Convex methods for classification

Peter Bartlett
Department of Statistics and Computer Science Division
UC Berkeley

Joint work with
Sylvain Arlot, Mike Jordan, Jon McAuliffe, Mikhail Traskin.

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

The Pattern Classification Problem

- i.i.d. $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ from $\mathcal{X} \times \{\pm 1\}$.
- Use data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose $f_n : \mathcal{X} \to \mathbb{R}$ with small risk,

$$R(f) = \Pr\left(\mathrm{sign}(f(X)) \neq Y\right) = \mathbf{E}\ell_f(X, Y),$$

where ℓ_f is the 0-1 loss:

$$\ell_f(x, y) = \begin{cases} 1 & \text{if } y \neq \mathrm{sign}(f(x)), \\ 0 & \text{otherwise.} \end{cases}$$

The Pattern Classification Problem

- Natural approach: minimize empirical risk,

$$\hat{R}(f) = \hat{\mathbf{E}}\ell_f = \frac{1}{n}\sum_{i=1}^{n}\ell_f(X_i, Y_i).$$

- Often computationally intractable...
- An alternative approach:

  Replace 0-1 loss, ℓ, with a convex surrogate, ϕ.

Large Margin Algorithms

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

Large Margin Algorithms

- Adaboost:
  - $\mathcal{F} = \mathrm{span}(\mathcal{G})$ for a VC-class $\mathcal{G}$,
  - $\phi(\alpha) = \exp(-\alpha)$,
  - Minimizes $\hat{R}_\phi(f)$ using greedy basis selection, line search:

$$f_{t+1} = f_t + \alpha_{t+1} g_{t+1},$$

$$\hat{R}_\phi(f_t + \alpha_{t+1} g_{t+1}) = \min_{\alpha \in \mathbb{R}, g \in \mathcal{G}} \hat{R}_\phi(f_t + \alpha g).$$

Large Margin Algorithms

- Support vector machines:
  - $\mathcal{F}$ = ball in reproducing kernel Hilbert space, $\mathcal{H}$.
  - $\phi(\alpha) = \max(0, 1-\alpha)$.
  - Algorithm minimizes $\hat{R}_\phi(f) + \lambda \|f\|_{\mathcal{H}}^2$.
    This is equivalent to a quadratic program:

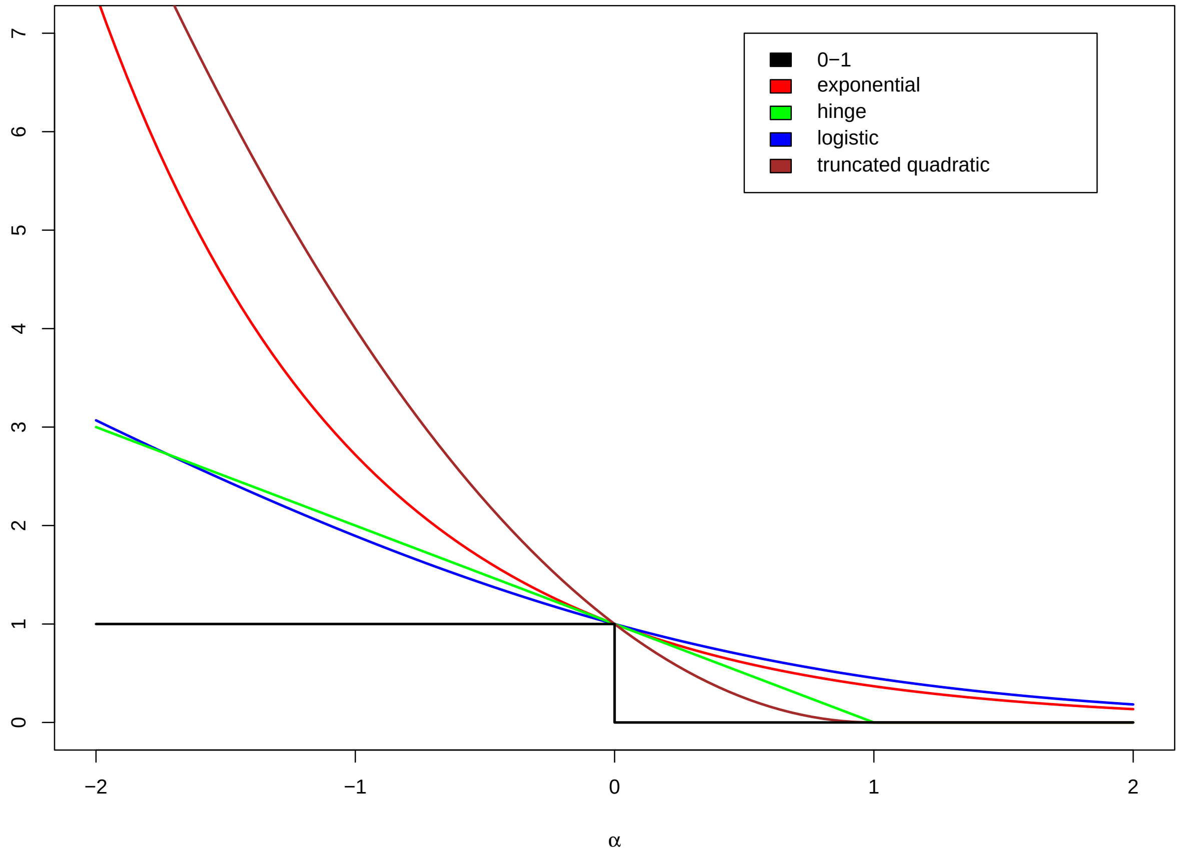
$$\begin{aligned} \min \quad & \xi' 1 + \lambda \alpha' K \alpha \\ \text{s.t.} \quad & 1 - \xi \leq \text{diag}(y) K \alpha, \\ & \xi \geq 0, \end{aligned}$$

