Convexity, Classification, and Risk Bounds

Peter Bartlett
UC Berkeley

Joint work with Jon McAuliffe and Mike Jordan.

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

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,
$$R(f_n) = \Pr\left(\mathrm{sign}(f_n(X)) \neq Y\right) = \mathbf{E}\ell(Y, f(X)).$$
- Natural approach: minimize empirical risk,
$$\hat{R}(f) = \hat{\mathbf{E}}\ell(Y, f(X)) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$
- Often intractable...
- Replace 0-1 loss, ℓ, with a convex surrogate, ϕ.

Large Margin Algorithms

- Consider the margins, $Y f(X)$.
- Define a margin cost function $\phi : \mathbb{R} \to \mathbb{R}^+$.
- Define the **ϕ-risk** of $f : \mathcal{X} \to \mathbb{R}$ as $R_\phi(f) = \mathbf{E}\phi(Y f(X))$.
- 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**,
$$\hat{R}_\phi(f) = \hat{\mathbf{E}}\phi(Y f(X)) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i f(X_i)),$$
or a regularized version.)

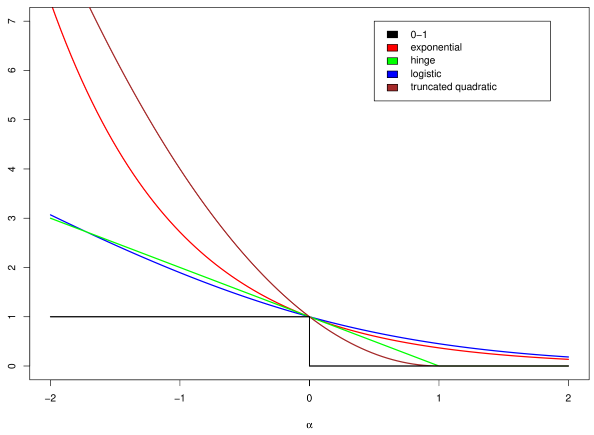
Large Margin Algorithms

- Adaboost:
  - $\mathcal{F} = \mathrm{span}(\mathcal{G})$ for a VC-class $\mathcal{G}$,
  - $\phi(\alpha) = \exp(-\alpha)$,
  - Minimizes $\hat{R}_\phi(f)$ using greedy basis selection, line search.
- Support vector machines with 2-norm soft margin.
  - $\mathcal{F} =$ ball in reproducing kernel Hilbert space, $\mathcal{H}$.
  - $\phi(\alpha) = (\max(0, 1 - \alpha))^2$.
  - Algorithm minimizes $\hat{R}_\phi(f) + \lambda\|f\|_{\mathcal{H}}^2$.

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
    $\phi(\alpha) = \log(1 + \exp(-2\alpha))$.

Large Margin Algorithms



Statistical Consequences of Using a Convex Cost

- Bayes risk consistency? For which ϕ?
  - (Lugosi and Vayatis, 2003), (Mannor, Meir and Zhang, 2002): regularized boosting.
  - (Zhang, 2003), (Steinwart, 2003): SVM.
  - (Jiang, 2003): boosting with early stopping.

Statistical Consequences of Using a Convex Cost

- How is risk related to ϕ-risk?
  - (Lugosi and Vayatis, 2003), (Steinwart, 2003): asymptotic.
  - (Zhang, 2003): comparison theorem.
- Convergence rates? With low noise?
  - (Tsybakov, 2001): empirical risk minimization.

Overview

- Relating excess risk to excess ϕ-risk.
  - ψ-transform: best possible bound.
  - proof idea.
  - conditions on ϕ.
- Low noise.
  - tighter bounds.
  - fast estimation rates.

