Statistical Methods in Artificial Intelligence
CSE471 - Monsoon 2015 : Lecture 04

Avinash Sharma

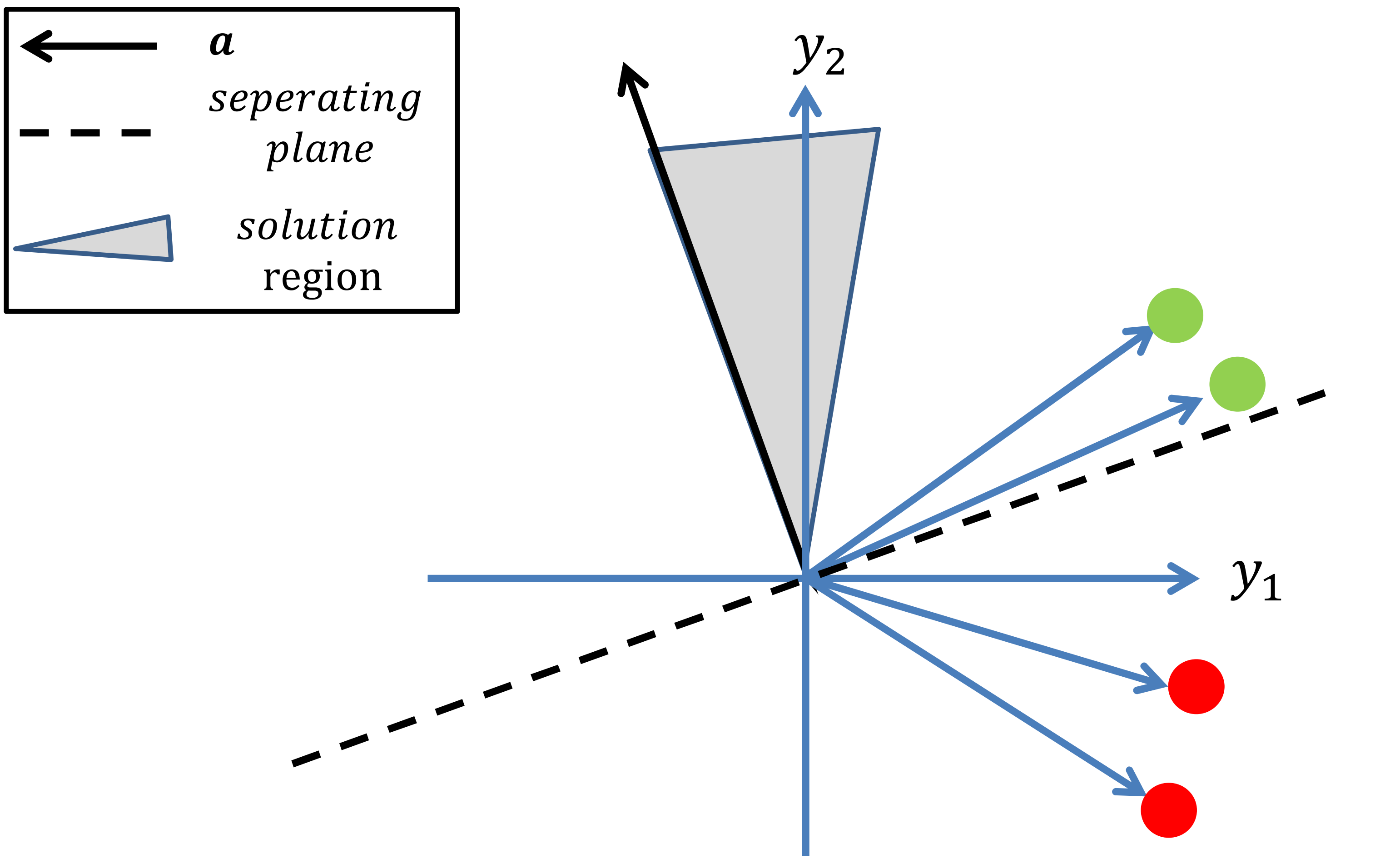
CVIT, IIIT Hyderabad

Lecture 04: Plan

- Recap
- Learning LDF: Basic Gradient Descend
- Perceptron Criterion Function
- Batch Perceptron
- Single Sample Perceptron
- Relaxation Procedures
- Non Separable Behavior
- Minimum Squared Error Procedures
- LMS and Ho-Kashyap Procedures

Two-Category Linearly Separable Case

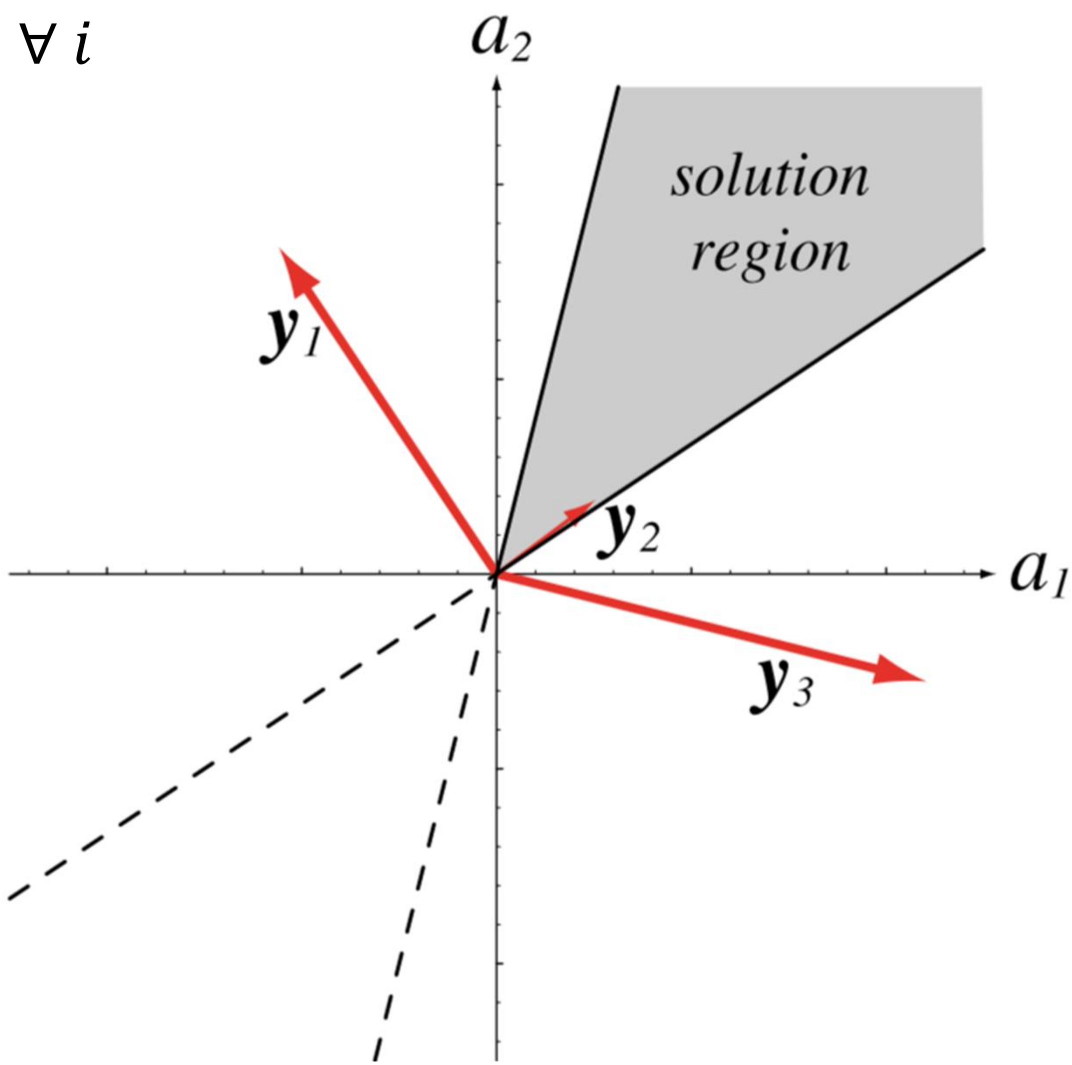
$$g(\boldsymbol{X}) = \boldsymbol{a}^T\boldsymbol{Y} = \sum_{i=1}^{\hat{d}} a_i y_i \quad = \begin{cases} > 0 \quad (+ve) & class\ A \\ < 0 \quad (-ve) & class\ B \\ = 0\ Decision\ Boundary & \end{cases}$$



