Section 8
Convex Optimisation 2

Lagrangian

Consider a general optimization problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$
$$\text{subject to } h_i(\boldsymbol{x}) \leq 0,\ i = 1, \cdots, m,$$
$$\ell_j(\boldsymbol{x}) = 0,\ j = 1, \cdots, r.$$

The objective function f needs not to be convex. Of course we pay special attention to the convex case.

Definition 8.1 (Lagrangian)

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} u_i h_i(\boldsymbol{x}) + \sum_{j=1}^{r} v_j \ell_j(\boldsymbol{x}).$$

Here $\boldsymbol{u} \in \mathbb{R}^m$, $\boldsymbol{v} \in \mathbb{R}^r$, and $\boldsymbol{u} \geq \mathbf{0}$.

Lagrange Dual Function

unconstrained opt. problem respect to x.

Definition 8.2 (Lagrange Dual Function)

$$g(\boldsymbol{u}, \boldsymbol{v}) := \min_{\boldsymbol{x} \in \mathbb{R}^n} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = \min_{\boldsymbol{x} \in \mathbb{R}^n} \left[f(\boldsymbol{x}) + \sum_{i=1}^m u_i h_i(\boldsymbol{x}) + \sum_{j=1}^r v_j \ell_j(\boldsymbol{x}) \right].$$

- For every feasible $\boldsymbol{x}$ ($\boldsymbol{x} \in \mathcal{X}$), $L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \leq f(\boldsymbol{x})$

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{x}) + \underbrace{\sum_{i=1}^m u_i h_i(\boldsymbol{x})}_{\leq 0} + \underbrace{\sum_{j=1}^r v_j \ell_j(\boldsymbol{x})}_{=0}.$$

- Let $\mathcal{X}$ denote the primal feasible set.

$$\boxed{g(\boldsymbol{u}, \boldsymbol{v}) = \min_{\boldsymbol{x} \in \mathbb{R}^n} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \leq \min_{\boldsymbol{x} \in \mathcal{X}} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \leq f(\boldsymbol{x}).} \quad (13)$$

Concavity of Lagrange Dual Function

Lemma 8.3

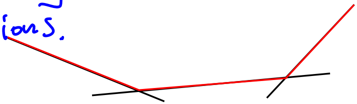
The Lagrange dual function

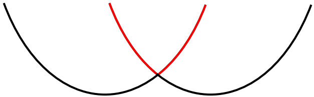
$$g(\boldsymbol{u},\boldsymbol{v}) = \min_{\boldsymbol{x}\in\mathbb{R}^n} L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) = \min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) + \sum_{i=1}^{m} u_i h_i(\boldsymbol{x}) + \sum_{j=1}^{r} v_j \ell_j(\boldsymbol{x})$$

is concave in $(\boldsymbol{u},\boldsymbol{v})$.

Lemma 8.4

- *Let* $f_\alpha(\boldsymbol{x})$ *be concave functions. Then* $g(\boldsymbol{x}) = \inf_\alpha\ f_\alpha(\boldsymbol{x})$ *is concave.*
- *Let* $\underbrace{f_\alpha(\boldsymbol{x})}_{\text{many functions.}}$ *be convex functions. Then* $g(\boldsymbol{x}) = \sup_\alpha\ f_\alpha(\boldsymbol{x})$ *is convex.*





Proofs

Proof of Lemma 8.4: For any $\lambda \in [0, 1]$,

assume f_α is concave.

$$
\begin{aligned}
g(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) &= \inf_\alpha \; f_\alpha(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \\
&\geq \inf_\alpha \; \lambda f_\alpha(\boldsymbol{x}) + (1-\lambda) f_\alpha(\boldsymbol{y}) \\
&\geq \lambda \inf_\alpha \; f_\alpha(\boldsymbol{x}) + (1-\lambda) \inf_\alpha \; f_\alpha(\boldsymbol{y}).
\end{aligned}
$$

Proof of Lemma 8.3: For any given $\boldsymbol{x}$, $L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$ is linear in $(\boldsymbol{u}, \boldsymbol{v})$, and hence concave in $(\boldsymbol{u}, \boldsymbol{v})$. The minimum of concave functions is concave based on Lemma 8.4.

Lagrange Dual Problem

Given the primal problem

$$\begin{aligned} \min_{\boldsymbol{x}} \ & f(\boldsymbol{x}) \\ \text{subject to } & h_i(\boldsymbol{x}) \le 0,\ i = 1, \cdots, m, \\ & \ell_j(\boldsymbol{x}) = 0,\ j = 1, \cdots, r. \end{aligned}$$

Its Lagrange dual problem is

$$\boxed{\max_{\boldsymbol{u} \in \mathbb{R}^m,\ \boldsymbol{v} \in \mathbb{R}^r} g(\boldsymbol{u}, \boldsymbol{v}),\ \text{subject to } \boldsymbol{u} \ge \boldsymbol{0}.}$$

Concave function.

Weak and Strong Duality

Weak duality: the dual optimal value $g^\star$ satisfies

$$f^\star \geq g^\star.$$

This is a direct consequence of (13).
Strong duality is referred to as the case that

$$\boxed{f^\star = g^\star.}$$

the optinum value of the two problems are equal

Slater's condition: if the primal is a convex problem (i.e., f and g_i's are convex and ℓ_j's are affine), and there exists at least one strictly feasible $\boldsymbol{x} \in \mathbb{R}^n$ satisfying

$$h_i(\boldsymbol{x}) < 0,\ \forall i \in [m],\ \text{and}\ \ell_j(\boldsymbol{x}) = 0,\ \forall j \in [r],$$

then strong duality holds. (Proof is omitted.)

Karush-Kuhn-Tucker conditions

Given the optimization problem

convex functions.

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}) \\ &\text{subject to } h_i(\boldsymbol{x}) \le 0,\ i = 1, \cdots, m, \\ &\qquad\qquad \ell_j(\boldsymbol{x}) = 0,\ i = 1, \cdots, r. \end{aligned}$$

the optimum point satisfies the KKT conditions.

The Karush-Kuhn-Tucker (KKT) conditions are:

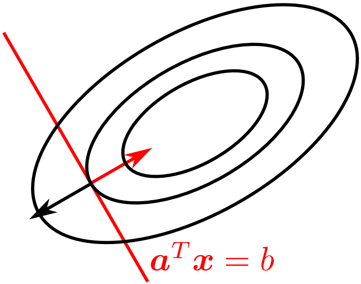
- $\mathbf{0} \in \partial f(\boldsymbol{x}) + \sum_{i=1}^{m} u_i \partial h_i(\boldsymbol{x}) + \sum_{j=1}^{r} v_j \partial \ell_j(\boldsymbol{x})$. (stationarity)
- $u_i h_i(\boldsymbol{x}) = 0$, $\forall i$. ← most useful!!! (complementary slackness)
- $h_i(\boldsymbol{x}) \le 0$, $\ell_j(\boldsymbol{x}) = 0$, $\forall i$, $\forall j$. ← constraint. (primal feasibility)
- $u_i \ge 0$, $\forall i$. (dual feasibility)

← Lagrange multipliers

KKT conditions are

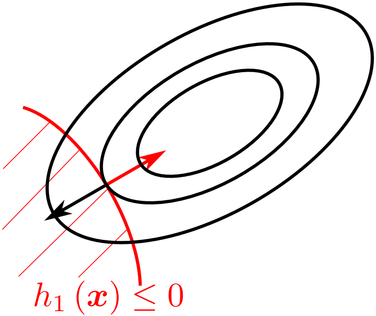
- Always sufficient.
- Necessary under strong duality.

