UNIVERSITY AT ALBANY
STATE UNIVERSITY OF NEW YORK

CSI 436/536 (Spring 2025)
Machine Learning
Lecture 11: Support Vector Machines

Chong Liu
Department of Computer Science

Mar 3, 2025

Announcement

- HW 2 due today.
- HW 1 and 2 review this Wednesday.
- Midterm exam next Monday.
- Midterm presentation next Wednesday.

Recap: Regularization

- Linear regression
  - Solving the Least Square problem {with GD, SGD and direct solver}
- Regularization
  - Controls the parameter complexity of the fitted function
  - Prevents overfitting!
  - Different regularization: L-2 (most popular) and L-1
- Case study: Predict House Price
  - Effect of regularization on training test and test error
  - Regularization path (Effect of regularization on coefficients)

Today

- Move back to binary classification problem
    - Spam email / non-spam email

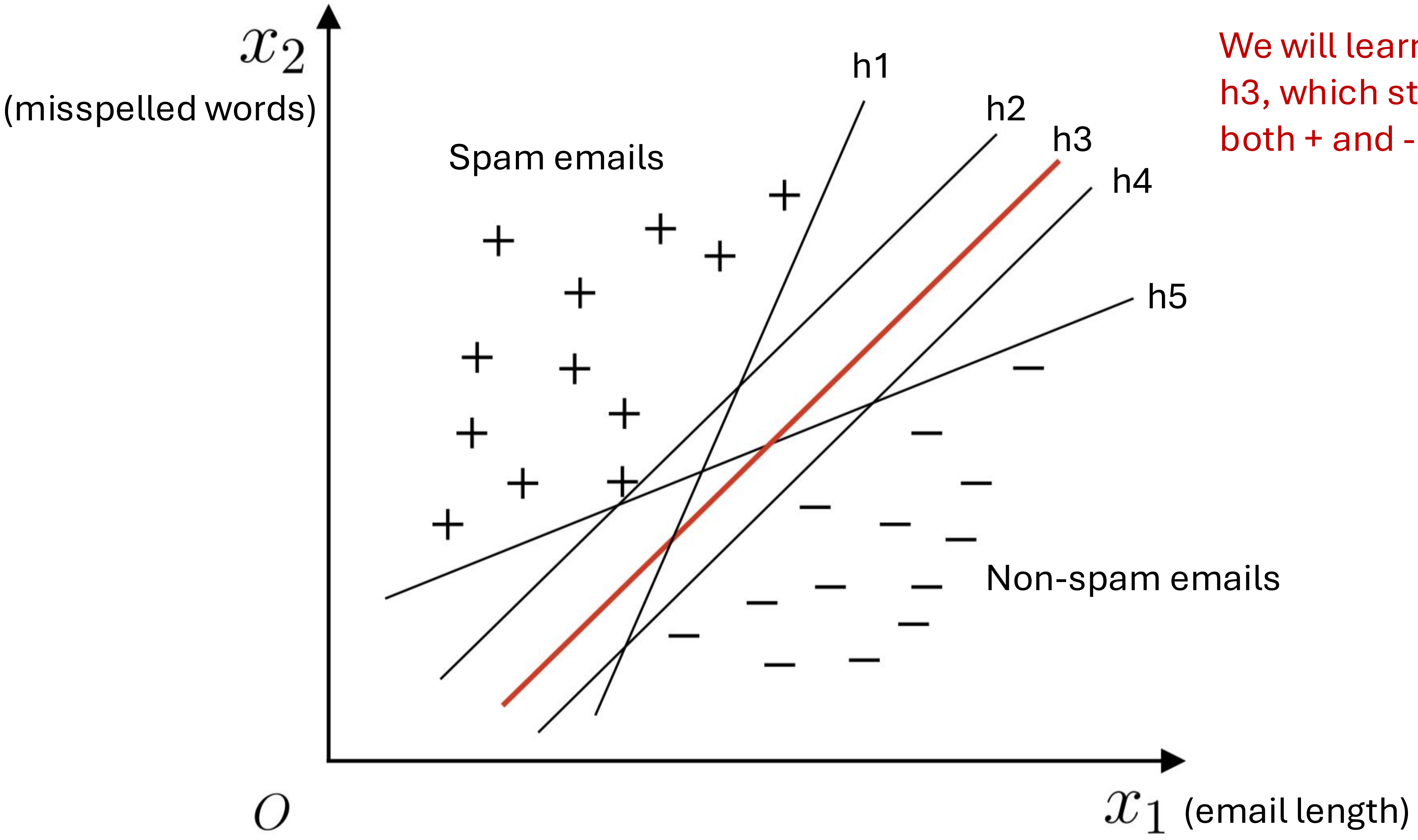
- Margin

- Support Vector Machines

- Warning: While without any proof, today’s lecture will be very technical. Feel free to interrupt me at any point to ask questions.

