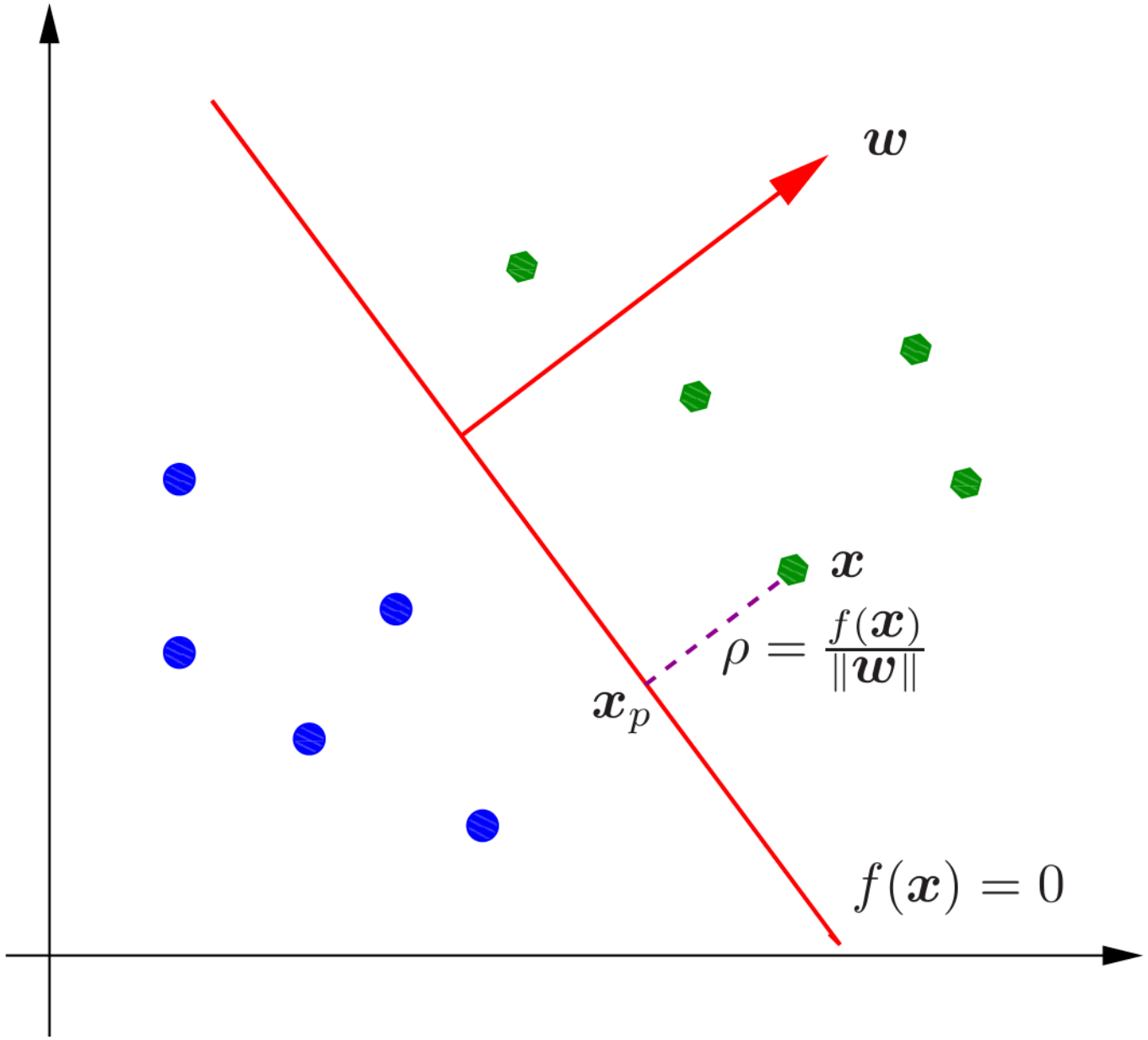
# Lecture 10
# Support Vector Machine (SVM)

- Separating Hyperplane and Binary Classification
- Margin and maximum margin classifier
- Support Vector
- Advanced topics
- Math foundations

# Separating Hyperplane: Geometric View



Consider $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$, where a decision boundary is given by the hyperplane $f(\boldsymbol{x}) = 0$.

