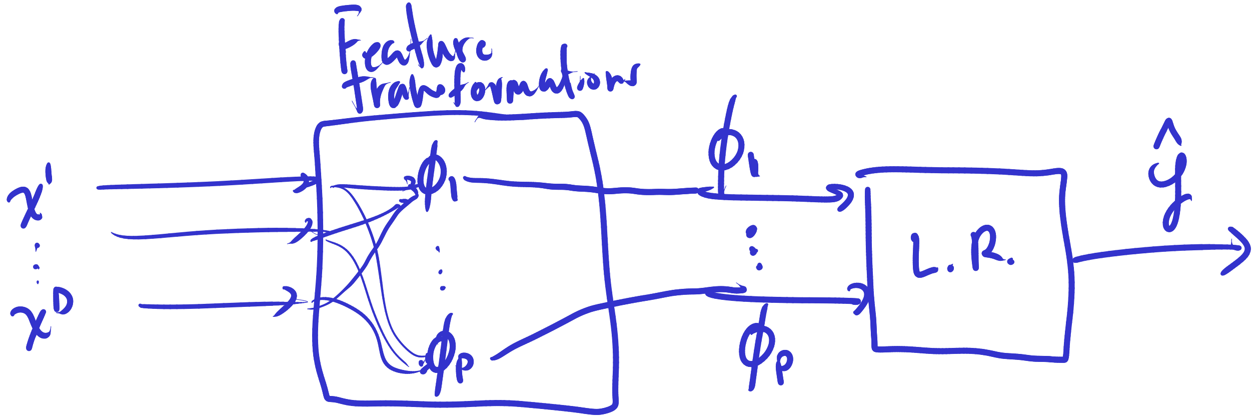
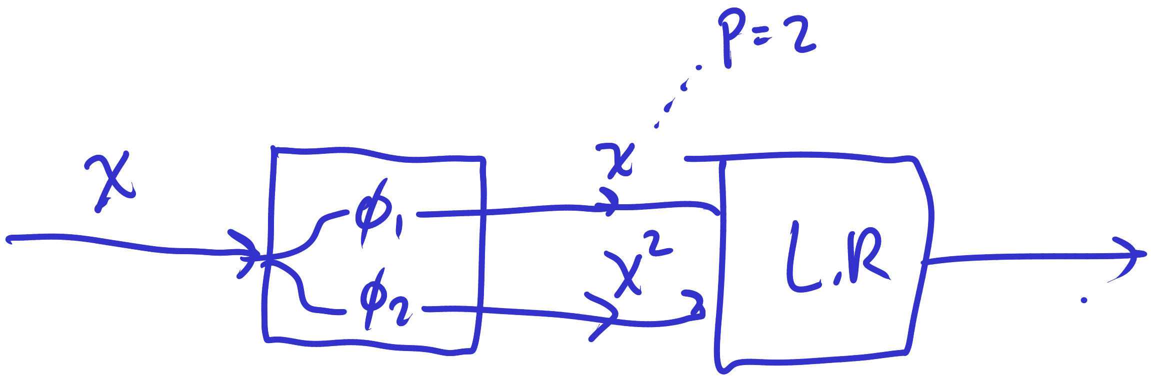
Statistics and Data Science for Engineers
E178 / ME276DS

Linear regression
Part 2

Recap


Feature transformations
x^1
x^D
ϕ_1
ϕ_p
L.R.
$\hat{y}$


P=2
x
ϕ_1
ϕ_2
x^2
L.R

Features

Feature : $\phi : \mathbb{R}^D \longrightarrow \mathbb{R}$. ... nonlinear transformation.

Takes an input sample and produces a number.

e.g.

$D=1$: $\phi_1(x) = x^2$

$D=2$: $\phi_1(x^1, x^2) = e^{x^1} \sin x^2$

Example: Quadratic fit

D=1

P=2,

trivial ... $\phi_1(x) = x$

$\phi_2(x) = x^2$

$$\Rightarrow$$

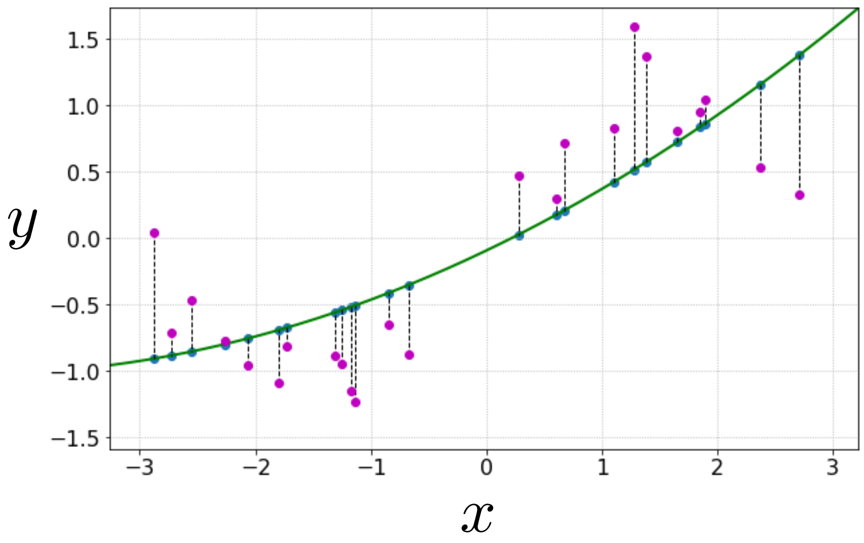
$$\begin{aligned} h(x_i; \theta_0, \underline{\theta}_1) &= \theta_0 + \phi_1(x_i)\theta_1 + \phi_2(x_i)\theta_2 \\ &= \theta_0 + x_i\theta_1 + x_i^2\theta_2 \end{aligned}$$

Least squares solution:

$$\widehat{\theta}_0 = -0.0935$$

$$\widehat{\theta}_1 = 0.4174$$

$$\widehat{\theta}_2 = 0.0467$$



In general: $h(x_i; \theta_0, \underline{\theta}_1) = \theta_0 + \phi_1(x_i)\theta_1 + \ldots + \phi_P(x_i)\theta_P$ ..... P features.

Data transformation: $\Phi = [\phi_1(\mathbf{X}) \quad \ldots \quad \phi_P(\mathbf{X})]$

P > D .

P < D .