where $y = (Y_1, \ldots, Y_n)$,
$K_{i,j} = k(X_i, X_j)$,
$\hat{f}(x) = \sum_{i=1}^n \alpha_i k(X_i, x)$,
and k is the reproducing kernel of $\mathcal{H}$.

Large Margin Algorithms

- Many other variants
  - Neural net classifiers
  - L2Boost, LS-SVMs
  - Logistic regression

Large Margin Algorithms



Overview

1. Classification with convex loss.
2. Universal consistency of large margin algorithms.
3. Classification problems with low noise.

Overview

1. Classification with convex loss.
   - Impact of replacing 0-1 loss with convex loss?
   - What ϕ are suitable for classification?
   - Relationship between risk and ϕ-risk?
2. Universal consistency of large margin algorithms.
3. Classification problems with low noise.

Definitions and Facts

$R(f) = \Pr\left(\operatorname{sign}(f(X)) \neq Y\right)$ $\quad R^* = \inf_f R(f)$ $\quad$ risk

$R_\phi(f) = \mathbb{E}\phi(Yf(X))$ $\quad R^*_\phi = \inf_f R_\phi(f)$ $\quad$ ϕ-risk

$\eta(x) = \Pr(Y = 1|X = x)$ $\quad$ conditional probability

$f^*(x) = \operatorname{sign}(2\eta(x) - 1)$ $\quad$ Bayes decision rule.

Notice: $R_\phi(f) = \mathbb{E}\left(\mathbb{E}\left[\phi(Yf(X))|X\right]\right)$, and conditional ϕ-risk is:

$$\mathbb{E}\left[\phi(Yf(X))|X = x\right] = \eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)).$$

Definitions

$$H(\eta) = \inf_{\alpha \in \mathbb{R}} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right)$$

$$H^{-}(\eta) = \inf_{\alpha:\alpha(2\eta-1)\leq 0} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right).$$

Definition: We say that ϕ is **classification-calibrated** if, for $\eta \neq 1/2$,

$$H^{-}(\eta) > H(\eta).$$

i.e., pointwise optimization of conditional ϕ-risk leads to the correct sign.

The ψ transform

Definition: Given convex ϕ, define $\psi : [0, 1] \to [0, \infty)$ by

$$\psi(\theta) = H^{-}\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right).$$

(The definition is a little more involved for non-convex ϕ.)

The Relationship between Excess Risk and Excess ϕ-risk

Theorem: [with Mike Jordan and Jon McAuliffe]

1. For any P and f, $\quad \psi(R(f) - R^*) \leq R_\phi(f) - R^*_\phi$.
2. For $|\mathcal{X}| \geq 2$, $\epsilon > 0$ and $\theta \in [0, 1]$, there is a P and an f with

$$R(f) - R^* = \theta$$

$$\psi(\theta) \leq R_\phi(f) - R^*_\phi \leq \psi(\theta) + \epsilon.$$

3. The following conditions are equivalent:
   (a) ϕ is classification calibrated.
   (b) $\psi(\theta_i) \to 0$ iff $\theta_i \to 0$.
   (c) $R_\phi(f_i) \to R^*_\phi$ implies $R(f_i) \to R^*$.

