Fat-Shattering Dimension and Majorizing Measures: two ways to quantify complexity

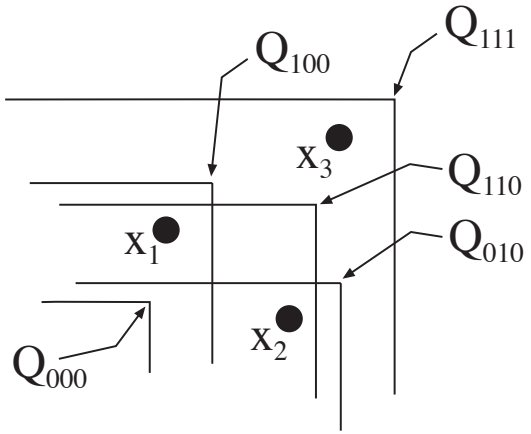
David Pollard
Statistics Department
Yale University
http://www.stat.yale.edu/~pollard

Talk given at the
Workshop on *Simplicity and Likelihood*
Yale University
November 2009
http://cowles.econ.yale.edu/conferences/simplicity/

VC classes of sets

$\mathcal{Q}$ = all quadrants $(-\infty, a] \times (-\infty, b]$ in $\mathbb{R}^2$

- How many patterns picked out from $\{x_1, x_2, x_3\}$?



$$V = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

- For all x_1, x_2, x_3 cannot get all 2^3 patterns: V has < 8 rows.
  For some x_1, x_2 can get all 2^2 patterns.

- VCdim$(\mathcal{Q}) = 2$

- [packing numbers for $\mathcal{Q}$ as subset of $\mathcal{L}^1(P)$, for each probability measure P on $\mathbb{R}^2$]
  If $Q_1, \ldots, Q_N \in \mathcal{Q}$ and

$$P(Q_i \Delta Q_{i'}) > \epsilon \qquad \text{for } 1 \le i < i' \le N$$

  then

$$N \le \text{constant} \times (1/\epsilon)^2$$

- Hard to prove with exponent 2 (I think).
  Easy to prove with exponent 4.

- Haussler (1995) got constant $\times (1/\epsilon)^{\text{VCdim}}$ for general VC classes of sets.

- Fat shattering?

ϵ-shattering: Kearns and Schapire (1994, §6)
=
fat-shattering: Bartlett, Long, and Williamson (1996);
Anthony and Bartlett (2002, §11.3)
≈
(not) stable: Talagrand (1987a);
shatter at levels α, β: Talagrand (1996a);
??? Talagrand (1984);
Fremlin??

- $\mathcal{F}$ = a set of real-valued functions on a set $\mathcal{X}$
  $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$
  levels: $\alpha = (\alpha_1, \ldots, \alpha_n)$ and $\beta = (\beta_1, \ldots, \beta_n)$ with $\beta_j \geq \alpha_j$

- $\mathcal{F}$ picks out pattern $\mathbb{K} \subseteq \{1, 2, \ldots, n\}$ at levels α and β means:
  There exists $f_{\mathbb{K}} \in \mathcal{F}$ such that

$$f_{\mathbb{K}}(x_j) \begin{cases} \geq \beta_j & \text{if } j \in \mathbb{K} \\ \leq \alpha_j & \text{if } j \in \mathbb{K}^c \end{cases}$$

- $\mathcal{S}(x, \alpha, \beta, \mathcal{F}) =$ number of distinct patterns picked out $\leq 2^n$

- $\mathcal{F}$ shatters x at levels α and β means: $\mathcal{S}(x, \alpha, \beta, \mathcal{F}) = 2^n$

- $\mathrm{sdim}(\epsilon, \mathcal{F})$ is largest n such that $\mathcal{F}$ shatters some x in $\mathcal{X}^n$ at some levels α and β with $\beta_j \geq \alpha_j + 2\epsilon$ for all j

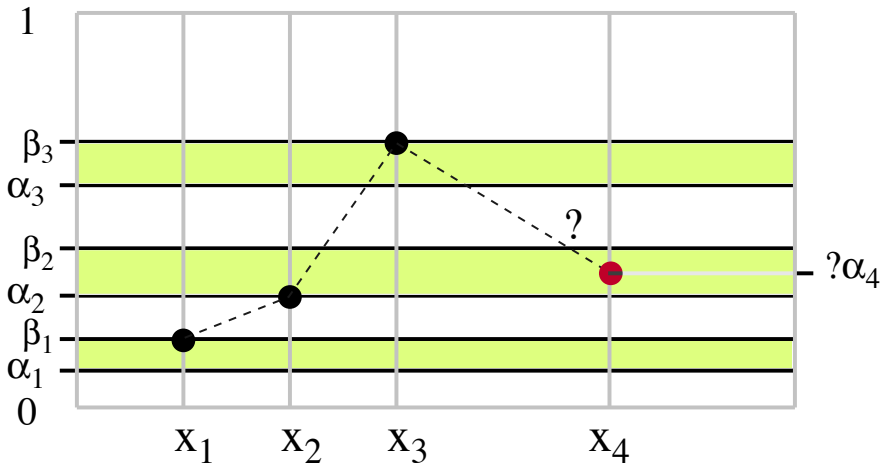
Example:

$\mathcal{F} =$ all increasing functions f on $\mathbb{R}$ with $0 \le f \le 1$

- Why must we have

$$0 \le \alpha_1 \le \cdots \le \alpha_3 + 2\epsilon \le \beta_3 \le \alpha_4 \ldots \quad ?$$

- Why is $\mathrm{sdim}(\epsilon, \mathcal{F}) \le (2\epsilon)^{-1}$?



