Binary classification

Observation: \( X \in \mathcal{X} \), distribution \( \mu \)

Label/Class: \( Y \in \{-1, +1\} \)

Distribution: \( P = (\mu, \eta) \), \( \mu = P_X \)

Regression function:

\[
\eta(x) = \mathbb{P}\{Y = 1 | X = x\}
\]

Data sample:

\[
D_n = \{(X_1, Y_1), ..., (X_n, Y_n)\} \text{ i.i.d.}
\]

Classifier:

\[
g_n(X) \in \{-1, +1\}
\]
Error measure:

\[ L(g_n) = \mathbb{P}\{g_n(X) \neq Y \mid D_n\} = \mathbb{E}\mathbb{I}\{Y \cdot g_n(X) < 0\} \]

Bayes classifier and Bayes error:

\[ g^* = \arg \min_{g} L(g) = 2\mathbb{I}\{\eta > 1/2\} - 1 \]
\[ L^* = L(g^*) = \mathbb{E}\{\min(\eta(X), 1 - \eta(X))\} \]
Efficient (non-parametric) algorithms

- **Support Vector Machines (Vapnik, 1995)**
  1. ‘kernel trick’: map the data into a Kernel Hilbert Space where data can be (almost) linearly separated.
  2. large margin hyperplane: convex optimization with quadratic constraints.
Efficient (non-parametric) algorithms

- **Support Vector Machines (Vapnik, 1995)**
  1. 'kernel trick': map the data into a Kernel Hilbert Space where data can be (almost) linearly separated.
  2. large margin hyperplane: convex optimization with quadratic constraints.

- **Boosting (Freund, 1990 - Freund and Schapire, 1996)**
  1. start with a base class $\mathcal{H}$ of weak classifiers.
  2. then construct iteratively a combination of weak classifiers which decreases the empirical error.
AdaBoost

- **Initialization:** $\Pi_1$ uniform distribution on $D_n$

- **Step** $t$: extract $(h_t, w_t) \in \mathcal{H} \times \mathbb{R}$
  - classifier $h_t$ minimizes empirical error $\epsilon_t$ weighted by $\Pi_t$
  - weight $w_t$ given by:
    $$w_t = \frac{1}{2} \log \left( \frac{1 - \epsilon_t}{\epsilon_t} \right)$$
  - update: $\Pi_{t+1}(i) \propto \exp(-w_t Y_i h_t(X_i)) \Pi_t(i)$

- **Output (step $T$):** $g_n(X) = \text{sgn} \left( \sum_{t=1}^{T} w_t h_t(X) \right)$. 
Putting aside ... 

- geometric intuitions, or
- particular dynamics of the algorithm,

... both can be seen as ...

... penalized convex risk minimization procedures.
General setting

Estimator:
\[ f_n(X) \in \mathbb{R} \]

Classifier:
\[ g_n(X) = \text{sgn}(f_n(X)) \in \{-1, +1\} \]

Natural criterion (classification risk):
\[ L(f) = L(g) = \mathbb{E}\mathbb{I}\{Y \cdot f(X) < 0\} \]

‘Practical’ criterion (\(\phi\)-risk):
\[ A(f) = \mathbb{E}_\phi(-Yf(X)) \]
Examples of cost functions

- **NEURAL NETWORKS**: $\phi$ sigmoid
- **ADABOOST**: $\phi(u) = e^u$
- **LOGIT**: $\phi(u) = \log(1 + e^u)$
- **SVM**: $\phi(u) = (1 + u)_+$
Minimal conditions

Assume (for now):

- Infinite sample (distribution of $(X, Y)$ is known!)
- Candidates: no restriction

Question: which cost function?

Ideas:

- let $\phi$ such that $L(f) \leq A(f)$,
- $\phi$ convex
Calibrated costs (1)

Conditional $\phi$-risk:

$$\mathbb{E}(\phi(-Yf(X)) \mid X = x) = \eta(x)\phi(-f(x)) + (1 - \eta(x))\phi(f(x))$$

Let

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

We have:

$$A^* = \mathbb{E} H(\eta(X)) = \inf_f A(f)$$

Let

$$f^*_\phi(X) = \arg\min_f A(f)$$

For classification purpose, one should take $\phi$ such that:

$$f^*_\phi \cdot (2\eta - 1) > 0$$
Calibrated costs (2)

It is equivalent to assume that the infimum of

$$C_\eta(\alpha) = \eta \phi(-\alpha) + (1 - \eta) \phi(\alpha)$$

is reached over the set \( \{\alpha : \alpha(2\eta - 1) > 0\} \).