Definitions and Facts

$$\begin{aligned} R(f) &= \Pr\left(\operatorname{sign}(f(X)) \neq Y\right) && \text{Risk,} \\ R^* &= \inf_f R(f) && \text{Bayes risk,} \\ \eta(x) &= \Pr(Y = 1 | X = x) && \text{conditional probability.} \end{aligned}$$

- η defines an optimal classifier:

$$R^* = R(\operatorname{sign}(\eta(x) - 1/2)).$$

- Excess risk of $f : \mathcal{X} \to \mathbb{R}$ is

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

Definitions

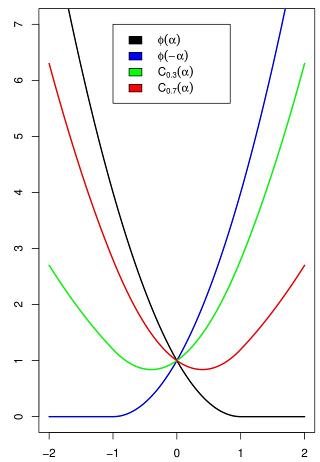
$$\begin{aligned} \text{Risk:} \qquad & R(f) = \Pr\left(\operatorname{sign}(f(X)) \neq Y\right). \\ \phi\text{-Risk:} \qquad & R_\phi(f) = \mathbf{E}\phi(Yf(X)). \end{aligned}$$

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

Conditional ϕ-risk:

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

Conditional ϕ-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

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

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

Conditional ϕ-risk:

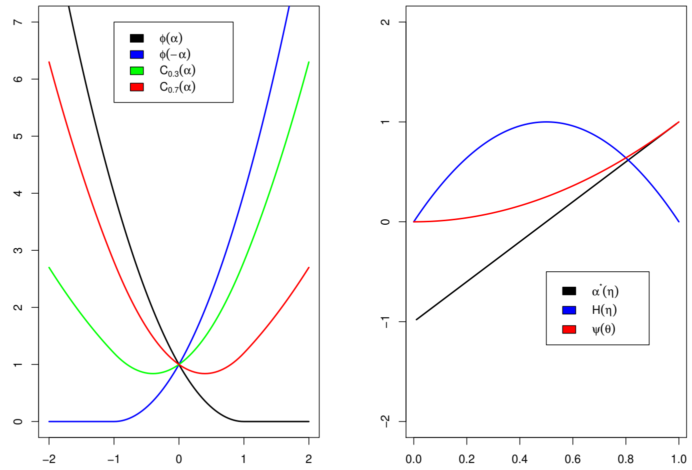
$$\mathbf{E}\left[\phi(Y f(X)) | X = x\right] = \eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)).$$

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

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

$$R_\phi^* = \mathbf{E}H(\eta(X)).$$

Optimal Conditional ϕ-risk: Example



Definitions

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

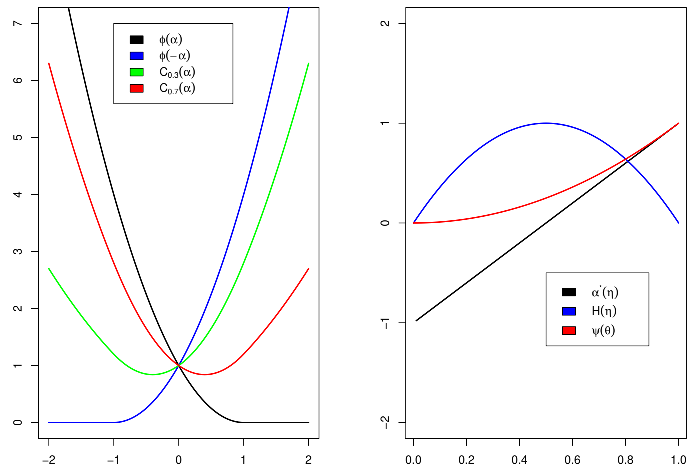
$$H(\eta) = \inf_{\alpha \in \mathbb{R}} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right).$$

Optimal conditional ϕ-risk with incorrect sign:

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

Note: $\qquad H^-(\eta) \geq H(\eta) \qquad H^-(1/2) = H(1/2).$

Example: $H^-(\eta) = \phi(0)$



Definitions

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

Definition: 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. (c.f. Lin (2001))