Classification-calibrated ϕ

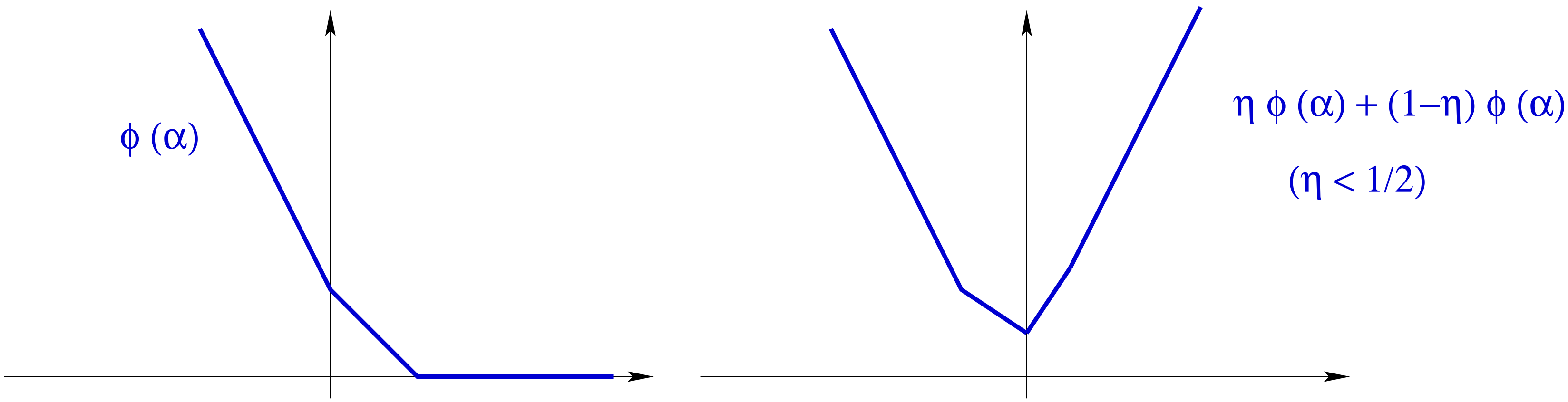
Theorem: If ϕ is convex,

$$\phi \text{ is classification calibrated} \Leftrightarrow \begin{cases} \phi \text{ is differentiable at } 0 \\ \phi'(0) < 0. \end{cases}$$

Classification-calibrated ϕ

Theorem: If ϕ is convex,

$$\phi \text{ is classification calibrated} \Leftrightarrow \begin{cases} \phi \text{ is differentiable at } 0 \\ \phi'(0) < 0. \end{cases}$$



Classification with convex loss

Bartlett, Jordan and McAuliffe, *Convexity, classification, and risk bounds*.

See also:

Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization.

Steinwart, How to compare different loss functions and their risks.

Overview

1. Classification with convex loss.
2. Universal consistency of large margin algorithms.
   - AdaBoost.
3. Classification problems with low noise.

Universal Consistency

- Assume: i.i.d. data, $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ from from $\mathcal{X} \times \mathcal{Y}$ (with $\mathcal{Y} = \{\pm 1\}$).
- Consider a method $f_n = A((X_1, Y_1), \ldots, (X_n, Y_n))$, e.g., $f_n = \mathtt{AdaBoost}((X_1, Y_1), \ldots, (X_n, Y_n), t_n)$.

Definition: We say that the method is universally consistent if, for all distributions P,

$$R(f_n) \stackrel{a.s}{\to} R^*,$$

where R is the risk and R^* is the Bayes risk:

$$R(f) = \Pr(Y \neq \mathrm{sign}(f(X))), \qquad R^* = \inf_f R(f).$$

The Approximation/Estimation Decomposition

Consider an algorithm that chooses

$$f_n = \arg\min_{f \in \mathcal{F}_n} \hat{R}_\phi(f) \qquad \text{or} \qquad f_n = \arg\min_{f \in \mathcal{F}} \left(\hat{R}_\phi(f) + \lambda_n \Omega(f) \right).$$

($\hat{R}_\phi(f)$ is empirical ϕ-risk, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$, and Ω is regularization.)

