Machine Learning Principles

Class5 : Sept. 15

Linear Regression II

Instructor: Diana Kim

Today's Lecture

- **Linear Regression**

  (1) Basic Overview (blackboard)

  - from modeling to learning

  - MLE & MAP

  (2) Convex Optimization Theory

  - necessary & sufficient condition for optimality

  - equality constraint problem

  - inequality constraint problem

  - three interpretations of MMSE with regularization

[1] What is the regression problem?

- Learning the function f
  to predict continuous y
  given the value of M dimensional input data $(x_1, x_{2,}, \dots . x_{m,})$

$$y = f(x_1, x_{2,}, \dots . x_{m.})$$

(functional relation between x and y)

[2] What is the regression problem?

$y = ax + b$ [linear]

$y = ax^3 + bx^2 + c$ [non-linear]

$y = a\exp\{-\frac{(x-\mu_1)^2}{2\sigma_1}\} + b\exp\{-\frac{(x-\mu_2)^2}{2\sigma_2}\} + c\exp\{-\frac{(x-\mu_3)^2}{2\sigma_3}\}$[non-linear]

[3] What is the linear regression problem? (linear representation)

$$y = ax + b \quad \longleftrightarrow \quad y = \begin{bmatrix} x & 1 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix}$$

$$y = ax^3 + bx^2 + c \quad \longleftrightarrow \quad y = \begin{bmatrix} x^3 & x^2 & 1 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

$$y = a \exp\{-\frac{(x-\mu_1)^2}{2\sigma_1}\} + b \exp\{-\frac{(x-\mu_2)^2}{2\sigma_2}\} + c \exp\{-\frac{(x-\mu_3)^2}{2\sigma_3}\}$$

Regression modeling can be expressed
as a linear combination of parameters and data features, hence the name is **Linear Regression.**

[8] What is the linear regression problem? (intuitive way of learning)

$$y = ax + b$$

$$y = ax^3 + bx^2 + c$$

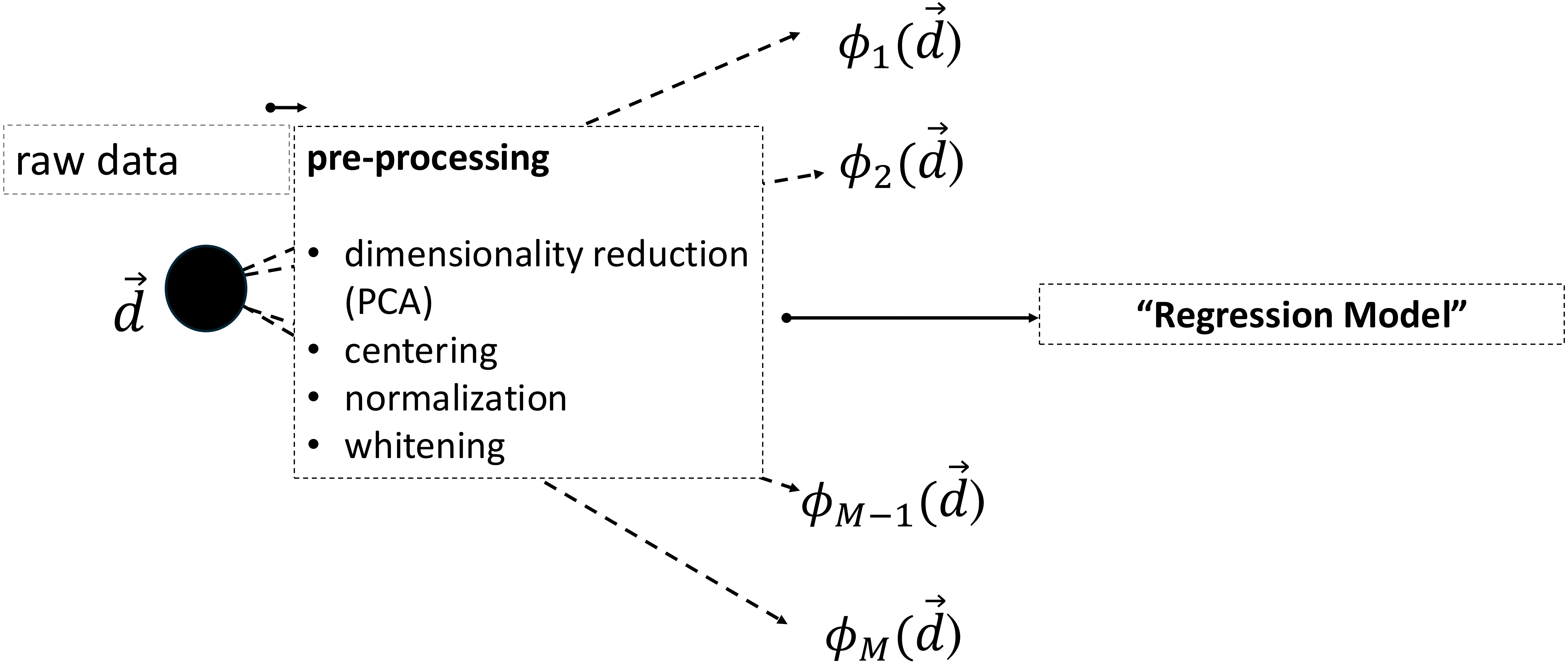
$$y = a \exp\{-\frac{(x-\mu_1)^2}{2\sigma_1}\} + b \exp\{-\frac{(x-\mu_2)^2}{2\sigma_2}\} + c \exp\{-\frac{(x-\mu_3)^2}{2\sigma_3}\}$$

Q: how could we learn the a, b, c?

