

Prediction Algorithms: Complexity, Concentration, and Convexity

Peter Bartlett

Division of Computer Science and Department of Statistics
UC Berkeley

slides at <http://www.stat.berkeley.edu/~bartlett/talks>

1

Handwritten Character Recognition

X = grey-scale image of a character

$Y \in \{ 'a,' 'b,' 'c,' \dots, 'z,' '0,' '1,' \dots, '9' \}$

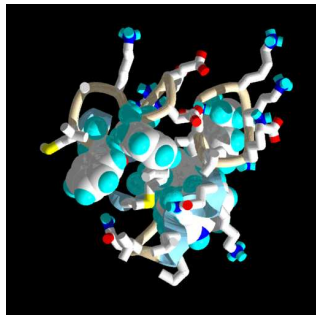


2

Predicting Protein Function

X = amino acid sequence, protein family alignments, protein-protein interactions, gene expression profiles, ...

Y = protein function (metabolism, energy, cell cycle, transcription, ...)



3

Prediction Problems

- Data: i.i.d. $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$ from $\mathcal{X} \times \mathcal{Y}$.
- Loss function: $\ell : \mathcal{Y}^2 \rightarrow \mathbb{R}^+, \ell(\hat{y}, y) = \text{cost of mistake}$.
- Use data $(X_1, Y_1), \dots, (X_n, Y_n)$ to choose $f : \mathcal{X} \rightarrow \mathcal{Y}$ with small **risk**,

$$R(f) = \mathbf{E}\ell(f(X), Y).$$

Often choose f from a fixed class \mathcal{F} .

Three key issues:

- Approximation error
- Estimation error
- Computation

4

Prediction Algorithms

1. Kernel methods:

Choose f from a reproducing kernel Hilbert space, \mathcal{F} .

2. Boosting methods:

Choose f from the span of a fixed dictionary: $\mathcal{F} = \text{span}(\mathcal{G})$.

Both families of prediction algorithms use a function class \mathcal{F} that is

- large (typically infinite-dimensional),
- convex,
- linearly parameterized.

5

Outline

• Prediction algorithms:

- Kernel methods
- Boosting methods

• Estimation error:

- Complexity of function class
- Convexity

• Convexity and approximation error

• Dependent data.

6

Prediction Algorithms: Kernel Methods

Use a subset \mathcal{F} of a *reproducing kernel Hilbert space* \mathcal{H} , with norm $\|\cdot\|_{\mathcal{H}}$.

Choose $f \in \mathcal{F}$ to minimize

$$\underbrace{\mathbf{E}_n \ell(f(X), Y)}_{\text{empirical risk}} + \underbrace{\lambda \|f\|_{\mathcal{H}}^2}_{\text{regularization}}$$

where $\mathbf{E}_n \ell(f(X), Y) = \frac{1}{n} \sum_{i=1}^n \ell(f(X_i), Y_i)$.

The regularization term encourages “simple” functions.

7

Prediction Algorithms: Kernel Methods

Example: Linear kernel, \mathcal{H} = linear functions on \mathbb{R}^d .

Then $f \in \mathcal{H}$ has $f(x) = \theta^T x$, and $\|f\|_{\mathcal{H}} = \|\theta\|_2$.

Choose $\theta \in \mathbb{R}^d$ to minimize

$$\underbrace{\mathbf{E}_n \ell(\theta^T X, Y)}_{\text{empirical risk}} + \underbrace{\lambda \|\theta\|^2}_{\text{regularization}}$$

8

Prediction Algorithms: Kernel Methods

Recall: A reproducing kernel Hilbert space (RKHS) of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ is a Hilbert space \mathcal{H} with a “reproducing kernel,” $k : \mathcal{X}^2 \rightarrow \mathbb{R}$, that is, a symmetric function satisfying

