CS 461: Machine Learning Principles

Class 6: Sept. 23

Error Decomposition & Regularized Regression

(Optimization Theory)

Instructor: Diana Kim

Regression in the Last Class

Summary: Train and Test Procedures for Regression Problem

- A set of hypothetical basis functions (polynomial / Gaussian)
- We have data points $((d_1, y_1), (d_2, y_2), (d_3, y_3), \ldots, ((d_N, y_N))$

Train

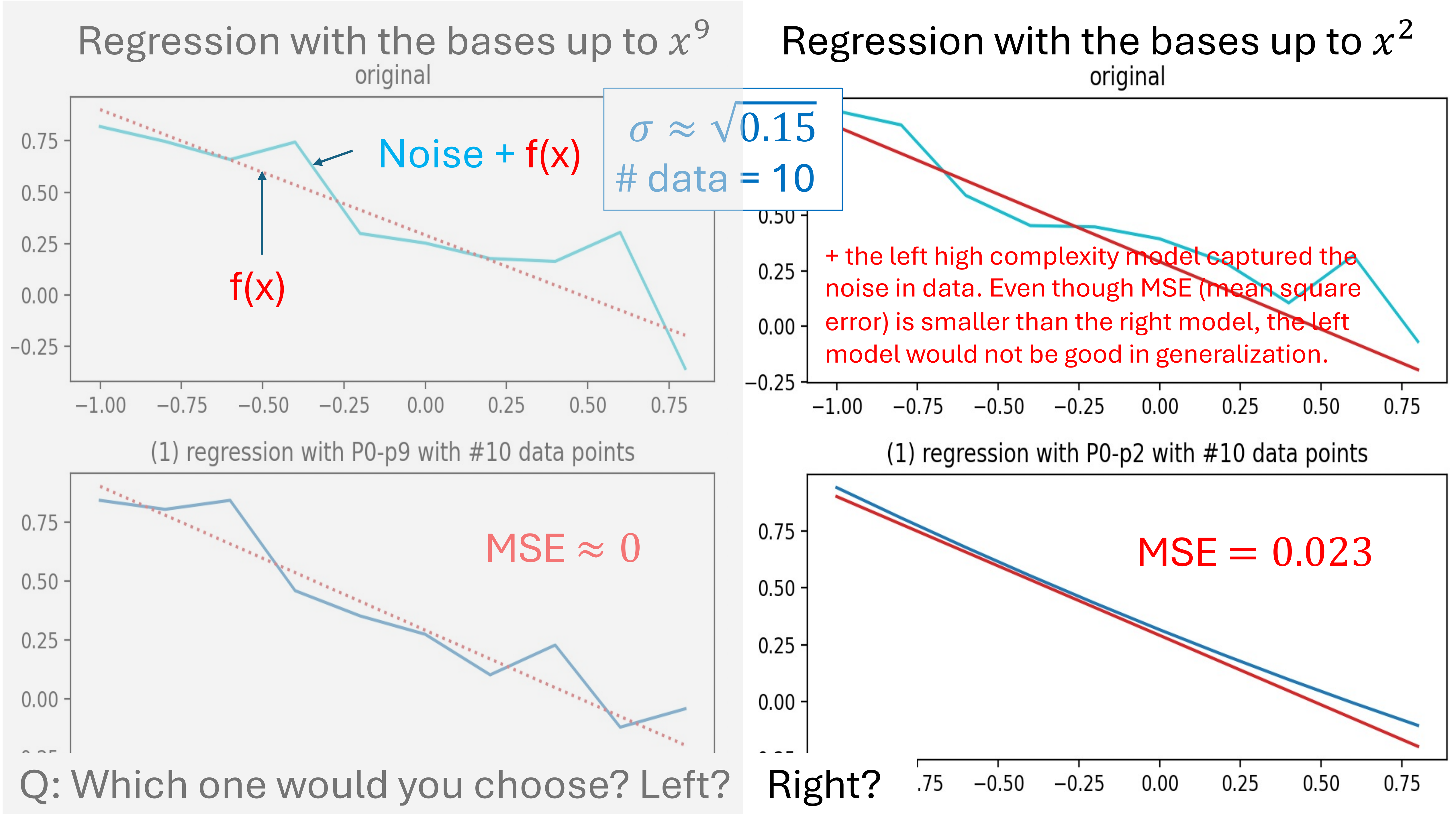
- Train data PCA for dim-reduction (high-dimensional data) and whitening. (save the PCA blocks computed with training set)
- Normal equation to estimate $\vec{w}$

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

Test

- Test data PCA for dim-reduction and whitening with the same block used in training!!!! (important)
- Compute error between the groud turth y and prediction y': $\|y - y'\|^2$

Overfitting



Regression with the bases up to x^9
original
Noise + f(x)
f(x)
$\sigma \approx \sqrt{0.15}$
data = 10
(1) regression with P0-p9 with #10 data points
MSE ≈ 0
Regression with the bases up to x^2
original
+ the left high complexity model captured the noise in data. Even though MSE (mean square error) is smaller than the right model, the left model would not be good in generalization.
(1) regression with P0-p2 with #10 data points
MSE = 0.023
Q: Which one would you choose? Left? Right?