Solution:

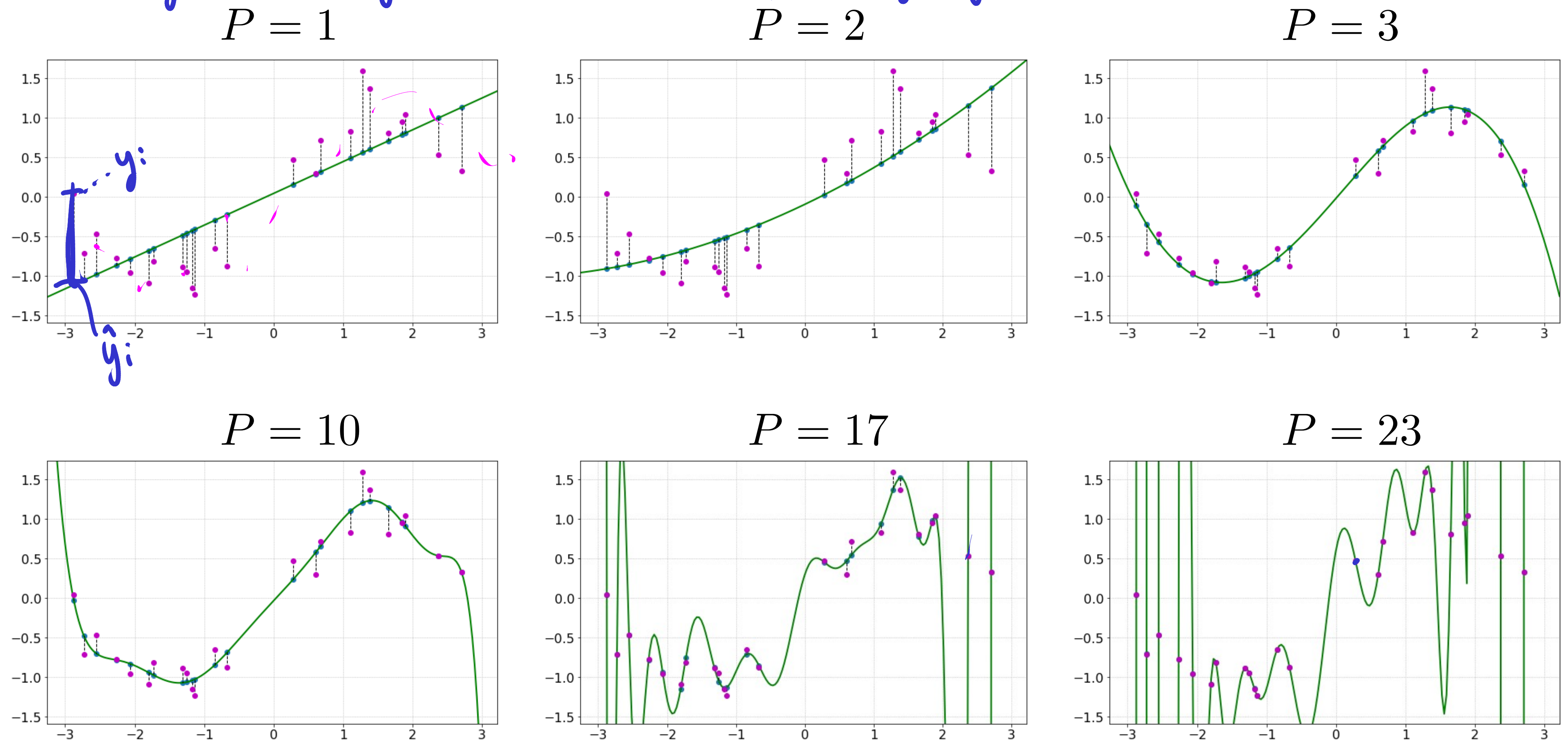
$\hat{\underline{\theta}}_1 = (\Phi_c^T \Phi_c)^{-1} \Phi_c^T Y_c$

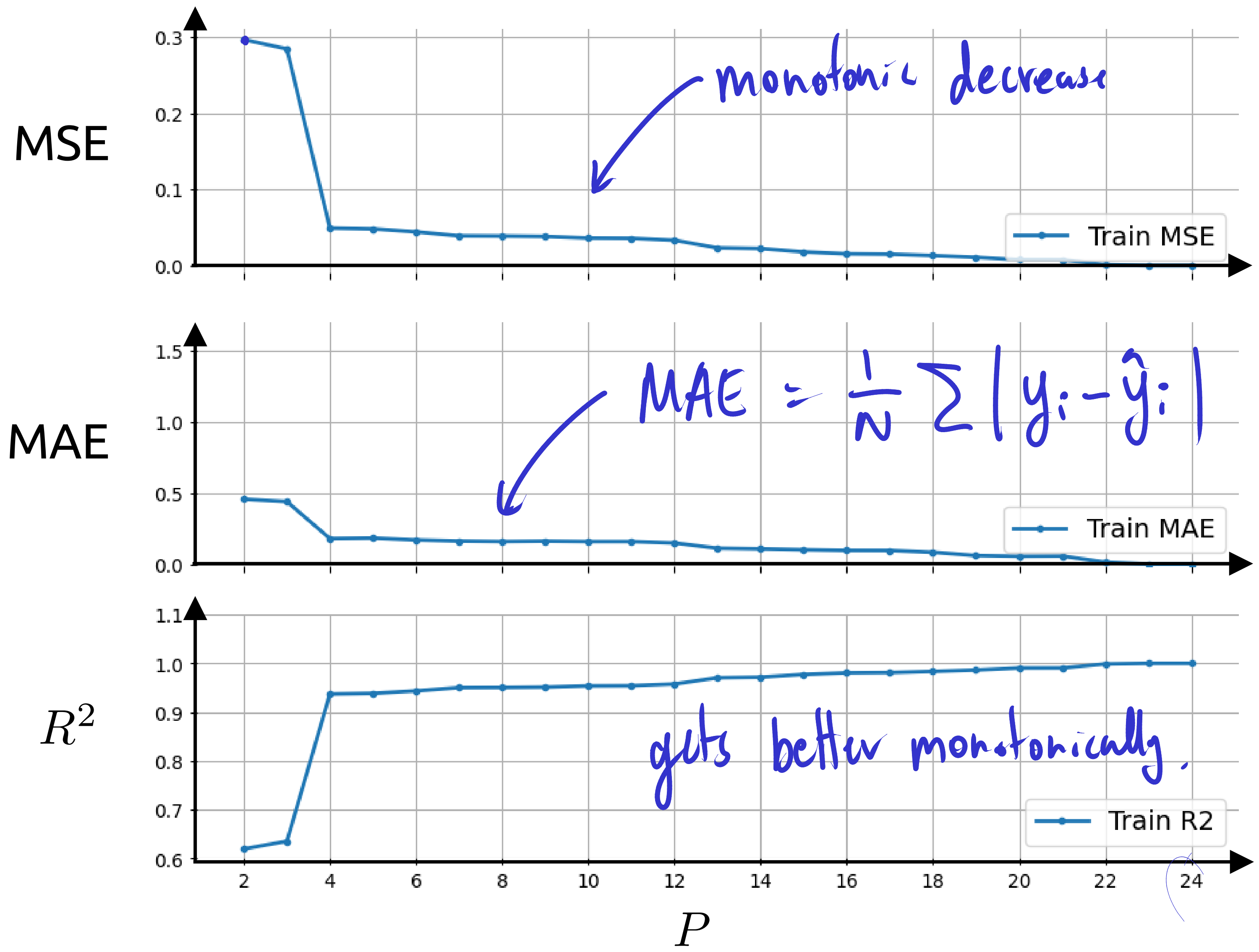
$\hat{\underline{\theta}}_0 = \hat{\mu}_Y - \hat{\mu}_\Phi \cdot \hat{\underline{\theta}}_1$

