Large Margin Classifiers: Convexity and Classification

Peter Bartlett
Division of Computer Science and Department of Statistics
UC Berkeley

Joint work with
Mike Collins, Mike Jordan, David McAllester,
Jon McAuliffe, Ben Taskar, Ambuj Tewari.

slides at http://www.cs.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, $Yf(X)$.
- 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.)

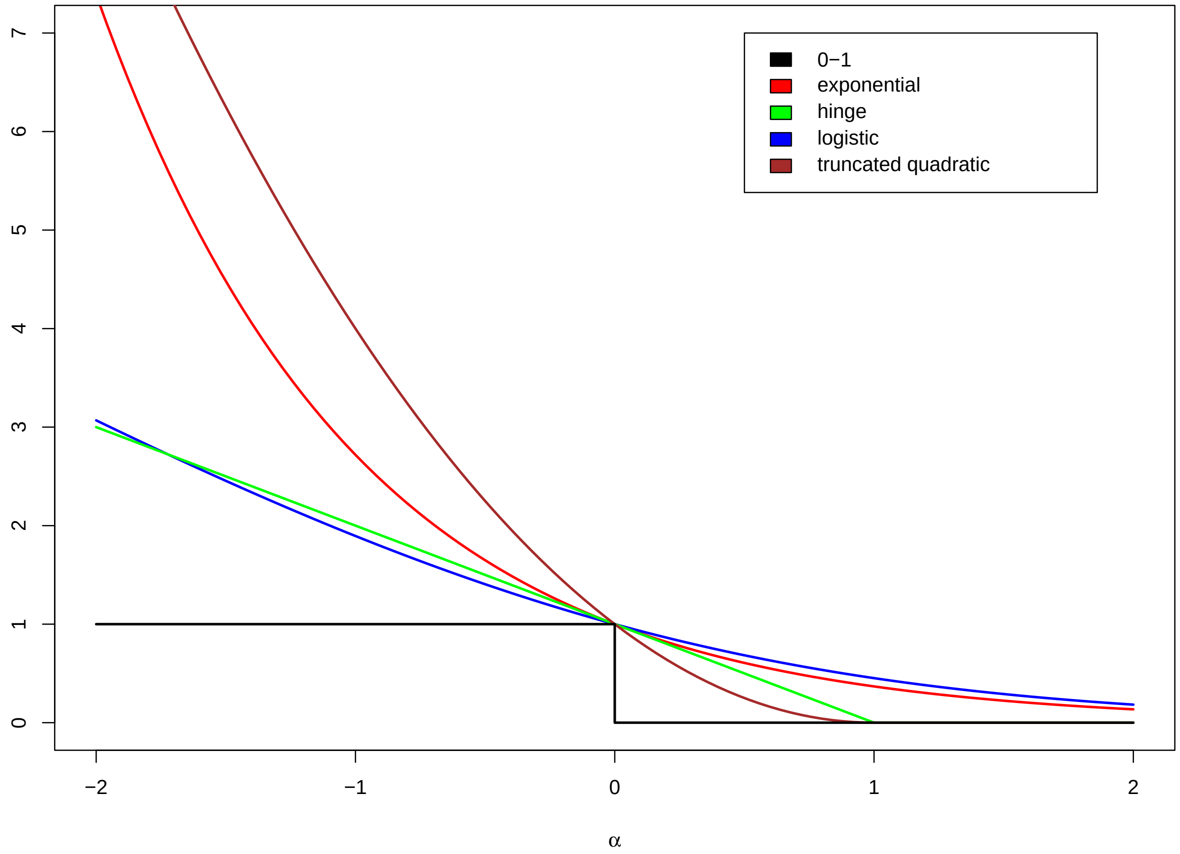
Large Margin Algorithms

- 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.
- 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, 2004), (Mannor, Meir and Zhang, 2002): regularized boosting.
  - (Zhang, 2004), (Steinwart, 2003): SVM.
  - (Jiang, 2004): boosting with early stopping.

Statistical Consequences of Using a Convex Cost

- How is risk related to ϕ-risk?
  - (Lugosi and Vayatis, 2004), (Steinwart, 2003): asymptotic.
  - (Zhang, 2004): comparison theorem.
- Convergence rates? With low noise?
  - (Tsybakov, 2001): empirical risk minimization.
- Estimating conditional probabilities?
- Multiclass?

Overview

- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition and universal consistency.
- Convergence rates: low noise.
- Kernel classifiers: sparseness versus probability estimation.
- Structured multiclass classification.

Definitions and Facts

$$R(f) = \Pr\left(\operatorname{sign}(f(X)) \neq Y\right) \qquad \text{Risk,}$$

$$R^* = \inf_f R(f) \qquad \text{Bayes risk,}$$

$$\eta(x) = \Pr(Y = 1 | X = x) \qquad \text{conditional probability.}$$

- η defines an optimal classifier:

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

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

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

Definitions

Risk: $R(f) = \Pr\left(\mathrm{sign}(f(X)) \neq Y\right).$

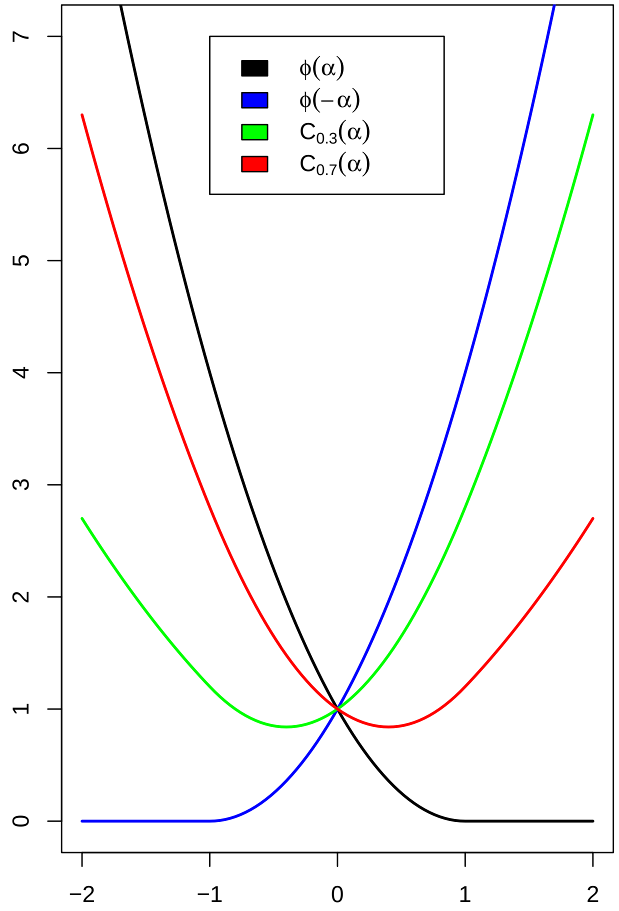
ϕ-Risk: $R_\phi(f) = \mathbb{E}\phi(Yf(X)).$

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

Conditional ϕ-risk:

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

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

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

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

Conditional ϕ-risk:

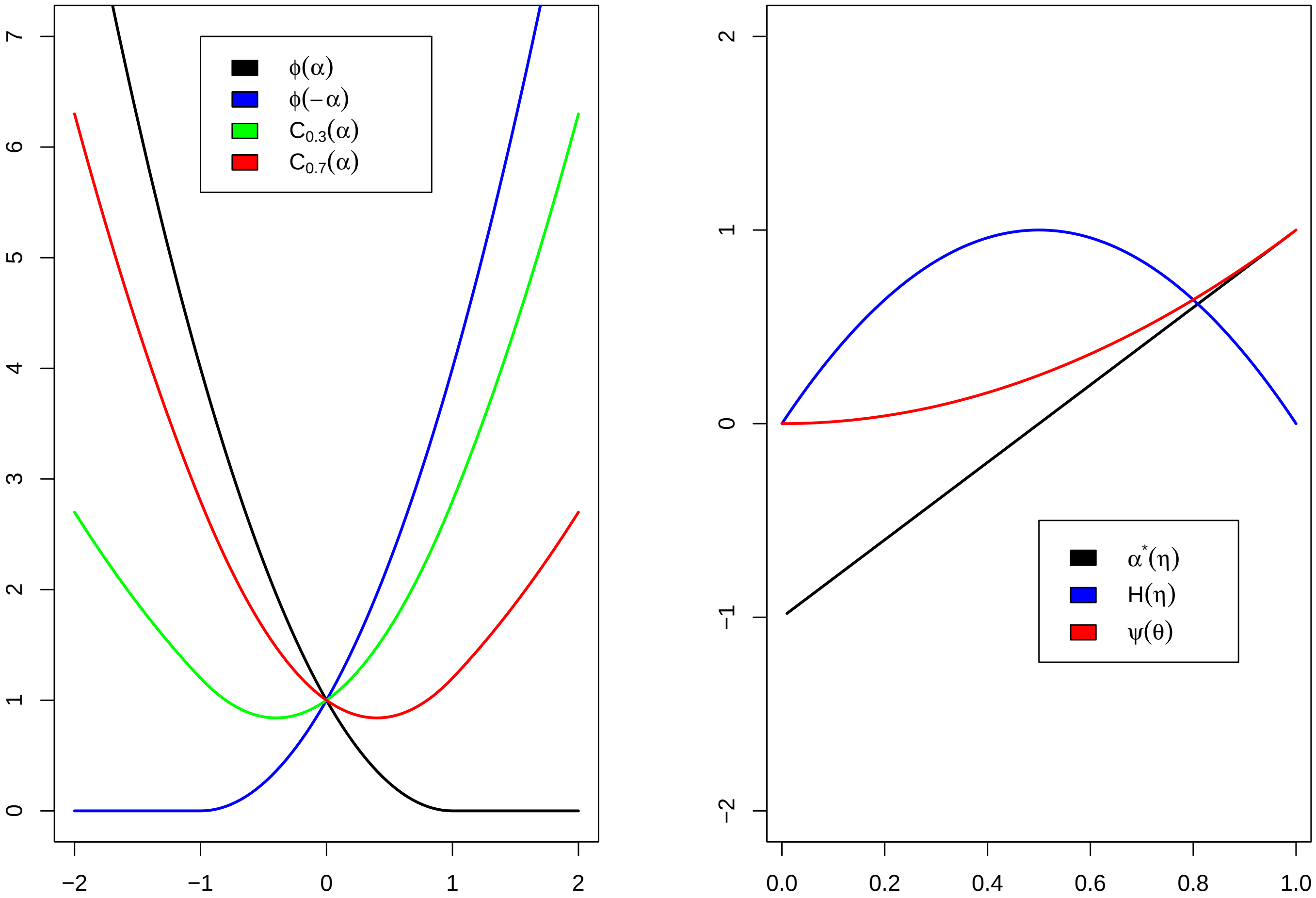
$$\mathbb{E}\left[\phi(Yf(X))|X = x\right] = \eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)).$$

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

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

$$R^*_\phi = \mathbb{E}H(\eta(X)).$$

Optimal Conditional ϕ-risk: Example



Definitions

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

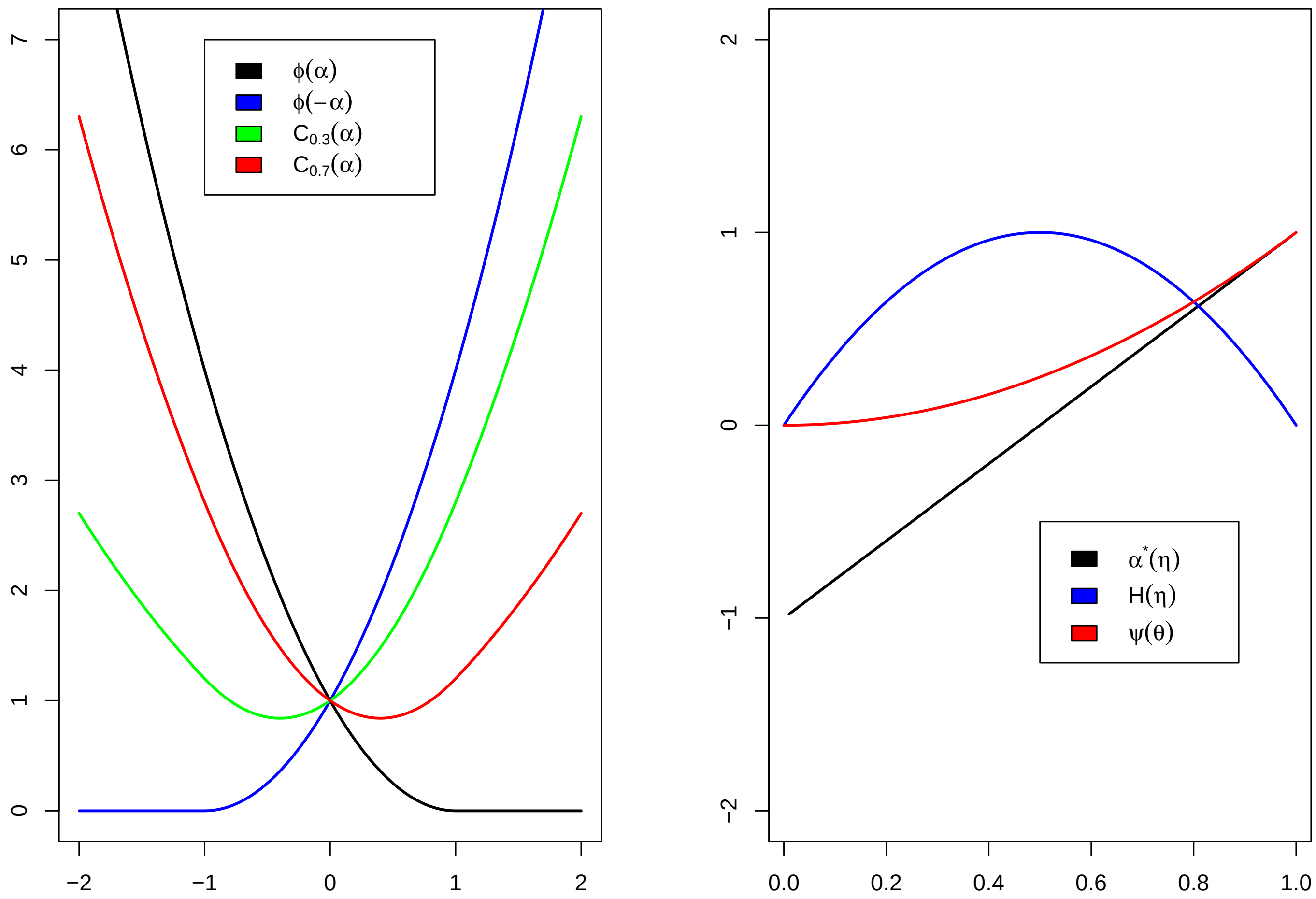
$$H(\eta) = \inf_{\alpha \in \mathbb{R}} \left(\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha) \right).$$