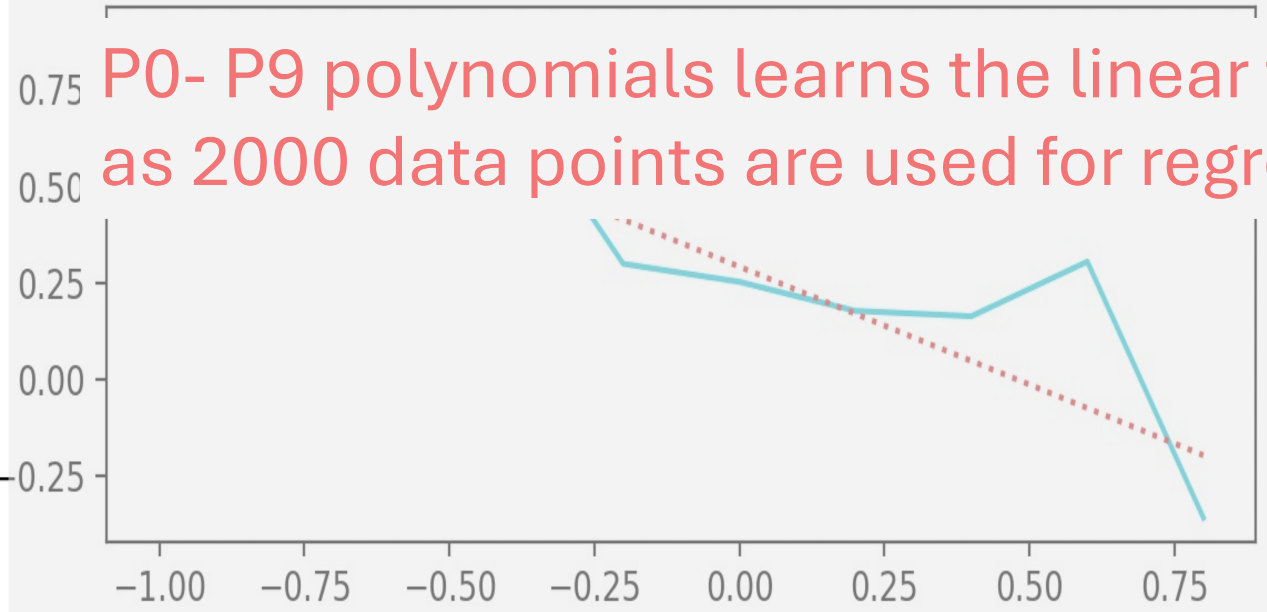
- **Overfitting**

When model complexity is too high relatively to # training data, then the complex model fits to noise. One phenomenon for overfitting we can observe is the large gap between train vs. test. For example) train error ≈ 0

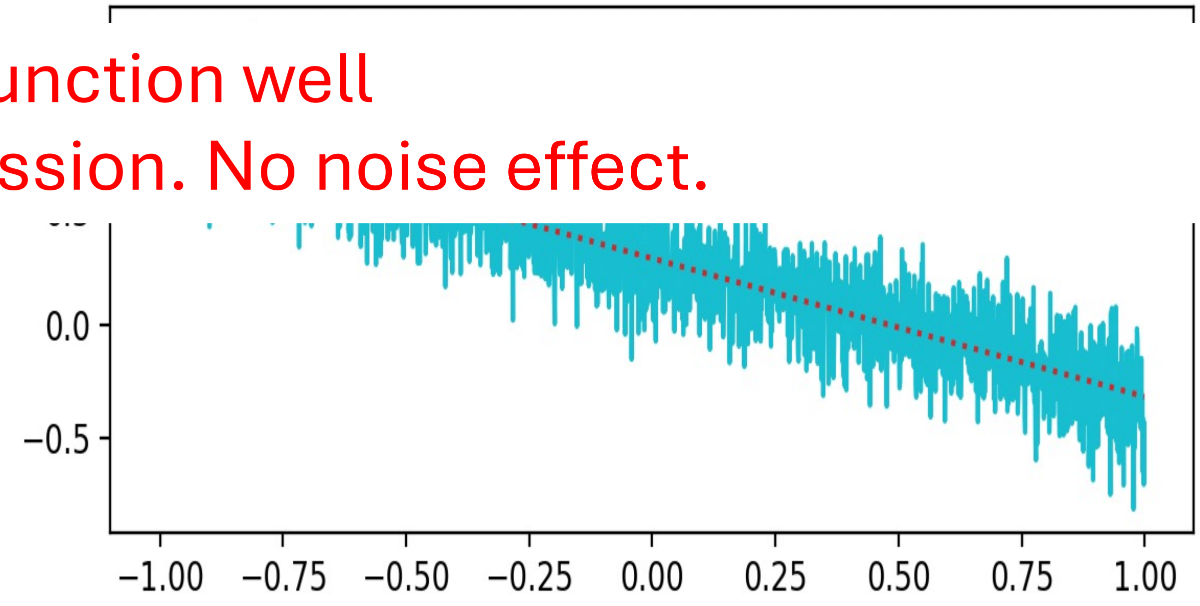
Q:
why if we have enough training data the overfitting phenomenon disappear?
$\sum_{i=1}^{\infty} \varepsilon_i = 0$ noise effect diminishes.

P0- P9 polynomials learns the linear function well as 2000 data points are used for regression. No noise effect.