$\mathbf{\Phi}$ $\quad$ $\mathbf{Y}$

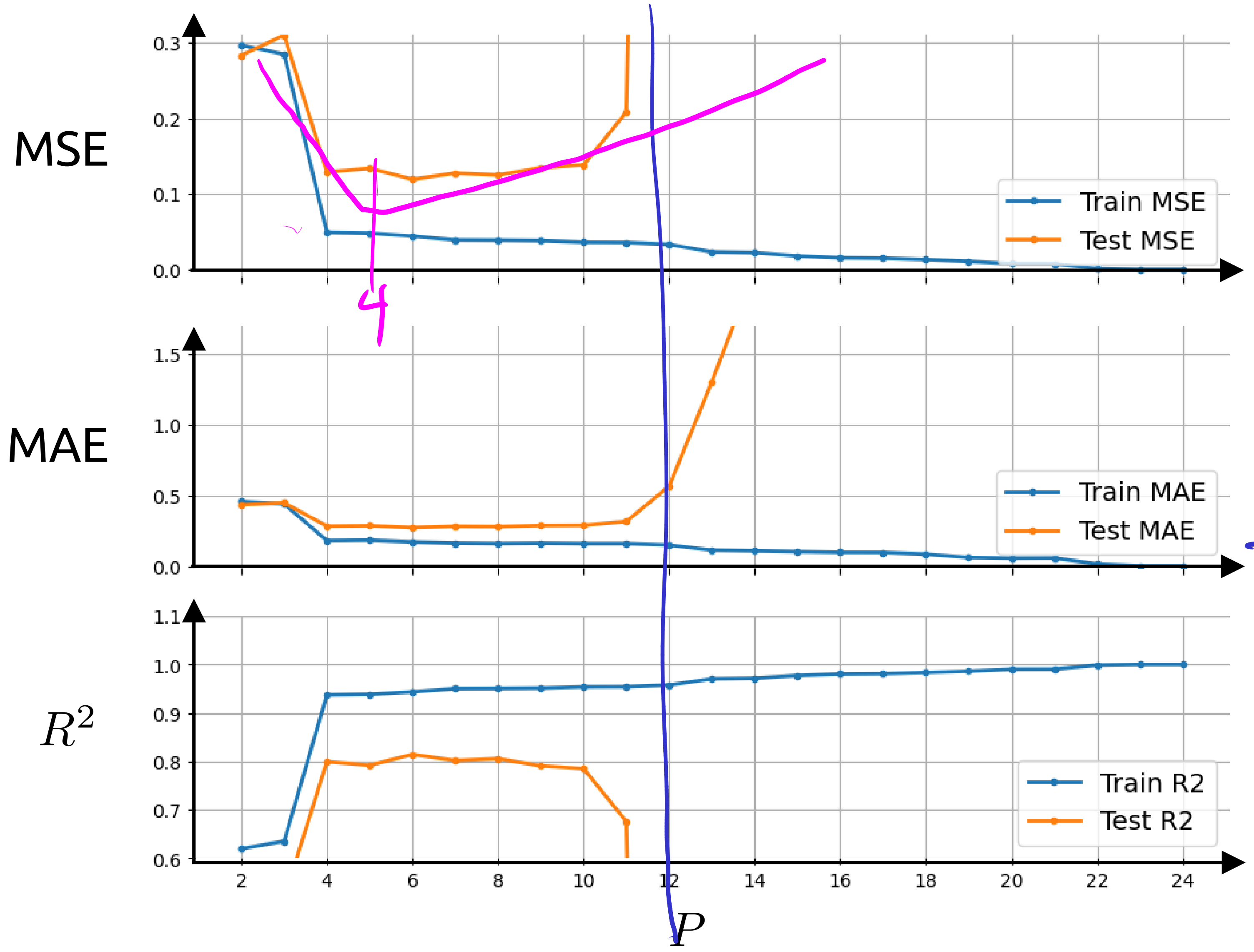
ϕ_1	ϕ_2	ϕ_3	$\cdots$	ϕ_P	y
0.247746	36.0	266.0	0.247746	88.019654	57.000823
0.179340	37.0	365.0	0.179340	27.211935	22.471227
0.956807	21.0	151.0	0.956807	97.456012	56.357366
0.869653	43.0	437[illegible]	[illegible]869653	43.203221	34.6[illegible]5
0.825345	47.0	160.0	0.825345	98.933930	64.426163
0.331114	321.0	371.0	0.331114	8.257917	14.218720
0.765523	17.0	364.0	0.765523	96.696783	59.123769
0.956807	21.0	151.0	0.956807	97.456012	52.983470

minimize $\sum (y_i - \hat{y}_i)^2$ = $MSE = \frac{1}{N} \sum (y_i - \hat{y}_i)^2$





MSE
MAE
R^2
P
Train MSE
Train MAE
Train R2
monotonic decrease
$MAE = \frac{1}{N}\sum |y_i - \hat{y}_i|$
gets better monotonically.
0.0
0.1
0.2
0.3
0.5
1.0
1.5
0.6
0.7
0.8
0.9
1.1
2
4
6
8
10
12
14
16
18
20
22
24