Geometric Intuition: Equality Constraints



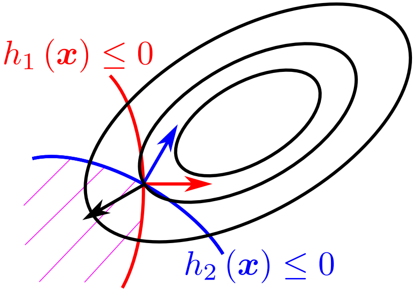
$\partial f(\boldsymbol{x})$ is a linear combination of $\partial \ell_j(\boldsymbol{x})$'s.

Geometric Intuition: One Inequality Constraint



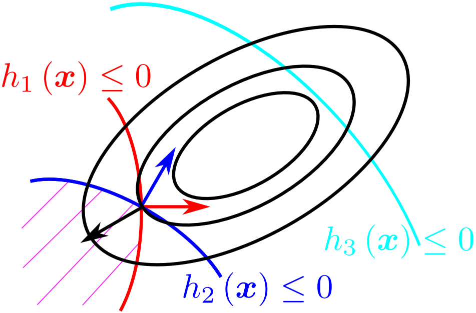
$$\partial f(\boldsymbol{x}) + u_1 \partial h_1(\boldsymbol{x}) = \mathbf{0}.$$
$$h_1(\boldsymbol{x}) = 0.$$

Geometric Intuition: Inequality Constraints



$$\partial f(\boldsymbol{x}) + \sum_{i=1}^{2} u_i \partial h_i(\boldsymbol{x}) = \mathbf{0}.$$
$$h_1(\boldsymbol{x}) = 0,\ h_2(\boldsymbol{x}) = 0.$$

Geometric Intuition: Inequality Constraints



$$\partial f(\boldsymbol{x}) + \sum_{i=1}^{3} u_i \partial h_i(\boldsymbol{x}) = \mathbf{0}.$$

$$h_1(\boldsymbol{x}) = 0,\ h_2(\boldsymbol{x}) = 0,$$

$h_3(\boldsymbol{x}) < 0$ but $u_3 = 0$ so that $u_3 h_3(\boldsymbol{x}) = 0$. !!!

Sufficiency

If $\boldsymbol{x}^\star, \boldsymbol{u}^\star, \boldsymbol{v}^\star$ satisfy the KKT conditions, then $\boldsymbol{x}^\star$ and $\boldsymbol{u}^\star, \boldsymbol{v}^\star$ are primal and dual solutions.

If $\boldsymbol{x}^\star, \boldsymbol{u}^\star, \boldsymbol{v}^\star$ satisfy the KKT conditions, then

$$\begin{aligned} g\left(\boldsymbol{u}^\star, \boldsymbol{v}^\star\right) &= f\left(\boldsymbol{x}^\star\right) + \sum_{i=1}^{m} u_i^\star h_i\left(\boldsymbol{x}^\star\right) + \sum_{j=1}^{r} v_i^\star \ell_i\left(\boldsymbol{x}^\star\right) \\ &= f\left(\boldsymbol{x}^\star\right), \end{aligned}$$

where the first equality follows from stationarity, and the second follows from complementary slackness. This equality suggests the duality gap is zero. Hence, $\boldsymbol{x}^\star$, $\boldsymbol{u}^\star$and $\boldsymbol{v}^\star$ are primal and dual optimal.

Necessity

Suppose that the strong duality holds and that $\boldsymbol{x}^\star$ and $\boldsymbol{u}^\star$, $\boldsymbol{v}^\star$ are primal and dual solutions. Then $\boldsymbol{x}^\star$, $\boldsymbol{u}^\star$, $\boldsymbol{v}^\star$ satisfy the KKT conditions.

Due to the strong duality, one has

$$
\begin{aligned}
f\left(\boldsymbol{x}^\star\right) &= g\left(\boldsymbol{u}^\star, \boldsymbol{v}^\star\right) \\
&= \min_{\boldsymbol{x}\in\mathbb{R}^n}\ f\left(\boldsymbol{x}\right) + \sum_{i=1}^{m} u_i^\star h_i\left(\boldsymbol{x}\right) + \sum_{j=1}^{r} v_j^\star \ell_j\left(\boldsymbol{x}\right) \\
&\le f\left(\boldsymbol{x}^\star\right) + \sum_{i=1}^{m} u_i^\star h_i\left(\boldsymbol{x}^\star\right) + \sum_{j=1}^{r} v_j^\star \ell_j\left(\boldsymbol{x}^\star\right) \\
&\le f\left(\boldsymbol{x}^\star\right).
\end{aligned}
$$

In other words, all the inequalities are actually equalities.

Quadratic Programming with Equality Constraints

Let $\boldsymbol{Q} \succeq 0$.

$$\min_{\boldsymbol{x}} \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{x} \text{ subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}.$$

By KKT conditions, $\boldsymbol{x}$ is the minimizer if and only if

$$\begin{bmatrix} \boldsymbol{Q} & \boldsymbol{A}^T \\ \boldsymbol{A} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{c} \\ \boldsymbol{0} \end{bmatrix},$$

where the first set of linear equations come from the stationarity and the second set follows from the primal feasibility.

The optimal $\boldsymbol{x}^\star$ can be obtained by solving the linear inverse problem.

Water Filling

channel capacity of the channel.

$$\min_{\boldsymbol{x}} \; -\sum_{i=1}^{n} \log\left(\alpha_i + x_i\right) \text{ subject to } \boldsymbol{x} \geq \boldsymbol{0},\ \boldsymbol{1}^T \boldsymbol{x} = 1.$$

$L = -\sum_i \log(\alpha_i + x_i) + \sum_i u_i(-x_i) + v(1^T x - 1)$.

By KKT conditions,

- $-1/\left(\alpha_i + x_i\right) - u_i + v = 0$, $\forall i$
- $u_i x_i = 0$, $\forall i$
- $\boldsymbol{x} \geq 0$, $\boldsymbol{1}^T \boldsymbol{x} = 1$, $\boldsymbol{u} \geq 0$.

Eliminate $\boldsymbol{u}$. The first two conditions become

$$1/\left(\alpha_i + x_i\right) \leq v, \text{ and } x_i\left(v - 1/\left(\alpha_i + x_i\right)\right) = 0, \ \forall i.$$

Therefore, the solution:

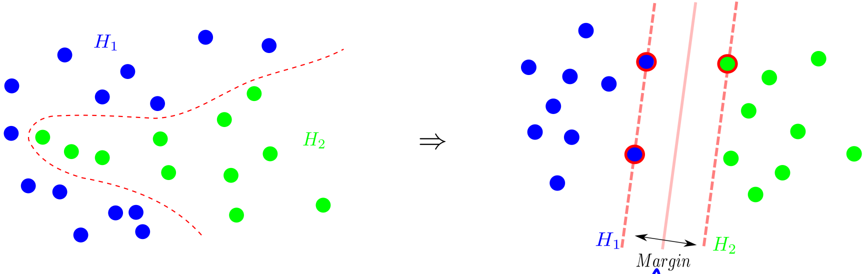
$$\boxed{x_i = \max\left(0, 1/v - \alpha_i\right)}$$

where v is chosen such that

$$\sum_{i=1}^{n} \max\left(0, 1/v - \alpha_i\right) = 1.$$

Section 9
Support Vector Machine

Idea of SVM



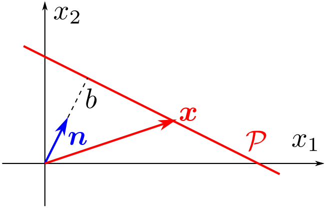
The hyperplane may not be unique.
We pick the one with the largest margin.

A Hyperplane

A hyperplane in $\mathbb{R}^n$ can be defined using its normal vector $\boldsymbol{n} \in \mathbb{R}^n$:

$$\boxed{\mathcal{P} = \left\{ \boldsymbol{x} : \quad \boldsymbol{n}^T \boldsymbol{x} = b \right\}.}$$

- Usually we assume $\|\boldsymbol{n}\|_2 = 1$.