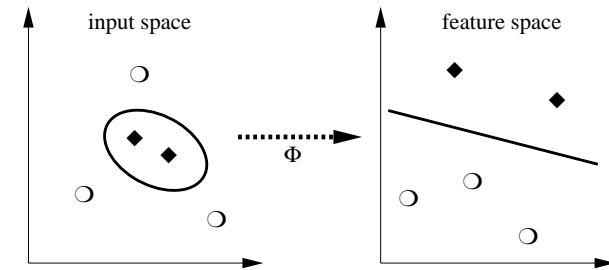
1. $k(x, \cdot) \in \mathcal{H}$,
2. $\langle k(x, \cdot), f \rangle = f(x)$ (the reproducing property)
3. $\overline{\text{span}\{k(x, \cdot) : x \in \mathcal{X}\}} = \mathcal{H}$ (k spans \mathcal{H})

9

Prediction Algorithms: Kernel Methods

Any kernel k corresponds to an inner product in a **feature** space,

$$k(a, b) = \langle \Phi(a), \Phi(b) \rangle.$$



picture: (Schölkopf and Smola, 2002)

10

Prediction Algorithms: Kernel Methods

Example: Linear kernel, $k(a, b) = a^T b$.

Example: Gaussian kernel, $k(a, b) = \exp\left(-\frac{\|a - b\|^2}{2\sigma^2}\right)$.

The corresponding RKHS is infinite-dimensional.

We can think of it as the set of linear combinations of Gaussian bump functions.

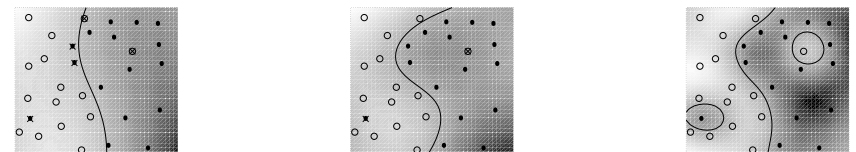
11

Prediction Algorithms: Kernel Methods

Choose $f \in \mathcal{F}$ to minimize

$$\underbrace{\mathbf{E}_n \ell(f(X), Y)}_{\text{empirical risk}} + \underbrace{\lambda \|f\|_{\mathcal{H}}^2}_{\text{regularization}}$$

The regularization term encourages “simple” functions:



picture: (Schölkopf and Smola, 2002)

12

Kernel Methods: Finite representation

Theorem: For any function $L : \mathbb{R}^n \rightarrow \mathbb{R}$ e.g., $(f(X_i)) \mapsto \mathbf{E}_n \ell(f(X), Y)$
and any increasing function $\Omega : \mathbb{R} \rightarrow \mathbb{R}$ e.g., id
if $f^* \in \mathcal{H}$ minimizes

$$J(f) = L(f(x_1), \dots, f(x_n)) + \Omega(\|f\|_{\mathcal{H}}^2)$$

then it can be written

$$f^*(\cdot) = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$$

for some $\alpha_1, \dots, \alpha_n \in \mathbb{R}$.

(Kimeldorf and Wahba, 1971)

13

Kernel Methods: Finite representation

Because of this result, we can restrict our attention to the **Gram matrix**,
 $K \in \mathbb{R}^{n \times n}$,

$$K_{ij} = k(x_i, x_j).$$

Indeed, for $f \in \text{span}\{k(x_1, \cdot), \dots, k(x_n, \cdot)\}$, we have

$$(f(x_1), \dots, f(x_n))^T = K\alpha$$

$$\|f\|_{\mathcal{H}}^2 = \alpha^T K\alpha,$$

and we can write

$$J(\alpha) = L(K\alpha) + \Omega(\alpha^T K\alpha).$$

14

Kernel Methods: Finite representation

- We need only compute $k(x_i, x_j)$, we do not need to compute the transformations $\Phi(x_i)$.
- We can have a rich (even infinite-dimensional) \mathcal{H} , but we always have a finite-dimensional optimization problem.
- Often (if ℓ is suitably chosen), the solution vector α is sparse.

15

Kernel Methods: Applications

Information retrieval: discriminating documents

Kernel, evaluated on documents 1 and 2 (d_1, d_2):

$$k(d_1, d_2) = \sum_w \Phi_w(d_1)\Phi_w(d_2),$$

where the sum is over all words w , and $\Phi_w(d_1)$ is the number of times word w appears in document d_1 (appropriately weighted).

