

Support Vector Machines

cs 556

Today's Lecture

- Vector Algebra, Formulation, Margin
- SVM for Linear Separable Case
- Non-separable Case, Penalties

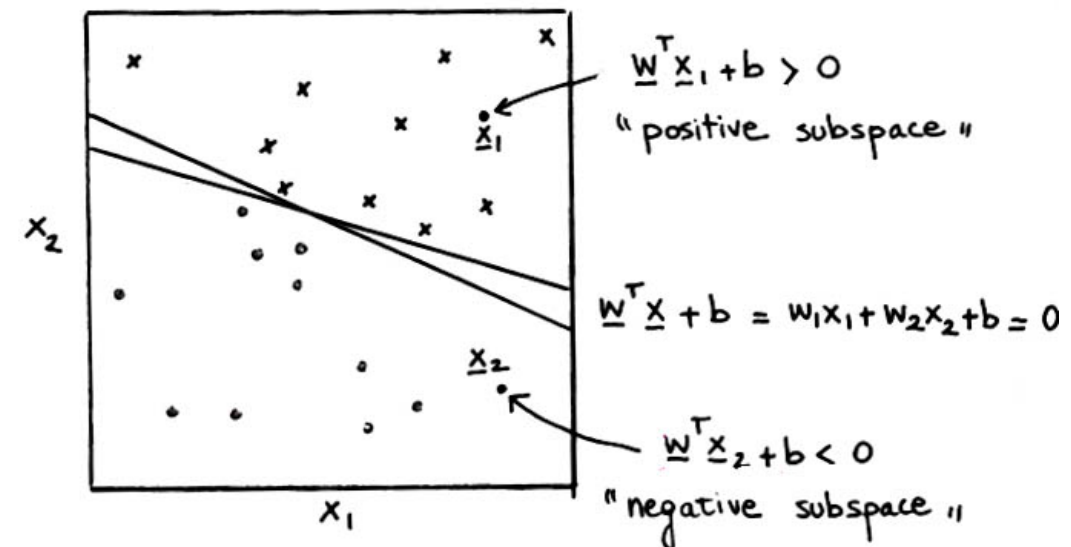
Linear Classifier

- Linear classifiers construct **linear decision boundaries** (hyperplanes) that try to separate the data into different classes as well as possible.
- Classification rule of the **Perceptron algorithm**:

Input: $x \in \mathbb{R}^d$

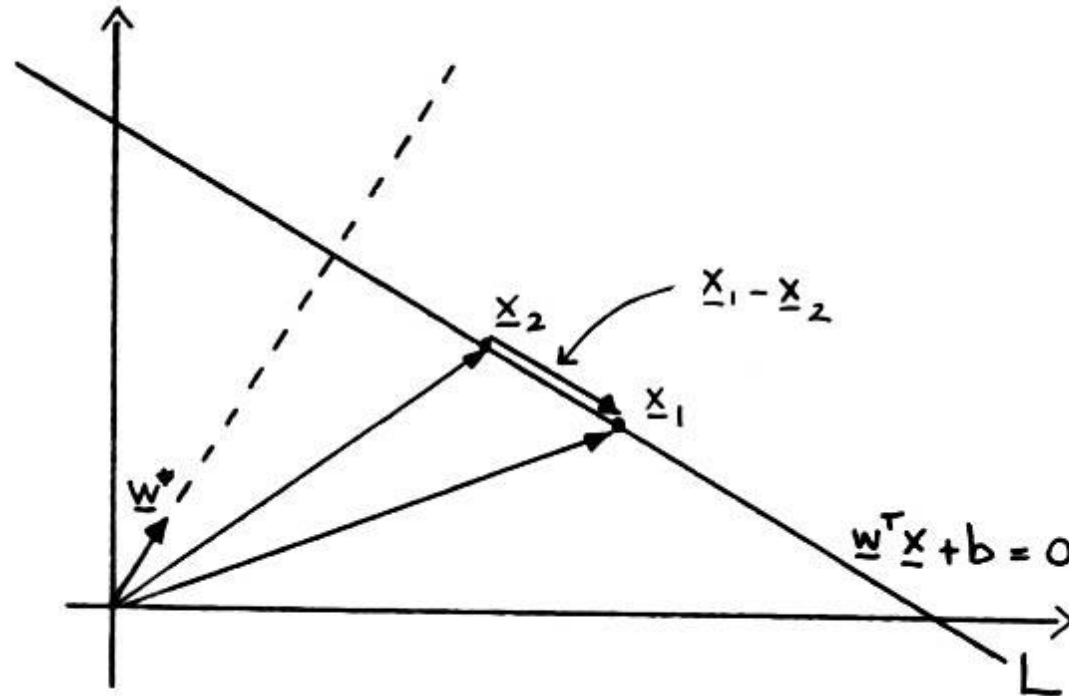
Output: $\text{sign}(w^T x + b)$

- The classifier computes a linear combination of the input features and return the sign.



Some Vector Algebra

- Hyperplane $L: f(x) = w^T x + b = 0$ when $x \in \mathbb{R}^d$, $f(x)$ is a linear boundary.



Some Vector Algebra: Property 1

- Consider any two points x_1 and x_2 , lying on hyperplane L :

$$\begin{aligned} w^T x_1 + b &= 0 \\ w^T x_2 + b &= 0 \end{aligned} \rightarrow w^T (x_1 - x_2) = 0$$

- Since $w^T (x_1 - x_2) = w \cdot (x_1 - x_2) = 0$, the two vectors w and $x_1 - x_2$ are orthogonal vectors.

$$w^* = \frac{w}{\|w\|}$$

is the vector normal to the surface of L .

- Note 1: All vectors here are column vectors.
- Note 2: Dot product (inner product) of two vectors $a \cdot b = a^T b = \|a\| \times \|b\| \times \cos \alpha$, where α is the angle between a and b .

