

# Asymptotic properties of convex optimization methods for multiclass classification

**Peter Bartlett**

Computer Science Division and Department of Statistics  
UC Berkeley

Joint work with Ambuj Tewari.

<http://www.cs.berkeley.edu/~bartlett>

## The Pattern Classification Problem

- i.i.d.  $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$  from  $\mathcal{X} \times \mathcal{Y}$ ,  
 $|\mathcal{Y}| < \infty$ , for example,  $\mathcal{Y} = \{\pm 1\}$ .
- Use data  $(X_1, Y_1), \dots, (X_n, Y_n)$  to choose  $\hat{f} : \mathcal{X} \rightarrow \mathcal{Y}$  with small  
risk,  $R(\hat{f}) = \mathbb{E}\ell(Y, \hat{f}(X))$  e.g.  $= \Pr(\hat{f}(X) \neq Y)$ .
- Natural approach: minimize **empirical risk**,

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

- Often computationally intractable...
- Replace 0-1 loss,  $\ell$ , with a convex surrogate,  $\phi$ .

## Large Margin Algorithms: Two Class Case

- Suppose  $Y \in \{\pm 1\}$ ,  $\hat{f} : \mathcal{X} \rightarrow \mathbb{R}$ . Define

$$R(\hat{f}) = \Pr(\text{sign}(\hat{f}(X)) \neq Y) = \mathbb{E}\ell(Y, \hat{f}(X)).$$

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

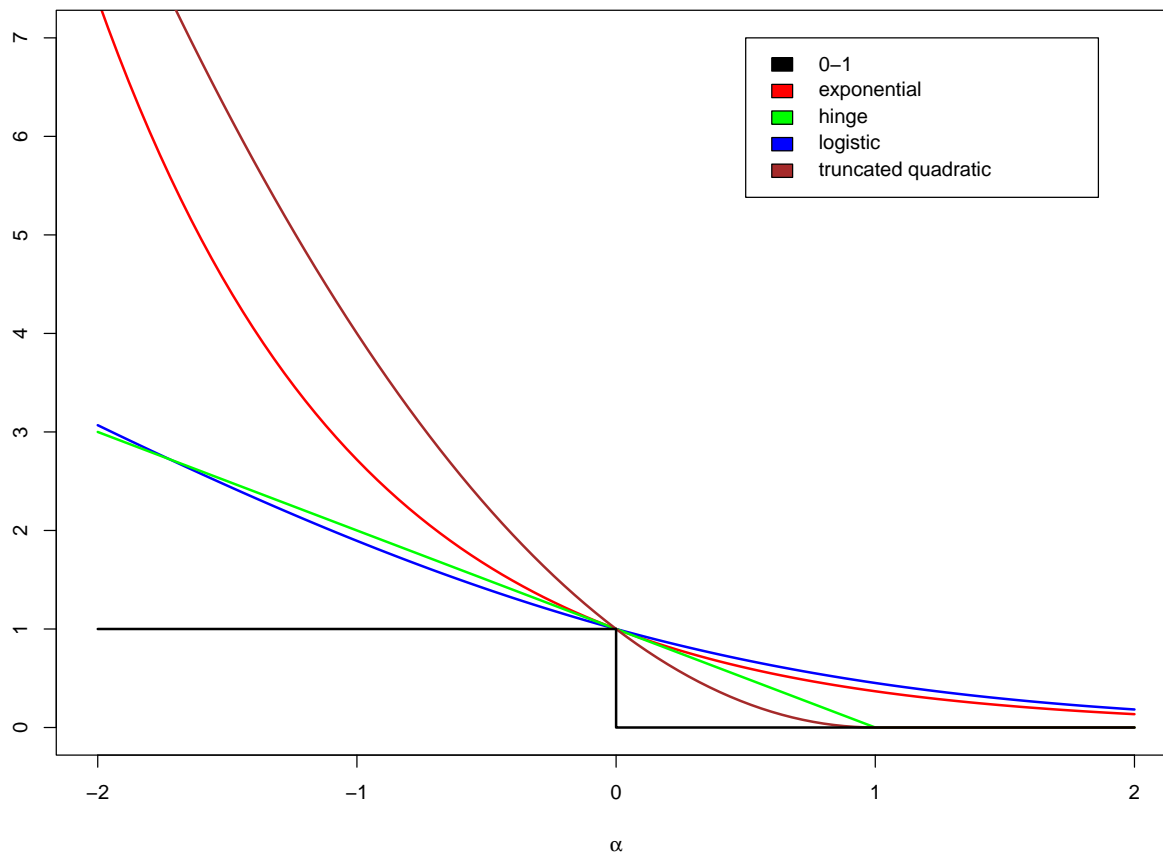
$$\hat{R}_\phi(f) = \hat{\mathbb{E}}_n \phi(Y f(X)) = \frac{1}{n} \sum_{i=1}^n \phi(Y_i f(X_i)),$$