Two-Category Linearly Separable Case

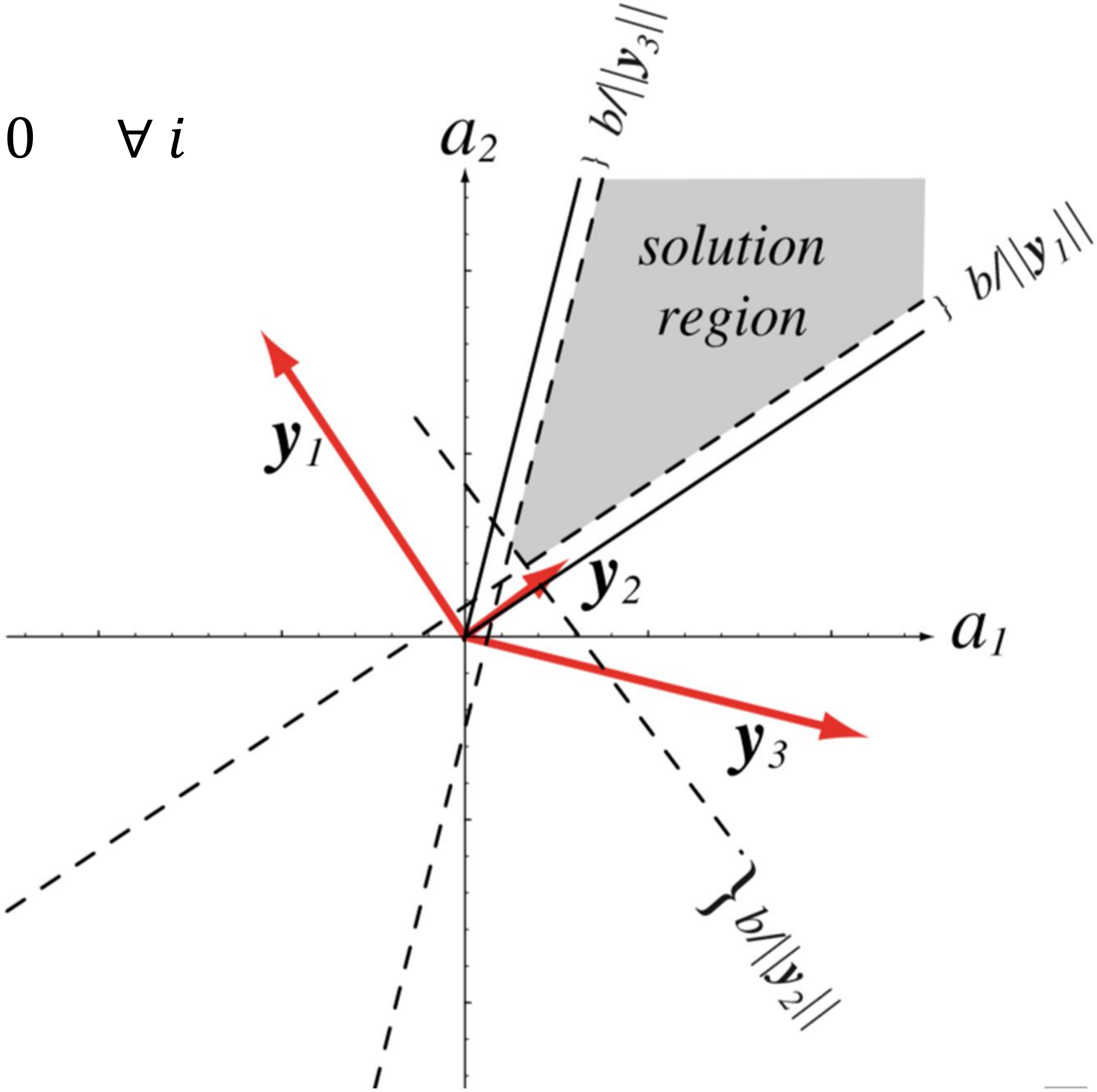
$$\boldsymbol{a}^T \boldsymbol{y}_i > 0 \quad \forall\, i$$

Data vector



Two-Category Linearly Separable Case

$$\boldsymbol{a}^T \boldsymbol{y}_i \geq b > 0 \quad \forall\, i$$



Learning LDF: Basic Gradient Descend

- Define a scalar function $J(\boldsymbol{a})$ which captures classification error for specific boundary plane described by parameter $\boldsymbol{a}$

- Minimize $J(\boldsymbol{a})$ using ***gradient descent***.
  - Start with arbitrary value of $\boldsymbol{a}(1)$ for $k = 1$.
  - Iteratively refine estimate of $\boldsymbol{a}$:

    $$\boldsymbol{a}(k+1) = \boldsymbol{a}(k) - \eta(k)\nabla J(\boldsymbol{a}(k))$$

- η is the positive scale factor known as ***learning rate***
  - A too small η makes the convergence very slow
  - A too large η can diverge due to overshooting correct.