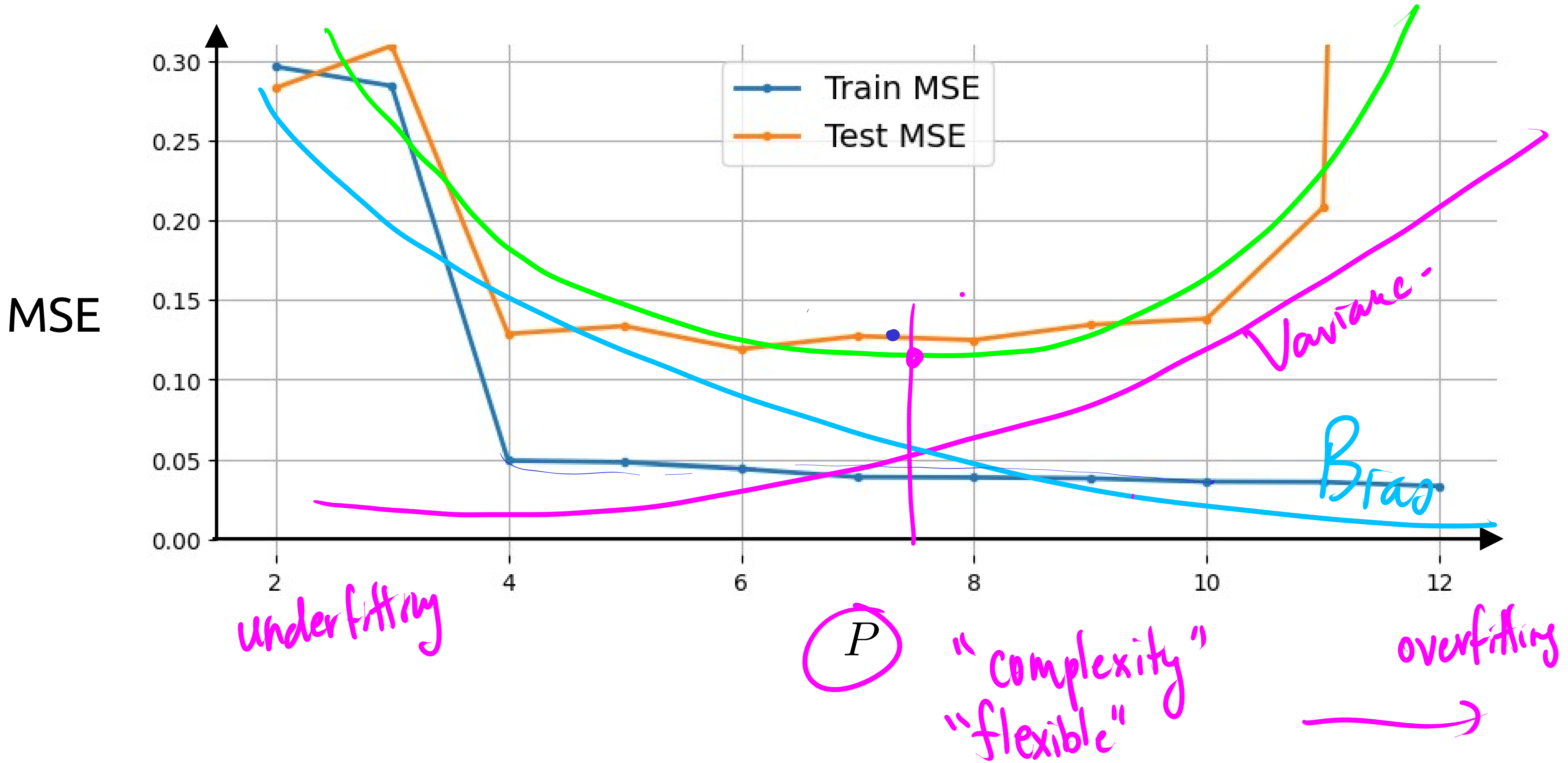
Test data provide a better (unbiased) estimate of the true performance of the model.

cross-validation also give an unbiased estimate of the true performance.

Aside: Bias/Variance tradeoff

Bias/Variance decomp.

MSE = Var + Bias^2.



Metrics :

- MSE
- RMSE
- MAE
- MAPE
- R^2

... will not detect overfitting.

Aside: Metrics that penalize complexity

penalizes P

$$\text{AIC} = \frac{1}{\hat{\sigma}^2}\text{MSE}_{\text{train}} + \frac{2}{N}P$$

... Akaike information criterion

$$\text{BIC} = \frac{1}{\hat{\sigma}^2}\text{MSE}_{\text{train}} + \frac{\log(N)}{N}P$$

... Bayesian information criterion

Cp ... Mallow's statistic.

where

$$\hat{\sigma} = \frac{1}{N-1-P}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

for the "full" model

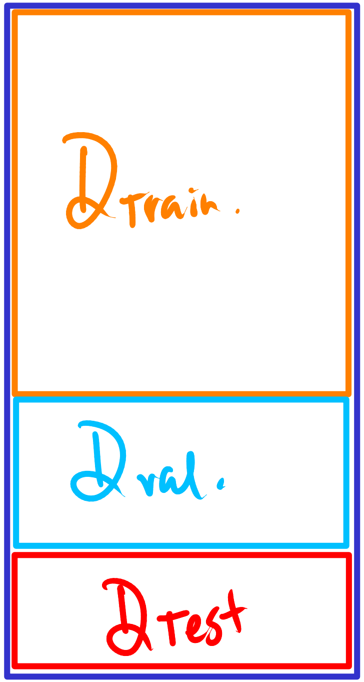
General feature selection (a.k.a. Model selection)

1. Subset selection.

2. Regularization.

Validation data:

$\mathcal{D} =$



Best subset selection

Set of all features

$$\mathcal{P} = \{ \phi_1 , \phi_2 , \phi_3 , \quad \cdots \quad , \phi_{P-1} , \phi_P \}$$

P

subsets of size 1

k=1 P

$S_1 = \{\phi_1\}$ → train a model, evaluate → ℓ

$S_2 = \{\phi_2\}$ → ℓ

⋮ → ℓ

$S_p = \{\phi_p\}$.

Subsets size 2

k=2

$S_{p+1} = \{\phi_1, \phi_2\}$.

$S_{p+2} = \{\phi_1, \phi_3\}$.

⋮

Best ℓ

size K $\binom{K}{p}$

$\vdots$

$K=p$ size p

$P=10 \longrightarrow 1024$

$P=20 \longrightarrow 10^6$

$P=30 \longrightarrow 10^9$

2^p

p things, how many subsets are there?

