MS&E 214
Optimization via Case Studies

Week 4: Pattern classification

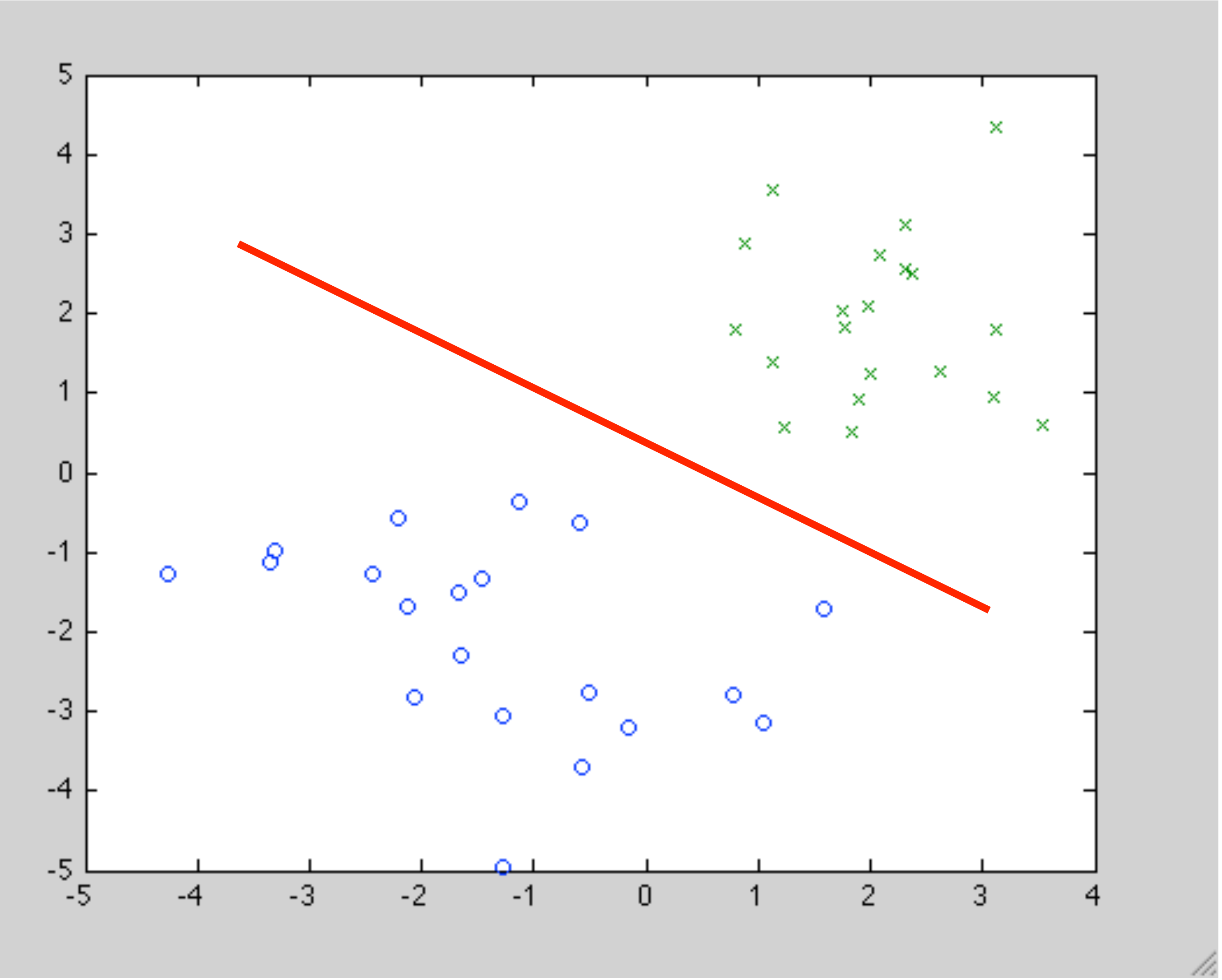
Ashish Goel

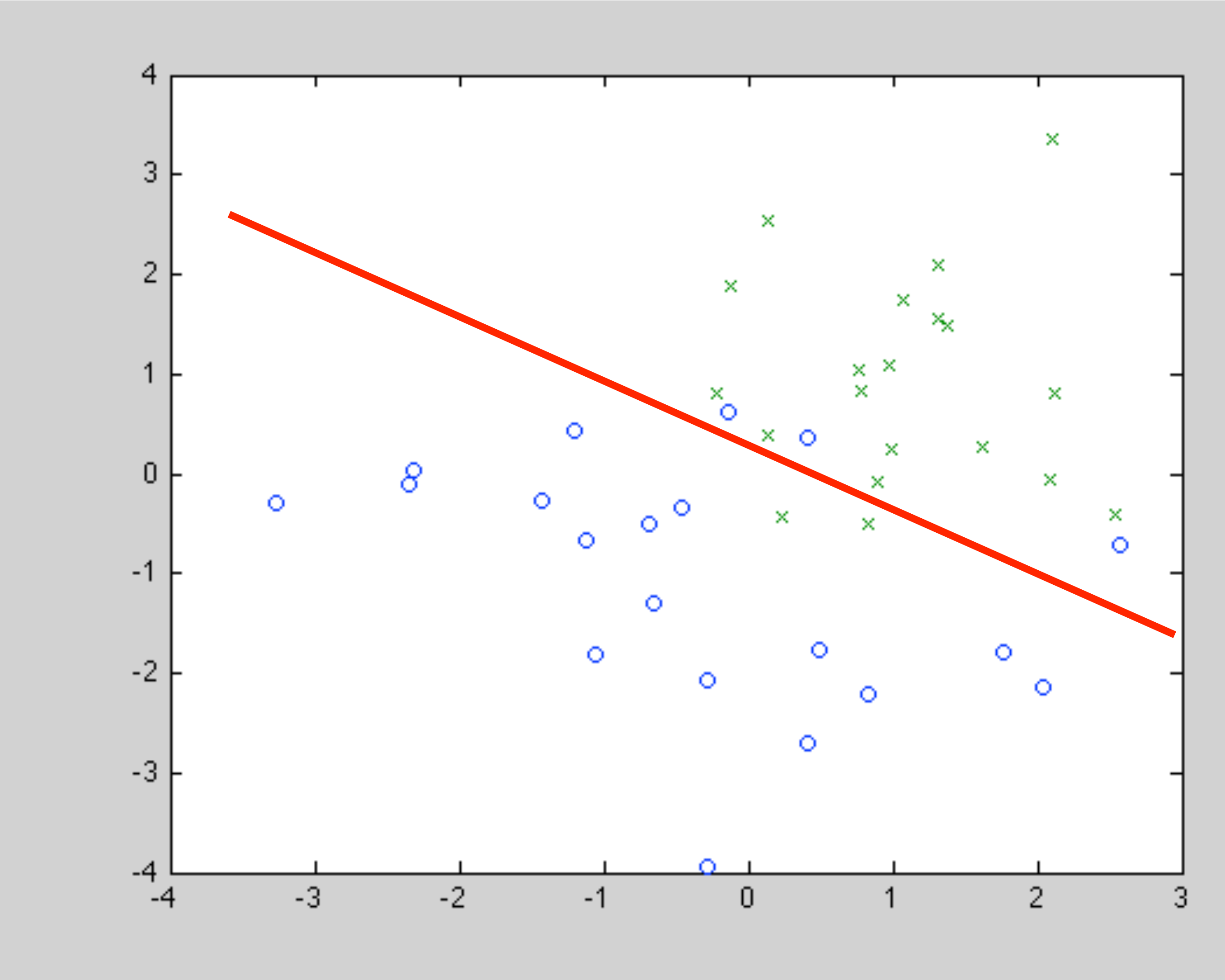
(based on slides by Professor Benjamin Van Roy)

Roadmap

- We are on module 2: Machine Learning
- Two major canonical problems:
  - Regression: Already completed
    - Used the Max-Min trick for absolute value
    - Introduced Quadratic programming
  - Binary Classification: Next
    - Using the max-min trick
    - Using quadratic programming
- Class goals after this:
  - Comfortable with Basic LPs, Min-Max/Max-Min trick, Quadratic Programming
  - Understand the optimization behind basic ML algorithms
- Discussion Point: Why study optimization when there are libraries available for all this?