Definitions

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

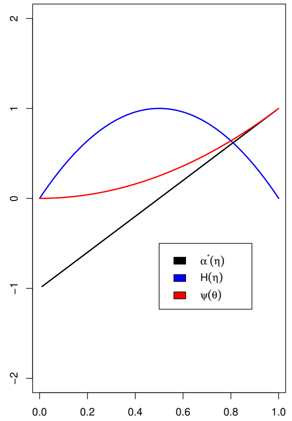
$$\text{epi}(g^{**}) = \overline{\text{co}}(\text{epi}(g)),$$
$$\text{epi}(g) = \{(x,y) : x \in [0,1],\ g(x) \le y\}.$$

ψ-transform: Example

- ψ is the best convex lower bound on $\tilde{\psi}(\theta) = H^-((1+\theta)/2) - H((1+\theta)/2)$, the excess conditional ϕ-risk when the sign is incorrect.
- $\psi = \tilde{\psi}^{**}$ is the biconjugate of $\tilde{\psi}$,

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

- ψ is the functional convex hull of $\tilde{\psi}$.



The Relationship between Excess Risk and Excess ϕ-risk

Theorem:

1. For any P and f, $\psi(R(f) - R^*) \le R_\phi(f) - R^*_\phi$.
2. This bound cannot be improved.
3. Near-minimal ϕ-risk implies near-minimal risk precisely when ϕ is classification-calibrated.

The Relationship between Excess Risk and Excess ϕ-risk

Theorem:

1. For any P and f, $\quad \psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^*$.
2. This bound cannot be improved:
   For $|\mathcal{X}| \geq 2$, $\epsilon > 0$ and $\theta \in [0, 1]$, there is a P and an f with
$$R(f) - R^* = \theta$$
$$\psi(\theta) \leq R_\phi(f) - R_\phi^* \leq \psi(\theta) + \epsilon.$$
3. Near-minimal ϕ-risk implies near-minimal risk precisely when ϕ is classification-calibrated.

The Relationship between Excess Risk and Excess ϕ-risk

Theorem:

1. For any P and f, $\quad \psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^*$.
2. This bound cannot be improved.
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 $R(f_i) \to R^*$.

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

Excess Risk Bounds: Proof Idea

Recall:

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

Thus,

$$\begin{aligned}
&\psi(R(f) - R^*) \qquad\qquad (\psi \text{ convex}, \psi(0) = 0)\\
&\leq \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \psi\left(|2\eta(X) - 1|\right)\right)\\
&\leq \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \tilde{\psi}\left(|2\eta(X) - 1|\right)\right)\\
&= \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \left(H^-(\eta(X)) - H(\eta(X))\right)\right)\\
&\leq \mathbf{E}\left(\phi(Y f(X)) - H(\eta(X))\right)\\
&= R_\phi(f) - R_\phi^*.
\end{aligned}$$

Excess Risk Bounds: Proof Idea

Recall:

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

Thus,

$$\begin{aligned}
&\psi(R(f) - R^*) && (\psi \le \tilde{\psi}) \\
&\le \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \psi\left(|2\eta(X) - 1|\right)\right) \\
&\le \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \tilde{\psi}\left(|2\eta(X) - 1|\right)\right) \\
&= \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \left(H^-(\eta(X)) - H(\eta(X))\right)\right) \\
&\le \mathbf{E}\left(\phi(Yf(X)) - H(\eta(X))\right) \\
&= R_\phi(f) - R_\phi^*.
\end{aligned}$$

Excess Risk Bounds: Proof Idea

Recall:

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

Thus,

$$\begin{aligned}
&\psi(R(f) - R^*) && (\text{definition of } \tilde{\psi}) \\
&\le \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \psi\left(|2\eta(X) - 1|\right)\right) \\
&\le \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \tilde{\psi}\left(|2\eta(X) - 1|\right)\right) \\
&= \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \left(H^-(\eta(X)) - H(\eta(X))\right)\right) \\
&\le \mathbf{E}\left(\phi(Yf(X)) - H(\eta(X))\right) \\
&= R_\phi(f) - R_\phi^*.
\end{aligned}$$