Basic Gradient Descend Algorithm

1. Initialize $\boldsymbol{a}, \theta$ (threshold), $\eta(.), k = 0$

2. do $k = k + 1$

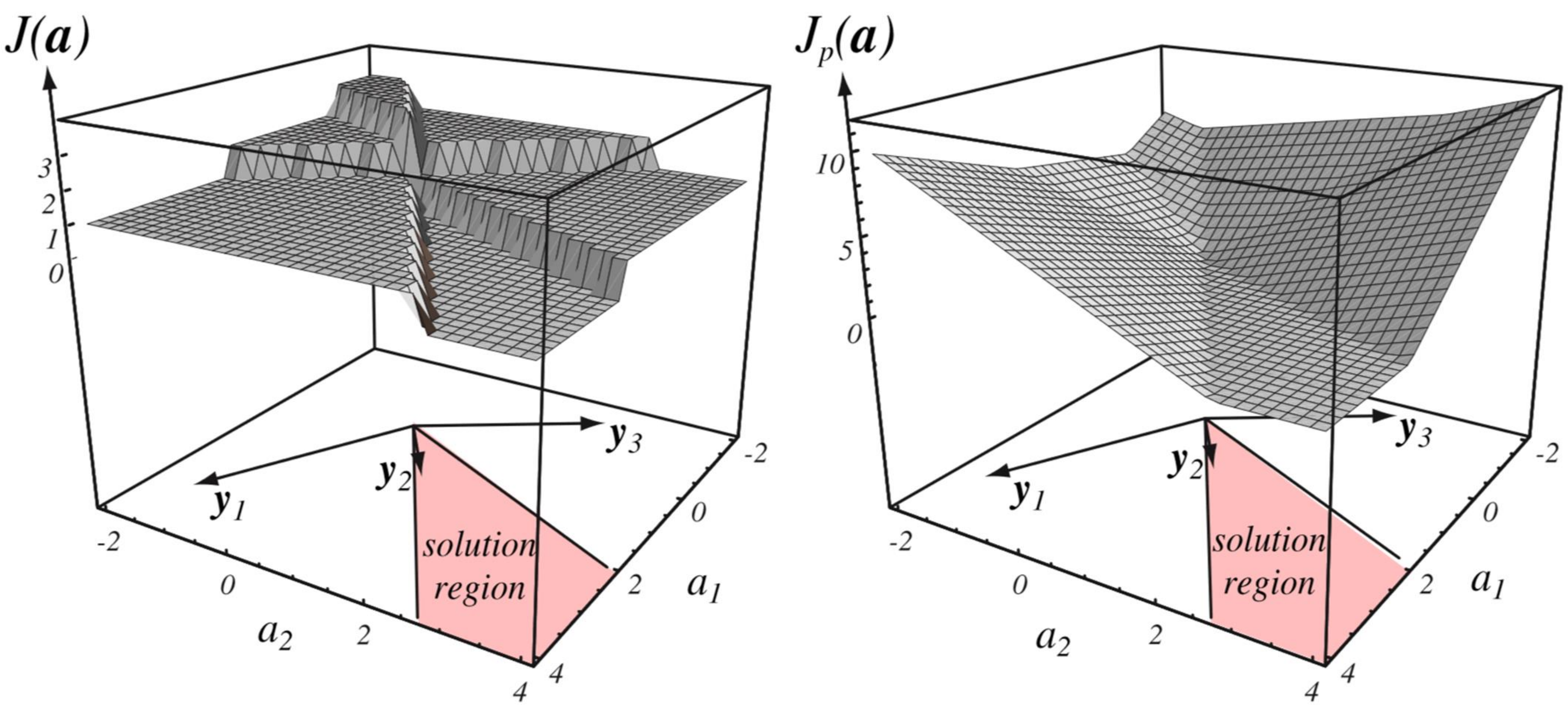
$$\boldsymbol{a}(k+1) = \boldsymbol{a}(k) - \eta(k)\nabla J(\boldsymbol{a}(k))$$

3. untill $|\eta(k)\nabla J(\boldsymbol{a}(k))| < \theta$

4. return $\boldsymbol{a}$

Perceptron Criterion Function

- Discrete v/s continuous function



Perceptron Criterion Function

- Discrete v/s continuous function
- $J_p(\boldsymbol{a}) = \sum_{\boldsymbol{y}\in\mathcal{Y}}(-\boldsymbol{a}^T\boldsymbol{y})$ where $\mathcal{Y}$ **is the set of misclassified samples**
- $J_p(\boldsymbol{a})$ is proportional to to the sum of the distances from all the miss classified samples to the decision boundary.
- Derivative of $J_p(\boldsymbol{a})$

$$\nabla J_p(\boldsymbol{a}) = \sum_{\boldsymbol{y}\in\mathcal{Y}}(-\boldsymbol{y})$$

- $\boldsymbol{a}(k+1) = \boldsymbol{a}(k) + \eta(k)\sum_{\boldsymbol{y}\in\mathcal{Y}_k}\boldsymbol{y}$

Batch Perceptron

1. Initialize $\boldsymbol{a}, \theta$ (threshold), $\eta(.), k = 0$

2. do $k = k + 1$

$$\boldsymbol{a}(k+1) = \boldsymbol{a}(k) + \eta(k) \sum_{\boldsymbol{y} \in \mathcal{Y}_k} \boldsymbol{y}$$

3. untill $\left|\eta(k) \sum_{\boldsymbol{y} \in \mathcal{Y}_k} (-\boldsymbol{y})\right| < \theta$