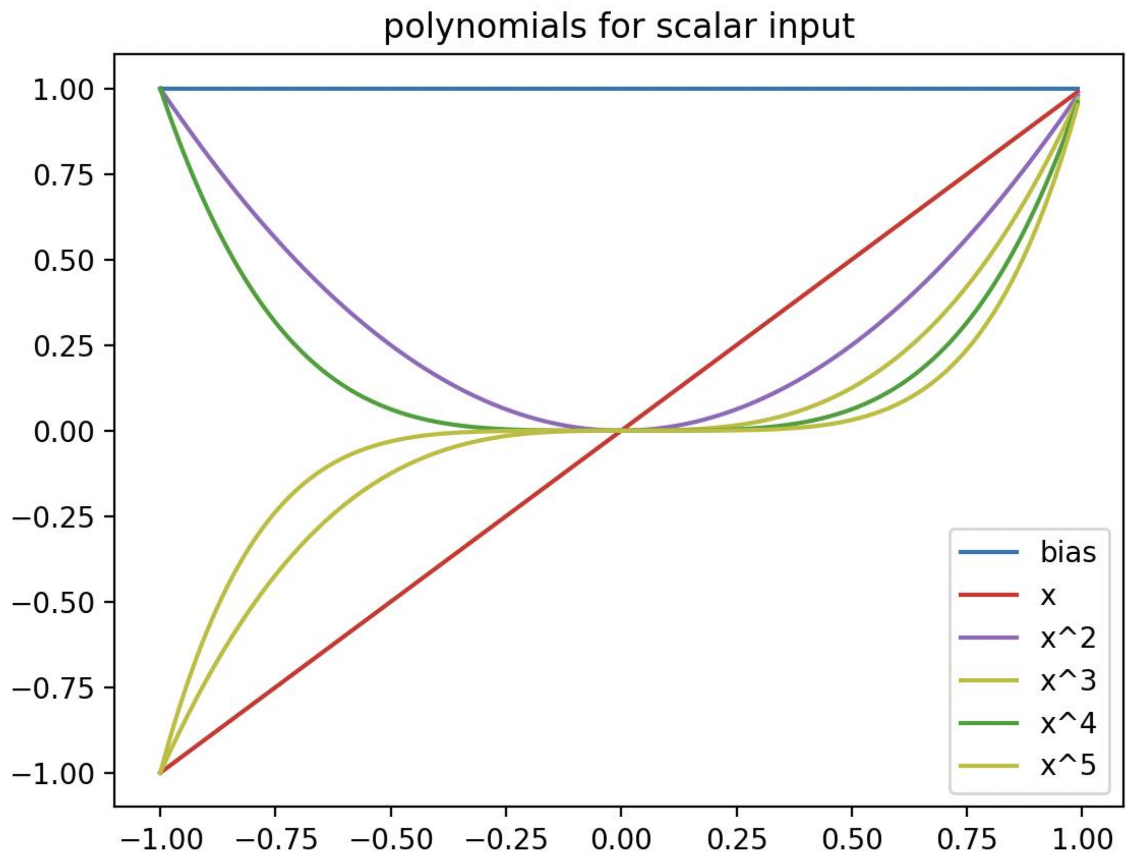
Basis Functions (Feature Functions)

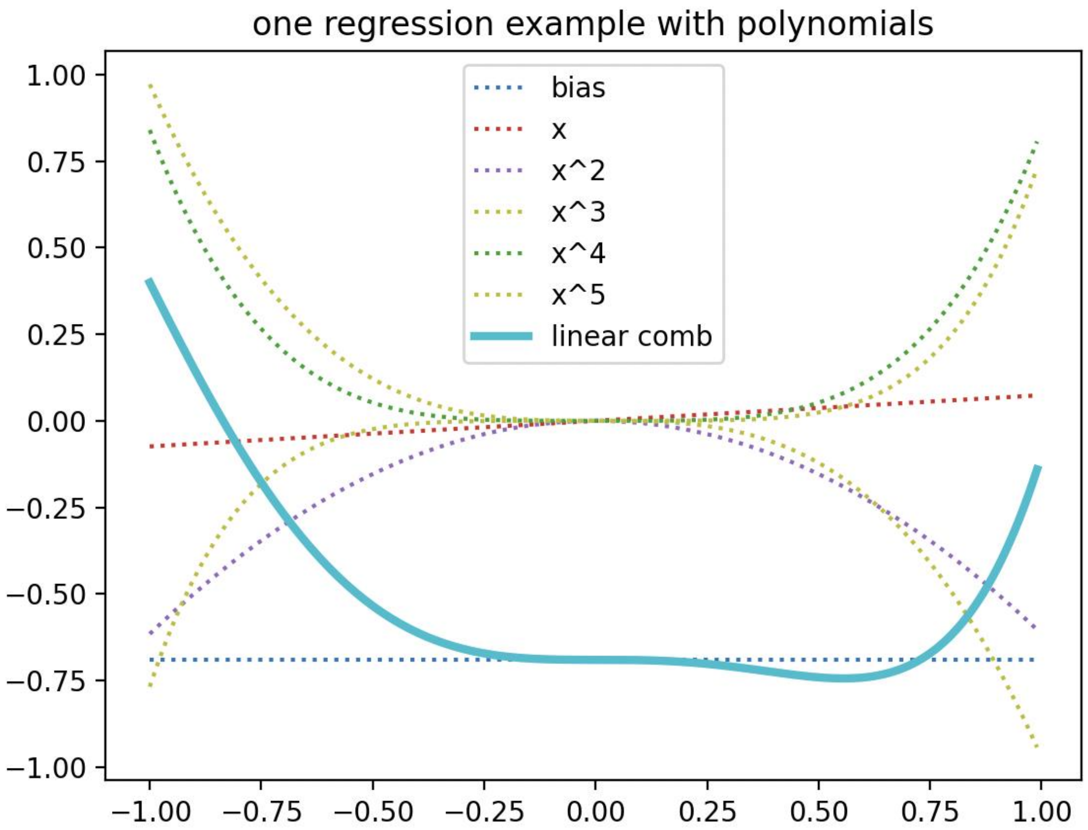
-a set of functions sharing the same domain with the raw data
-elementary functions to describe a function we target

[1] Basis Functions (a set of function on the space of raw data)



[2] Basis Functions (polynomial expansion: scalar)



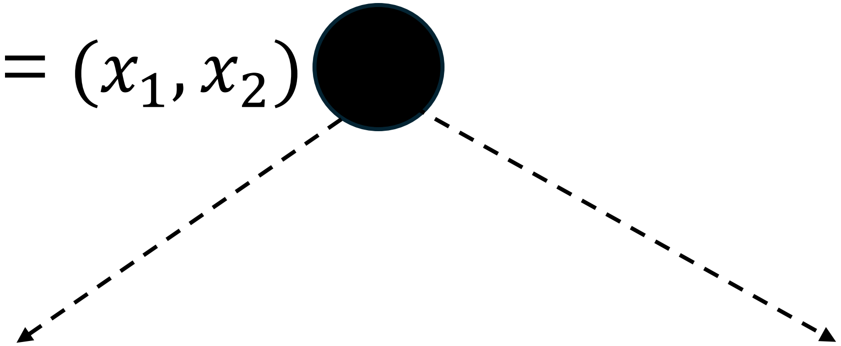


- This example shows the case when input is scalar.
- Q: what if input is a 2D vector? How would you draw the plot?

