Greenslopes Seminar: Introduction to Support Vector Machines (SVMs)

Sofya Chepushtanova

Department of Mathematics
Colorado State University

January 19, 2012

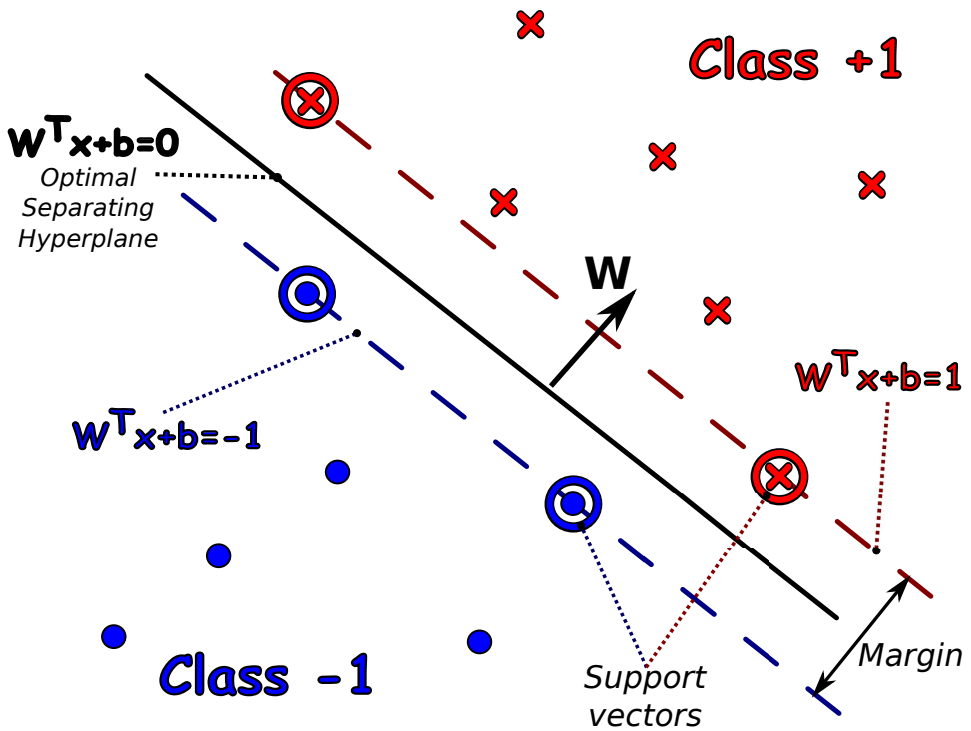
History

- An SVM creates a model which assigns an object to one of the data classes.
- First mentioned in 1979 by Vladimir Vapnik in *Estimation of Dependences Based on Empirical Data*
- Attracted attention in the 90's: *The Nature of Statistical Learning Theory* by Vapnik, 1995
- Since then SVMs, or large margin classifiers, have been widely used demonstrating good performance in different applications

SVM: Linearly Separable Case

Consider a binary classification problem:

- m vectors $x_i \in \mathbb{R}^N$, each from either class $+1$ or -1.
- m labels $d_i = \{-1, +1\}$.



- Separating hyperplane $P = \{x : w^T x + b = 0\}$, $w \in \mathbb{R}^N$ is the normal to P, b is the bias.
- There are many hyperplanes! We prefer the one for which the smallest perpendicular distance to a training sample is maximized.

SVM: Linearly Separable Case

Distance from the hyperplane P to a sample x? (we use the Euclidean norm $\|\cdot\| = \|\cdot\|_2$)

$x = v + r\frac{w}{\|w\|}$

$w^T x + b = w^T(v + r\frac{w}{\|w\|}) + b$

$w^T x + b = (w^T v + b) + r\|w\|$

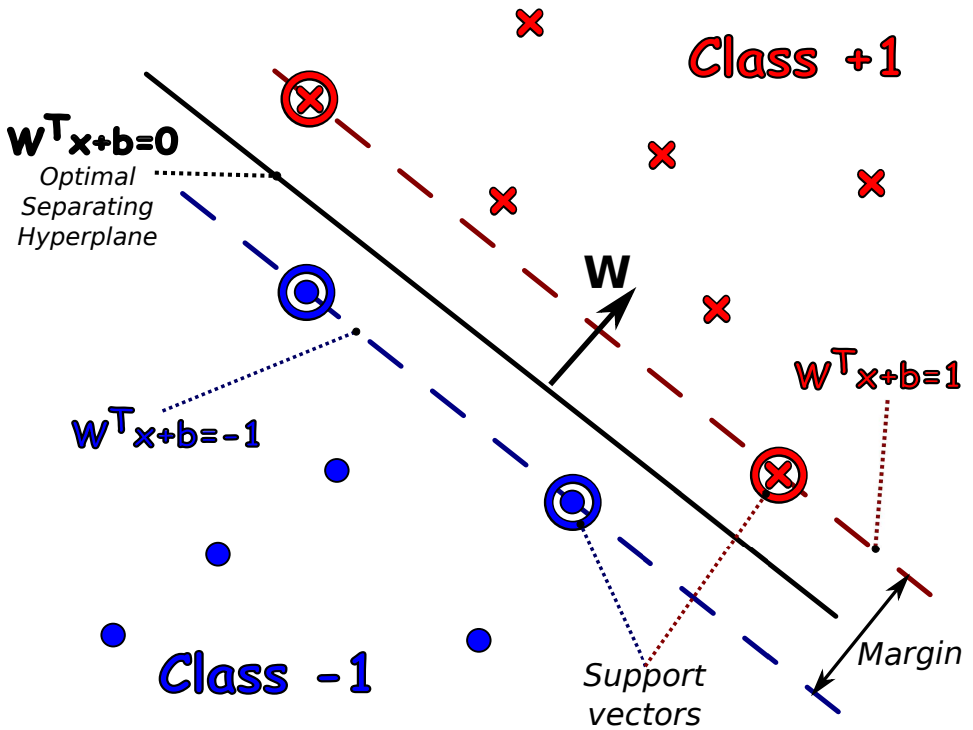
$w^T x + b = 0 + r\|w\| \Rightarrow d(x, P) = r = \frac{w^T x + b}{\|w\|}$.

- $\forall x_i, d_i(w^T x_i + b) > 0$ (recall: $d_i = \{-1, +1\}$).
- Want: optimal hyperplane $\underset{w,b}{argmax}(\underset{1 \le i \le m}{min}\ d(x_i, P))$.
- The argument in decision function $f(x) = sgn(w^T x + b)$ is invariant under rescaling $w \to \lambda w, b \to \lambda b \Rightarrow$ constraint $d_i(w^T x_i + b) \ge 1$ for $\forall i$.

SVM: Linearly Separable Case

Support vectors (SVs)

x_i that lie on the canonical hyperplanes $|w^T x_i + b| = 1$



Margin

Distance between canonical hyperplanes = $|\frac{1}{\|w\|} - \frac{-1}{\|w\|}| = \frac{2}{\|w\|}$.