Some Vector Algebra: Property 2

- For any point x_0 on L :

$$w^T x_0 + b = 0$$

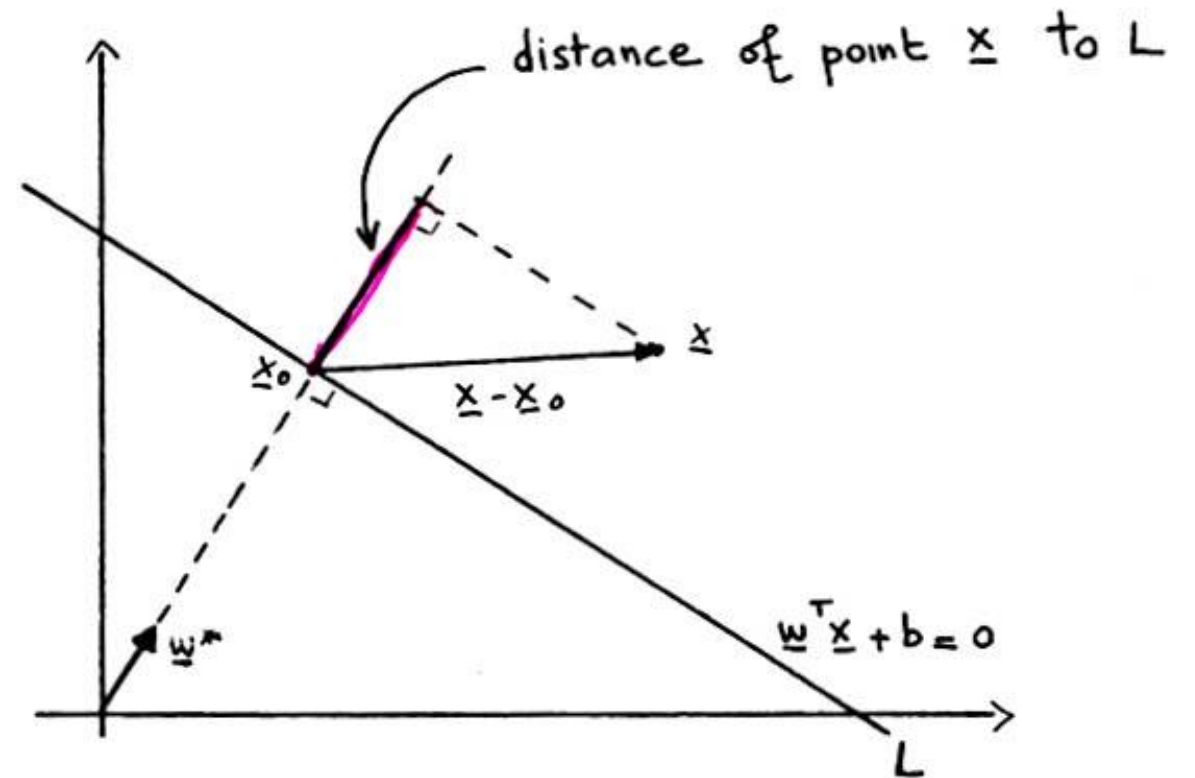
Thus:

$$w^T x_0 = -b$$

Some Vector Algebra: Property 3

- The signed distance of any point x to L is:

$$\begin{aligned} & (w^*)^T (x - x_0) \\ &= \frac{w^T}{\|w\|} (x - x_0) \\ &= \frac{1}{\|w\|} (w^T x - w^T x_0) \\ &= \frac{1}{\|w\|} (w^T x + b) \\ &= \frac{f(x)}{\|w\|} \end{aligned}$$



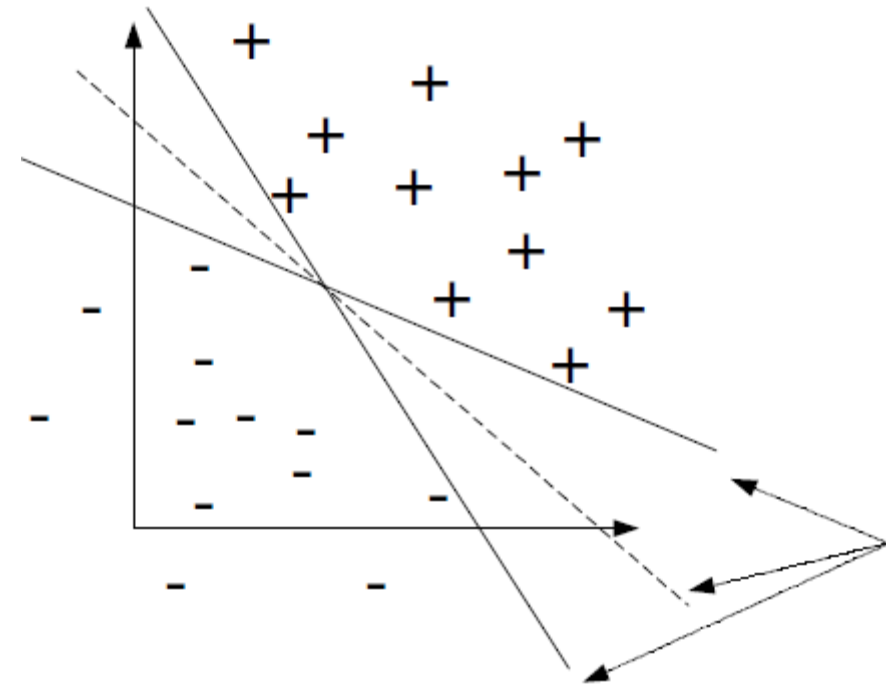
Limitations of Perceptron Algorithm

- When the data are **linearly separable**, there are many solutions.
 - The result depends on the starting values of the parameters and the order of data samples.
 - The finite number of steps to convergence can be very large: the smaller the gap between the two classes, the longer the time to find it.
- When the data are **not linearly separable**, the algorithm will not converge, and cycles develop. The cycles can be long, and therefore hard to detect.