  The projection $\|\mathrm{Proj}\,(\boldsymbol{x}, \mathrm{span}\,(\boldsymbol{n}))\|_2 = b$.

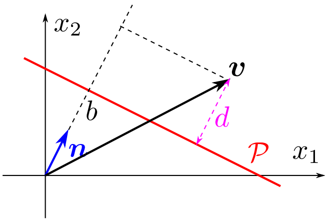
- If $\|\boldsymbol{n}\|_2 \neq 1$, then

$$\mathcal{P} = \left\{ \boldsymbol{x} : \quad \boldsymbol{n}^T \boldsymbol{x} = b \right\} = \left\{ \boldsymbol{x} : \quad \boldsymbol{n}^T \boldsymbol{x} / \|\boldsymbol{n}\|_2 = b / \|\boldsymbol{n}\|_2 \right\}.$$

Distance to a Hyperplane

Define a hyperplane $\mathcal{P} = \left\{ \boldsymbol{x} : \quad \boldsymbol{n}^T \boldsymbol{x} = b \right\}$ where $\|\boldsymbol{n}\|_2 = 1$.
Let $\boldsymbol{v}$ be an arbitrary point.
The distance between $\boldsymbol{v}$ and $\mathcal{P}$ is given by

$$d = d\left(\boldsymbol{v}, \mathcal{P}\right) = \left| \boldsymbol{n}^T \boldsymbol{v} - b \right|. \tag{14}$$



When $\|\boldsymbol{n}\|_2 \neq 1$,

$$d = \left| \frac{\boldsymbol{n}^T}{\|\boldsymbol{n}\|_2} \boldsymbol{v} - \frac{b}{\|\boldsymbol{n}\|_2} \right| = \frac{\left| \boldsymbol{n}^T \boldsymbol{v} - b \right|}{\|\boldsymbol{n}\|_2}. \tag{15}$$

minimize this

SVM: Separate Points from Two Different Classes

Given training dataset $\{\boldsymbol{x}_i, y_i\}$ where the labels $y_i \in \{-1, 1\}$, want to find $\boldsymbol{\beta}$ and b s.t.

$$\boldsymbol{\beta}^T \boldsymbol{x}_i + b \geq +1 \quad \text{for } y_i = +1,$$
$$\boldsymbol{\beta}^T \boldsymbol{x}_i + b \leq -1 \quad \text{for } y_i = -1.$$

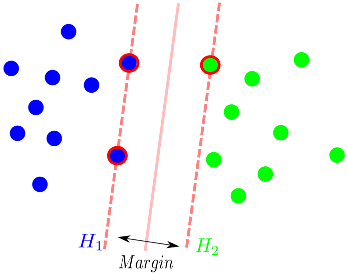
or equivalently

$$y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b\right) - 1 \geq 0, \quad \forall i.$$

In other words, find a hyperplane $\left\{\boldsymbol{x} : \boldsymbol{\beta}^T \boldsymbol{x} - b\right\}$ s.t.

- Distance from one class to the hyperplane is $1/\left\|\boldsymbol{\beta}\right\|_2$.
- Distance between the two classes (along direction $\boldsymbol{\beta}$) is $2/\left\|\boldsymbol{\beta}\right\|_2$.

SVM: Best Separation



SVM: a convex optimization problem:

$$\min_{\boldsymbol{\beta},b} \quad \frac{1}{2}\|\boldsymbol{\beta}\|_2^2,$$
$$\text{subject to} \quad 1 - y_i\left(\boldsymbol{\beta}^T\boldsymbol{x}_i + b\right) \le 0. \tag{16}$$

Lagrange Dual Problem of SVM

Lagrangian of the SVM primal optimization problem:

$$L = \frac{1}{2} \|\boldsymbol{\beta}\|^2 + \sum_i \lambda_i \left(1 - y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b\right)\right), \tag{17}$$

where $\lambda_i \geq 0$.

Lagrange Dual Problem

$$\max_{\boldsymbol{\lambda}} \quad \underbrace{\min_{\boldsymbol{\beta}, b} \; L}_{\text{Lagrange dual function } L_D}$$

The Dual Function

To solve $\min_{\boldsymbol{\beta},b} L$, set $\partial L/\partial \boldsymbol{\beta}$ and $\partial L/\partial b$ to zero:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \boldsymbol{\beta} - \sum_i \lambda_i y_i \boldsymbol{x}_i = 0 \;\Rightarrow\; \boldsymbol{\beta} = \sum_i \lambda_i y_i \boldsymbol{x}_i. \tag{18}$$

$$\frac{\partial L}{\partial b} = \sum_i \lambda_i y_i = 0. \tag{19}$$

Substitute (18) and (19) into the Lagrangian (17). It holds that

lagrange dual function.

$$\begin{aligned} L_D &= \sum \lambda_i - \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 = \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \\ &= -\frac{1}{2}\boldsymbol{\lambda}^T \boldsymbol{K} \boldsymbol{\lambda} + \mathbf{1}^T \boldsymbol{\lambda}, \end{aligned} \tag{20}$$

where $K_{i,j} = y_i \boldsymbol{x}_i^T \boldsymbol{x}_j y_j$.

is known.

The Dual Problem

The dual problem becomes:

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & -\frac{1}{2}\boldsymbol{\lambda}^T \boldsymbol{K} \boldsymbol{\lambda} + \mathbf{1}^T \boldsymbol{\lambda}, \\ \text{subject to} \quad & \lambda_i \geq 0, \quad \forall i, \\ & \sum_i \lambda_i y_i = 0. \end{aligned} \tag{21}$$

The KKT Condition

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \boldsymbol{\beta} - \sum_i \lambda_i y_i \boldsymbol{x}_i = 0, \tag{22}$$

$$\frac{\partial L}{\partial b} = \sum_i \lambda_i y_i = 0, \tag{23}$$

$$1 - y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b\right) \leq 0, \tag{24}$$

$$\lambda_i \geq 0, \tag{25}$$

$$\underline{\lambda_i \left(1 - y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b\right)\right) = 0.} \tag{26}$$

second bullet
of KKT conditions.

SVM Classifier: Support Vectors

If I can derive this slide I will have understood the concept.

Condition (26) implies

$$\begin{cases} \text{if } \lambda_i \neq 0 & \text{then } 1 = y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b\right), \\ \text{if } 1 \neq y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b\right) & \text{then } \lambda_i = 0. \end{cases}$$

Hence from (22),

$$\boldsymbol{\beta} = \sum_{i \in \mathcal{I}} \lambda_i y_i \boldsymbol{x}_i, \quad \mathcal{I} = \left\{ i : \ y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b\right) = 1 \quad (\text{or } \lambda_i \neq 0) \right\}.$$

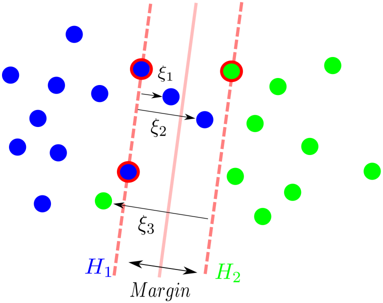
$\lambda_i \neq 0$ only for boundary points.

For a new test data $\boldsymbol{x}^{\text{new}}$,

$$y^{\text{new}} = \text{sign}\left(\sum_{i \in \mathcal{I}} \lambda_i y_i \boldsymbol{x}_i^T \boldsymbol{x}^{\text{new}} + b \right).$$

The classifier only uses the boundary points (sparsity!).

SVM for Overlapping Classes



Primal Problem for Overlapping Classes

The constraints:

$$\boldsymbol{\beta}^T \boldsymbol{x}_i + b \geq +1 - \xi_i \quad \text{for } y_i = +1,$$
$$\boldsymbol{\beta}^T \boldsymbol{x}_i + b \leq -1 + \xi_i \quad \text{for } y_i = -1,$$

where $\xi_i \geq 0$, $\forall i$. for more data points $\xi_i = 0$.