- Midterm exam overview

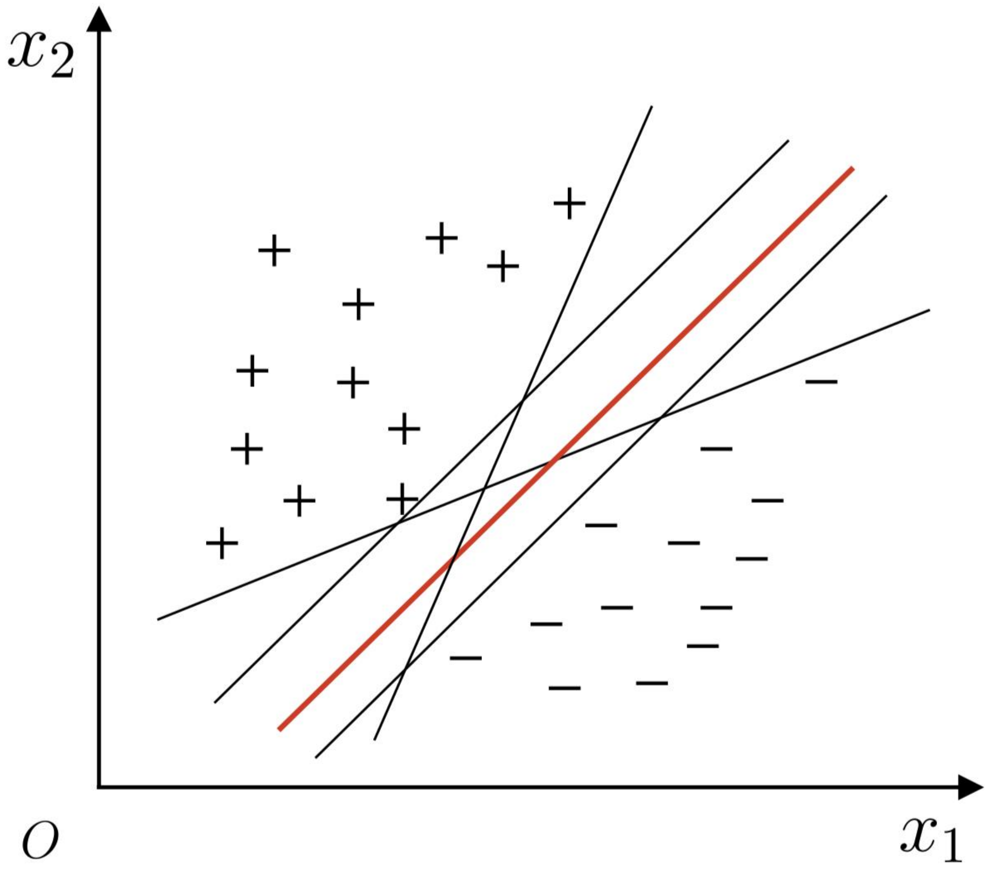
Discussion: which is the best classifier?



We will learn how to train h3, which stays away from both + and -.

Linear classification

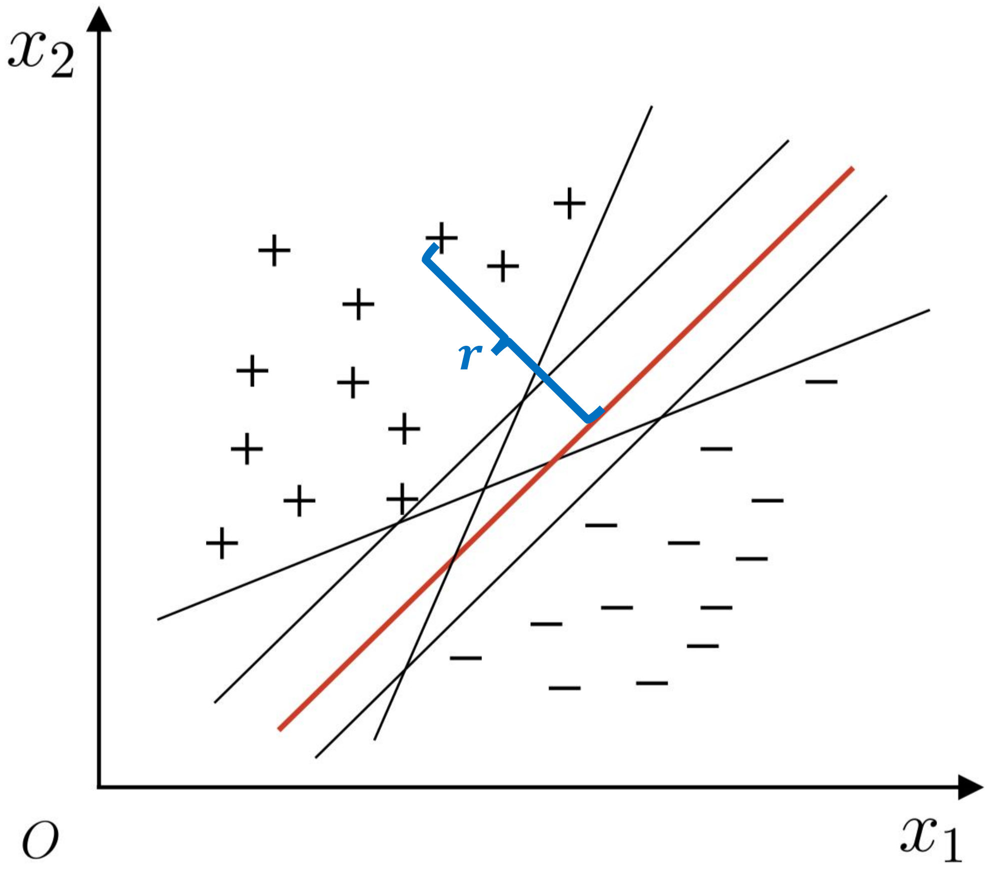
- Input: $x = [x_1, x_2] \in R^2$
- Output: $y \in \{1, -1\}$
- Data: n data points
- Decision line:
  - $w^T x + b = 0$
  - $w \in R^2, b \in R$ are parameters
  - In-class exercise: Rewrite $x_2 = x_1 - 5$ in $w^T x + b = 0$ form.



