CS 461: Machine Learning Principles

Class 12: Oct. 14

SVM: Support Vector Machine

Instructor: Diana Kim

- LDA (GDA)
- Logistic Regression
- Perceptron

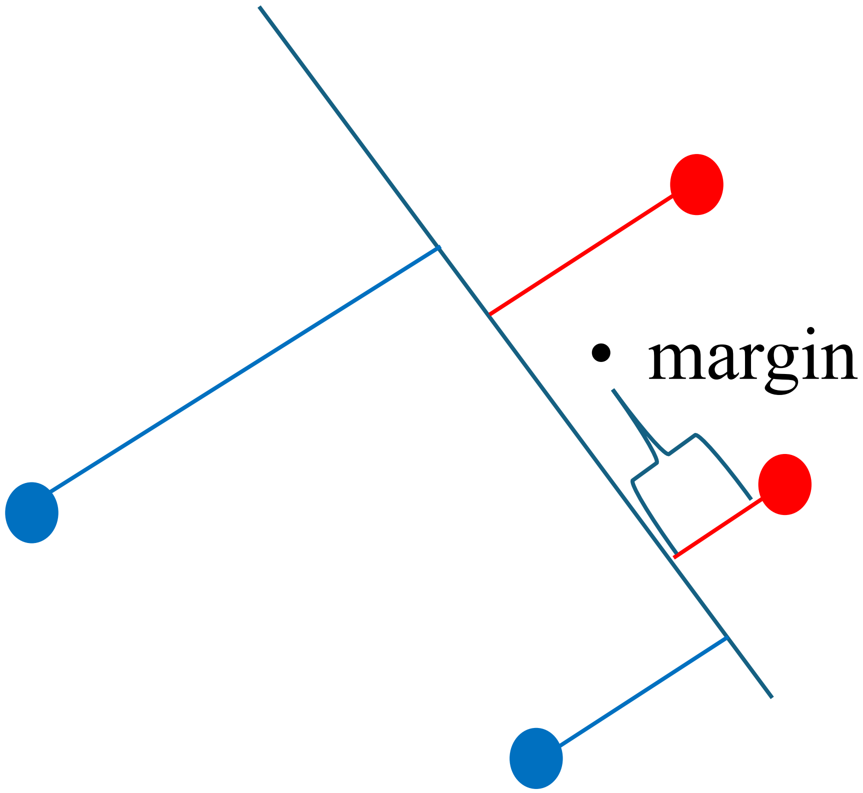
→

All define
A linear decision boundary
on the space of $\phi(x)\ or\ x$.

Today, we are going to learn
how to compute a decision boundary that gives the maximum margin.

Preliminary (2) the Concept of Margin

Margin: the smallest distance
between the decision boundary and any of the samples



- margin

- Motivation of learning maximum margin classifier

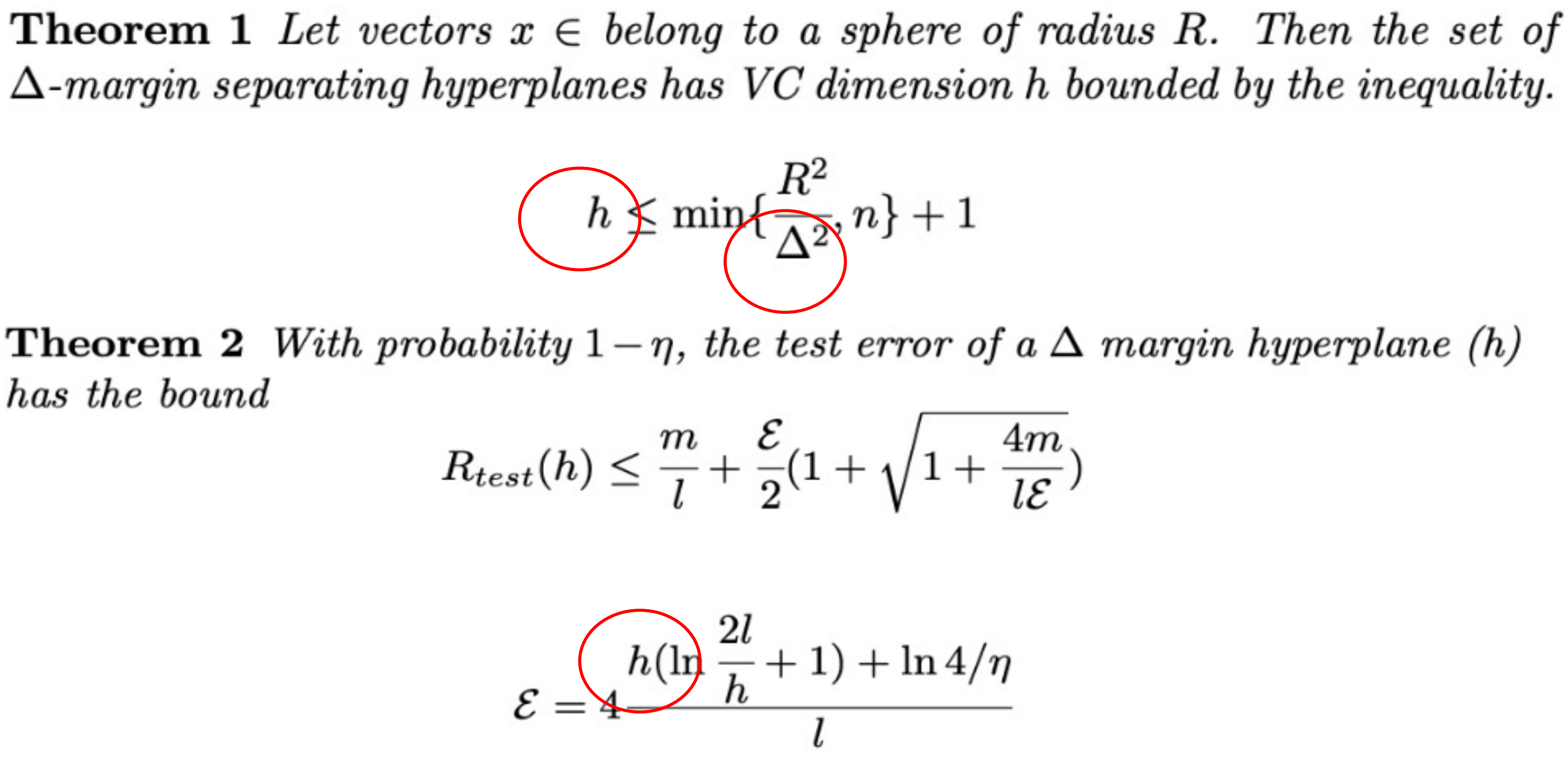
The large margin decision boundary on train set
will be more beneficial for generalization (unseen data performance)
than small margin cases.

- Robust to small perturbations of the data. (no change in classification)
- Robust to small perturbations in the estimation of the decision surface

Learning theory proves

From The nature of statistical learning theory by Vladmir N. Vapnik

- Large margin implies bounded Vapnik–Chervonenkis (VC) dimension.
- Bounded VC dimension implies high probability of low generalization error.



Theorem 1 *Let vectors $x \in$ belong to a sphere of radius R. Then the set of Δ-margin separating hyperplanes has VC dimension h bounded by the inequality.*

$$h \leq \min\{\frac{R^2}{\Delta^2}, n\} + 1$$

Theorem 2 *With probability $1 - \eta$, the test error of a Δ margin hyperplane (h) has the bound*

$$R_{test}(h) \leq \frac{m}{l} + \frac{\mathcal{E}}{2}(1 + \sqrt{1 + \frac{4m}{l\mathcal{E}}})$$

$$\mathcal{E} = 4\frac{h(\ln\frac{2l}{h} + 1) + \ln 4/\eta}{l}$$

Okay, a large margin classifier will be beneficial for generalization,
How about the methods we have learned so far?