SVM Primal Problem:

$$\begin{aligned} \min_{\boldsymbol{\beta}, b, \boldsymbol{\xi}} \quad & \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 + C \left(\sum_i \xi_i \right)^k \\ \text{subject to} \quad & 1 - \xi_i - y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b \right) \leq 0, \\ & -\xi_i \leq 0, \quad \forall i, \end{aligned}$$

where $C > 0$ is a constant and k is a positive integer. Usually $k = 1$.

Dual Function

The Lagrangian

$$L = \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 + C \sum \xi_i + \sum \lambda_i \left(1 - \xi_i - y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i - b\right)\right) - \sum u_i \xi_i,$$

where $\lambda_i \geq 0, \; u_i \geq 0$ are Lagrange multipliers.

The dual function

$$L_D = \min_{\boldsymbol{\beta}, b, \boldsymbol{\xi}} \; L.$$

To find the dual function

$$\frac{dL}{d\boldsymbol{\beta}} = 0 \quad \Rightarrow \quad \boldsymbol{\beta} = \sum \lambda_i y_i \boldsymbol{x}_i.$$

$$\frac{dL}{db} = 0 \quad \Rightarrow \quad \sum \lambda_i y_i = 0.$$

$$\frac{dL}{d\boldsymbol{\xi}} = 0 \quad \Rightarrow \quad C - \lambda_i - u_i = 0 \quad \Rightarrow \quad \lambda_i = C - u_i \leq C.$$

The Dual Problem

The dual problem:

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & \sum \lambda_i - \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 = -\frac{1}{2} \boldsymbol{\lambda}^T \boldsymbol{K} \boldsymbol{\lambda} + \mathbf{1}^T \boldsymbol{\lambda} \\ \text{subject to} \quad & 0 \le \lambda_i \le C, \\ & \sum \lambda_i y_i = 0, \end{aligned}$$

where $K_{i,j} = y_i \boldsymbol{x}_i^T \boldsymbol{x}_j y_j$.
The only difference is that now λ_i's are upper bounded by C.

Again, only boundary points are involved.

$$\boldsymbol{\beta} = \sum_{i \in \mathcal{I}} \lambda_i y_i \boldsymbol{x}_i, \quad \mathcal{I} = \{i : \quad \lambda_i \neq 0\},$$

which comes from the KKT condition $\lambda_i \left(1 - \xi_i - y_i \left(\boldsymbol{\beta}^T \boldsymbol{x}_i + b\right)\right) = 0$.

The General Case

- Two classes $\Rightarrow$ multiple classes
  - Regression
- Data space $\Rightarrow$ feature space
  Define a kernel function $\varphi: \mathbb{R}^n \rightarrow \mathcal{H}$ and work on the space of $\varphi(\boldsymbol{x}_i)$.

In SVM, what really matters is $\boldsymbol{x}_i^T \boldsymbol{x}_j$.
In the general case (kernel method), what matters is

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi^T(\boldsymbol{x}_i)\,\varphi(\boldsymbol{x}_j).$$

Example of nonlinear features:

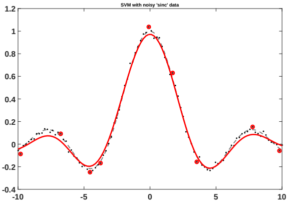
- $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \left(\boldsymbol{x}^T \boldsymbol{y}\right)^2 = (x_1 y_1 + x_2 y_2)^2 = x_1^2 y_2^2 + 2 x_1 y_1 x_2 y_2 + x_2^2 y_2^2$.
  - $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \varphi^T(\boldsymbol{x})\,\varphi(\boldsymbol{y})$ with $\varphi(\boldsymbol{x}) = \left[x_1^2, \sqrt{2} x_1 x_2, x_2^2\right]^T$.
- $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 / 2\sigma^2\right)$. σ small $\Rightarrow$ watch the nearest neighbours.
  - $\varphi(\boldsymbol{x})$ has infinite dimension.

SVM for Regression

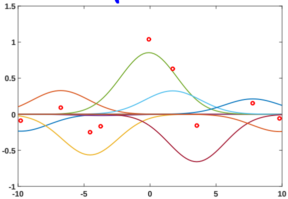
Regression problem: find $\boldsymbol{\beta}$ and b s.t.

$$
\begin{aligned}
y_i = f(\boldsymbol{x}_i) &= \boldsymbol{\beta}^T \varphi(\boldsymbol{x}_i) + b \\
&= \sum_j \lambda'_j \underbrace{\varphi^T(\boldsymbol{x}_j)\,\varphi(\boldsymbol{x}_i)} + b \\
&= \sum_j \lambda'_j \underbrace{\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)} + b.
\end{aligned}
$$

We do not care about φ, but about κ, the inner product.



=



The Primal Optimization Problem

Let $\epsilon > 0$ be the error tolerance. Then one has

$$\begin{aligned} \min \quad & \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 \\ \text{subject to} \quad & \left| y_i - \boldsymbol{\beta}^T \varphi(\boldsymbol{x}_i) - b \right| \le \epsilon. \end{aligned}$$

The constraints are equivalent to

$$\begin{aligned} y_i - \boldsymbol{\beta}^T \varphi(\boldsymbol{x}_i) - b \le \epsilon, \\ \boldsymbol{\beta}^T \varphi(\boldsymbol{x}_i) + b - y_i \le \epsilon. \end{aligned}$$

Now if we allow additional noise, represented by $\xi_i \ge 0$ and $\xi_i^\star \ge 0$. Then

$$\begin{aligned} \min \quad & \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 + C \sum (\xi_i + \xi_i^\star) \\ \text{subject to} \quad & y_i - \boldsymbol{\beta}^T \varphi(\boldsymbol{x}_i) - b \le \epsilon + \xi_i, \\ & \boldsymbol{\beta}^T \varphi(\boldsymbol{x}_i) + b - y_i \le \epsilon + \xi_i^\star, \\ & -\xi_i \le 0, \quad -\xi_i^\star \le 0. \end{aligned}$$

Lagrangian

$$
\begin{aligned}
L = \frac{1}{2}\left\|\boldsymbol{\beta}\right\|_2^2 + C\sum\left(\xi_i + \xi_i^\star\right) - \sum_i \Big(& u_i\xi_i + \sum u_i^\star\xi_i^\star \Big) \\
& + \lambda_i\left(y_i - \boldsymbol{\beta}^T\varphi\left(\boldsymbol{x}_i\right) - b - \epsilon - \xi_i\right) \\
& + \lambda_i^\star\left(\boldsymbol{\beta}^T\varphi\left(\boldsymbol{x}_i\right) + b - y_i - \epsilon - \xi_i^\star\right),
\end{aligned}
$$

where $u_i, u_i^\star, \xi_i, \xi_i^\star \geq 0$ are Lagrange multiplier. To minimize L,

$$
\begin{aligned}
dL/d\boldsymbol{\beta} = \mathbf{0} \quad & \Rightarrow \quad \boldsymbol{\beta} = \sum_i\left(\lambda_i - \lambda_i^\star\right)\varphi\left(\boldsymbol{\xi}_i\right), \\
dL/db = \mathbf{0} \quad & \Rightarrow \quad \sum\lambda_i = \sum\lambda_i^\star, \\
dL/d\xi_i = 0,\ dL/d\xi_i^\star = 0 \quad & \Rightarrow \quad \lambda_i \leq C,\ \lambda_i^\star \leq C.
\end{aligned}
$$

The Dual Problem

The objective function of the dual problem

$$L_D = -\epsilon \sum (\lambda_i + \lambda_i^\star) + y_i \sum (\lambda_i - \lambda_i^\star)$$
$$- \underbrace{\frac{1}{2} \sum_{i,j} (\lambda_i - \lambda_i^\star) \left(\lambda_j - \lambda_j^\star\right) \kappa (\boldsymbol{x}_i, \boldsymbol{x}_j),}_{\|\boldsymbol{\beta}\|_2^2}$$

where $\kappa (\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi^T (\boldsymbol{x}_i) \varphi (\boldsymbol{x}_j)$.