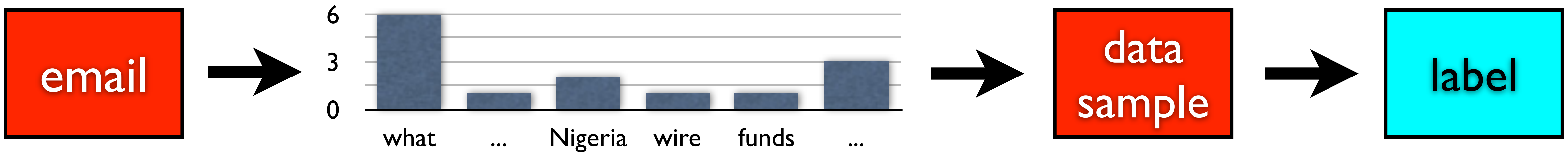
Pattern Classification



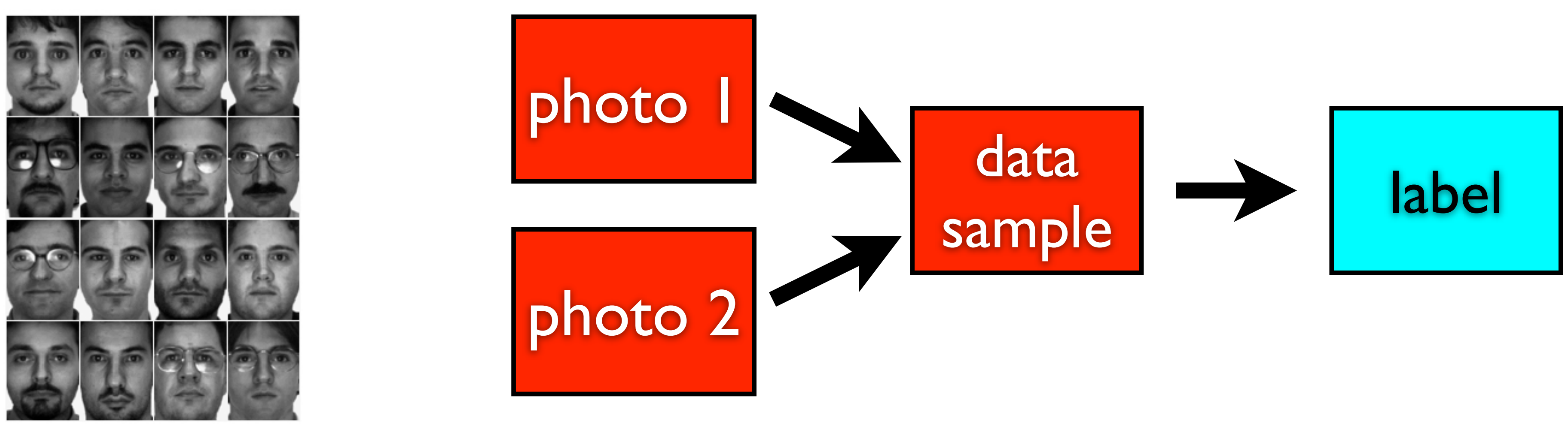


Examples

- Spam detection



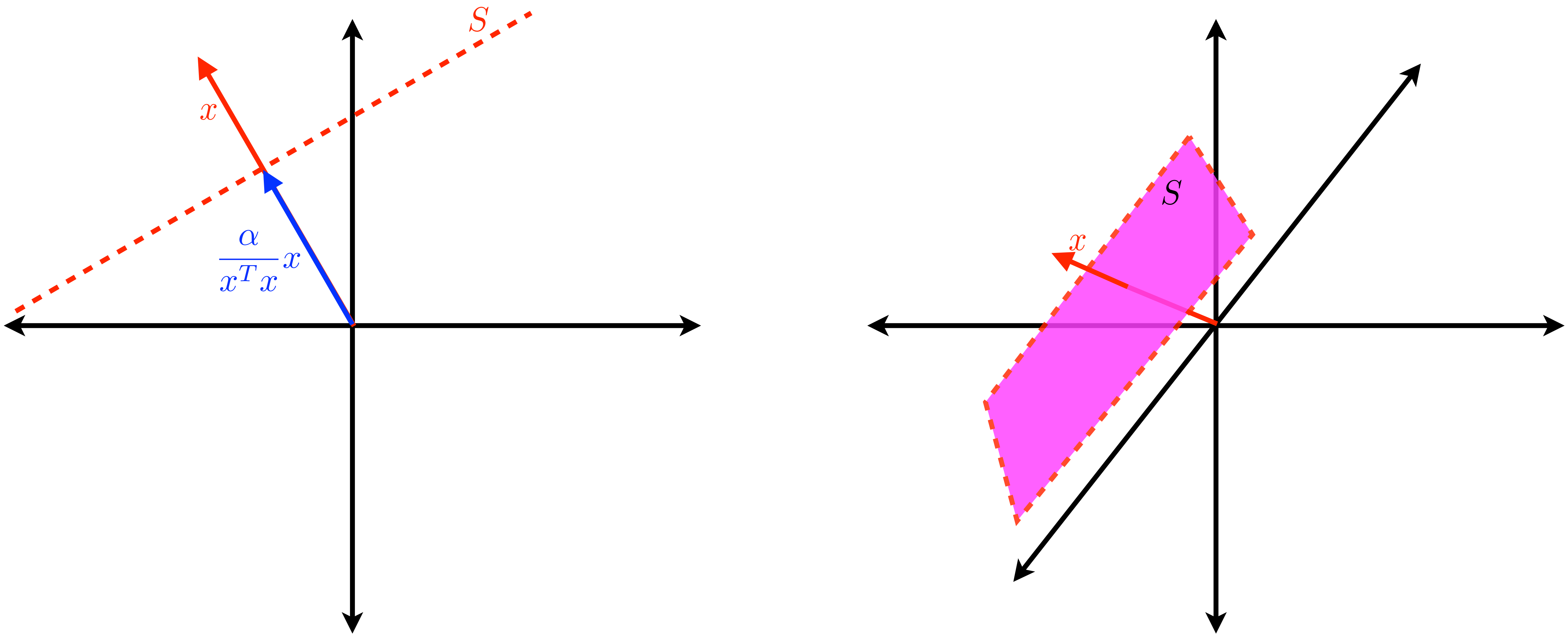
- Face recognition