The Linearly Separable Case

Limitations of Perceptron Algorithm

- When the data are **linearly separable**, there are **many solutions**, and which one is found depends on the starting values of the parameters and the order of data samples.
- There can be an infinite number of hyperplanes that achieve 100% accuracy on training data.
- Which hyperplane is the **optimal** with respect to the accuracy on test data?
- Solution: **adding additional constraints to the separating hyperplane.**

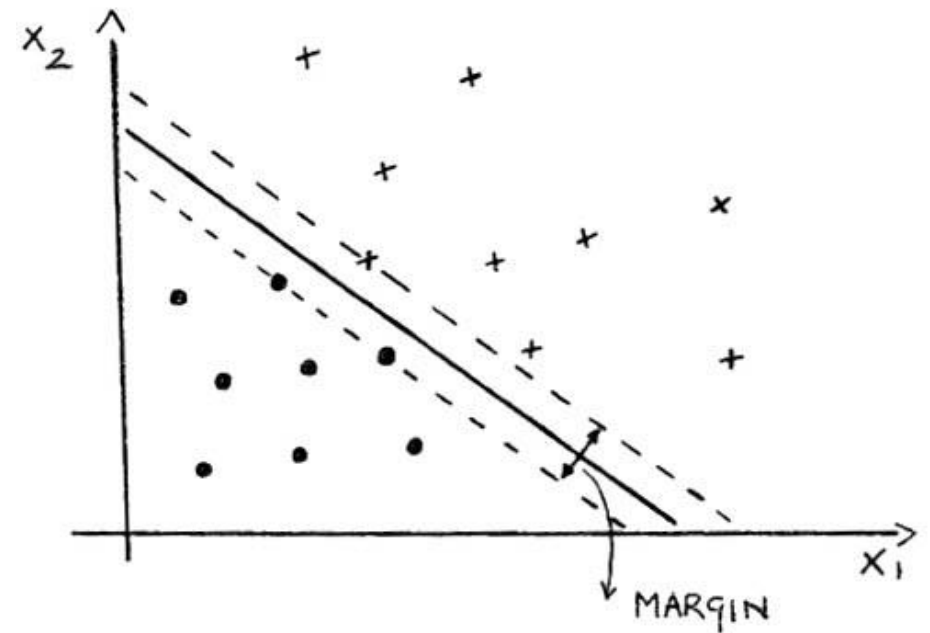


Three possible separations

Largest Margin Hyperplanes

Goal: Find the hyperplane that separates the two classes and maximizes the distance **to the closest points from either class**.

- Such distance is called **margin**.
- The added constraint:
 - Provide a unique solution to the separating hyperplane problem.
 - Maximizing the margin between the two classes on the training data gives better classification performance on test data.



Training Data

- For two classes:

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$

$$x_i \in R^d$$

$$y_i \in \{-1, +1\}$$

- We need to formalize **the largest margin criterion**.

Formulation

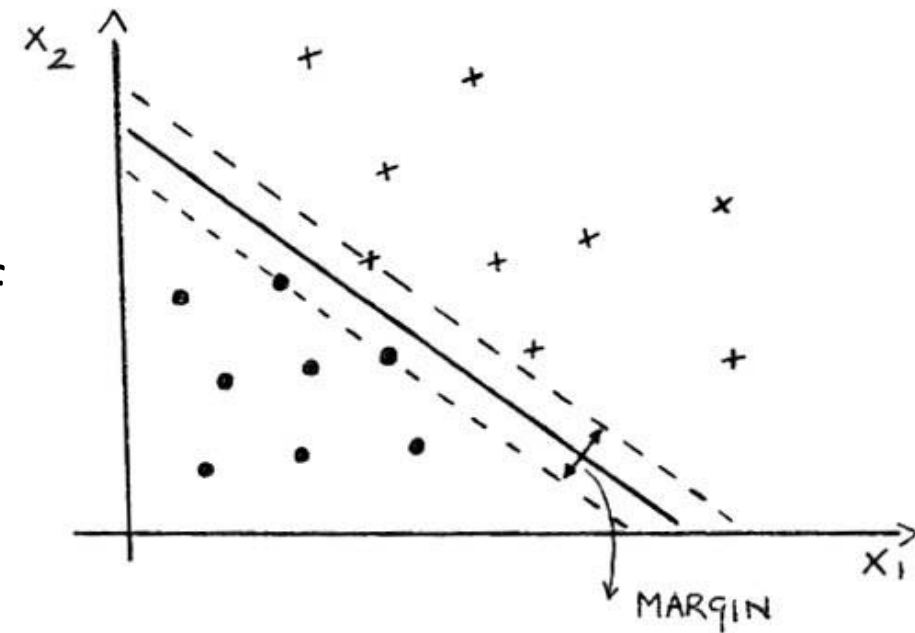
- Consider the following optimization problem:

$$\begin{aligned} & \max_{w,b} 2C \\ & \text{subject to } \frac{1}{\|w\|} y_i (w^T x_i + b) \geq C, i = 1, \dots, N \end{aligned}$$

- Remember Property 3: The signed distance of any point x to L is:

$$\frac{1}{\|w\|} (w^T x_i + b)$$

- Thus, the set of conditions above ensure that all the training data are **at least at distance C from the decision boundary**.



Formulation

- The optimization problem:

$$\max_{w,b} 2C \quad \text{subject to} \quad \frac{1}{\|w\|} y_i(w^T x_i + b) \geq C, i = 1, \dots, N$$

- We seek the largest C and associated parameters.
- We can rewrite the above conditions as:

$$y_i(w^T x_i + b) \geq C \|w\|$$

- Since $w^T x + b = 0$ and $c(w^T x + b) = 0$ define the same plane, we can arbitrarily normalize $\|w\| = \frac{1}{C}$.

Formulation

- The optimization problem:

$$\begin{aligned} & \max_{w,b} 2C \\ & \text{subject to } \frac{1}{\|w\|} y_i(w^T x_i + b) \geq C, i = 1, \dots, N \end{aligned}$$

- By **normalizing** $\|w\| = \frac{1}{C}$, the original maximization problem is equivalent to:

$$\begin{aligned} & \min_{w,b} \frac{1}{2} \|w\|^2 \\ & \text{subject to } y_i(w^T x_i + b) \geq 1, i = 1, \dots, N \end{aligned}$$

- The constraints define an empty margin around the linear decision boundary of **thickness** $\frac{2}{\|w\|}$. We choose w, b to maximize its thickness.
- This is a convex quadratic optimization problem subject to linear constraints and there is a unique minimum.