In a variety of document categorization tasks, support vector machines with this kernel gave significant improvements over naive Bayes, k -nearest neighbour, Rocchio (a leading IR algorithm), and decision trees. (Joachims, 1998)

16

Kernel Methods: Applications

Bioinformatics: predicting protein function

Kernel, evaluated on proteins 1 and 2 (p_1, p_2):

$$k(p_1, p_2) = \sum_i \mu_i k_i(p_1, p_2),$$

where the μ_i are convex coefficients and each k_i is a kernel defined using different information about the protein (protein family alignments, protein-protein interactions, genetic interactions, co-participation in protein complexes, gene expression profiles, sequence comparison with known genes).

In predicting the function of yeast proteins, this approach gave significant improvements over standard methods. (Lanckriet et al, 2003)

17

Outline

- Prediction algorithms:
 - Kernel methods
 - Boosting methods
- Estimation error:
 - Complexity of function class
 - Convexity
- Convexity and approximation error
- Dependent data.

18

Prediction Algorithms: Boosting

\mathcal{F} = linear combinations of functions from a fixed dictionary.

History: Schapire (1990) showed that the performance of a *weak* learning algorithm for classification (i.e., slightly better than chance) can be boosted by forming a **committee** of such classifiers.

Subsequently, a particular boosting algorithm—AdaBoost (Freund and Schapire, 1996)—was found to work well in practice with, e.g., decision trees.

19

A Generic Boosting Algorithm

Greedy convex optimization over linear combinations

set $f_0 = 0$

for $t = 1, \dots, T$

choose $\alpha_t \in \mathbb{R}$ and $g_t \in \mathcal{G}$ to minimize

$$A(f_{t-1} + \alpha_t g_t).$$

set $f_t = f_{t-1} + \alpha_t g_t$

return f_T .

(A is a convex objective function).

20

A Generic Boosting Algorithm

Greedy convex optimization over **convex** combinations

set $f_0 = 0$

for $t = 1, \dots, T$

choose $\alpha_t \in [0, 1]$ and $g_t \in \mathcal{G}$ to minimize

$$A((1 - \alpha_t)f_{t-1} + \alpha_t g_t).$$

set $f_t = (1 - \alpha_t)f_{t-1} + \alpha_t g_t$

return f_T .

(A is a convex objective function).

21

AdaBoost

For two-class classification ($Y \in \{\pm 1\}$), AdaBoost uses

$$\mathcal{F} = \text{span}(\mathcal{G}),$$

$$A(f) = \mathbf{E}_n \exp(-Y f(X)).$$

Regularization:

- small number of steps, or
- constrain size of coefficient vector, $\|\alpha\|_1$.

22

Boosting Methods: Applications

Natural language processing (parsing, part of speech tagging):

In analysing Wall Street Journal and web text, boosting gives significant accuracy improvements over previous state-of-the-art (maximum entropy) methods. (Collins, 2000, 2002)

AT&T spoken-dialogue system:

Use text from a speech recognition system to identify the type of request, so the call can be appropriately directed. (\mathcal{G} = indicator functions of keywords.) (Rochery et al, 2002)

23

Outline

- Prediction algorithms:
 - Kernel methods
 - Boosting methods
- Estimation error:
 - Complexity of function class
 - Convexity
- Convexity and approximation error
- Dependent data.

24

Estimation Error, Concentration and Complexity

Aim: Use $(X_1, Y_1), \dots, (X_n, Y_n)$ to choose $f_n \in \mathcal{F}$ to minimize risk, $\mathbf{E}\ell(f_n(X), Y)$.

Estimation error:

$$\mathbf{E}\ell(f_n(X), Y) - \underbrace{\inf_{f \in \mathcal{F}} \mathbf{E}\ell(f(X), Y)}_{\text{Best possible in } \mathcal{F}}$$

Depends on:

- Sample size, n
- Complexity of \mathcal{F}
- Algorithm (in particular, regularization)

25

Estimation Error

Bounds on estimation error

- show how to **measure complexity**, and hence
- motivate **regularization** schemes

For example, if we know that, uniformly over $f \in \mathcal{F}$,

$$\mathbf{E}\ell(f(X), Y) \leq \mathbf{E}_n \ell(f(X), Y) + c_n \Omega(f),$$

then minimizing $\mathbf{E}_n \ell(f(X), Y) + c_n \Omega(f)$ corresponds to obtaining the best upper bound.