- Medical diagnosis

Hyperplanes

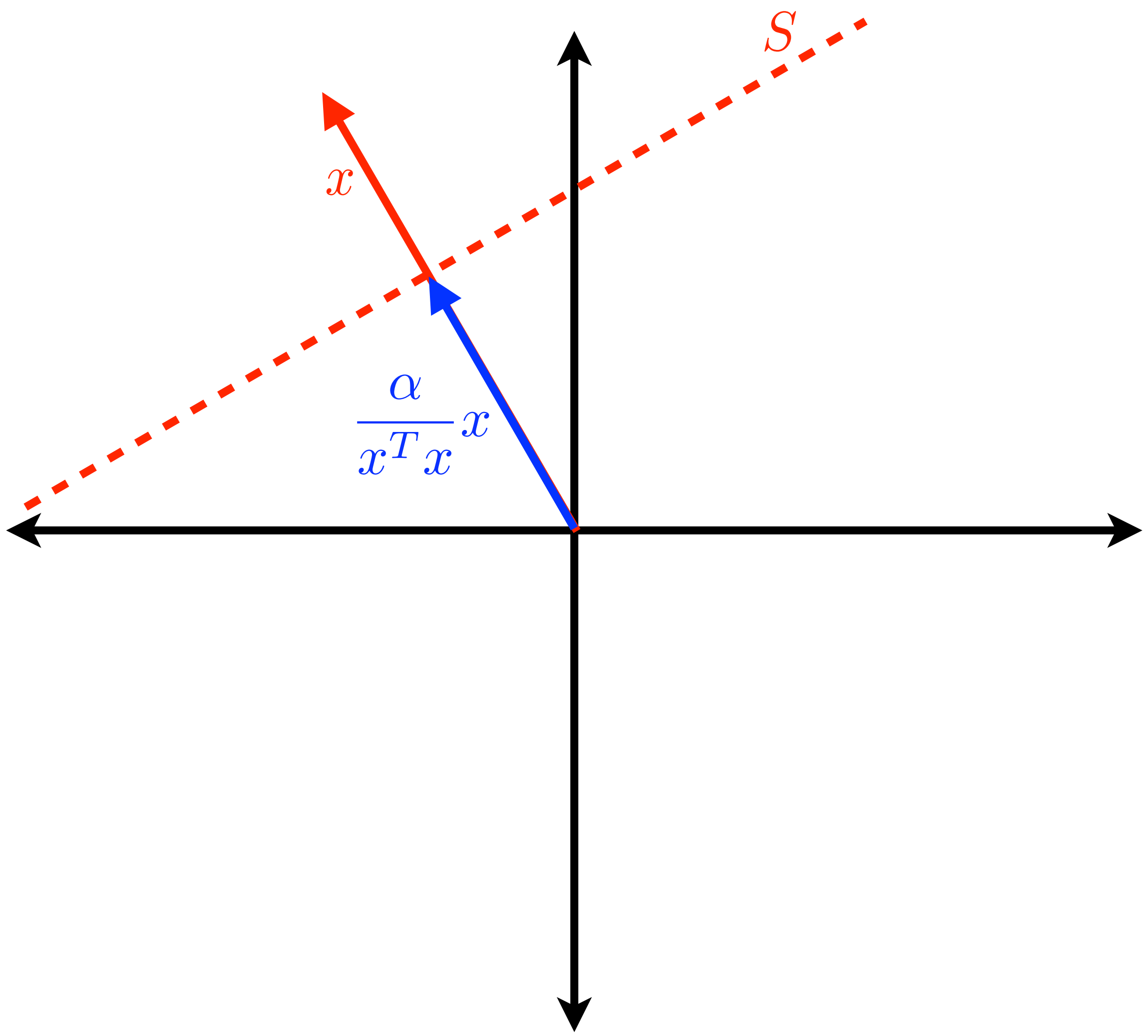
$$S = \{\mathbf{z} \mid \mathbf{x}^T \mathbf{z} = \alpha\} \quad \text{where} \quad \mathbf{x} \in \Re^N, \alpha \in \Re$$