1. The weight vector $\boldsymbol{w}$ is normal to the hyperplane.
2. $\rho = \frac{f(\boldsymbol{x})}{\|\boldsymbol{w}\|}$.

Why? $\Rightarrow$ See next slide!

1. Suppose $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two points lying on the hyperplane, leading to

$$\begin{aligned} \boldsymbol{w}^\top \boldsymbol{x}_1 + b &= 0, \\ \boldsymbol{w}^\top \boldsymbol{x}_2 + b &= 0. \end{aligned}$$

   It follows from these equations that we have $\boldsymbol{w}^\top(\boldsymbol{x}_1 - \boldsymbol{x}_2) = 0$ which implies that $\boldsymbol{w}$ is normal to the hyperplane.

2. Express $\boldsymbol{x} = \boldsymbol{x}_p + \rho\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$. Note that $f(\boldsymbol{x}_p) = 0$. Then we have

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{w}^\top \boldsymbol{x} + b \\ &= \boldsymbol{w}^\top \left( \boldsymbol{x}_p + \rho\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right) + b \\ &= \underbrace{\boldsymbol{w}^\top \boldsymbol{x}_p + b}_{0} + \rho\|\boldsymbol{w}\|. \end{aligned}$$

# Binary Classification

- Training sample: $\mathcal{D} = \{(\boldsymbol{x}_t, y_t)\}_{t=1}^{N}$.
  - Instances: $\boldsymbol{x}_t \in \mathbb{R}^m$.
  - Labels: $y_t \in \{+1, -1\}$.
- Prediction function: $h : \mathcal{X} \mapsto \{+1, -1\}$, $\quad h(\boldsymbol{x}) = \text{sign}\,(f(\boldsymbol{x}))$.
- Linear classification considers a linear discriminant function which has the form

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b,$$

where $(\boldsymbol{w}, b) \in \mathbb{R}^m \times \mathbb{R}$ (weight vector, bias) are the parameters that control the function.

# Margin

## Definition

The functional margin of an example $(\boldsymbol{x}_t, y_t)$ with respect to a hyperplane $\boldsymbol{w}$ is defined as

$$\begin{aligned} \rho_t &= y_t f(\boldsymbol{x}_t) \\ &= y_t(\boldsymbol{w}^\top \boldsymbol{x}_t + b). \end{aligned}$$

## Definition

The geometric margin is the functional margin for the normalized linear function $\frac{\boldsymbol{w}^\top}{\|\boldsymbol{w}\|}\boldsymbol{x} + \frac{b}{\|\boldsymbol{w}\|}$, i.e.,

$$\rho = y(f(\boldsymbol{x})/\|\boldsymbol{w}\|).$$

# Canonical Form

For $\lambda \neq 0$, $(\lambda \boldsymbol{w}, \lambda b)$ describes the same hyperplane as $(\boldsymbol{w}, b)$, i.e.,