Can give performance guarantees ('oracle inequalities') for *complexity regularization* schemes of this kind.

26

Rademacher Averages

Definition: For sample size n , the **Rademacher average** of \mathcal{F} is

$$R_n(\mathcal{F}) = \mathbf{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i),$$

where $\epsilon_1, \dots, \epsilon_n$ are independent uniform $\{\pm 1\}$ (Rademacher) random variables.

Intuition: Measures how well some $f \in \mathcal{F}$ can match the random direction $(\epsilon_1, \dots, \epsilon_n) \in \{\pm 1\}^n$.

27

Estimation Error and Rademacher Averages

Theorem: For $\mathcal{F} \subseteq [0, 1]^{\mathcal{X}}$ and for any P on \mathcal{X} , with probability at least $1 - e^{-x}$ over X_1, \dots, X_n ,

$$\sup_{f \in \mathcal{F}} (\mathbf{E}f - \mathbf{E}_n f) \leq 2R_n(\mathcal{F}) + \sqrt{\frac{x}{2n}}.$$

(Proof: Concentration of $\sup_{f \in \mathcal{F}} (\mathbf{E}f - \mathbf{E}_n f)$ about its expectation; symmetrization.)

28

Estimation Error and Rademacher Averages

We are usually interested in **loss classes**,

$$\ell_{\mathcal{F}} = \{\ell_f : f \in \mathcal{F}\} \quad \text{with} \quad \ell_f(x, y) = \ell(f(x), y).$$

Then $\mathbf{E}\ell_f$ is the *risk*, and $\mathbf{E}_n\ell_f$ is the *empirical risk*.

For an Lipschitz loss ℓ (with Lipschitz constant L), we have

$$R_n(\ell_{\mathcal{F}}) \leq LR_n(\mathcal{F}),$$

so we can restrict attention to $R_n(\mathcal{F})$.

29

Estimation Error and Rademacher Averages

For **kernel classes**, $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \lambda\}$,

$$R_n(\mathcal{F}) \leq \lambda \frac{\mathbf{E}\sqrt{\text{trace}(K)}}{n},$$

where K is the Gram matrix (inner products between training data pairs).

Thus, the *empirical risk minimizer* $\hat{f} = \arg \min_{f \in \mathcal{F}} \mathbf{E}_n\ell_f$ satisfies

$$\mathbf{E}\ell_{\hat{f}} - \inf_{f \in \mathcal{F}} \mathbf{E}\ell_f \leq 2L\lambda \frac{\mathbf{E}\sqrt{\text{trace}(K)}}{n} + \sqrt{\frac{c}{n}}.$$

(There is an analogous data-dependent bound that holds with high probability over the data.)

(Bartlett, Bousquet, Mendelson, 2002)

30

Estimation Error and Rademacher Averages

For **boosting algorithms**, with

$$\mathcal{F} = \left\{ \sum_{i=1}^k \alpha_i g_i : \|\alpha\|_1 \leq \lambda, g_i \in \mathcal{G} \right\},$$

we have

$$R_n(\mathcal{F}) = \lambda R_n(\mathcal{G}).$$

(Koltchinskii and Panchenko, 2002)

31

Estimation Error and Rademacher Averages

For **sigmoid neural networks**, with

$$\mathcal{F} = \left\{ x \mapsto \sum_{i=1}^k w_i \sigma(v_i^T x) : \|w\|_1 \leq B, \|v_i\|_1 \leq B \right\},$$

where $\sigma : \mathbb{R} \rightarrow [-1, 1]$ is Lipschitz (with constant 1),

$$R_n(\mathcal{F}) \leq B^2 \sqrt{\frac{2 \log(\dim(\mathcal{X}))}{n}}.$$

32

Outline

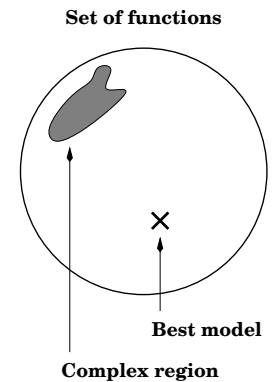
- Prediction algorithms:
 - Kernel methods
 - Boosting methods
- Estimation error:
 - Complexity of function class
 - Convexity
- Convexity and approximation error
- Dependent data.