Excess Risk Bounds: Proof Idea

Recall:

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

Thus,

$$\begin{aligned}
&\psi(R(f) - R^*) && (H^- \text{ minimizes conditional } \phi\text{-risk}) \\
&\le \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \psi\left(|2\eta(X) - 1|\right)\right) \\
&\le \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \tilde{\psi}\left(|2\eta(X) - 1|\right)\right) \\
&= \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \left(H^-(\eta(X)) - H(\eta(X))\right)\right) \\
&\le \mathbf{E}\left(\phi(Yf(X)) - H(\eta(X))\right) \\
&= R_\phi(f) - R_\phi^*.
\end{aligned}$$

Excess Risk Bounds: Proof Idea

Recall:

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

Thus,

$$\begin{aligned}
&\psi(R(f) - R^*) && (\text{definition of } R_\phi) \\
&\le \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \psi\left(|2\eta(X) - 1|\right)\right) \\
&\le \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \tilde{\psi}\left(|2\eta(X) - 1|\right)\right) \\
&= \mathbf{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \left(H^-(\eta(X)) - H(\eta(X))\right)\right) \\
&\le \mathbf{E}\left(\phi(Yf(X)) - H(\eta(X))\right) \\
&= R_\phi(f) - R_\phi^*.
\end{aligned}$$

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 θ_1 and θ_2 so that $\psi(\theta_i) = \tilde{\psi}(\theta_i)$ and $\theta \in \mathrm{co}\{\theta_1, \theta_2\}$.
   Set $\eta(x_1) = (1+\theta_1)/2$ and $\eta(x_2) = (1+\theta_2)/2$.
   Again, each inequality is clearly tight.

Classification-calibrated ϕ

Theorem: If ϕ is convex,

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

Theorem: If ϕ is classification calibrated,
$\exists \gamma > 0, \forall \alpha \in \mathbb{R}$,

$$\gamma \phi(\alpha) \geq \mathbf{1}[\alpha \leq 0].$$

Overview

- Relating excess risk to excess ϕ-risk.
  - ψ-transform: best possible bound.
  - proof idea.
  - conditions on ϕ.
- Low noise.
  - tighter bounds.
  - fast estimation rates.

Low Noise

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

$$\Pr\left(f(X)(\eta(X) - 1/2) < 0\right) \leq c\left(R(f) - R^*\right)^{\beta}.$$

Recall:
$$\begin{aligned} R(f) - R^* &= \mathbf{E}\left(\mathbf{1}[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)]\,|2\eta(X) - 1|\right) \\ &\leq P\left(\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right). \end{aligned}$$

Hence, $0 \leq \beta \leq 1$.

- $\beta = 0$ is no constraint.
- $\beta = 1$ iff $\Pr\left(0 < |2\eta(X) - 1| < 1/c\right) = 0$.
- $\Pr\left(0 < |2\eta(X) - 1| < \epsilon\right) \leq C\epsilon^{\alpha}$ implies noise exponent $\beta = \dfrac{\alpha}{1+\alpha}$.

Low Noise

- Tsybakov considered empirical risk minimization.
  (But ERM is typically hard)
- With:
  - the noise assumption,
  - the Bayes classifier in the function class

  the empirical risk minimizer has (true) risk converging suprisingly quickly to the minimum. (Tsybakov, 2001)

Risk Bounds with Low Noise

Theorem: If P has noise exponent $0 \leq \beta \leq 1$,
then there is a $c > 0$ such that for any $f : \mathcal{X} \to \mathbb{R}$,

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

Notice that we only improve the rate, since the convexity of ψ implies

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

Risk Bounds with Low Noise

Note: Minimizing R_ϕ adapts to noise exponent:
lower noise implies closer relationship between risk and ϕ-risk.

Proof idea

Split $\mathcal{X}$:

1. Low noise region ($|\eta(X) - 1/2| > \epsilon$): bound risk using noise assumption.
2. High noise ($\leq \epsilon$): bound risk as before.