4. return $\boldsymbol{a}$

Single Sample Perceptron

1. Initialize $\boldsymbol{a}, k = 0$

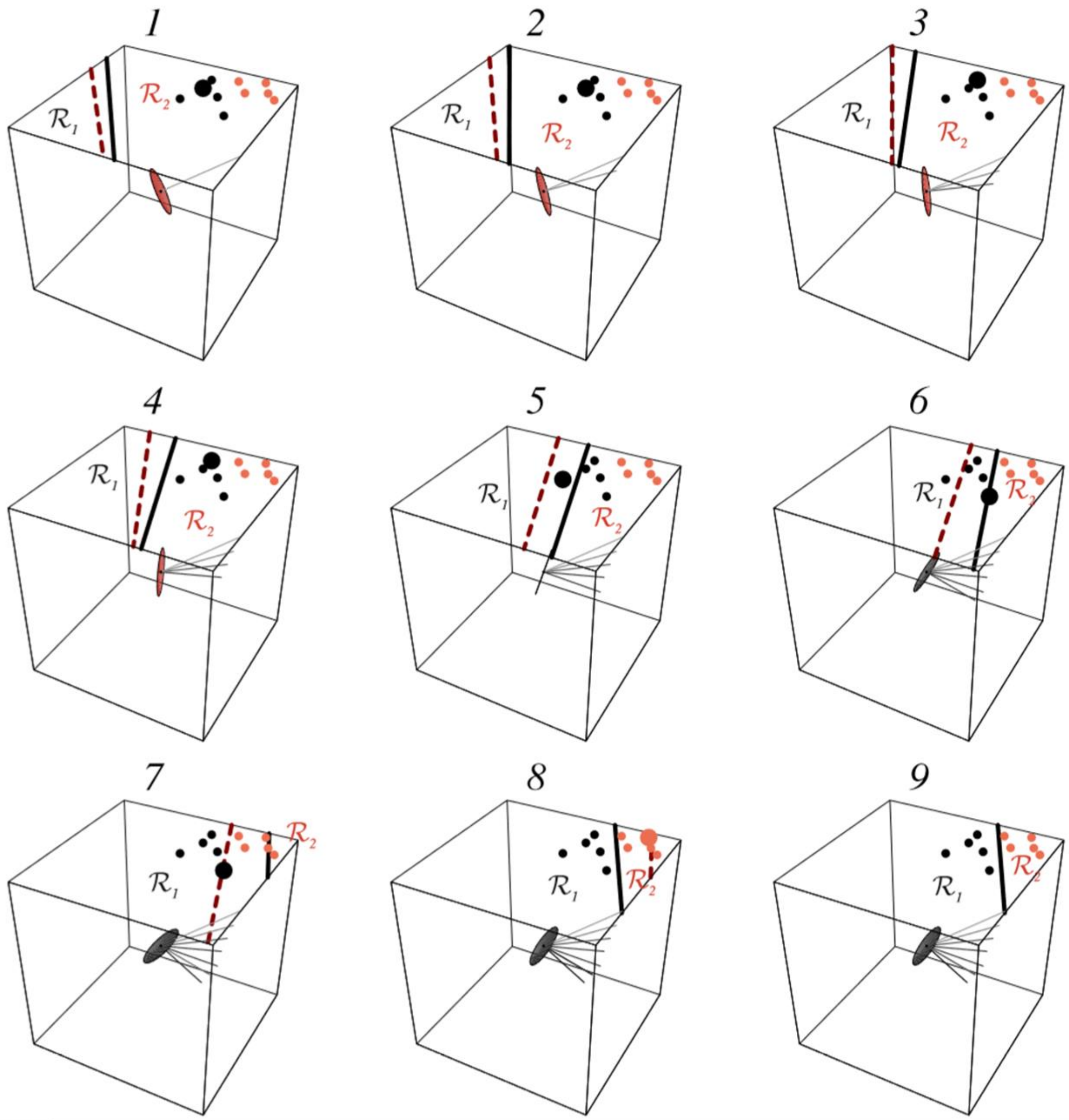
2. do $k = mod(k + 1, n)$

$$\boldsymbol{a} = \boldsymbol{a} + \boldsymbol{y}^k$$

3. untill all patterns are correctly classified

4. return $\boldsymbol{a}$

Single Sample Perceptron



Relaxation Procedures

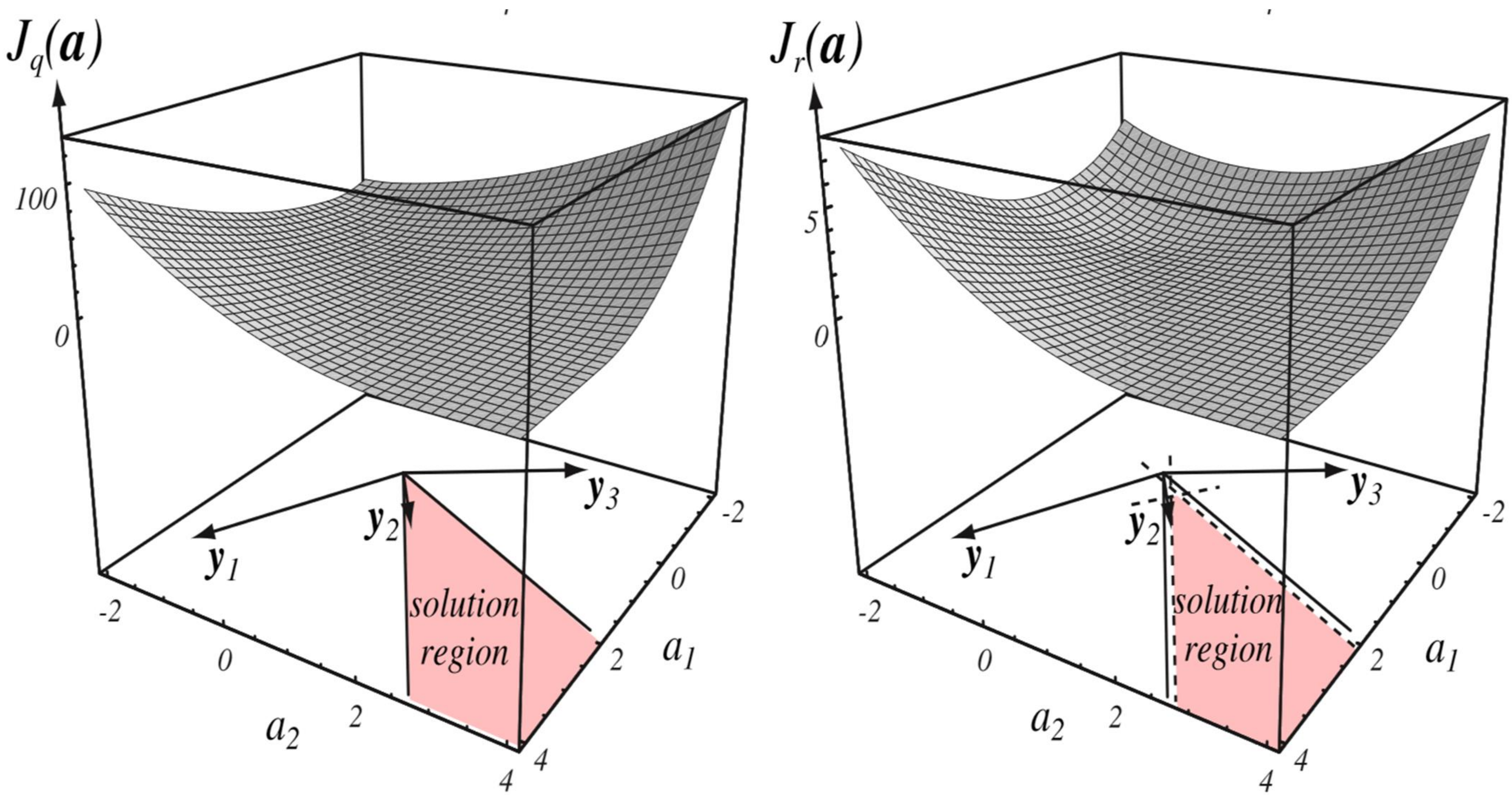
- These are broader class of criterion functions and associated minimization methods.