SVM: Linearly Separable Case

Optimization problem

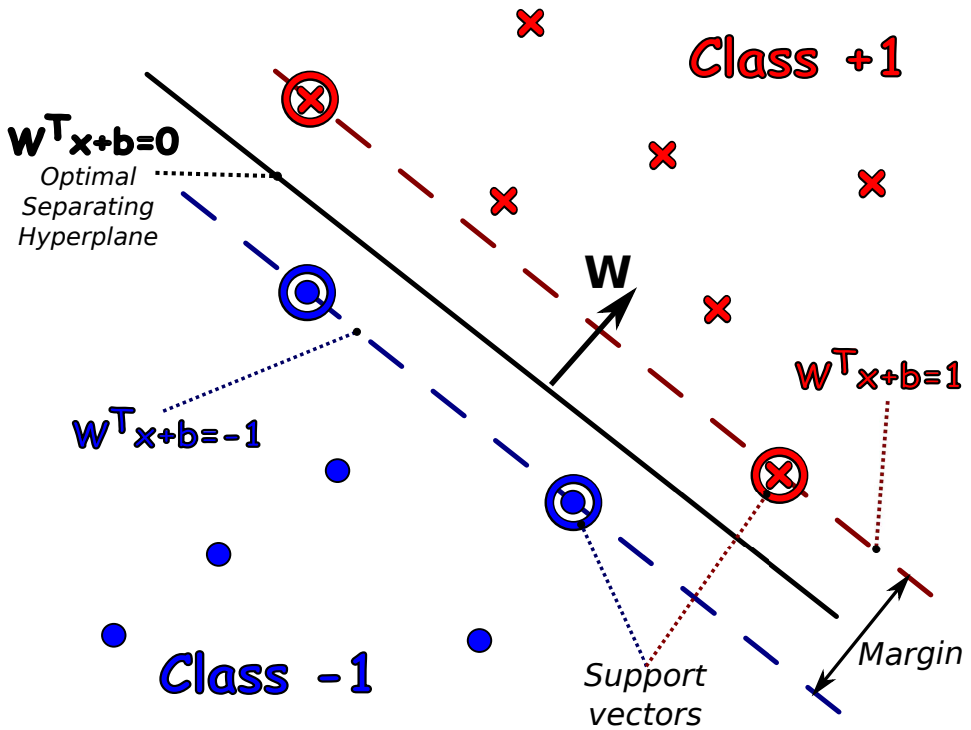
$$\min_{w,b} \quad \frac{\|w\|_2^2}{2}$$

subject to $d_i(w^T x_i + b) \geq 1,$
$i = 1, \ldots, m$

In block matrix form:

$$\min_{w,b} \quad \frac{\|w\|_2^2}{2}$$

subject to $D(Xw + be) \geq e$



The decision function $f(x) = sgn(w^T x + b)$ assigns a class label to a testing sample x.

Nonseparable Case: Soft-Margin SVM

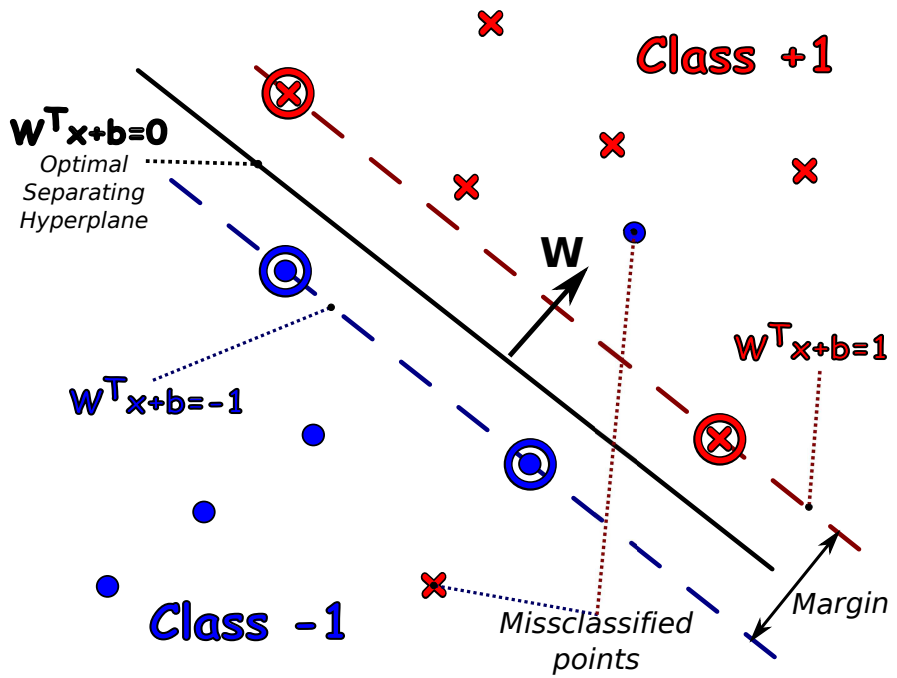
Misclassification penalty term:

Optimization problem

$$\min_{w,b} \quad \frac{\|w\|_2^2}{2} + C\sum_{i=1}^{m}\xi_i$$
subject to $d_i(w^T x_i + b) \geq 1 - \xi_i$,
$\xi_i \geq 0, i = 1, \ldots, m.$

In block matrix form:

$$\min_{w,b,\xi} \quad \frac{\|w\|_2^2}{2} + Ce^T\xi$$
subject to $D(Xw + be) \geq e - \xi$,
$\xi \geq 0.$



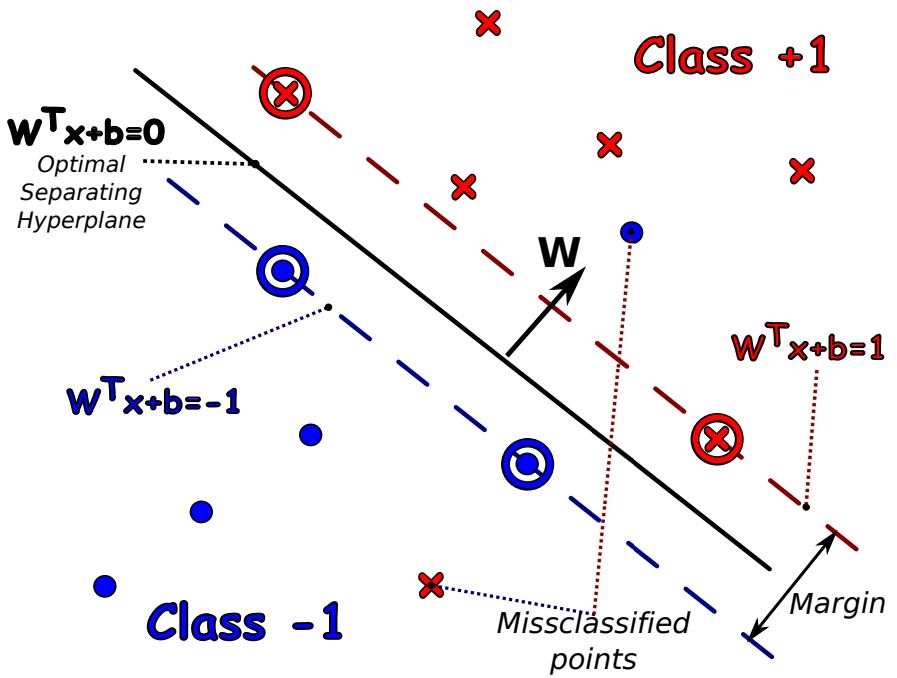
Nonseparable Case: Soft-Margin SVM

Primal Problem

$$\min_{w,b,\xi} \quad \frac{\|w\|_2^2}{2} + Ce^T\xi$$
subject to $D(Xw + be) \geq e - \xi$,
$\xi \geq 0$.

Dual Problem

$$\max_{\alpha} \quad e^T\alpha - \frac{1}{2}\alpha^T DXX^T D\alpha$$
subject to $e^T D\alpha = 0$,
$0 \leq \alpha \leq Ce$.