[3] Basis Functions (polynomial expansion:2d)

[quadratic polynomial example for 2-d raw data]

$\vec{x} = (x_1, x_2)$

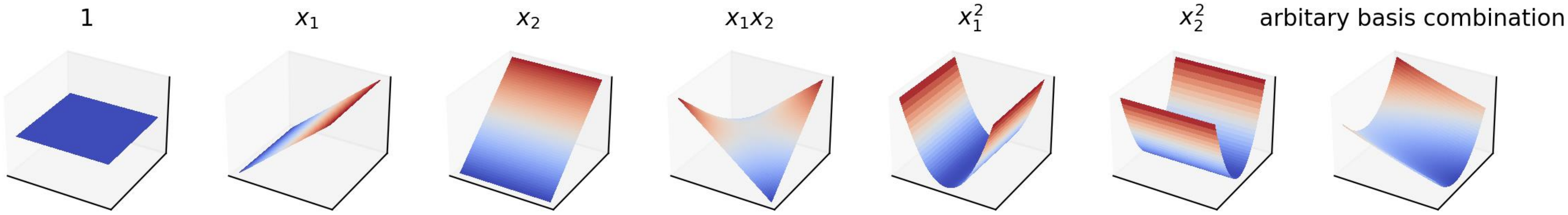


$$\left(1, x_1, x_2, x_1 x_2, x_1^2, x_2^2\right)$$

	1	x_2	${x_2}^2$
1	1	x_2	${x_2}^2$
x_1	x_1	$x_1 x_2$	×
${x_1}^2$	${x_1}^2$	×	×

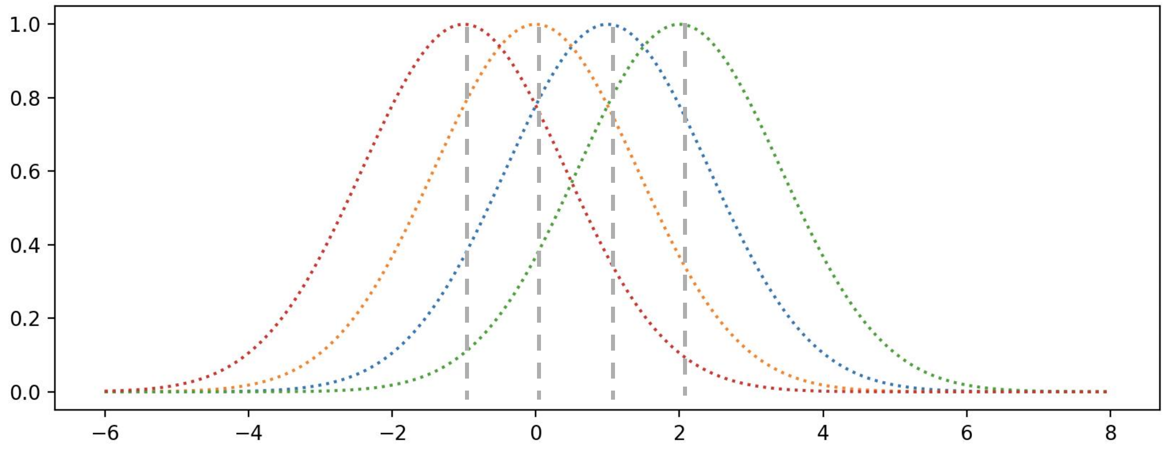
[4] Basis Functions (polynomial expansion: 2d)

[six quadratic polynomial basis functions for 2-d raw data]



[5] Basis Functions (Gaussian basis function/ Radial Basis Function)

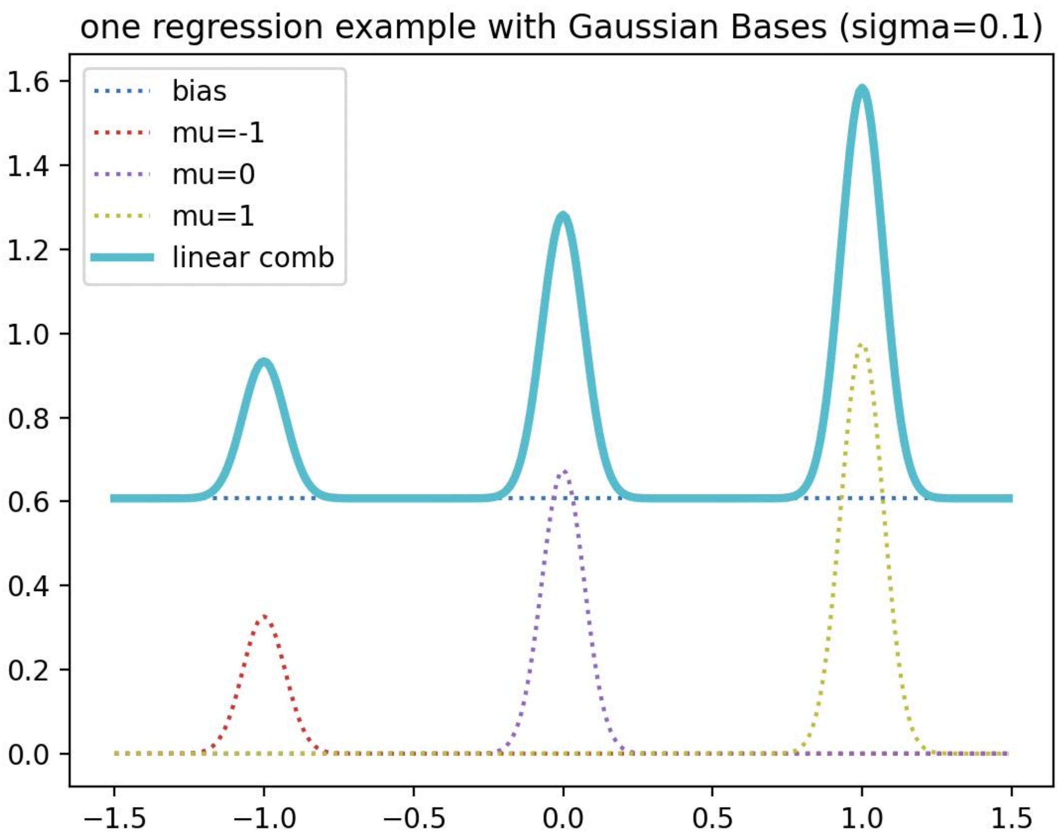
$$\phi_j = \exp\{-\frac{(x-\mu_j)^2}{2\sigma^2}\}$$