Lagrange Multipliers

- We introduce the **Lagrange multipliers** $\alpha_i \geq 0, i = 1, \dots, N$,
- One for each of the inequality constraints.
- Recall the rule:
 - For constraints of the form $C_i \geq 0$, the constraint equations are **multiplied by** Lagrange multipliers and **subtracted from** the objective function, to form the Lagrangian.

Lagrange Multipliers

- Lagrange multipliers allow us to take the constraints within the function to be minimized.
- Two reasons for doing this:
 - The constraints will be **replaced** by constraints on the Lagrange multipliers themselves, which are easier to handle.
 - In the new formulation of the problem, the training data will only appear **in the form of dot products between vectors**. This is a crucial property which will allow us to generalize the procedure to **the nonlinear case**.

Lagrange Multipliers: Primal Form

- We then obtain the Lagrangian: (also called **primal form**):

$$L_p = \frac{1}{2} \|w\|^2 - \sum_{i=1}^N \alpha_i [y_i (w^T x_i + b) - 1]$$

- We now **minimize** L_p with respect to w and b :

$$\min_{w,b} \max_{\alpha_i \geq 0} L_p$$

- This indicates that this is the **primal form** of the optimization problem.
- We will actually solve the primal optimization problem by solving the **dual of the original problem**, since they provide the same solution.

Dual Form

- The solution to the dual form provides a lower bound to the solution of the primal form.
- What is the dual form?

$$\max_{\alpha_i \geq 0} \min_{w, b} L_p$$

- Setting the derivatives to zero gives:

$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1}^N \alpha_i y_i x_i = 0 \quad \Rightarrow \quad w = \sum_{i=1}^N \alpha_i y_i x_i \quad (1)$$

$$\frac{\partial L_p}{\partial b} = - \sum_{i=1}^N \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^N \alpha_i y_i = 0 \quad (2)$$

Dual Form

- Substituting Eq. (1) and (2) in L_p gives:

$$\begin{aligned} L_D &= \frac{1}{2} \left(\sum_{i=1}^N \alpha_i y_i x_i \right) \left(\sum_{k=1}^N \alpha_k y_k x_k \right) - \sum_{i=1}^N \alpha_i \left[y_i \left(x_i^T \left(\sum_{k=1}^N \alpha_k y_k x_k \right) + b \right) - 1 \right] \\ &= \frac{1}{2} \sum_{i=1}^N \sum_{k=1}^N \alpha_i \alpha_k y_i y_k x_i^T x_k - \sum_{i=1}^N \sum_{k=1}^N \alpha_i \alpha_k y_i y_k x_i^T x_k - b \sum_{i=1}^N \alpha_i y_i + \sum_{i=1}^N \alpha_i \\ &= \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{k=1}^N \alpha_i \alpha_k y_i y_k x_i^T x_k \end{aligned}$$

Subject to $\alpha_i \geq 0$

The Lagrangian Dual Form

- The solution is obtained by maximizing L_D with respect to the α_i .
- The solution must satisfy the conditions:

$$w = \sum_{i=1}^N \alpha_i y_i x_i$$

$$\sum_{i=1}^N \alpha_i y_i = 0$$

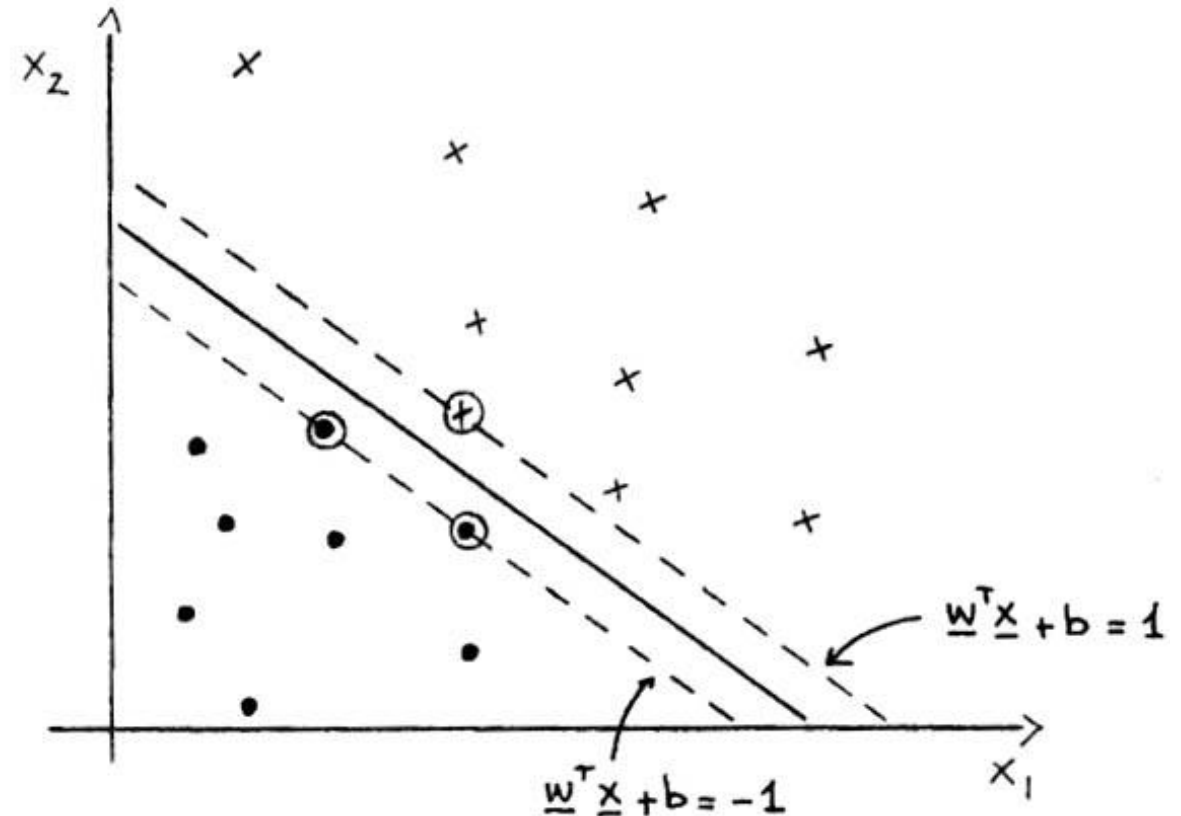
$$\alpha_i \geq 0$$

$$\alpha_i [y_i (w^T x_i + b) - 1] = 0 \quad \forall i = 1, \dots, N$$

Dual Form

$$\alpha_i [y_i(w^T x_i + b) - 1] = 0 \\ \forall i = 1, \dots, N$$

- If $\alpha_i > 0$, then $y_i(w^T x_i + b) = 1$, that is x_i is on the boundary of the margin.
- If $y_i(w^T x_i + b) > 1$, x_i is not on the boundary of the margin, and $\alpha_i = 0$.