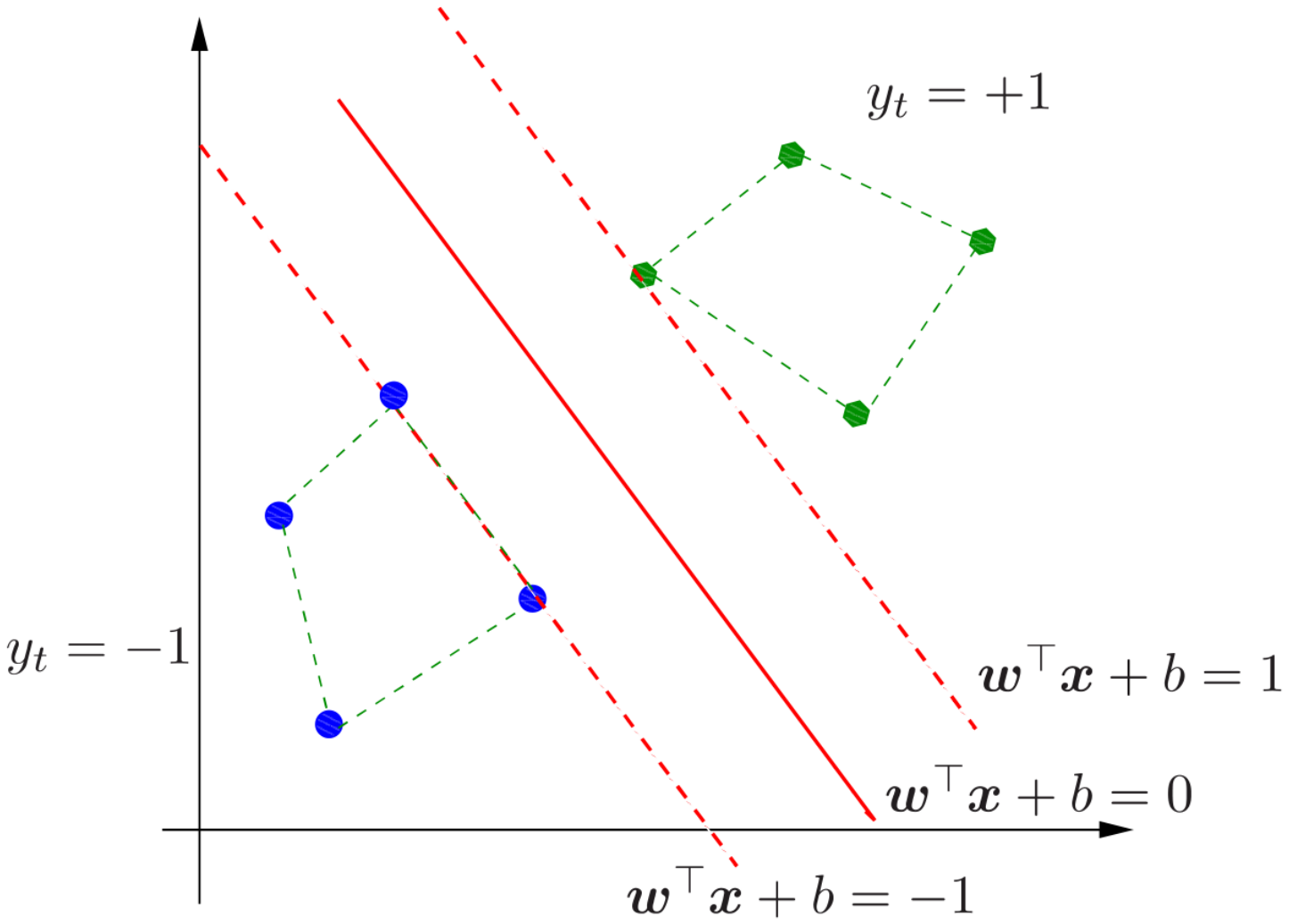
$$\{\boldsymbol{x} \,|\, \boldsymbol{w}^\top \boldsymbol{x} + b = 0\} = \{\boldsymbol{x} \,|\, \lambda \left(\boldsymbol{w}^\top \boldsymbol{x} + b\right) = 0\}.$$

## Definition

The hyperplane is in canonical form if

$$\min_{\boldsymbol{x}_t \in \mathcal{X}} \left|\boldsymbol{w}^\top \boldsymbol{x}_t + b\right| = 1.$$