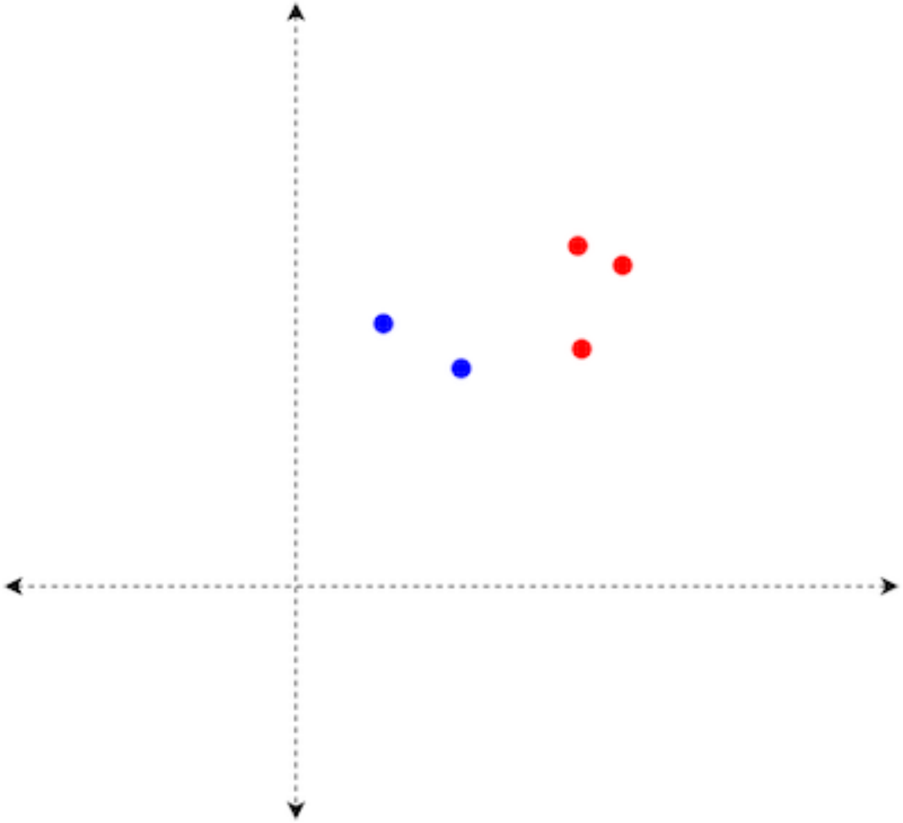
- LDA
- Logistic Regression
- Perceptron

→ All define
linear decision boundaries
on the space of $\phi(x)\ or\ x$.

Q: Which one considered margin as defining its decision boundary?

+ margin does not matter in LDA and Perceptron but the nature of Logistic Regression algorithm promotes the larger margin.

Logistic Regression



$$P(t|w) = \prod_{n=1} \sigma(w^t x_n)^{t_n} (1 - \sigma(w^t x_n))^{1-t_n}$$

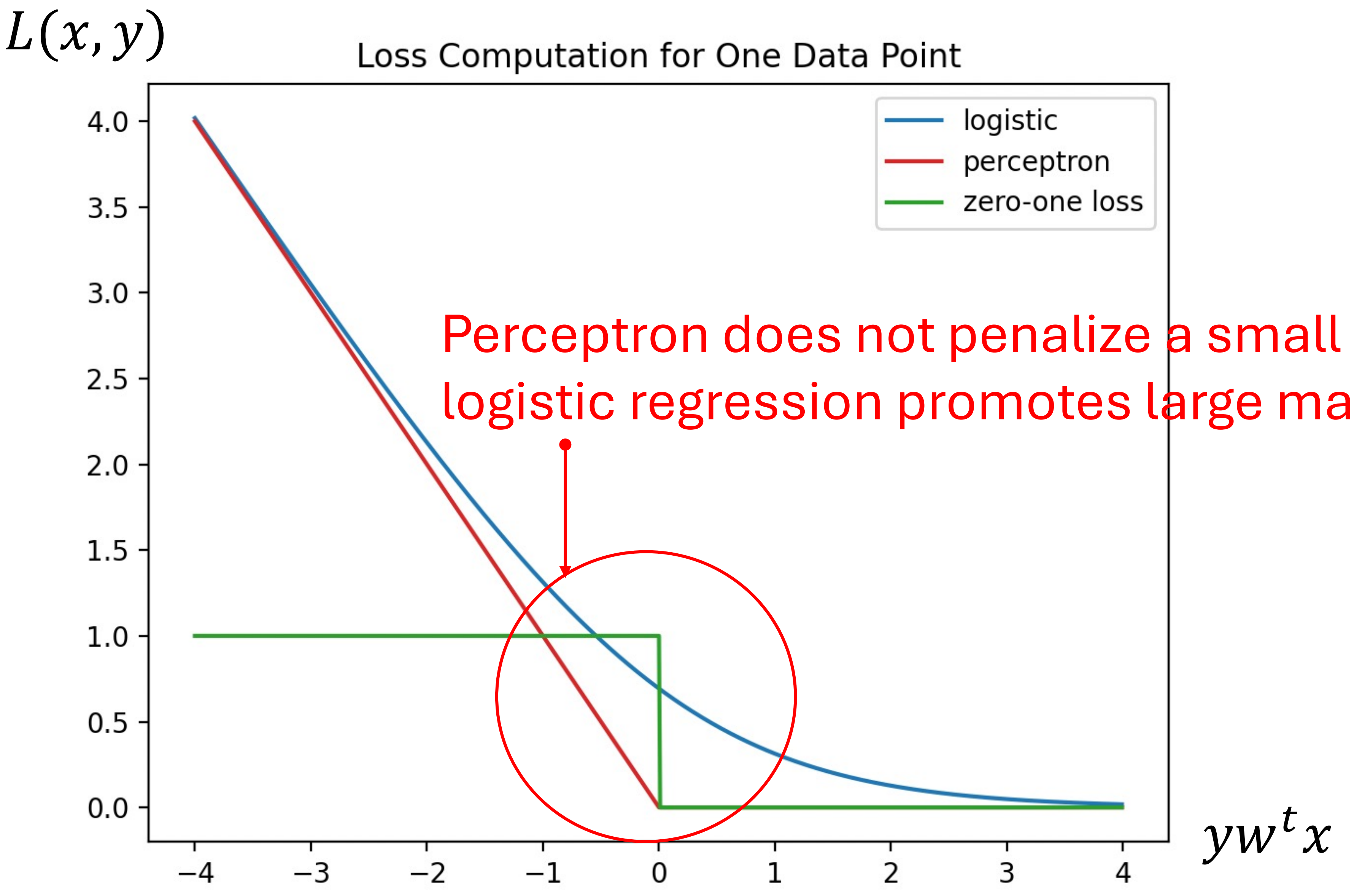
$$J(w) = -\ln P(t|w) = \sum_{n=1}^{N} -t_n \ln \sigma(w^t x_n) - (1 - t_n) \ln (1 - \sigma(w^t x_n))$$

By the n$ature\ of\ J(w)$ in training:

$$\sigma(w^t x_n) \simeq 1 \ (x_n \text{ is positive sample})$$
$$\sigma(w^t x_n) \simeq 0 \ (x_n \text{ is negative sample})$$

The larger margin is preferred.