Dual Form

- The solution vector w is: $w = \sum_{i=1}^N \alpha_i y_i x_i$. Thus: The solution is defined as a linear combination of those x_i for which $\alpha_i > 0$.
- Such x_i are the points **on the boundary of the margin**. They are called **SUPPORT VECTORS**. We have three support vectors in the above example.
- To obtain the value of b : solve $\alpha_i [y_i (w^T x_i + b) - 1] = 0$ for any of the support vectors.
- The largest margin hyperplane gives a function: $f(x) = w^T x + b$ for classifying new observations $\hat{y} = \text{sign}(f(x))$.

Observations

- The **support vectors** are the critical elements of the training set. They lie closest to the decision boundary.
- **Only the support vectors affect the solution** - If all other training points were removed (or moved around, but so as not to cross the margin), and training was repeated, the same separating hyperplane would be found.
- However, the identification of the support vectors **requires the use of all the training data**.
- Although none of the training observations fall within the margin (by construction), this will not necessarily be the case of test data. (The intuition is that a large margin on the training data indicates a good separation of the two classes and therefore a good separation on the test data as well).

Summary so far

- Training data: $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N), x_i \in R^d, y_i \in \{-1, +1\}$
- **When the two classes are linearly separable**, we can find a function $f(x) = w^T x + b$ with $y_i f(x_i) > 0 \forall i$.
- In particular, we can find the **hyperplane** that creates **the largest margin** between the training points of two classes.

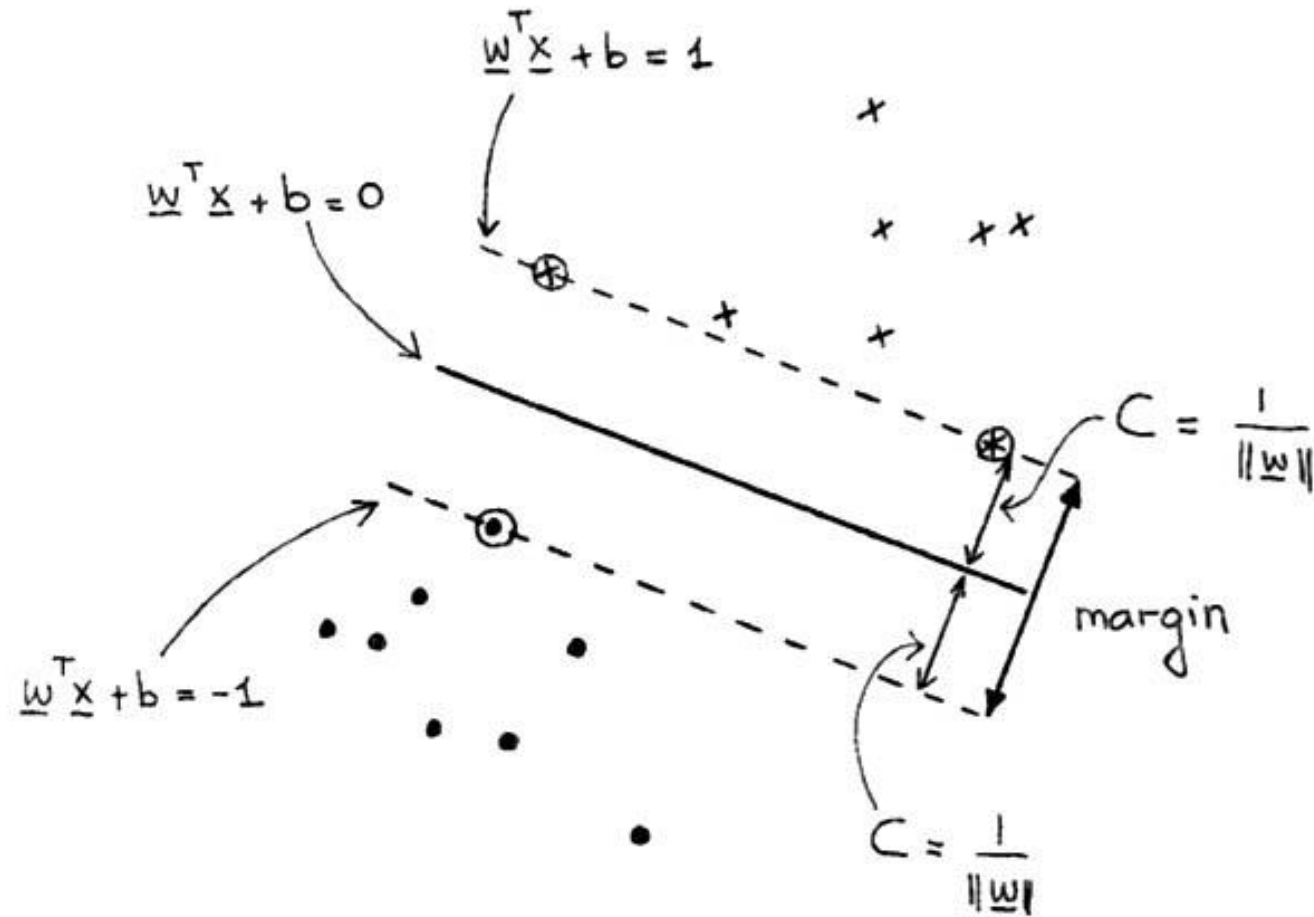
- The optimization problem captures this concept

$$\max_{w,b} 2C \quad \text{subject to} \quad \frac{1}{\|w\|} y_i (w^T x_i + b) \geq C, i = 1, \dots, N$$

- It can be more conveniently rewritten as below where $C = \frac{1}{\|w\|}$

$$\max_{w,b} \frac{1}{2} \|w\|^2 \quad \text{subject to} \quad y_i (w^T x_i + b) \geq 1, i = 1, \dots, N$$

Geometric Perspective



The Non-separable Case

The Non-separable Case

- Suppose now the classes overlap. We can still maximize C , but allow for some points to be on the wrong side of the margin.