The optimizatoin constraints are

$$\sum (\lambda_i - \lambda_i^\star) = 0,$$
$$0 \le \lambda_i, \lambda_i^\star \le C.$$

KKT Condition and Support Vectors

Part of the KKT condition is that $\forall i$,

$$\begin{cases} \lambda_i \left(y_i - \boldsymbol{\beta}^T \varphi\left(\boldsymbol{x}_i\right) - b - \epsilon - \xi_i \right) = 0, \\ \lambda_i^\star \left(\boldsymbol{\beta}^T \varphi\left(\boldsymbol{x}_i\right) + b - y_i - \epsilon - \xi_i^\star \right) = 0. \end{cases}$$

- Interior points: $\left| y_i - \boldsymbol{\beta}^T \varphi\left(\boldsymbol{x}_i\right) - b \right| < \epsilon + \xi_i$.
  - Both λ_i and $\lambda_i^\star$ are zero.
- Boundary points: $\left| y_i - \boldsymbol{\beta}^T \varphi\left(\boldsymbol{x}_i\right) - b \right| = \epsilon + \xi_i$.
  - One of λ_i and $\lambda_i^\star$ is zero.
  - $\lambda_i \neq \lambda_i^\star$.

The Standard Form

Let $\gamma_i = \lambda_i$ and $\gamma_{i+n} = \lambda_i^\star$ (Merge $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^\star$ into a single vector). The dual problem becomes

$$\min_{\boldsymbol{\gamma}} \quad \frac{1}{2}\boldsymbol{\gamma}^T \boldsymbol{Q} \boldsymbol{\gamma} + \boldsymbol{v}^T \boldsymbol{\gamma},$$

$$\text{subject to} \quad 0 \le \gamma_i \le C, \quad \sum_{i=1}^{n} \gamma_i - \sum_{i=n+1}^{2n} \gamma_i = 0.$$

The boundary points are given by $\mathcal{I} = \{i : \gamma_i - \gamma_{i+n} \neq 0\}$.

For a new data point $\boldsymbol{x}^{\text{new}}$, the regression is

$$\boxed{f\left(\boldsymbol{x}^{\text{new}}\right) = \sum_i \left(\gamma_i - \gamma_{i+n}\right) \kappa\left(\boldsymbol{x}_i, \boldsymbol{x}^{\text{new}}\right) + b.}$$

Section 10
Gaussian Distribution

Statistical learning.

Gaussian Random Vectors

A random vector $\boldsymbol{X} \in \mathbb{R}^n$ is Gaussian distributed $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if its pdf is given by

$$p(\boldsymbol{x}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathcal{S}_+^n$ (the set of $n \times n$ symmetric positive semidefinite matrices).

Here, we have assumed that $\boldsymbol{\Sigma}$ is invertible (of full rank).

Gaussian Random Vectors: Characteristic Function

$$\text{PDF} \underset{\text{Inverse Fourier Transform}}{\overset{\text{Fourier Transform}}{\rightleftharpoons}} \text{Characteristic function } \mathrm{E}\left[e^{i\langle \boldsymbol{\lambda}, \boldsymbol{X}\rangle}\right].$$

$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if

$$\mathrm{E}\left[e^{i\langle \boldsymbol{\lambda}, \boldsymbol{X}\rangle}\right] = \exp\left(i\langle \boldsymbol{\lambda}, \boldsymbol{\mu}\rangle - \frac{1}{2}\boldsymbol{\lambda}^T \boldsymbol{\Sigma} \boldsymbol{\lambda}\right).$$

It is well defined even when $\boldsymbol{\Sigma}$ is not invertible.

Quadratic form in both the pdf
and characteristic function.

Affine Transformation

Lemma 10.1

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Then for any* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ *and* $\boldsymbol{b} \in \mathbb{R}^m$,

$$\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\right).$$

Proof:

$$\begin{aligned}
\mathrm{E}\left[e^{i\langle\boldsymbol{\lambda},\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b}\rangle}\right] &= \mathrm{E}\left[e^{i\left\langle\boldsymbol{A}^T\boldsymbol{\lambda},\boldsymbol{X}\right\rangle+i\langle\boldsymbol{\lambda},\boldsymbol{b}\rangle}\right] \\
&= \exp\left(i\left\langle\boldsymbol{A}^T\boldsymbol{\lambda},\boldsymbol{\mu}\right\rangle - \frac{1}{2}\left(\boldsymbol{A}^T\boldsymbol{\lambda}\right)^T\boldsymbol{\Sigma}\left(\boldsymbol{A}^T\boldsymbol{\lambda}\right)\right)e^{i\langle\boldsymbol{\lambda},\boldsymbol{b}\rangle} \\
&= \exp\left(i\left\langle\boldsymbol{\lambda},\boldsymbol{A}^T\boldsymbol{\mu}+\boldsymbol{b}\right\rangle - \frac{1}{2}\boldsymbol{\lambda}^T\left(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\right)\boldsymbol{\lambda}\right).
\end{aligned}$$

Gaussian Conditioning Lemma

Lemma 10.2

Let $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.
Let $\boldsymbol{X}_{\mathcal{A}}$ *and* $\boldsymbol{X}_{\mathcal{B}}$ *be two subvectors of* $\boldsymbol{X}$, *i.e.,* $\boldsymbol{X} = \left[\boldsymbol{X}_{\mathcal{A}}^T, \boldsymbol{X}_{\mathcal{B}}^T\right]^T$.
Let $\boldsymbol{K} := \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{K}_{\mathcal{A}\mathcal{A}} & \boldsymbol{K}_{\mathcal{A}\mathcal{B}} \\ \boldsymbol{K}_{\mathcal{B}\mathcal{A}} & \boldsymbol{K}_{\mathcal{B}\mathcal{B}} \end{bmatrix}$ *be the precision matrix.*
Then $\boldsymbol{X}_{\mathcal{A}} | \boldsymbol{X}_{\mathcal{B}} \sim P_{\boldsymbol{X}_{\mathcal{A}}|\boldsymbol{X}_{\mathcal{B}}} = \mathcal{N}\left(-\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{X}_{\mathcal{B}}, \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\right)$. *In other words,*

$$\boldsymbol{X}_{\mathcal{A}} = -\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{X}_{\mathcal{B}} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\right)$ *is independent of* $\boldsymbol{X}_{\mathcal{B}}$.

Remark: $\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1} \neq \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}$.