Fast Convergence Rates for Large Margin Classifiers

- Minimize empirical ϕ-risk over a class $\mathcal{F}$:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}_\phi(f).$$

- $R(\hat{f}) - R^*$ decreases with $R_\phi(\hat{f}) - R_\phi^*$.
  (Faster decrease with low noise.)
- How rapidly does $R_\phi(\hat{f})$ converge?

Fast Convergence Rates for Large Margin Classifiers

Assume that ϕ satisfies

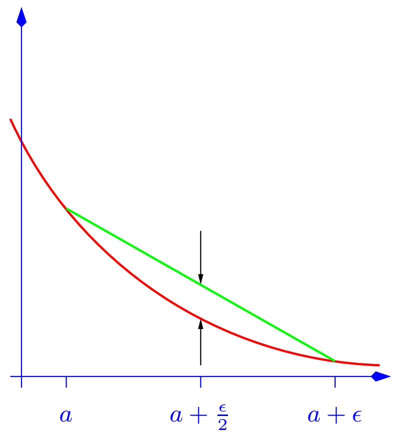
1. A Lipschitz condition:

$$\text{for all } a, b \in \mathbb{R}, |\phi(a) - \phi(b)| \leq L|a - b|.$$

2. A strict convexity condition: the modulus of convexity of ϕ satisfies $\delta_\phi(\epsilon) \geq \epsilon^r$, where

$$\delta_\phi(\epsilon) = \inf \left\{ \frac{\phi(\alpha_1) + \phi(\alpha_2)}{2} - \phi\left(\frac{\alpha_1 + \alpha_2}{2}\right) : |\alpha_1 - \alpha_2| \geq \epsilon \right\}.$$

Modulus of Convexity



Fast Convergence Rates for Strictly Convex ϕ, Convex $\mathcal{F}$

Theorem: Suppose that:

- ϕ is Lipschitz with constant L.
- ϕ has modulus of convexity $\delta_\phi(\epsilon) \geq \epsilon^r$. (Set $\alpha = \max(1, 2 - 2/r)$.)
- $\mathcal{F}$ is a convex set of uniformly bounded functions.
- $\mathcal{F}$ is finite dimensional $(\sup_P \log \mathcal{N}(\epsilon, \mathcal{F}, L_2(P)) \leq d \log(1/\epsilon))$.

Then with probability at least $1 - \delta$, the minimizer $\hat{f} \in \mathcal{F}$ of $\hat{R}_\phi$ satisfies

$$R_\phi(\hat{f}) - \inf_{f \in \mathcal{F}} R_\phi(f) \leq c \left(\frac{d \log n + \log(1/\delta)}{n} \right)^{1/\alpha}.$$

Fast Convergence Rates for Strictly Convex ϕ, Convex $\mathcal{F}$

The key idea:

Strict convexity ensures that the variance of the excess ϕ-loss is controlled.

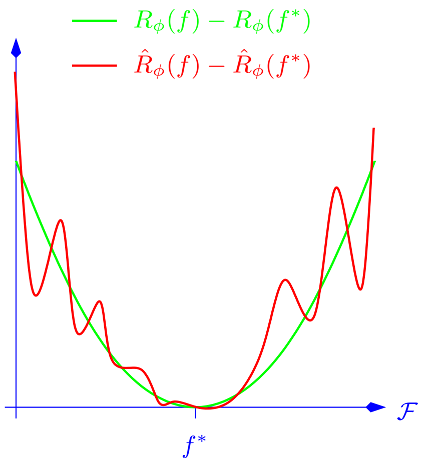
Define $f^* = \arg\min_{f \in \mathcal{F}} R_\phi(f)$.
For $f \in \mathcal{F}$, define the excess ϕ-loss as

$$g_f(x, y) = \phi(yf(x)) - \phi(yf^*(x)).$$

Theorem: If ϕ is Lipschitz with constant L and uniformly convex with modulus of convexity $\delta_\phi(\epsilon) \geq \epsilon^r$, then for any f in a convex set $\mathcal{F}$,

$$\mathbf{E} g_f^2 \leq L^2 \mathbf{E} (f - f^*)^2 \leq L^2 \left(\frac{\mathbf{E} g_f}{2} \right)^{\min(1, 2/r)}.$$