Margin: min distance of data point to line

- Any data point:
  - $x \in R^2$
- Any line:
  - $w^T x + b = 0$
- Margin:
  - $r = \frac{|w^T x + b|}{||w||}$

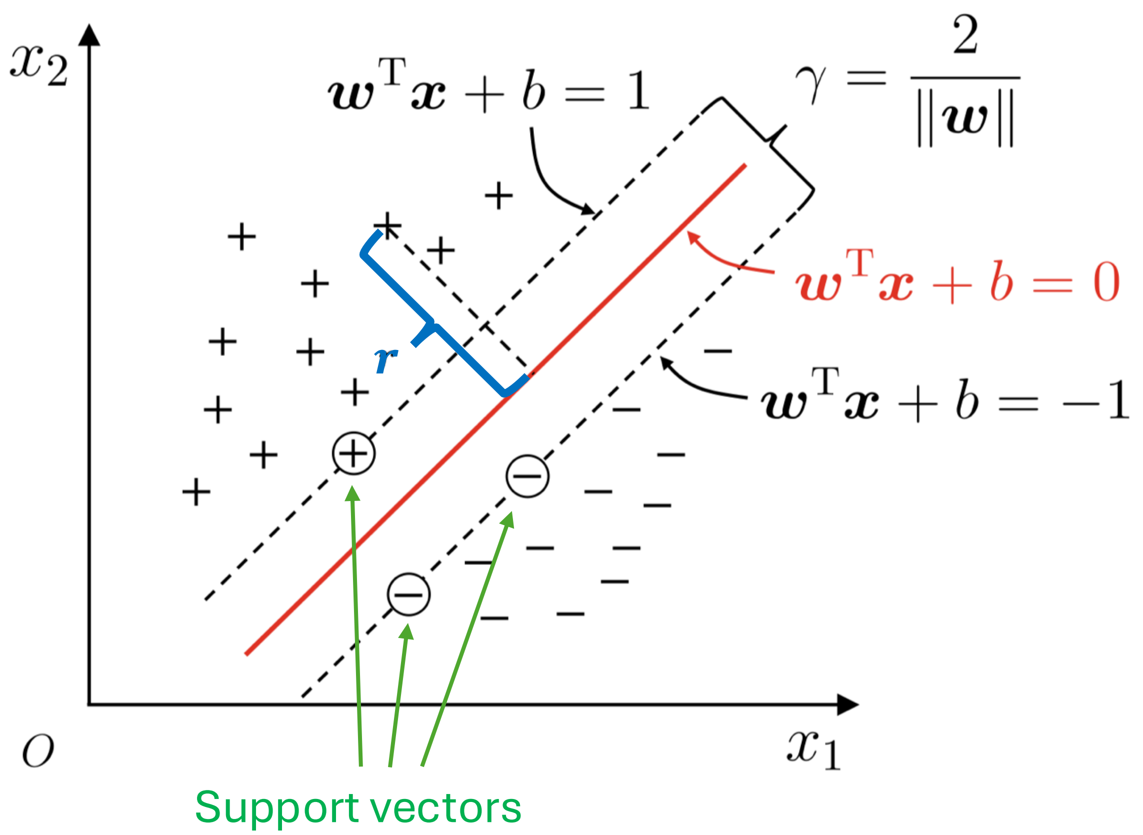
- Red line: We want to learn a max-margin classifier!



Max-margin classifier

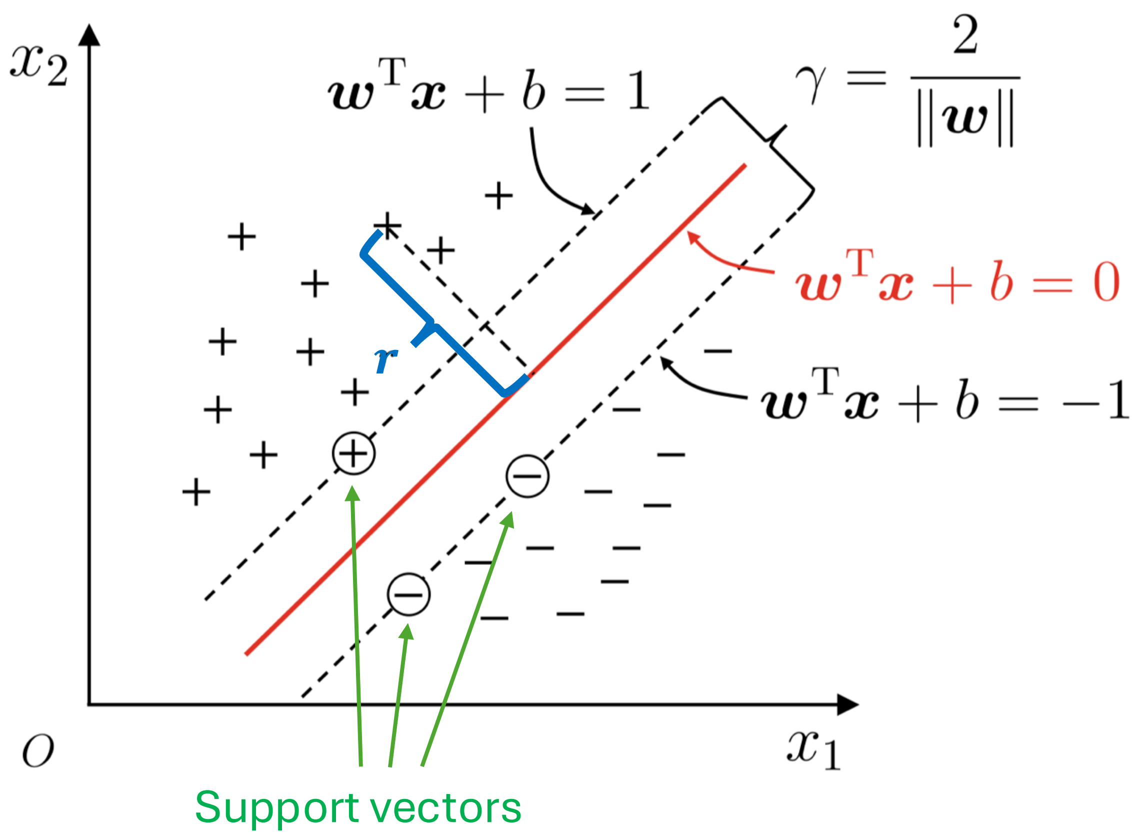
- Discussion: by maximizing margin, which data points are important?