Nonseparable Case: Soft-Margin SVM

Support vectors

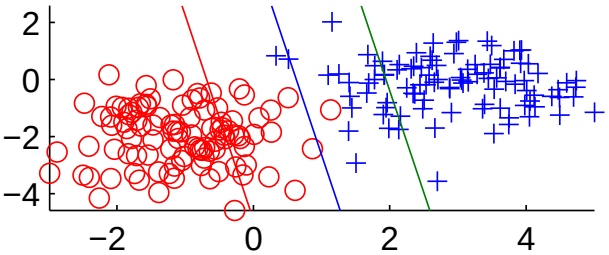
- The solution is given by $w = \sum_{i=1}^{m} \alpha_i d_i x_i$ (b can be computed using w and any training point).
- On-boundary SVs: $\xi_i = 0$ and $0 \leq \alpha_i \leq C$.
- Off-boundary SVs: $\xi_i > 0$ and $\alpha_i = C$.
- Non-SVs: $\xi_i = 0$ and $\alpha_i = 0$.

Parameter C

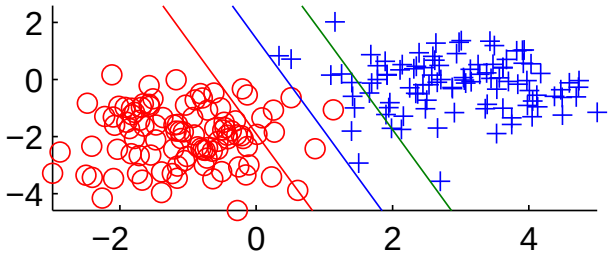
- $C = \infty$ corresponds to the hard-margin SVM (separable case).
- C is finite: higher C produces less errors, margin shrinks; smaller C maximizes the margin, more errors.
- C can be found by cross validation procedure.

Example: paramater C for toy data

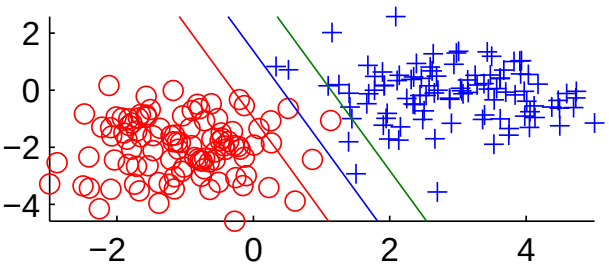
C = 0.03125



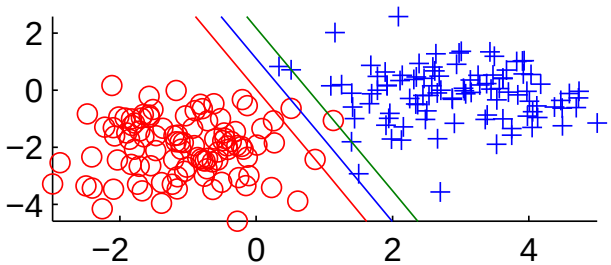
C = 0.125



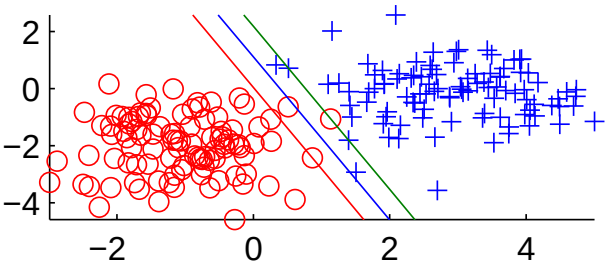
C = 0.5



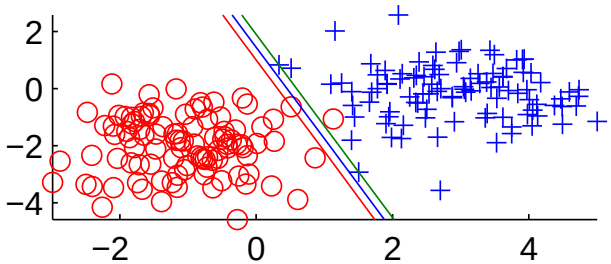
C = 2



C = 4



C = 1024

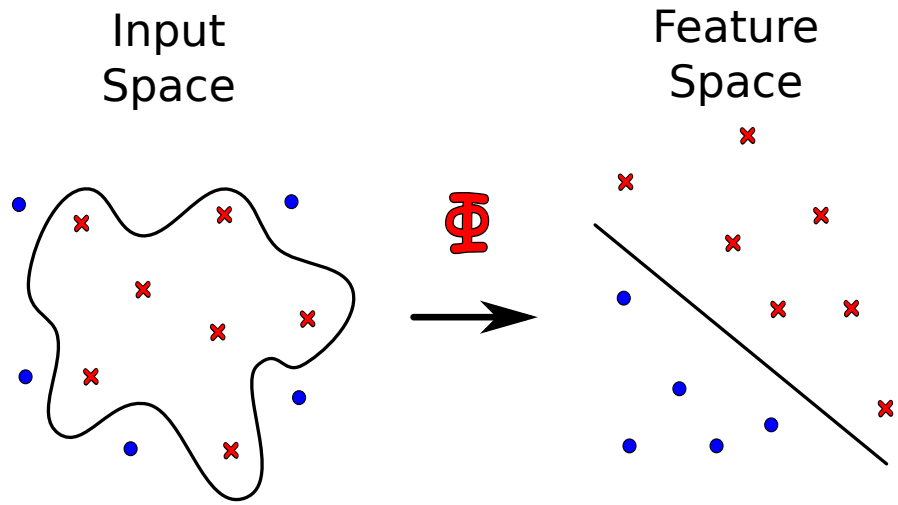


Nonlinear SVM: Kernel Trick

- $\Phi : x \in \mathbb{R}^N \mapsto \Phi(x) \in \mathbb{R}^{N'}, N' > N.$
- Kernel function $K_{ij} = K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$.

Dual Problem

$$\max_\alpha \quad e^T\alpha - \frac{1}{2}\alpha^T DKD\alpha$$
subject to $e^T D\alpha = 0$,
$0 \leq \alpha \leq Ce$.



- the decision function is $f(x) = sgn(\sum_i^m \alpha_i d_i K(x_i, x) + b)$.
- RBF $K(x_i, x) = exp(-\gamma\|x_i - x\|^2)$,
  polynomial $K(x_i, x) = (x_i^T x + 1)^n$.

Implementation and Applications

An SVM problem can be solved using

- Straightforward way: a quadratic programming (QP) solver (expensive)
- Decomposition (chunking) methods: decompose the problem into subproblems, each of which is small to use a QP solver
- Newton method
- Conjugate gradient method
- Primal dual interior point methods (I am using)
- SMO: sequential minimal optimization

Application Domains:

- Pattern recognition in plenty of fields (e.g. text classification, bioinformatics, medical data, image processing)
- SVM regression
- Feature selection

Binary SVMs can be extended to multiclass SVMs

Feature Selection Based on L_1-Norm SVM

- Feature selection is the technique of selecting a subset of relevant features for building robust models.
- The L_2-norm gives non-sparse decision function $f(x) = w^T x + b$.
- The L_0-norm (number of nonzero components) results in NP-hard optimization problem.

- L_p-norm: $\|x\|_p = (\sum_{i=1}^{N} |x_i|^p)^{\frac{1}{p}}$.

- The L_1-norm is known to generate sparse solutions $\Rightarrow$ L_1-norm optimization sets irrelevent weights to zero, thus providing an automatic feature selection method.