# Canonical Optimal Hyperplane



# Margin $= \frac{1}{\|\boldsymbol{w}\|}$

The geometric margin in previous slide is given by

$$\begin{aligned}
\rho &= \frac{1}{2}\left\{\frac{f(\boldsymbol{x}^+)}{\|\boldsymbol{w}\|} - \frac{f(\boldsymbol{x}^-)}{\|\boldsymbol{w}\|}\right\} \\
&= \frac{1}{2}\frac{1}{\|\boldsymbol{w}\|}\left\{\underbrace{\boldsymbol{w}^\top \boldsymbol{x}^+ - \boldsymbol{w}^\top \boldsymbol{x}^-}_{2}\right\} \\
&= \frac{1}{\|\boldsymbol{w}\|}.
\end{aligned}$$

Thus, maximizing margin is equivalent to minimizing the norm of the weight vector in the discriminant function.

# Max Margin Classifier: Primal Form

## Primal Problem

$$\begin{aligned} \min_{\boldsymbol{w},b} \quad & \mathcal{J}(\boldsymbol{w}) = \tfrac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{subject to} \quad & y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1, \quad i = 1, \ldots, N \end{aligned}$$

## Proposition

*Given a linearly separable training sample* $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, *the hyperplane* $\boldsymbol{w}^\top \boldsymbol{x}_i + b = 0$ *that solves the above optimization problem realizes the maximal margin hyperplane with geometric margin* $\rho = 1/\|\boldsymbol{w}\|$.

# Primal Lagrangian

Primal Lagrangian $\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha})$ is given by

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{N} \alpha_i \left(1 - y_i \left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right)\right).$$

$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0$ and $\frac{\partial \mathcal{L}}{\partial b} = 0$ yield respectively $\boxed{\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i}$ and $\boxed{\sum_{i=1}^{N} \alpha_i y_i = 0}$. Substitute these relations into the primal form, leading to

$$\mathcal{L} = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j + \sum_{i=1}^{N} \alpha_i.$$

# Max Margin Classifier: Dual Form

## Dual Problem

$$\max_{\boldsymbol{\alpha}} \quad \mathcal{G}(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j + \sum_{i=1}^{N}\alpha_i,$$

$$\text{subject to} \quad \sum_{i=1}^{N}\alpha_i y_i = 0,$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, N.$$

## Proposition

*Suppose that $\boldsymbol{\alpha}^*$ solves the dual problem, given a linearly separable training sample $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$. Then the weight vector $\boldsymbol{w}^* = \sum_{i=1}^N \alpha_i^* y_i \boldsymbol{x}_i$ realizes the maximal margin hyperplane with geometric margin $\rho = 1/\|\boldsymbol{w}\|$.*

# Support Vectors

Note that $\boxed{\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i}$, i.e., $\boldsymbol{w}$ is a linear combination of training data points $\boldsymbol{x}_i$.

It follows from KKT complimentary slackness condition that we have

$$\alpha_i \left[1 - y_i \left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right)\right] = 0.$$

- $y_i \left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \neq 1$ ($x_i$ is not support vector) $\Rightarrow \alpha_i = 0$ ($x_i$ is irrelevant).
- $\alpha_i \neq 0 \Rightarrow y_i \left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) = 1$ ($x_i$ is support vector).

Only support vectors influence the computation of $\boldsymbol{w}$.

# Soft Margin Classifier: $L_2$ Norm

In cases where data are not linearly separable, the optimization problem cannot be solved as the primal has an empty feasible region and the dual an unbounded objective function.
To sidestep this problem, we introduce slack variables $\zeta_i \geq 0$ to allow the margin constraints to be violated:

$$y_i \left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1 - \zeta_i, \quad i = 1, \ldots, N.$$

Primal Problem

$$\begin{aligned} \min_{\boldsymbol{w}, b, \boldsymbol{\zeta}} \quad & \mathcal{J}(\boldsymbol{w}, \boldsymbol{\zeta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{N} \zeta_i^2, \\ \text{subject to} \quad & y_i \left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1 - \zeta_i, \quad i = 1, \ldots, N. \end{aligned}$$