33

Convexity and Estimation Error

A bound on $\sup_{f \in \mathcal{F}} (\mathbf{E}l_f - \mathbf{E}_n l_f)$ ensures that functions with small empirical risk ($\mathbf{E}_n l_f$) have small risk ($\mathbf{E}l_f$). But this can be conservative.

Some part of the space of functions might be complex, but if it is far from the functions of interest (ie: those that fit the data well), it is unlikely to be important.



34

Convexity and Estimation Error

Consider instead

$$\begin{aligned} \mathcal{L}_{\mathcal{F}} &= \{l_f - l_{f^*} : f \in \mathcal{F}\} \\ &= \{(x, y) \mapsto l(f(x), y) - l(f^*(x), y) : f \in \mathcal{F}\}, \end{aligned}$$

where $f^* = \arg \min_{f \in \mathcal{F}} \mathbf{E}l_f$.

If

- l is Lipschitz, uniformly convex (say, has quadratic modulus of convexity)
- \mathcal{F} is convex,

then for all $f \in \mathcal{F}$, $\text{var}(l_f - l_{f^*}) \leq c \mathbf{E}(l_f - l_{f^*})$.

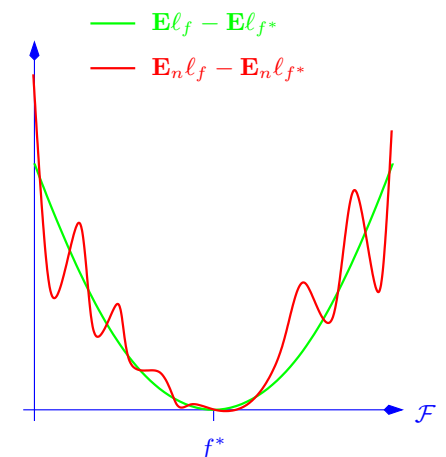
(Bartlett, Jordan, McAuliffe, 2003)

35

Convexity and Estimation Error

Even if $\mathbf{E}_n l_f$ fluctuates a lot, if $\mathbf{E}l_f$ is close to the minimal $\mathbf{E}l_{f^*}$, the fluctuations of $\mathbf{E}_n l_f$ and $\mathbf{E}_n l_{f^*}$ are strongly correlated.

Thus, $\hat{f} = \arg \min_{f \in \mathcal{F}} \mathbf{E}_n l_f$ has risk $\mathbf{E}l_{\hat{f}}$ converging quickly to $\mathbf{E}l_{f^*}$.



36

Convexity and Estimation Error

Theorem: For

- bounded, Lipschitz ℓ with modulus of convexity $\delta(\epsilon) \geq c\epsilon^2$,
- convex \mathcal{F} ,

if $\hat{f} \in \mathcal{F}$ minimizes $\mathbf{E}_n \ell_f$, then

$$\mathbf{E} \ell_{\hat{f}} \leq \mathbf{E} \ell_{f^*} + \epsilon^* + \frac{c}{n},$$

provided

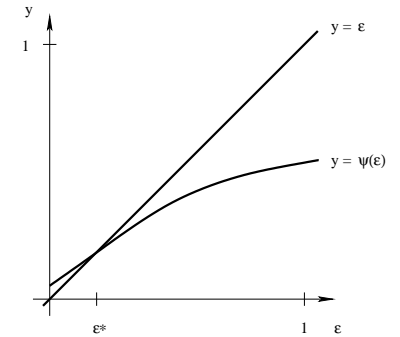
$$\epsilon^* \geq c R_n \underbrace{\{f \in \mathcal{F} : \mathbf{E} \ell_f \leq \mathbf{E} \ell_{f^*} + \epsilon^*\}}_{\text{local Rademacher averages}} = \Psi(\epsilon^*).$$

(Bartlett, Bousquet, Mendelson, 2002); see also (Koltchinskii and Panchenko, 2000), (Massart, 2000), (Lugosi and Wegkamp, 2003)

37

Convexity and Estimation Error

The local Rademacher averages can be *much* smaller than the Rademacher averages of the full class ($\Psi(1)$).