Feature Selection Based on L_1-Norm SVM

- Mangasarian (1999), Pedroso and Murata (2001):

If L_p-norm is used to measure distance between points and the separating hyperplane, the margin is equal to $\frac{2}{\|w\|_q}$, where L_p and L_q are dual ($1/p + 1/q = 1$).

- L_∞-norm ($\|x\|_\infty = max_i\{|x_i|\}$) is dual to L_1-norm.
- Use the L_∞-norm to determine the distance $\Rightarrow$ minimize $\|w\|_1$ (and vice versa).

Feature Selection Based on L_1-Norm SVM

L_1-Norm SVM Optimization Problem

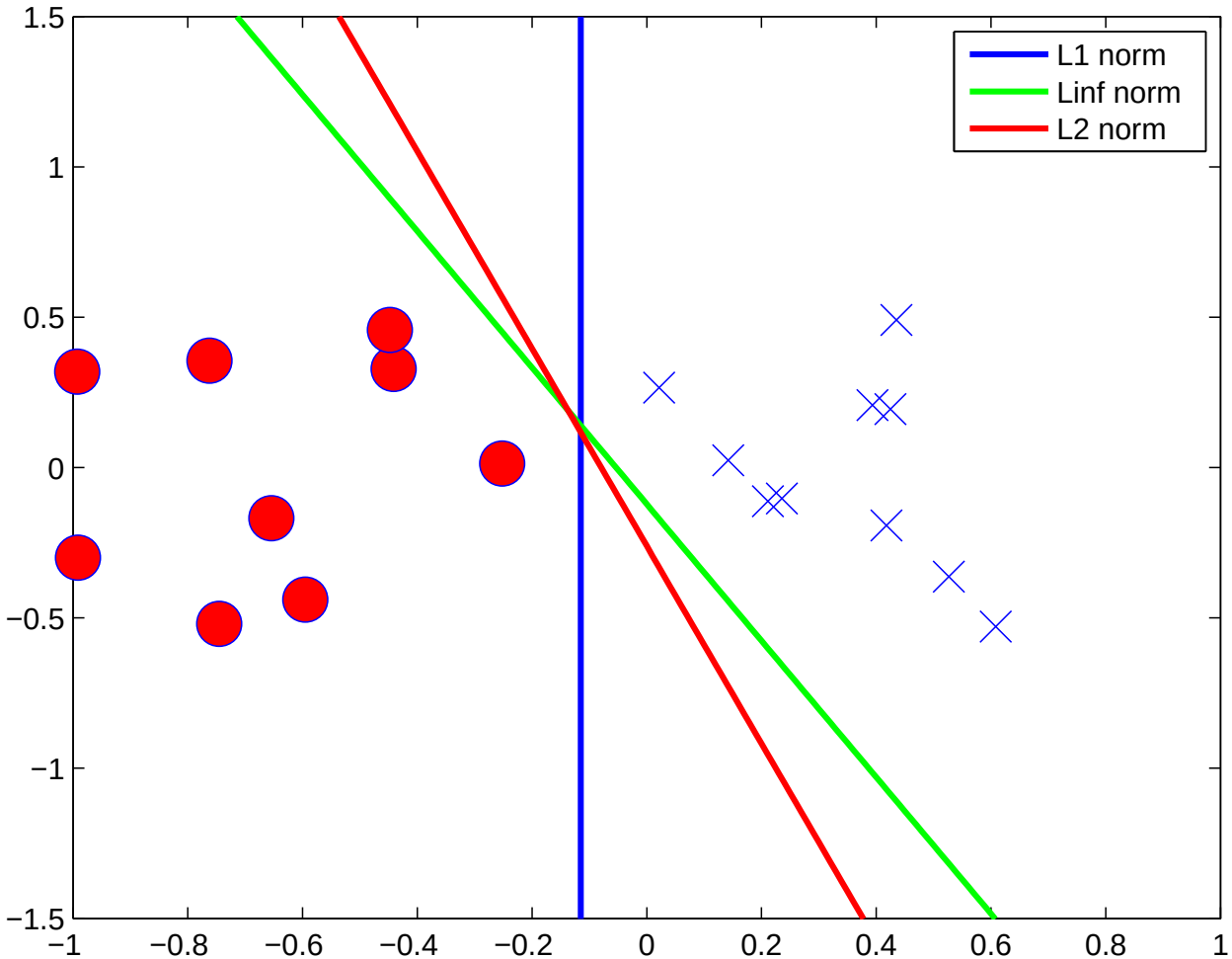
$$\begin{aligned} \min_{w,b,\xi} \|w\|_1 + Ce^T\xi \\ \text{subject to } D(Xw + be) \geq e - \xi \\ \xi \geq 0 \end{aligned}$$

This problem can be converted to the following

Linear Program (LP)

$$\begin{aligned} \min_{w,b,\xi,y} \quad e^T y + Ce^T\xi \\ \text{subject to } D(Xw + be) \geq e - \xi \\ -w + y \geq 0 \\ w + y \geq 0 \\ \xi \geq 0 \end{aligned}$$

Different Norms SVMs for 2D separable problem



Optimization Method

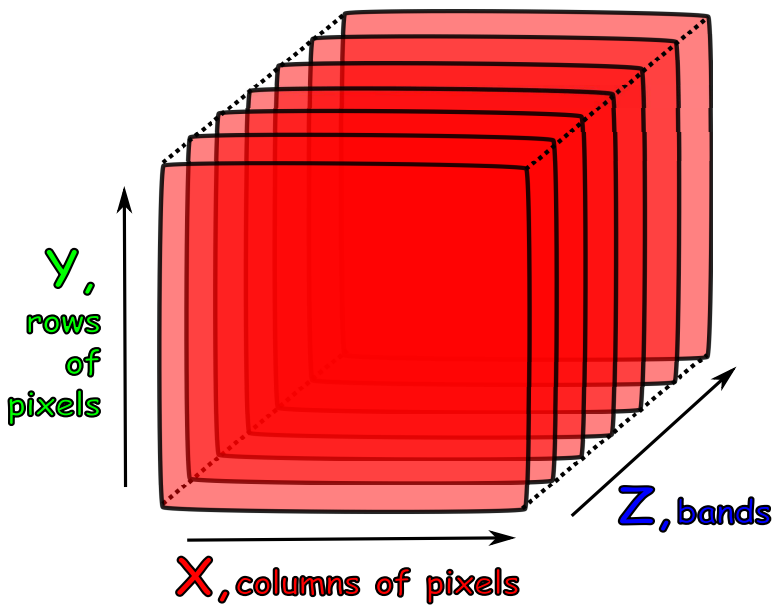
- Linear program (LP) can be solved by a primal-dual interior-point method, namely, path-following algorithm, that searches the optimal solution while moving in the interior of the feasible set.
- LP $\Rightarrow$ linear system $\Rightarrow$ Newton's method.
- Solution is obtained for both the primal and dual problems simultaneously.
- For large scale problems one can use decomposition techniques.

Hyperspectral Data

- Datacube with spatial information in the $x - y$ plane, and spectral information in the z-direction.
- Each pixel is a vector $x \in \mathbb{R}^N$, where N is the number of spectral bands/wavelengths.