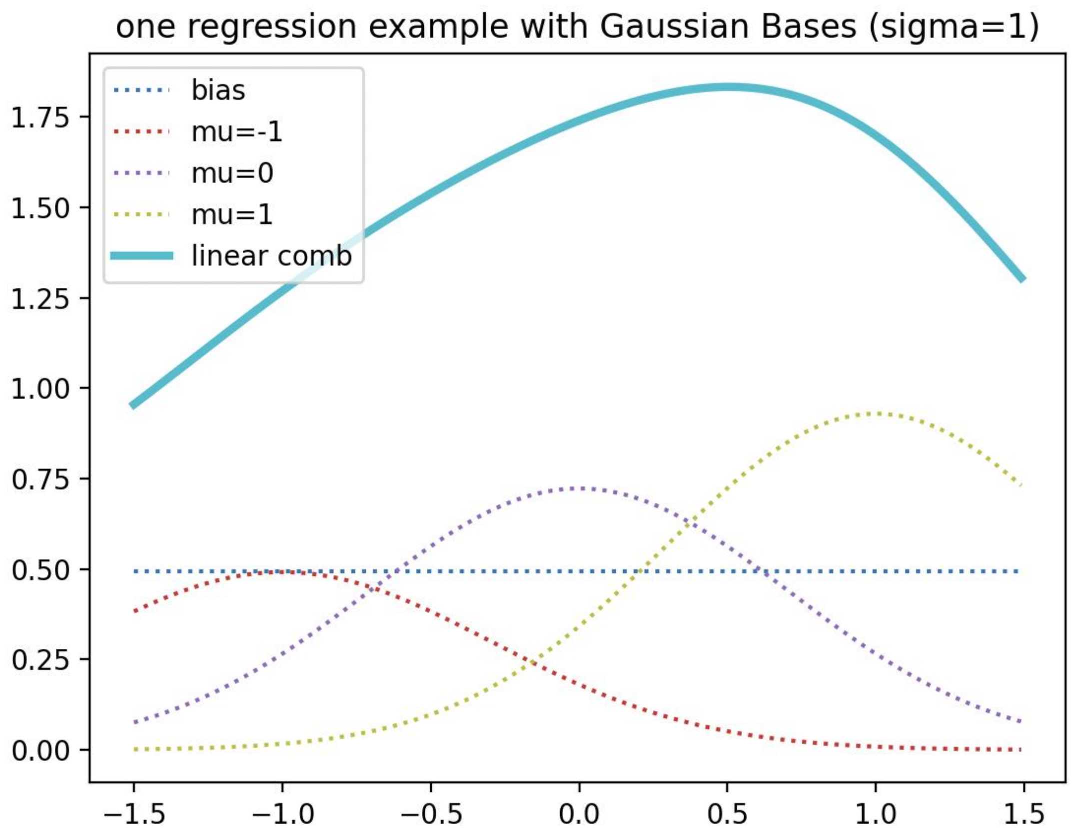


(-1 0 1 2)

Q: the locations of μ_j? (dense / sparse)
Q: the magnitude of σ^2?
(small: local and spiky vs. large: global and smooth)

[6] Basis Functions (Gaussian basis function)





The magnitude of sigma determines the influence over other neighboring Gaussian functions.

[7] Basis Functions (Polynomial vs. Gaussian)

Polynomial	**RBF**
use when data has a global structure	use when data has local structures.
a polynomial affect the target function globally.	a single RBF is in charge of the local prediction.
complexity : (1) # degree of polynomial	complexity : (1) the number of RBFs (2) the magnitude of variance
N/A	▪ dense data: small variance with many RBFs ▪ sparse data: large variance with fewer RBFs

- Linear Regression by MLE

  Learning by Minimum Mean Square Error (MMSE)

[0] Recall slide **: Learning, MLE vs. MAP

Frequentist vs. Bayes Estimation

- $w* = argmax\ P(D|w)$: **M**aximum **L**iklihood **E**stimation (MLE)
- $w* = argmax\ p(w|D) = \frac{p(D|w)p(w)}{p(D)}$: **M**aximum **A** **P**osteriori Estimation (**MAP**)

Frequentist assumes w (parameter) as fixed values and perform MLE to estimate the parameters. MLE can be interpreted as a special case of MAP when the prior density $p(w)$ is uniform.

[1] Linear Regression by MLE (data matrix $\Phi(x)$)

data dimension: D
raw data

data: N

x_1
x_2
…
…
…
…
x_N

feature dimension: M
data matrix $\Phi(x)$

$\phi_1(x_1)$	$\phi_2(x_1)$	…	$\phi_M(x_1)$
$\phi_1(x_2)$	$\phi_2(x_2)$	…	$\phi_M(x_2)$
…	…		
…	…		
…	…		
…	…		
$\phi_1(x_N)$	$\phi_2(x_N)$	…	$\phi_M(x_N)$

[2] Linear Regression by MLE (data)

data dimension: D
raw data

feature dimension: M
data matrix $\Phi(x)$

$\vec{y}$

data: N

x_1
x_2
…
…
…
…
x_N

$\phi_1(x_1)$	$\phi_2(x_1)$	…	$\phi_M(x_1)$
$\phi_1(x_2)$	$\phi_2(x_2)$	…	$\phi_M(x_2)$
…	…		
…	…		
…	…		
…	…		
$\phi_1(x_N)$	$\phi_2(x_N)$	…	$\phi_M(x_N)$

y_1
y_2
…
…
…
…
y_N

[3] Linear Regression by MLE (data density)

$$y = f(x) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$y = \overrightarrow{\Phi(x)}^t \cdot \vec{w} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$P(Y|\Phi, \vec{X}, \sigma^2) \sim \mathcal{N}(\Phi \cdot \vec{W}, \sigma^2)$$

when data samples i.i.d $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

- Q: the distribution of $p\left(\vec{y}\middle|\overrightarrow{w}, \overrightarrow{X}, \Phi\right)$?

[4] Linear Regression by MLE (MLE optimization problem)

$$
\begin{aligned}
\vec{W}* &= \arg\max_{\vec{W}} \prod_{n=1}^{N} P(Y_n|\Phi, \vec{W}, \sigma^2) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}^N} \exp - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (Y_n - \Phi(X_n)^t \vec{W})^2 \\
&= \frac{1}{\sqrt{2\pi\sigma^2}^N} \exp - \frac{1}{2\sigma^2} ||\vec{Y} - \Phi \cdot \vec{W}||^2 \\
\vec{W}* &= \arg\min_{w} ||\vec{Y} - \Phi \cdot \vec{W}||^2
\end{aligned}
$$

MLE becomes
Minimum Mean Square Error Problem.