- Suppose $0 \le f \le 1$ for all f in $\mathcal{F}$
  and $X_1, X_2, \cdots \sim$ iid P

- $$\Delta_n := \sup_{\mathcal{F}} |n^{-1} \sum_{i \le n} f(X_i) - \int f \, dP| = \sup_{\mathcal{F}} |P_n f - P f|$$

- Talagrand (1987a) and Talagrand (1996a):
  failure of $\Delta_n \to 0$ a.s. (uniform SLLN)
  iff
  (roughly) $\exists$ subset A with $PA > 0$ for which (almost) every sample from (nonatomic) $P(\cdot \mid A)$ can be shattered at some fixed levels α and β (with $\alpha_j \equiv \alpha_1$ and $\beta_j \equiv \beta_1$)

Mendelson and Vershynin (2003)

- If $0 \le f \le 1$ for all f in $\mathcal{F}$
  and P is a probability measure and $\alpha \ge 1$
  and $f_1, \ldots, f_N$ in $\mathcal{F}$ with

$$\int |f_i - f_{i'}|^\alpha dP > (c_\alpha \epsilon)^\alpha \qquad \text{for } 1 \le i < i' \le N$$

  (a packing assumption) then

$$N \le (C_\alpha/\epsilon)^{6\,\mathrm{sdim}(\epsilon,\mathcal{F})}$$

  for known constants c_α and C_α.

- Good consequences for bounding Δ_n
  when $X_1, X_2, \cdots \sim$ iid P if

$$\int_0^1 \sqrt{\mathrm{sdim}(\epsilon, \mathcal{F}) \log(1/\epsilon)}\, d\epsilon < \infty$$

M&V method:

- For some sample $x_1, \ldots, x_m$ from P,
  with $m =$ something involving $\log N$ and ϵ,
  discretize:

$$V[i,j] := \lfloor f_i(x_j)/\epsilon \rfloor$$

  to get an $N \times m$ matrix V with entries in $S_p = \{0, 1, \ldots, p\}$
  for $p = \lfloor 1/\epsilon \rfloor$.

- For a good realization $x_1, \ldots, x_m$ get

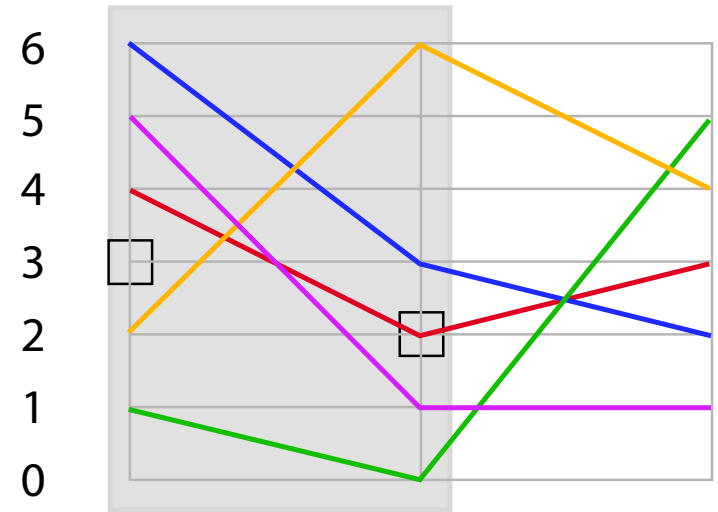
  ♡ $\quad m^{-1} \sum_{j \le m} |V[i,j] - V[i',j]|^\alpha > C'_\alpha \qquad$ for $1 \le i < i' \le N$

  for some magic constant C'_α

- Key question:
  For how many distinct $(\mathbb{J}, z)$ with $\mathbb{J} \subseteq S_p$ and $z \in \mathbb{Z}^{\mathbb{J}}$
  can V "surround" $(\mathbb{J}, z)$ in the sense:
  for each $\mathbb{K} \subseteq \mathbb{J}$ there is an $i_{\mathbb{K}}$ for which

$$V[i_{\mathbb{K}}, j] \begin{cases} \geq z_j + 1 & \text{if } j \in \mathbb{K} \\ \leq z_j - 1 & \text{if } j \in \mathbb{J} \backslash \mathbb{K} \end{cases} \quad ?$$

$$V = \begin{bmatrix} 6 & 3 & 2 \\ 5 & 1 & 1 \\ 4 & 2 & 3 \\ 1 & 0 & 5 \\ 2 & 6 & 4 \end{bmatrix}$$



$\mathbb{J} = (1,2) \quad z = (3,2)$

- Answer:
  If $\heartsuit$ holds then

$$\#\text{ distinct } (\mathbb{J}, z) \text{ surrounded by } V \text{ is } \geq \sqrt{N} - 1.$$

Majorizing measures

- Long and glorious history:
  Fernique (1975),
  . . . ,
  Talagrand (1987b), Talagrand (1990),
  Ledoux and Talagrand (1991, Chapter 11),
  Talagrand (1996b), Talagrand (2001), Talagrand (2005),
  Kwapień and Rosiński (2004),
  Bednorz (2006), Bednorz (2007), . . .