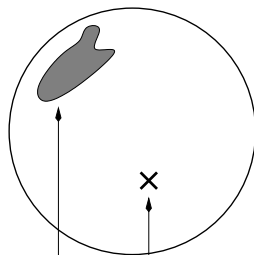


38

Convexity and Estimation Error

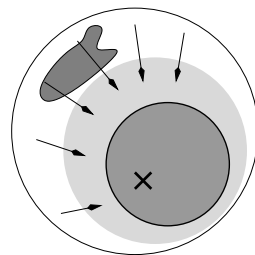
The intuition:

Set of functions



Complex region

Best model



If complexity of the set is sufficiently small, can confidently shrink the set.

39

Convexity and Estimation Error

For **kernel classes**, $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \lambda\}$, we can choose

$$\epsilon^* \geq c\lambda \sqrt{\frac{\sum_{i=1}^n \min(\mu_i/n, \epsilon^*)}{n}},$$

where μ_i are the eigenvalues of the Gram matrix K (the matrix of inner products between training data pairs).

For rapidly decreasing eigenvalues (e.g., Gaussian kernel), this gives a faster convergence rate.

40

Convexity and Estimation Error

For boosting algorithms, with

$$\mathcal{F} = \left\{ \sum_{i=1}^k \alpha_i g_i : \|\alpha\|_1 \leq \lambda, g_i \in \mathcal{G} \right\}$$

and $\dim(\mathcal{G}) = d$, we can choose

$$\epsilon^* = c(\lambda + 1)n^{-(d+2)/(2d+2)}.$$

Again, we get a **faster rate** of convergence of $\mathbf{E}l_{\hat{f}}$ to $\mathbf{E}l_{f^*}$.

41

Convexity and Estimation Error

Summary:

If

- l is strictly convex,
 - \mathcal{F} is convex,
- then $\mathbf{E}l_{\hat{f}} - \mathbf{E}l_{f^*}$ decreases rapidly,

where $\hat{f} = \arg \min_{f \in \mathcal{F}} \mathbf{E}_n l_f$, and $f^* = \arg \min_{f \in \mathcal{F}} \mathbf{E} l_f$.

In particular, the rate can be faster than $n^{-1/2}$ under these conditions.

There are converse results.

42

Convexity and Estimation Error: Converse Result

Theorem: If

- $l(\hat{y}, y) = (\hat{y} - y)^2$, and
- \mathcal{F} is *not convex* (i.e., for some $P_{\mathcal{X}}$ on \mathcal{X} , the closure of \mathcal{F} in $L_2(P_{\mathcal{X}})$ is not convex)

then

- for some $B > 0$,
 - for every algorithm for choosing \hat{f}
 - for every n ,
- there is a probability distribution P on $\mathcal{X} \times [-B, B]$ with

$$\mathbf{E}l_{\hat{f}} - \mathbf{E}l_{f^*} \geq \frac{c}{\sqrt{n}}.$$

(Lee, Bartlett, Williamson, 1998)

43

Outline

- Prediction algorithms:
 - Kernel methods
 - Boosting methods
- Estimation error:
 - Complexity of function class
 - Convexity
- Convexity and approximation error
- Dependent data.

44

Convexity and Approximation

Thus, if ℓ is convex, \mathcal{F} is non-convex, then replacing \mathcal{F} with $\text{co}(\mathcal{F})$ leads to

- Computational cost never much worse.
- Approximation error at least as good.
- Estimation error not much worse (e.g., ℓ quadratic, \mathcal{F} simple).

What about a non-convex cost function ℓ ?

45

Large Margin Classifiers

- In two-class classification ($y \in \{\pm 1\}$), we have

$$\ell(\hat{y}, y) = \mathbf{1}[\text{sign}(\hat{y}) \neq y].$$

- Minimizing empirical risk is often intractable.
- We replace the 0-1 loss, ℓ , with a convex surrogate, ϕ .
- Can efficiently minimize empirical average over many natural function classes.
- This is the approach used by AdaBoost, support vector machines, logistic regression, sigmoid neural networks, . . .
- What is the impact of this computational convenience?