[5] Linear Regression by MLE (Convex MMSE)

$$J(\vec{W}) = ||\vec{Y} - \Phi \cdot \vec{W}||^2$$
$$\nabla J(W) = -2\Phi^t \cdot \vec{Y} + 2\Phi^t \cdot \Phi \cdot \vec{W}$$
$$\nabla^2 J(W) = 2\Phi^t \cdot \Phi \geq 0$$

- $\mathrm{J}(\overrightarrow{W})$ is convex so the $\overrightarrow{W*}$ $s.t\ \nabla\mathrm{J}(\overrightarrow{W*}) = 0$ will be the optimal solution.

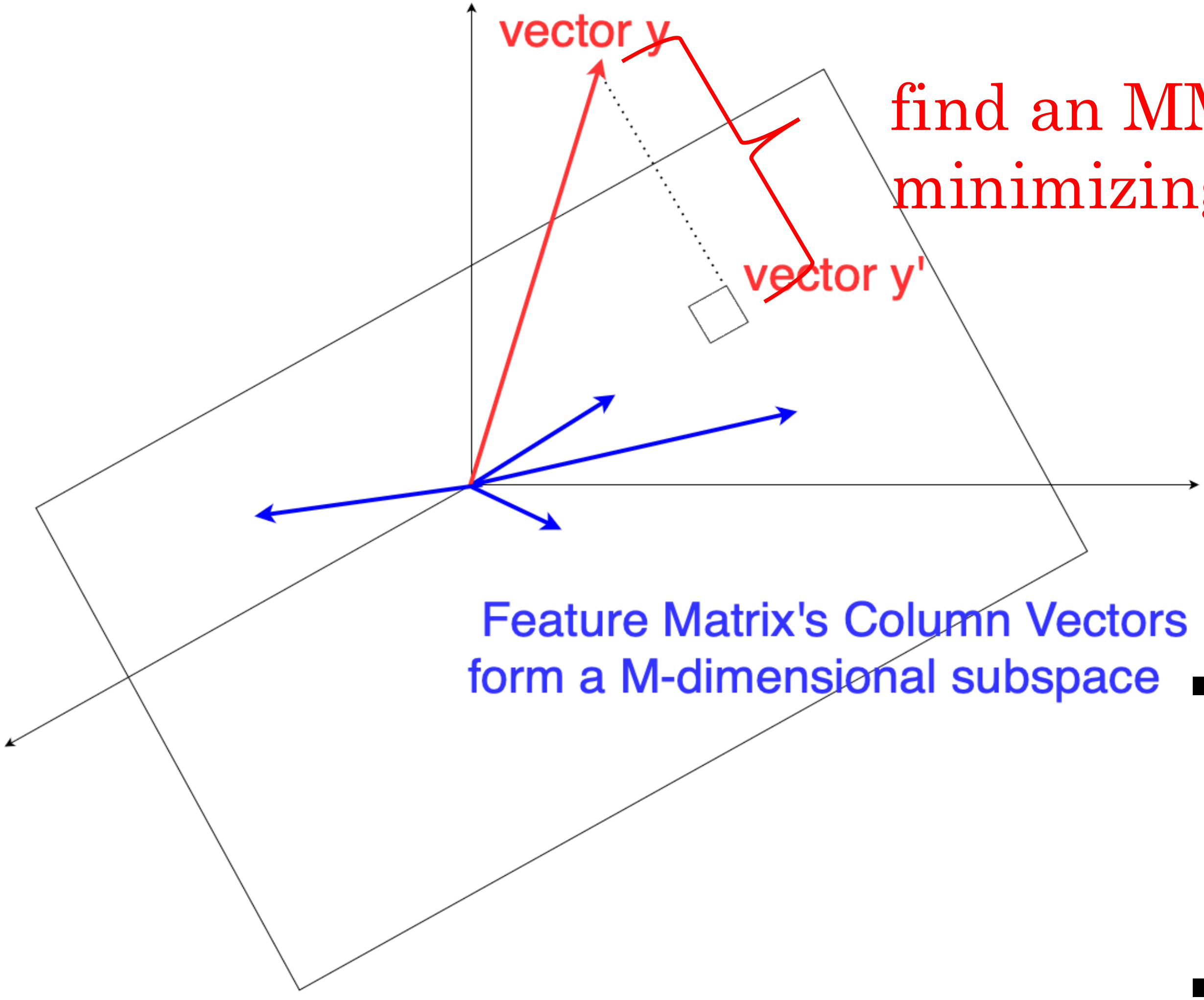
[6] Linear Regression by MLE (Normal Equation)

$$\nabla J(W) = -2\Phi^t \cdot \vec{Y} + 2\Phi^t \cdot \Phi \cdot \vec{W} = 0$$

$$\boxed{\Phi^t \cdot \Phi \cdot \vec{W} = \Phi^t \cdot \vec{Y}}$$

Normal Equation

[7] Linear Regression by MLE (Geometric Interpretation of MMSE)



find an MMSE solution
minimizing $||y - y'||^2$

- the overdetermined system $\Phi w = y$ might not have a solution, so we find an approximated solution instead. ($\Phi^t \Phi w = \Phi^t y$)
- the approximated solution is an MMSE.

[8] Linear Regression by MLE (solving Normal Equation)

$$\Phi(\vec{x}) \cdot \vec{w} = \vec{y}$$

- no solution (over determined equation)

$$\Phi(\vec{x})^t \cdot \Phi(\vec{x}) \cdot \vec{w} = \Phi(\vec{x})^t \cdot \vec{y}$$

- projection to column space (approximated)
- exist solution (one / infinite many solution)

$$\vec{w} = (\Phi(\vec{x})^t \cdot \Phi(\vec{x}))^\dagger \cdot \Phi(\vec{x})^t \cdot \vec{y}$$

- by computing the pseudo-inverse,
  find a solution in the approximated space

[9] Linear Regression by MLE (Spectral Decomposition)

$$
\begin{aligned}
\vec{W}* &= (\Phi^t \cdot \Phi)^\dagger \cdot \Phi^t \cdot \vec{Y} \\
&= V \cdot \Lambda^\dagger \cdot V^t \cdot V \Lambda^{1/2} E^t \vec{Y} \\
&= V \cdot \Lambda^{-1/2} E^t \vec{Y}
\end{aligned}
$$

- Pseudo inverse provides a generalized solution regardless of $\Phi^t \Phi$ is singular / non-singular.