- The primal Lagrangian is

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\zeta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{2}\sum_{i=1}^{N}\zeta_i^2 + \sum_{i=1}^{N}\alpha_i\left(1 - \zeta_i - y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right)\right).$$

- $\frac{\partial \mathcal{J}(\boldsymbol{w},\boldsymbol{\zeta})}{\partial \boldsymbol{w}} = 0$ leads to $\boldsymbol{w} = \sum_{i=1}^{N}\alpha_i y_i \boldsymbol{x}_i$.
- $\frac{\partial \mathcal{J}(\boldsymbol{w},\boldsymbol{\zeta})}{\partial b} = 0$ leads to $\sum_{i=1}^{N}\alpha_i y_i = 0$.
- $\frac{\partial \mathcal{J}(\boldsymbol{w},\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} = 0$ leads to $\boldsymbol{\zeta} = \frac{\boldsymbol{\alpha}}{C}$.
- Substitute these relations into the primal Lagrangian to obtain the dual Lagrangian function:

$$\mathcal{G}(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j - \frac{1}{2C}\sum_{i=1}^{N}\alpha_i^2 + \sum_{i=1}^{N}\alpha_i.$$

# Soft Margin Classifier: Dual Form

## Dual Problem

$$\max_{\boldsymbol{\alpha}} \quad \mathcal{G}(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \left(\boldsymbol{x}_i^\top \boldsymbol{x}_j + \frac{1}{C}\delta_{ij}\right) + \sum_{i=1}^{N}\alpha_i,$$

$$\text{subject to} \quad \sum_{i=1}^{N}\alpha_i y_i = 0,$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, N.$$

# Soft Margin Classifier: $L_1$ Norm

## Primal Problem

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\zeta}} \quad & \mathcal{J}(\boldsymbol{w},\boldsymbol{\zeta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\zeta_i, \\ \text{subject to} \quad & y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1-\zeta_i, \quad i=1,\ldots,N, \\ & \zeta_i \geq 0, \quad i=1,\ldots,N. \end{aligned}$$

The primal Lagrangian is

$$\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\zeta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{2}\sum_{i=1}^{N}\zeta_i^2 + \sum_{i=1}^{N}\alpha_i\left(1-\zeta_i - y_i\left(\boldsymbol{w}^\top\boldsymbol{x}_i + b\right)\right) - \sum_{i=1}^{N}\beta_i\zeta_i.$$

# Soft Margin Classifier: Dual Form

## Dual Problem

$$\max_{\boldsymbol{\alpha}} \quad \mathcal{G}(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j + \sum_{i=1}^{N}\alpha_i,$$

$$\text{subject to} \quad \sum_{i=1}^{N}\alpha_i y_i = 0,$$

$$0 \leq \alpha_i \leq C, \quad i = 1, \ldots, N.$$

# Regularization View

Consider a linear discriminant function

$$f(\boldsymbol{x}_i) = \boldsymbol{w}^\top \boldsymbol{x}_i + b,$$

where labels are $y_i \in \{1, -1\}$

Regularized learning problem involves:

$$\min_f \underbrace{\frac{1}{N}\sum_{i=1}^{N} V\left(y_i, f(\boldsymbol{x}_i)\right)}_{\text{empirical risk}} + \lambda \|\boldsymbol{w}\|^2,$$

where $V(y_i, f(\boldsymbol{x}_i))$ is a loss function.

# Why Regularization

- The problem of approximating a function from sparse data is ill-posed and a classical way to solve it is regularization.
- For a finite set of training examples the search for the best model or approximating function has to be constrained on an approximately small hypothesis space (which can be thought of as a space of model).
- If the space is too large, models can be found which will fit exactly the data but will have a poor generalization performance , that is poor predictive capabiility on new data.
- Vapnik's theory characterize these concepts in terms of capacity of a set of functions and capacity control depending on the training data.
- In the case of regularization, a form of capacity control leads to choosing an optimal regularization parameter $\lambda$.