- Support vectors:
    - Data points closest to red line.
    - Only support vectors affect the training process.
    - Support vector machines (SVM) == max-margin classifier



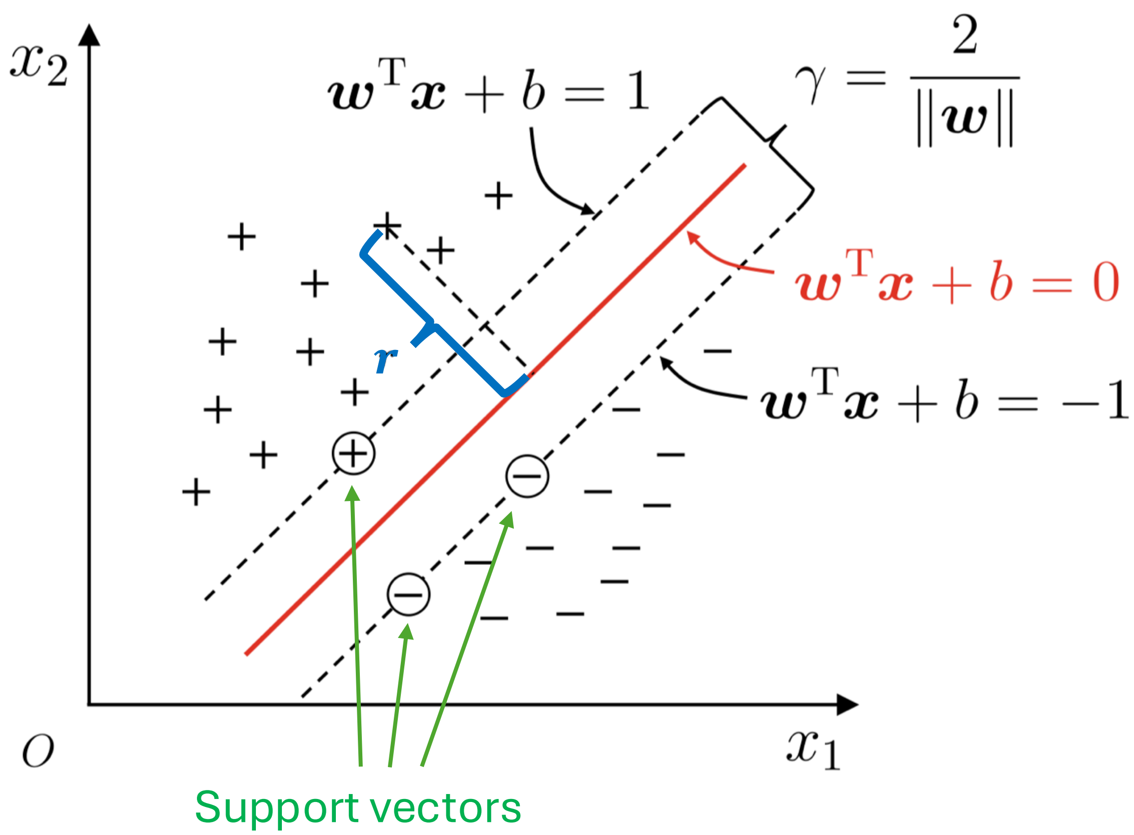
How to train max-margin classifier?

- Assumption:
  - Linearly separable data points
- Recap: Linear classifier
  - If $y = 1, w^T x + b > 0$
  - If $y = -1, w^T x + b < 0$
- Key idea of SVM:
  - If $y = 1, w^T x + b \geq 1$
  - If $y = -1, w^T x + b \leq -1$
  - Why? Support vectors are only data points that matter.