[10] Linear Regression by MLE (Spectral Decomposition)

- invertible (Rank M)

- invertible (Rank M) but close to singular (very small eigenvalues)

- non – invertible (Rank $<M$)

$$\begin{aligned} \vec{W}* &= (\Phi^t \cdot \Phi)^\dagger \cdot \Phi^t \cdot \vec{Y} \\ &= V \cdot \Lambda^\dagger \cdot V^t \cdot V \Lambda^{1/2} E^t \vec{Y} \\ &= V \cdot \Lambda^{-1/2} E^t \vec{Y} \end{aligned}$$

result in very large coefficients

- increase sensitivity to error
- symptom of collinearity
- better to drop one of the high correlated axes

- Linear Regression by MAP

  Learning by Minimum Mean Square Error (MMSE) + Regularization

[0] Recall slide **: Learning, MLE vs. MAP

Frequentist vs. Bayes Estimation

- $w* = argmax\ P(D|w)$: **M**aximum **L**liklihood **E**stimation (MLE)
- $w* = argmax\ p(w|D) = \frac{p(D|w)p(w)}{p(D)}$: **M**aximum **A** **P**osteriori Estimation (**MAP**)

Frequentist assumes w (parameter) as fixed values and perform MLE to estimate the parameters. MLE can be interpreted as a special case of MAP when the prior density $p(w)$ is uniform.

[1] Linear Regression by MAP (optimization problem formulation)

$$\vec{W}* = \arg\max_{\vec{W}} \prod_{n=1}^{N} P(Y_n | \Phi, \vec{W}, \sigma^2) \cdot P(\vec{W})$$

$$= \arg\max_{\vec{W}} \frac{1}{\sqrt{2\pi\sigma^2}^N} \exp - \frac{1}{2\sigma^2} ||\vec{Y} - \Phi \cdot \vec{W}||^2 \cdot \frac{1}{\sqrt{2\pi\lambda}} \exp - \frac{||\vec{W}||^2}{2\lambda}$$

$$= \arg\min_{\vec{W}} \frac{1}{2\sigma^2} ||\vec{Y} - \Phi \cdot \vec{W}||^2 + \frac{||W||^2}{2\lambda}$$

$$= \arg\min_{\vec{W}} ||\vec{Y} - \Phi \cdot \vec{W}||^2 + \frac{||\vec{W}||^2}{\lambda'}$$

$$\lambda' = \frac{\lambda}{\sigma^2}$$

[2] Linear Regression by MAP (optimization problem formulation)

- Regression <u>without</u> prior (MLE)

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

- Regression <u>with</u> prior (MAP)

$$\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2 + \lambda^* (||\vec{w}||^2)$$

$\lambda^* = 1/\lambda$

+ different variances (small variance (w) , large lambda)

[3] Linear Regression by MAP (solving the optimization problem)

$$\nabla J(W) = -2\Phi^t \cdot \vec{Y} + 2\Phi^t \cdot \Phi \cdot \vec{W} + 2\lambda * \cdot \vec{W} = 0$$

$$\leftrightarrow \Phi^t \cdot \Phi \cdot \vec{W} + \lambda * \cdot \vec{W} = \Phi^t \cdot \vec{Y}$$

$$\leftrightarrow V \begin{bmatrix} \lambda_1 + \lambda* & 0 & ... & 0 \\ 0 & \lambda_2 + \lambda* & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & ... & \lambda_m + \lambda* \end{bmatrix} V^t \cdot \vec{W} = V^t \lambda^{1/2} E^t \vec{Y}$$

$$\leftrightarrow \vec{W} = V \begin{bmatrix} \frac{1}{\lambda_1 + \lambda*} & 0 & ... & 0 \\ 0 & \frac{1}{\lambda_2 + \lambda*} & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & ... & \frac{1}{\lambda_m + \lambda*} \end{bmatrix} E^t \vec{Y}$$

[by the λ^* we can avoid the case parameters gets too large.]

[4] Linear Regression by MAP

$$\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2 + \lambda^*(||\vec{w}||^2)$$

The MMSE with regularization can be
Translated into convex optimization problem.

[5] Linear Regression by MAP (as an optimization problem)

- regression <u>without</u> constraint

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

- regression <u>with</u> constraint (regularization)

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

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

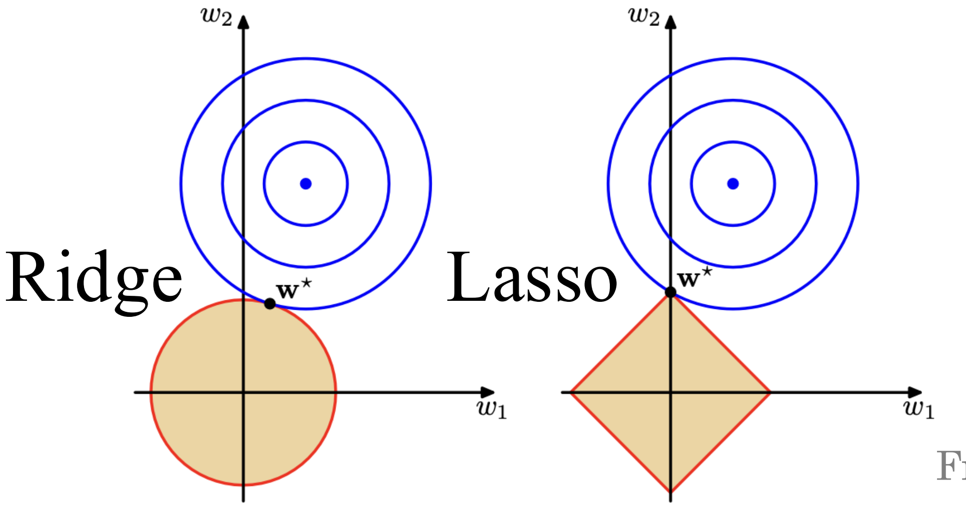
$$\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2 + \lambda^*(||\vec{w}||^2)$$

(Lagrangian form of constrained MMSE objective)

[6] Linear Regression by MAP (Ridge and Lasso Regression)

$$\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2 + \lambda^*(||\vec{w}||^2)$$ [Ridge regularization]

$$\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2 + \lambda^*(||\vec{w}||)$$ [Lasso regularization]



- the constraints regulate

the magnitude of *w* (parameters), so the model complexity. Lasso gives a sparse solution.

From Bishop Chap Figure 3.4

- Optimization Theory:

  Solving a convex optimization problem by using a Langrangian function

[1] Local and Global Minimum (why optimization theory?)