We can decompose the excess risk estimate as

$$\begin{aligned}\psi\left(R(f_n) - R^*\right) &\leq R_\phi(f_n) - R_\phi^* \\ &= \underbrace{R_\phi(f_n) - \inf_{f \in \mathcal{F}_n} R_\phi(f)}_{\text{estimation error}} + \underbrace{\inf_{f \in \mathcal{F}_n} R_\phi(f) - R_\phi^*}_{\text{approximation error}}.\end{aligned}$$

The Approximation/Estimation Decomposition

$$\psi\left(R(f_n) - R^*\right) \le R_\phi(f_n) - R_\phi^*$$
$$= \underbrace{R_\phi(f_n) - \inf_{f\in\mathcal{F}_n} R_\phi(f)}_{\text{estimation error}} + \underbrace{\inf_{f\in\mathcal{F}_n} R_\phi(f) - R_\phi^*}_{\text{approximation error}} .$$

- Approximation and estimation errors are in terms of R_ϕ, not R.
- Like a regression problem.
- With a rich class and appropriate regularization, $R_\phi(f_n) \to R_\phi^*$. (e.g., $\mathcal{F}_n$ gets large slowly, or $\lambda_n \to 0$ slowly.)
- Universal consistency ($R(f_n) \to R^*$) iff ϕ is classification calibrated.

Example: Universal Consistency of SVMs

For a Reproducing Kernel Hilbert Space $\mathcal{H}$, choose

$$f_n = \arg\min_{f \in \mathcal{H}} \left(\hat{R}_\phi(f) + \lambda_n \|f\|_{\mathcal{H}}^2 \right),$$

or $$f_n = \arg\min_{f \in \mathcal{H}_n} \hat{R}_\phi(f) \qquad \text{with } \mathcal{H}_n = \{f \in \mathcal{H} : \lambda_n \|f\|_{\mathcal{H}}^2 \le 1\}.$$

Then $$\psi\left(R(f_n) - R^*\right) \le \underbrace{R_\phi(f_n) - \inf_{f \in \mathcal{H}_n} R_\phi(f)}_{\text{estimation error}} + \underbrace{\inf_{f \in \mathcal{H}_n} R_\phi(f) - R_\phi^*}_{\text{approximation error}}.$$

If $\mathcal{H}$ is large (e.g., a Gaussian kernel on $\mathbb{R}^d$), $\inf_{f \in \mathcal{H}_n} R_\phi(f) \to R_\phi^*$.

For $\lambda_n \to 0$ (suitably slowly), $|\hat{R}_\phi(f_n) - R_\phi(f_n)| \stackrel{a.s}{\to} 0$.

In that case, $R_\phi(f_n) \stackrel{a.s}{\to} R_\phi^*$, and universal consistency follows.

(Steinwart, 2005)

Universal Consistency: AdaBoost?

- For SVMs, the regularization term keeps f_n small, which is essential for the uniform convergence result: $|\hat{R}_\phi(f_n) - R_\phi(f_n)| \xrightarrow{a.s} 0$.
- AdaBoost?

AdaBoost

Sample, $S_n = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times \{\pm 1\})^n$

Number of iterations, T

Class of basis functions, $\mathcal{G}$

function AdaBoost(S_n, T):

$f_0 := 0$

for t from $1, \ldots, T$

$$(\alpha_t, g_t) := \arg \min_{\alpha \in \mathbb{R}, g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \exp\left(-y_i \left(f_{t-1}(x_i) + \alpha g(x_i)\right)\right)$$

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

return f_T

Previous results: Regularized versions

Instead, we could consider a regularized version of AdaBoost:

1. Minimize $\hat{R}_\phi(f)$ over $\mathcal{F}_n = \gamma_n \mathrm{co}(\mathcal{G})$, the scaled convex hull of $\mathcal{G}$.
2. Minimize

$$\hat{R}_\phi(f) + \lambda_n \|f\|_*,$$

over $\mathrm{span}(\mathcal{G})$, where $\|f\|_* = \inf\{\gamma : f \in \gamma \mathrm{co}(\mathcal{G})\}$.

For suitable choices of the parameters (γ_n and λ_n), these algorithms are universally consistent. (Lugosi and Vayatis, 2004), (Zhang, 2004)

Also bounded step size. (Zhang and Yu, 2005), (Bickel, Ritov, Zakai, 2006)

Previous results: 'Process consistency'

Theorem: [Jiang, 2004]
For a (suitable) basis class defined on $\mathbb{R}^d$, and for all probability distributions P satisfying certain smoothness assumptions, there is a sequence t_n such that $f_n =$`AdaBoost(`S_n, t_n`)` satisfies

$$R(f_n) \stackrel{a.s.}{\longrightarrow} R^*.$$

Universal consistency of AdaBoost

Theorem: [with Mikhail Traskin]

If
$$d_{VC}(\mathcal{G}) < \infty,$$
$$R_\phi^* = \lim_{\lambda\to\infty} \inf \left\{ R_\phi(f) : f \in \lambda \mathrm{co}(\mathcal{G}) \right\},$$
$$t_n \to \infty$$
$$t_n = O(n^{1-\alpha}) \qquad \text{for some } \alpha > 0,$$

then AdaBoost is universally consistent.