46

Large Margin Classifiers

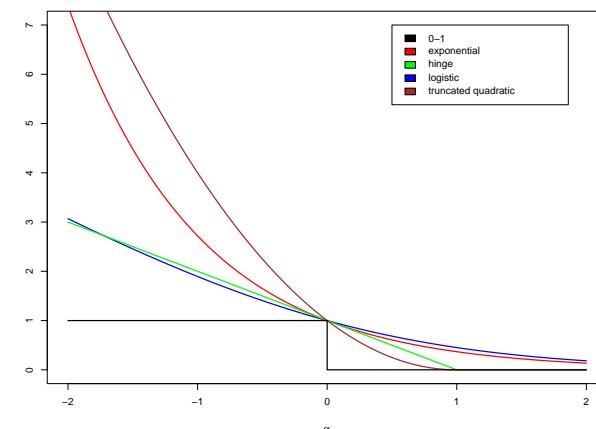
- Consider the margins, $Yf(X)$.
- Define a margin cost function $\phi : \mathbb{R} \rightarrow \mathbb{R}^+$.
- Define the ϕ -risk of $f : \mathcal{X} \rightarrow \mathbb{R}$ as $R_\phi(f) = \mathbf{E}\phi(Yf(X))$.
- Choose $f \in \mathcal{F}$ to minimize ϕ -risk.
(e.g., use data, $(X_1, Y_1), \dots, (X_n, Y_n)$, to minimize **empirical ϕ -risk**,

$$\hat{R}_\phi(f) = \hat{\mathbf{E}}\phi(Yf(X)) = \frac{1}{n} \sum_{i=1}^n \phi(Y_i f(X_i)),$$

or a regularized version.)

47

Large Margin Classifiers



48

Consequences of Using a Convex Cost

$$R(f) = \Pr(\text{sign}(f(X)) \neq Y) \quad R^* = \inf_f R(f) \quad (\text{Bayes risk})$$

$$R_\phi(f) = \mathbf{E}\phi(Yf(X)) \quad R_\phi^* = \inf_f R_\phi(f) \quad (\text{optimal } \phi\text{-risk})$$

Theorem: For any ϕ , there is a function ψ such that for all P and f ,

$$\psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^*.$$

(Bartlett, Jordan, McAuliffe, 2003)

49

Consequences of Using a Convex Cost

For convex ϕ , the ψ -transform is given by

$$\psi(\theta) = \inf_{\alpha \leq 0} \left(\frac{1+\theta}{2} \phi(\alpha) + \frac{1-\theta}{2} \phi(-\alpha) \right) - \inf_{\alpha} \left(\frac{1+\theta}{2} \phi(\alpha) + \frac{1-\theta}{2} \phi(-\alpha) \right).$$

- This gives the best possible upper bound (it cannot be improved anywhere).
- For convex ϕ , $R_\phi(f_n) \rightarrow R_\phi^*$ always implies $R(f_n) \rightarrow R^*$ if and only if $\phi'(0)$ exists and is negative.

50

ψ -transform: Example

$$\psi(\theta) = \inf_{\alpha \leq 0} \left(\frac{1+\theta}{2} \phi(\alpha) + \frac{1-\theta}{2} \phi(-\alpha) \right) - \inf_{\alpha} \left(\frac{1+\theta}{2} \phi(\alpha) + \frac{1-\theta}{2} \phi(-\alpha) \right).$$

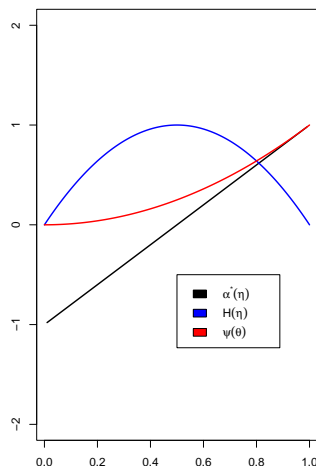
e.g., two-norm soft-margin SVM:

$$\phi(\alpha) = (\max(1 - \alpha, 0))^2.$$

first term = 1.

second term = $H((1 + \theta)/2)$, with

$$H(\eta) = 4\eta(1 - \eta).$$



51

Outline

- Prediction algorithms:
 - Kernel methods
 - Boosting methods
- Estimation error:
 - Complexity of function class
 - Convexity
- Convexity and approximation error
- Dependent data.