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

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,

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

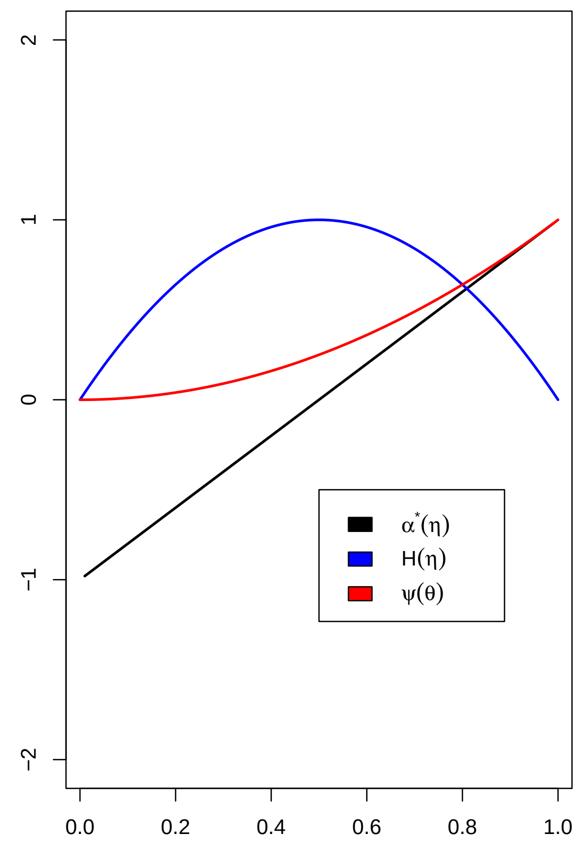
ψ-transform: Example

- 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 $\tilde{\psi}$,

$$\mathrm{epi}(\psi) = \overline{\mathrm{co}}(\mathrm{epi}(\tilde{\psi})),$$

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

- 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, $\psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^*$.
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.
- $\psi(\theta) \le \tilde{\psi}(\theta) = H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)$.

Excess Risk Bounds: Proof Idea

Recall:

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

Thus,

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

Excess Risk Bounds: Proof Idea

Recall:

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

Thus,

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

Excess Risk Bounds: Proof Idea

Recall:

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

Thus,

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

Excess Risk Bounds: Proof Idea

Recall:

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

Thus,

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

Excess Risk Bounds: Proof Idea

Recall:

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

Thus,

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

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

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

Overview

- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition and universal consistency.
- Convergence rates: low noise.
- Kernel classifiers: sparseness versus probability estimation.
- Structured multiclass classification.

The Approximation/Estimation Decomposition

Algorithm chooses

$$f_n = \arg \min_{f \in \mathcal{F}_n} \hat{E}_n R_\phi(f) + \lambda_n \Omega(f).$$

We can decompose the excess risk estimate as

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

The Approximation/Estimation Decomposition

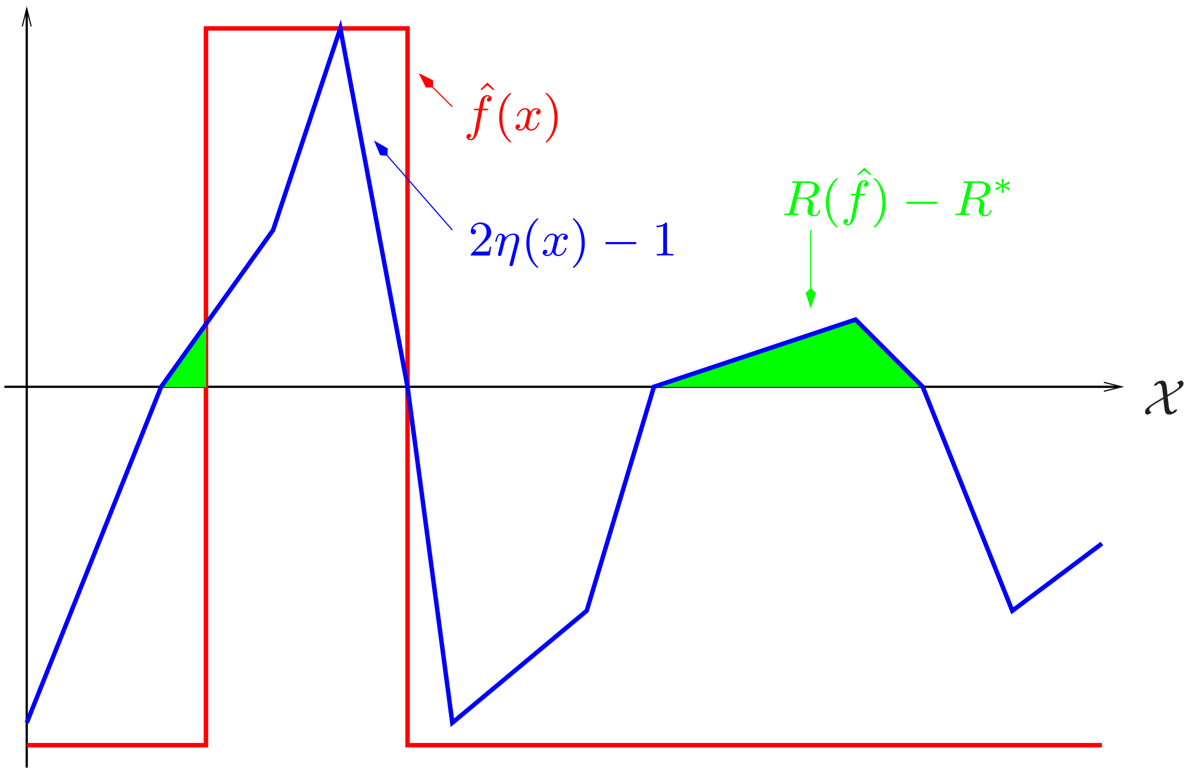
$$\psi\left(R(f_n) - R^*\right) \leq R_\phi(f_n) - R^*_\phi$$
$$= \underbrace{R_\phi(f_n) - \inf_{f \in \mathcal{F}_n} R_\phi(f)}_{\text{estimation error}} + \underbrace{\inf_{f \in \mathcal{F}_n} R_\phi(f) - R^*_\phi}_{\text{approximation error}} .$$

- 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.

Overview

- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition and universal consistency.
- Convergence rates: low noise.
- Kernel classifiers: sparseness versus probability estimation.
- Structured multiclass classification.

Low Noise



Low Noise

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

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

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

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

where $\beta = \dfrac{\alpha}{1+\alpha}$.

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

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 α,
then there is a $c > 0$ such that for any $f : \mathcal{X} \to \mathbb{R}$,

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

where $\beta = \dfrac{\alpha}{1+\alpha} \in [0, 1]$.