Matrix Identities

- Block matrix inverse (BMI)

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} & \left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1} \end{bmatrix}. \tag{27}$$

- Woodbury matrix identity (WMI)

$$\left(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V}\right)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\left(\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}. \tag{28}$$

Proof of Gaussian Conditioning Lemma

By Bayes rule, $p\left(\boldsymbol{x}_{\mathcal{A}}|\boldsymbol{x}_{\mathcal{B}}\right) = p\left(\boldsymbol{x}_{\mathcal{A}}, \boldsymbol{x}_{\mathcal{B}}\right)/p\left(\boldsymbol{x}_{\mathcal{B}}\right)$. Then

$$\begin{aligned} \ln p\left(\boldsymbol{x}_{\mathcal{A}}|\boldsymbol{x}_{\mathcal{B}}\right) &= \ln p\left(\boldsymbol{x}_{\mathcal{A}}, \boldsymbol{x}_{\mathcal{B}}\right) - \ln p\left(\boldsymbol{x}_{\mathcal{B}}\right) \\ &= c - \frac{1}{2}\boldsymbol{x}_{\mathcal{A}}^T \boldsymbol{K}_{\mathcal{A}\mathcal{A}} \boldsymbol{x}_{\mathcal{A}} - \boldsymbol{x}_{\mathcal{A}}^T \boldsymbol{K}_{\mathcal{A}\mathcal{B}} \boldsymbol{x}_{\mathcal{B}} - \frac{1}{2}\boldsymbol{x}_{\mathcal{B}}^T \left(\boldsymbol{K}_{\mathcal{B}\mathcal{B}} - \boldsymbol{\Sigma}_{\mathcal{B}\mathcal{B}}^{-1}\right) \boldsymbol{x}_{\mathcal{B}}, \end{aligned}$$

where c is a constant. By (27),

$$\boldsymbol{\Sigma}_{\mathcal{B}\mathcal{B}}^{-1} = \boldsymbol{K}_{\mathcal{B}B} - \boldsymbol{K}_{\mathcal{B}\mathcal{A}} \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1} \boldsymbol{K}_{\mathcal{A}\mathcal{B}}.$$

One has

$$\ln p\left(\boldsymbol{x}_{\mathcal{A}}|\boldsymbol{x}_{\mathcal{B}}\right) = c - \frac{1}{2}\left(\boldsymbol{x}_{\mathcal{A}} + \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{x}_{\mathcal{B}}\right)^T \boldsymbol{K}_{\mathcal{A}\mathcal{A}} \left(\boldsymbol{x}_{\mathcal{A}} + \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{x}_{\mathcal{B}}\right).$$

That is, $\boldsymbol{X}_{\mathcal{A}}|\boldsymbol{X}_{\mathcal{B}} \sim \mathcal{N}\left(-\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{X}_{\mathcal{B}}, \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\right)$.

A Signal Processing Application

The problem:
Given

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{W},$$

where $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_x)$ and $\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_w)$.

Given observation $\boldsymbol{y}$, want to find $\hat{\boldsymbol{x}} = f(\boldsymbol{y})$ s.t. the mean squared error $\mathrm{E}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2\right]$ is minimized (MMSE solution).

Fact: The general MMSE solution is given by

$$\hat{\boldsymbol{x}} = \mathrm{E}[\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}] = \int \boldsymbol{x} \cdot p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})\, d\boldsymbol{x}.$$

Hence for Gaussian random variables, Gaussian conditioning lemma can be used.

Finding the MMSE Solution

1. $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{W}$ is Gaussian distributed $\mathcal{N}\left(\mathbf{0}, \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^T + \boldsymbol{\Sigma}_w\right)$.
2.
$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \underbrace{\begin{bmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_x\boldsymbol{A}^T \\ \boldsymbol{A}\boldsymbol{\Sigma}_x & \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^T + \boldsymbol{\Sigma}_w \end{bmatrix}}_{\boldsymbol{\Sigma}}\right).$$
3. Find the precision matrix from $\boldsymbol{\Sigma}$:
$$\boldsymbol{K} = \begin{bmatrix} \boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1} \\ -\boldsymbol{\Sigma}_w^{-T}\boldsymbol{A} & \text{sth} \end{bmatrix}$$
4. $\boldsymbol{X}|\boldsymbol{Y} \sim \mathcal{N}\left(-\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{Y}, \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\right)$ by Gaussian Conditioning Lemma.

   We use the conditional mean as the estimate $\hat{\boldsymbol{x}}$:
$$\hat{\boldsymbol{x}} = \left(\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{y}. \tag{29}$$
$$\boldsymbol{\Sigma}_{X|Y} = \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1} = \left(\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{A}\right)^{-1}. \tag{30}$$

Calculation of The $\boldsymbol{K}$ Matrix

$$\begin{aligned}
\boldsymbol{K}_{\mathcal{A}\mathcal{A}} &\overset{\text{BMI(27)}}{=} \left(\boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_x \boldsymbol{A}^T \left(\boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^T + \boldsymbol{\Sigma}_w\right)^{-1} \boldsymbol{A}\boldsymbol{\Sigma}_x\right)^{-1} \\
&\overset{\text{WMI(28)}}{=} \left(\left(\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{A}\right)^{-1}\right)^{-1} \\
&= \boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{A}.
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{K}_{\mathcal{A}\mathcal{B}} &\overset{\text{BMI(27)}}{=} -\boldsymbol{\Sigma}_x^{-1}\left(\boldsymbol{\Sigma}_x\boldsymbol{A}^T\right)\left(\boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^T + \boldsymbol{\Sigma}_w - \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_x\boldsymbol{A}^T\right)^{-1} \\
&= -\boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}.
\end{aligned}$$

Hence $\boldsymbol{\Sigma}_{X|Y} = \left(\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{A}\right)^{-1}$ and $\boldsymbol{L} = \boldsymbol{\Sigma}_{X|Y}\boldsymbol{A}^T\boldsymbol{\Sigma}_w^{-1}$.

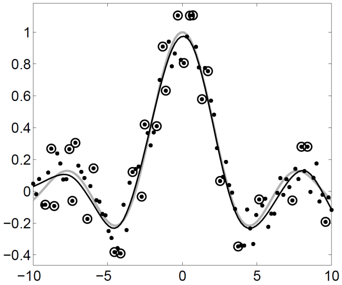
Section 11
Sparse Bayesian Learning

Didn't do it.

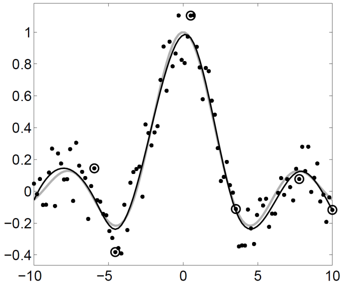
Sparse Bayesian Learning (SBL) via Relevance Vector Machine (RVM)

M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," J. Mach. Learn. Res., JMLR.org, 2001, 1, 211-244.

Examples from Tipping's paper:
To approximate the sinc function



Support vector approx.

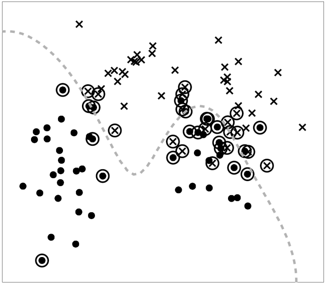


Relevance vector approx.

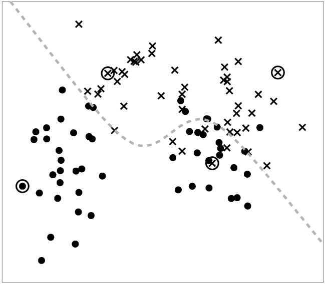
Sparse Bayesian Learning (SBL) via Relevance Vector Machine (RVM)

M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," J. Mach. Learn. Res., JMLR.org, 2001, 1, 211-244.

Examples from Tipping's paper:
Classify two classes



SVM



RVM

RVM: A Hierarchical Gaussian Model

Target: Solve the sparse linear inverse problem:

$$\boldsymbol{y} = \boldsymbol{\Phi} \boldsymbol{x} + \boldsymbol{w}.$$

Signal model: signal components are i.i.d. Gaussian

$$p\left(X_i|\alpha_i\right) = \mathcal{N}\left(X_i; 0, \alpha_i^{-1}\right).$$

- $\alpha_i > 0$: precision ($\alpha_i^{-1} = \sigma_i^2$, for mathematical convenience)
- $\alpha_i = \infty$: $X_i \sim \mathcal{N}\left(0, 0\right)$ and $X_i = 0$ with prob. one.
- $\alpha_i \in \mathbb{R}^+$: $X_i \sim \mathcal{N}\left(0, \alpha_i^{-1}\right)$ and $X_i \neq 0$ with prob. one.