Loss Comparison for One Data Point (Objective Function)

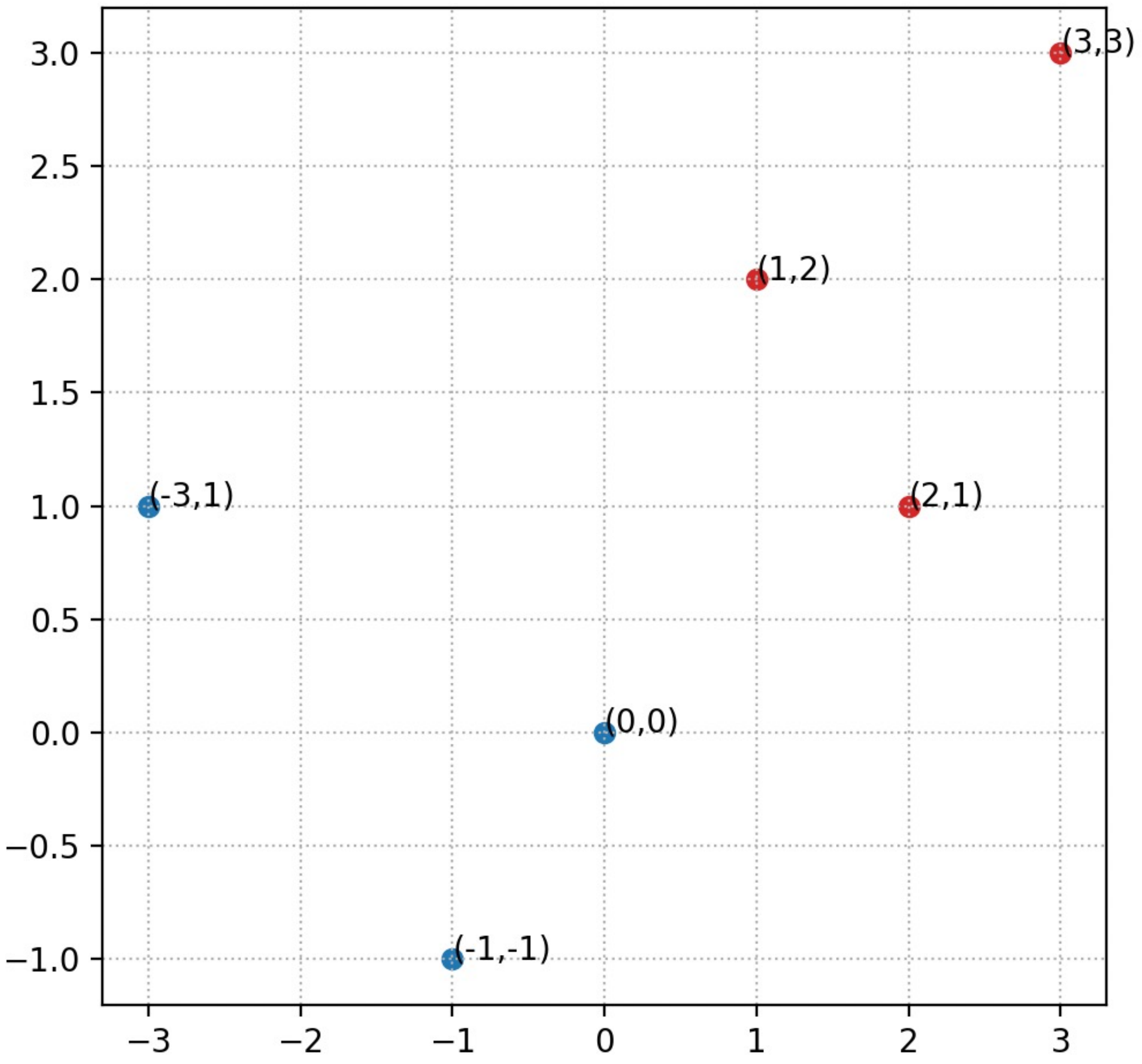


Perceptron does not penalize a small margin.
logistic regression promotes large margin.

SVM (**S**upport **V**ector **M**achine)

This is the method to compute the Maximum Margin Boundary

Example) Finding a Maximum Margin Classifier (2-D)



1. Pick one blue and one red sample

2. Compute a hyper plane by the two points.

3. Compute the margin for the plane

4. Repeat 1.2,3 for other points combinations.

5. Pick the plane gives the maximum margin.

- SVM Problem formulation on Linearly Separable Data:

  given $\mathcal{D} : \{(x, y) : x \in \mathbb{R}^M \; and \; t \in \{-1, 1\}\}$
  which is linearly separable,
  how can we find the maximum margin classifier?

SVM Problem Formulation

$$w*, b* = \arg\max_{w,b} \min_{n} \frac{t_n(w^t x_n + b)}{||w||}$$

$$\Leftrightarrow w*, b* = \arg\max_{w,b} \frac{\Delta}{||w||}$$

$$\text{subject to} \quad t_n(w^t x_n + b) \geq \Delta \quad \forall n$$

$$\Delta \geq 0$$

$$\Leftrightarrow w*, b* = \arg\max_{w,b} \frac{1}{||w||}$$

$$\text{subject to} \quad t_n(w^t x_n + b) \geq 1 \quad \forall n$$

Translation of the Original Problem to Constrained Optimization Problem

$$\Leftrightarrow w*, b* = \boxed{\arg\max_{w,b} \frac{1}{||w||}}$$

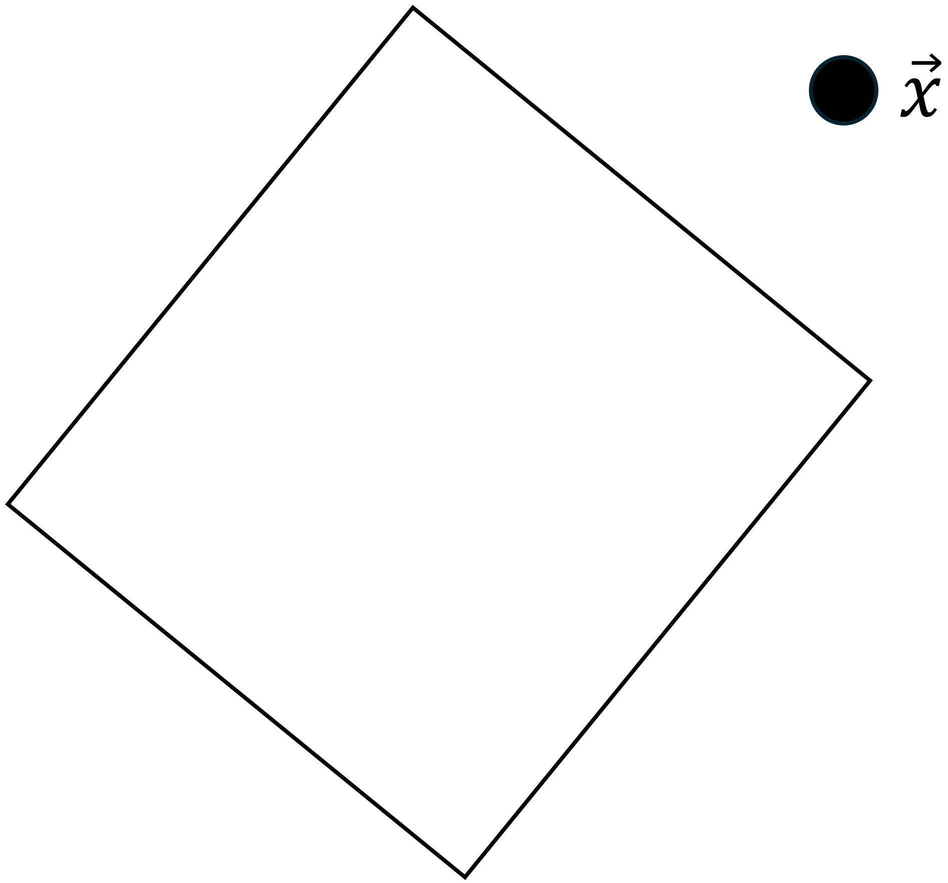
$$\text{subject to} \quad t_n(w^t x_n + b) \geq 1 \quad \forall n$$

$$\Leftrightarrow w*, b* = \boxed{\arg\min_{w,b} \frac{1}{2}||w||^2}$$

$$\text{subject to} \quad t_n(w^t x_n + b) \geq 1 \quad \forall n$$

Preliminary (1)

Distance from a Point ($\vec{x}$) to a Hyperplane $\vec{w}^t \cdot \vec{x} = 0$

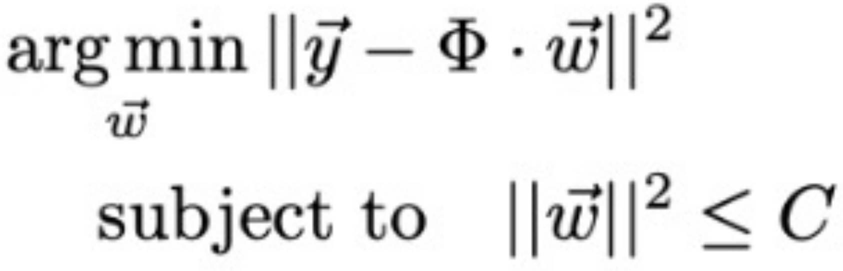


+ distance: $\frac{|w^t x|}{||w||}$

Lagrangian Function for the Constrained Primal Problem

[The slides from the previous classes]

Our Optimization Problem from Regression



$$\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2$$

$$\text{subject to} \quad ||\vec{w}||^2 \leq C$$

We would like to control the effective model complexity by constraining the magnitude of parameter.