Best subset selection

for $k = 1 \ldots P$:

for $\mathcal{A}_\kappa$ in $\{$all k-sized subsets of $\mathcal{P}\}$:

$\hat{\theta}_\kappa = \text{train}(\mathcal{A}_\kappa, \mathcal{D}_{\text{train}})$

$\ell_\kappa = \text{perf}(\mathcal{A}_\kappa, \hat{\theta}_\kappa, \mathcal{D}_{\text{val}})$

$\kappa^* = \text{argbest}(\{\ell_\kappa\})$

$\mathcal{S}_k = \mathcal{A}_{\kappa^*}$

$\mathcal{S}^* = \text{best of } \{\mathcal{S}_k\}_P$

$\hat{\theta}^* = \text{train}(\mathcal{S}^*, \mathcal{D}_{\text{train}})$

$\ell^* = \text{perf}(\mathcal{S}^*, \hat{\theta}^*, \mathcal{D}_{\text{test}})$

Forward stepwise selection

Assume $\mathcal{S}_k \subset \mathcal{S}_{k+1}$

$$\# \text{evaluations} = \sum_{i=1}^{P} i = \frac{P(P-1)}{2}.$$

$$\mathcal{P} = \{ \phi_1 , \phi_2 , \phi_3 , \quad \cdots \quad , \phi_{P-1} , \phi_P \}$$

$$\mathcal{S}_0 = \{\}$$

$$\mathcal{S}_1 = \{ \quad \phi_3 \quad \} \quad P \text{ evaluations.}$$

$$\mathcal{S}_2 = \{ \quad \phi_3 \quad \phi_5 \quad \} \quad P-1 \text{ evaluations.}$$

$$\vdots$$

$$\mathcal{S}_P = \{ \phi_1 , \phi_2 , \phi_3 , \quad \cdots \quad , \phi_{P-1} , \phi_P \}$$

Forward stepwise selection

set minus.

all unused features in stage k-1

$\mathcal{S}_0 = \{\}$

for $k = 1 \ldots P$:

for $\kappa, \phi_p \in$ enumerate$(\mathcal{P} \backslash \mathcal{S}_{k-1})$

$\mathcal{A}_\kappa = \mathcal{S}_{k-1} \cup \phi_p$

$\hat{\theta}_\kappa = \text{train}(\mathcal{A}_\kappa, \mathcal{D}_{\text{train}})$

$\ell_\kappa = \text{perf}(\mathcal{A}_\kappa, \hat{\theta}_\kappa, \mathcal{D}_{\text{val}})$

$\kappa^* = \text{argbest}(\{\ell_\kappa\})$

$\mathcal{S}_k = \mathcal{A}_{\kappa^*}$

$\mathcal{S}^* = \text{best of } \{\mathcal{S}_k\}_P$

$\hat{\theta}^* = \text{train}(\mathcal{S}^*, \mathcal{D}_{\text{train}})$

$\ell^* = \text{perf}(\mathcal{S}^*, \hat{\theta}^*, \mathcal{D}_{\text{test}})$

```
curlyS = [set() for i in range(P+1)]
ellk = np.full(P+1,np.inf)

for k in range(1,P+1):

    curlyA = [set() for i in range(P-k+1)]
    ellkappa = np.full(P-k+1,np.inf)

    for kappa, phip in enumerate(curlyP-curlyS[k-1]):
        curlyA[kappa] = curlyS[k-1].union({phip})
        theta0hat, theta1hat = train( curlyA[kappa] , Dtrain)
        ellkappa[kappa] = perf(curlyA[kappa], theta0hat, theta1hat,
                               Dvalidate)

    kappastar = ellkappa.argmin()
    curlyS[k] = curlyA[kappastar]
    ellk[k] = ellkappa[kappastar]

kstar = ellk.argmin()
Sstar = curlyS[kstar]
theta0star, theta1star = train(Sstar, Dtrain)
ellstar = perf(Sstar, theta0star, theta1star, Dtest)

# Store the results
f_ellk = ellk
f_ellstar = ellstar
f_kstar = kstar
```

Backward stepwise selection

Assume $\underline{\mathcal{S}_k \subset \mathcal{S}_{k+1}}$

$$\mathcal{S}_P = \{ \phi_1 , \phi_2 , \phi_3 , \quad \cdots \quad , \phi_{P-1} , \phi_P \}$$

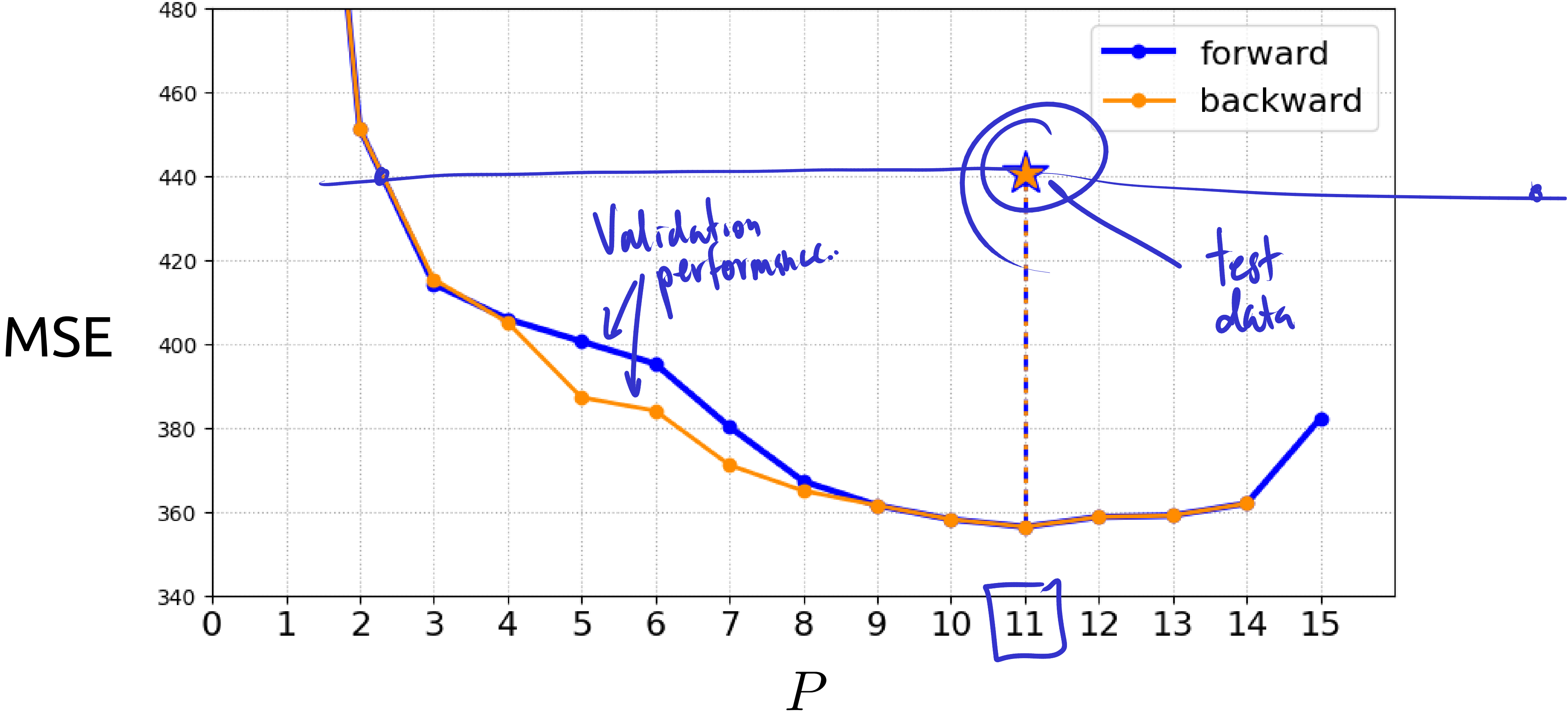
$$\mathcal{S}_{P-1} = \{ \phi_1 , \cancel{\phi_2} , \phi_3 , \quad \cdots \quad , \phi_{P-1} , \phi_P \}$$