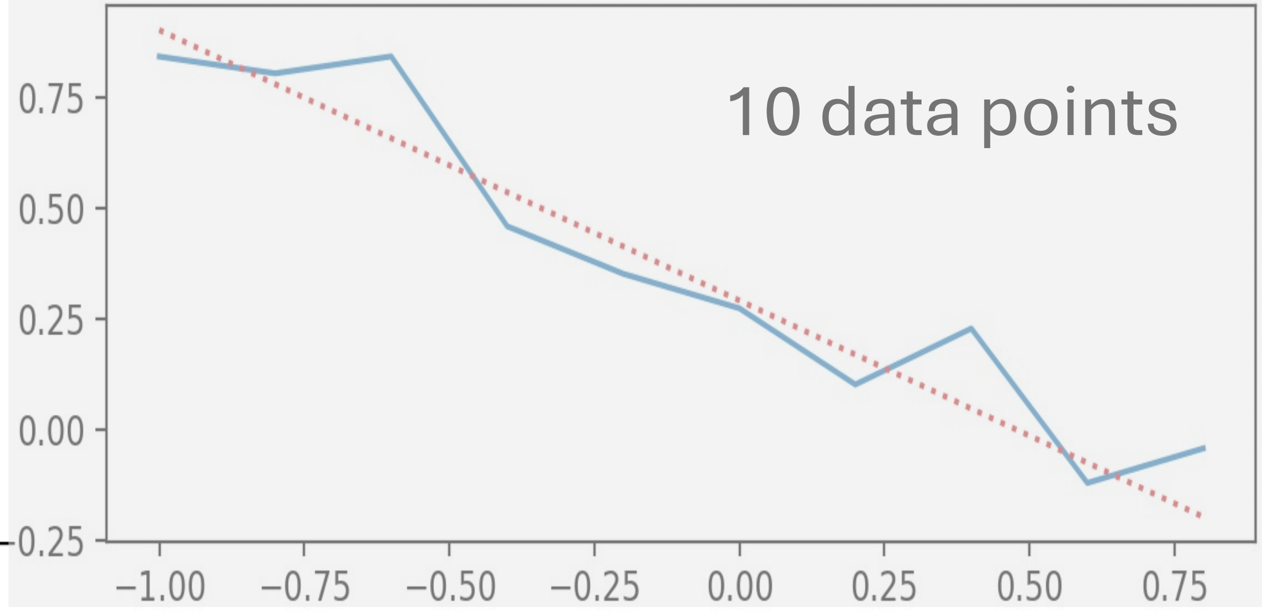
original



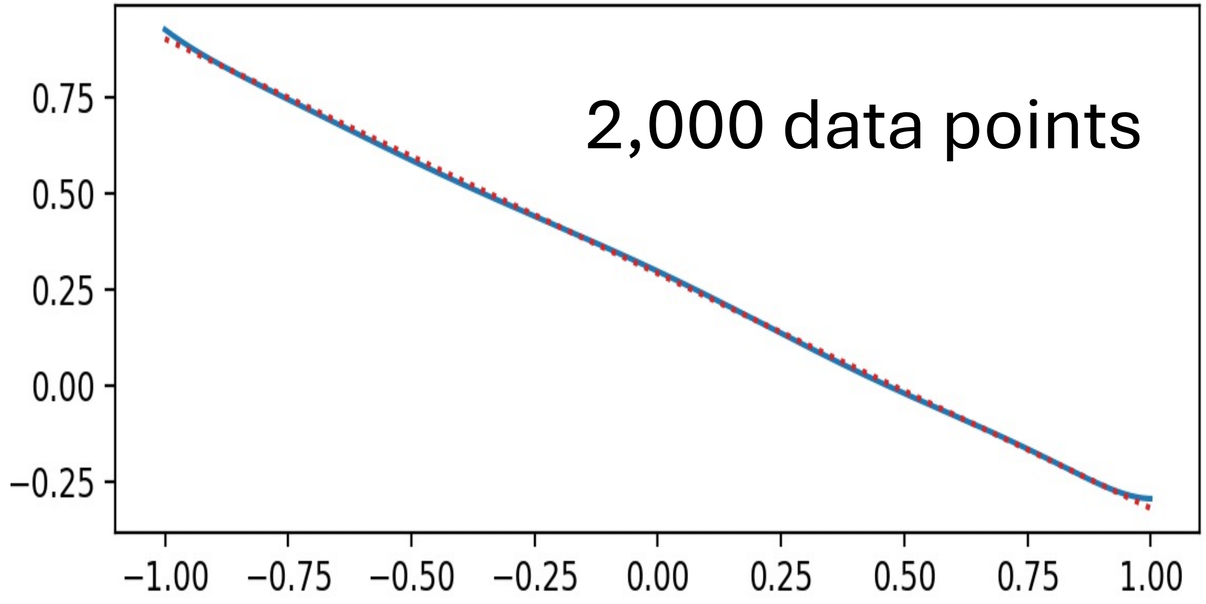
original



(1) regression with P0-p9 with #10 data points



(1) regression with P0-p9 with #2000 data points



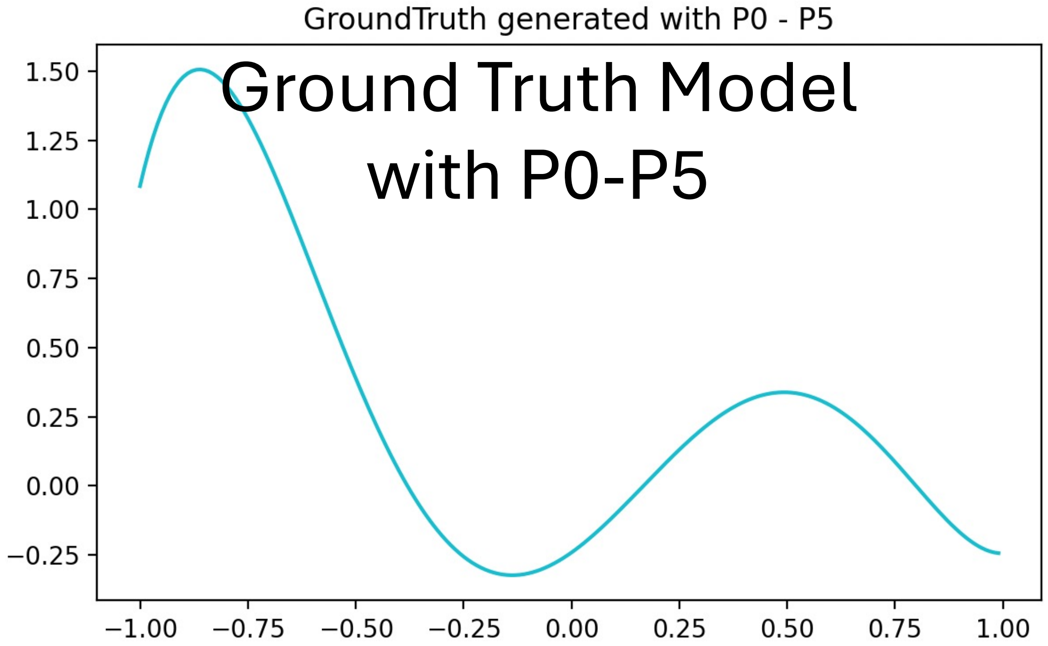
- **Overfitting**

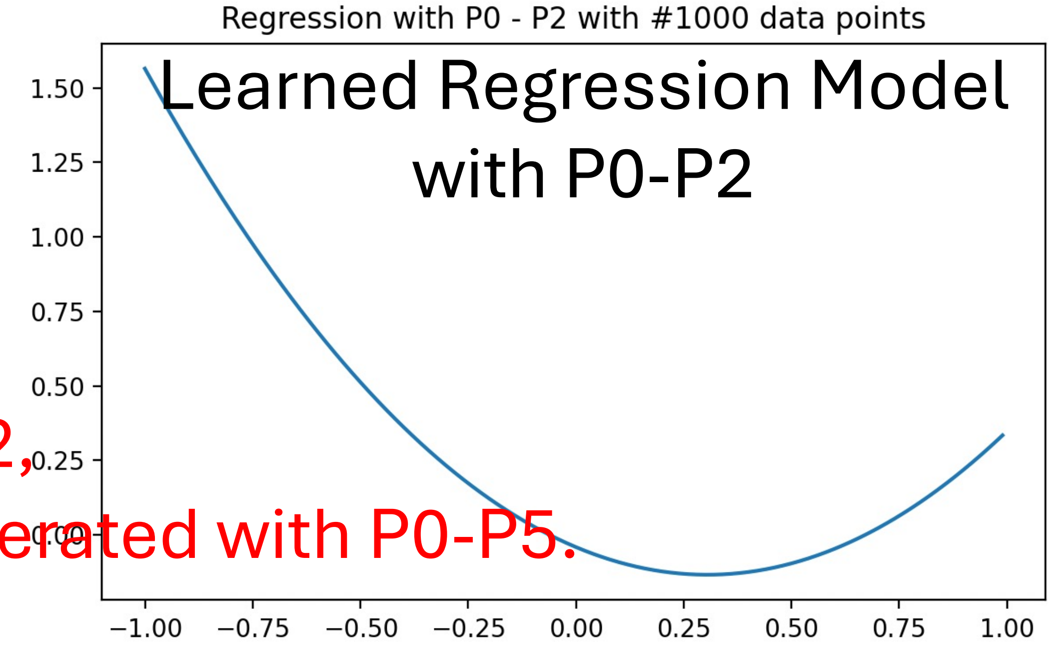
When model complexity is too high relatively to # training data,
then the complex model fits to noise. One phenomenon for overfitting we can observe is the large gap between train vs. test. For example) train error $\approx$ 0

Q:
Then, how could we prevent the overfitting as we have limited data?
Do we need to reduce the number of features?
this can be a way but need to check the underfitting possibility.

