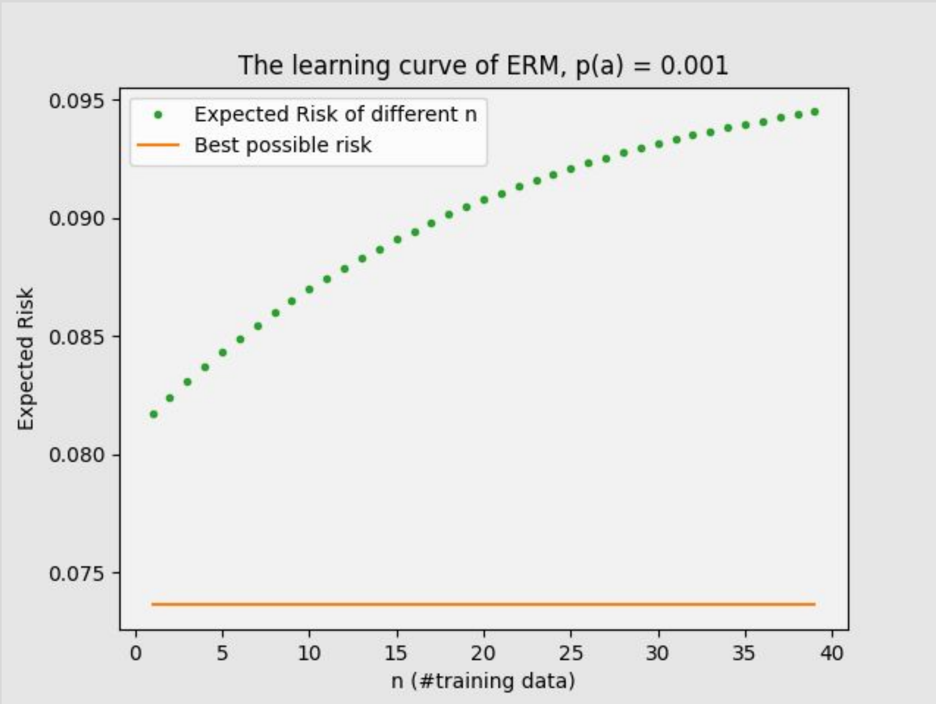
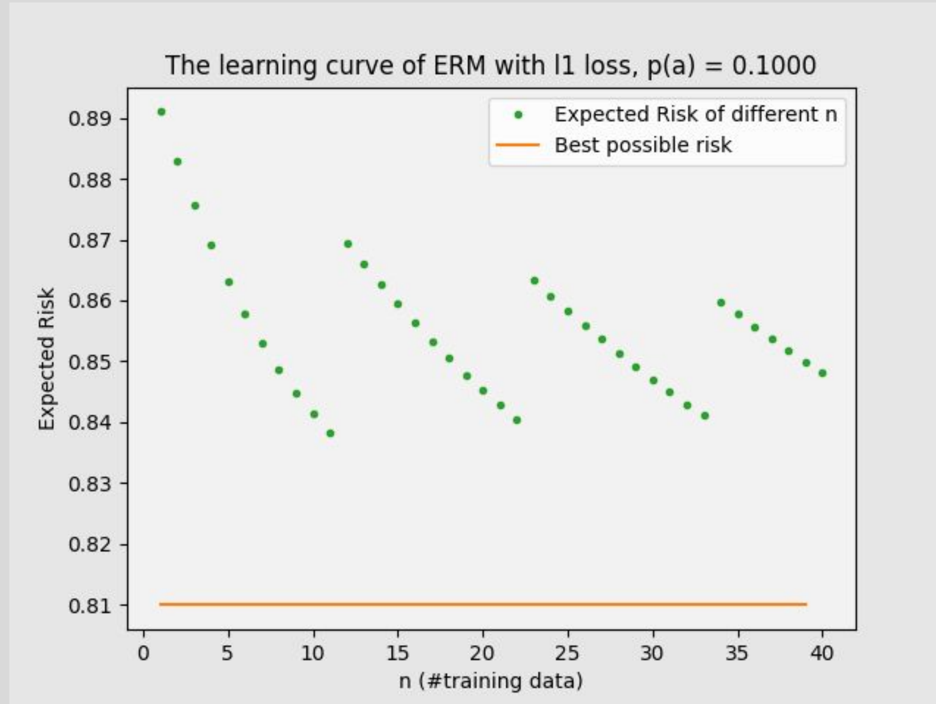
More Data: Better or Worse

Zhiyi Chen Supervisors: Prof.dr. Marco Loog, Tom Viering

Why do the *learning curves** have unexpected behavior ?