P-1 evaluations.

$$\mathcal{S}_{P-2} = \{ \phi_1 , \cancel{\phi_2} , \phi_3 , \quad \cdots \quad , \phi_{P-1} , \phi_P \}$$

$$\vdots$$

$$\mathcal{S}_0 = \{ \}$$


MSE
480
460
440
420
400
380
360
340
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
P
forward
backward
Validation performance.
test data

Parameter shrinkage, a.k.a. Regularization

penalty on the size of the parameters. θ.

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \left(\underbrace{\sum_{i=1}^{N} L(y_i, \hat{y}_i)}_{\text{loss minimization.}} + \underbrace{\lambda R(\theta)} \right)$$

balancing parameter.

$$\boxed{\lambda \geq 0}$$

λ small $\longrightarrow$ overfitting

λ large $\longrightarrow$ less overfitting.

R can be a norm function.
$\rightarrow$ 1 norm (absolute values) LASSO.
$\rightarrow$ 2 norm (Euclidean) Ridge.

Ridge regression

(linear regression

$$\hat{\theta}_{\text{ridge}} = \underset{\theta_0 \ldots \theta_P}{\operatorname{argmin}} \left(\overbrace{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}^{L_2.} + \lambda \overbrace{\sum_{j=1}^{P} \theta_j^2}^{\text{slope parameters.}} \right)$$

$$= \underset{\theta_0, \underline{\theta}_1}{\operatorname{argmin}} \underbrace{\left(\left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|_2^2 + \lambda \left\| \underline{\theta}_1 \right\|_2^2 \right)}_{\text{matrix notation.}}$$

$$\theta = \begin{bmatrix} \cancel{\theta_0} \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_P \end{bmatrix}$$

intercept

$\underline{\theta}_1$

slope parameters.

always invertible

$\lambda > 0.$

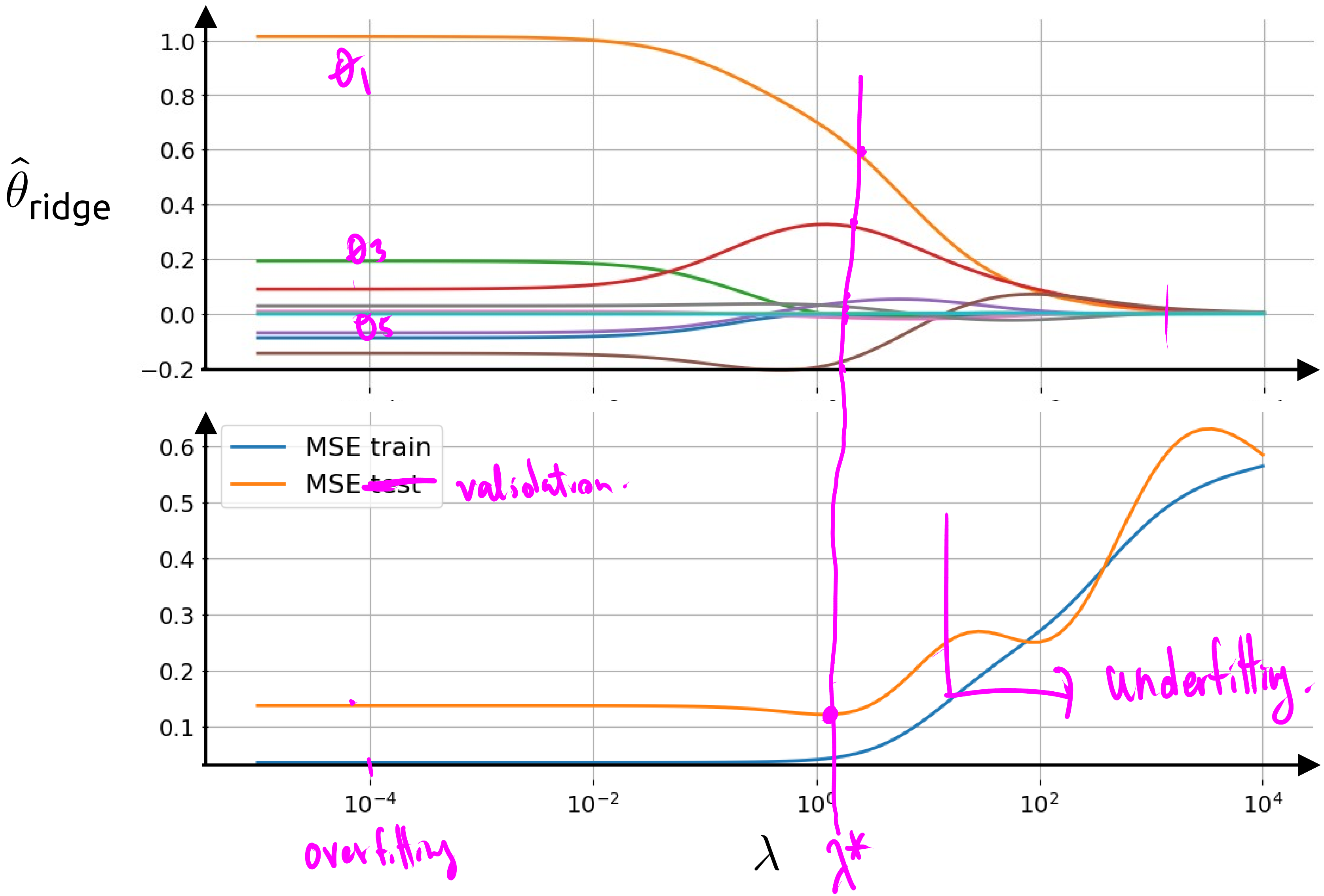
Solution:

$$\hat{\theta}_0 = \hat{\mu}_Y - \hat{\mu}_X \underline{\hat{\theta}}_1$$

identity matrix.