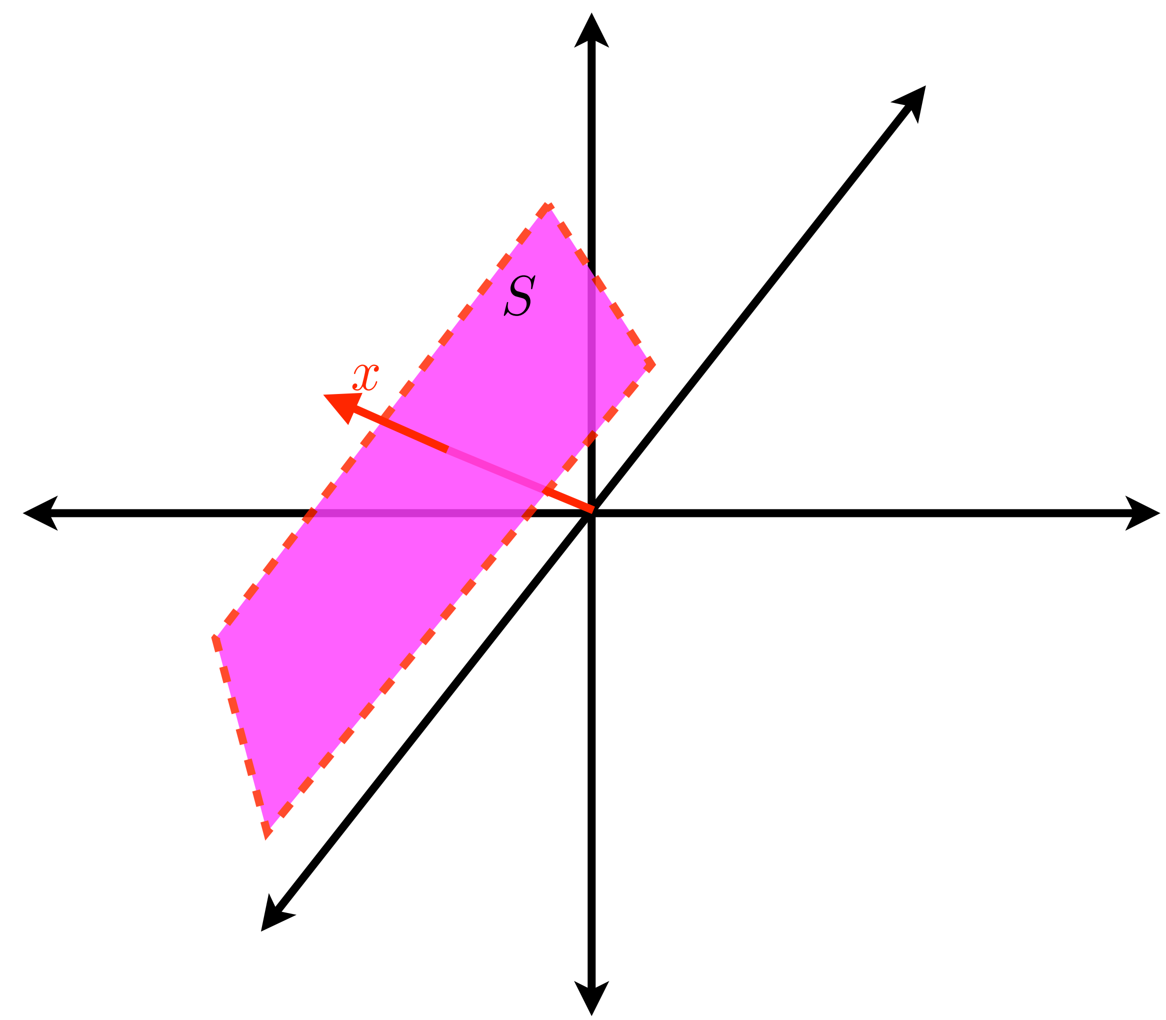


Scaling (x, α) does not change the hyperplane

Hyperplanes

$$S = \{\mathbf{z} \mid \mathbf{x}^T\mathbf{z} = \alpha\} \quad \text{where} \quad \mathbf{x} \in \Re^N, \alpha \in \Re$$



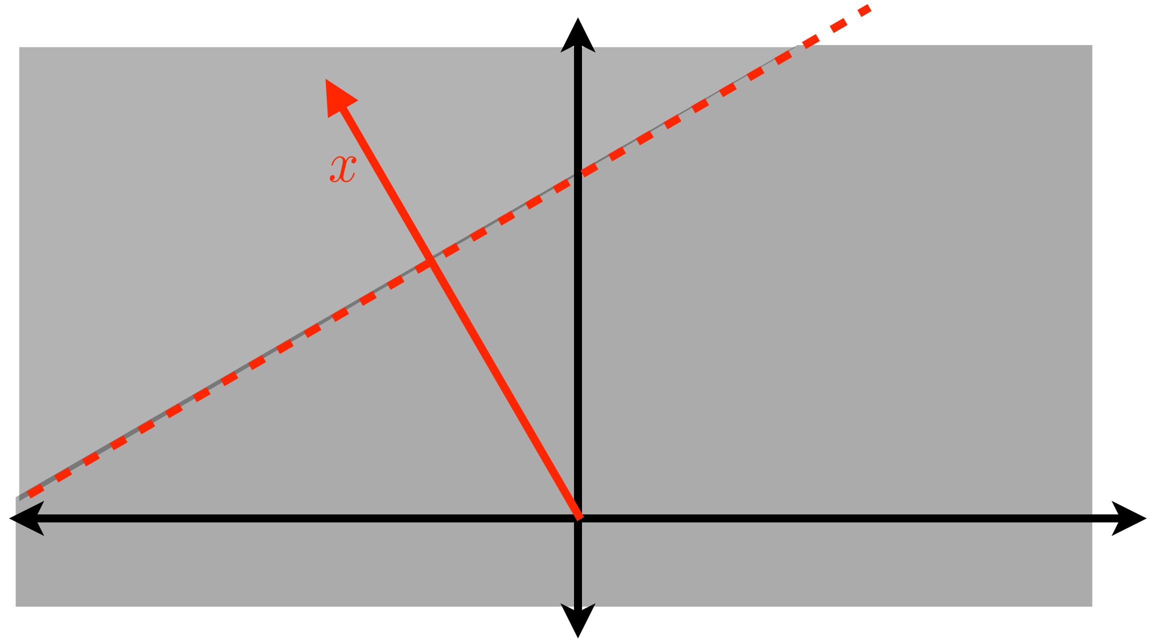


Preview: (x, α) will be the decision variables, and z will represent points in the data set

Linear Separation of Data

- A hyperplane defines two half spaces

$$\{\mathbf{z} \mid \mathbf{x}^T\mathbf{z} \geq \alpha\} \quad \text{and} \quad \{\mathbf{z} \mid \mathbf{x}^T\mathbf{z} \leq \alpha\}$$