- Using *ERM**, Same distribution, but More data
- Hypothesis Class: $\mathcal{H} = \{h(x) = \beta x | \beta \in \mathbb{R}\}$





What is a *learning curve*?

- # training samples vs. generalization performance
- More training data? Better and worthy?

What is *ERM* (Empirical Risk Minimizer)?

- Empirical Risk:

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(h(x_i), y_i)$$

- Best hypothesis for training data

$$\mathcal{A}_{erm}(S^n) = \arg\min_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(h(x_i), y_i)$$

Problem Setting Ⅰ

- Learner: ERM, L2 loss

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} (h(x_i) - y_i)^2$$

- Distribution:

$$a = (1,1), b = (\tfrac{1}{10}, 1)$$
$$P(a) = 0.001, P(b) = 0.999$$

Problem Setting Ⅱ

- Learner: ERM, L1 loss

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} |(h(x_i) - y_i)|$$

- Distribution:

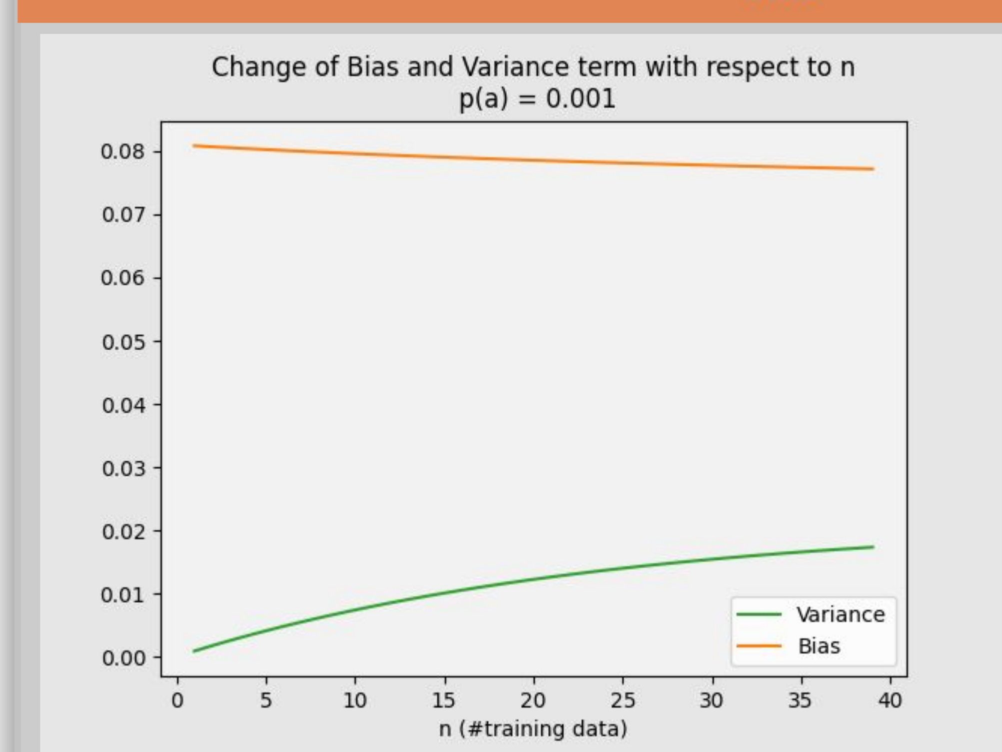
$$a = (1,1), b = (\tfrac{1}{10}, 1)$$
$$P(a) = 0.1, P(b) = 0.9$$

Performance metrics :

$$\mathbb{E}_{S^n}\mathbb{E}_{(x,y)}(\hat{\beta}x - y)^2$$

Bias-Variance trade-off :

$$\mathbb{E}_{S^n}\mathbb{E}_{(x,y)}(\hat{h}x - y)^2 = \underbrace{\mathbb{E}_x x^2 Var_{S^n}(\hat{h})}_{\text{variance}} + \underbrace{\mathbb{E}_x(\mathbb{E}_{S^n}\hat{h}x - 1)^2}_{\text{bias}}$$