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.
  - $\mathcal{F} = \text{ball}$  in reproducing kernel Hilbert space,  $\mathcal{H}$ .
  - $\phi(\alpha) = (1 - \alpha)_+ = \max\{0, 1 - \alpha\}$ .
  - Algorithm minimizes  $\hat{R}_\phi(f) + \lambda \|f\|_{\mathcal{H}}^2$ .

# Large Margin Algorithms



## Universal consistency

$$\begin{array}{lll} R(f) = \Pr(\text{sign}(f(X)) \neq Y) & R^* = \inf_f R(f) & \text{risk} \\ R_\phi(f) = \mathbb{E}\phi(Y f(X)) & R_\phi^* = \inf_f R_\phi(f) & \phi\text{-risk} \end{array}$$

**Theorem:** If  $\phi$  is convex, the following conditions are equivalent:

1.  $R_\phi(f_i) \rightarrow R_\phi^*$  implies  $R(f_i) \rightarrow R^*$ .
2.  $\phi$  is differentiable at 0, and  $\phi'(0) < 0$ .

## Multiclass large margin methods ( $|\mathcal{Y}| > 2$ )

Two broad categories of multiclass classification methods:

- Combine several binary classifiers. (e.g., error correcting output codes)
- Minimize a cost function defined on a vector space.

We will focus on methods in the second category.

Particularly interested in methods for problems with  $|\mathcal{Y}|$  very large.


## Overview

- Multiclass pattern classification problems and methods.
- Formulation of multiclass large margin methods.
- Characterization of consistency: classification calibration.
- Checking classification calibration: admissibility of projections.
- Examples.