Forming a Lower Bound to $f(x)$ by Lagrangian Function

$$||\vec{y}-\Phi\cdot\vec{w}||^2+\lambda(||\vec{w}||^2-C)\leq||\vec{y}-\Phi\cdot\vec{w}||^2$$

$$\boxed{f(x)+\lambda g(x)}\leq f(x)$$

Lagrangian function $L(x,\lambda)$

- $f(x)$ *is the original objective functon*
- *g(x) is the inequality constraint*
- $\lambda\geq 0$
- $g(x)\leq 0$

Karush-**K**hun-**T**ucker Necessary condition
KKT condition defines $x*$ in relation to a certain $\lambda*$.

- Optimization problem

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \end{aligned}$$

- Let $x*$be a local minimal of the primal problem and $L(x,\lambda) = f(x) + \lambda g(x)$
- then there exist unique Lagrangian multiplier $\lambda*$ s.t

$$\begin{aligned} \nabla_x L(x*, \lambda*) &= 0 \\ \lambda* = 0 \quad \text{if} \quad g(x*) &< 0 \\ \lambda* = \text{positive} \quad \text{if} \quad g(x*) &= 0 \end{aligned}$$

Setting up the Lagrangian Function for SVM Primary Problem

- [Primary Problem]

$$w*, b* = \arg\min_{w,b} \frac{1}{2}||w||^2$$

$$\text{subject to} \quad t_n(w^t x_n + b) \geq 1 \quad \forall n$$

- [Lagrangian Function]

$$L(w, b, \lambda_{n=1}^{N}) = \frac{1}{2}||w||^2 - \sum_{n=1}^{N} \lambda_n \cdot t_n \cdot (w^t x_n + b) - \lambda_n$$

The first KKT Condition gives $\boxed{\vec{w}*} = \sum_{n=1}^{N} \boxed{\lambda_n^*} \cdot t_n \cdot \vec{x_n}$

[1] $\nabla_x L(x*, \lambda*) = 0$

$$\nabla_w L(w, b) = \vec{w} - \sum_{n=1}^{N} \lambda_n \cdot t_n \cdot \vec{x_n}$$

$$\nabla_b L(w, b) = \sum_{n=1}^{N} \lambda_n \cdot t_n$$

[2,3] $\text{The second and third KKT} = \begin{cases} \lambda*_n = 0 & \text{if} \quad w*^t x_n - 1 > 0 \\ \lambda*_n > 0 & \text{if} \quad w*^t x_n - 1 = 0 \end{cases}$

The first KKT condition gives $\vec{w}* = \sum_{n=1}^{N} \lambda_n^* \cdot t_n \cdot \vec{x_n}$

And the relation defines the lower bound function $L'(\lambda *)$

- $L' \leq L$

$$\boxed{L'(\lambda*) = L(w* = \sum_{n=1}^{N} \lambda_n^* \cdot t_n \cdot \vec{x_n}, \lambda*) \leq L(w, \lambda*)}$$

$$L'(\lambda*) = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n^* \cdot t_n \lambda_m^* \cdot t_m \cdot \vec{x_n}^t \vec{x_m} - \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n^* \cdot t_n \lambda_m^* \cdot t_m \cdot \vec{x_n}^t \vec{x_m} - \sum_{n=1}^{N} \lambda_n^* \cdot t_n \cdot b + \sum_{n=1}^{N} \lambda_n$$

$$\boxed{L'(\lambda*) = -\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n^* \cdot t_n \lambda_m^* \cdot t_m \boxed{\cdot \vec{x_n}^t \vec{x_m}} + \sum_{n=1}^{N} \lambda_n}$$

- *L' is the dual functon of the primal problem.*
  now we need to solve *$\lambda *$ by the dual problem.*

SVM
Primal & Dual Problem

- Primal

$$w*, b* = \underset{w,b}{\arg\min} \frac{1}{2}||w||^2$$

$$\text{subject to} \quad t_n(w^t x_n + b) \geq 1 \quad \forall n$$

- Dual

$$\lambda*_{n=1}^{N} = \underset{\lambda*}{\arg\max} \sum_{n=1}^{N} \lambda *_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n^* \cdot t_n \lambda_m^* \cdot t_m \cdot \kappa(x_n, x_m)$$

$$\text{subject to} \quad \lambda_n^* \geq 0 \quad \forall n$$

$$\sum_{n=1}^{N} \lambda_n^* \cdot t_n = 0$$

SVM: Support Vector Machine $y(x)$

$$\lambda*_{n=1}^{N} = \arg\max_{\lambda*} \sum_{n=1}^{N} \lambda *_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n^* \cdot t_n \lambda_m^* \cdot t_m \cdot \kappa(x_n, x_m)$$

$$\text{subject to} \quad \lambda_n^* \geq 0 \quad \forall n$$

$$\sum_{n=1}^{N} \lambda_n^* \cdot t_n = 0$$

Once we got $\lambda *$,

$we\ can\ build\ an\ SVM\ classifier.$

$$y(x) = \sum_{n=1}^{N} \lambda *_n \, t_n \kappa(x_n, x) + b$$

$$\vec{w}* = \sum_{n=1}^{N} \lambda_n^* \cdot t_n \cdot \vec{x_n}$$

SVM: Support Vector Machine $y(x)$

$$y(x) = \sum_{n=1}^{N} \lambda *_n t_n \kappa(x_n, x) + b$$

- Classification by computing the linear combination of inner product between x train data points.

- # of parameters are not fixed like $M,$
  where M is the dimension of feature space,
  # of parameters for λ is same as # of data points.

- Q: All the data points matter in $y(x)$?