Variances increases too fast
Bias decreases too slow

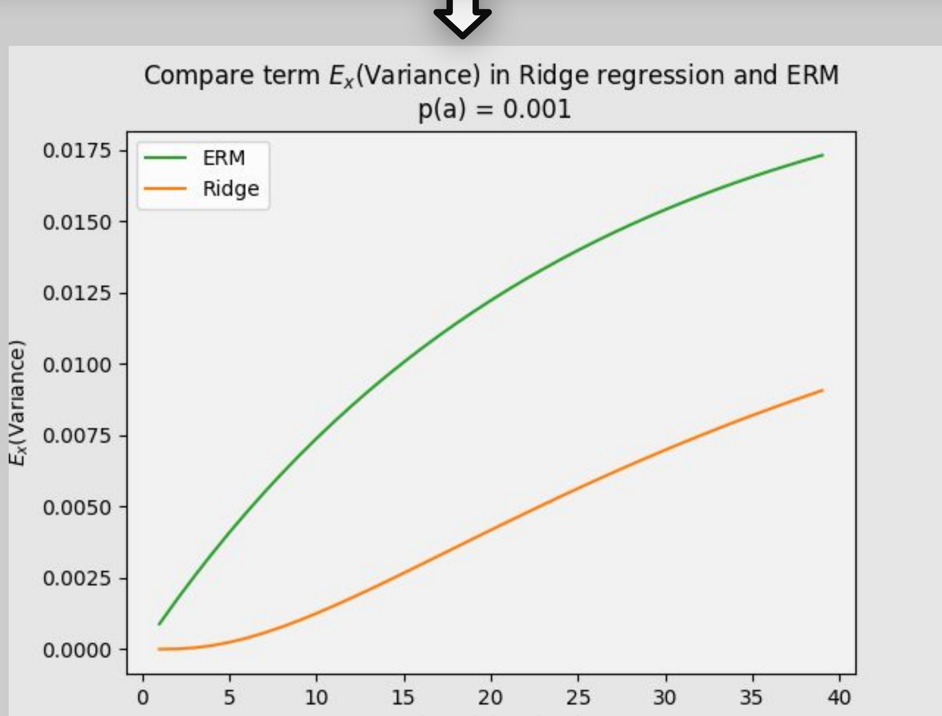
Increasing learning Curve?

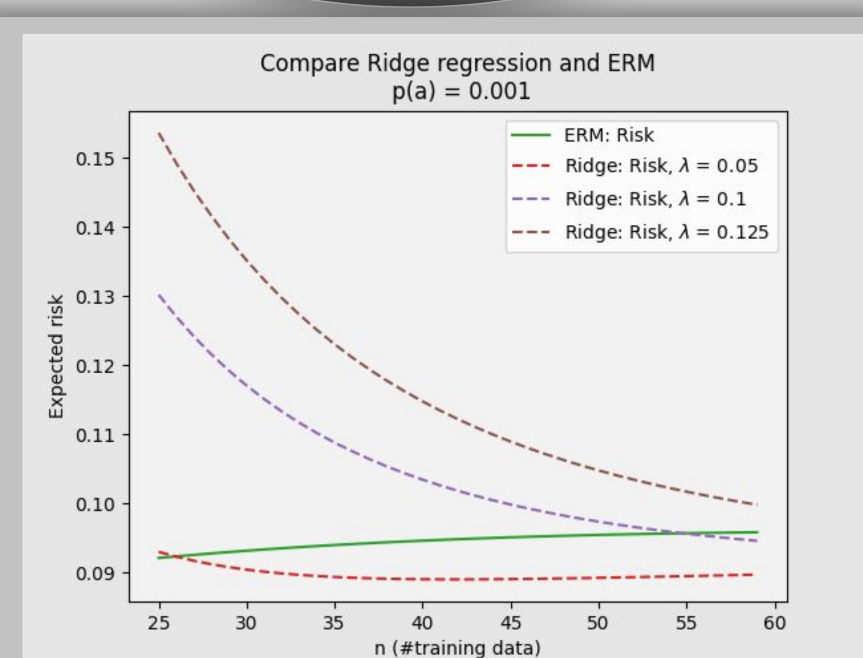
⇩

Ridge Regression

$$\mathcal{A}_{ridge}(S^n) = \arg\min_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(h(x_i), y_i) + \lambda||\beta||$$

⇩





Why does variance increase?