Signal Estimation with Known $\boldsymbol{\alpha}$

If we know $\boldsymbol{\alpha}$, then the posterior of $\boldsymbol{x}$ can be derived by Bayes' rule

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{x})\, p(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y})} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\mu} = \sigma_w^{-2} \boldsymbol{\Sigma} \boldsymbol{\Phi}^T \boldsymbol{y},\ \boldsymbol{\Sigma} = \left(\boldsymbol{A} + \sigma_w^{-2} \boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1},\ \boldsymbol{A} = \mathrm{diag}(\boldsymbol{\alpha}).$$

See (29) and (30) for details.

Set $\hat{\boldsymbol{x}} = \boldsymbol{\mu}$. This is the MMSE solution as well.

Signal Recovery with Unknown $\boldsymbol{\alpha}$

The key is to estimate $\boldsymbol{\alpha}$:

- $\boldsymbol{Y} \sim \mathcal{N}\left(\mathbf{0}, \sigma_w^2 \boldsymbol{I} + \boldsymbol{\Phi} \boldsymbol{A}^{-1} \boldsymbol{\Phi}^T\right)$
- Linear combination of Gaussian is Gaussian ($\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$).
- Maximum likelihood (ML) estimate:

  $$\mathcal{L}(\boldsymbol{\alpha}) = \log p\left(\boldsymbol{y} | \boldsymbol{\alpha}, \sigma_w^2\right) = -\frac{1}{2}\left[N \log 2\pi + \log |\boldsymbol{C}| + \boldsymbol{y}^T \boldsymbol{C}^{-1} \boldsymbol{y}\right],$$

  where $\boldsymbol{C} = \sigma_w^2 \boldsymbol{I} + \boldsymbol{\Phi} \boldsymbol{A}^{-1} \boldsymbol{\Phi}^T$.
- Marginal likelihood maximization: For a given i, we solve

  $$\max_{\alpha_i} \mathcal{L}(\boldsymbol{\alpha})$$

  by fixing all α_j's, $j \neq i$.

Marginal Likelihood Maximization (1)

$$\begin{aligned} \boldsymbol{C} &= \sigma^2 \boldsymbol{I} + \boldsymbol{\Phi} \boldsymbol{A}^{-1} \boldsymbol{\Phi} \\ &= \sigma^2 \boldsymbol{I} + \sum_{j \neq i} \alpha_j^{-1} \boldsymbol{\phi}_j \boldsymbol{\phi}_j^T + \alpha_i^{-1} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T \\ &= \boldsymbol{C}_{-i} + \alpha_i^{-1} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T. \end{aligned}$$

One has

$$\begin{aligned} |\boldsymbol{C}| &= |\boldsymbol{C}_{-i}| \cdot \left| \boldsymbol{I} + \alpha_i^{-1} \boldsymbol{C}_{-i}^{-1/2} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T \boldsymbol{C}_{-i}^{-1/2} \right| \\ &= |\boldsymbol{C}_{-i}| \cdot \left| 1 + \alpha_i^{-1} \boldsymbol{\phi}_i^T \boldsymbol{C}_{-i}^{-1} \boldsymbol{\phi}_i \right|. \end{aligned}$$

The last line comes from $\left| \boldsymbol{I} + \boldsymbol{M}\boldsymbol{M}^T \right| = \left| \boldsymbol{I} + \boldsymbol{M}^T \boldsymbol{M} \right|$, which can be easily verified by the SVD of $\boldsymbol{M}$.

$$\boldsymbol{C}^{-1} \overset{\text{WMI(28)}}{=} \boldsymbol{C}_{-i}^{-1} - \frac{\boldsymbol{C}_{-i}^{-1} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T \boldsymbol{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\phi}_i^T \boldsymbol{C}_{-i}^{-1} \boldsymbol{\phi}_i}.$$

Marginal Likelihood Maximization (2)

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\alpha}) &= -\tfrac{1}{2}\left[N\log 2\pi + \log|\boldsymbol{C}| + \boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{y}\right], \\
&= c - \tfrac{1}{2}\Big[\log|\boldsymbol{C}_{-i}| + \boldsymbol{y}^T\boldsymbol{C}_{-i}^{-1}\boldsymbol{y} \\
&\qquad - \log\alpha_i + \log\left(\alpha_i + \boldsymbol{\phi}_i^T\boldsymbol{C}_{-i}^{-1}\boldsymbol{\phi}_i\right) - \tfrac{\left(\boldsymbol{\phi}_i^T\boldsymbol{C}_{-i}^{-1}\boldsymbol{y}\right)^2}{\alpha_i+\boldsymbol{\phi}_i^T\boldsymbol{C}_{-i}^{-1}\boldsymbol{\phi}_i}\Big] \\
&= \mathcal{L}(\boldsymbol{\alpha}_{-i}) + \tfrac{1}{2}\left[\log\alpha_i - \log(\alpha_i + s_i) - \tfrac{q_i^2}{\alpha_i+s_i}\right] \\
&= \mathcal{L}(\boldsymbol{\alpha}_{-i}) + \ell(\alpha_i),
\end{aligned}
$$

where for simplification of expressions, we have defined

$$
s_i := \boldsymbol{\phi}_i^T\boldsymbol{C}_{-i}^{-1}\boldsymbol{\phi}_i, \text{ and } q_i := \boldsymbol{\phi}_i^T\boldsymbol{C}_{-i}^{-1}\boldsymbol{y}.
$$

Set the derivative of $\ell(\alpha_i)$ to zero. One obtains the closed form solution for the optimal α_i:

$$
\alpha_i = \begin{cases} \frac{s_i^2}{q_i^2 - s_i} & \text{if } q_i^2 > s_i, \\ \infty & \text{if } q_i^2 \le s_i, \end{cases}
$$

Algorithms for RVM

A sequential algorithm: [Tipping & Faul, 2003]

In each iteration:

- Scan $i \in [N]$ and find the i that $\mathcal{L}(\alpha_i^*) - \mathcal{L}(\alpha_i)$ is maximized.
- Set $\alpha_i = \alpha_i^*$.
- Update $\mathcal{L}$ and other parameters (preparation for the next iteration).

Connections to Greedy Algorithms

- Similar to greedy algorithms
  - The sequential algorithm is very similar to OMP.
  - "Subspace pursuit" type of RVM: [Karseras & D., 2013]
  - Performance guarantees: *sufficient* conditions based on mutual coherence or RIP [Karseras & D., 2013]
- Different from greedy algorithms
  - Noise variance is considered.
  - Statistical information: estimation quality information

A Demo

Track global Ozone density

- Spatial sparsity:
  - Each frame is sparse under DCT transform
- Temporal sparsity: correlations across days
  - Support innovation.
  - Nonzero component magnitude innovation.

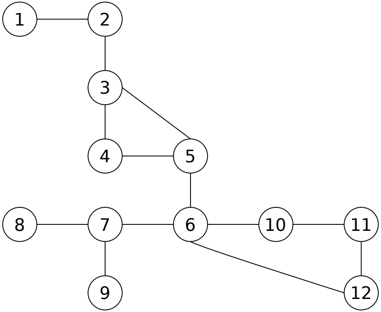
Let $\boldsymbol{x}^t$ be the DCT coefficient vector at the t-th day:

$$\boldsymbol{x}^t = \boldsymbol{x}^{t-1} + \boldsymbol{u}^t,$$
$$\boldsymbol{y}^t = \boldsymbol{A}\boldsymbol{D}\boldsymbol{x}^t + \boldsymbol{w}^t,$$

where $\boldsymbol{u}^t \sim \prod \mathcal{N}\left(0, \alpha_i^{-1}\right)$, $\boldsymbol{A}$ is the sampling matrix, and $\boldsymbol{w}^t$ is the white Gaussian noise.