- **Underfitting**

When model complexity
is not enough for the ground truth model,
no way to learn even when we have
enough data. (no hope to learn)





Underfitting Example:
data is 1,000 but
as the bases functions are limited by P0-P2,
no way to capture the original function generated with P0-P5.

Q: Then, how could we avoid the overfitting as we have limited data?
Do we need to reduce the number of features?

We can maintain the complexity to avoid underfitting.
Instead, we can add “Regularization Block”.

Q: Then, how could we avoid the overfitting as we have limited data?
Do we need to reduce the number of features?

We can maintain the complexity to avoid underfitting.
Instead, we can add “Regularization Block”.

+ It will limit the effective model complexity.”
+ It will allow complex models to be trained on data sets of limited size
without sever overfitting."

Regularized Regression

Formulation of Constrained Regression to Control Model Complexity

$$||\vec{w}|| \leq C \quad or \quad ||\vec{w}||^2 \leq C$$

Formulation of Constrained MMSE Objective

- Regression <u>without</u> any constraint

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

- Regression <u>with</u> the constraints (regularizations)

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

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

Lagrangian Form of Constrained MMSE Objective

- Regression <u>without</u> constraint

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

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

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

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

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

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

Computing the Optimal Solution

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

- the optimal $\vec{w}$ is computed by $\nabla J(\vec{w}) = 0$ for the strict convexity of the objective function J(w)

$$J(\vec{w}) = ||\vec{y} - \Phi \cdot \vec{w}||^2$$

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

$$\nabla J(\vec{w}) = -2 \cdot \Phi^t \cdot (\vec{y} - \Phi \cdot \vec{w}) = 0$$

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

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

Computing the Optimal Solution

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

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

$$= \left\| \left(\begin{bmatrix} \Phi \\ \sqrt{\lambda * I} \end{bmatrix} \cdot \vec{w} - \begin{bmatrix} \vec{y} \\ 0 \end{bmatrix} \right) \right\|^2$$

Same !
Just we have an extended data matrix.

$$\begin{bmatrix} \Phi \\ \sqrt{\lambda * I} \end{bmatrix}^t \cdot \begin{bmatrix} \Phi \\ \sqrt{\lambda * I} \end{bmatrix} \cdot \vec{w} = \begin{bmatrix} \Phi \\ \sqrt{\lambda * I} \end{bmatrix}^t \cdot \begin{bmatrix} \vec{y} \\ 0 \end{bmatrix}$$

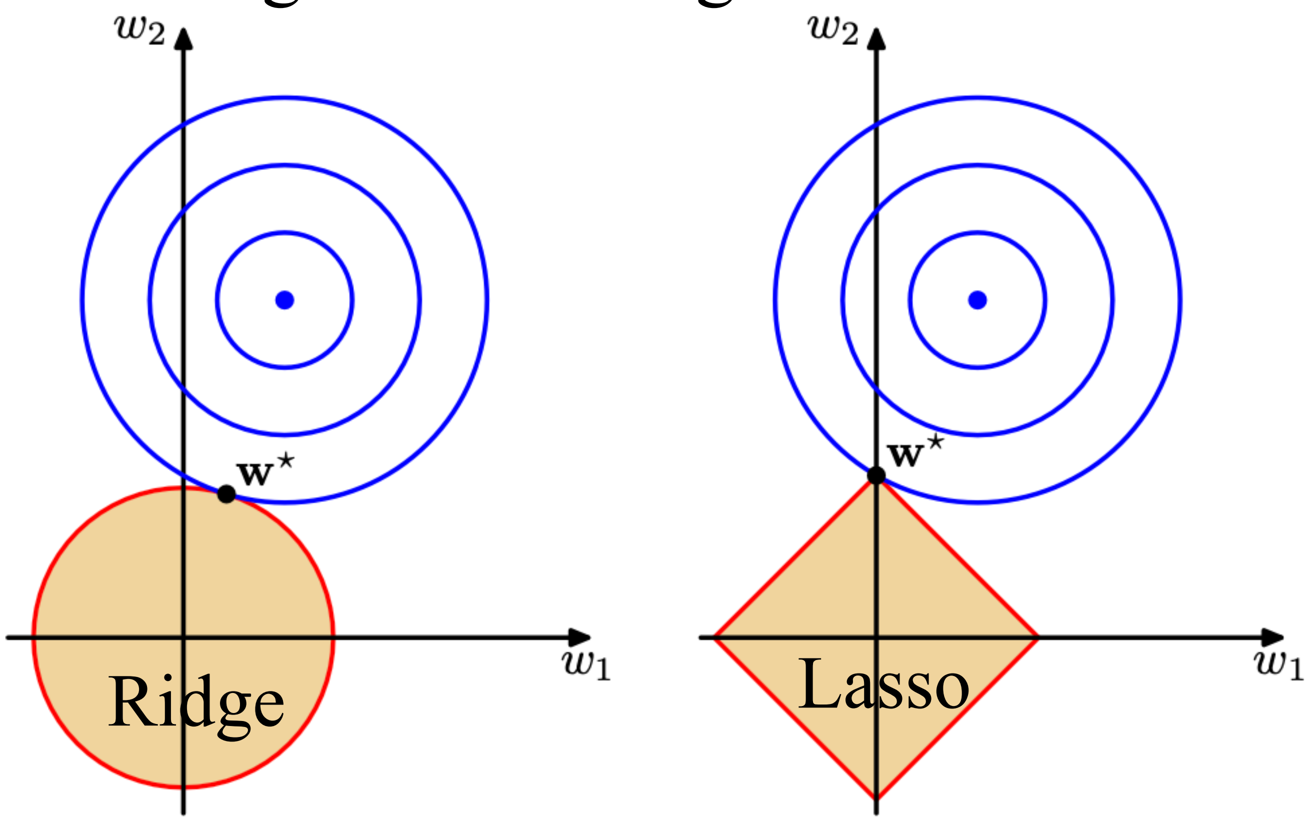
$$(\Phi^t \cdot \Phi + \lambda^* \cdot I)\vec{w} = \Phi^t \cdot \vec{y}$$

$$(V \Lambda V^t + V \lambda^* I V^t)\vec{w} = \Phi^t \cdot \vec{y}$$

Q: Do we need to worry about the singular case?
the matrix is invertible! (all positive eigenvalues)

$$(V \begin{bmatrix} \lambda_1 + \lambda^* & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 + \lambda^* & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \lambda_M + \lambda^* \end{bmatrix} V^t)\vec{w} = \Phi^t \cdot \vec{y}$$