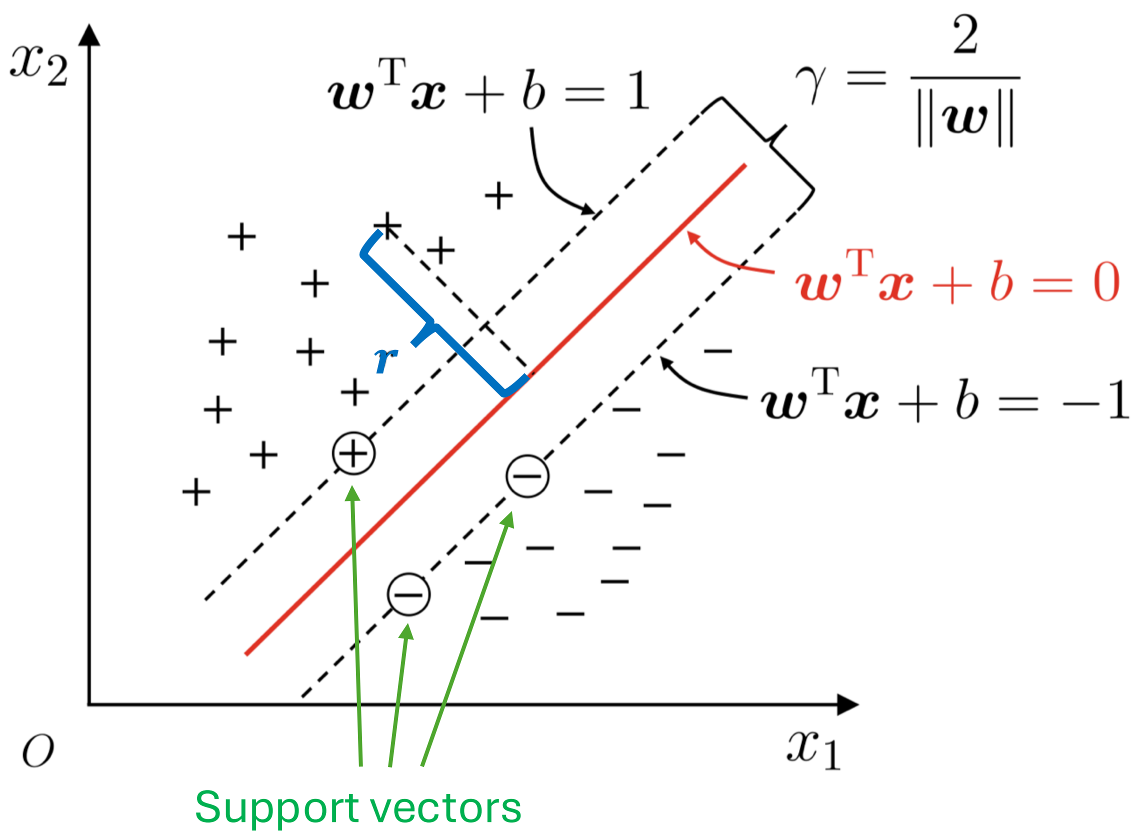
How to train max-margin classifier?

- Key idea of SVM:
  - If $y = 1, w^T x + b \geq 1$
  - If $y = -1, w^T x + b \leq -1$
- Recap: Margin for any data point x
  - $r = \frac{|w^T x + b|}{||w||}$
- Total margin between support vectors:
  - $\gamma = \frac{2}{||w||}$



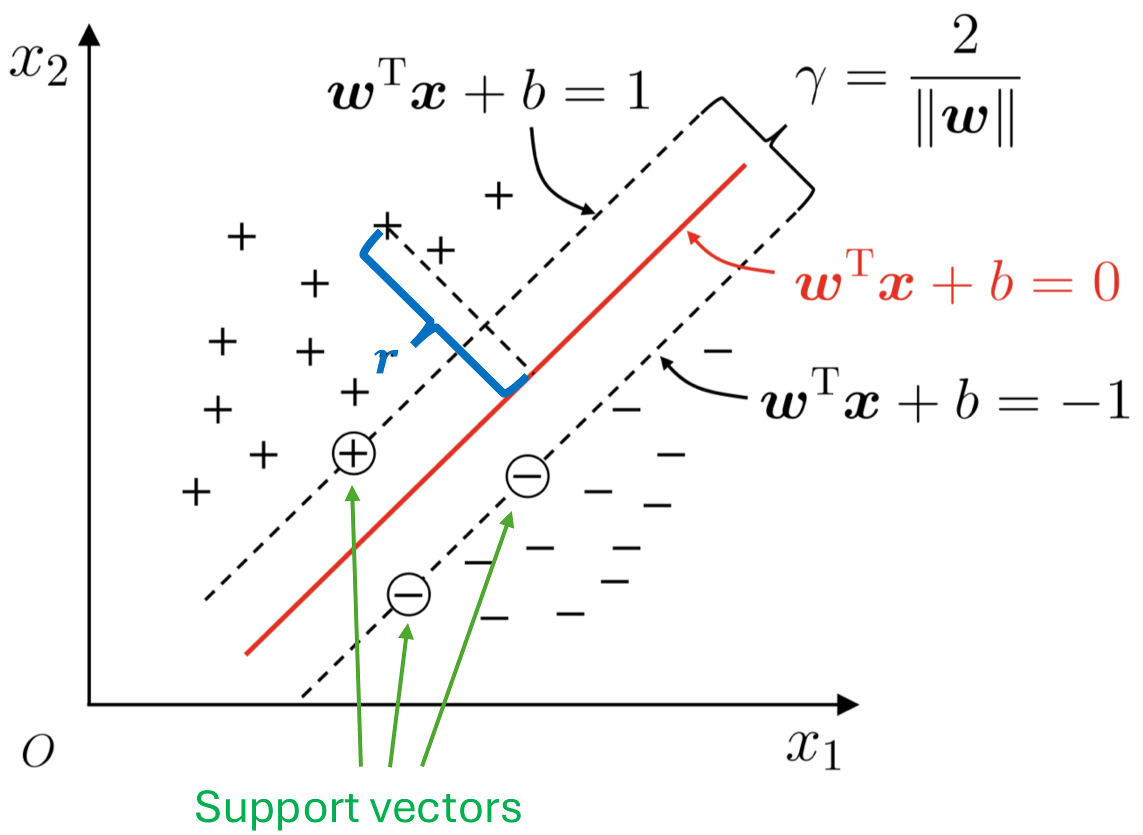
How to train max-margin classifier?