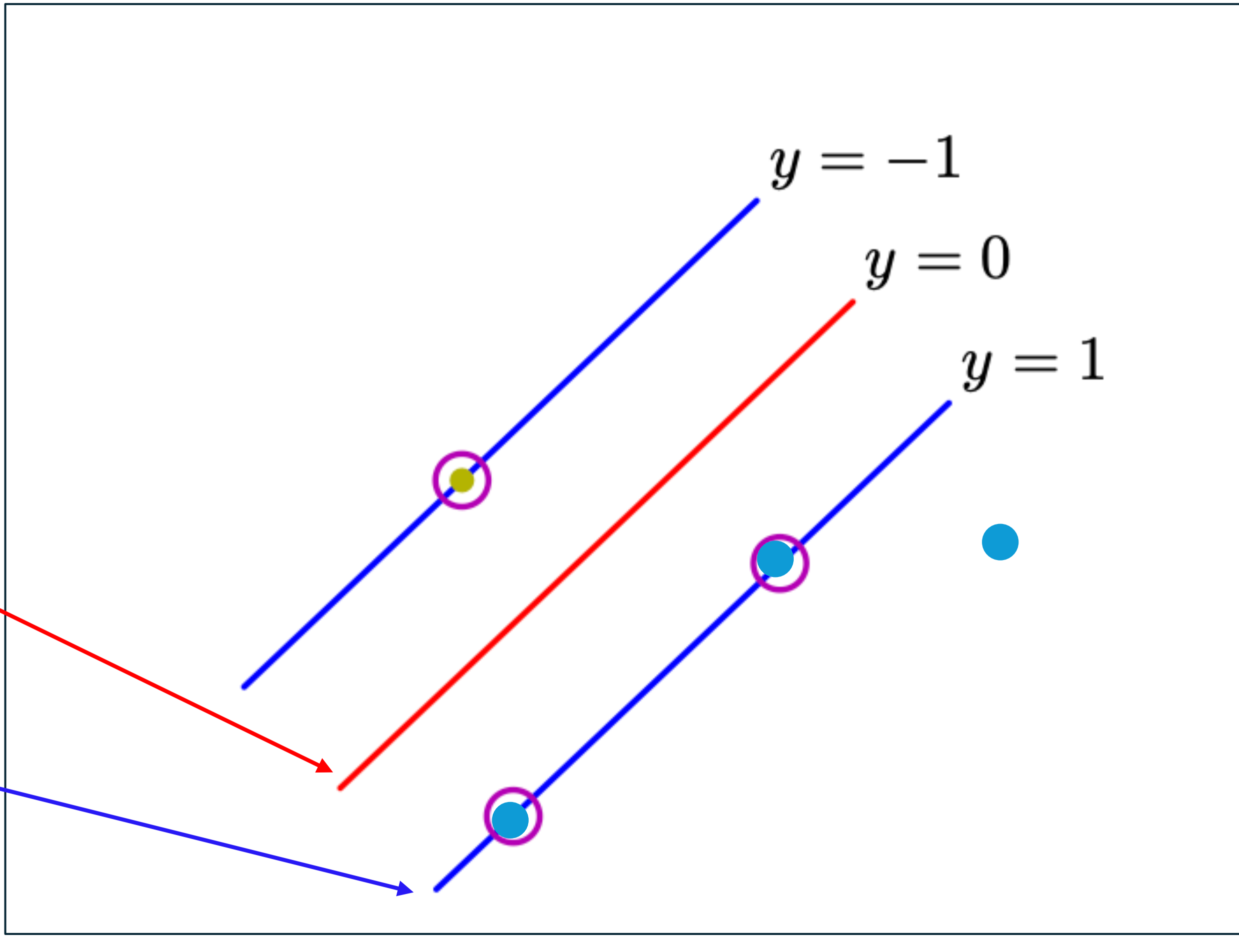
Not all training data points are not used to define $y(x)$.

$$y(x) = \sum_{n=1}^{N} \lambda *_n t_n \kappa(x_n, x) + b$$

From the second and third KKT condition,
We can see that only the data samples on the margin matter.
The machine is named "Sparse Vector Machine"

$$\text{The second and third KKT} = \begin{cases} \lambda *_n = 0 & \text{if} \quad w *^t x_n - 1 > 0 \\ \lambda *_n > 0 & \text{if} \quad w *^t x_n - 1 = 0 \end{cases}$$

From Bishop Figure 7.1



The Outcomes of Linear Kernel SVM

(1) Decision hyperplane

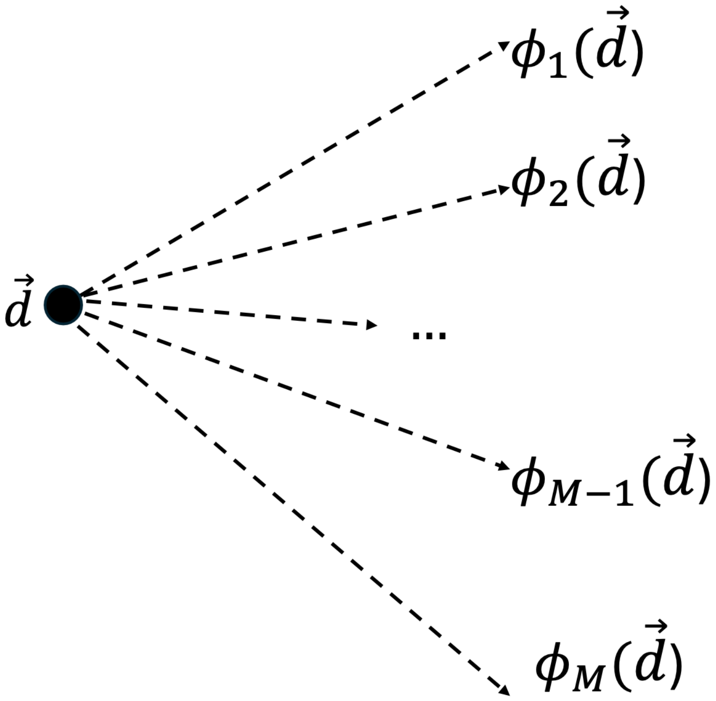
(2) Support Vectors (Samples)

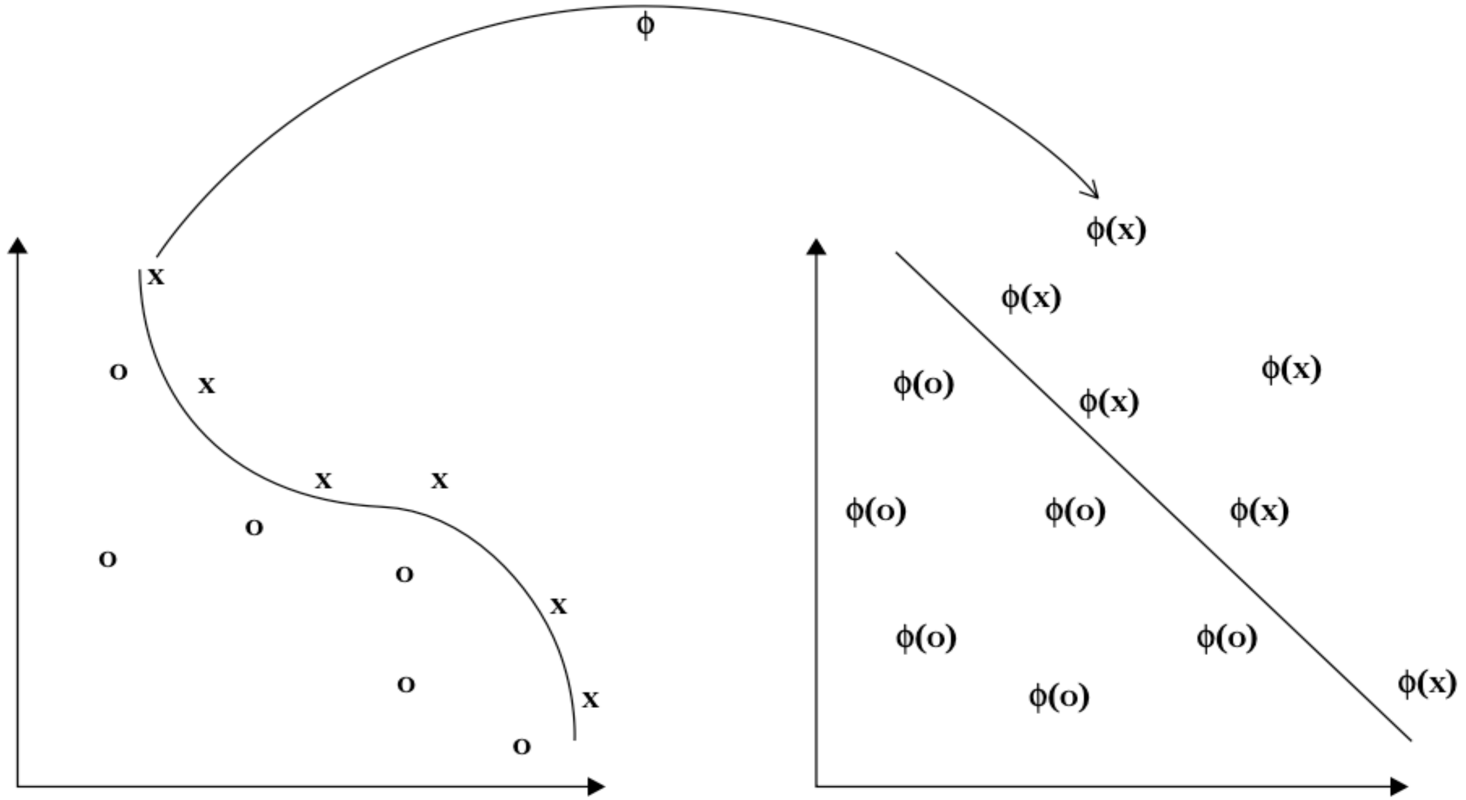
(3) Margins

Initially we build SVM based on the assumption “linearly separable data”
What if the data is not linearly separable?