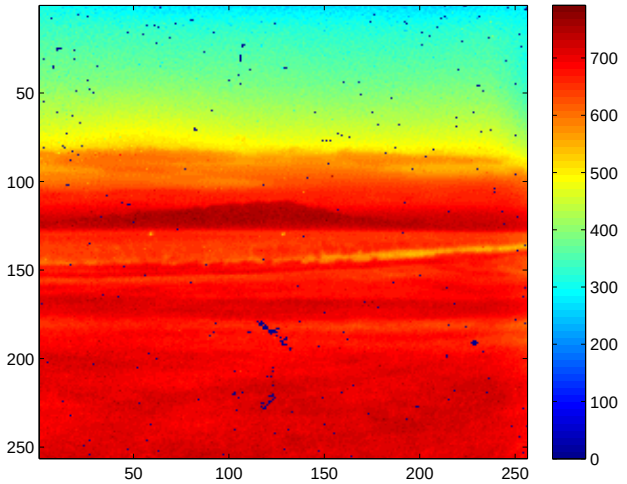
Band selection:

Select a minimal and effective subset of bands useful for detection problem.



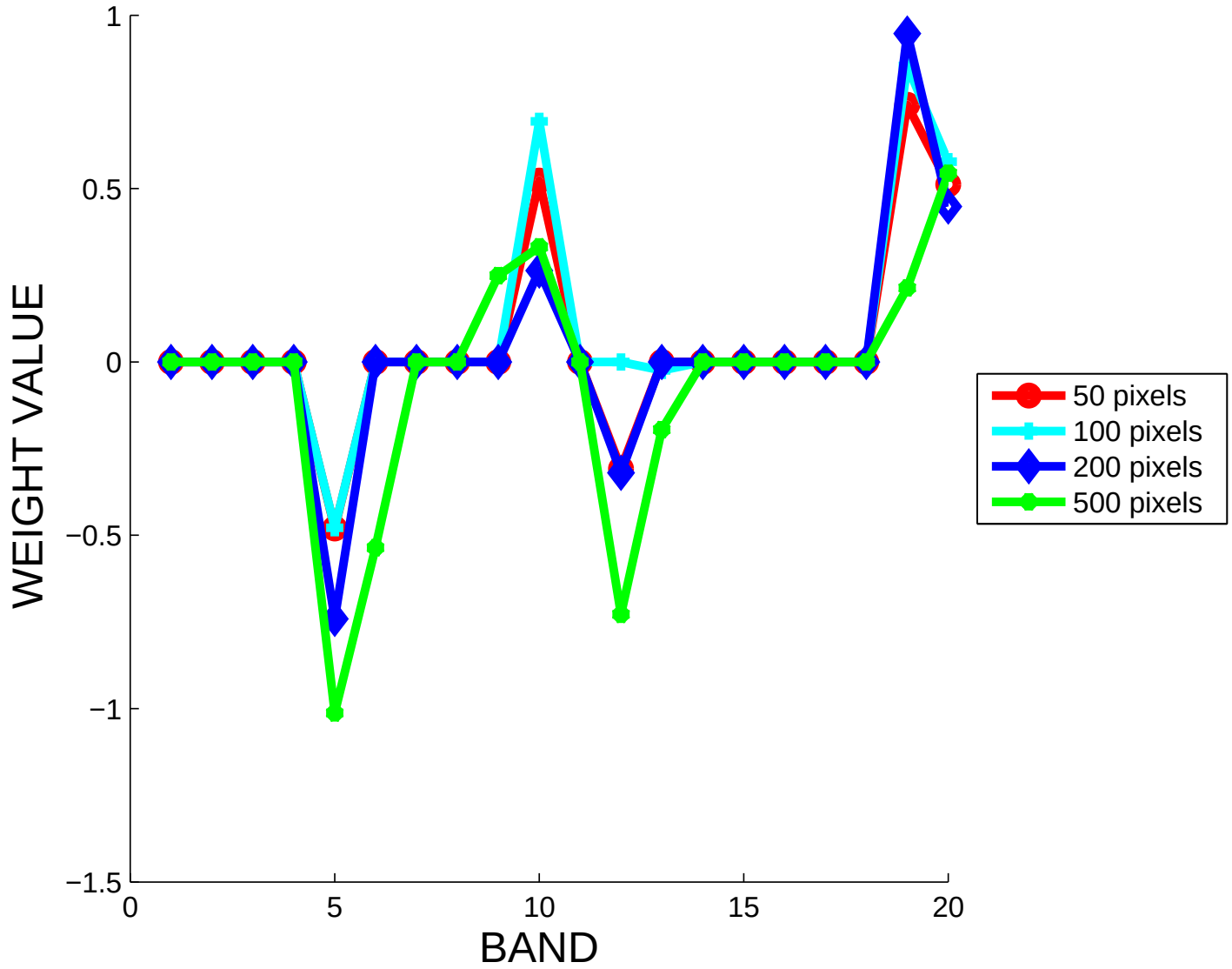
Febry-Pérot Data Set

- Interferometer collects 20 images from different wavelengths during a single scanning.
- Each image is composed of 256×256 pixels.