- We need to modify the constraints we had for the separable case:

$$\frac{1}{\|w\|} y_i (w^T x_i + b) \geq C, i = 1, \dots, N$$

- To achieve this goal, we define N **slack variables**:

$$\xi_1, \xi_2, \dots, \xi_N$$

- Then a natural way to modify the constraints above is:

$$\frac{1}{\|w\|} y_i (w^T x_i + b) \geq C(1 - \xi_i), i = 1, \dots, N$$

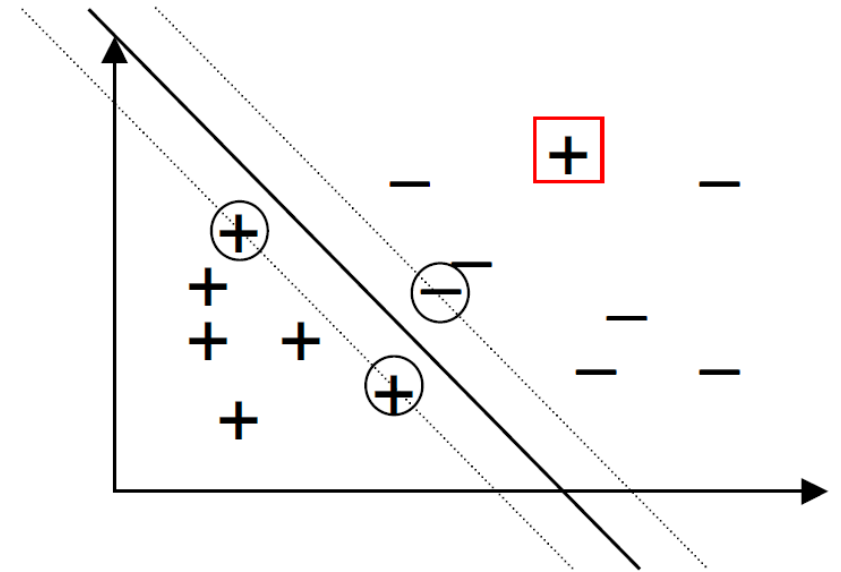
$$\text{with } \xi_i \geq 0, \forall i, \sum_{i=1}^N \xi_i \leq \text{Constant}$$

The Non-separable Case

$$\frac{1}{\|w\|} y_i (w^T x_i + b) \geq C(1 - \xi_i), i = 1, \dots, N$$

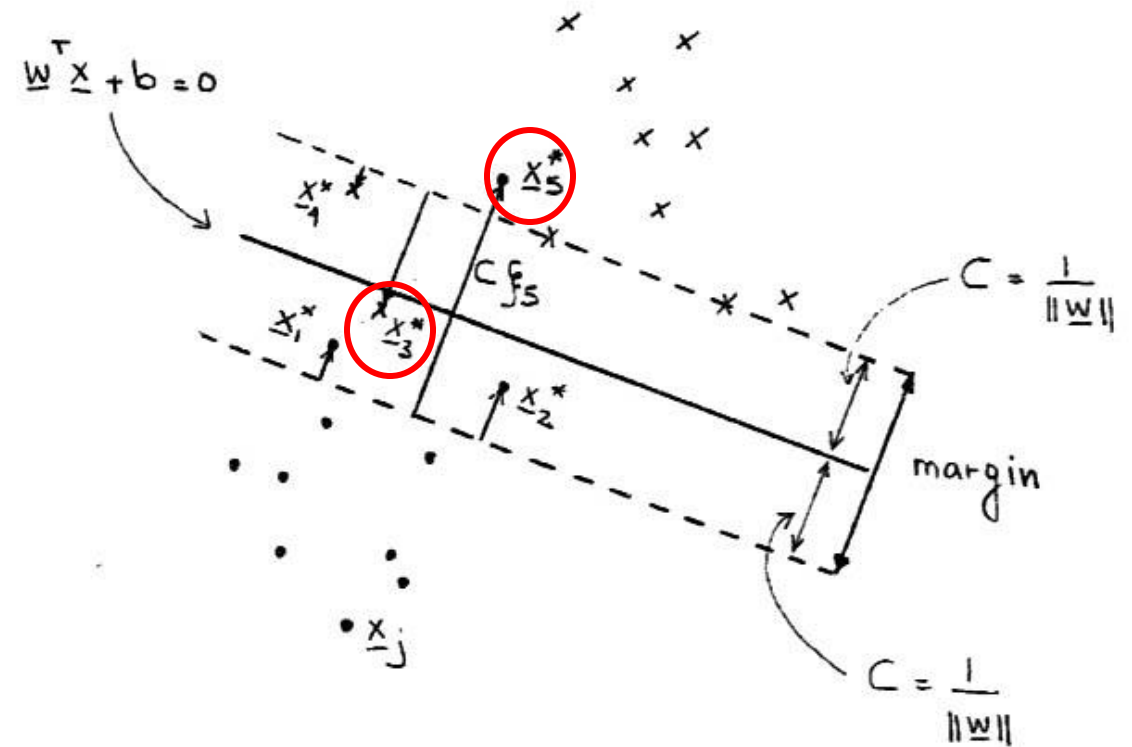
with $\xi_i \geq 0, \forall i, \sum_{i=1}^N \xi_i \leq \text{Constant}$

- Idea of the formulation: ξ_i is the proportional amount by which the prediction $f(x_i)$ is on the wrong side of the margin.
- Note: $C(1 - \xi_i) = C - C\xi_i$



Slack Variables

- The points $(x_1^*, x_2^*, x_3^*, x_4^*, x_5^*)$ are on the wrong side of their margin.
- Point x_i^* is on the wrong side of its margin by an **amount** $C\xi_i$
- Point x_j^* on the correct side have $\xi_j = 0$
- Misclassification occurs when $\xi_i > 1 \Rightarrow C(1 - \xi_i) < 0$, e.g., points x_3^* and x_5^* are misclassified by the given boundary.