- $J_q(\boldsymbol{a}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} (\boldsymbol{a}^T \boldsymbol{y})^2$

- Problems:
  - convergence to boundary
  - dominated by the longest sample vector

- $J_r(\boldsymbol{a}) = \frac{1}{2} \sum_{\boldsymbol{y} \in \mathcal{Y}} \frac{(\boldsymbol{a}^T \boldsymbol{y} - b)^2}{\|\boldsymbol{y}\|^2}$ and $\nabla J_r(\boldsymbol{a}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} \frac{\boldsymbol{a}^T \boldsymbol{y} - b}{\|\boldsymbol{y}\|^2} \boldsymbol{y}$

Relaxation Procedures



Single Sample Relaxation with Margin

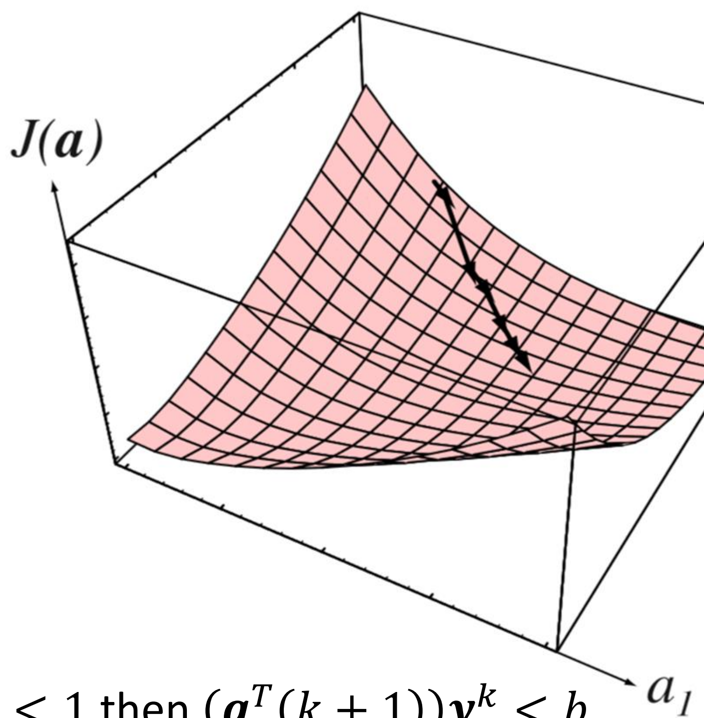
- Single Sample relaxation with margin

1. Initialize $\boldsymbol{a}, \eta(\cdot), k = 0$
2. do $k = mod(k+1, n)$

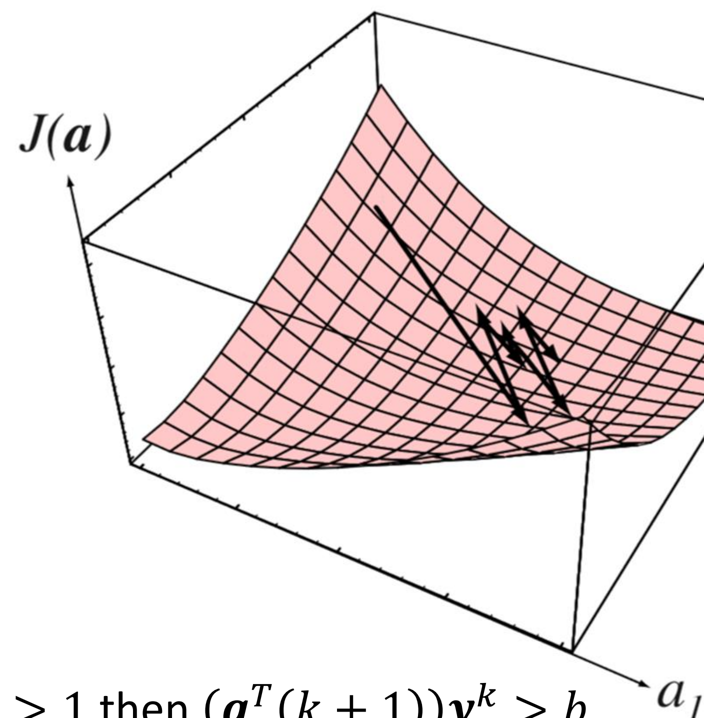
   if $\boldsymbol{a}^T \boldsymbol{y}^k \le b$ then $\boldsymbol{a} = \boldsymbol{a} + \eta(k)\frac{(b-\boldsymbol{a}^T\boldsymbol{y}^k)^2}{\|\boldsymbol{y}^k\|^2}\boldsymbol{y}^k$
3. untill $\boldsymbol{a}^T \boldsymbol{y}^k > b$ for all $\boldsymbol{y}^k$
4. return $\boldsymbol{a}$

Over/Under-Relaxation

- $r(k) = \frac{(b - \boldsymbol{a}^T \boldsymbol{y}^k)^2}{\|\boldsymbol{y}^k\|^2} \boldsymbol{y}^k$



$\eta < 1$ then $(\boldsymbol{a}^T(k+1))\boldsymbol{y}^k < b$



$\eta > 1$ then $(\boldsymbol{a}^T(k+1))\boldsymbol{y}^k > b$

Non Separable Behavior

- Error Correcting Procedures
- Generalization to unseen test data not guaranteed
- Fails to handle non-separable case
- Many Heuristic exists to handle non-separable cases:
  - Forced termination of loop
  - Annealing of η with increasing k

Minimum Squared Error Procedures

- MSE consider all samples instead of just miss-classified ones.