- Goal
  - Positive samples in first half space, Negative samples in second half space, nothing on the boundary
  - This general classification technique is known as a Perceptron

Linear Program for Separable Data

- Data
  - Positive samples $\mathbf{u}^1, \ldots, \mathbf{u}^M \in \Re^N$
  - Negative samples $\mathbf{v}^1, \ldots, \mathbf{v}^K \in \Re^N$
- Linear program

$$
\begin{array}{lrcll}
\min & 0 & & & \\
\text{s.t.} & -\mathbf{x}^T\mathbf{u}^m + \alpha & \leq & 0 & \forall m = 1, \ldots, M \\
& \mathbf{x}^T\mathbf{v}^k - \alpha & \leq & 0 & \forall k = 1, \ldots, K
\end{array}
$$

- What is wrong with this LP?

Linear Program for Separable Data

It is always infeasible

The objective function is always 0

It has a trivial solution regardless of whether the data is separable

All of the above

$$
\begin{aligned}
\min \quad & 0 \\
\text{s.t.} \quad & -\mathbf{x}^T\mathbf{u}^m + \alpha \leq 0 \qquad \forall m = 1, \ldots, M \\
& \mathbf{x}^T\mathbf{v}^k - \alpha \leq 0 \qquad \forall k = 1, \ldots, K
\end{aligned}
$$

- What is wrong with this LP?

Linear Program for Separable Data

It is always infeasible

The objective function is always 0

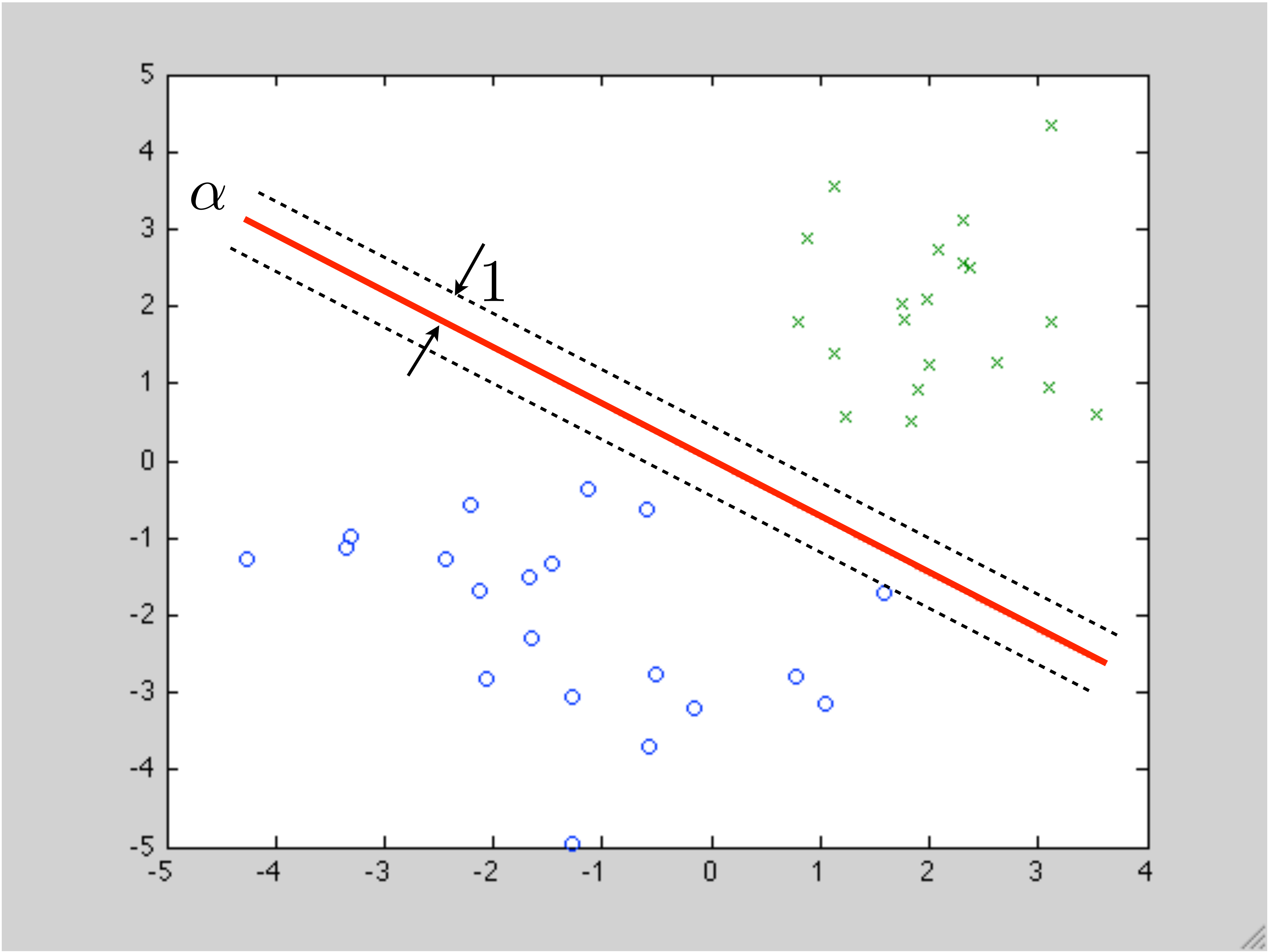
It has a trivial solution regardless of whether the data is separable