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

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

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

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

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

Fast Convergence Rates for Large Margin Classifiers

Assume that ϕ satisfies

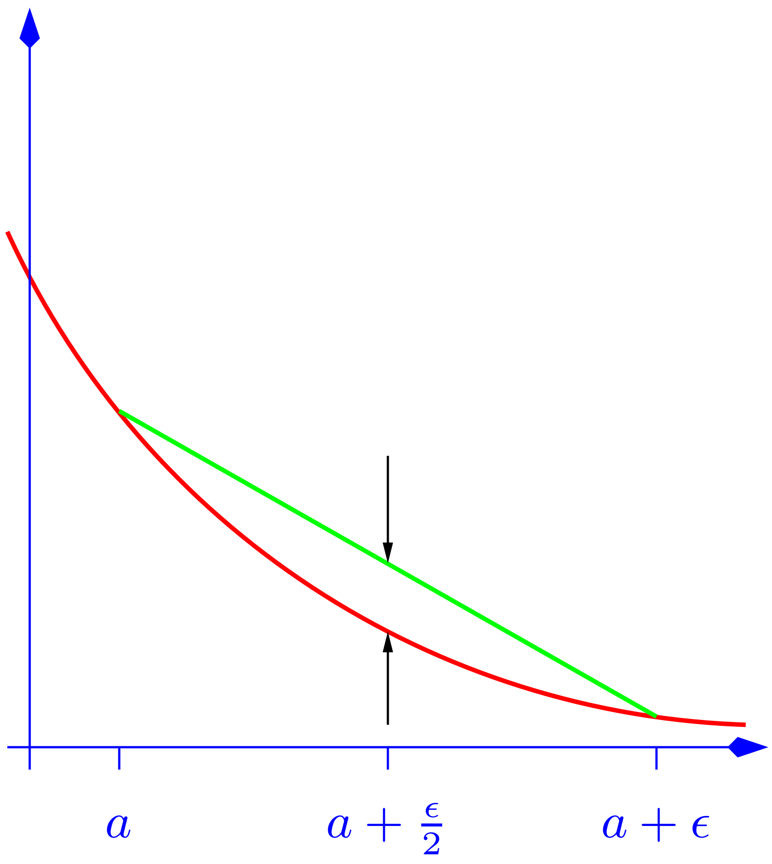
1. A Lipschitz condition:

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

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.

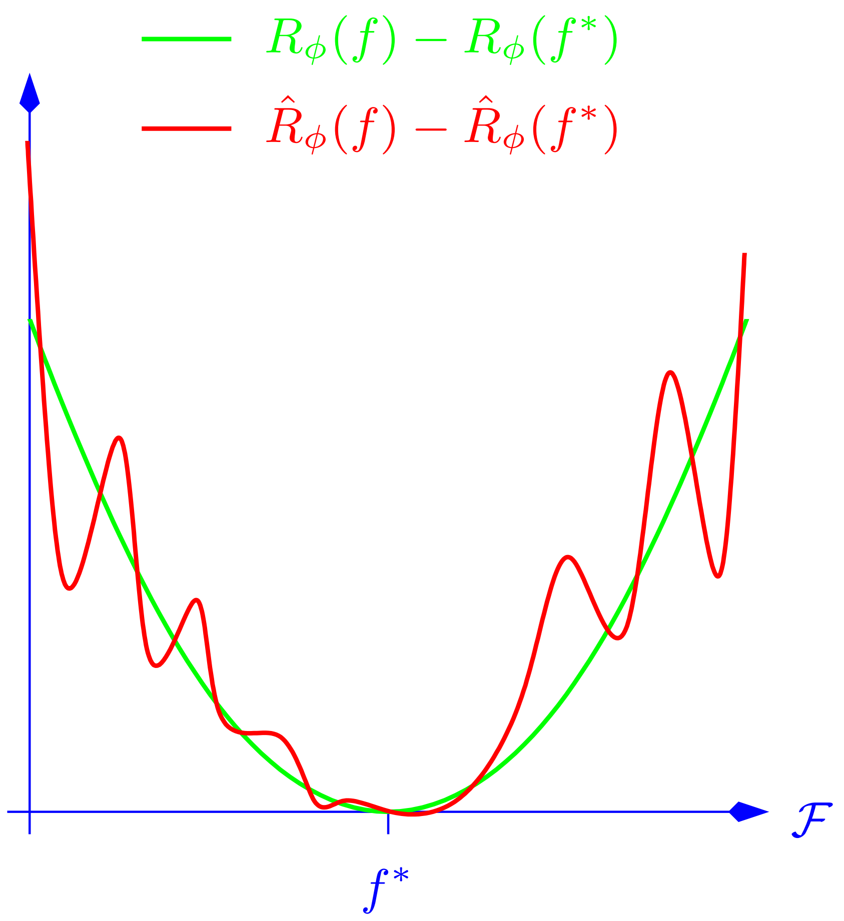
Define $f^* = \arg\min_{f \in \mathcal{F}} R_\phi(f)$.
For $f \in \mathcal{F}$, define the excess ϕ-loss as

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

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 $\mathcal{F}$,

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

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* α if there is a $c > 0$ such that every $f : \mathcal{X} \to \{\pm 1\}$ has

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

where $\beta = \dfrac{\alpha}{1+\alpha} \in [0, 1]$.

This is the variance condition:

- Bayes classifier is in $\mathcal{F}$; set $f^* = \mathrm{sign}(\eta - 1/2)$.
- $\mathbb{E}g_f^2 = \Pr\left(f(X)(\eta(X) - 1/2) < 0\right)$.
- $\mathbb{E}g_f = R(f) - R^*$.
- $\implies$ Assumption is equivalent to $\mathbb{E}g_f^2 \leq c\left(\mathbb{E}g_f\right)^{\beta}$. 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 $\delta_\phi(\alpha) \geq \alpha^2$,
- noise exponent $= \infty$ (that is, $|\Pr(Y = 1|X) - 1/2| \geq c_1$), and
- $\mathcal{F}$ is d-dimensional,

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

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

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

Summary: Large Margin Classifiers

- Relating excess risk to excess ϕ-risk:
  - ψ relates excess risk to excess ϕ-risk.
  - Best possible.
- The approximation/estimation decomposition and universal consistency.
- Convergence rates: low noise.
  - Tighter bound on excess risk.
  - Fast convergence of ϕ-risk for strictly convex ϕ.

Overview

- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition and universal consistency.
- Convergence rates: low noise.
- Kernel classifiers: sparseness versus probability estimation.
- Structured multiclass classification.

Kernel Methods for Classification

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

where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), with norm $\|\cdot\|$, and $\lambda_n > 0$ is a regularization parameter.

Example:

$$\text{L1-SVM:} \quad \phi(\alpha) = (1-\alpha)_+$$

$$\text{L2-SVM:} \quad \phi(\alpha) = \left((1-\alpha)_+\right)^2.$$

Kernel Methods for Classification

$$\left.\begin{array}{l}\text{support of } P \text{ in } \{x : k(x,x) \le B\}.\\ \lambda_n \to 0, \text{ suitably slowly.}\\ \phi \text{ locally Lipschitz.}\end{array}\right\} \Rightarrow \quad R_\phi(f_n) \to \inf_{f \in \mathcal{H}} R_\phi(f).$$

RKHS suitably rich $\Rightarrow$ $\inf_{f \in \mathcal{H}} R_\phi(f) = R_\phi^*.$

ϕ classification calibrated $\Rightarrow$ $R(f_n) \to R^*.$

i.e., a universal kernel, suitable ϕ, appropriate regularization schedule
$\Rightarrow$ universal consistency.

e.g., (Steinwart, 2001)

Overview

- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition and universal consistency.
- Convergence rates: low noise.
- Kernel classifiers
  - probability estimation
  - sparseness
- Structured multiclass classification.

Estimating Conditional Probabilities

Can we use a large margin classifier,

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

to estimate the conditional probability $\eta(x) = \Pr(Y = 1 | X = x)$?
Does $f_n(x)$ give information about $\eta(x)$, say, asymptotically?

- Confidence-rated predictions are of interest for many decision problems.
- Probabilities are useful for combining decisions.