We then have:

$$H(\eta) = \min_{\alpha \in \mathbb{R}} C_\eta(\alpha)$$

Let

$$H^-(\eta) = \min_{\alpha : \alpha(2\eta - 1) \leq 0} C_\eta(\alpha)$$
Calibrated costs (3)

Definition
We shall say that \( \phi \) is classification-calibrated if:

\[
\forall \eta \neq 1/2, \quad H^-(\eta) - H(\eta) > 0.
\]

Proposition
Let \( \phi \) be a convex cost function. Then \( \phi \) is calibrated iff it is differentiable at 0 and \( \phi'(0) > 0 \).
Function $\psi$

**Definition**

For any positive cost function $\phi$, we introduce the function $\tilde{\psi}$ defined over $[0, 1]$ by

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

We call $\psi$-transform of the cost function $\phi$ the convexified of $\tilde{\psi}$:

$$
\psi = \text{conv} \tilde{\psi}.
$$

**Theorem**

*For any positive cost function $\phi$, any measurable function $f$, and any distribution $P = (\mu, \eta)$:*

$$
\psi(L(f) - L^*) \leq A(f) - A^*.
$$
Consistency and rate reduction

**Theorem**

Let $\phi$ be a calibrated cost function. Then, for any sequence $(f_n)$ and any distribution $(\mu, \eta)$:

$$\lim_{n \to \infty} A(f_n) = A^* \Rightarrow \lim_{n \to \infty} L(f_n) = L^*$$

**Theorem**

If $\phi$ is convex and calibrated, then its $\psi$-transform is convex and we have:

$$\psi(\theta) = \phi(0) - H \left( \frac{1 + \theta}{2} \right).$$
Examples (1)

- **SIGMOID**: $\phi(u) = 1 + \tanh(u)$
  
  \[ H(\eta) = 2 \min\{\eta, 1 - \eta\} \]
  
  \[ \psi(\theta) = \theta \]

- **ADABOOST**: $\phi(u) = e^u$
  
  \[ H(\eta) = 2 \sqrt{\eta(1 - \eta)} \]
  
  \[ \psi(\theta) = 1 - \sqrt{1 - \theta^2} \]
Examples (2)

- **LOGISTIC REGRESSION:** \( \phi(u) = \log(1 + e^u) \)
  \[
  H(\eta) = -\eta \log \eta - (1 - \eta) \log(1 - \eta)
  \]
  \[
  \psi(\theta) = \frac{1}{2} \log(1 - \theta^2) + \frac{\theta}{2} \log \left( \frac{1 + \theta}{1 - \theta} \right)
  \]

- **SVM:** \( \phi(u) = (1 + u)_+ \)
  \[
  H(\eta) = \min\{\eta, 1 - \eta\}
  \]
  \[
  \psi(\theta) = \frac{\theta}{2}
  \]
Excess error and convex risk

- **SIGMOID, SVM:**

\[
L(f) - L(f^*) \leq c(A(f) - A(f^*)) .
\]

\[\rightarrow\] no rate reduction!

- **ADABOOST, LOGISTIC REGRESSION:**

\[
L(f) - L(f^*) \leq c(A(f) - A(f^*))^{1/2} .
\]

\[\rightarrow\] rate reduction by a power of 2...
Estimators: $\mathcal{F} = \text{span}\{K(x, \cdot) : x \in \mathcal{X}\}$ with $K$ symmetric positive kernel.

$$f \in \mathcal{F} \Rightarrow \exists (\alpha_i), (x_i) : f = \sum_{i=1}^{+\infty} \alpha_i K(x_i, \cdot)$$

$$\|f\|_K = \sum_{i=1}^{+\infty} \sum_{j=1}^{+\infty} \alpha_i \alpha_j K(x_i, x_j)$$

Cost function: $\phi(u) = (1 + u)_+$
**Estimators:** $\mathcal{F} = \text{span}(\mathcal{H})$ with $\mathcal{H}$ symmetric class of ‘simple’ classifiers.

$$f \in \mathcal{F} \Rightarrow \exists (w_i) > 0, (g_i) : f = \sum_{i=1}^{+\infty} w_i g_i$$

$$\|f\|_{\mathcal{H}} = \sum_{i=1}^{+\infty} w_i$$

**Cost function:** $\phi(u) = \exp(u), \logit(u)$
Let $\mathcal{F}_\lambda = \{ f \in \mathcal{F} : \|f\| \leq \lambda \}$.