Geometric Interpretation of Ridge / Lasso Regression



- as λ getting bigger
  the constraint range getting smaller!
- Q: which one gives a sparse solution?

From Bishop Chap Figure 3.4

+ the constraints regulate the magnitude of W, so the model complexity. Lasso gives sparse solution.

Summary: Train, Val, Test Procedures for **Ridge** Regression Problem

- A set of hypothetical basis functions (polynomial / Gaussian)
- We have data points $((d_1, y_1), (d_2, y_2), (d_3, y_3), \ldots, ((d_N, y_N))$

Train

- Train data PCA for dim-reduction (high-dimensional data) and whitening. (save the PCA blocks computed with training set)
- Normal equation with various $\lambda *$ to estimate W

Val

+ we should not use test data to determine a regularization parameter.

- Val data PCA for dim-reduction and whitening with the same block used in training.
- Compute error between the groud turth y and prediction y': $||y - y'||^2$
- Choose the best $\lambda *$

Summary: Train, Val, Test Procedures for **Ridge** Regression Problem

Train

- Train data PCA for dim-reduction (high-dimensional data) and whitening. (save the PCA blocks computed with training set)
- Normal equation with various $\lambda *$ to estimate W

Val

- Val data PCA for dim-reduction and whitening with the same block used in training.
- Compute error between the groudturth y and prediction y': $\|y - y'\|^2$
- Choose the best $\lambda *$

Test

- Val data PCA for dim-reduction and whitening with the same block used in training.
- Compute error between the groudturth y and prediction y': $\|y - y'\|^2$

Cross Validation
Feature/ Model Selection ($\lambda *$)

+ this will be covered in the next class.

Optimization Theory

Local and Global Minimum

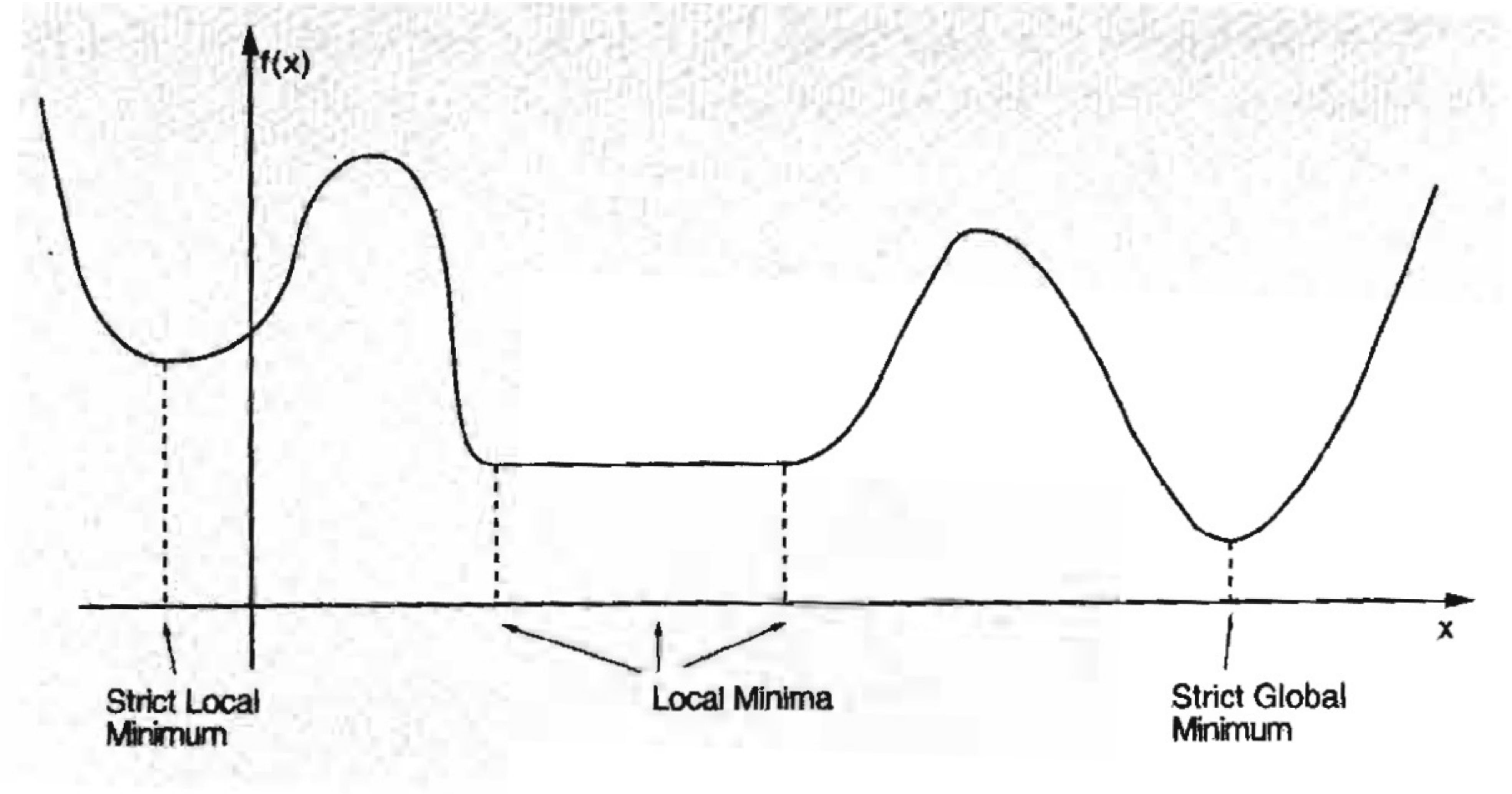
The Definition of <u>Local</u> Minimum $x*$