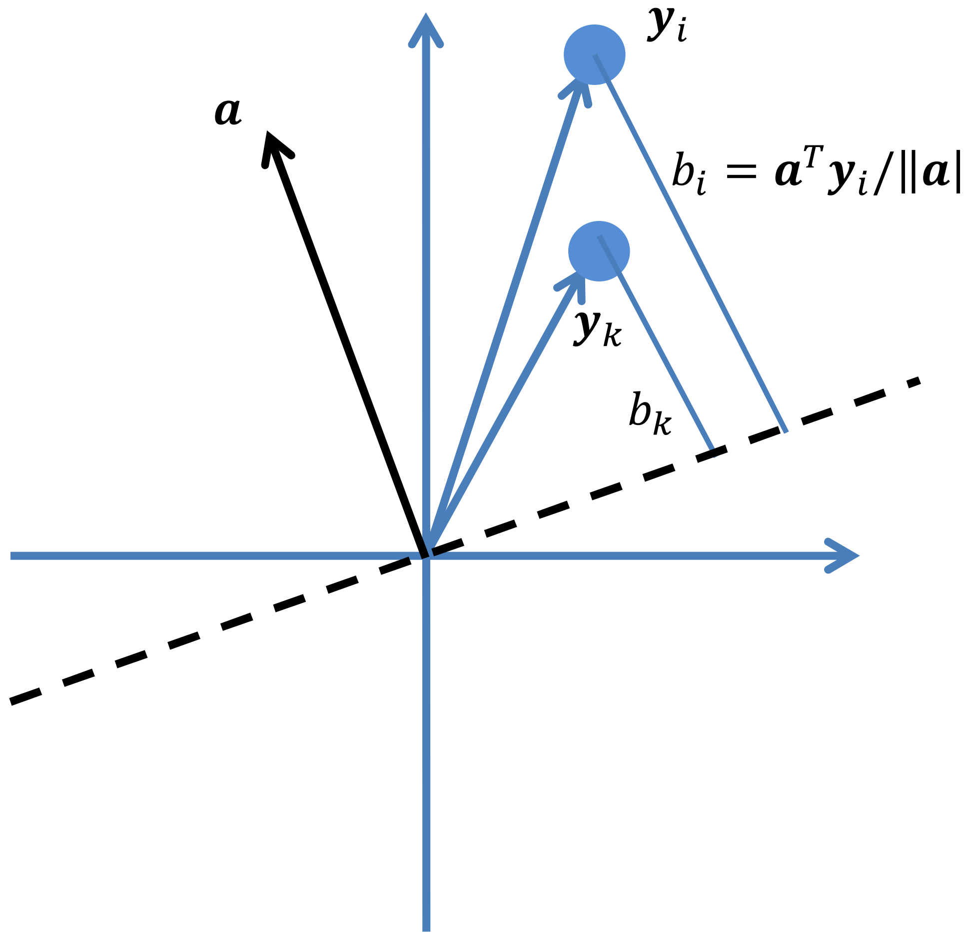
- Moved from problem of finding solution to a set of linear inequalities to a set of linear equations, i.e., $\boldsymbol{a}^T \boldsymbol{y}_i = b_i$ instead of $\boldsymbol{a}^T \boldsymbol{y}_i > 0$

Minimum Squared Error Procedures

- MSE consider all samples instead of just miss-classified ones.

$\boldsymbol{a}^T\boldsymbol{y}_i > 0$ for all samples $\boldsymbol{y}_i$

$\boldsymbol{a}^T\boldsymbol{y}_i = b_i$ for all samples $\boldsymbol{y}_i$
where $b_i > 0$



Minimum Squared Error Procedures

- $\boldsymbol{Y} = [\boldsymbol{y}_1 \quad \cdots \quad \boldsymbol{y}_n]^T$ be the set of all data points where $\boldsymbol{y}_i = [y_{i0} \quad \cdots \quad y_{id}]^T \in \mathbb{R}^{\widehat{\boldsymbol{d}}=(d+1)}$
- Let $\boldsymbol{a} = [a_0 \quad \cdots \quad a_d]^T$ and $\boldsymbol{b} = [b_1 \quad \cdots \quad b_n]^T$
- $\boldsymbol{Y}\boldsymbol{a} = \boldsymbol{b}$ (over-determined problem as $n \gg \hat{d}$)
- $\boldsymbol{a} = \boldsymbol{Y}^{-1}\boldsymbol{b}$ not possible (Y is rectangular and possibly singular)
- No exact solution ! We look for approximate solution.

Minimum Squared Error Procedures

- $\boldsymbol{e} = \boldsymbol{Y}\boldsymbol{a} - \boldsymbol{b}$ (Error definition)
- $J(\boldsymbol{a}) = \|\boldsymbol{e}\|^2 = \|\boldsymbol{Y}\boldsymbol{a} - \boldsymbol{b}\|^2$
- $\nabla J(\boldsymbol{a}) = 2\boldsymbol{Y}^T(\boldsymbol{Y}\boldsymbol{a} - \boldsymbol{b})$=0
- $\boldsymbol{Y}^T\boldsymbol{Y}\boldsymbol{a} = \boldsymbol{Y}^T\boldsymbol{b}$
- $\boldsymbol{a} = (\boldsymbol{Y}^T\boldsymbol{Y})^{-1}\boldsymbol{Y}^T\boldsymbol{b} = \boldsymbol{Y}^\dagger\boldsymbol{b}$