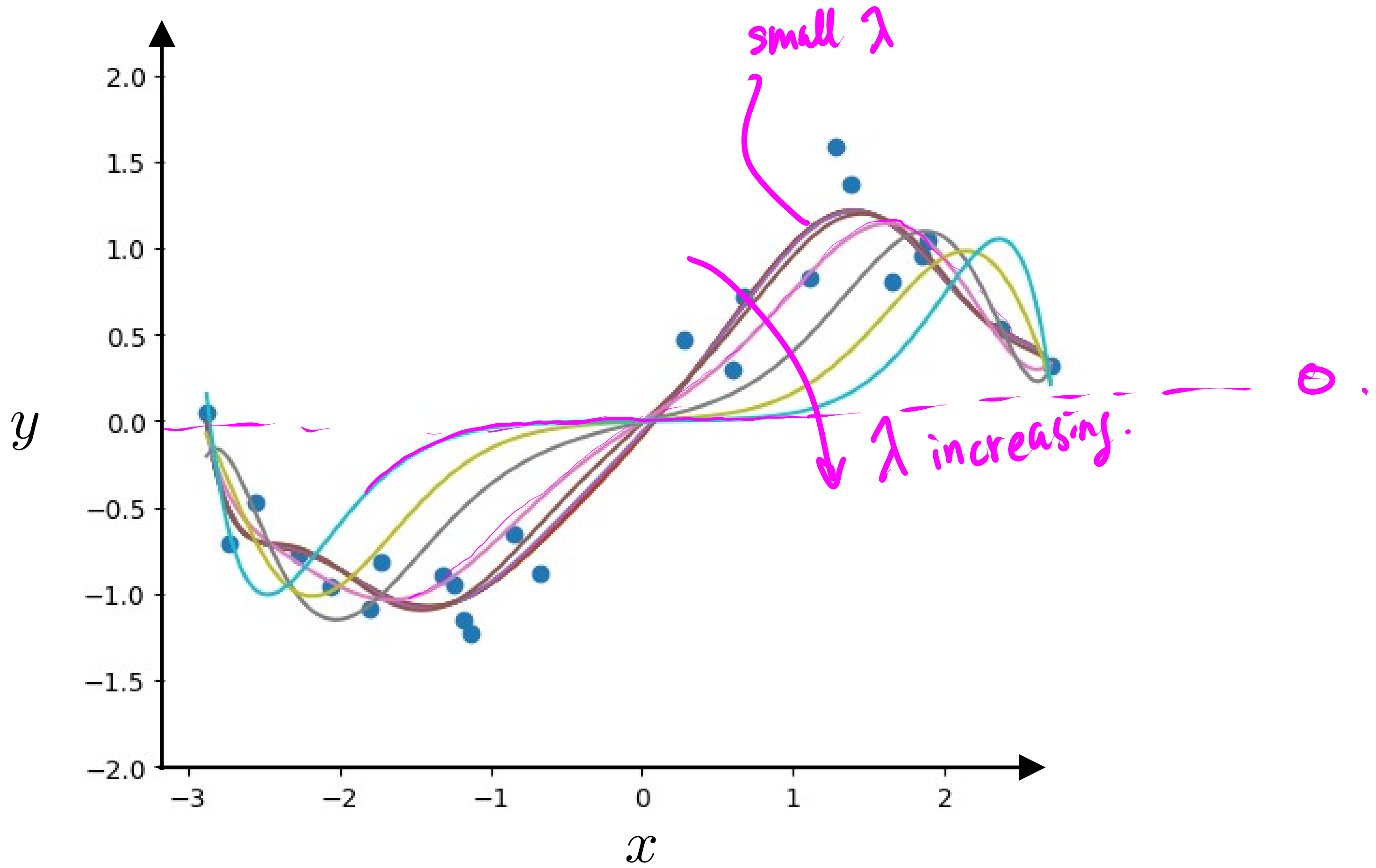
$$\underline{\hat{\theta}}_1 = \big(\underbrace{(\mathbf{X}^c)^T \mathbf{X}^c}_{\text{Before}} + \underbrace{\lambda I}_{\text{assumed invertible.}}\big)^{-1} (\mathbf{X}^c)^T \mathbf{Y}^c$$

$$\underline{\hat{\theta}}_1 = \left((X^c)^T X^c + \lambda I \right)^{-1} (X^c)^T Y^c$$



$\hat{\theta}_{\text{ridge}}$
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
θ1
θ3
θ5
MSE train
MSE ~~test~~
validation.
0.6
0.5
0.4
0.3
0.2
0.1
10^{-4}
10^{-2}
10^{0}
10^{2}
10^{4}
λ
λ^*
overfitting
underfitting.