Universal consistency of AdaBoost

Theorem:

If
$$d_{VC}(\mathcal{G}) < \infty,$$
$$R_\phi^* = \lim_{\lambda\to\infty} \inf\left\{R_\phi(f) : f \in \lambda\mathrm{co}(\mathcal{G})\right\},$$
$$t_n \to \infty$$
$$t_n = O(n^{1-\alpha}) \qquad \text{for some } \alpha > 0,$$

then AdaBoost is universally consistent.

Idea of proof:

Uniform convergence of clipped t_n-combinations. Clipping does not greatly increase $\hat{R}_\phi$. Then $\hat{R}_\phi(f_{t_n})$ approaches best in an ℓ_*-ball. Then uniform convergence over ℓ_*-balls.

Overview

1. Classification with convex loss.
2. Universal consistency of large margin algorithms.
3. Classification problems with low noise.

Low Noise

The difficulty of a binary classification problem is determined by the probability that $\eta(X) = \Pr(Y = 1|X)$ is near $1/2$.

Most favorable case:
for some $c > 0$, $\Pr\left(0 < |2\eta(X) - 1| < c\right) = 0$.

Low Noise

Definition: [Tsybakov] The distribution P on $\mathcal{X} \times \{\pm 1\}$ has *noise exponent* $0 \leq \alpha < \infty$ if there is a $c > 0$ such that

$$\Pr\left(0 < |2\eta(X) - 1| < \epsilon\right) \leq c\epsilon^{\alpha}.$$

- Tsybakov considered empirical risk minimization in binary classification.
- Under the noise assumption, if the Bayes classifier is in the function class, the risk of the empirical risk minimizer converges suprisingly quickly to the minimum.

Overview

1. Classification with convex loss.
2. Universal consistency of large margin algorithms.
3. Classification problems with low noise.
   - Large margin classifiers exploit low noise.
   - Adaptivity to low noise.

Risk Bounds with Low Noise: Convex Losses

Low noise improves the comparison inequality: (Bartlett, Jordan, McAuliffe)

$$c\left(R(f)-R^*\right)^\beta \psi\left(\frac{\left(R(f)-R^*\right)^{1-\beta}}{2c}\right) \leq R_\phi(f)-R_\phi^*,$$

where $\beta = \dfrac{\alpha}{1+\alpha} \in [0,1]$. (Consider, for example, $\alpha = \infty$.)

- Strictly convex loss ϕ (e.g., AdaBoost's exponential loss)

  $\Rightarrow \psi$ strictly convex $\Rightarrow$ strict improvement.

Risk Bounds with Low Noise: Convex Losses

Example: Suppose that ϕ has quadratic modulus of convexity, $\Pr\left(0 < |2\eta(X) - 1| < c\right) = 0$, and $\hat{f}$ minimizes $\hat{R}_\phi$ over a finite-dimensional function class $\mathcal{F}$. Then

$$\mathbf{E}R(\hat{f}) - R^* \leq C\left(\inf_{f \in \mathcal{F}} R_\phi(f) - R_\phi^* + \frac{\log n}{n}\right).$$

- Striking: the fluctuations in $\hat{R}(\hat{f})$ are of the order of $1/\sqrt{n}$ in this case.
- Note that the algorithm minimizes the empirical ϕ-risk as before, but now an improvement in the noise exponent gives an improvement in the rate.

Low Noise: Small Variance

Definition: The distribution P on $\mathcal{X} \times \{\pm 1\}$ has *noise exponent* $0 \leq \alpha < \infty$ if there is a $c > 0$ such that

$$\Pr\left(0 < |2\eta(X) - 1| < \epsilon\right) \leq c\epsilon^{\alpha}.$$

- Equivalently, there is a c such that for every $f : \mathcal{X} \to \{\pm 1\}$,

$$\Pr\left(f(X) \neq f^*(X)\right) \leq c\left(R(f) - R^*\right)^{\beta}$$

$$\Leftrightarrow \quad \mathbf{E}\left(\ell_f - \ell_{f^*}\right)^2 \leq c\left(\mathbf{E}\left(\ell_f - \ell_{f^*}\right)\right)^{\beta},$$

where f^* is the Bayes decision rule and $\beta = \dfrac{\alpha}{1+\alpha}$.