Section 12
Gaussian Graphic Model

Motivation: Gaussian Graphic Model



Encoding the conditional dependencies between n random variables $X_1, \cdots, X_n$ by a graph.

Correlation and Conditional Independence

Sneeze — Catch Cold — Weather Change

Observation: "Weather Change" and "Sneeze" are correlated.

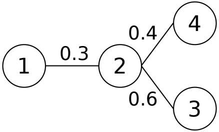
- "Weather Change" and "Catch Cold" are highly correlated.
- "Catch Cold" and "Sneeze" are highly correlated.

However, given the status of "Catch Cold", "Weather Change" and "Sneeze" are independent.

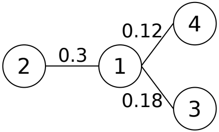
- Given that "Catch Cold" is false, "Sneeze" is likely to be false, independent of whether "Weather Change" is true or not.
- Given that "Catch Cold" is true, "Sneeze" is likely to be true, independent of whether "Weather Change" is true or not.

Other Examples

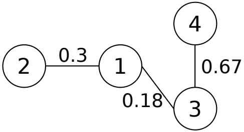
Suppose that $\rho\left(X_1, X_2\right) = 0.3$, $\rho\left(X_1, X_3\right) = 0.18$, and $\rho\left(X_1, X_4\right) = 0.12$.
Suppose that on one day, $X_2 \uparrow 0.2$, $X_3 \downarrow 0.1$, and $X_4 \downarrow 0.5$.
Find the expected change of X_1.



$$\mathrm{E}\left[\Delta X_1\right] = 0.2 \times 0.3 = 0.06.$$

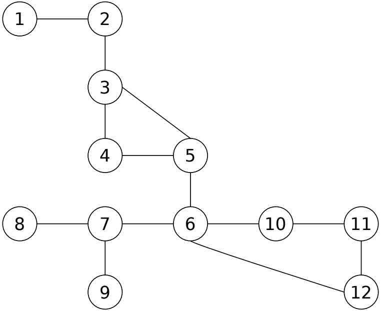


$$\begin{aligned}\mathrm{E}\left[\Delta X_1\right] &= 0.2 \times 0.3 - 0.1 \times 0.18 - 0.5 \times 0.12 \\ &= -0.018.\end{aligned}$$



$$\begin{aligned}\mathrm{E}\left[\Delta X_1\right] &= 0.2 \times 0.3 - 0.1 \times 0.18 \\ &= 0.042.\end{aligned}$$

Nondirected Graphical Model



The distribution of the Gaussian random vector $\boldsymbol{X} = [X_1, \cdots, X_n]^T$ is a graphic model according to the graph g if

$$\text{for all } a: \; X_a \perp \{X_b: \; b \notin \mathrm{ne}(a), \; b \neq a\} \text{ given } \{X_c: \; c \in \mathrm{ne}(a)\}.$$

Or, given X_c's, $c \in \mathrm{ne}(a)$, X_a and X_b's are independent for all b not in the neighborhood.

Consequence of Gaussian Conditioning

Recall the Gaussian conditioning lemma (Lemma 10.2).
Let $\boldsymbol{K}$ be the precision matrix of $\boldsymbol{X}$.

Corollary 12.1

For any $a \in [n]$,

$$X_a = -\sum_{b:\ b \neq a} \frac{K_{ab}}{K_{aa}} X_b + \epsilon_a,$$

where $\epsilon_a \sim \mathcal{N}\left(0, K_{aa}^{-1}\right)$ *is independent of* $\{X_b : \ b \neq a\}$.

Proof: Apply Lemma 10.2 with $A = \{a\}$ and $B = [n] \setminus \{a\} = A^c$.

Remark: Find the neighboring points.

Conditional Correlation

Corollary 12.2

$$\text{cor}\left(X_a, X_b | \boldsymbol{X}_{\mathcal{C}}\right) = -\frac{K_{ab}}{\sqrt{K_{aa}K_{bb}}}.$$

Proof: From Gaussian Conditioning (Lemma 10.2), it holds that

$$\text{cov}\left(\boldsymbol{X}_{\{a,b\}} | \boldsymbol{X}_{\mathcal{C}}\right) = \begin{bmatrix} K_{aa} & K_{ab} \\ K_{ba} & K_{bb} \end{bmatrix}^{-1} = \frac{1}{K_{aa}K_{bb} - K_{ab}^2} \begin{bmatrix} K_{bb} & -K_{ba} \\ -K_{ab} & K_{aa} \end{bmatrix}.$$

Plug this formula into the definition of conditional correlation. Corollary 12.2 is proved.

Remark: Find the correlation between neighboring points.

Estimate the Precision Matrix

From the definition $\boldsymbol{K} = \boldsymbol{\Sigma}^{-1}$, the computation seems straightforward. However, the commonly used fact

$$\frac{1}{m} \sum \left(\boldsymbol{X} - \bar{\boldsymbol{X}} \right) \left(\boldsymbol{X} - \bar{\boldsymbol{X}} \right)^T \to \boldsymbol{\Sigma} \tag{31}$$

is based on the assumption that n is fixed and $m \to \infty$.

In reality, we may not have sufficient data m. Hence (31) may not be applicable.

Assumption: $\boldsymbol{K}$ is sparse.

Estimation via Regression (1)

Define the matrix $\boldsymbol{\Theta}$ by $\theta_{ab} = -K_{ab}/K_{bb}$ for $b \neq a$ and $\theta_{aa} = 0$. Then Corollary 12.1 implies

$$\mathrm{E}\left[X_a | X_b : \ b \neq a\right] = \sum_b \theta_{ba} X_b.$$

Hence we need to find θ_{ba}'s $(b \neq a)$ to minimize

$$\mathrm{E}\left[\left(X_a - \sum_b \theta_{ba} X_b\right)^2\right].$$

Or in matrix format

$$\hat{\boldsymbol{\Theta}} = \arg \min_{\boldsymbol{\Theta} \in \Theta} \mathrm{E}\left[\left\|\boldsymbol{X} - \boldsymbol{\Theta}^T \boldsymbol{X}\right\|_2^2\right],$$

where $\Theta = \{\boldsymbol{\Theta} : \ \mathrm{diag}\left(\boldsymbol{\Theta}\right) = \mathbf{0}\}$.

Estimation via Regression (2)

The objective function can be rewritten as

$$\mathrm{E}\left[\left\|\boldsymbol{X} - \boldsymbol{\Theta}^T \boldsymbol{X}\right\|_2^2\right] \approx \frac{1}{m} \sum \left(\boldsymbol{x} - \boldsymbol{\Theta}^T \boldsymbol{x}\right)^T \left(\boldsymbol{x} - \boldsymbol{\Theta}^T \boldsymbol{x}\right)$$

$$= \frac{1}{m} \left\| \begin{bmatrix} \boldsymbol{x}_{(1)}^T \\ \vdots \\ \boldsymbol{x}_{(m)}^T \end{bmatrix} - \begin{bmatrix} \boldsymbol{x}_{(1)}^T \\ \vdots \\ \boldsymbol{x}_{(m)}^T \end{bmatrix} \boldsymbol{\Theta} \right\|_F^2$$

$$= \frac{1}{m} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{\Theta}\right\|_F^2 .$$

Note that the $\boldsymbol{X}$ on this slide is the data matrix and the $\boldsymbol{X}$ on previous slides are random vectors.

Estimation via Regression (3)

The overall optimization problem:

$$\min_{\boldsymbol{\Theta} \in \Theta} \quad \frac{1}{m} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{\Theta}\|_F^2 + \lambda \sum_{a \neq b} |\theta_{ab}|,$$

Or

$$\min_{\boldsymbol{\Theta} \in \Theta} \quad \frac{1}{m} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{\Theta}\|_F^2 + \lambda \sum_{a < b} \sqrt{\theta_{ab}^2 + \theta_{ba}^2}.$$