- **Fixed model:**

  $$
  \min_{f \in \mathcal{F}_\lambda} A_n(f) := \frac{1}{n} \sum_{i=1}^{n} \phi(-Y_if(X_i))
  $$

  $$
  \hat{f}_n^\lambda = \arg\min_{f \in \mathcal{F}_\lambda} A_n(f)
  $$

- **Model selection:**

  $$
  \min_{\lambda > 0} \left\{ \min_{f \in \mathcal{F}_\lambda} \frac{1}{n} \sum_{i=1}^{n} \phi(-Y_if(X_i)) + \text{pen}(\lambda) \right\}
  $$

  $$
  \hat{f}_n = \arg\min_{\lambda > 0} A_n(\hat{f}_n^\lambda) + \text{pen}(\lambda)
  $$
Questions

- consistency ?
- which penalty ?
- rate of convergence for $L(\hat{f}_n) - L^*$ ?
Assumptions on the cost

Assume that $\phi$ satisfies the following conditions:

- twice differentiable
- strictly increasing
- strictly convex
- $\phi(0) = 1$
- for all $x$, the constant:

$$L_\phi = \max_{x \in \mathbb{R}} \left( \frac{2(\phi'(x) + \phi'(-x))}{\frac{\phi''}{\phi'}(x) + \frac{\phi''}{\phi'}(-x)} - (\phi(x) + \phi(-x)) \right)$$

is bounded.
Rate of convergence for boosting

**Theorem**

Let:

- \( \mathcal{H} \) such that \( \text{VCdim}(\mathcal{H}) = V < \infty \)
- \( \mathcal{F} = \text{span}(\mathcal{H}) \)
- \( \hat{f}_n \) be the 'appropriate' penalized estimator (see below...),

Assume that the distribution of \((X, Y)\) satisfies:

\[
\exists \lambda_0 > 0 : \inf_{f \in \mathcal{F}_{\lambda_0}} A(f) = A(f^*).
\]

Then, for all \( n \), with probability \( > 1 - 1/n^2 \), we have:

\[
L(\hat{f}_n) - L^* \leq Cn^{-1/4} \left( \frac{V+2}{V+1} \right),
\]

where \( C = C(\eta, \mathcal{H}, \phi) \).
Penalty calibration

- for general cost function $\phi$:
  \[
  \text{pen}(\lambda) \propto c(V)((L_\phi + 2)\phi(\lambda))^{\frac{1}{V+1}} (\lambda \phi'(\lambda))^{\frac{V}{V+1}} n^{-\frac{1}{2} \frac{V+2}{V+1}}
  \]

- $\phi = \exp$: $L_\phi = 0$
  \[
  \text{pen}(\lambda) \propto e^{\lambda} \lambda^{\frac{V}{V+1}} n^{-\frac{1}{2} \frac{V+2}{V+1}}
  \]

- $\phi = \logit$: $L_\phi = 2 - 2 \log 2$
  \[
  \text{pen}(\lambda) \propto \lambda n^{-\frac{1}{2} \frac{V+2}{V+1}}
  \]
Refinement

Assumption (Margin condition)
There exists $\alpha \in [0, 1]$ and $\beta > 0$ such that:

$$\mathbb{P}[g_f(X) \neq g^*(X)] \leq \beta (L(f) - L^*)^\alpha .$$

Corollary
The previous rate of convergence can then be improved to:

$$L(\hat{f}_n) - L^* \leq Cn^{-\frac{1}{2(2-\alpha)}}\left(\frac{V+2}{V+1}\right) .$$
Example 1

Consider one-dimensional observations \((X \in [0, 1])\)

- let \(\mathcal{H} = \{1D \text{ - decision stumps}\},\)
- let \(\phi = \exp\) or \(\logit,\)
- let \(\hat{f}_n\) be our penalized estimator

Also assume the regression function \(\eta\) is such that:

- \(\exists b > 0\) such that \(b < \eta(X) < 1 - b\) a.s.
- \(\text{TV}(\eta) < \infty.\)

Then, for all \(n\), with probability \(> 1 - 1/n^2\), we have:

\[
L(\hat{f}_n) - L^* \leq Cn^{-\frac{1}{3}},
\]

where \(C = C(\text{TV}(\eta), b).\)
Example 2

Consider $d$-dimensional observations $(X \in [0, 1]^d)$

- let $\mathcal{H} = \{\text{decision stumps}\}$,
- let $\phi = \text{logit}$,
- let $\hat{f}_n$ be our penalized estimator

Assume an additive logistic regression model for $\eta$:

$$
\log \left( \frac{\eta(x)}{1 - \eta(x)} \right) = \sum_{i=1}^{d} f_i(x^{(i)}) ,
$$

and $\text{TV}(f_i) < \infty$. Then, for all $n$, with probability $> 1 - 1/n^2$, we have:

$$
L(\hat{f}_n) - L^* \leq C \left( \sqrt{d \log d} \right) n^{-\frac{1}{4}} \left( \frac{V_d + 2}{V_d + 1} \right) ,
$$

where $V_d \propto \log d$ and $C$ is a constant.
Price of convexity?

- not in the complexity control!
  \[\rightarrow\] Lipschitz cost functions will do...