Low Noise: Small Variance

Suppose that, for some g^* (think ℓ_{f^*}) and for all g (think ℓ_f),

$$b\left(\sqrt{\mathrm{Var}(g - g^*)}\right) \leq \mathbf{E}\left(g - g^*\right),$$

where b is a convex, increasing function.

- The variance of the excess loss is bounded in terms of its expectation.
- As the risk, $\mathbf{E}g$, approaches the optimal risk, $\mathbf{E}g^*$, the loss g becomes more correlated with g^*.
- This ensures that the excess risk $\mathbf{E}\left(\hat{g} - g^*\right)$ for the empirical minimizer $\hat{g}$, converges quickly.

Local Low Noise and Model Selection

Suppose that we wish to do model selection over a nested hierarchy,

$$F_1 \subseteq F_2 \subseteq \cdots \subseteq F_m \subseteq \cdots$$

Tsybakov's low noise assumption bounds the variance of the excess loss of all functions. Instead, suppose we have convex increasing b_m for which

$$\text{for all } m \text{ and all } f \in F_m, \qquad b_m\left(\sqrt{\mathrm{Var}(\ell_f - \ell_{f^*})}\right) \leq \mathbf{E}\left(\ell_f - \ell_{f^*}\right).$$

This allows for the local condition to be favorable,
even when the best global condition is weak.

For instance, for some small model F_m, we might have small variance of excess loss, that is, the best functions tend to agree with the Bayes rule.

Local Low Noise and Model Selection

Consider penalization-based model selection schemes:

$$
\begin{aligned}
\text{empirical minimizer in } F_m: &\qquad \hat{f}_m = \arg\min_{f \in F_m} \hat{\mathbf{E}} \ell_f, \\
\text{selected model:} &\qquad \hat{m} = \arg\min_m \left(\hat{\mathbf{E}} \ell_{\hat{f}_m} + \mathrm{pen}(m) \right), \\
\text{estimator:} &\qquad \hat{f} = \hat{f}_{\hat{m}}.
\end{aligned}
$$

Ideally, the penalty $\mathrm{pen}(m)$ would approximate the difference between the risk and the empirical risk of $\hat{f}_m$.

Local Low Noise and Model Selection

Theorem: [with Sylvain Arlot] There is a penalty $\mathrm{pen}(m)$ such that for all b_m, with probability at least $1 - e^{-t}$, $R(\hat{f}) - R^*$ is no more than

$$C \inf_m \left(\inf_{f \in F_m} (R(f) - R^*) + \mathrm{pen}(m) + b_m^* \left(\sqrt{\frac{ct}{n}} \right) + \frac{t}{n} \right),$$

where b_m^* is the convex conjugate of b_m:
$b_m^*(x) = \sup\{xy - b_m(y) : y \geq 0\}$.

- The penalties are *local Rademacher averages*.

See (Massart, 2000), (Lugosi and Wegkamp, 2004), (Bartlett, Bousquet and Mendelson, 2005), *(Koltchinskii, 2006)*

- This model selection scheme *adapts* to the b_m.

Local Low Noise and Model Selection

For example, suppose that

- $\mathrm{VCdim}(F_m) = V_m$,
- every $f \in F_m$ satisfies the local low noise condition

$$\Pr(f \neq f^*) \leq \frac{1}{h_m} \left(R(f) - R^* \right).$$

Then this model selection scheme satisfies the oracle inequality

$$\mathbf{E} R(\hat{f}) - R^* \leq c \inf_m \left(\inf_{f \in F_m} \left(R(f) - R^* \right) + \frac{V_m \log n}{h_m n} \right).$$

This is optimal up to a factor of $\log n$.

Convex methods for classification

1. Classification with convex loss.
   - Large margin methods.
   - Classification-calibrated ϕ: minimization of R_ϕ minimizes R.
2. Universal consistency of large margin algorithms.
   - AdaBoost, stopped after $t_n = O(n^{1-\alpha})$, is universally consistent.
3. Classification problems with low noise.
   - Large margin methods exploit low noise.
   - Penalization-based model selection methods that are adaptive to local low noise conditions.

slides at http://www.stat.berkeley.edu/∼bartlett