- SVM on Linearly Separable Data:
  given $\mathcal{D} : \{(x, y): x \ \in \ \mathbb{R}^M \ and\ t \in \{-1, 1\}\}$
  which is linearly separable,
  how can we find the maximum margin classifier?

We know the use of a well-designed high dimensional feature map can make the data separable!





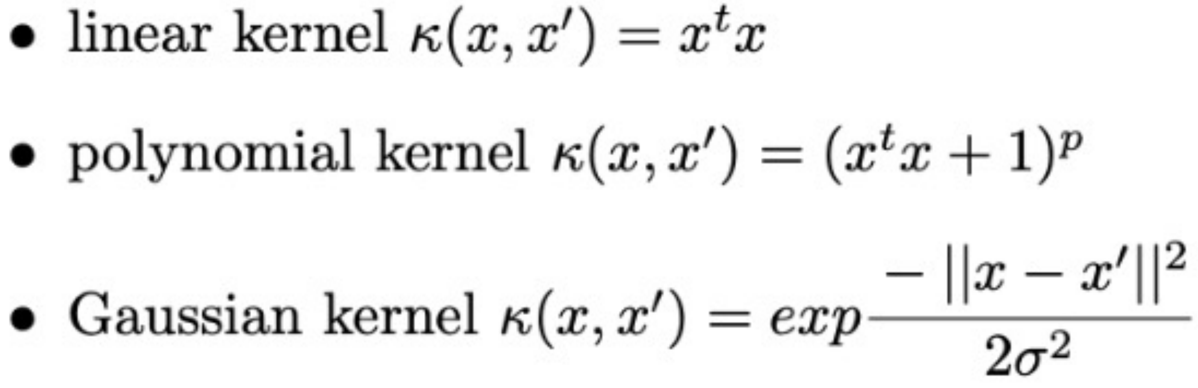
From Kernel Methods for Pattern Analysis by John Shawe-Talyor

Fig. 2.1. The function ϕ embeds the data into a feature space where the nonlinear pattern now appears linear. The kernel computes inner products in the feature space directly from the inputs.

We know the use of a well-designed high dimensional feature map can make the data separable!
Based on the problem formulation for SVM in the previous slides,
Q: do we need an explicit feature map design?

+ we don’t need explicit feature design, the dual function only needs kernel functions.

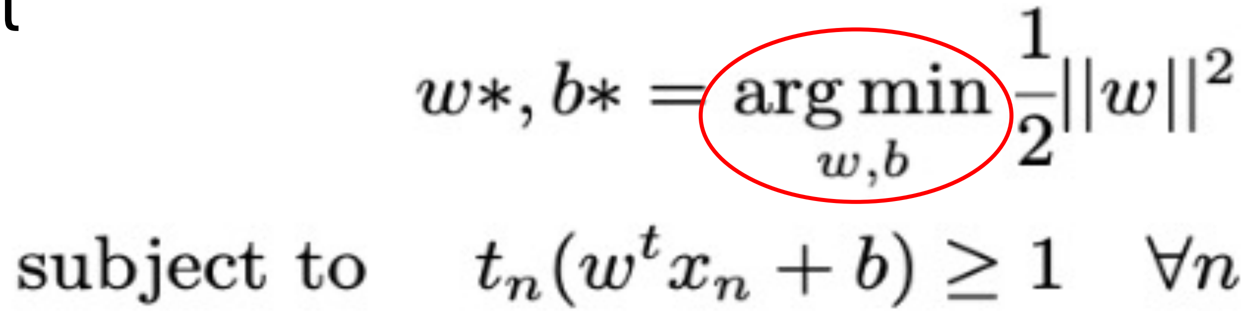
In SVM process, what parts will be changed
if we apply different kernels other than the linear kernel?



- linear kernel $\kappa(x, x') = x^t x$
- polynomial kernel $\kappa(x, x') = (x^t x + 1)^p$
- Gaussian kernel $\kappa(x, x') = exp \frac{- ||x - x'||^2}{2\sigma^2}$

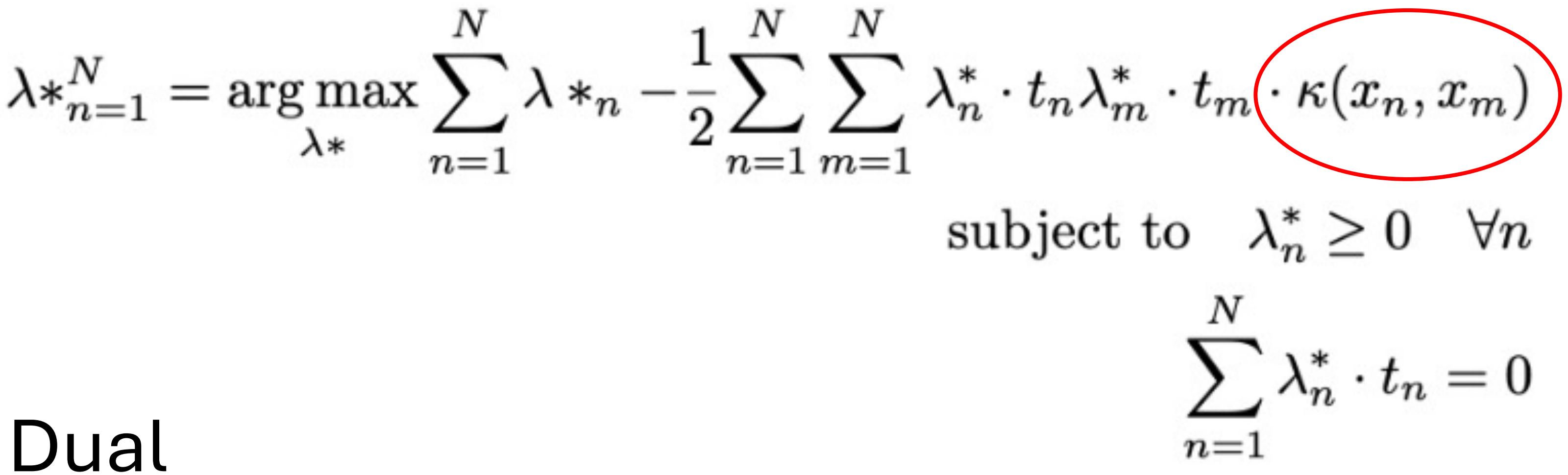
SVM
Primal & Dual Problem: dual problem will be changed!

- Primal



$$w*, b* = \arg\min_{w,b} \frac{1}{2}||w||^2$$

$$\text{subject to} \quad t_n(w^t x_n + b) \geq 1 \quad \forall n$$



$$\lambda*_{n=1}^{N} = \arg\max_{\lambda*} \sum_{n=1}^{N} \lambda *_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n^* \cdot t_n \lambda_m^* \cdot t_m \cdot \kappa(x_n, x_m)$$

$$\text{subject to} \quad \lambda_n^* \geq 0 \quad \forall n$$

$$\sum_{n=1}^{N} \lambda_n^* \cdot t_n = 0$$

- Dual

SVM: Final support vector machine $y(x)$ will be changed!