Ridge regularized linear regression



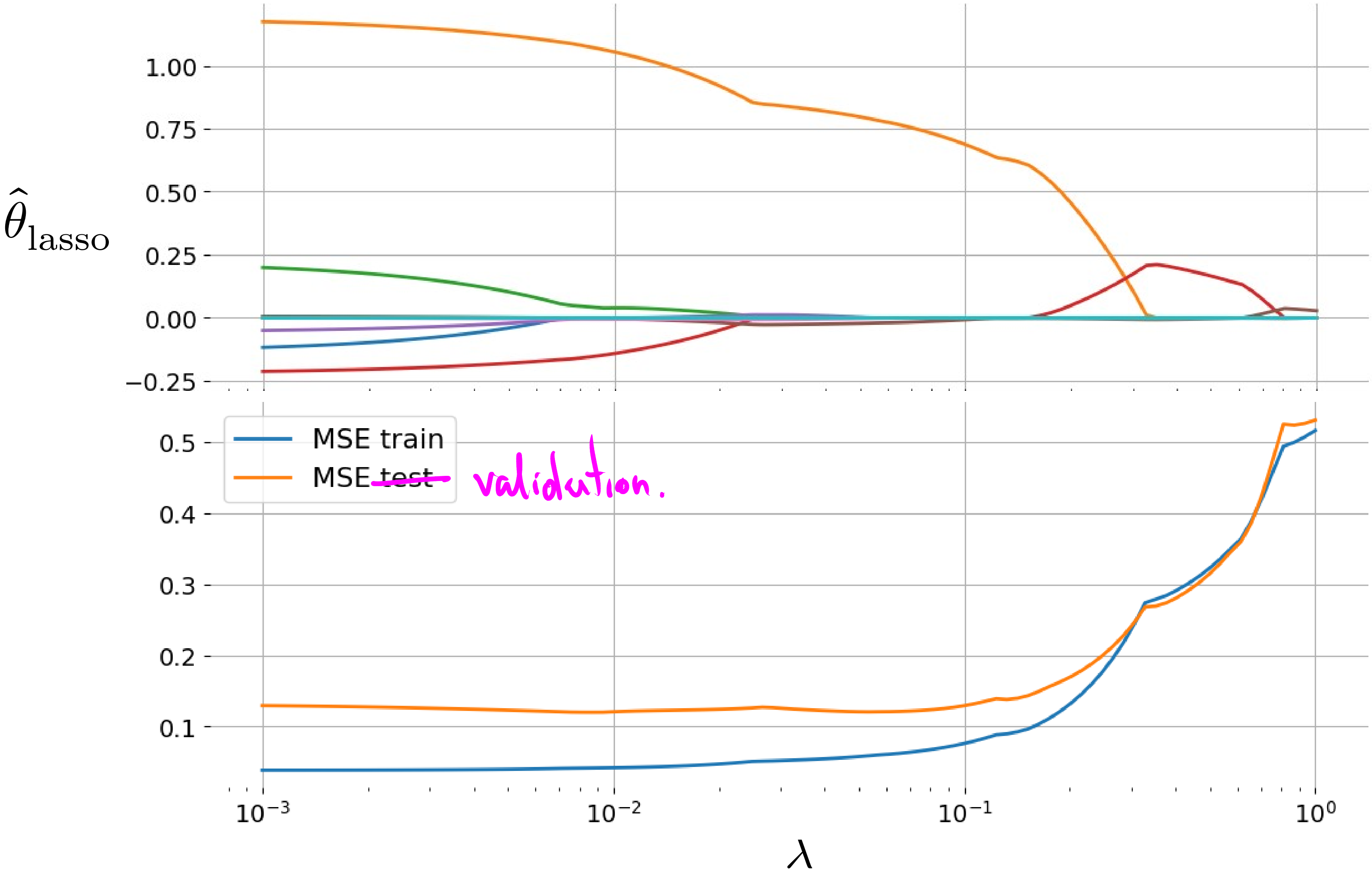
Lasso regression

$$\hat{\theta}_{\text{lasso}} = \underset{\theta_0 \ldots \theta_P}{\operatorname{argmin}} \left(\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} |\theta_j| \right)$$

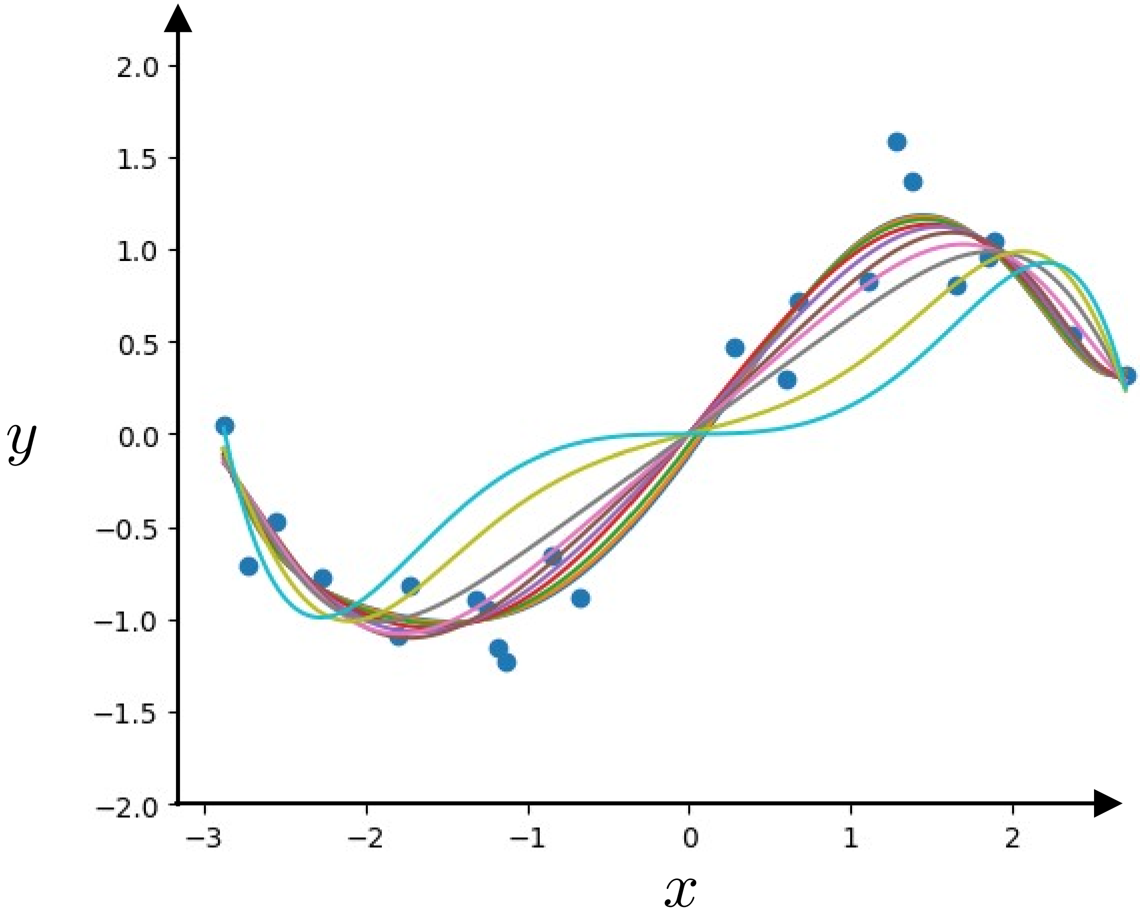
$$= \underset{\theta_0, \underline{\theta}_1}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|_2^2 + \lambda \left\| \underline{\theta}_1 \right\|_1 \right)$$

1 norm.

Creates sparse models. several θ's = 0.


$\widehat{\theta}_{\text{lasso}}$
1.00
0.75
0.50
0.25
0.00
−0.25
MSE train
MSE ~~test~~ validation.
0.5
0.4
0.3
0.2
0.1
10^{-3}
10^{-2}
10^{-1}
10^{0}
λ

Lasso regularized linear regression



A different perspective on Ridge and Lasso

$$\hat{\theta}_{\text{ridge}} = \underset{\theta_0, \underline{\theta}_1}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|^2 + \lambda \left\| \underline{\theta}_1 \right\|_2^2 \right)$$

Lagrange multipliers

$$= \underset{\theta_0, \underline{\theta}_1}{\operatorname{argmin}} \left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|^2$$

$$\text{s.t.} \quad \left\| \underline{\theta}_1 \right\|_2 \leq t$$

depends on λ

$$\hat{\theta}_{\text{lasso}} = \underset{\theta_0, \underline{\theta}_1}{\operatorname{argmin}} \left(\left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|^2 + \lambda \left\| \underline{\theta}_1 \right\|_1 \right)$$

$$= \underset{\theta_0, \underline{\theta}_1}{\operatorname{argmin}} \left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|^2$$

$$\text{s.t.} \quad \left\| \underline{\theta}_1 \right\|_1 \leq t$$