A geometric perspective

Slack Variables

- The condition $\sum_{i=1}^N \xi_i \leq \text{Constant}$ bounds the sum $\sum_{i=1}^N \xi_i$.
- Thus, it bounds the total proportional amount by which predictions fall on the wrong side of their margin.
- Since misclassification occur when $\xi_i > 1$ (in this case $y_i f(x_i) < 0$, bounding $\sum_{i=1}^N \xi_i < k$, bounds the **total number of training misclassifications** at k .
- So, for the non-separable case, we have the optimization problem:

$$\max_{w,b} 2C \quad \text{subject to} \quad \frac{1}{\|w\|} y_i (w^T x_i + b) \geq C(1 - \xi_i), i = 1, \dots, N$$
$$\text{with } \xi_i \geq 0, \forall i, \sum_{i=1}^N \xi_i \leq \text{Constant}$$

Slack Variables

- Similar to the separable case, we define $C = \frac{1}{\|w\|}$ and rewrite the above maximization problem in the equivalent form:

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

subject to $y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, \dots, N$

with $\xi_i \geq 0, \forall i, \sum_{i=1}^N \xi_i \leq \text{Constant}$

- We have obtained a quadratic optimization problem with linear constraints. We will solve it using Lagrange multipliers.

Lagrange Multipliers for Slack Variables

- First, one more step: we have seen that the condition $\sum_{i=1}^N \xi_i \leq \text{Constant}$, bounds the number of training misclassifications.
- We can incorporate this condition into the objective function by adding an extra cost for errors:

$$\min_{w,b} \frac{\|w\|^2}{2} + \gamma \sum_{i=1}^N \xi_i$$

subject to $y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, \dots, N$

with $\xi_i \geq 0, \forall i$

here, γ is a parameter to be chosen by the user. A larger γ corresponds to assigning a higher penalty to errors.

Lagrange Multipliers for Slack Variables

$$\begin{aligned} \min_{w,b} \quad & \frac{\|w\|^2}{2} + \gamma \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, \dots, N \\ & \text{with } \xi_i \geq 0, \forall i \end{aligned}$$

- Introducing the Lagrange multipliers α_i and μ_i (one for each constraint), gives the following Lagrange (primal) function:

$$L_p = \frac{1}{2} \|w\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i(w^T x_i + b) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i$$

- Our objective is:

$$\min_{w,b,\xi_i} L_p$$

Lagrange Multipliers for Slack Variables

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

$$\text{subject to } y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, \dots, N$$

$$\xi_i \geq 0, \forall i$$

$$\sum_{i=1}^N \xi_i \leq \text{Constant}$$



$$L_p = \frac{1}{2} \|w\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i(w^T x_i + b) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i$$

Lagrange Multipliers Solution

- Setting the respective derivatives to zero gives:

$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1}^N \alpha_i y_i x_i = 0 \quad \Rightarrow w = \sum_{i=1}^N \alpha_i y_i x_i \quad (3)$$

$$\frac{\partial L_p}{\partial b} = -\sum_{i=1}^N \alpha_i y_i = 0 \quad \Rightarrow \sum_{i=1}^N \alpha_i y_i = 0 \quad (4)$$

$$\frac{\partial L_p}{\partial \xi_i} = \gamma - \alpha_i - \mu_i, \forall i \quad \Rightarrow \alpha_i = \gamma - \mu_i, \forall i \quad (5)$$

along with the positivity constraints $\alpha_i, \mu_i, \xi_i \geq 0, \forall i$

- Substituting Eq. (3), (4), (5) in L_p , we obtain the so called **dual objective function**:

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j x_i^T x_j$$

where L_D gives a lower bound on the objective function $\frac{1}{2} \|w\|^2 + \gamma \sum_{i=1}^N \xi_i$

Deriving the Dual Form

$$\begin{aligned}
 & \frac{1}{2} \left(\sum_{i=1}^N \alpha_i y_i x_i \right)^T \left(\sum_{j=1}^N \alpha_j y_j x_j \right) + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i y_i x_i^T w^T - \sum_{i=1}^N \alpha_i y_i b + \sum_i \alpha_i (1 - \xi_i) - \sum_{i=1}^N \mu_i \xi_i \\
 &= \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j - \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j + \gamma \sum_{i=1}^N \xi_i - \underbrace{b \sum_{i=1}^N \alpha_i y_i}_{=0} + \sum_i \alpha_i (1 - \xi_i) - \sum_{i=1}^N \mu_i \xi_i \\
 &= -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j + \gamma \sum_{i=1}^N \xi_i + \sum_{i=1}^N \alpha_i - \sum_i \alpha_i \xi_i - \sum_i \mu_i \xi_i \\
 &= \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j + \underbrace{\sum_i \overbrace{(\gamma - \mu_i)}^{=\alpha_j} \xi_i}_{=0} - \sum_i \alpha_i \xi_i \\
 &= \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j
 \end{aligned}$$

Lagrange Multipliers Solution

- Thus: the solution is obtained by maximizing L_D w.r.t the α_i , subject to:

$$\sum_{i=1}^N \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq \gamma$$

- The solution must satisfy the conditions:

- $w = \sum_{i=1}^N \alpha_i y_i x_i$ (6)

- $\sum_{i=1}^N \alpha_i y_i = 0$ (7)

- $\alpha_i = \gamma - \mu_i, \forall i$ (8)

- $\alpha_i [y_i (w^T x_i + b) - (1 - \xi_i)] = 0, \forall i$ (9)

- $\mu_i \xi_i = 0, \forall i$ (10)