- but in the condition on the constant \( L_\phi \)
  \[\rightarrow\] needed for the risk communication assumption to hold...
General argument - Notations

- Set:
  \[ \Delta(f, f^*) = A(f) - A(f^*) \]

- Let:
  \[ \ell(f) = \phi(-Yf(X)) \quad \text{and then} \quad A(f) = \mathbb{E}\ell(f) \]

- Define a pseudo-distance:
  \[ d^2(f, g) = \mathbb{E}(\ell(f) - \ell(g))^2 \]

- Set:
  \[ \bar{f} = \arg\min_{\lambda} A(f) \quad \text{for} \quad f \in \mathcal{F}_\lambda \]

- Set also:
  \[ \hat{f} = \arg\min_{\lambda} A_n(f) \quad \text{for} \quad f \in \mathcal{F}_\lambda \]

- Eventually:
  \[ \Delta(\bar{f}, f^*) = \text{approximation error} \]
General argument (1)

1. **Fixed model** \((F_\lambda)\):

\[
\Delta(\hat{f}, f^*) \leq \Delta(\bar{f}, f^*) + (\mathbb{P} - \mathbb{P}_n)(\ell(\hat{f}) - \ell(\bar{f}))
\]

2. **By Talagrand’s concentration inequality:**

For any \(K > 0\) for some constant \(c\), with probability \(1 - e^{-x}\), for all \(f \in F_\lambda\),

\[
(\mathbb{P} - \mathbb{P}_n)(\ell(f) - \ell(\bar{f})) \leq K^{-1}d^2(f, \bar{f}) + Kc\frac{x}{n}
\]

3. **Then, use:**

\[
d^2(f, \bar{f}) \leq 2\left(d^2(f, f^*) + d^2(\bar{f}, f^*)\right)
\]
General argument (2)

- We have, with probability $1 - e^{-x}$,
  \[
  \Delta(\hat{f}, f^*) \leq \Delta(\bar{f}, f^*) + 2K^{-1} \left( d^2(\hat{f}, f^*) + d^2(\bar{f}, f^*) \right) + K \left( c_1 r^* + c_2 \frac{x}{n} \right)
  \]

- Assume **Risk communication** holds:
  \[
  \forall f \in \mathcal{F}_\lambda, \quad d^2(f, f^*) \leq C \Delta(f, f^*)
  \]
  for some constant $C$.

- Finally, with probability $1 - e^{-x}$,
  \[
  \Delta(\hat{f}, f^*) \leq \left( \frac{1 + 2K^{-1}C}{1 - 2K^{-1}C} \right) \Delta(\bar{f}, f^*) + \left( \frac{K}{1 - 2K^{-1}C} \right) \left( c_1 r^* + c_2 \frac{x}{n} \right)
  \]
Risk communication

Assumption (Risk communication)

\[ \forall f \in \mathcal{F}_\lambda, \quad \mathbb{E}(\ell(f) - \ell(f^*))^2 \leq C\mathbb{E}(\ell(f) - \ell(f^*)) \]

for some constant \( C = C(\lambda) \).

Lemma

Under the previous assumptions on \( \phi \), we have, \( \forall f \in \mathcal{F}_\lambda \),

\[ d^2(f, f^*) \leq (\phi(\lambda) + \phi(-\lambda) + L_\phi)\Delta(f, f^*) \]
Risk communication for the hinge loss

Lemma

Let $\phi(u) = (1 + u)_+$. Assume

- $\exists \eta_0 > 0 : |\eta(X) - 1/2| \geq \eta_0$ a.s.
- $\exists \kappa : \forall f, \quad \|f\|_\infty \leq \kappa$.

Then,

$$\forall f, \quad d^2(f, f^*) \leq (\kappa + \eta_0^{-1}) \Delta(f, f^*)$$.
Lemma

If $\mathcal{F}$ contains the indicators of all the hyper-rectangles of $\mathbb{R}^d$, then:

$$\lim_{\lambda \to +\infty} \inf_{f \in \mathcal{F}_\lambda} A(f) = A^*.$$

**Question:** Given a class $\mathcal{F}$, characterize the distributions of $(X, Y)$ satisfying:

$$\exists \lambda_0 > 0 : \inf_{f \in \mathcal{F}_{\lambda_0}} A(f) = A^* ?$$
Approximation error

Lemma

If $\mathcal{F}$ contains the indicators of all the hyper-rectangles of $\mathbb{R}^d$, then:

$$\lim_{\lambda \to +\infty} \inf_{f \in \mathcal{F}_\lambda} A(f) = A^*$.$$

Question: Given a class $\mathcal{F}$, characterize the distributions of $(X, Y)$ satisfying:

$$\exists \lambda_0 > 0 : \inf_{f \in \mathcal{F}_{\lambda_0}} A(f) = A^*?$$

Answer:

- YES! for additive logistic regression models with BV components.
- Other cases: ???