Estimating Conditional Probabilities

If ϕ is convex, we can write

$$
\begin{aligned}
H(\eta) &= \inf_{\alpha \in \mathbb{R}} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha) \right) \\
&= \eta\phi(\alpha^*(\eta)) + (1-\eta)\phi(-\alpha^*(\eta)),
\end{aligned}
$$

where $\alpha^*(\eta) = \arg\min_{\alpha} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha) \right) \subset \mathbb{R} \cup \{\pm\infty\}$.

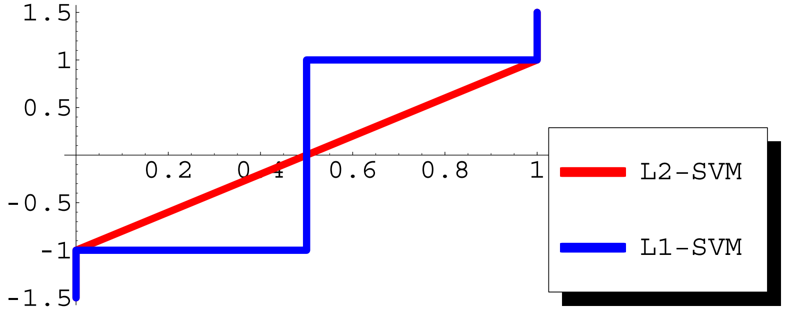
Recall:

$$
\begin{aligned}
R_\phi^* &= \mathbb{E}H(\eta(X)) = \mathbb{E}\phi(Y\alpha^*(\eta(X))) \\
\eta(x) &= \Pr(Y = 1 | X = x).
\end{aligned}
$$

Estimating Conditional Probabilities

$$\alpha^*(\eta) = \arg\min_{\alpha} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right) \subset \mathbb{R} \cup \{\pm\infty\}.$$

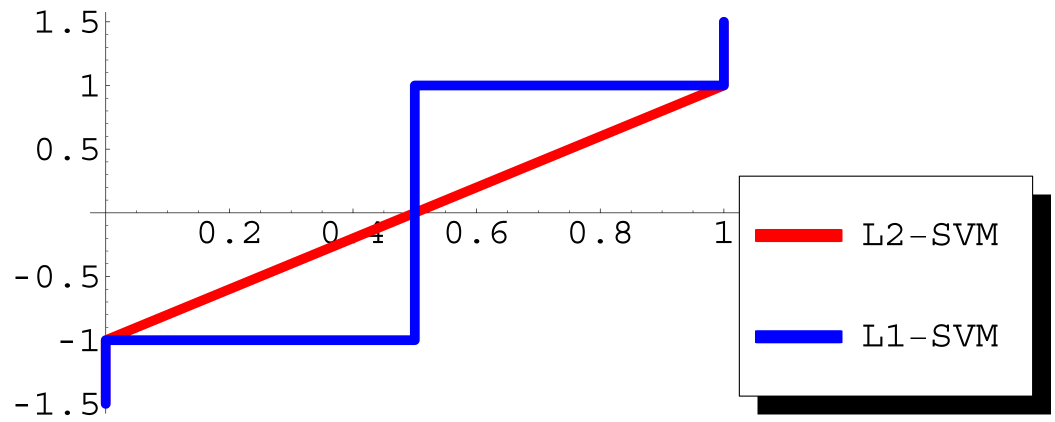
Examples of $\alpha^*(\eta)$ versus $\eta \in [0,1]$:



L2-SVM: $\phi(\alpha) = ((1-\alpha)_+)^2$

L1-SVM: $\phi(\alpha) = (1-\alpha)_+$.

Estimating Conditional Probabilities



If $\alpha^*(\eta)$ is not invertible, that is, there are $\eta_1 \neq \eta_2$ with

$$\alpha^*(\eta_1) \cap \alpha^*(\eta_2) \neq \emptyset,$$

then there are distributions P and functions f_n with $R_\phi(f_n) \to R_\phi^*$ but $f_n(x)$ cannot be used to estimate $\eta(x)$.

e.g., $f_n(x) \to \alpha^*(\eta_1) \cap \alpha^*(\eta_2)$. Is $\eta(x) = \eta_1$ or $\eta(x) = \eta_2$?

Overview

- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition and universal consistency.
- Convergence rates: low noise.
- Kernel classifiers: sparseness versus probability estimation
- Structured multiclass classification.

Sparseness

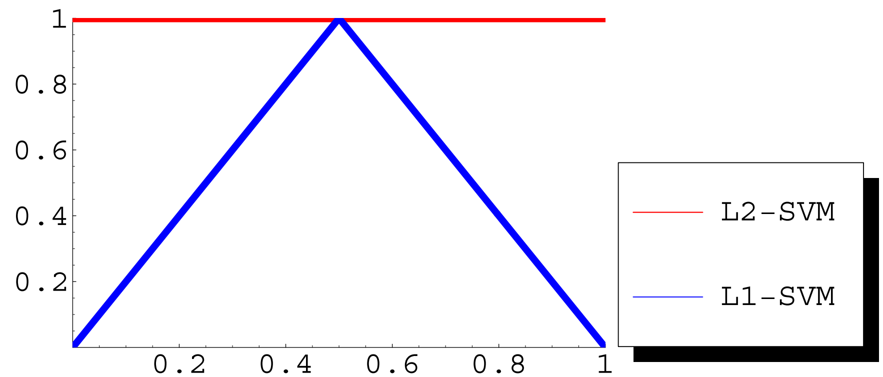
- Representer theorem: solution of optimization problem can be represented as:

$$f_n(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)\ .$$

- Inputs x_i with $\alpha_i \neq 0$ are called *support vectors* (SV's).
- Sparseness (number of support vectors $\ll n$) means faster evaluation of the classifier.

Sparseness: Steinwart's results

- For L1 and L2-SVM, Steinwart proved that the asymptotic fraction of SV's is $\mathbb{E}G(\eta(X))$ (under some technical assumptions).
- The function $G(\eta)$ depends on the loss function used:



- L2-SVM doesn't produce sparse solutions (asymptotically) while L1-SVM does.
- Recall: L2-SVM can estimate η while L1-SVM cannot.

Sparseness versus Estimating Conditional Probabilities

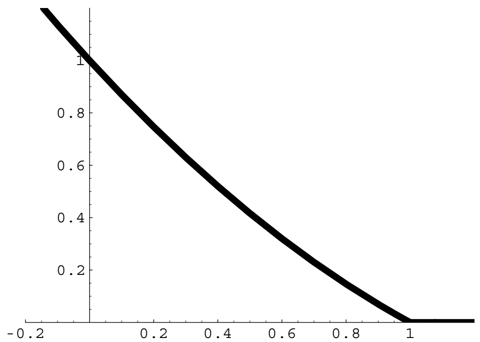
The ability to estimate conditional probabilities always causes loss of sparseness:

- Lower bound of the asymptotic fraction of data that become SV's can be written as $\mathbb{E}G(\eta(X))$.
- $G(\eta)$ is 1 throughout the region where probabilities can be estimated.
- The region where $G(\eta) = 1$ is an interval centered at $1/2$.