## Optical Character Recognition

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

$Y$  = sequence of characters



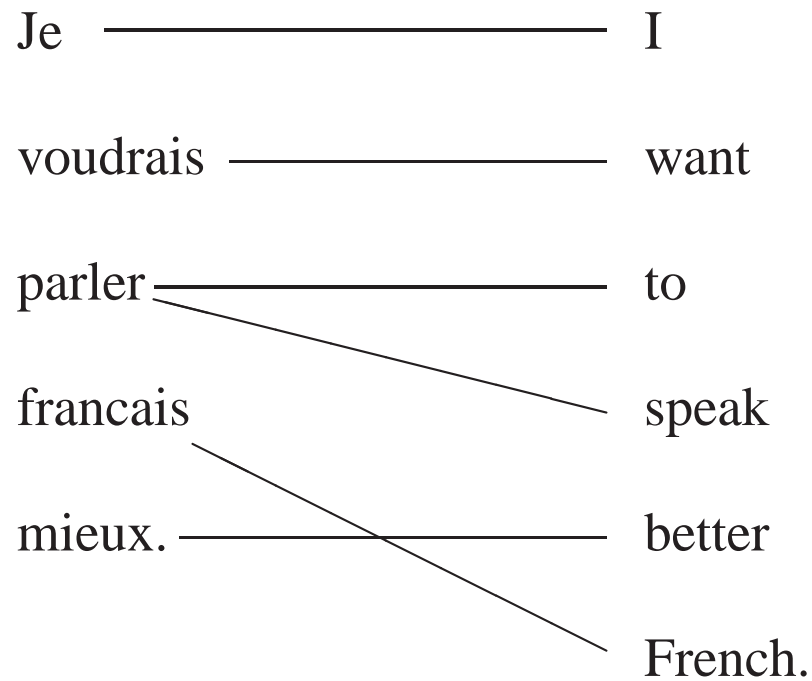
*This is an example of*

This is an example of

## Word Alignment in Machine Translation

$X$  = a sentence in the two languages

$Y$  = alignment

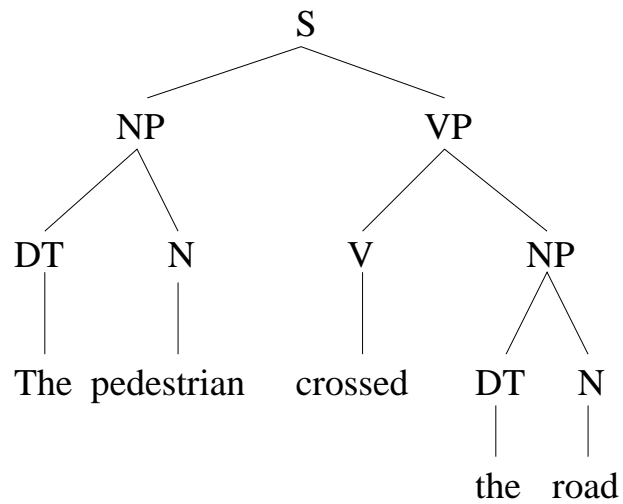


# Parsing

$X$  = sentence

$Y$  = parse tree

The pedestrian crossed the road.



## Structured multiclass pattern classification problems

Key issue:  $|\mathcal{Y}|$  is huge (exponential in number of characters/words).

**A convex optimization approach:**

(Taskar *et al*, 2004)

Choose  $f \in \mathcal{F} \subseteq \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$  to minimize

$$\frac{1}{n} \sum_{i=1}^n \max_{\hat{y}} (\ell(Y_i, \hat{y}) - (f(X_i, Y_i) - f(X_i, \hat{y})))_+ + \Omega_n(f),$$

where  $(x)_+ = \max\{x, 0\}$ .

Then predict  $\arg \max_y f(X, y)$ .

This is a generalization of the SVM optimization.

## Structured multiclass pattern classification problems

Suppose  $y$  decomposes into (small) parts, e.g.,

- configurations of cliques in Markov random field.
- rule-location pairs in PCFG.

If  $\ell$  is linear in this decomposition,

that is,  $\ell(Y_i, \hat{y}) = \sum_r \ell_r(r(Y_i), r(\hat{y}))$  and each  $r$  has small range, and similarly for  $f(x, \cdot)$ , then there are efficient algorithms for this optimization:

Maximum margin Markov networks: Taskar et al.

Constraint sampling: Tsochantaridis et al.

Exponentiated gradient: Bartlett, Collins, Taskar, McAllester.

## Overview

- Multiclass pattern classification problems and methods.
- Formulation of multiclass large margin methods.
- Characterization of consistency: classification calibration.
- Checking classification calibration: admissibility of projections.
- Examples.

## Multiclass large margin methods

$$\mathcal{Y} = \{1, \dots, K\}.$$

Think of a classifier as a vector valued function  $\mathbf{f} : \mathcal{X} \mapsto \mathbb{R}^K$ .

For suitable loss functions  $\Psi_y : \mathbb{R}^K \rightarrow \mathbb{R}_+$ , pick  $\hat{\mathbf{f}}$  by minimizing

$$\frac{1}{n} \sum_{i=1}^n \Psi_{y_i}(\mathbf{f}(x_i)) + \Omega_n(\mathbf{f}).$$

Predict label using  $\arg \min_{y \in \mathcal{Y}} \Psi_y(\hat{\mathbf{f}}(x))$ .