- Key idea of SVM:
  - If $y = 1, w^T x + b \geq 1$
  - If $y = -1, w^T x + b \leq -1$
- Total margin between support vectors:
  - $\gamma = \frac{2}{||w||}$
- Optimization problem of SVM:
  - $\max_{w,b} \frac{2}{||w||}$
  - s.t. $y_i(w^T x_i + b) \geq 1, i = 1, \ldots, n$



How to train max-margin classifier?

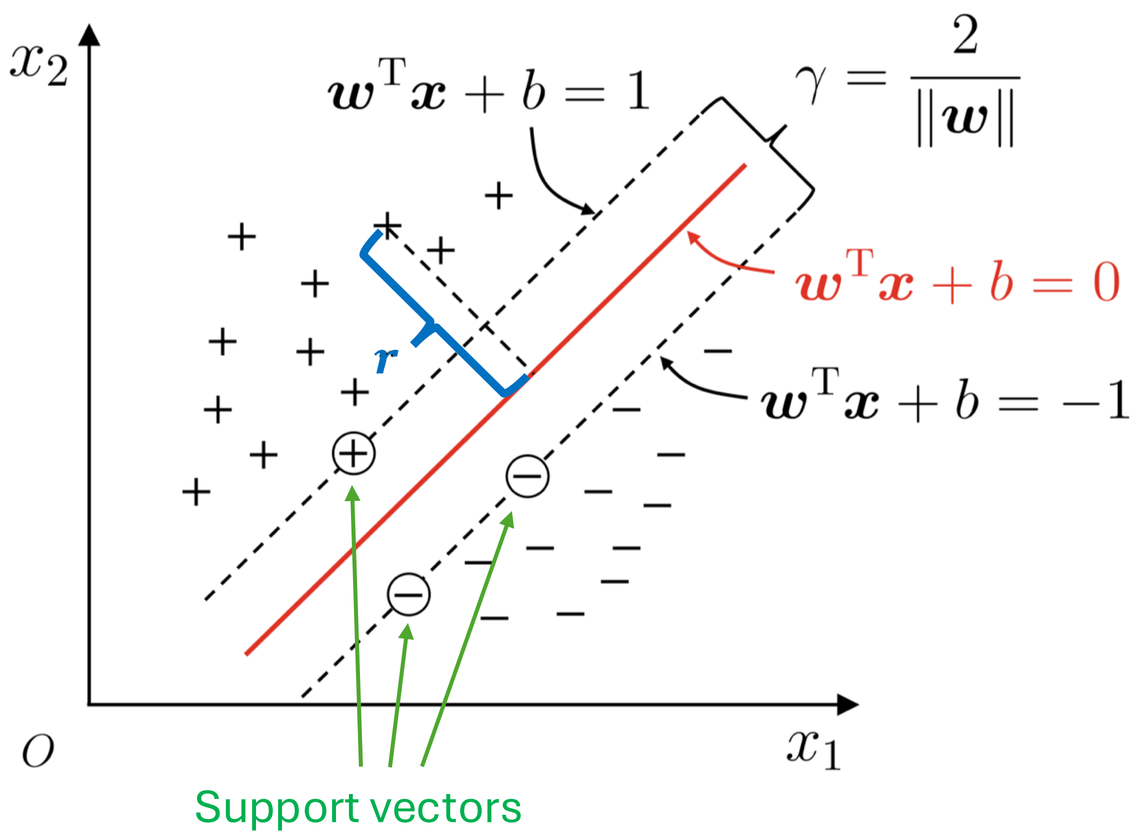
- Optimization problem of SVM:
  - $\max_{w,b} \frac{2}{||w||}$
  - $\text{s.t. } y_i(w^T x_i + b) \geq 1, i = 1, \ldots, n$
- Equivalent optimization problem:
  - $\min_{w,b} \frac{1}{2}||w||$
  - $\text{s.t. } y_i(w^T x_i + b) \geq 1, i = 1, \ldots, n$
  - Quadratic programming problem
    - Can be solved using some optimization tools, e.g., CPLEX.



How to train max-margin classifier?

- Equivalent optimization problem:
  - $\min_{w,b} \frac{1}{2}||w||$
  - s.t. $y_i(w^T x_i + b) \geq 1, i = 1, \ldots, n$

- In-class exercise:
  - Write the optimization problem with three support vectors: $(6, 2)\ +, (7,1)\ -, (8,2)\ -$.