Since $Var(X_n) = \frac{1}{n}Var(x))$

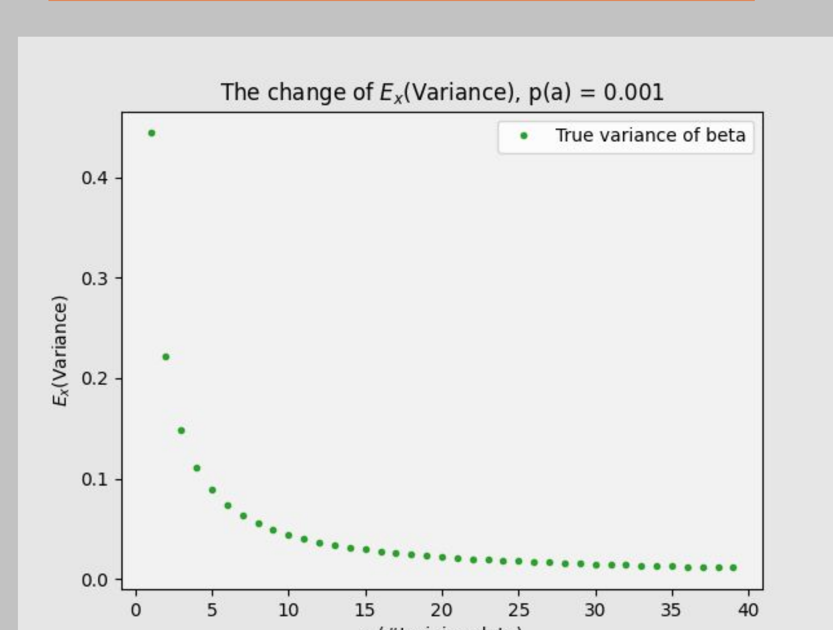
Not conforming with *linear model** ?

$$Y = \beta X + \epsilon, \text{ where } \mathbb{E}\epsilon = 0$$

Four-point distribution:

- $a_1 = [1, \frac{3}{2}]$ $\frac{1}{2}p_a$ $a_2 = [1, \frac{1}{2}]$ $\frac{1}{2}p_a$ $b_1 = [\frac{3}{4}, \frac{5}{4}]$ $\frac{1}{2}p_b$ $b_2 = [\frac{3}{4}, \frac{1}{4}]$ $\frac{1}{2}p_b$
- Fits linear model

$$Y = \beta X + \epsilon$$
$$\beta = 1, \mathbf{R}_X = \left\{1, \frac{3}{4}\right\}$$
$$\mathbf{R}_\epsilon = \left\{-\frac{1}{2}, \frac{1}{2}\right\}, P\left(\epsilon = \pm\frac{1}{2}\right) = \frac{1}{2}$$



Result of ERM :

$$\hat{\beta} = \begin{cases} \frac{y_b}{x_b} & \text{if } n_a x_a - n_b x_b < 0 \\ \frac{y_a}{x_a} & \text{else} \end{cases}$$

Best possible hypothesis :

$$\beta = \begin{cases} \frac{y_b}{x_b} & \text{if } P(a)x_a - P(b)x_b < 0 \\ \frac{y_a}{x_a} & \text{else} \end{cases}$$

Performance metrics :