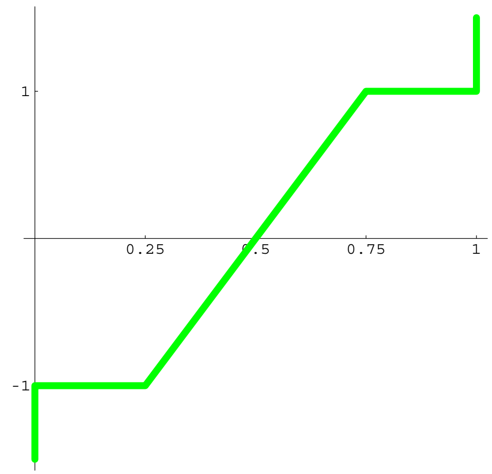
Example

- Steinwart's lower bound on the asymptotic fraction of SV's:
$$\Pr[\, 0 \notin \partial\phi(Y\alpha^*(\eta(X)))\,]$$

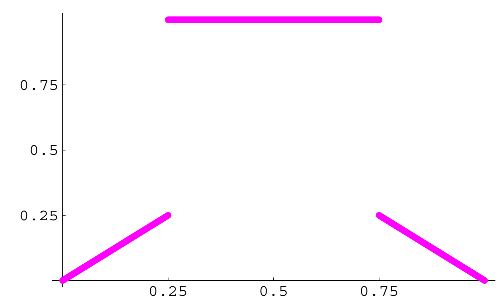
- Write the lower bound as $\mathbb{E}G(\eta(X))$ where
$$G(\eta) = \eta \mathbf{1}\left[0 \notin \partial\phi(\alpha^*(\eta))\right] + (1-\eta)\mathbf{1}\left[0 \notin \partial\phi(-\alpha^*(\eta))\right]$$



$\frac{1}{3}((1-t)_+)^2 + \frac{2}{3}(1-t)_+$



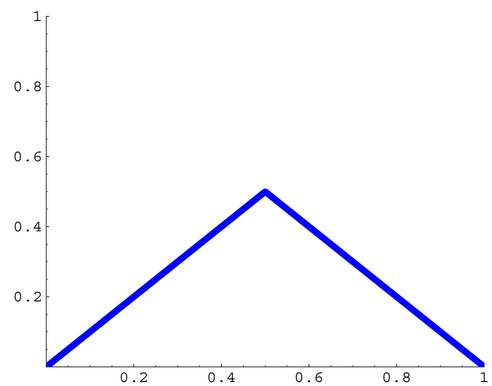
$\alpha^*(\eta)$ vs. η

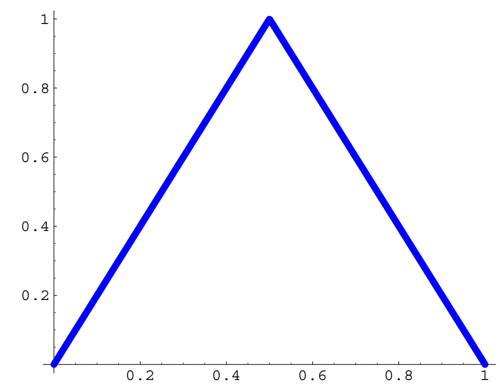


$G(\eta)$ vs. η

Sparseness vs. Estimating Probabilities

- In general, $G(\eta)$ is 1 on an interval around 1/2; outside that interval, $G(\eta) = \min\{\eta, 1-\eta\}$.
- We know this gives a loose lower bound for L1-SVM:





- Sharp bound can be derived for loss functions of the form:

$$\phi(t) = h((t_0 - t)_+)$$

where h is convex, differentiable and $h'(0) > 0$.

Asymptotically Sharp Result

- Recall that our classifier can be expressed as $\sum_i \alpha_i k(\cdot, x_i)$ and let $\#SV = |\{i : \alpha_i \neq 0\}|$.
- If the kernel k is *analytic* and *universal* (and the underlying P_X is continuous and non-trivial), then for a regularization sequence $\lambda_n \to 0$ sufficiently slowly:

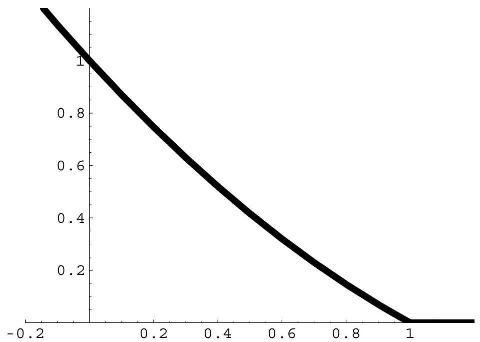
$$\frac{\#SV}{n} \xrightarrow{P} \mathbb{E}G(\eta(X))$$

where

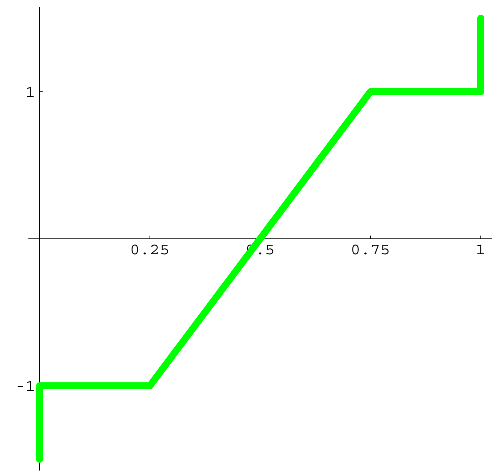
$$G(\eta) = \begin{cases} \eta/\gamma & 0 \leq \eta \leq \gamma \\ 1 & \gamma < \eta < 1 - \gamma \\ (1-\eta)/\gamma & 1 - \gamma \leq \eta \leq 1 \end{cases}$$

Example again

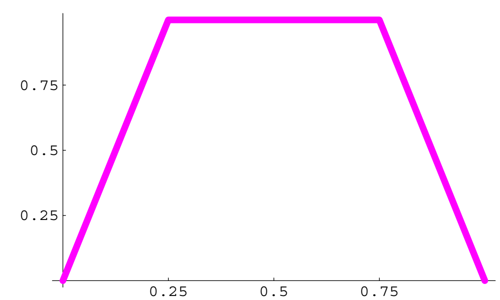
- γ is given by $\frac{-\phi'(t_0)}{-\phi'(t_0)-\phi'(-t_0)}$ and $\alpha^*(\eta)$ is invertible in the interval $(\gamma, 1-\gamma)$.
- Below $h(t) = \frac{1}{3}t^2 + \frac{2}{3}t$, $-\phi'(1) = \frac{2}{3}$, $-\phi'(-1) = 2$ and hence $\gamma = \frac{1}{4}$.



$\frac{1}{3}((1-t)_+)^2 + \frac{2}{3}(1-t)_+$



$\alpha^*(\eta)$ vs. η



$G(\eta)$ vs. η

Overview

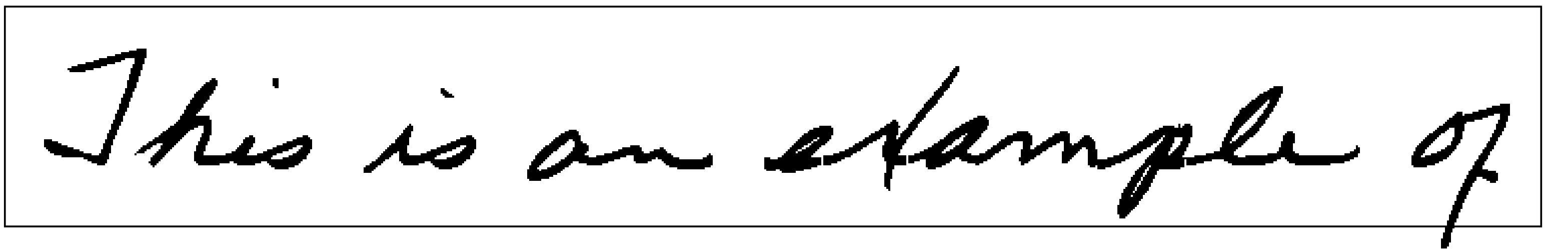
- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition and universal consistency.
- Convergence rates: low noise.
- Kernel classifiers
  - No sparseness where $\alpha^*(\eta)$ is invertible.
  - Can design ϕ to trade off sparseness and probability estimation.
- Structured multiclass classification.