## Multiclass large margin methods

A few methods of this kind from the literature:

$$(x_+ = \max\{0, x\})$$

|  | $\Psi_{y_i}(\mathbf{f}(x_i))$  |
|--|--|
| Vapnik; Weston and Watkins;<br>Bredensteiner and Bennett | $\sum_{y' \neq y_i} (f_{y'}(x_i) - f_{y_i}(x_i) + 1)_+$                                      |
| Crammer and Singer; Taskar et al                         | $\max_{y' \neq y_i} (f_{y'}(x_i) - f_{y_i}(x_i) + 1)_+$                                      |
| Lee, Lin and Wahba                                       | $\sum_{y' \neq y_i} (1 + f_{y'}(x_i))_+$<br>with sum-to-zero constraint, $\sum_y f_y(x) = 0$ |

All predict label using  $\arg \max_{y \in \mathcal{Y}} f_y(x) = \arg \min_{y \in \mathcal{Y}} \Psi_y(\mathbf{f}(x))$ .

## Different behaviors

- For  $K = 2$ , all methods are equivalent and universally consistent.
- But they have different behaviors for  $K > 2$ .
  - Lee, Lin and Wahba's is consistent.
  - The other two are not.
- This led us to investigate consistency of a general class of methods of which all of these are special cases.

## General Framework

- Pointwise constraint on  $\mathbf{f}$ ,  $\forall x, \mathbf{f}(x) \in \mathcal{C}$  for some  $\mathcal{C} \subseteq \mathbb{R}^K$ .

| $\Psi_y(\mathbf{f})$ :                | $\mathcal{C}$ :   |
|---------------------------------------|---|
| $\sum_{y' \neq y} \phi(f_y - f_{y'})$ | $\mathbb{R}^K$  |
| $\max_{y' \neq y} \phi(f_y - f_{y'})$ | $\mathbb{R}^K$  |
| $\sum_{y' \neq y} \phi(-f_{y'})$      | $\left\{ \mathbf{z} \in \mathbb{R}^K : \sum_{i=1}^K z_i = 0 \right\}$ |

- $\phi(x) = (1 - x)_+$  gives us our three example methods but other  $\phi$  have been proposed also.

## $\Psi$ -risk

Define  $\mathcal{F} = \{\mathbf{f} : \forall x, \mathbf{f}(x) \in \mathcal{C}\}$ .

$$\Psi\text{-risk: } R_{\Psi}(\mathbf{f}) = \mathbb{E}\Psi_y(\mathbf{f}(x)),$$

$$\text{optimal } \Psi\text{-risk: } R_{\Psi}^* = \inf_{\mathbf{f} \in \mathcal{F}} R_{\Psi}(\mathbf{f}).$$

Since  $\mathbf{f}$  enters into the  $\Psi$ -risk definition only through  $\Psi$ , we assume that we predict labels using

$$\text{pred}(\Psi_1(\mathbf{f}(x)), \dots, \Psi_K(\mathbf{f}(x)))$$

for some  $\text{pred} : \mathbb{R}^K \mapsto \mathcal{Y}$ .

## $\Psi$ -risk

Since

$$R_{\Psi}(\mathbf{f}) = \mathbb{E} \Psi_y(\mathbf{f}(x)) = \mathbb{E}_x \mathbb{E}_{y|x} \Psi_y(\mathbf{f}(x)),$$

we can write

$$\begin{aligned} R_{\Psi}^* &= \inf_{\mathbf{f} \in \mathcal{F}} R_{\Psi}(\mathbf{f}) = \mathbb{E}_x \inf_{f \in \mathcal{C}} \langle p(x), \Psi(f) \rangle \\ &= \mathbb{E}_x \inf \{ \langle p(x), z \rangle : z = (\Psi_1(f), \dots, \Psi_K(f)), f \in \mathcal{C} \} \\ &= \mathbb{E}_x \inf \left\{ \langle p(x), z \rangle : z \in \underbrace{\text{conv} \{ (\Psi_1(f), \dots, \Psi_K(f)) : f \in \mathcal{C} \}}_S \right\}, \end{aligned}$$

where  $p(x) \in \Delta^K$  is the vector of conditional probabilities,  
 $p(x) = (p_1(x), \dots, p_K(x))$ ,  $p_y(x) = P(Y = y|X = x)$ .