In an ML problem, we need to solve an optimization problem, finding local / global minimum (suboptimal/optimal): MLE / MAP

- regression <u>without</u> constraint

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

- regression <u>with</u> constraint (regularization)

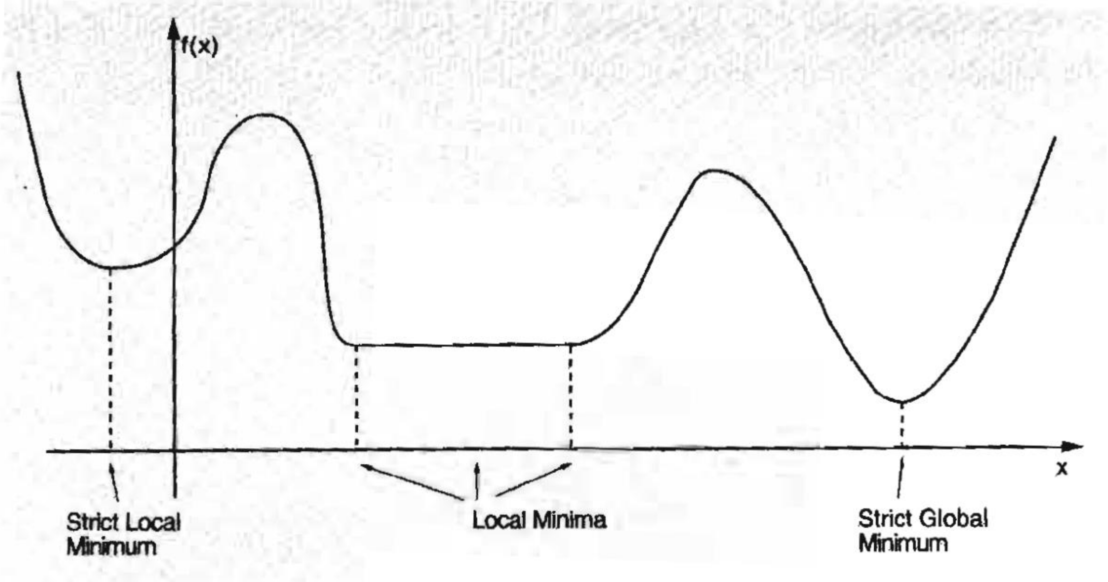
$$\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2$$
$$\text{subject to} \quad ||\vec{w}||^2 \leq C$$

Local Minimum x^*

$$f(x*) \leq f(x), \quad \exists \epsilon \quad s.t \quad ||x - x*|| < \epsilon \quad \forall x$$

Global Minimum x^*

$$f(x*) \leq f(x) \quad \forall x$$



[3] Local and Global Minimum (necessary conditions for local minimum)

- By Taylor series — if x^* is a local optimal, then the Taylor approximation is non-negative:

$$f(x*+\Delta x) - f(x*) \approx \nabla f(x*)^t \Delta x + \frac{1}{2}\Delta x^t \nabla^2 f(x*) \Delta x \geq 0$$

- Two necessary conditions for optimality

$$\nabla f(x*) = 0$$

$$\Delta x^t \boxed{\nabla^2 f(x*)} \Delta x^t \geq 0$$

[Hessian matrix at x^* is locally positive semidefinite: a convex /ball shape]

- Equality Constraint Problem

[1] Equality Constraint Problem (example)

$$\min_x \quad f(x)$$
$$\text{s.t.} \quad h_i(x) = 0 \quad i = 1, ..., m$$

ex] $$\min_x \quad x_1 + x_2$$
$$\text{s.t.} \quad x_1^2 + x_2^2 = 2$$

[2] Equality Constraint Problem (necessary conditions)

$$\min_x \quad f(x)$$
$$\text{s.t.} \quad h_i(x) = 0 \quad i = 1, ..., m$$

- **Condition1**: let x^* be a local minimum of f s.t $h_i(x) = 0$ and $\nabla h_i(x^*)\ldots \nabla h_i(x^*)$ are linearly independent. then, there exist a unique vector $\lambda^* = (\lambda_1^*, \lambda_2^*, \ldots. \lambda_m^*)$ s.t

$$\nabla f(x*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x*) = 0$$

- **Condition2**: $y^t\{\nabla^2 f(x*) + \sum_{i=1}^{n} \lambda_i^* \nabla^2 h_i(x*))\}y \geq 0 \quad y \in V(x*)$

$$V(x*) = \{y | \nabla h_i(x*)^t y = 0 \quad \forall i = 1, ..., m\}$$

[3] Equality Constraint Problem (Lagrangian Multiplier Theorem)

$$\min_x \quad f(x)$$
$$\text{s.t.} \quad h_i(x) = 0 \quad i = 1, \ldots, m$$

- Condition1: let x^* be a local minimum of f s.t $h_i(x) = 0$ and $\nabla h_i(x^*) \ldots \nabla h_i(x^*)$ are linearly independent. then, there exist a unique vector $\lambda^* = (\lambda_1^*, \lambda_2^*, \ldots \lambda_m^*)$ s.t

$$\nabla f(x*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x*) = 0$$

- Condition2: $y^t \{\nabla^2 f(x*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x*))\} y \geq 0 \quad y \in V(x*)$

$$V(x*) = \{y | \nabla h_i(x*)^t y = 0 \quad \forall i = 1, \ldots, m\}$$

Q: What if we define a new function? $L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x)$

[4] Equality Constraint Problem (Lagrangian function)

- Lagrangian function/ unconstrained function

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x)$$

- two necessary optimality conditions for $L(x,\lambda)$

$$\begin{cases} \nabla f(x*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x*) = 0 \qquad h_i(x*) = 0 \quad \forall i = 1, 2..., m \\ \\ y^t\{\nabla^2 f(x*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x*))\}y \geq 0 \quad y \in V(x*) \end{cases}$$

$$V(x*) = \{y | \nabla h_i(x*)^t y = 0 \quad \forall i = 1, ..., m\}$$

[5] Equality Constraint Problem (Lagrangian function)

[Lagrangian Function/ unconstrained function]

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i \, h_i(x)$$

The necessary conditions for the unconstrained function (Lagrangain) gives the optimal solutions to the original constrained problem.
Therefore, we solve the necessary conditions of the Lagrangian function.

[6] Equality Constraint Problem (Lagrangian example)

Consider the problem

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2}\left(x_1^2 + x_2^2 + x_3^2\right) \\ &\text{subject to} \quad x_1 + x_2 + x_3 = -3. \end{aligned}$$

Q: Lagrangian function?

- Inequality Constraint Problem

[1] Inequality Constraint Problem (necessary conditions)

$$
\begin{aligned}
\min_{x} \quad & f(x) \\
\text{s.t.} \quad & g_i(x) \leq 0 \quad i = 1, ..., m
\end{aligned}
$$

* pay attention! the direction of inequality!