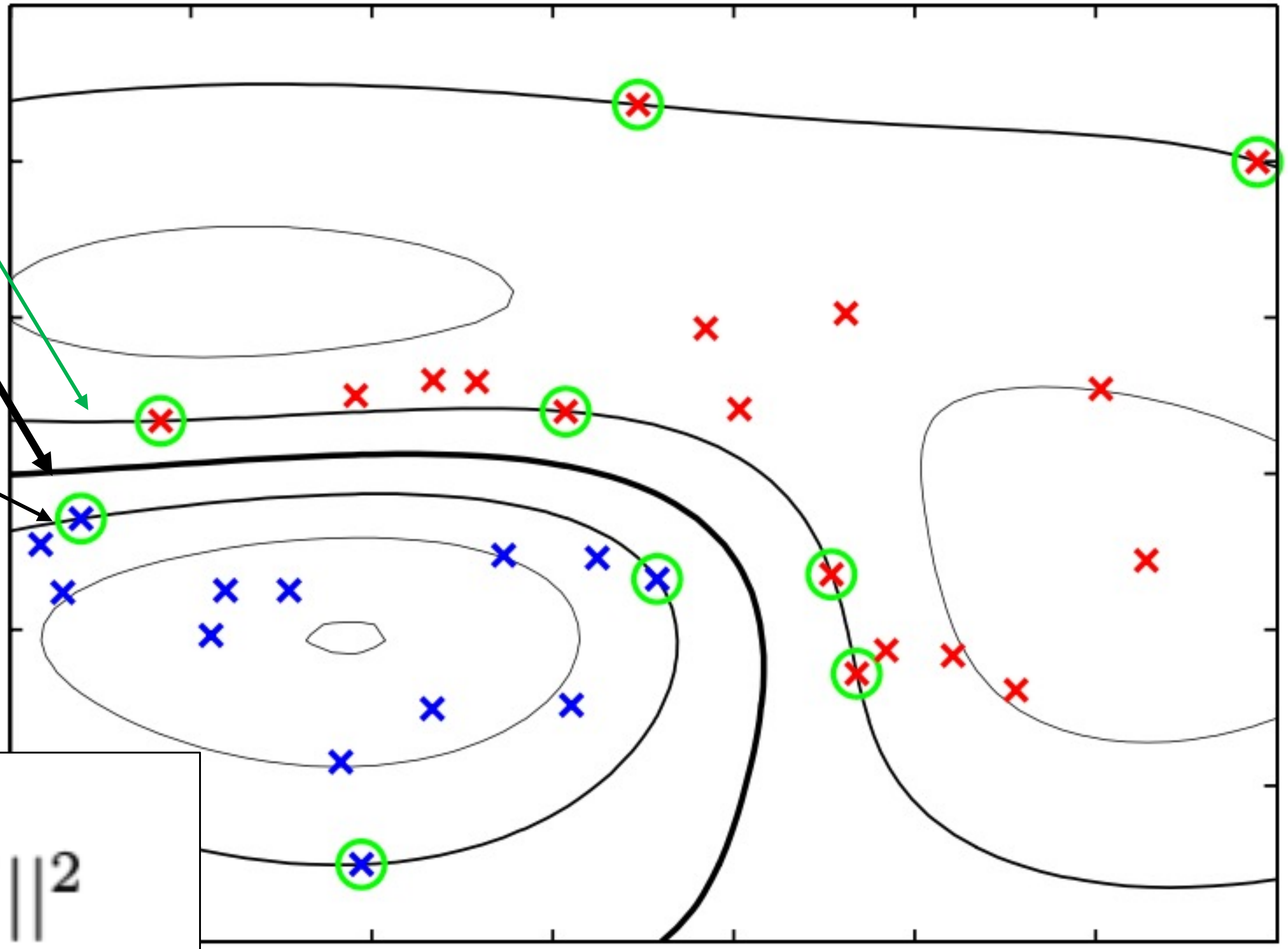
$$y(x) = \sum_{n=1}^{N} \lambda *_n t_n \kappa(x_n, x) + b$$

The Outcomes of Gaussian Kernel SVM

(1) **Decision hyperplane**

(2) Support Vectors (Samples)

(3) Margins



Gaussian Kernel

$$\kappa(x, x') = \exp \frac{-||x - x'||^2}{2\sigma^2}$$

Q1) How Gaussian Kernel SVM finds the optimal solution in the infinite dimensional space with limited data points?

+ for the previous methods like logistic regression, the limited data points in high dimensional modeling cause overfitting. (large variance) . The algorithm finds a decision boundary based on all training data points in the feature space, so total number of data and how they are shaped matter

+ However, for SVM, the optimal decision boundary is defined by a subset of training samples (support vectors) so it is less sensitive to # data points. Of course, different training sets can give different hyperplanes but would not much differ because the large margin constraint is robust to perturbation by data.

+ it is true that SVM build a classifier by using limited data points,
but the # support vectors is controlled by the sigma and #support vectors is related to the complexity.

+ sigma controls the complexity of SVM we need the proper value of sigma depending of your goal.

Q2) How about the sigma?

- $\sigma \to 0$

All data points become support vectors
no meaningful decision region.
The SVM machine has very spiky RBMs.

- $\sigma \to \infty$

One data point become support vector.
No meaningful decision region
The SVM machine has a flat RBM (like uniform)

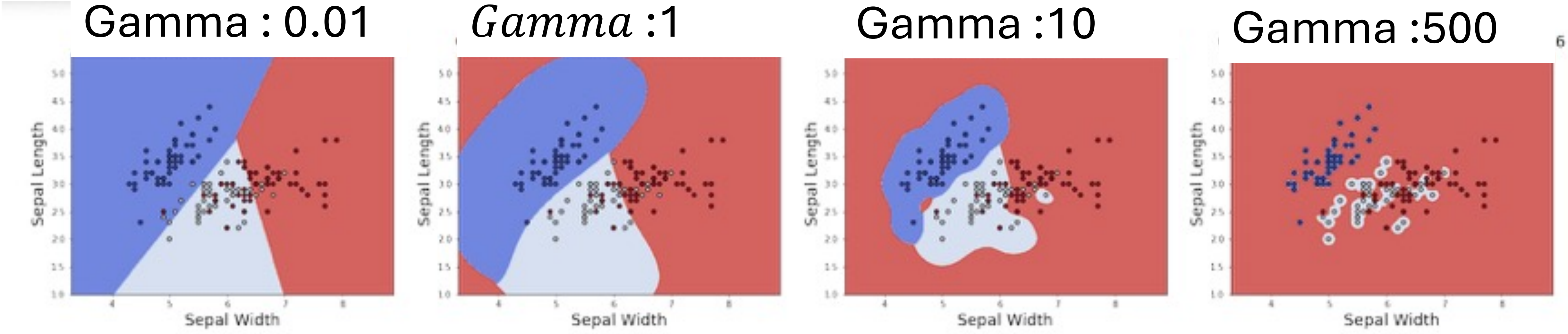
Gaussian Kernel

$$\kappa(x, x') = \exp \frac{-||x - x'||^2}{2\sigma^2}$$

The Effect of Gamma
on the number of Support vectors, Decision boundary, and Margins

$$y(x) = \sum_{n=1}^{N} \lambda *_n t_n \kappa(x_n, x) + b$$

+ each data sample becomes a support vector


Gamma : 0.01

Gamma :1

Gamma :10

Gamma :500

From https://www.kaggle.com/code/gorkemgunay/understanding-parameters-of-svm

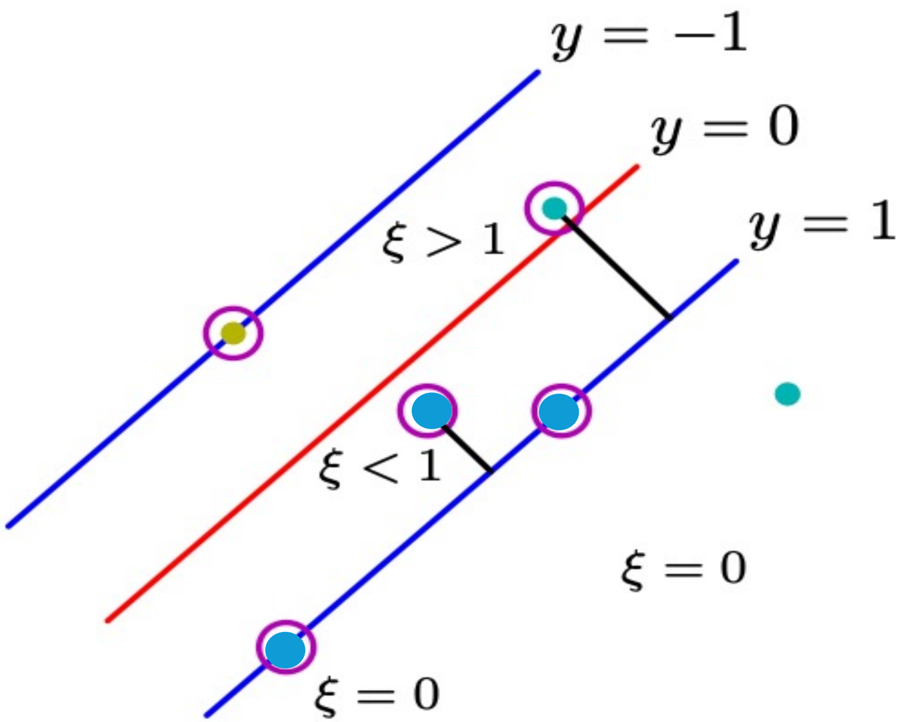
+Based on the figure,
the three RBFs by the three support vectors, so the three decision boundaries are drawn.