52

Dependent Data

We have assumed that $(X_1, Y_1), \dots, (X_n, Y_n)$ are **independent and identically distributed**.

Consider

$$\begin{aligned} X_t &= (u_t, u_{t-1}, u_{t-2}, \dots) \\ Y_t &= f(u_t, u_{t-1}, u_{t-2}, \dots) + \epsilon_t. \end{aligned}$$

Clearly, the (X_i, Y_i) are **dependent**, and the concentration results can no longer be applied.

53

Dependent Data

However, if

- the process is **ergodic** and **rapidly mixing**, and
 - every $f \in \mathcal{F}$ has a **fading memory** property,
- then we can relate the behaviour of an empirical risk minimization algorithm to its behaviour with i.i.d. data.

[See, e.g., (Nobel and Dembo, 1993), (Yu, 1994), (Weyer, 2000), (Vidyasagar and Karandikar, 2001,2002).]

- Data-dependent error estimates in this setting?

54

Complexity, Concentration, and Convexity

- Prediction algorithms:
Choose $\hat{f} \in \mathcal{F}$ to minimize empirical risk, $\mathbf{E}_n \ell_f$.
 - Kernel methods
 - Boosting methodsConvex, infinite-dimensional \mathcal{F} ; convex ℓ .
- Estimation error: $\mathbf{E} \ell_{\hat{f}} - \inf_{f \in \mathcal{F}} \mathbf{E} \ell_{f^*}$.
 - Bound using Rademacher averages
 - Implies complexity penalty:
 - * norm in RKHS, kernel matrix eigenvalues,
 - * norm of boosting coefficients.

55

Complexity, Concentration, and Convexity

- If \mathcal{F} and ℓ are convex, local Rademacher averages give better error estimates.
- Benefits of replacing a nonconvex \mathcal{F} with its convex hull.
- Effect of replacing a nonconvex ℓ with a convex surrogate: large margin classifiers.
- Dependent data.

56

References

- [1] P. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, University of California at Berkeley, 2003.
- [2] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. Technical report, University of California at Berkeley, 2003.
- [3] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. (<http://www.stat.berkeley.edu/~bartlett/papers/bm-em-03.pdf>), 2003.
- [4] Michael Collins. Discriminative reranking for natural language parsing. In Pat Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*, Stanford University, Stanford, CA, USA, June 29–July 2, 2000, pages 175–182. Morgan Kaufmann, 2000.
- [5] Michael Collins. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 489–496, 2002.
- [6] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.
- [7] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In Claire Nedellec and Céline Rouveirol, editors, *Machine Learning: ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Proceedings*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer, 1998.
- [8] Rajeeva L. Karandikar and M. Vidyasagar. Rates of uniform convergence of empirical means with mixing processes. *Statistics and Probability Letters*, 58(3):297–307, 2002.
- [9] George S. Kimeldorf and Grace Wahba. Some results on Tchebycheff spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.
- [10] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1):1–50, 2002.
- [11] Vladimir I. Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1), 2002.

- [12] G. R. G. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In *Proceedings of the 2003 Pacific Symposium on Biocomputing (PSB)*, 2003. In press.
- [13] Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.
- [14] Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.
- [15] Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. manuscript (available at <http://www.econ.upf.es/~lugosi/penaltynew.ps>), 2002.
- [16] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX:245–303, 2000.
- [17] S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.
- [18] A. B. Nobel and A. Dembo. A note on uniform laws of averages for dependent processes. *Statistics and Probability Letters*, 17:169–172, 1993.
- [19] M. Rochery, R. Schapire, M. Rahim, N. Gupta, G. Riccardi, S. Bangalore, H. Alshawi, and S. Douglas. Combining prior knowledge and boosting for call classification in spoken language dialogue. In *Proceedings of the 2002 International Conference on Acoustics, Speech and Signal Processing*, 2002.
- [20] Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.
- [21] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
- [22] M. Vidyasagar and R. L. Karandikar. A learning theory approach to system identification and stochastic adaptive control. In *IFAC Symposium on Adaptation and Learning*, 2001.
- [23] E. Weyer. Finite sample properties of system identification of ARX models under mixing conditions. *Automatica*, 36(9):1291–1299, 2000.
- [24] B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22:94–116, 1994.