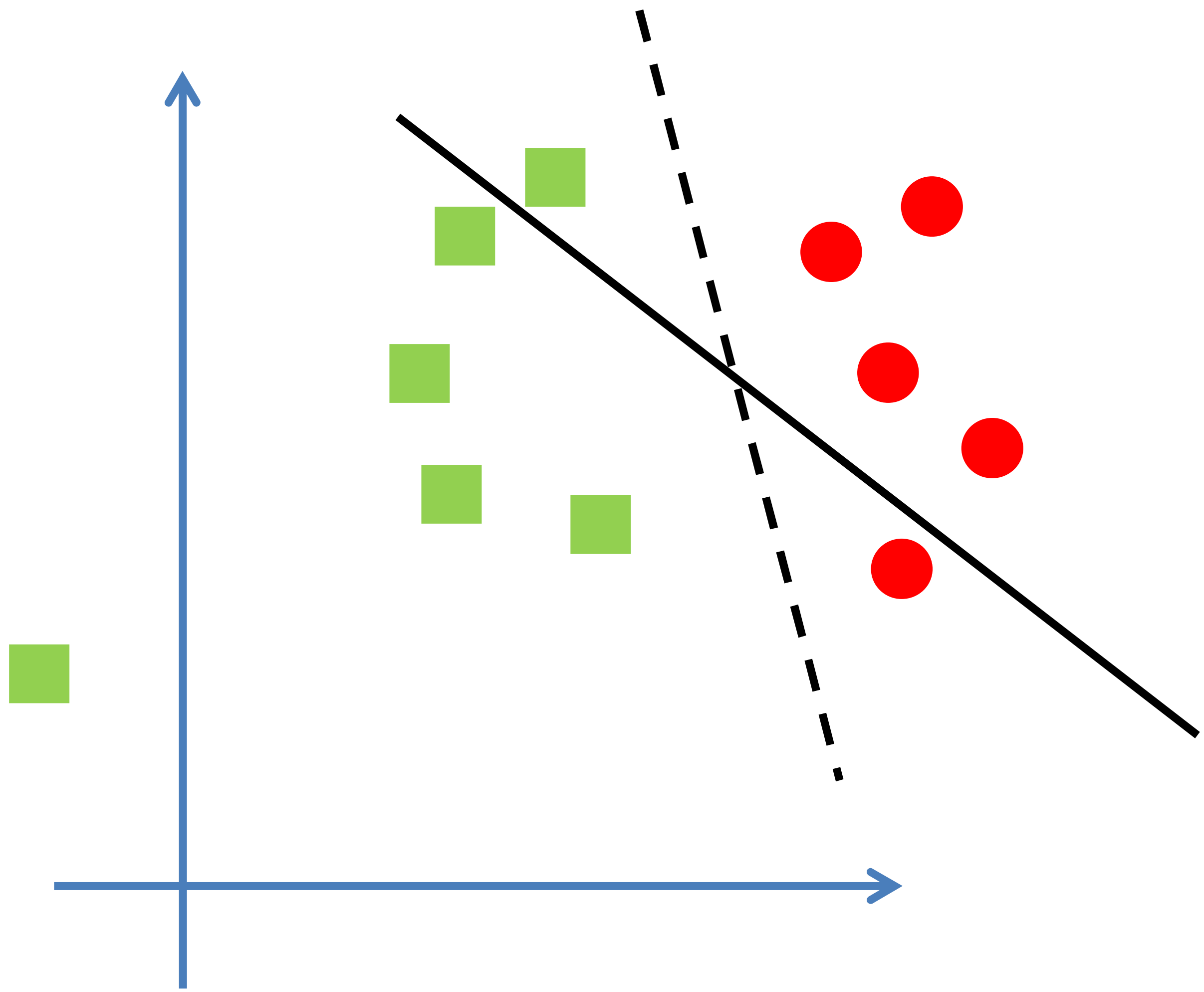
Minimum Squared Error Procedures

- $(\boldsymbol{Y}^T\boldsymbol{Y})$ is a square matrix and often non-singular and hence invertible.
- There could be many solutions for weight vector $\boldsymbol{a}$ based on choice of vector $\boldsymbol{b}$.
- A separating hyperplane is guaranteed if $(\boldsymbol{Ya} > \boldsymbol{0})$, i.e., $\forall i \ (\boldsymbol{a}^T\boldsymbol{y}_i) > 0$
- However, we have $\boldsymbol{Ya} \approx \boldsymbol{b}$, i.e., $\boldsymbol{Ya} = \boldsymbol{b} + \boldsymbol{\varepsilon}$.
- In practice, some entries of $\boldsymbol{b}$ can be negative if $|b_i| < |\varepsilon_i|$ and $\varepsilon_i < 0$.

Minimum Squared Error Procedures

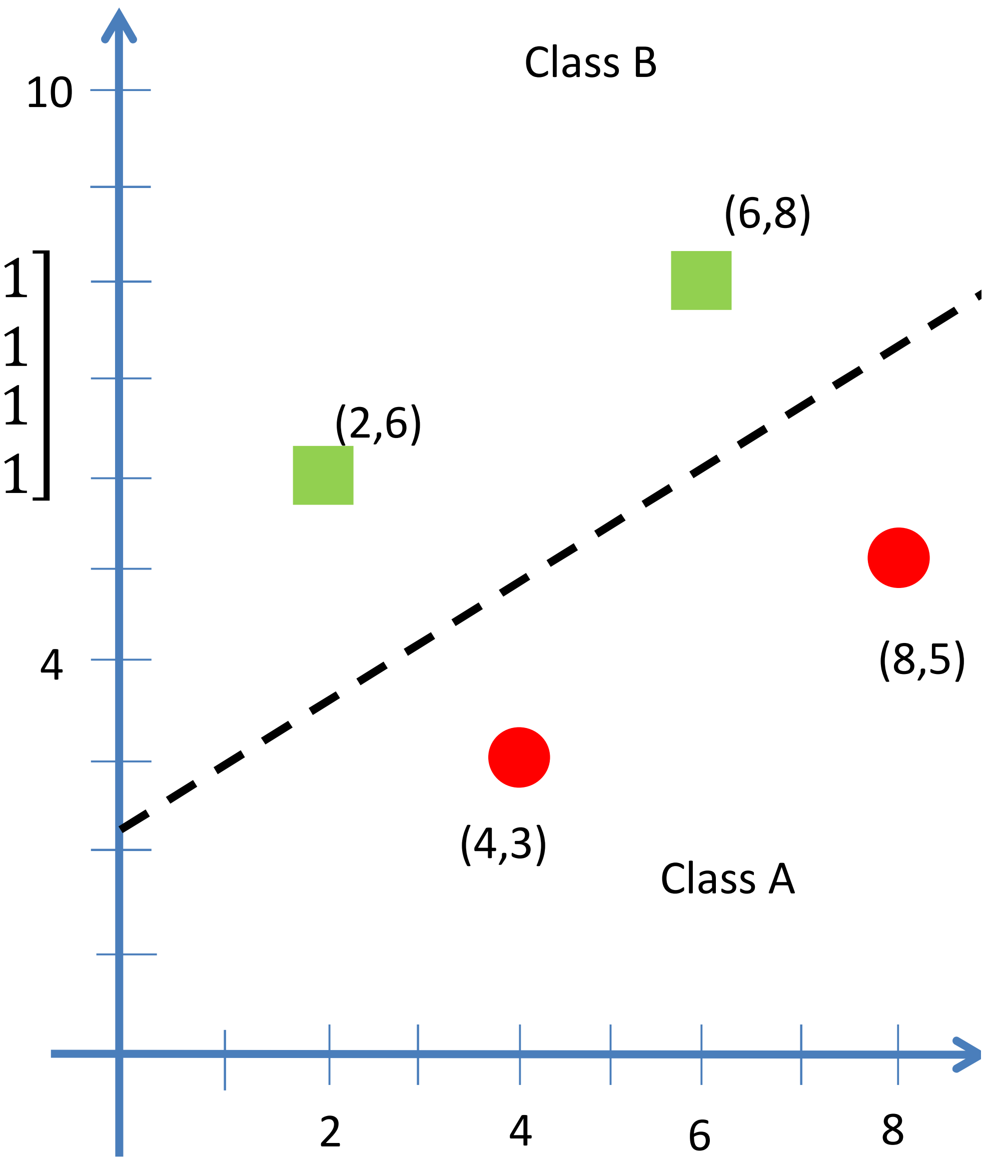
- Thus, even in linearly separable case, least square solution $\boldsymbol{a}$ *might not yield a separating hyperplane* but a ***reasonable*** one.
- An arbitrary scaling of $\boldsymbol{b}$ to overcome the -$\boldsymbol{\varepsilon}$ values is ***not helpful*** as it translates to scaling up the $\boldsymbol{a}$ vector.
- However, relative difference in elements of $\boldsymbol{b}$ can help in improving the classification, especially to handle the case of outlier data points.