## Definitions

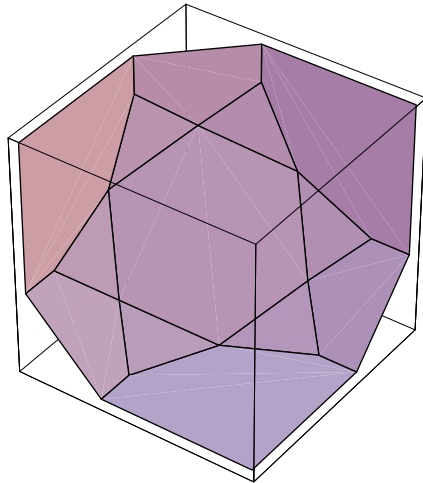
$$p_y(x) = P(Y = y|X = x),$$

$$p(x) = (p_1(x), \dots, p_K(x))$$

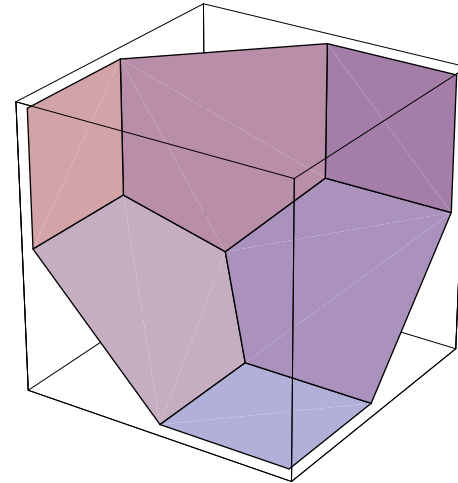
$$\mathcal{F} = \{\mathbf{f} : \forall x, \mathbf{f}(x) \in \mathcal{C}\}$$

$$\mathcal{S} = \text{conv} \{(\Psi_1(f), \dots, \Psi_K(f)) : f \in \mathcal{C}\}.$$

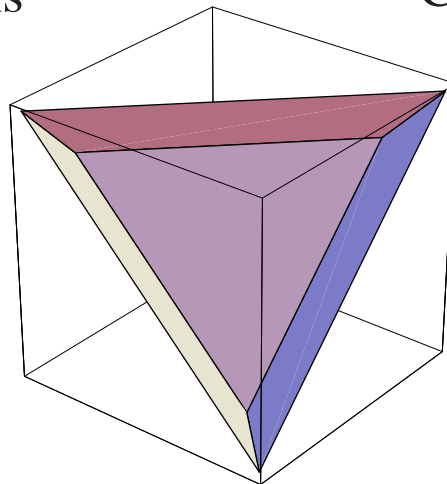
# Pictures of boundary of $S$



Weston & Watkins



Crammer & Singer



Lee, Lin & Wahba

## Consistency

Optimizing a cost function  $\Psi$  will lead to a consistent method provided that, for all probability distributions and all sequences  $\{\mathbf{f}^{(n)}\}$ ,

$$R_{\Psi}(\mathbf{f}^{(n)}) \rightarrow R_{\Psi}^* \implies R(\mathbf{f}^{(n)}) \rightarrow R^*.$$

$$R_{\Psi}^* = \mathbb{E}_x \inf_{z \in \mathcal{S}} \langle p(x), z \rangle. \quad R^* = 1 - \mathbb{E}_x \max_{y \in \mathcal{Y}} p_y(x).$$

- Minimizing  $\langle p, z \rangle$  over  $\mathcal{S}$  should lead to a  $z$  that allows us to determine the index of one of the maximum coordinates of  $p$ .

