Classification consistency and surrogate empirical risk minimization

Jon McAuliffe
Statistics Department
University of Pennsylvania,
Wharton School

Joint work with Peter Bartlett and Mike Jordan,
UC Berkeley Statistics and CS

28 April 2006
The binary 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 \(f_n : \mathcal{X} \rightarrow \mathbb{R}\) with small risk,

\[
R(f_n) = \Pr(\text{sign}(f_n(X)) \neq Y) = \mathbb{E}\ell(Y, f(X)).
\]

- Natural approach: minimize empirical risk,

\[
\hat{R}(f) = \hat{\mathbb{E}}\ell(Y, f(X)) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)).
\]

- Often intractable...
- Replace 0-1 loss, \(\ell\), with a convex surrogate, \(\phi\).
Large margin algorithms

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

\[
\hat{R}_\phi(f) = \hat{\mathbb{E}} \phi(Y f(X)) = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i f(X_i)),
\]

or a regularized version.)
Large margin algorithms

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

- **Support vector machines** with 2-norm soft margin.  
  - $\mathcal{F} =$ ball in reproducing kernel Hilbert space, $\mathcal{H}$.  
  - $\phi(\alpha) = (\max(0, 1 - \alpha))^2$.  
  - Minimizes $\hat{R}_\phi(f) + \lambda \| f \|^2_{\mathcal{H}}$ (using dual QP).
Large margin algorithms

• Many other variants
  
  – Neural net classifiers
    \[ \phi(\alpha) = \max(0, (0.8 - \alpha)^2). \]
  
  – Support vector machines with 1-norm soft margin
    \[ \phi(\alpha) = \max(0, 1 - \alpha). \]
  
  – L2Boost, LS-SVMs
    \[ \phi(\alpha) = (1 - \alpha)^2. \]
  
  – Logistic regression (negative Bernoulli likelihood, aka binomial deviance)
    \[ \phi(\alpha) = \log(1 + \exp(-\alpha)). \]
Large margin algorithms

![Graph showing various loss functions including 0-1, exponential, hinge, logistic, and truncated quadratic.](image)

**Consistency and Surrogate ERM**
Statistical consequences of using a convex loss

• Bayes risk consistency? For which $\phi$?
• How is 0-1 risk related to $\phi$-risk?
Overview

- Relating excess risk to excess $\phi$-risk.
  - $\psi$-transform: best possible bound.
  - proof idea.
  - conditions on $\phi$. 
Definitions and facts

\[ R(f) = \Pr(\text{sign}(f(X)) \neq Y) \quad \text{Risk,} \]
\[ R^* = \inf_f R(f) \quad \text{Bayes risk,} \]
\[ \eta(x) = \Pr(Y = 1|X = x) \quad \text{conditional probability.} \]

- \( \eta \) defines an optimal classifier:

\[ R^* = R(\text{sign}(\eta(\cdot) - 1/2)). \]

- **Excess risk** of \( f : \mathcal{X} \to \mathbb{R} \) is

\[ R(f) - R^* = \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] |2\eta(X) - 1| \right). \]
Definitions

Risk: \[ R(f) = \Pr \left( \text{sign}(f(X)) \neq Y \right). \]
\[ \phi\text{-Risk: } R_\phi(f) = \mathbb{E}_\phi(Yf(X)). \]

\[ R_\phi(f) = \mathbb{E} \left( \mathbb{E} \left[ \phi(Yf(X)) | X \right] \right). \]

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

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

\[ C_{0.3}(\alpha) = 0.3\phi(\alpha) + 0.7\phi(-\alpha) \]

\[ C_{0.7}(\alpha) = 0.7\phi(\alpha) + 0.3\phi(-\alpha) \]
Definitions

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

\[ R_\phi(f) = \mathbb{E}_\phi(Yf(X)) \]
\[ R^*_\phi = \inf_f R_\phi(f) \quad \text{optimal } \phi\text{-risk} \]

Conditional \( \phi\)-risk:

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

Optimal conditional \( \phi\)-risk for \( \eta \in [0, 1] \):

\[ H(\eta) = \inf_{\alpha \in \mathbb{R}} (\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha)). \]

\[ R^*_\phi = \mathbb{E}H(\eta(X)). \]
Optimal conditional $\phi$-risk: example
Definitions

Optimal conditional $\phi$-risk for $\eta \in [0, 1]$:

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

Optimal conditional $\phi$-risk with incorrect sign:

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

Note: $H^- (\eta) \geq H(\eta)$ \hspace{1cm} $H^- (1/2) = H(1/2)$. 
Example: $H^-(\eta) = \phi(0)$
Definitions

\[ H(\eta) = \inf_{\alpha \in \mathbb{R}} (\eta \phi(\alpha) + (1 - \eta)\phi(-\alpha)) \]

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

**Defn.** \( \phi \) is **classification-calibrated** if, for \( \eta \neq 1/2 \),

\[ H^-(\eta) > H(\eta). \]

i.e., pointwise optimization of conditional \( \phi \)-risk leads to the correct sign.
(c.f. Lin (2001))
Definitions

Defn. Given \( \phi \), define \( \psi : [0, 1] \rightarrow [0, \infty) \) by \( \psi = \tilde{\psi}^{**} \), where

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

Here, \( g^{**} \) is the Fenchel-Legendre biconjugate of \( g \),

\[
epi(g^{**}) = \overline{\co}(\epi(g)),
\]

\[
epi(g) = \{(x, y) : x \in [0, 1], \ g(x) \leq y\}.
\]
$\psi$-transform

- $\psi$ is the best convex lower bound on

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

the excess conditional $\phi$-risk when the sign is incorrect.

- $\psi = \tilde{\psi}^{**}$ is the biconjugate of $\tilde{\psi}$,

$$\text{epi}(\psi) = \overline{\cap}(\text{epi}(\tilde{\psi})),
\text{epi}(\psi) = \{(\alpha, t) : \alpha \in [0, 1], \psi(\alpha) \leq t\}.$$

- $\psi$ is the functional convex hull of $\tilde{\psi}$.
\(\psi\)-transform: example

\[\begin{align*}
\phi(\alpha) & \quad \phi(-\alpha) \\
C_{0.3}(\alpha) & \quad C_{0.7}(\alpha)
\end{align*}\]

\[\begin{align*}
\alpha & \quad \eta, \theta
\end{align*}\]
The relationship between excess risk and excess $\phi$-risk

**Theorem:**

1. For any $P$ and $f$, $\psi(R(f) - R^*) \leq R_\phi(f) - R^*_\phi$.

2. This bound cannot be improved.

3. Near-minimal $\phi$-risk implies near-minimal risk precisely when $\phi$ is classification-calibrated.
The relationship between excess risk and excess $\phi$-risk

Theorem:

1. For any $P$ and $f$, $\psi(R(f) - R^*) \leq R_\phi(f) - R^*_\phi$.

2. This bound cannot be improved:
   For $|\mathcal{X}| \geq 2$, $\epsilon > 0$ and $\theta \in [0, 1]$, there are $P, f$ with
   \[
   R(f) - R^* = \theta
   \]
   \[
   \psi(\theta) \leq R_\phi(f) - R^*_\phi \leq \psi(\theta) + \epsilon.
   \]

3. Near-minimal $\phi$-risk implies near-minimal risk precisely when $\phi$ is classification-calibrated.
The relationship between excess risk and excess $\phi$-risk

Theorem:

1. For any $P$ and $f$, $\psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^*$.

2. This bound cannot be improved.

3. The following conditions are equivalent:

   (a) $\phi$ is classification calibrated.
   (b) $\psi(\theta_i) \to 0$ iff $\theta_i \to 0$.
   (c) $R_\phi(f_i) \to R_\phi^*$ implies $R(f_i) \to R^*$. 
Excess risk bounds: proof idea

Facts:

- $H(\eta), H^-(\eta)$ are symmetric about $\eta = 1/2$.
- $H(1/2) = H^-(1/2)$, hence $\psi(0) = 0$.
- $\psi(\theta)$ is convex.
- $\psi(\theta) \leq \tilde{\psi}(\theta) = H^-(\frac{1+\theta}{2}) - H\left(\frac{1+\theta}{2}\right)$. 
Excess risk bounds: proof idea

Recall:

\[ R(f) - R^* = \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] |2\eta(X) - 1| \right). \]

Thus,

\[
\psi(R(f) - R^*) \leq \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] \psi(|2\eta(X) - 1|) \right) \leq \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] \tilde{\psi}(|2\eta(X) - 1|) \right) = \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] (H^{-}(\eta(X)) - H(\eta(X))) \right) \leq \mathbb{E} (\phi(Yf(X)) - H(\eta(X))) \leq R_\phi(f) - R_\phi^*.
\]
Excess risk bounds: proof idea