All of the above

$$
\begin{aligned}
\min \quad & 0 \\
\text{s.t.} \quad & -\mathbf{x}^T\mathbf{u}^m + \alpha \leq 0 \qquad \forall m = 1, ..., M \\
& \mathbf{x}^T\mathbf{v}^k - \alpha \leq 0 \qquad \forall k = 1, ..., K
\end{aligned}
$$

- What is wrong with this LP?

Linear Program for Separable Data



Linear Program for Separable Data

- Data
  - Positive samples $\mathbf{u}^1, \ldots, \mathbf{u}^M \in \Re^N$
  - Negative samples $\mathbf{v}^1, \ldots, \mathbf{v}^K \in \Re^N$
- Linear program

$$
\begin{array}{llll}
\min & 0 & & \\
\text{s.t.} & -\mathbf{x}^T\mathbf{u}^m + \alpha + 1 & \leq 0 & \forall m = 1, \ldots, M \\
& \mathbf{x}^T\mathbf{v}^k - \alpha + 1 & \leq 0 & \forall k = 1, \ldots, K
\end{array}
$$

-

Linear Program for Separable Data

$$
\begin{array}{llll}
\min & 0 & & \\
\text{s.t.} & -\mathbf{x}^T\mathbf{u}^m + \alpha + 1 & \leq 0 & \forall m = 1, ..., M \\
 & \mathbf{x}^T\mathbf{v}^k - \alpha + 1 & \leq 0 & \forall k = 1, ..., K
\end{array}
$$

- Why did we choose "1"? What would go wrong with another choice?

The number "1" stands for unity, and is perfect

The separation must be integral

Since x and alpha are scalable, the constant doesn't matter

Since x and alpha are scalable, the constant doesn't matter as long as it is positive

Linear Program for Separable Data

$$
\begin{array}{llcll}
\min & 0 & & & \\
\text{s.t.} & -\mathbf{x}^T\mathbf{u}^m + \alpha + 1 & \leq & 0 & \forall m = 1, \ldots, M \\
 & \mathbf{x}^T\mathbf{v}^k - \alpha + 1 & \leq & 0 & \forall k = 1, \ldots, K
\end{array}
$$

- Why did we choose "1"? What would go wrong with another choice?

The number "1" stands for unity, and is perfect

The separation must be integral

Since x and alpha are scalable, the constant doesn't matter

Since x and alpha are scalable, the constant doesn't matter as long as it is positive

What if Data is not Completely Separable

- Data
  - Positive samples $\mathbf{u}^1, \dots, \mathbf{u}^M \in \Re^N$
  - Negative samples $\mathbf{v}^1, \dots, \mathbf{v}^K \in \Re^N$
- Linear program

Impose "Penalty"
(if positive)

$$\begin{array}{lrcll} \min & 0 & & & \\ \text{s.t.} & -\mathbf{x}^T\mathbf{u}^m + \alpha + 1 & \leq & 0 & \forall m = 1, \dots, M \\ & \mathbf{x}^T\mathbf{v}^k - \alpha + 1 & \leq & 0 & \forall k = 1, \dots, K \end{array}$$

- But what penalty?

Notation Clarification

- We will use the following terms:
  - Violation
    - Just the value of the LHS

Impose "Penalty" (if positive)

Violation

0

$$-\mathbf{x}^T\mathbf{u}^m + \alpha + 1 \leq 0 \qquad \forall m = 1, \ldots, M$$

$$\mathbf{x}^T\mathbf{v}^k - \alpha + 1 \leq 0 \qquad \forall k = 1, \ldots, K$$

    - This should be negative for the point to be "correctly classified"
    - Can not minimize the "violation" directly, since we don't want a point that is correctly classified to contribute a "negative amount" to the objective function
    - If violation is positive, then the point is incorrectly classified and there should be a penalty
  - Penalty: This must be 0 when the violation is "negative"