# Loss Functions

- $L_2$ Regularization Networks

$$V(y_i, f(\boldsymbol{x}_i)) = (y_i - f(\boldsymbol{x}_i))^2.$$

- Support Vector Regression

$$V(y_i, f(\boldsymbol{x}_i)) = |y_i - f(\boldsymbol{x}_i)|_\epsilon,$$

where $|\cdot|_\epsilon$ is epsilon-insensitive norm.

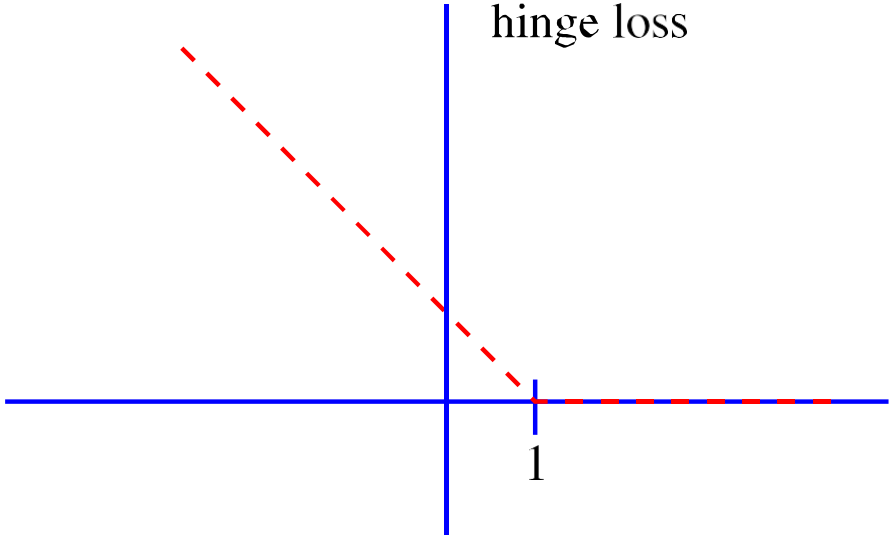
- Support Vector Machine

$$V(y_i, f(\boldsymbol{x}_i)) = [1 - yf(\boldsymbol{x})]_+,$$

where $[1 - yf(\boldsymbol{x})]_+$ is hinge loss (see the next slide).

SVM arises by considering hinge loss in the regularization framework:

$$V(y, f(\boldsymbol{x})) = [1 - yf(\boldsymbol{x})]_+ = \max(1 - yf(\boldsymbol{x}), 0).$$



The problem is not differentiable, so introduce slack variables $\xi_i$ to make the problem easier to work with

$$\begin{aligned} \min_{f} \quad & \frac{1}{N}\xi_i + \lambda\|\boldsymbol{w}\|^2, \\ \text{subject to} \quad & y_i f(\boldsymbol{x}_i) \geq 1 - \xi_i, \quad i = 1, \ldots, N, \\ & \xi_i \geq 0, \end{aligned}$$

which is equivalent to soft-margin SVM.

# Regularization in RKHS

Regularized learning problem involves:

$$\min_{f \in \mathcal{H}} \underbrace{\frac{1}{N} \sum_{i=1}^{N} V\left(y_i, f(\boldsymbol{x}_i)\right)}_{\text{empirical risk}} + \lambda \|f\|_K^2,$$

where $\|f\|_K^2$ is a norm in a Reproducing Kernel Hilbert Space $\mathcal{H}$ define by the positive definite function $\kappa$.
Introducing slack variables $\xi_i$,

$$\begin{aligned} \min_{f} \quad & \frac{1}{N}\xi_i + \lambda \|f\|_K^2, \\ \text{subject to} \quad & y_i f(\boldsymbol{x}_i) \geq 1 - \xi_i, \quad i = 1, \ldots, N, \\ & \xi_i \geq 0. \end{aligned}$$