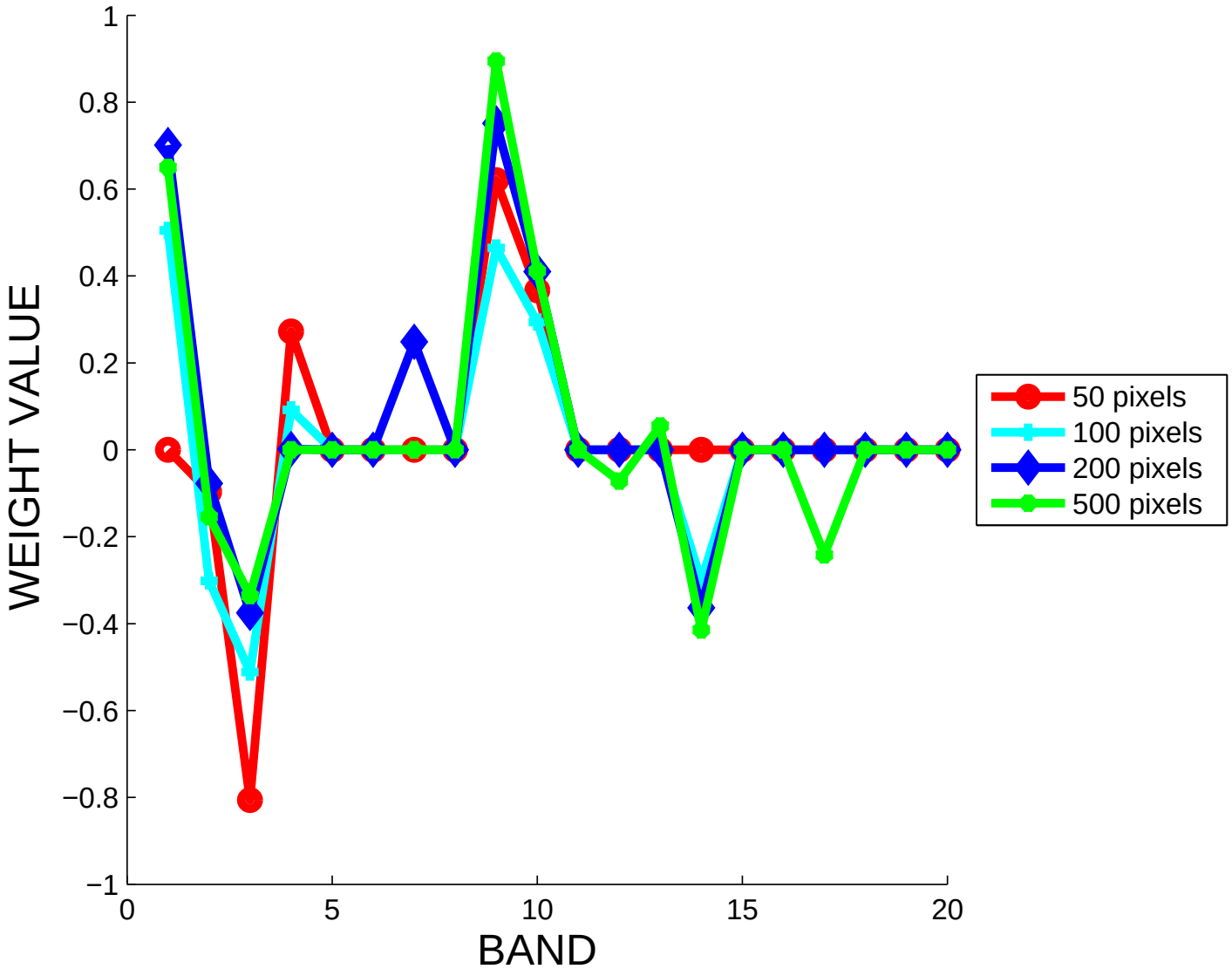


- Three chemicals (classes): Glacial Acetic Acid (GAA), Triethyl Phosphate (TEP), and Methyl Salicylate (MeS).
- Data pre-processing: background removal; pixels from each cube are clustered into groups; Each pixel x is standardized (has zero mean and unit variance).
- This resulted in a data set of $\sim$ 12700 pixels GAA, $\sim$ 13200 pixels MeS, and $\sim$ 12000 pixels TEP, each pixel living in $\mathbb{R}^{20}$.