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

The Definition of <u>Global</u> Minimum $x*$

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

Strict Local and Global Minimum



The Necessary Conditions for Optimality

- approximation by Taylor series

$if\ x^*\ is\ a\ local\ minima$

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

- Two Necessary Condition for optimality

$$\begin{cases}\nabla f(x*)=0\\ \Delta x^t\nabla^2 f(x*)\Delta x^t\geq 0\end{cases}$$

positive semi-definite

The Sufficient Conditions for Optimality

- Two Sufficient Condition for Optimality

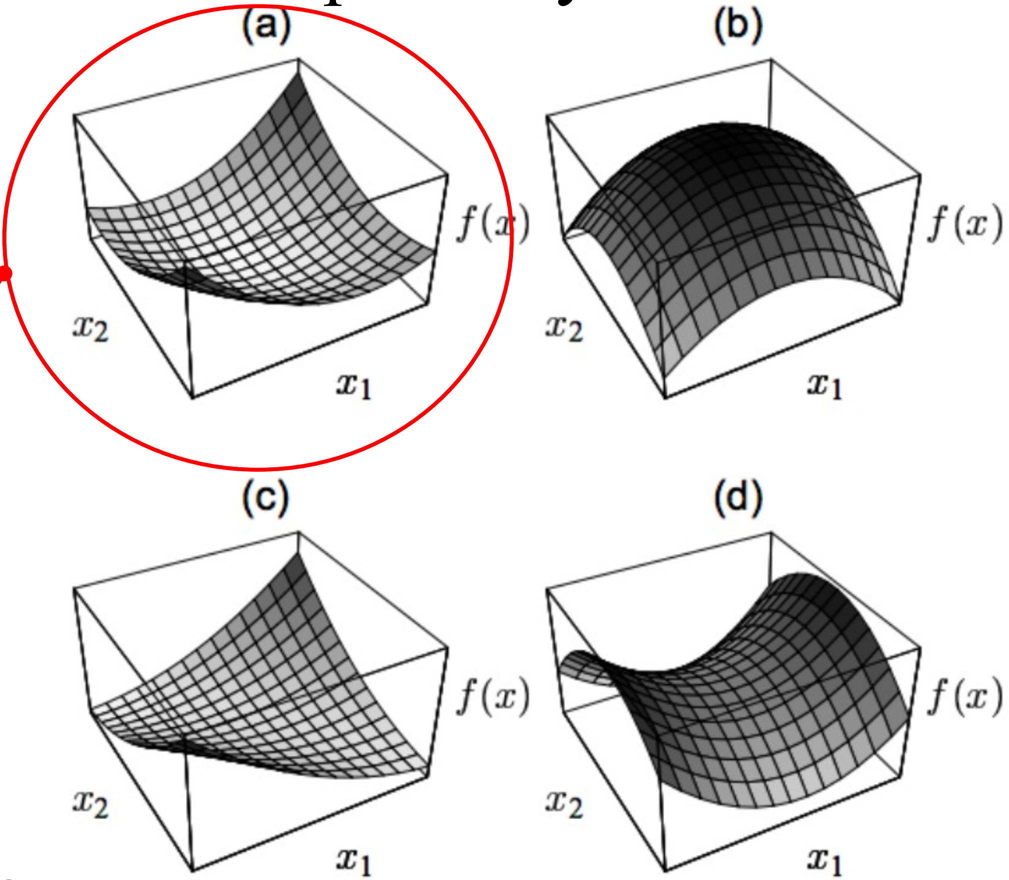
$$\begin{cases} \nabla f(x*) = 0 \\ \Delta x^t \nabla^2 f(x*) \Delta x^t > 0 \end{cases}$$

Hessian is positive definite:
for any direction, f strictly increases!
x^* is strict local minimum.

For the case (a): positive definite,

$\nabla f(x*) = 0$ becomes sufficient for optimality.

positive definite



From Bishop Chap. 8

Figure 8.6: The quadratic form $f(\boldsymbol{x}) = \boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}$ in 2d. (a) $\mathbf{A}$ is positive definite, so f is convex. (b) $\mathbf{A}$ is negative definite, so f is concave. (c) $\mathbf{A}$ is positive semidefinite but singular, so f is convex, but not strictly. Notice the valley of constant height in the middle. (d) $\mathbf{A}$ is indefinite, so f is neither convex nor concave. The stationary point in the middle of the surface is a saddle point. From Figure 5 of [She94].

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 function to $f(x)$: 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-Khun-Tucker Necessary condition
KKT condition defines $x*$ in relation to a certain $\lambda*$.

- Optimization problem

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

- Then, there exist unique Lagrangian multiplier $\lambda*$ s.t

$$\boxed{\nabla_x L(x*, \lambda*) = 0}$$

$$\lambda* = 0 \quad \text{if} \quad g(x*) < 0$$

$$\lambda* = \text{positive} \quad \text{if} \quad g(x*) = 0$$

This is the necessary condition for the minimum of $L(x, \lambda*)$.

The original optimization problem can be solve by minimizing its Lagrangain function.

$$x* = \arg\min_x f(x)$$

$$\text{subject to} \quad g(x) \leq 0$$

$$\Updownarrow$$

Two optimization problems will fine the same x*!
(need to check positive definite condition for L)

$$x* = \arg\min_x L(x, \lambda*)$$

The original optimization problem
can be solve by minimizing its Lagrangain function.

$$x* = \arg\min_x f(x)$$

$$\text{subject to} \quad g(x) \leq 0$$

$$\Updownarrow$$

if $L(x, \lambda *)$ is strict convex,
then we can find $x *$ $by\ \nabla_x L(x *, \lambda *) = 0$

$$x* = \arg\min_x L(x, \lambda*)$$

- $L(x*, \lambda*) = f(x*)$

$$x* = \arg\min_x f(x)$$

$$\text{subject to} \quad g(x) \leq 0$$

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

$$f(x*) + \lambda^* g(x*) = f(x*)$$

$$x* = \arg\min_x L(x, \lambda*)$$

$$\because \quad \lambda* = 0 \quad \text{if} \quad g(x*) < 0$$

$$\lambda* = \text{positive} \quad \text{if} \quad g(x*) = 0$$

Back to the Optimization Problem from Regression

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

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

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

this can be removed! + constant term would not affect the optimal solution.

- according to C we define, optimal Lagrangian $\lambda *$ will be different!
- Constant addition/subtraction won't change $x *$.