Apply the representer theorem

$$f^*(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \kappa(\boldsymbol{x}, \boldsymbol{x}_i),$$

leading to

$$\|f\|_K^2 = \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha},$$

where $\boldsymbol{K} = [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)] \in \mathbb{R}^{N \times N}$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^\top \in \mathbb{R}^N$.
Then we have the primal form for SVM

$$\begin{aligned} \min_{f \in \mathcal{H}} \quad & \frac{1}{N} \xi_i + \lambda \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha}, \\ \text{subject to} \quad & y_i \left( \sum_{j=1}^{N} \alpha_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) + b \right) \geq 1 - \xi_i, \quad i = 1, \ldots, N, \\ & \xi_i \geq 0, \end{aligned}$$

where an unregularized bias term $b$ is added.

# Quick Overview of Constrained Optimization

# Constrained Optimization

Consider a primal form of constrained optimization problem:

$$\begin{aligned} \text{minimize} \quad & \mathcal{J}(\boldsymbol{w}), \\ \text{subject to} \quad & g_i(\boldsymbol{w}) \leq 0, \quad i = 1, \ldots, M, \\ & h_j(\boldsymbol{w}) = 0, \quad j = 1, \ldots, L. \end{aligned}$$

Lagrangian is given by

$$\boxed{\mathcal{L}(\boldsymbol{w}, \boldsymbol{\nu}, \boldsymbol{\lambda}) = \mathcal{J}(\boldsymbol{w}) + \sum_{i=1}^{M} \nu_i g_i(\boldsymbol{w}) + \sum_{j=1}^{L} \lambda_j h_j(\boldsymbol{w})},$$

where $\nu_i > 0$ for $i = 1, \ldots, M$ and $\lambda_j$ for $j = 1, \ldots, L$ are unrestricted in sign.

# Karush-Kuhn-Tucker (KKT) Necessary Conditions

1. Optimality

$$\nabla \mathcal{L} = \nabla \mathcal{J}(\boldsymbol{w}) + \sum_{i=1}^{M} \nu_i \nabla g_i(\boldsymbol{w}) + \sum_{j=1}^{L} \lambda_j \nabla h_j(\boldsymbol{w}) = 0.$$

2. Feasibility

$$\begin{aligned} g_i(\boldsymbol{w}) &\leq 0, \quad i = 1, \ldots, M \\ h_j(\boldsymbol{w}) &= 0, \quad j = 1, \ldots, L. \end{aligned}$$

3. Complementary slackness

$$\nu_i g_i(\boldsymbol{w}) = 0, \quad i = 1, \ldots, M \quad (\nu_i > 0).$$

# Lagrangian Dual Problem

Lagrangian dual function is defined as the minimum value of the Lagrangian over $\boldsymbol{w}$:

$$\mathcal{G}(\boldsymbol{\nu}, \boldsymbol{\lambda}) = \inf_{\boldsymbol{w} \in \mathcal{W}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\nu}, \boldsymbol{\lambda}).$$

The Lagrangian dual problem has the form:

$$\begin{array}{ll} \text{maximize} & \mathcal{G}(\boldsymbol{\nu}, \boldsymbol{\lambda}), \\ \text{subject to} & \nu_i \geq 0, \quad i = 1, \ldots, M. \end{array}$$

Weak duality: $\mathcal{J}(\boldsymbol{w}) \geq \mathcal{G}(\boldsymbol{\nu}, \boldsymbol{\lambda})$ if $\boldsymbol{w}$ is a feasible solution of the primal problem and $(\boldsymbol{\nu}, \boldsymbol{\lambda})$ is a feasible solution of the dual problem.