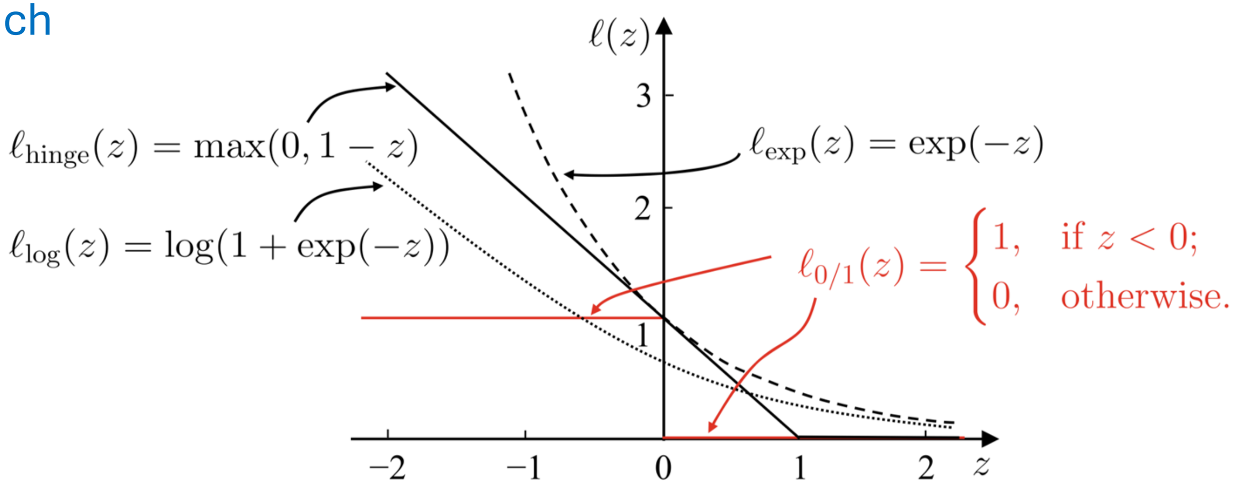
Take a deeper look at optimization problem

- Equivalent optimization problem:
  - $\min_{w,b} \frac{1}{2}||w||$
  - $\text{s.t. } y_i(w^T x_i + b) \geq 1, i = 1, \ldots, n$
- In-class exercise: Write the Lagrange function
  - $\min_{w,b} \boxed{\frac{1}{2}||w||} + \sum_{i=1}^{n} \boxed{\lambda_i} \boxed{(1 - y_i(w^T x_i + b)})$

Related to a New surrogate loss function - Hinge loss!

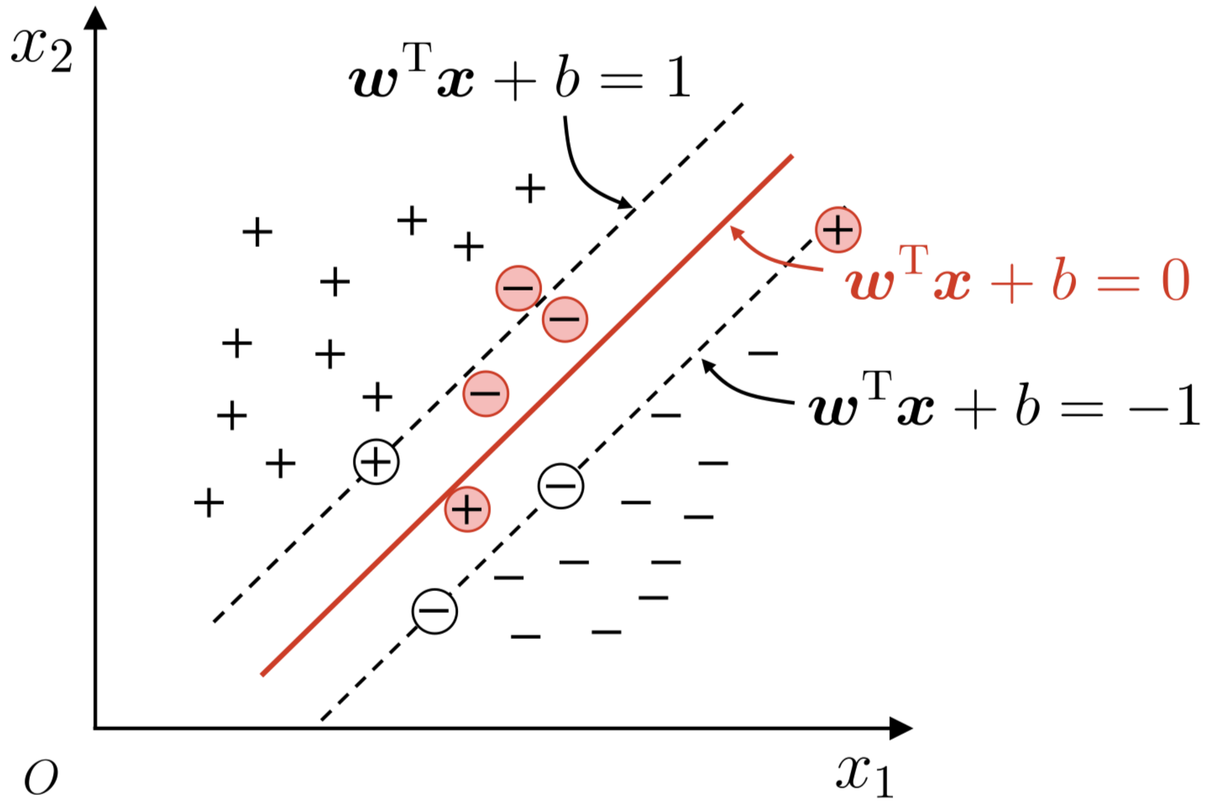
L-2 regularization on parameter!

Weight of each data point



What if data points are not linearly separable?

- $y_i(w^T x_i + b) \geq 1$ is violated.
- What can we do?
  - Key idea: we give some tolerance.
- New constraint:
  - $y_i(w^T x_i + b) \geq 1 - \xi$
  - $\xi > 0$
  - Discussion: what happens when ξ is very large / small?