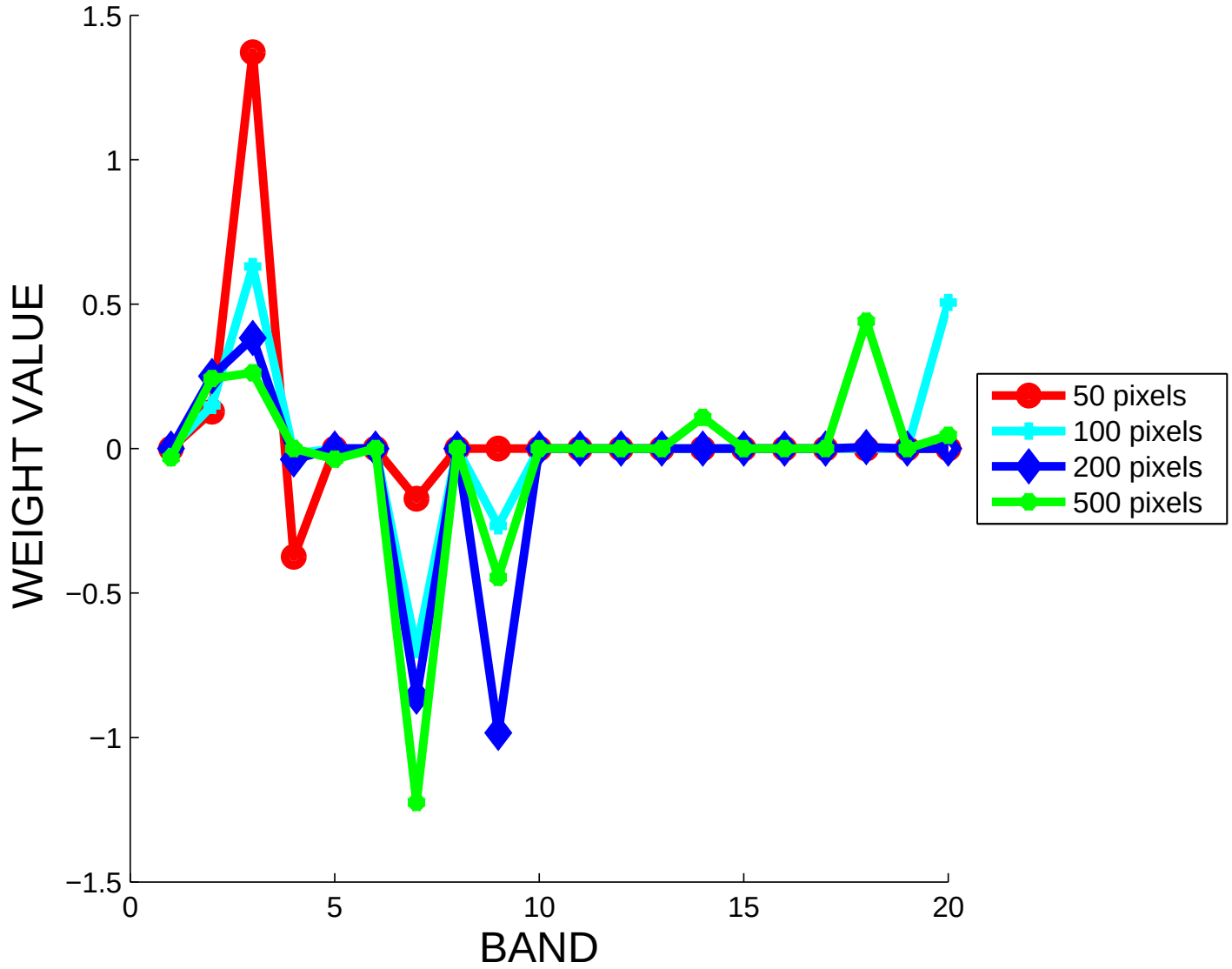
GAA and MeS Data Sets: Weights



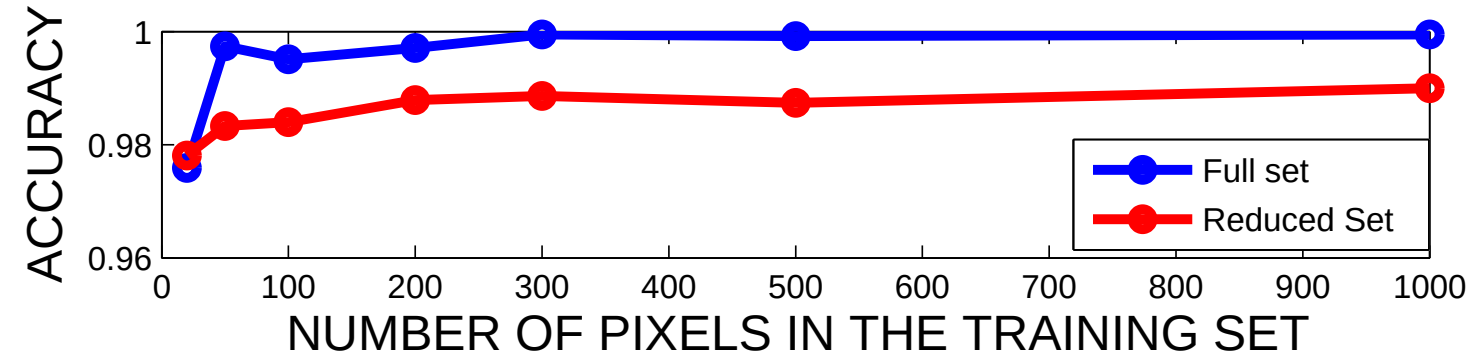
GAA and TEP Data Sets: Weights

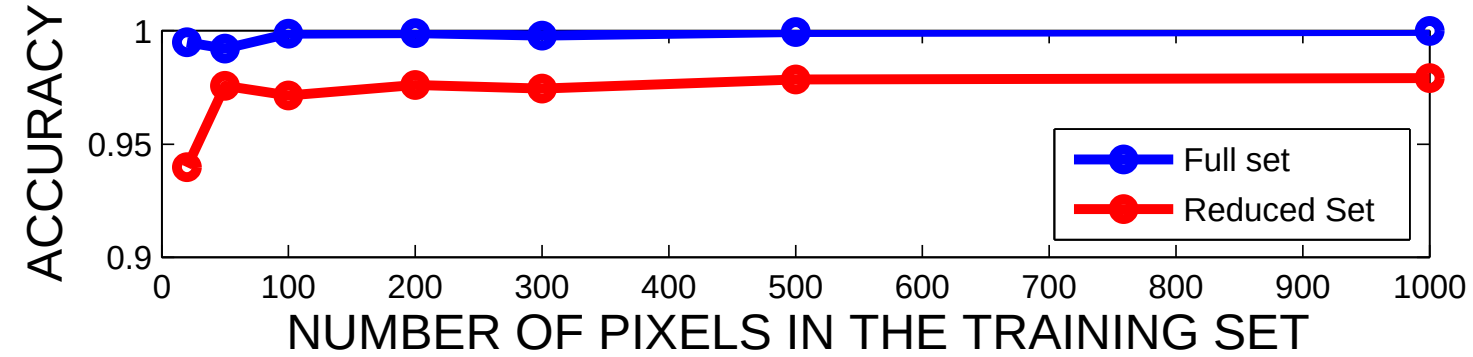


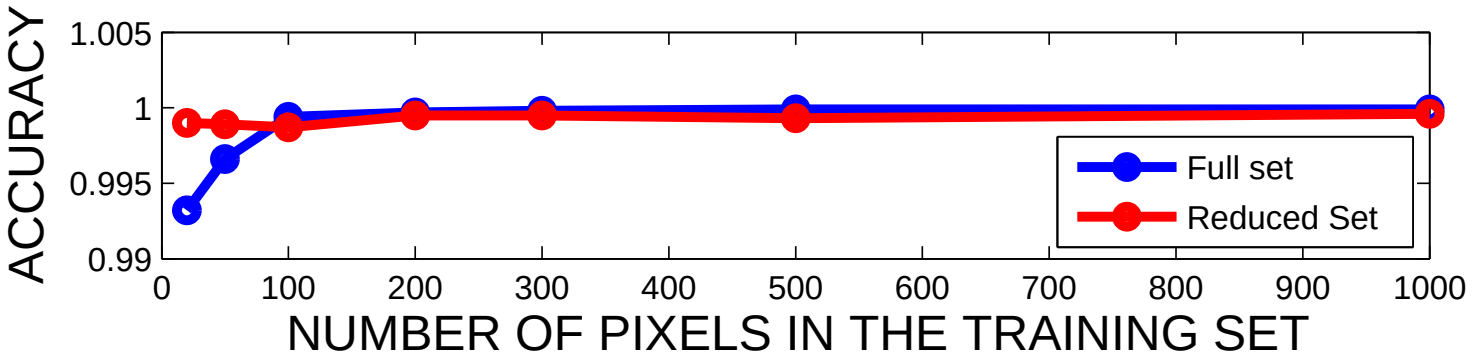
MeS and TEP Data Sets: Weights



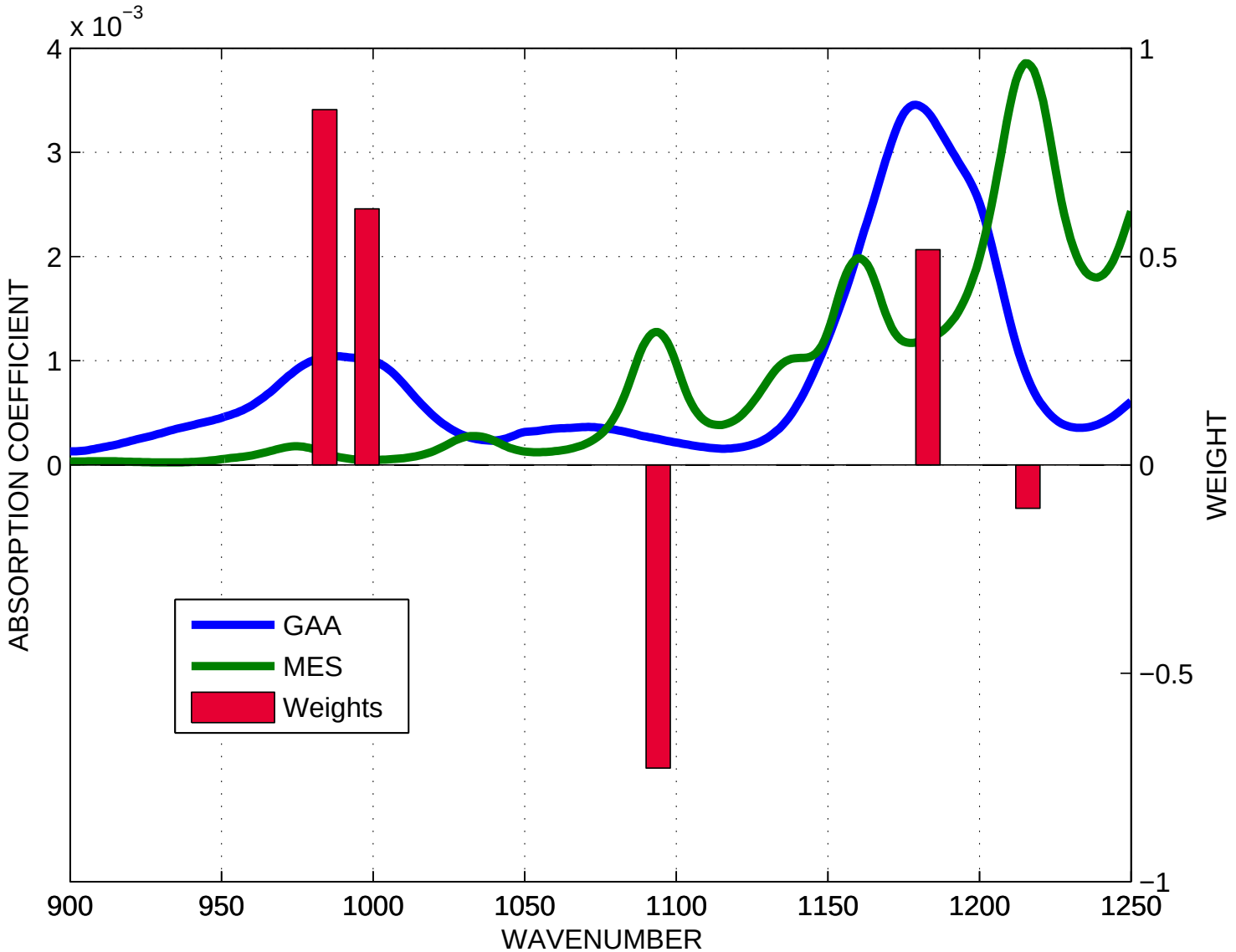
Classification Accuracy on Testing Data: Full Set of Bands vs. Reduced Set of Bands