[2] Inequality Constraint Problem (necessary conditions)

$$\min_x \quad f(x)$$
$$\text{s.t.} \quad g_i(x) \leq 0 \quad i = 1, ..., m$$

- two possible cases for x^*

  (1) x^* inside of the manifold by $g_i(x) < 0$

  (2) x^* on the boundary of $g_i(x) = 0$

[3] Inequality Constraint Problem (necessary conditions)

$$\min_x \quad f(x)$$
$$\text{s.t.} \quad g_i(x) \leq 0 \quad i = 1, \ldots, m$$

- two possible cases and their necessary conditions

(1) x^* inside of the manifold by $g_j(x) < 0$

$$\rightarrow \nabla f(x*) = 0$$

(2) x^* on the boundary of $g_h(x) = 0$ there exist λ_h

X

$$\rightarrow \nabla f(x*) + \lambda_h \nabla g_h(x*) = 0$$

Q: sign λ_h ?

[4] Inequality Constraint Problem (KKT necessary conditions)

Let x^* be a local minimum of the problem

[**K**arush–**K**uhn–**T**ucker conditions]

$$\min_x \quad f(x)$$
$$\text{s.t.} \quad g_i(x) \leq 0 \quad i = 1, \ldots, m$$

important to set the inequality this form (less than or equal to)!

Then, there exist λ_i, $i = 1, \ldots, m$ such that

- stationary condition

(1) $$\nabla f(x*) + \sum_{i=1}^{m} \lambda_i \nabla g_m(x*) = 0$$

- complementary slackness condition $\lambda_i \cdot g_i(x^*) = 0$

(2) $$\begin{cases} \lambda_j \geq 0 & j = 1, \ldots, r \\ \lambda_j = 0 & \forall j \notin A(x*) \end{cases}$$

$A(x^*)$ is the set of active constraints at x^*

[5] Inequality Constraint Problem (KKT necessary conditions)

Let x^*be a local minimum of the problem

$$\min_x \quad f(x)$$
$$\text{s.t.} \quad g_i(x) \leq 0 \quad i = 1, \ldots, m$$

Then, there exist $\lambda_i,\ i = 1, \ldots, m$ such that

(1) $\nabla f(x*) + \sum^{m} \lambda_i \nabla g_m(x*) = 0$

(2) $\begin{cases} \lambda_j \geq 0 & j = 1, \ldots, r \\ \lambda_j = 0 & \forall j \notin A(x*) \end{cases}$

(3) $g(x^*) \leq 0$

- Primary feasibility

[6] Inequality Constraint Problem (example 1)

Consider the problem

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2}\left(x_1^2 + x_2^2 + x_3^2\right) \\ &\text{subject to} \quad x_1 + x_2 + x_3 \leq -3. \end{aligned}$$

Then for a local minimum x^*, the first order necessary condition [cf. Eq. (3.47)] yields

$$x_1^* + \mu^* = 0,$$

$$x_2^* + \mu^* = 0,$$

$$x_3^* + \mu^* = 0.$$

From Nonlinear Programming, Bertsekas Example 3.3.1

[7] Inequality Constraint Problem (example 2)

ex] solve the two-dimensional problem

$$\begin{aligned} \min_{x} \quad & (x-1)^2 + (y-1)^2 + xy \\ \text{s.t.} \quad & 0 \le x \le 1, \qquad 0 \le y \le 1 \end{aligned}$$

this problem will be covered during recitation.

- Regularization as an optimization problem

$$\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2 + \lambda^*(||\vec{w}||^2)$$

[1] Regularization by an optimization problem

In a ML problem, we need to solve an optimization problem, finding local / global minimum (suboptimal/optimal).

- regression without constraint

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

- regression with constraint (regularization)

$$\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2$$
$$\text{subject to} \quad ||\vec{w}||^2 \leq C$$

[2] Regularization by an optimization problem (Lagrangian form)

$$\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2 \quad \text{subject to} \quad ||\vec{w}||^2 \leq C \quad \longleftrightarrow \quad \arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2 + \lambda^*(||\vec{w}||^2 - C)$$

- according to C we define,
  optimal Lagrangian $\lambda *$ will be different!

- constant addition/subtraction won't change x^*

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

[3] Regularization by an optimization problem (Lagrangian form)

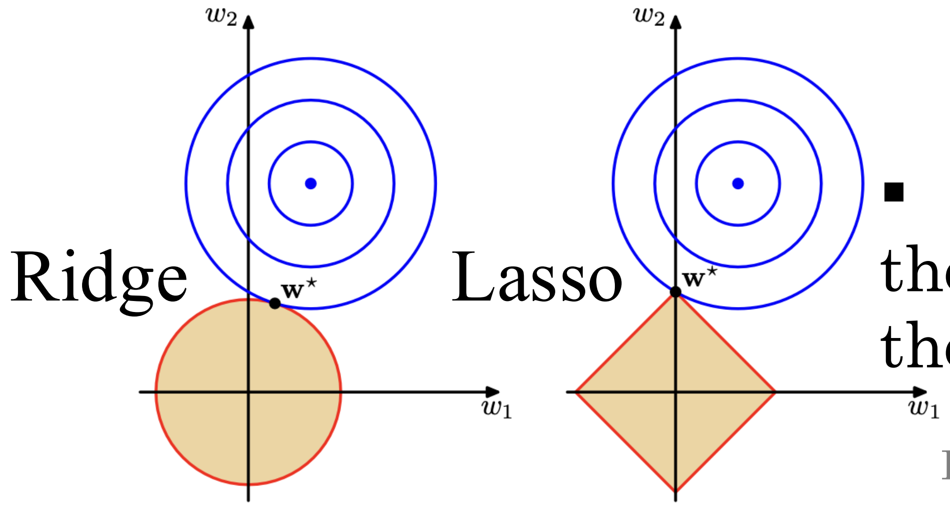
$$\left\{\begin{aligned} &\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2 \\ &\quad \text{subject to} \quad ||\vec{w}||^2 \leq C \\ \\ &\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2 + \lambda^*(||\vec{w}||^2) \end{aligned}\right.$$

- in regularized regression learning, we will change λ^* and test its performance to find a good λ^* (empirically)

[4] Regularization by an optimization problem (Ridge & Lasso)

$$\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2 + \lambda^*(||\vec{w}||^2)$$ [Ridge regularization]

$$\arg\min_{\vec{w}} ||\vec{y} - \Phi \cdot \vec{w}||^2 + \lambda^*(||\vec{w}||)$$ [Lasso regularization]



- the constraints regulate
the magnitude of *w* (parameters),
the model complexity. Lasso gives a sparse solution.

From Bishop Chap Figure 3.4