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

Structured Classification: Optical Character Recognition

$X =$ grey-scale image of a sequence of characters

$Y =$ sequence of characters



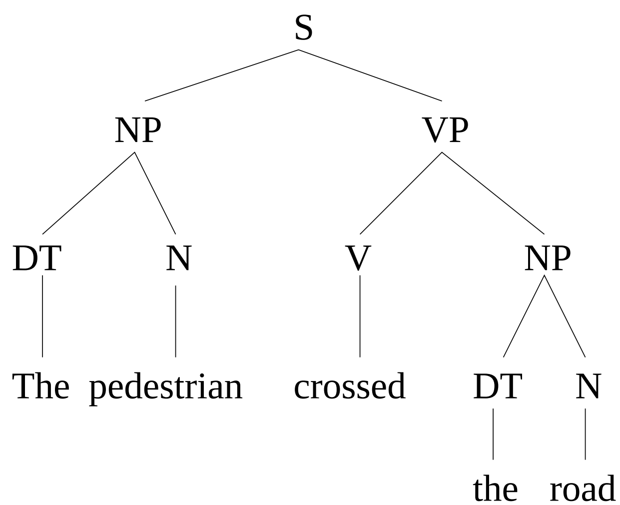
This is an example of

Structured Classification: Parsing

$X =$ sentence

$Y =$ parse tree

The pedestrian crossed the road.



Structured Classification

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

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

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

Structured Classification Problems

Key issue: $|\mathcal{Y}|$ is very large.

- OCR: exponential in number of characters
- parsing: exponential in sentence length

Generative Modelling:

- Split Y into parts/assume sparse dependencies.
  (e.g., graphical model; probabilistic context-free grammar.)
- Plug-in estimate:
  1. Simple model $\hat{p}(x, y; \theta)$ of $\Pr(Y = y | X = x)$.
  2. Use data to estimate parameters θ. (e.g., ML)
  3. Compute $\arg\max_{y \in \mathcal{Y}} \hat{p}(x, y; \theta)$. (e.g., dynamic programming)

Generative Model

If each factor is a log-linear model, we compute a linear discriminant:

$$\begin{aligned}\hat{y} &= \arg\max_{y \in \mathcal{Y}} \log(\hat{p}(x, y; \theta)) \\ &= \arg\max_{y \in \mathcal{Y}} \sum_i g_i(x, y)\theta_i.\end{aligned}$$

Structured Classification Problems: Sparse Representations

Suppose y naturally decomposes into parts:

$R(x, y)$ denotes the set of "parts" belonging to $(x, y) \in \mathcal{X} \times \mathcal{Y}$

$$G(x, y) = \sum_{r \in R(x,y)} g(x, r)$$

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} G(x, y)'\theta = \arg\max_{y \in \mathcal{Y}} \sum_{r \in R(x,y)} g(x, r)'\theta,$$

- e.g. Markov random fields. Parts are configurations for cliques.
- e.g. PCFGs. Parts are rule-location pairs (rules of grammar applied at specific locations in the sentence).

Large Margin Methods for Structured Classification

- Choose f as maximum of linear functions,

$$f(x) = \arg\max_{y \in \mathcal{Y}} G(x, y)'\theta,$$

  to minimize empirical ϕ-risk.

- e.g., Support Vector Machines:
  $\mathcal{Y} = \{\pm 1\}$, $\ell(\hat{y}, y) = 1[\hat{y} \neq y]$, $G(x, y) = yx$:
  Choose θ to minimize

$$\lambda \|\theta\|^2 + \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i X_i' \theta)_+ \ ,$$

  where $(x)_+ = \max\{x, 0\}$.)
  This is a quadratic program (QP).

Large Margin Classifiers

- For $\mathcal{Y} = \{\pm 1\}$, $\ell(\hat{y}, y) = 1[\hat{y} \neq y]$, and $G(x, y) = yx$,

$$
\begin{aligned}
(1 - 2Y_i X_i'\theta)_+ &= \max_{\hat{y}} \left(\ell(\hat{y}, Y_i) - (Y_i - \hat{y})X_i'\theta\right)_+ \\
&= \max_{\hat{y}} \left(\ell(\hat{y}, Y_i) - (G(X_i, Y_i)'\theta - G(X_i, \hat{y})'\theta)\right)_+ .
\end{aligned}
$$

- Think of $G(x, y)'\theta - G(x, \hat{y})'\theta$ as an upper bound on the loss $l(\hat{y}, y)$ that we'll incur when we choose the $\hat{y}$ that maximizes $G(x, \hat{y})'\theta$.

Large Margin Multiclass Classification

Choose θ to minimize

$$\lambda\|\theta\|^2 + \frac{1}{n}\sum_{i=1}^{n} \max_{\hat{y}} \left(\ell(\hat{y}, Y_i) - (G(X_i, Y_i)'\theta - G(X_i, \hat{y})'\theta)\right)_+$$

$$= \lambda\|\theta\|^2 + \frac{1}{n}\sum_{i=1}^{n} \max_{\hat{y}} \left(\ell(\hat{y}, Y_i) - G'_{i,\hat{y}}\theta\right)_+ ,$$

where $(x)_+ = \max\{x, 0\}$ and $G_{i,\hat{y}} = G(X_i, Y_i) - G(X_i, \hat{y})$.

- Suggested by Taskar et al, 2004.
- Quadratic program.

Large Margin Multiclass Classification

Primal problem:

$$\min_{\theta,\epsilon} \left(\frac{1}{2}\lambda\|\theta\|^2 + \frac{1}{n}\sum_i \epsilon_i\right)$$

Subject to the constraints:

$$\forall i, y \in \mathcal{Y}(X_i),$$

$$\theta' G_{i,y} \geq \ell(y, Y_i) - \epsilon_i$$

$$\forall i,\ \epsilon_i \geq 0$$

Dual problem:

$$\max_{\alpha} \left(C \sum_{i,y} \alpha_{i,y} \ell(y, Y_i) - \frac{C^2}{2} \sum_{i,y,j,z} \alpha_{i,y}\alpha_{j,z} G'_{i,y} G_{j,z} \right)$$

Subject to the constraints:

$$\forall i,\ \sum_y \alpha_{i,y} = 1$$

$$\forall i, y,\ \alpha_{i,y} \geq 0$$

Large Margin Multiclass Classification

Some observations:

- Quadratic program over $\alpha = (\alpha_{i,y})$, restricted to (n copies of) the probability simplex:

$$\begin{aligned} \max_{\alpha} \quad & Q(\alpha) \\ \text{s.t.} \quad & \alpha_i \in \Delta. \end{aligned}$$

- Number of variables is sum over data of number of possible labels. Very large: $n|\mathcal{Y}|$.

Exponentiated Gradient Algorithm

Exponentiated gradients:

$$\alpha^{(t+1)} = \arg\min_{\alpha} \left(D\left(\alpha, \alpha^{(t)}\right) + \eta \alpha' \nabla Q\left(\alpha^{(t)}\right)\right).$$

- D is Kullback-Liebler divergence.
- ∇Q term moves α in direction of decreasing Q.
- KL term constrains it to be close to $\alpha^{(t)}$.