  (List woefully incomplete)

- $\Psi : \mathbb{R}^+ \to \mathbb{R}^+$ convex, increasing, with $\Psi(0) = 0$
  eg. $\Psi_{gaus}(x) := \exp(x^2/2) - 1$ useful for Gaussian processes

- Stochastic process $\{Z_t : t \in T\}$ indexed by metric space (T, d) with $\|Z_s - Z_t\|_\Psi \leq d(s,t)$, that is,

$$\mathbb{P}\Psi\left(\frac{|Z_t - Z_s|}{d(s,t)}\right) \leq 1 \qquad \text{for all } s \neq t$$

- Probability measure μ on T is a majorizing measure if

$$\sup_{t \in T} \int_0^{\text{diam}(T)} \Psi^{-1}\left(\frac{1}{\mu B(t,r)}\right) dr < \infty$$

  where $B(t,r) =$ closed ball of radius r around t.

- (Talagrand 1987b) for Gaussian process

$$\mathbb{P} \sup_{t \in T} Z_t < \infty$$

  if and only if there exists a majorizing measure (using Ψ_{gaus})

- Actually, Talagrand gave explicit bounds.

One way to build a MM: (under mild conditions on Ψ)

- T_k is a maximal set of points separated by at least $\text{diam}(T)/2^k$, for $k = 1, 2, \ldots$
- $N_k := \#T_k$
- $\mu_k =$ uniform distribution on T_k, mass $1/N_k$ at each point
- $\mu = \sum_{k=1}^{\infty} 2^{-k}\mu_k$ is a MM if $\sum_{k=1}^{\infty} 2^{-k}\Psi^{-1}(N_k) < \infty$
- Other kinds of MM's exist and are useful

- Use MM to construct nested partitions for "chaining argument" (Talagrand 2005)
- Use MM as a local smoothing operator [Kwapień and Rosiński (2004), Bednorz (2006), Bednorz (2007)]; get bounds on supremum (and increments) of stochastic process
- Traditional chaining arguments seem to require "local complexity" the same everywhere in T
- MM gives control where "local complexity" differs from one part of T to another

Wild analogies and speculation

- Arguments for minimax rates of convergence of estimators often use Bayes argument with uniform prior on a maximal set of points separated by at least ϵ_n
- Some minimax arguments have a suggestive similarity to MM arguments
- MM like a (possibly nonuniform) Bayes prior?

References

Anthony, M. and P. Bartlett (2002). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

Bartlett, P. L., P. M. Long, and R. C. Williamson (1996). Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences 52*, 434–452.

Bednorz, W. (2006). A theorem on majorizing measures. *Annals of Probability 34*, 1771–1781.

Bednorz, W. (2007). Hölder continuity of random processes. *Journal of Theoretical Probability 20*, 917–934.

Fernique, X. (1975). Regularité des trajectoires des fonctions aléatoires gaussiennes. *Springer Lecture Notes in Mathematics 480*, 1–97.

Haussler, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory 69*, 217–232.

Kearns, M. J. and R. E. Schapire (1994). Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences 48*, 464–497.

Kwapień, S. and J. Rosiński (2004). Sample Hölder continuity of stochastic processes and majorizing measures. *Progress in Probability 58*, 155–163.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer.

Mendelson, S. and R. Vershynin (2003). Entropy and the combinatorial dimension. *Inventiones Mathematicae 152*, 37–55.

Talagrand, M. (1984). *Pettis Integral and Measure Theory*, Volume 307 of *Memoirs AMS*. American Mathematical Society.

Talagrand, M. (1987a). The Glivenko-Cantelli problem. *Annals of Probability 15*(3), 837–870.

Talagrand, M. (1987b). Regularity of gaussian processes. *Acta Mathematica 159*, 99–149.

Talagrand, M. (1990). Sample boundedness of stochastic processes under increment conditions. *Annals of Probability 18*(1), 1–49.

Talagrand, M. (1996a). The Glivenko-Cantelli problem, ten years later. *Journal of Theoretical Probability 9*(2), 371–384.

Talagrand, M. (1996b). Majorizing measures: The generic chaining (in special invited paper). *Annals of Probability 24*, 1049–1103.

Talagrand, M. (2001). Majorizing measures without measures. *The Annals of Probability 29*(1), 411–417.

Talagrand, M. (2005). *The Generic Chaining: Upper and lower bounds of stochastic processes*. Springer-Verlag.