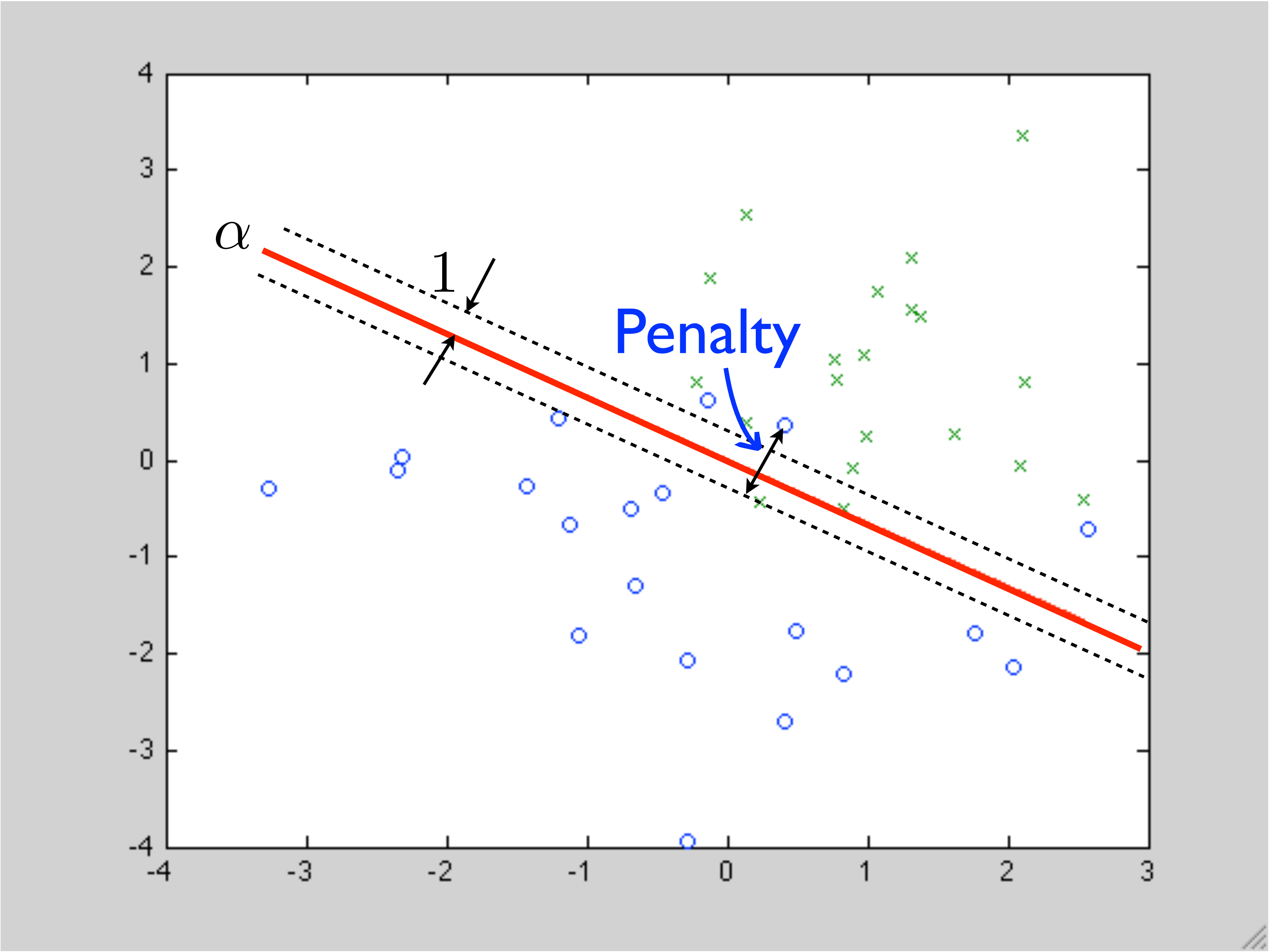
Inseparable Data

- Minimize number of misclassifications?
  - That is, penalty = 1 if violation > 0 and 0 otherwise
  - NP-complete

- Another natural penalty function: = violation when the violation is positive; zero otherwise (later: quadratic)
  - Penalty: $\max(0, -\mathbf{x}^T\mathbf{u}^m + \alpha + 1)$, $\max(0, \mathbf{x}^T\mathbf{v}^k - \alpha + 1)$
  - Objective:

$$\min\left(\sum_{m=1}^{M} \max(0, -\mathbf{x}^T\mathbf{u}^m + \alpha + 1) + \sum_{k=1}^{K} \max(0, \mathbf{x}^T\mathbf{v}^k - \alpha + 1)\right)$$

Inseparable Data



Linear Program for Inseparable Data

- Introduce additional decision variables:

$$\delta_m^+ \leftrightarrow \max(0, -\mathbf{x}^T \mathbf{u}^m + \alpha + 1)$$

$$\delta_k^- \leftrightarrow \max(0, \mathbf{x}^T \mathbf{v}^k - \alpha + 1)$$

- Linear program

$$\begin{aligned}
\min \quad & \left(\sum_{m=1}^{M} \delta_m^+ + \sum_{k=1}^{K} \delta_k^- \right) & & & \\
\text{s.t.} \quad & -\mathbf{x}^T \mathbf{u}^m + \alpha + 1 & \leq & \ \delta_m^+ & \forall m = 1, ..., M \\
& \mathbf{x}^T \mathbf{v}^k - \alpha + 1 & \leq & \ \delta_k^- & \forall k = 1, ..., K \\
& \delta_m^+ & \geq & \ 0 & \forall m = 1, ..., M \\
& \delta_k^- & \geq & \ 0 & \forall k = 1, ..., K
\end{aligned}$$

Breast Cancer Diagnosis

radius	**perimiter**	**area**	**...**	**class**
18	122.8	1001	...	malignant
...	...	...	...	...
...	...	...	...	...
20.6	132.9	1326	...	benign
...	...	...	...	...
...	...	...	...	...
19.7	130	1203	...	malignant