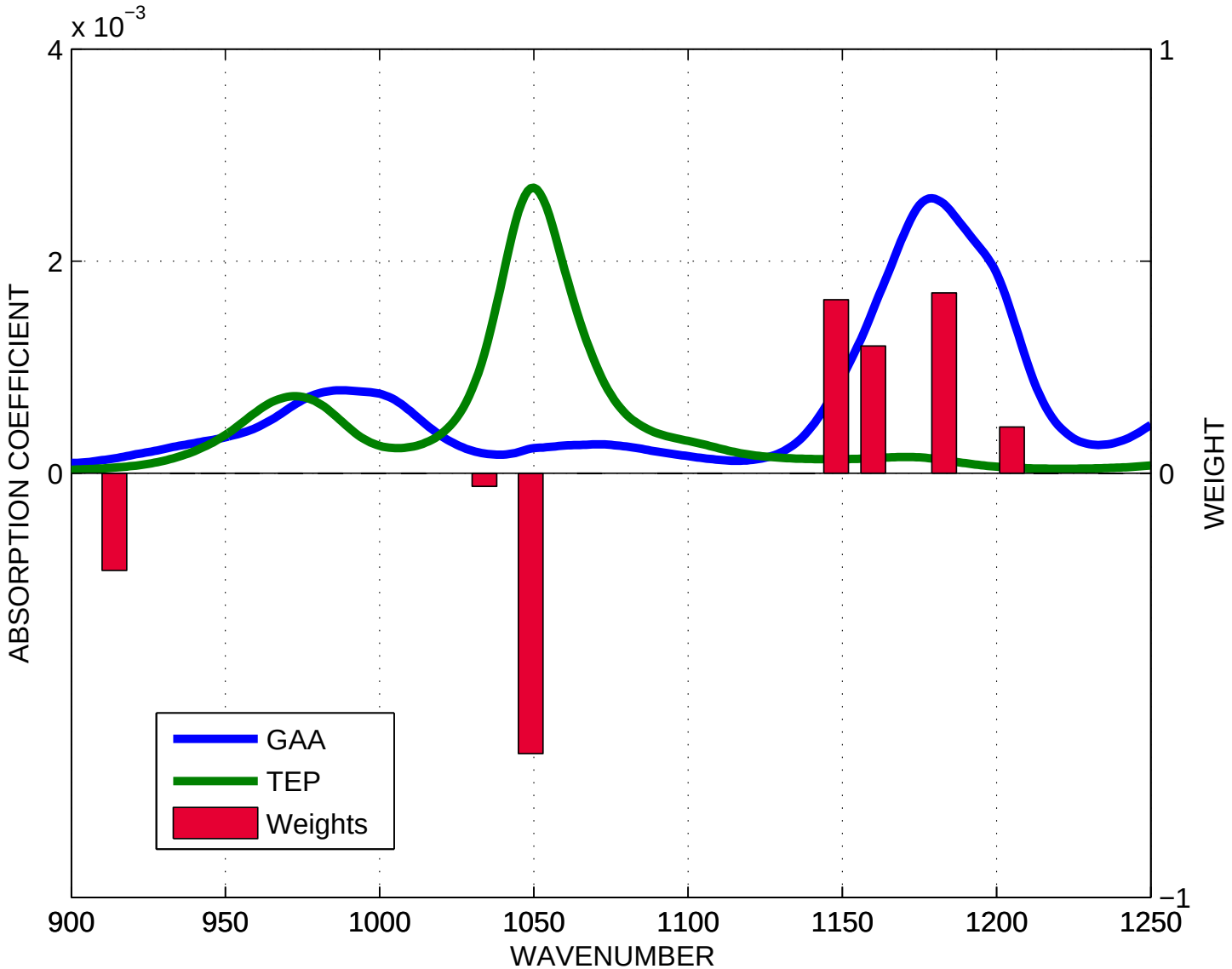




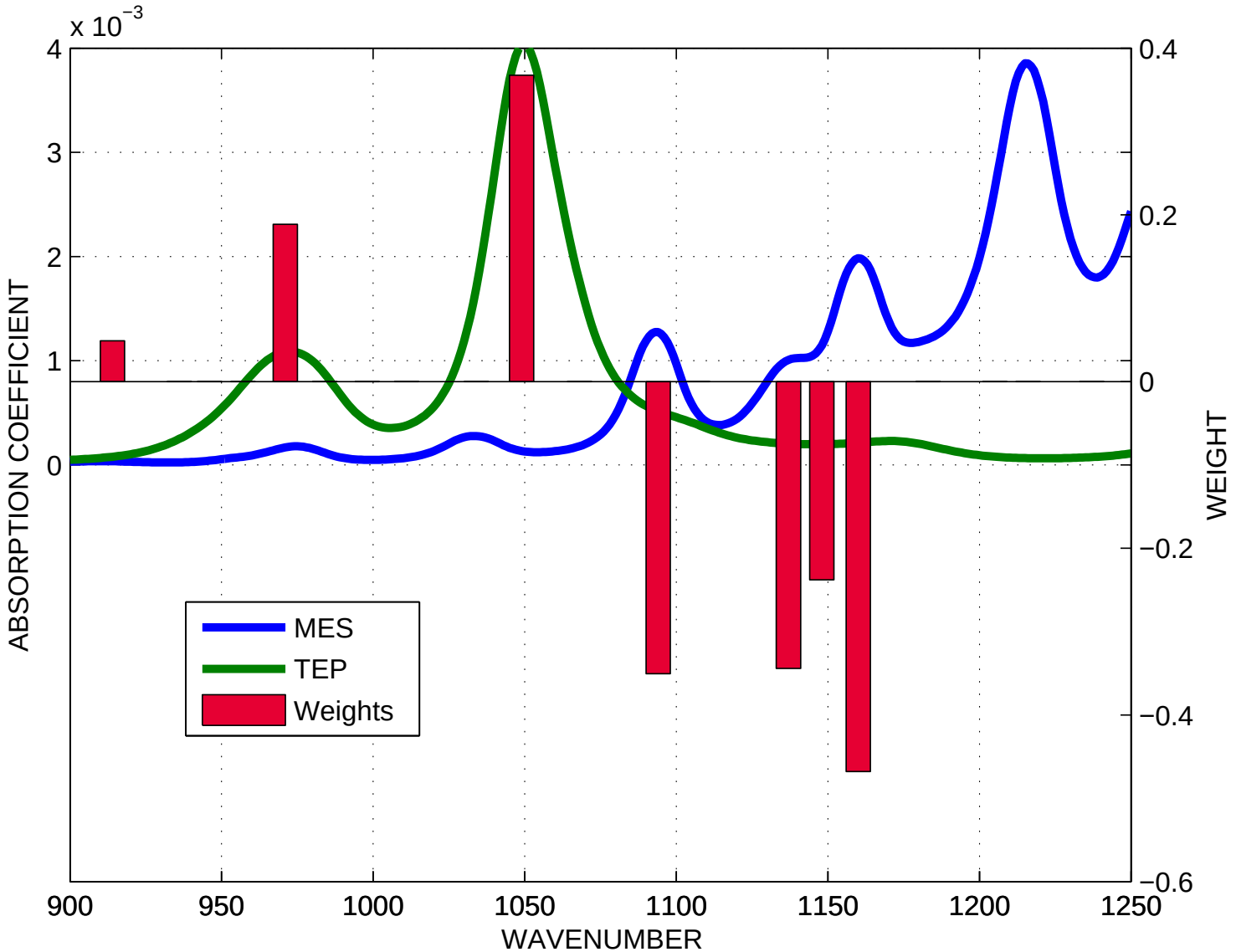
Reference Spectra: GAA and MeS



Reference Spectra: GAA and TEP



Reference Spectra: MeS and TEP



Sources of Information

- Tutorial: C.J.C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*, 1998
- Books on SVMs:
-
  - V. Vapnik, *Statistical Learning Theory*, 1998
  - N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, 2000
- Optimization Book: R. J. Vanderbei,*Linear Programming Foundations and Extensions*, 2008
- Google
- Me