Recall:

\[ R(f) - R^* = \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] |2\eta(X) - 1| \right). \]

Thus,

\[
\begin{align*}
\psi(R(f) - R^*) & \leq \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] \psi(|2\eta(X) - 1|) \right) \\
& \leq \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] \tilde{\psi}(|2\eta(X) - 1|) \right) \\
& = \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] (H^{-}(\eta(X)) - H(\eta(X))) \right) \\
& \leq \mathbb{E} (\phi(Yf(X)) - H(\eta(X))) \\
& = R_{\phi}(f) - R_{\phi}^*. 
\end{align*}
\]
Excess risk bounds: proof idea

Recall:

\[ R(f) - R^* = \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] |2\eta(X) - 1| \right). \]

Thus,

\[
\begin{align*}
\psi(R(f) - R^*) & \leq \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] \psi(|2\eta(X) - 1|) \right) \\
& \leq \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] \tilde{\psi}(|2\eta(X) - 1|) \right) \\
& = \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] (H^-(\eta(X)) - H(\eta(X))) \right) \\
& \leq \mathbb{E} (\phi(Yf(X)) - H(\eta(X))) \\
& = R_\phi(f) - R^*_\phi.
\end{align*}
\]
Excess risk bounds: proof idea

Recall:

\[ R(f) - R^* = \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] |2\eta(X) - 1| \right). \]

Thus,

\[
\psi(R(f) - R^*) \quad \quad (H^- \text{ minimizes conditional } \phi\text{-risk})
\]

\[
\leq \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] \psi(|2\eta(X) - 1|) \right)
\]

\[
\leq \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] \tilde{\psi}(|2\eta(X) - 1|) \right)
\]

\[
= \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] (H^-(\eta(X)) - H(\eta(X))) \right)
\]

\[
\leq \mathbb{E} (\phi(Yf(X)) - H(\eta(X)))
\]

\[
= R_\phi(f) - R^*_\phi.
\]
Excess risk bounds: proof idea

Recall:

\[ R(f) - R^* = \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] |2\eta(X) - 1| \right). \]

Thus,

\[ \psi(R(f) - R^*) \leq \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] \psi(|2\eta(X) - 1|) \right) \leq \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] \tilde{\psi}(|2\eta(X) - 1|) \right) = \mathbb{E} \left( 1 \left[ \text{sign}(f(X)) \neq \text{sign}(\eta(X) - 1/2) \right] (H^{-}(\eta(X)) - H(\eta(X))) \right) \leq \mathbb{E} (\phi(Yf(X)) - H(\eta(X))) = R_\phi(f) - R^*_\phi. \]
Excess risk bounds: proof idea

Converse:

1. If \( \tilde{\psi} \) is convex, \( \psi = \tilde{\psi} \).
   Fix \( P(x_1) = 1 \) and choose \( \eta(x_1) = (1 + \theta)/2 \).
   Each inequality is clearly tight.

2. If \( \tilde{\psi} \) is not convex:
   Choose \( \theta_1 \) and \( \theta_2 \) so that \( \psi(\theta_i) = \tilde{\psi}(\theta_i) \) and \( \theta \in \text{co}\{\theta_1, \theta_2\} \).
   Set \( \eta(x_1) = (1 + \theta_1)/2 \) and \( \eta(x_2) = (1 + \theta_2)/2 \).
   Again, each inequality is clearly tight.
Theorem: If $\phi$ is convex,

$\phi$ is classification calibrated $\iff \left\{\begin{array}{l}
\phi \text{ is differentiable at } 0 \\
\phi'(0) < 0.
\end{array}\right.$

Theorem: If $\phi$ is classification calibrated,

$\exists \gamma > 0, \forall \alpha \in \mathbb{R},$

$\gamma \phi(\alpha) \geq 1 [\alpha \leq 0].$
Summary: large margin classifiers

- Risk bounds relative to Bayes risk:
  - $\psi$ relates excess risk to excess $\phi$-risk.
  - Best possible.

- Additional results on fast rates with low noise:
  - Tighter bound on excess risk.
  - Fast convergence of $\phi$-risk for strictly convex $\phi$. 