## Classification Calibration

**Definition:**  $\mathcal{S} \subseteq \mathbb{R}_+^K$  is *classification calibrated* iff

$\exists \text{pred} : \mathbb{R}^K \rightarrow \{1, \dots, K\}$  such that  $\forall \mathbf{p} \in \Delta_K$ ,

$$\inf \left\{ \langle \mathbf{p}, \mathbf{z} \rangle : \mathbf{z} \in \mathcal{S}, p_{\text{pred}(\mathbf{z})} < \max_y p_y \right\} > \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle.$$

**Theorem:**  $\mathcal{S}$  is CC iff

$$\forall \{\mathbf{f}^{(n)}\} \text{ in } \mathcal{F}, \quad R_{\Psi}(\mathbf{f}^{(n)}) \rightarrow R_{\Psi}^* \quad \Rightarrow \quad R(\mathbf{f}^{(n)}) \rightarrow R^* .$$

## Consistency

- Consider an (informal) game where:
  - The opponent chooses a  $\mathbf{p} \in \Delta_K$  and reveals to us a sequence  $\mathbf{z}^{(n)}$  with  $\langle \mathbf{p}, \mathbf{z}^{(n)} \rangle \rightarrow \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle$
  - We output the sequence  $l_n = \text{pred}(\mathbf{z}^{(n)})$ .

We win if  $p_{l_n} = \max_y p_y$  ultimately.

- For consistency, there should be a  $\text{pred}$  such that we win irrespective of the choice of the opponent.

## Classification Calibration

Clearly:  $\mathcal{S} \subseteq \mathbb{R}_+^K$  is CC iff  $\exists$  pred such that  $\forall \mathbf{p} \in \Delta_K$  and all  $\{\mathbf{z}^{(n)}\}$  in  $\mathcal{S}$ ,

$$\langle \mathbf{p}, \mathbf{z}^{(n)} \rangle \rightarrow \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle$$

implies

$$p_{\text{pred}(\mathbf{z}^{(n)})} \rightarrow \max_y p_y.$$

## The predictor function

- Assume from now on that the set  $\mathcal{S}$  is convex and symmetric (symmetry means that all  $K$  classes are treated equally).

**Lemma:** If any pred satisfies

$$\forall \mathbf{p} \in \Delta_K, \inf \left\{ \langle \mathbf{p}, \mathbf{z} \rangle : \mathbf{z} \in \mathcal{S}, p_{\text{pred}(\mathbf{z})} < \max_y p_y \right\} > \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle,$$

then so does one satisfying  $z_{\text{pred}(\mathbf{z})} = \min_y z_y$ .

## Admissibility

Define the class  $\mathcal{N}(\mathbf{z})$  of non-negative normals to  $\mathcal{S}$  at  $\mathbf{z}$ ,

$$\mathcal{N}(\mathbf{z}) = \{p : \forall \mathbf{z}' \in \mathcal{S}, \langle \mathbf{z}' - \mathbf{z}, \mathbf{p} \rangle \geq 0\} \cap \Delta_K.$$

This is the set of  $\mathbf{p}$  for which  $\mathbf{z}$  minimizes  $\langle \mathbf{z}, \mathbf{p} \rangle$ .

**Definition:**  $\mathcal{S}$  is admissible if  $\forall \mathbf{z} \in \partial\mathcal{S}, \forall \mathbf{p} \in \mathcal{N}(\mathbf{z})$ , we have

$$\arg \min_y (z_y) \subseteq \arg \max_y (p_y).$$

## Necessary and sufficient condition

- Admissibility *weaker* than classification calibration.
- It is equivalent to the CC definition with the additional assumption of *boundedness* of the sequence  $\{\mathbf{z}^{(n)}\}$ .
- Necessary and sufficient condition is given by:

**Theorem** Let  $\mathcal{S} \subseteq \mathbb{R}_+^K$  be a symmetric convex set. Define the sets

$$\mathcal{S}^{(i)} = \{(z_1, \dots, z_i) : \mathbf{z} \in \mathcal{S}\}$$

for  $i \in \{2, \dots, K\}$ . Then  $\mathcal{S}$  is classification calibrated iff each  $\mathcal{S}^{(i)}$  is admissible.