Fast Convergence Rates for Strictly Convex ϕ



An Aside: Tsybakov's Condition Revisited

Definition: [Tsybakov] The distribution P on $\mathcal{X} \times \{\pm 1\}$ has *noise exponent* $0 \le \beta \le 1$ if there is a $c > 0$ such that every $f : \mathcal{X} \to \{\pm 1\}$ has

$$\Pr\left(f(X)(\eta(X) - 1/2) < 0\right) \le c\left(R(f) - R^*\right)^\beta.$$

This is the variance condition:

- Bayes classifier is in $\mathcal{F}$; set $f^* = \mathrm{sign}(\eta - 1/2)$.
- $\mathbf{E} g_f^2 = \Pr\left(f(X)(\eta(X) - 1/2) < 0\right)$.
- $\mathbf{E} g_f = R(f) - R^*$.
- $\Longrightarrow$ Assumption is equivalent to $\mathbf{E} g_f^2 \le c\left(\mathbf{E} g_f\right)^\beta$. Fast rates follow.

Risk Bounds with Low Noise: Examples

- Adaboost: $\phi(\alpha) = e^{-\alpha}$.
- SVM with 2-norm soft-margin penalty: $\phi(\alpha) = (\max(0, 1 - \alpha))^2$.
- Quadratic loss: $\phi(\alpha) = (1 - \alpha)^2$.

All of these satisfy:

- convex.
- classification calibrated.
- quadratic modulus of convexity, δ_ϕ.
- quadratic ψ.

Risk Bounds with Low Noise

Theorem: If ϕ has

- modulus of convexity $\delta_\phi(\alpha) \ge \alpha^2$,
- noise exponent $= 1$ (that is, $|\Pr(Y = 1|X) - 1/2| \ge c_1$), and
- $\mathcal{F}$ is d-dimensional,

then with probability at least $1 - \delta$, the minimizer $\hat{f}$ of $\hat{L}_\phi$ satisfies

$$R(\hat{f}) - R^* \le c\left(d\frac{\log n/\delta}{n} + \inf_{f \in \mathcal{F}} R_\phi(f) - R_\phi^*\right).$$

(And there are similar fast rates for larger classes.)

Summary: Large Margin Classifiers

- Risk bounds relative to Bayes risk:
  - ψ relates excess risk to excess ϕ-risk.
  - Best possible.
- Fast rates with low noise:
  - Tighter bound on excess risk.
  - Fast convergence of ϕ-risk for strictly convex ϕ.

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

References

[Bartlett, Jordan and McAuliffe(2003)] P. L. Bartlett, M. I. Jordan and J. D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, U.C. Berkeley [www.stat.berkeley.edu/tech-reports], 2003.

[Jiang(2003)] W. Jiang. Process consistency for Adaboost. *Annals of Statistics*, 2003 (To appear).

[Lee, Bartlett, and Williamson(1996)] W. S. Lee, P. L. Bartlett, and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

[Lin(2001)] Y. Lin. A note on margin-based loss functions in classification. Technical Report 1044r, Department of Statistics, University of Wisconsin, 2001.

[Lugosi and Vayatis(2003)] G. Lugosi and N. Vayatis. On the Bayes risk consistency of regularized boosting methods. *Annals of Statistics*, 2003 (To appear).

[Mannor, Meir, and Zhang(2002)] S. Mannor, R. Meir, and T. Zhang. The consistency of greedy algorithms for classification. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 319–333, 2002.

[Mendelson(2002)] S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.

[Steinwart(2002)] I. Steinwart. Consistency of support vector machines and other regularized classifiers. Technical Report 02-03, University of Jena, Department of Mathematics and Computer Science, 2002.

[Tsybakov(2001)] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. Technical Report PMA-682, Université Paris VI, 2001.

[Zhang(2003)] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 2003 (To appear).