Minimum Squared Error Procedures



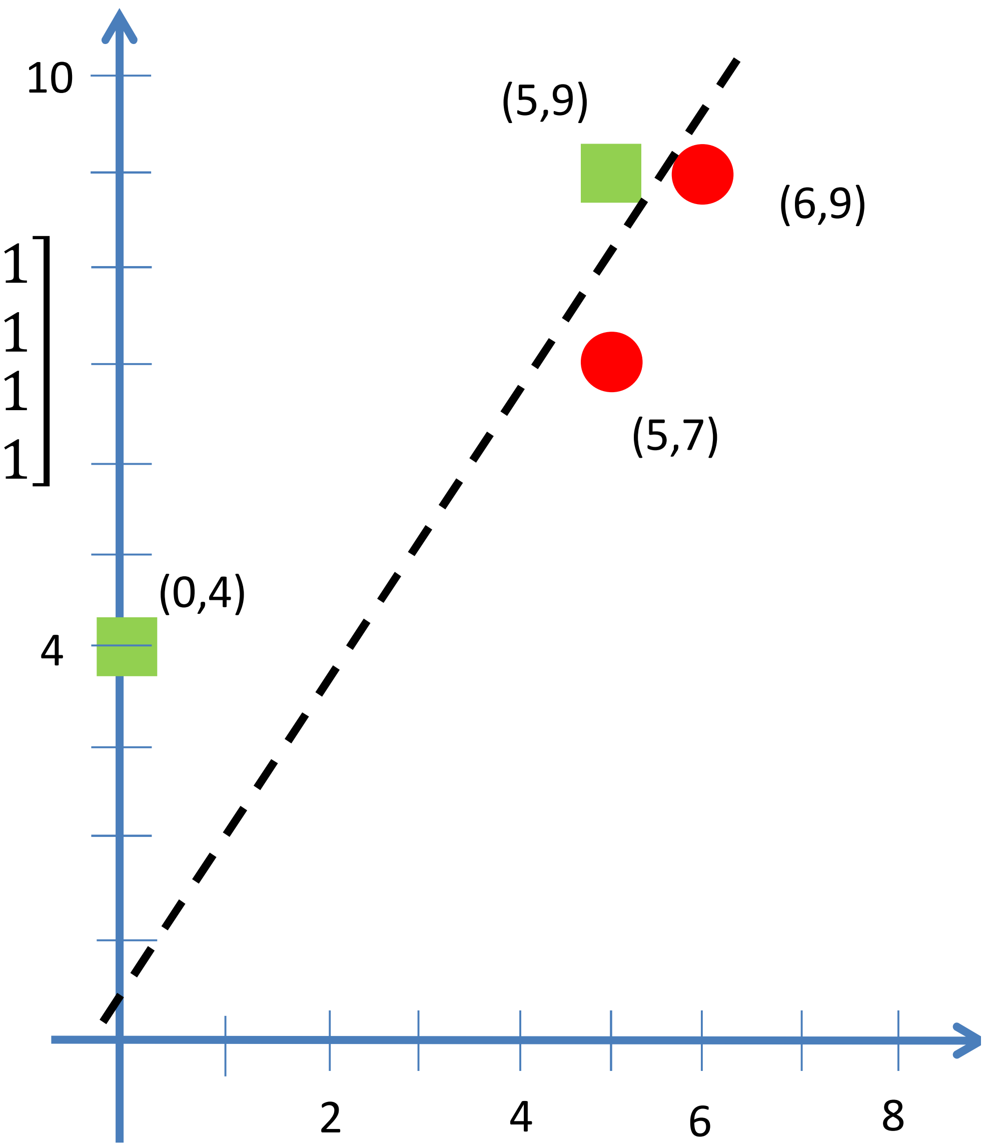
Example Walkthrough

- Class A: (8,5), (4,3)
- Class B: (2,6), (6,8)
- $\boldsymbol{Y}^T = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 8 & 4 & -2 & -6 \\ 5 & 3 & -6 & -8 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$
- $\boldsymbol{a} = \boldsymbol{Y}^{\dagger}\boldsymbol{b} = \begin{bmatrix} 1.5 \\ 0.25 \\ -0.5 \end{bmatrix}$
- $\boldsymbol{Y}\boldsymbol{a} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$



Example Walkthrough

- Class A: (6,9), (5,7)
- Class B: (5,9), (0,4)
- $\boldsymbol{Y}^T = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 6 & 5 & -5 & 0 \\ 9 & 7 & -9 & -4 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$
- $\boldsymbol{a} = \boldsymbol{Y}^\dagger \boldsymbol{b} = \begin{bmatrix} 2.66 \\ 1.04 \\ -0.94 \end{bmatrix}$
- $\boldsymbol{Ya} = \begin{bmatrix} 0.43 \\ 1.28 \\ 0.60 \\ 1.11 \end{bmatrix}$



Example Walkthrough

- Class A: (6,9), (5,7)
- Class B: (5,9),(0,10)
- $\boldsymbol{Y}^T = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 6 & 5 & -5 & 0 \\ 9 & 7 & -9 & -10 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$
- $\boldsymbol{a} = \boldsymbol{Y}^{\dagger}\boldsymbol{b} = \begin{bmatrix} 3.21 \\ 0.15 \\ -0.43 \end{bmatrix}$
- $\boldsymbol{Ya} = \begin{bmatrix} 0.19 \\ 0.91 \\ -0.04 \\ 1.16 \end{bmatrix}$