## Checking Admissibility

**Unique minimum:** If  $|\arg \min_y z_y| = 1$ , then for all  $\mathbf{p} \in \mathcal{N}(\mathbf{z})$ ,

$$\arg \min_y z_y \subseteq \arg \max_y p_y.$$

**Unique normal:** If  $\mathcal{N}(\mathbf{z}) = \{\mathbf{p}\}$ , then

$$\arg \min_y z_y \subseteq \arg \max_y p_y.$$

## Overview

- Multiclass pattern classification problems and methods.
- Formulation of multiclass large margin methods.
- Characterization of consistency: classification calibration.
- Checking classification calibration: admissibility of projections.
- Examples.

## Example 0: Two-class classification

For some convex  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ , define

$$\Psi_1(f) = \phi(f_1)$$

$$\Psi_{-1}(f) = \phi(f_{-1}) = \phi(-f_1) \quad (f_1 + f_{-1} = 0)$$

$$\mathcal{S} = \text{conv} \{(\phi(f_1), \phi(-f_1)) : f_1 \in \mathbb{R}\},$$

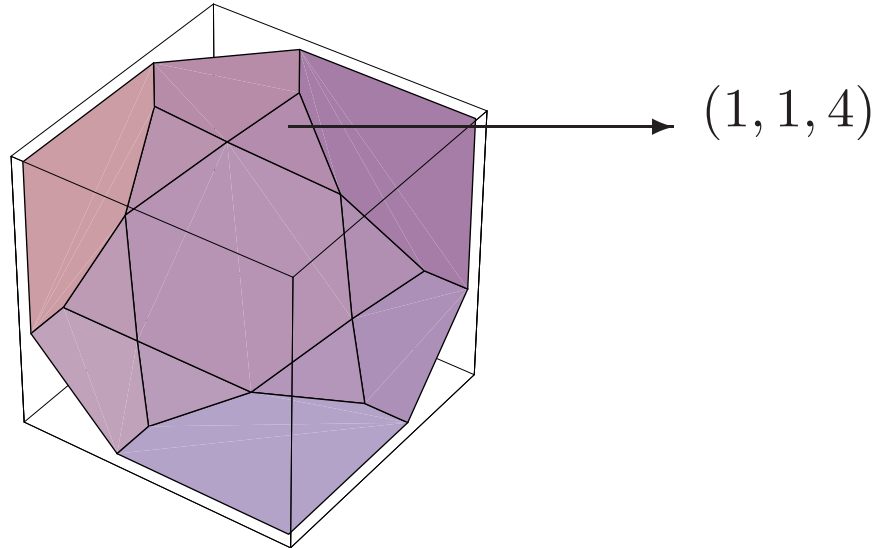
$|\arg \min_y z_y| > 1$  only at  $\phi(f_1) = \phi(-f_1)$ , and convexity implies

$$\frac{\phi(f_1) + \phi(-f_1)}{2} \geq \phi(0),$$

so  $\mathcal{S}$  is CC iff there is a unique normal at  $(\phi(0), \phi(0))$

iff  $\phi'(0)$  exists and is non-zero.

