinations and penalize by the sum of the weights

- Global
  $$R_n(\operatorname{conv}\mathcal{F}) = R_n(\mathcal{F})$$
  However entropy can be much larger

Convex Hull

- Local

$$\phi_n(\mathrm{conv}(\mathcal{F}), r) \leq \inf_{\epsilon>0} \left(2\phi_n(\mathcal{F}, \epsilon^2) + c\sqrt{rN(\mathcal{F}, \epsilon)/n}\right) ,$$

Proof idea: approximate convex hull by linear subspace (span of an ϵ-net)

$$r_n^*(\mathrm{conv}\,\mathcal{F}) \leq \inf_{\epsilon>0} \frac{c}{n} \exp\left(Knr_n^*(\mathcal{F}) \log^2 \frac{1}{\epsilon}\right) + 4\epsilon\sqrt{r_n^*(\mathcal{F})} .$$

- For VC classes $r_n^*(\mathcal{F}) = O(d/n)$, and

$$r_n^*(\mathrm{conv}\,\mathcal{F}) = O(n^{-\frac{1}{2}\frac{\alpha d+1}{\alpha d+2}})$$

for some constant $\alpha \geq 1$ (ideally $\alpha = 1$), with log factors

Reproducing Kernel Hilbert Space

- Motivation: kernel algorithms (SVM)

$$\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$$

- Global

$$R_n(\mathcal{F}) = \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i k(X_i, \cdot) \right\| \right] \leq \frac{1}{n} \sqrt{\sum_{i=1}^n k(X_i, X_i)}$$

- Gram matrix

$$K_{i,j} = k(X_i, X_j)$$

K positive semidefinite, $\sum k(X_i, X_i) = \operatorname{tr} K = \sum \lambda_i$

Reproducing Kernel Hilbert Space

- Local

$$\phi_n(\mathcal{F}, r) \leq \frac{c}{\sqrt{n}} \inf_{d \in \mathbb{N}} \left(\sqrt{rd} + \sqrt{\sum_{j>d} \lambda_j / n} \right)$$

Proof idea: approximation by a linear subspace (span of main eigenvectors)

- Radius

$$r_n^* \leq \frac{c}{n} \inf_{d \in \mathbb{N}} \left(d + \sqrt{\sum_{j>d} \lambda_j} \right)$$

$d = 0$ gives the trace bound
Exponential decay (e.g. $\lambda_i = ne^{-i}$) gives $1/n$ bound instead of $1/\sqrt{n}$.

Union of RKHS

- Motivation: automatic choice of the kernel

$$\mathcal{F} = \cup_{k \in \mathcal{K}} \{ f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1 \}$$

- Rademacher averages

$$R_n(\mathcal{F}) = \frac{1}{n} \mathbb{E} \left[\sup_{K \in \mathcal{K}} \sqrt{\sigma^t K \sigma} \right] \leq \frac{1}{n} \sqrt{\mathbb{E} \left[\sup_{K \in \mathcal{K}} \sigma^t K \sigma \right]}$$

$\Rightarrow$ Rademacher chaos

Union of RKHS

For one matrix

$$R_n(K) \le \frac{1}{n}\sqrt{\operatorname{tr} K}$$

Interesting classes of positive semidefinite matrices

- Convex hull

$$R_n(\mathcal{F}) \le \frac{c}{n}\sqrt{\log N \ \max \operatorname{tr} K}$$

- Quadratic hull

$$R_n(\mathcal{F}) \le \frac{c}{n}\left(\sum_{j=1}^{N} \operatorname{tr}^2 K_j + \left\|\bar{K}_j\right\|_2^2\right)^{1/4}$$

- Spectral classes (commuting matrices)

$$R_n(\mathcal{F}) \le \frac{c}{n}\sqrt{\log n \ \max \operatorname{tr} K}$$

Lipschitz Spaces

- Motivation: regularize by the Lipschitz norm of the functions

$$\min L_n(f) + \lambda \|f\|_L$$

- Use duality
  Predual = Arens-Eells space, functions with finite support with norm

$$\|f\| = \inf\{\sum |a_i| d(x_i, y_i) : f = \sum a_i(\mathbb{1}_{x_i} - \mathbb{1}_{y_i})\}$$

- Relates to matching theorems/transportation (Ajtai, Komlos, Tusnady) (Talagrand)

Lipschitz Spaces

For a Lipschitz ball

$$\mathcal{F} = \{f : \|f\|_L \leq 1\}$$

we have

- In $\mathbb{R}^d$, for $d \geq 3$,

$$R_n(\mathcal{F}) \leq n^{-1/d}$$

- In $\mathbb{R}^2$

$$R_n(\mathcal{F}) \leq \sqrt{\frac{\log n}{n}}$$

Proof idea: use majorizing measures (Talagrand) for $d = 2$ and a modification of Dudley's entropy bound for $d > 2$

$$R_n(\mathcal{F}) \leq \epsilon + \int_{\epsilon}^{\infty} H^{1/2}(\mathcal{F}, u)du$$

Entropy estimates of Lipschitz balls (e.g. Kolmogorov and Tihomirov)

Embedding of a Metric Space

- Motivation: large margin classification in metric spaces
- Isometric embedding into $C_b(\mathcal{X})$

$$x \mapsto \Phi_x := d(x, \cdot) - d(x_0, \cdot)$$

- The span of $\{\Phi_x : x \in \mathcal{X}\}$ can be completed into a Banach space with the supremum norm.
- One can define large margin hyperplanes and consider the unit ball $\mathcal{F}$ of the dual
- Result: geometry of $\mathcal{F}$ = geometry of $\mathcal{X}$

$$R_n(\mathcal{F}) = R_n(\mathcal{X})$$

where points in x are seen as evaluation functions defined on $\{\Phi_{X_i}\}$

Open problems

1. Improve convex hull estimates

2. Obtain capacity radius bounds for chaoses

3. Investigate interesting classes of matrices (with nice geometry)

4. Obtain capacity radius bounds for Lipschitz balls