Soft-Margin Support Vector Machines

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

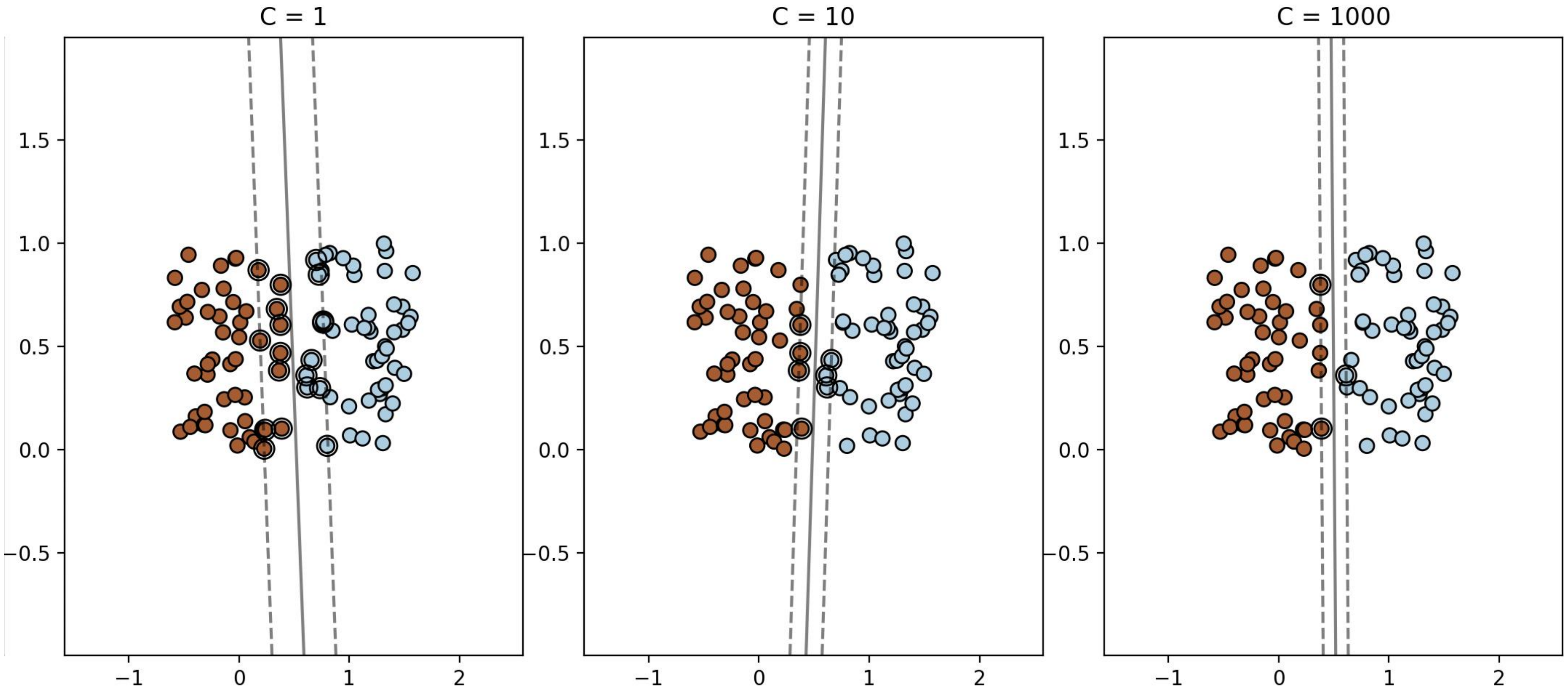
$$s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i \in [n]$$

$$\xi_i \geq 0 \quad \forall i \in [n]$$

Equivalent to minimizing ***Hinge losses:***

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \max\left[1 - y_i(\mathbf{w}^T\mathbf{x} + b), 0\right]$$

Hyperparameter C in **soft-margin SVM** and how they affect the margins and "support vectors".



As C increases, smaller tolerance and fewer soft-margin support vectors.

Checkpoint of Lecture 1-11

- Tasks of ML:
  - Classification (spam / non-spam email) and regression (house price)
- Philosophy of designing ML algorithms:
  - Regularization: Control the complexity of parameters
    - Prevent overfitting
    - Fun fact: L-2 regularization is associated with max margin classifier
  - Optimization: Toolbox of ML
    - ML problem => optimization problem
      - Direct solver, GD, SGD, and much more!
    - Minimize the loss / parameter complexity
    - Maximize the margin

Midterm exam

- What does the exam look like?
  - 80 min (3 - 4:20pm) on Mon Mar 10 at LC 4
  - Please arrive 5min earlier!
  - Closed-book exam
  - Given individually (not in groups!)
  - Counts 20% towards your final grades
  - No make-up exam
- What to bring?
  - Your pen only.
- What not to bring?
  - Your book, note, lecture slide, or cheat sheet.

What are you expected to know?

- Basic mathematical tools
  - In our math review (Lecture 2-4)
  - Linear algebra, calculus and optimization, probability and statistics

What are you expected to know?

- Basic concepts of machine learning
  - Classification and regression
  - Input space (feature space), output space (label space), hypothesis class
  - Confusion matrix of binary classification
  - Accuracy
  - Holdout / cross validation / hyperparameter
  - Problem of overfitting
  - Loss function
  - Linear model

What are you expected to know?

- Understanding how machine learning algorithms work
  - Why do we need surrogate loss in classification?
  - Why do we need SGD? Drawback of GD?
  - How to define a linear classifier / linear regression?
  - Why do we need SVM? Difference between linear classifier and SVM.
  - Why do we need regularization? How to apply it?

- Important tips:
  - Review HW 1 and 2 and all in-class exercise problems.