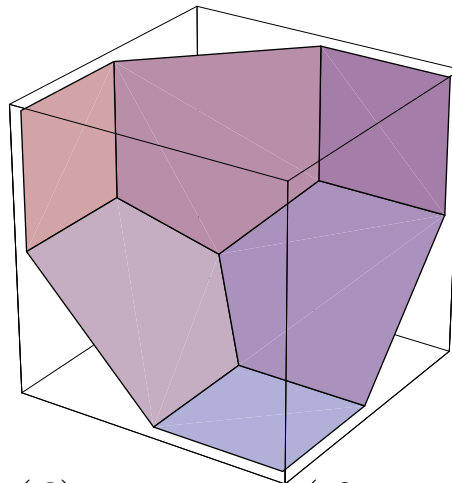
## Example 1: Weston and Watkins



$$\Psi_y(\mathbf{f}) = \sum_{y' \neq y} \phi(f_y - f_{y'})$$

- For  $\mathbf{z} = (1, 1, 4)$ ,  $\mathcal{N}(\mathbf{z})$  includes  $(1, 1, 0)$ ,  $(1, 1, 1)$ ,  $(2, 3, 1)$  and  $(3, 2, 1)$ . Now  $\arg \min(\mathbf{z}) = \{1, 2\}$  while  $\arg \max(2, 3, 1) = \{2\}$  which violates admissibility. The method is therefore not consistent (but choosing a suitable differentiable  $\phi$  will make it consistent).

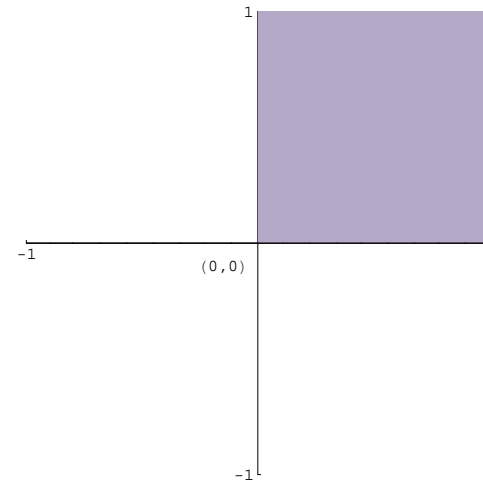
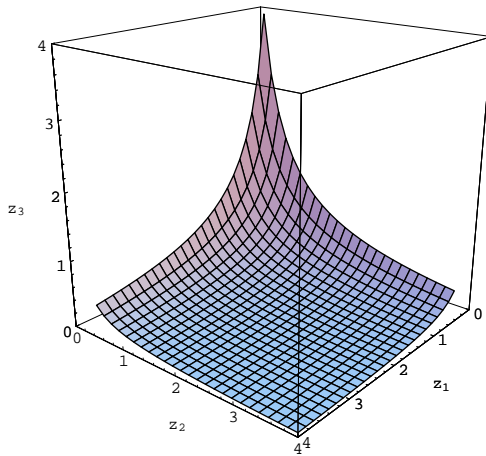
## Example 2: Crammer and Singer



$$\Psi_y(\mathbf{f}) = \max_{y' \neq y} \phi(f_y - f_{y'})$$

- For all  $\phi$  differentiable at 0, the set of normals at  $\mathbf{z} = (\phi(0), \phi(0), \phi(0))$  includes  $(0, \frac{1}{2}, \frac{1}{2})$ ,  $(\frac{1}{2}, 0, \frac{1}{2})$  and  $(\frac{1}{2}, \frac{1}{2}, 0)$ . Since  $\arg \min_y (z_y) = \{1, 2, 3\}$  and  $\arg \max_y ((0, \frac{1}{2}, \frac{1}{2})) = \{2, 3\}$ , admissibility is violated.
- Notice that  $\mathcal{N}(\mathbf{z})$  includes all  $p$  on the line from  $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$  to  $(0, \frac{1}{2}, \frac{1}{2})$ .

## Example 3: A smooth loss function



Boundary of the set  $\mathcal{S} = \mathcal{S}^{(3)}$

The set  $\mathcal{S}^{(2)}$

- $\Psi_y(\mathbf{f}) = \exp(-f_y)$  with  $K = 3$  and  $\sum_y f_y = 0$  gives  $\mathcal{S} = \{\mathbf{z} \in \mathbb{R}_+ : z_1 z_2 z_3 \geq 1\}$ .
- $\mathcal{S}$  is admissible,  $\mathcal{S}^{(2)}$  is not (origin has  $(0, 1)$  and  $(1, 0)$  as normals).

## **Asymptotic properties of multiclass methods**

- Multiclass pattern classification problems and methods.
- Formulation of multiclass large margin methods.
- Characterization of consistency: classification calibration.
- Checking classification calibration: admissibility of projections.
- Examples.