- $y_i (w^T x_i + b) - (1 - \xi_i) \geq 0, \forall i$ (11)

Lagrange Multipliers Solution

- From (6), the solution is $w = \sum_{i=1}^N \alpha_i y_i x_i$.
- From (9), $\alpha_i > 0$ when constraint (11) is exactly met.
- The points (x_i) with $\alpha_i > 0$ are the **SUPPORT VECTORS**. Two types:
 - Those for which $\xi_i = 0$: they lie on the edge of the margin. From (8) and (10): $0 < \alpha_i < \gamma$
 - Those for which $\xi_i > 0$: they have $\alpha_i = \gamma$ and they lie on the wrong side of their margin.
- To estimate b , we can use (9) with any of the support vectors with $\xi_i = 0$.

$$w = \sum_{i=1}^N \alpha_i y_i x_i \quad (6)$$

$$\sum_{i=1}^N \alpha_i y_i = 0 \quad (7)$$

$$\alpha_i = \gamma - \mu_i, \forall i \quad (8)$$

$$\alpha_i [y_i (w^T x_i + b) - (1 - \xi_i)] = 0, \forall i \quad (9)$$

$$\mu_i \xi_i = 0, \forall i \quad (10)$$

$$y_i (w^T x_i + b) - (1 - \xi_i) \geq 0, \forall i \quad (11)$$

Lagrange Multipliers Solution

- Once we have w and b , the decision function can be written as:

$$\hat{y} = \text{sign}(f(x)) = \text{sign}(w^T x + b)$$

- The tuning parameter of this procedure is γ . Its optimal value can be estimated via cross validation.

Recast as Unconstrained Optimization Problem

- A constrained optimization problem over w and ξ

$$\min_{w,b} \frac{\|w\|^2}{2} + \gamma \sum_{i=1}^N \xi_i$$

Subject to: $y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, \dots, N$

- The constraint can be written more concisely as: $y_i f(x_i) \geq 1 - \xi_i$
together with $\xi_i \geq 0$ is equivalent to

$$\xi_i = \max(0, 1 - y_i f(x_i))$$

- Hence the learning problem is equivalent to the **unconstrained** optimization problem over w :

$$\min_{w,b} \frac{\|w\|^2}{2} + \gamma \sum_{i=1}^N \max(0, 1 - y_i f(x_i))$$

Loss Function

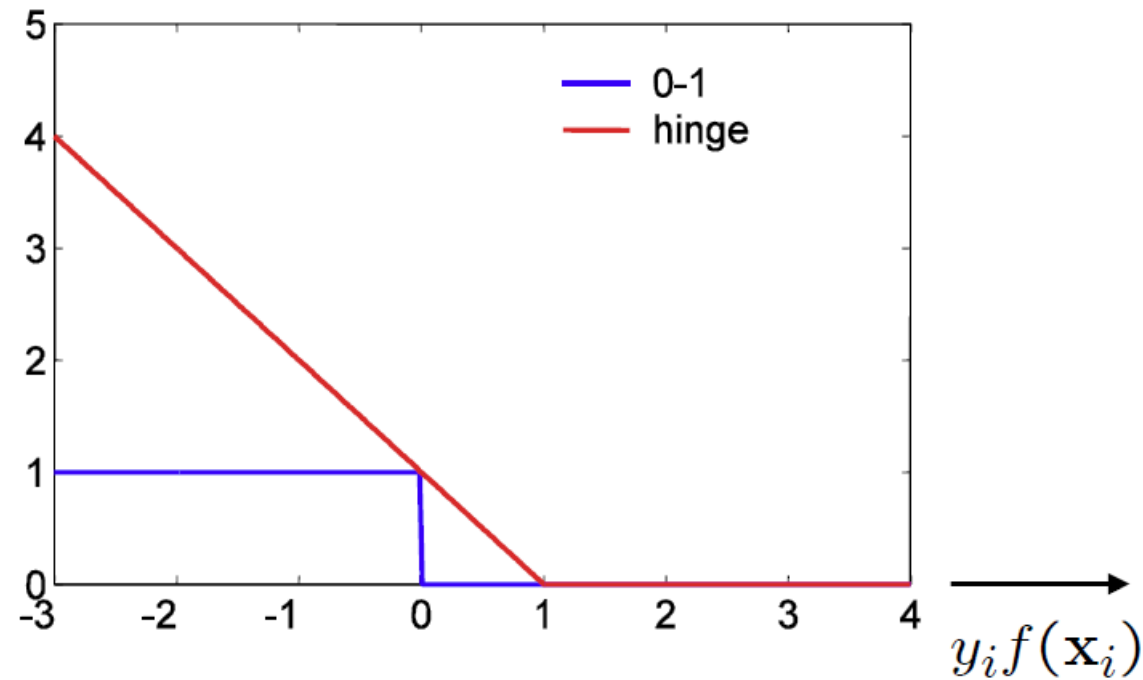
We can consider it as the minimization of a regularized error function.

$$\min_{w,b} \frac{\|w\|^2}{2} + \gamma \sum_{i=1}^N \underbrace{\max(0, 1 - y_i f(x_i))}_{\text{Hinge Loss}}$$

- $y_i f(x_i) > 1$: points outside margin. No contribution to loss.
- $y_i f(x_i) = 1$: points on margin. No contribution to loss (hard margin case)
- $y_i f(x_i) < 1$: points violates margin constraints. Contribute to loss.

Hinge Loss

Hinge loss vs 0-1 loss



- A convex approximation to the 0-1 loss.
- It is an upper bound on the 0-1 loss.
- Not differentiable, we need to compute the **sub**-gradient.

Optimization

- Solving the Quadratic Programming Problems
- Platt's sequential minimal optimization (SMO) algorithm
- Stochastic **sub**-gradient descent algorithms.

Readings

1. http://www.personal.psu.edu/sxj937/Notes/Lagrange_Multipliers.pdf
2. <http://cs229.stanford.edu/summer2020/cs229-notes3.pdf>

Summary of Today's Lecture

- Vector algebra, Formulation, Margin
- SVM for Linear Separable Case
- Non-separable Case, Penalties

Slides

- Credit: Ping Wang