Solution is

$$\alpha_{i,y}^{(t)} = \frac{\exp(\theta_{i,y}^{(t)})}{\sum_z \exp(\theta_{i,z}^{(t)})},$$

with $\theta^{(t+1)} = \theta^{(t)} - \eta \nabla Q(\alpha^{(t)})$.

Exponentiated Gradient Algorithm: Convergence

Theorem: For all $u \in \Delta$,

$$\frac{1}{T}\sum_{t=1}^{T} Q(\alpha^{(t)}) \leq Q(u) + \frac{D(u, \alpha^{(1)})}{\eta T} + c_{\eta,Q}\frac{Q(\alpha^{(1)})}{T}.$$

Exponentiated Gradient Algorithm with Parts

Suppose y naturally decomposes into parts:

$R(x, y)$ denotes the set of "parts" belonging to $(x, y) \in \mathcal{X} \times \mathcal{Y}$

$$G(x, y) = \sum_{r \in R(x,y)} g(x, r)$$

$$\ell(\hat{y}, y) = \sum_{r \in R(x,\hat{y})} L(r, y).$$

- e.g. Markov random fields. Parts are configurations for cliques.
- e.g. PCFGs. Parts are rule-location pairs (rules of grammar applied at specific locations in the sentence).

Exponentiated Gradient Algorithm with Parts

$$G(x, y) = \sum_{r \in R(x,y)} g(x, r)$$

$$\ell(\hat{y}, y) = \sum_{r \in R(x,\hat{y})} L(r, y).$$

- Like a factorization of $\Pr(Y|X)$, where log probabilities decompose as sums over parts.
- We require that loss decomposes in the same way.
  - e.g., Markov random field: $\ell(\hat{y}, y) = \sum_c L(\hat{y}_c, y_c)$.
  - e.g., PCFG: $\ell(\hat{y}, y) = \sum_r 1[r \text{ in } \hat{y}, \text{ not in } y]$.

Exponentiated Gradient Algorithm with Parts

In this case, Q can be expressed as a function of the “marginal” variables, $Q(\alpha) = \tilde{Q}(\mu)$, with

$$\mu_{i,r} = \sum_{y} \alpha_{i,y} 1[r \in R(x_i, y)].$$

Exponentiated gradient algorithm:

$$\mu_{i,r}^{(t)} = \sum_{y} \alpha_{i,y}^{(t)} \; 1[r \in R(x_i, y)]$$

$$\alpha_{i,y}^{(t)} = \frac{\exp(\sum_{r \in R(x_i,y)} \theta_{i,r}^{(t)})}{\sum_y \exp(\sum_{r \in R(x_i,y)} \theta_{i,r}^{(t)})}$$

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_{\mu} \tilde{Q}(\mu^{(t)}).$$

Exponentiated Gradient Algorithm: Sparse Representations

Efficiently computing μ from θ:

- Markov random field: Computing clique marginals from exponential family parameters.
- PCFG: Computing rule probabilities from exponential family parameters.

Exponentiated Gradient Algorithm: Convergence

Theorem: For all $u \in \Delta$,

$$\frac{1}{T}\sum_{t=1}^{T} Q(\alpha^{(t)}) \leq Q(u) + \frac{D(u, \alpha^{(1)})}{\eta T} + c_{\eta,Q}\frac{Q(\alpha^{(1)})}{T}.$$

Step 1:

For any $u \in \Delta$,

$$\eta Q(\alpha^{(t)}) - \eta Q(u) \leq D(u, \alpha^{(t)}) - D(u, \alpha^{(t+1)}) + D(\alpha^{(t)}, \alpha^{(t+1)}).$$

Follows from convexity of Q, definition of updates.

(Standard in analysis of online prediction algorithms.)

Exponentiated Gradient Algorithm: Convergence

Step 2:

$$D(\alpha^{(t)}, \alpha^{(t+1)}) = \sum_{i=1}^{n} \log \mathbf{E}\left[e^{\eta\left(X_i^{(t)} - \mathbf{E}X_i^{(t)}\right)}\right]$$

$$\leq \left(\frac{e^{\eta B} - 1 - \eta B}{B^2}\right) \sum_{i=1}^{n} \mathrm{var}(X_i^{(t)}),$$

where $\Pr\left(X_i^{(t)} = -\left(\nabla Q(\alpha^{(t)})\right)_{i,y}\right) = \alpha_{i,y}^{(t)}$.

Follows from definition of updates, Bernstein's inequality.

Exponentiated Gradient Algorithm: Convergence

Step 3a: For some $\theta \in [\theta^{(t)}, \theta^{(t+1)}]$,

$$\eta \sum_{i=1}^{n} \mathrm{var}(X_i^{(t)}) - \eta^2 (B + \lambda) \sum_{i=1}^{n} \mathrm{var}(X_{i,\theta}^{(t)}) \leq Q(\alpha^{(t)}) - Q(\alpha^{(t+1)}),$$

where $\Pr\left(X_{i,\theta}^{(t)} = -\left(\nabla Q(\alpha^{(t)})\right)_{i,y}\right) = \alpha(\theta)_{i,y}$.

- Variance of $X_i^{(t)}$ is first order term in Taylor series expansion *(in θ)* for Q.
- Variance of $X_{i,\theta}^{(t)}$ is second order term.
- B is infinity norm of centered version of ∇Q
- λ is largest eigenvalue of $\nabla^2 Q$.

Exponentiated Gradient Algorithm: Convergence

Step 3b: For all $\theta \in [\theta^{(t)}, \theta^{(t+1)}]$,

$$\mathrm{var}(X_{i,\theta}^{(t)}) \le e^{\eta B}\,\mathrm{var}(X_i^{(t)}).$$

Hence,

$$\sum_{i=1}^{n} \mathrm{var}(X_i^{(t)}) \le \frac{1}{\eta\left(1 - \eta(B+\lambda)e^{2\eta B}\right)} \left(Q(\alpha^{(t)}) - Q(\alpha^{(t+1)})\right).$$

Exponentiated Gradient Algorithm: Convergence

$$
\begin{aligned}
&\eta Q(\alpha^{(t)}) - \eta Q(u) \\
&\leq D(u, \alpha^{(t)}) - D(u, \alpha^{(t+1)}) + D(\alpha^{(t)}, \alpha^{(t+1)}) \\
&\leq D(u, \alpha^{(t)}) - D(u, \alpha^{(t+1)}) + \left(\frac{e^{\eta B} - 1 - \eta B}{B^2}\right) \sum_{i=1}^{n} \mathrm{var}(X_i^{(t)}) \\
&\leq D(u, \alpha^{(t)}) - D(u, \alpha^{(t+1)}) + c'_{\eta,Q} \left(Q(\alpha^{(t)}) - Q(\alpha^{(t+1)})\right).
\end{aligned}
$$

Theorem: For all $u \in \Delta$,

$$
\frac{1}{T} \sum_{t=1}^{T} Q(\alpha^{(t)}) \leq Q(u) + \frac{D(u, \alpha^{(1)})}{\eta T} + c_{\eta,Q} \frac{Q(\alpha^{(1)})}{T}.
$$

Large Margin Methods for Structured Classification

- Generative models
  - Markov random fields
  - Probabilistic context-free grammars
- Quadratic program for large margin classifiers
- Exponentiated gradient algorithm
- Convergence analysis

Overview

- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition and universal consistency.
- Convergence rates: low noise.
- Kernel classifiers: sparseness versus probability estimation.
- Structured multiclass classification.

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