$$\gamma = \frac{1}{\sigma^2}$$

The Gaussian kernel trick in SVM enables us to work on high dimensional space that is linearly separable and find the exact separation hyperplane on training data. However, this often results in a complex decision boundary fitting to training data. Now we modify SVM to allow the wrong side of the margin boundary to reduce complexity and for better generalization.

For each data points,
the constraint to define the margin is
relaxed by the slack variable $\xi_n \geq 0$

From Bishop Figure 7.3



- hard margin

$$t_n(w^t x_n + b) \geq 1 \quad \forall n$$

- soft margin

$$t_n(w^t x_n + b) \geq 1 - \xi n \quad \text{and} \quad \xi n \geq 0 \quad \forall n$$

- Hard Margin Case (Exact-Separable)

$$w*, b* = \arg\min_{w,b} \frac{1}{2}||w||^2$$

$$\text{subject to} \quad t_n(w^t x_n + b) \geq 1 \quad \forall n$$

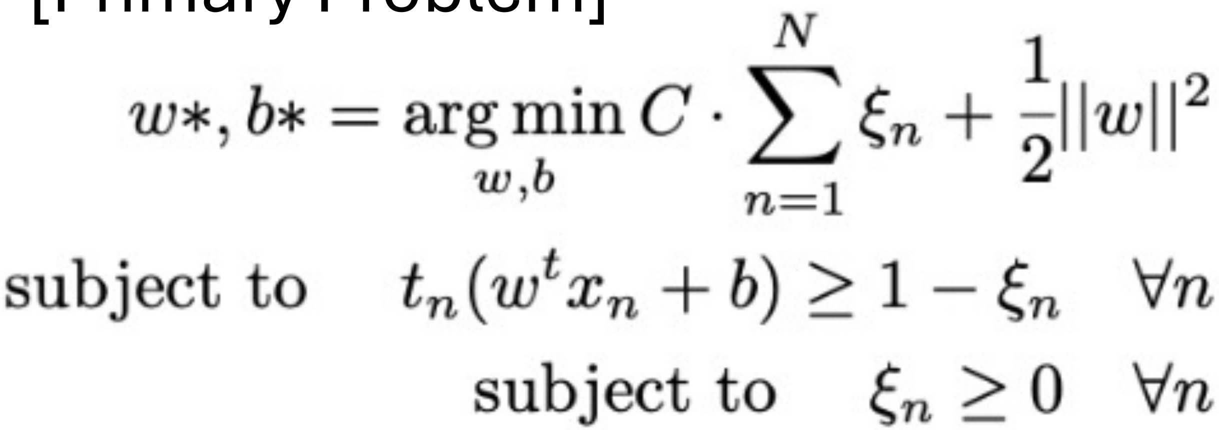
- Soft Margin Case (Non-Separable)

$$w*, b* = \arg\min_{w,b} C \cdot \sum_{n=1}^{N} \xi_n + \frac{1}{2}||w||^2$$

$$\text{subject to} \quad t_n(w^t x_n + b) \geq 1 - \xi_n \quad \forall n$$

$$\text{subject to} \quad \xi_n \geq 0 \quad \forall n$$

- [Primary Problem]

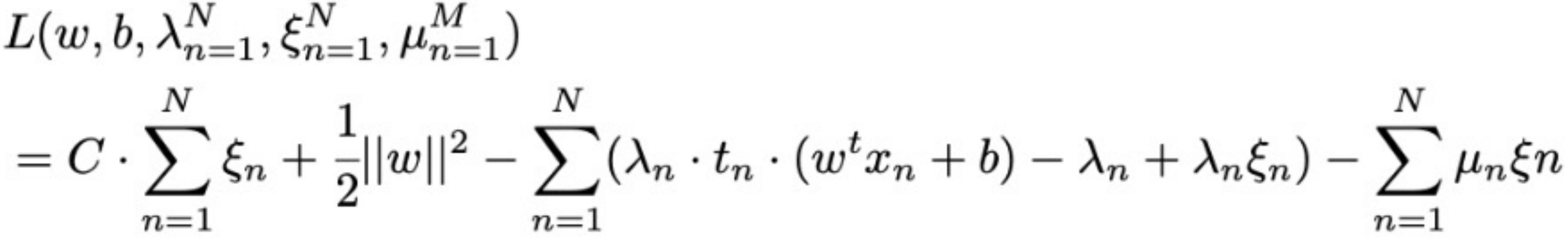


$$w*, b* = \arg\min_{w,b} C \cdot \sum_{n=1}^{N} \xi_n + \frac{1}{2}||w||^2$$

$$\text{subject to} \quad t_n(w^t x_n + b) \geq 1 - \xi_n \quad \forall n$$

$$\text{subject to} \quad \xi_n \geq 0 \quad \forall n$$

- [Lagrangian Function]



$$L(w, b, \lambda_{n=1}^{N}, \xi_{n=1}^{N}, \mu_{n=1}^{M})$$

$$= C \cdot \sum_{n=1}^{N} \xi_n + \frac{1}{2}||w||^2 - \sum_{n=1}^{N}(\lambda_n \cdot t_n \cdot (w^t x_n + b) - \lambda_n + \lambda_n \xi_n) - \sum_{n=1}^{N} \mu_n \xi n$$

KKT conditions

$$\nabla_w L(w, b, \lambda_{n=1}^{N}, \xi_{n=1}^{N}, \mu_{n=1}^{M}) = \vec{w} - \sum_{n=1}^{N} (\lambda_n \cdot t_n \cdot \vec{x_n})$$

$$\nabla_b L(w, b, \lambda_{n=1}^{N}, \xi_{n=1}^{N}, \mu_{n=1}^{M}) = \sum_{n=1}^{N} \lambda_n \cdot t_n$$

$$\nabla_{\xi_n} L(w, b, \lambda_{n=1}^{N}, \xi_{n=1}^{N}, \mu_{n=1}^{M}) = C - \lambda_n - \mu_n$$

$$\lambda_n \cdot \{t_n(w^t x_n + b) - 1 + \xi_n\} = 0$$

$$\mu_n \xi n = 0$$

Soft Margin SVM
Primal & Dual Problems

- Primal

$$w*, b* = \arg\min_{w,b} C \cdot \sum_{n=1}^{N} \xi_n + \frac{1}{2}||w||^2$$

$$\text{subject to} \quad t_n(w^t x_n + b) \geq 1 - \xi_n \quad \forall n$$

$$\text{subject to} \quad \xi_n \geq 0 \quad \forall n$$

- Dual

$$\lambda*_{n=1}^{N} = \arg\max_{\lambda*} \sum_{n=1}^{N} \lambda *_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n^* \cdot t_n \lambda_m^* \cdot t_m \cdot \kappa(x_n, x_m)$$

$$\boxed{\text{subject to} \quad 0 \leq \lambda_n^* \leq C}$$

$$\sum_{n=1}^{N} \lambda_n^* \cdot t_n = 0$$