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

Back to the Optimization Problem from Regression

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

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

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

+ now we don't have C term!
but C term is implicitly
controlled by λ^*.

- in Ridge Regression Learning,
  we will change $\lambda *$ *and* test its performance to find a good $\lambda *$.
- the change of $\lambda *$ implicitly changes C value.

Example Problem) Use KKT conditions to find an optimal solution.

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

Error Decomposition

- Bias
- Variance
- Intrinsic Noise

Empirical MSE error vs. Expected Loss

- Suppose we had a data set: $D = \{(\overrightarrow{x_i}, t_i)\}$
- Suppose we learned a regression model $y(\vec{x}; D)$ from D.

  (Suppose one algorithm is fixed, for example: regression)

- Empirical MSE error (this is the one we do.)

$$L = \frac{1}{N}\sum_{i}^{N}\{y(x_i; D) - t_i\}^2$$

- Expected MSE error, as we know the density $f(\vec{x}, t) = f(t|\vec{x})\, f(\vec{x})$

$$E[L] = \int_x \int_t \{y(x; D) - t)\}^2 f(t|x) f(x)\, \mathrm{d}t\, \mathrm{d}x$$

- Error Decomposition

$$E[L] = \int_x \int_t \{y(x; D) - t)\}^2 f(t|x) f(x) \, \mathrm{d}t \, \mathrm{d}x$$

$$\begin{aligned} E[L] &= \int_x \int_t (y(x; D) - h(x) + h(x) - t)^2 f(t|x) f(x) \, \mathrm{d}t \, \mathrm{d}x \\ &= \int_x \int_t (y(x; D) - h(x))^2 + (h(x) - t)^2 + 2(y(x; D) - h(x))(h(x) - t) \, \mathrm{d}t \, \mathrm{d}x \end{aligned}$$

$h(x)$ is the optimal function that minimizes the modeling error given data density $f(x, t)$

Zero!

ϵ

$$h(x) = \arg\min_{y(x)} \int_x \int_t \{y(x) - t)\}^2 f(t|x) f(x) \, \mathrm{d}t \, \mathrm{d}x$$

- Optimal MMSE $h(x) \;=\; E[T|x]$

Suppose y(x) is our model, data follows the density $f(t,x) = f(t|x) \cdot f(x)$

$$E[L] = \int_x \int_t (y(x) - t)^2 f(t|x) f(x)\, \mathrm{d}t\, \mathrm{d}x$$

$$\int_t (y(x) - t)^2 f(t|x) f(x)\, \mathrm{d}t > 0$$

$$J(y(x*)) = \int_t (y(x*) - t)^2 f(t|x*)\, \mathrm{d}t$$

$$\frac{\partial J}{\partial y(x*)} = \int_t (2y(x*)^{-}2 \cdot t) f(t|x*)\, \mathrm{d}t = 0$$

$$y(x*) = \int_t t f(t|x*)\, \mathrm{d}t = E[T|x*]$$

E[T|x] is the function of x.
An ideal MMSE regression model! It is computed from densities, not from data! But the densities are unknown, so we estimated E[T|x] by using MLE!

- Error Decomposition

$$E[L] = \int_x \int_t \{y(x;D) - t)\}^2 f(t|x) f(x)\,\mathrm{d}t\,\mathrm{d}x$$

$$E[L] = \int_x \int_t (y(x;D) - h(x) + h(x) - t)^2 f(t|x) f(x)\,\mathrm{d}t\,\mathrm{d}x$$

$$= \int_x \int_t \underbrace{(y(x;D) - h(x))^2}_{?} + \underbrace{(h(x) - t)^2}_{\epsilon} + \underbrace{2(y(x;D) - h(x))(h(x) - t)}_{\text{zero}}\,\mathrm{d}t\,\mathrm{d}x$$

$t = h(x) + \epsilon$

$$\int_x \int_t 2(y(x;D) - h(x))(h(x) - t)\,\mathrm{d}t\,\mathrm{d}x$$

$$\int_x 2(y(x;D) - h(x)) \int_t (h(x) - t)\,\mathrm{d}t\, f(t|x)\,\mathrm{d}t\, f(x)\,\mathrm{d}x$$

$$\int_x 2(y(x;D) - h(x))(E[T|x] - E[T|x]) f(x)\,\mathrm{d}x = 0$$

- Error Decomposition

$$E[L] = \int_x \int_t \{y(x;D) - t)\}^2 f(t|x) f(x)\,\mathrm{d}t\,\mathrm{d}x$$

$$\begin{aligned} E[L] &= \int_x \int_t (y(x;D) - h(x) + h(x) - t)^2 f(t|x) f(x)\,\mathrm{d}t\,\mathrm{d}x \\ &= \int_x \int_t \underbrace{(y(x;D) - h(x))^2}_{} + \underbrace{(h(x) - t)^2}_{\epsilon} + \underbrace{2(y(x;D) - h(x))(h(x) - t)}_{\text{zero}}\,\mathrm{d}t\,\mathrm{d}x \end{aligned}$$

$$\int_x \int_t (y(x;D) - h(x))^2 f(t|x) f(x)\,\mathrm{d}t\,\mathrm{d}x$$

$$\int_x (y(x;D) - E_D[y(x;D)] + E_D[y(x;D)] - h(x))^2 f(x)\,\mathrm{d}x$$

$$\int_x (\underbrace{y(x;D) - E_D[y(x;D)]}_{\text{Variance}})^2 + (\underbrace{E_D[y(x;D)] - h(x)}_{\text{Bias}})^2 f(x)\,\mathrm{d}x$$

Trade-off Relation

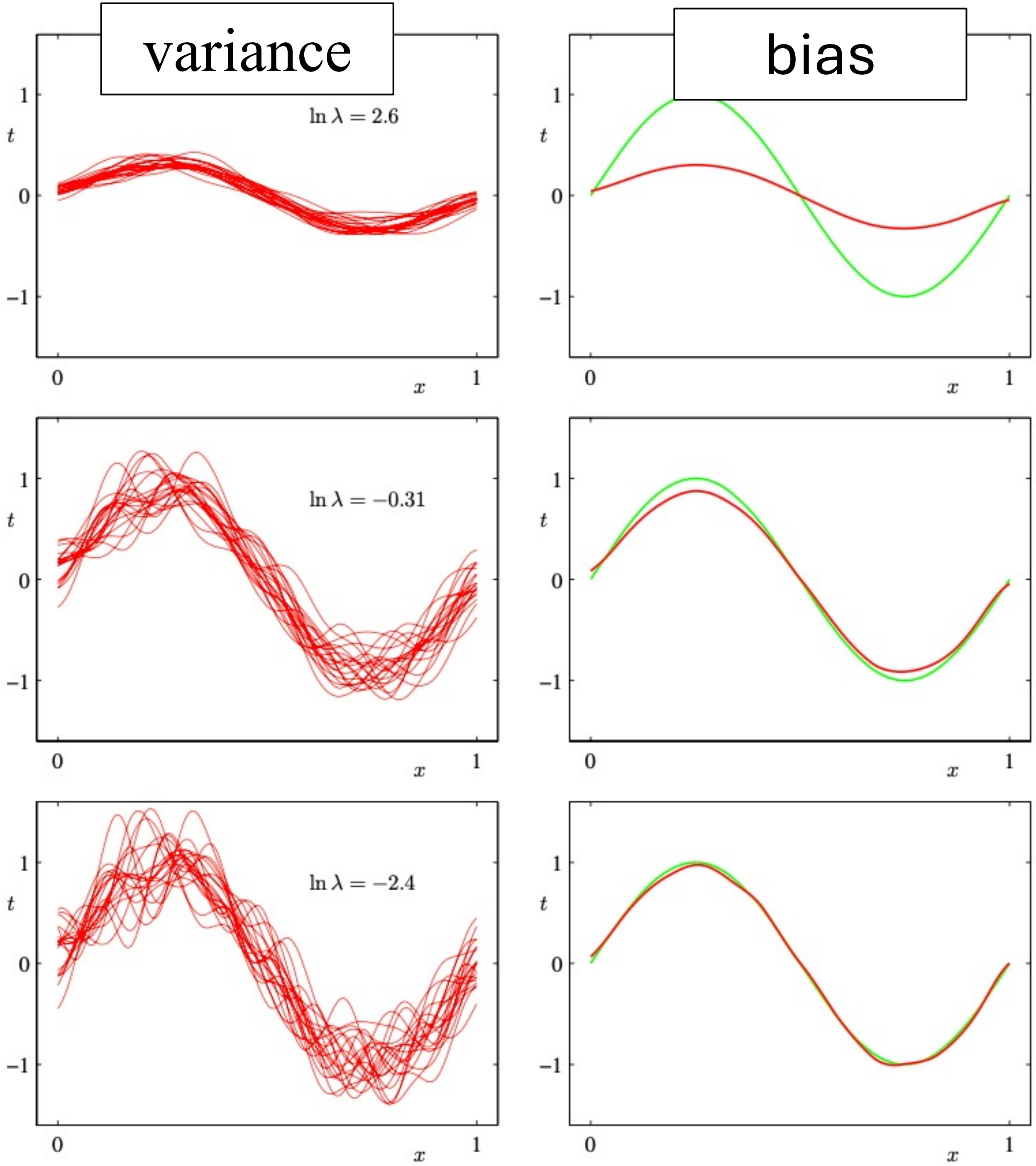
- Error Decomposition $E[L] = \text{Variance} + \text{Bias} + \text{Intrinstic Error}$

  - Intrinsic Error: $\int_x \int_t (E[T|x] - t)^2 f(t|x) f(x)\, \mathrm{d}t\, \mathrm{d}x$
  - Variance: $\int_x VAR_D[y(x;D)] f(x)\, \mathrm{d}x$
  - Bias: $\int_x \{E_D[y(x;D)] - E[T|x]\}^2 f(x)\, \mathrm{d}x$

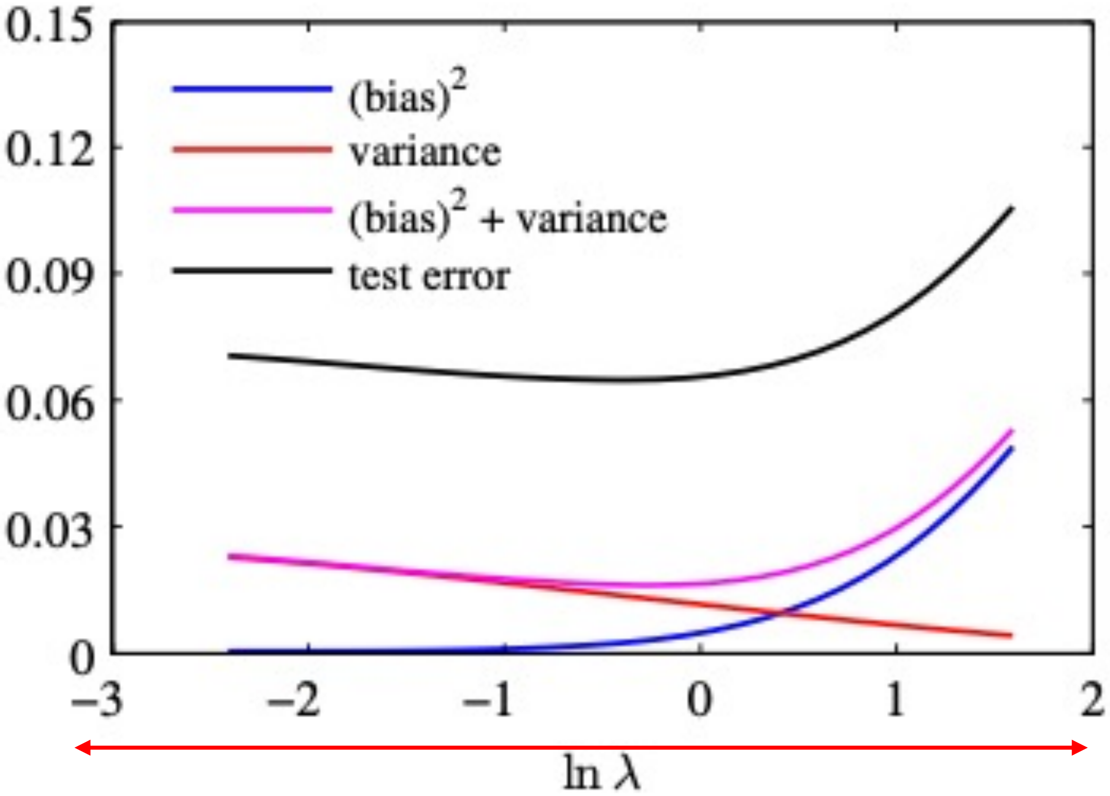
- Trade-Off between Variance & Bais

  Complex models : High Variance but Low Bias

  Simple model : Low Variance but High Bias



Trade off between variance and bias according to λ in Ridge Regression



High Complexity

Low Complexity