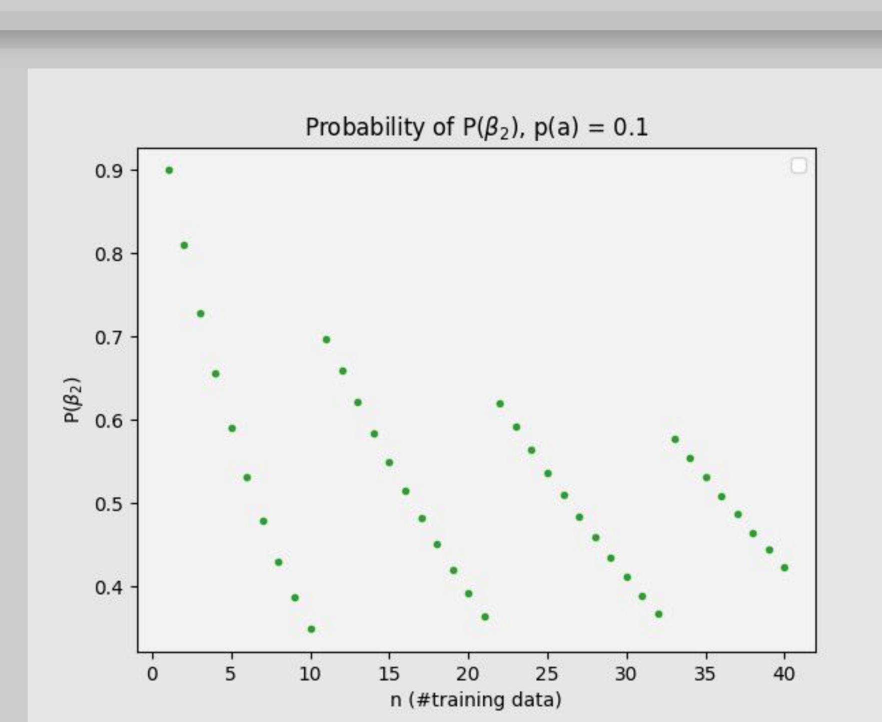
Assume $\hat{\beta}_1 = \frac{y_a}{x_a} = \beta$, $\hat{\beta}_2 = \frac{y_b}{x_b} \neq \beta$

Denote $P(\hat{\beta} = \hat{\beta}_1)$ as P_1, $P(\hat{\beta} = \hat{\beta}_2)$ as P_2

$$\mathbb{E}_{S^n}\mathbb{E}_{(x,y)}|\hat{\beta}x - y| = P_1 \cdot \mathbb{E}_{(x,y)}|\hat{\beta}_1(x) - y| + P_2 \cdot \mathbb{E}_{(x,y)}|\hat{\beta}_2(x) - y|$$

⇩

The *smaller* $\mathbf{P_2}$ is,
the *smaller* the risk is.



What causes the periodic behavior?

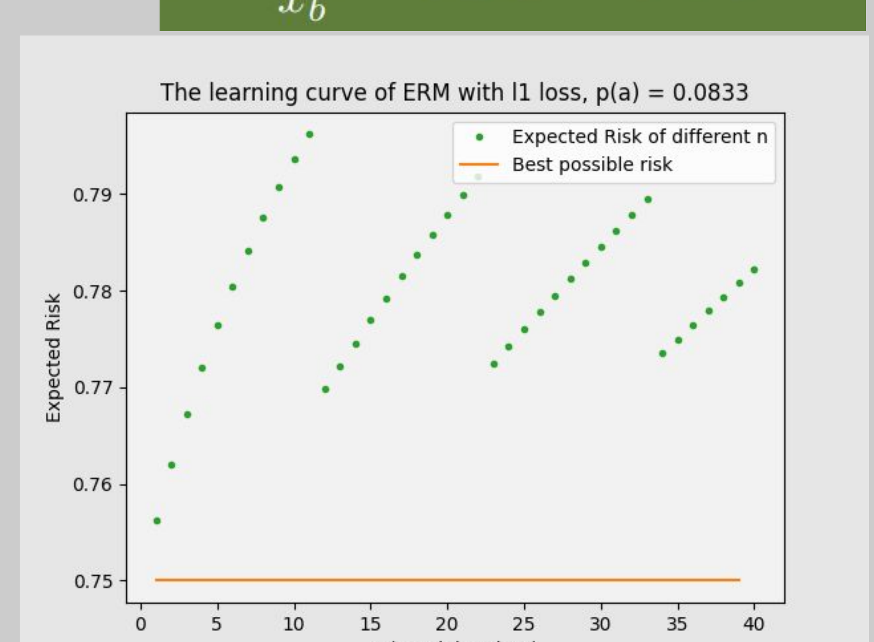
If $n_a x_a - n_b x_b < 0$, $\hat{\beta} = \frac{y_b}{x_b}$

Replace n_b with $n - n_a$, $n_a < \frac{1}{\frac{x_a}{x_b} + 1} n$

n_a, $n \in \mathbb{N}$, the possible value for n_a increases by 1 when n increases $\lceil x_a/x_b + 1 \rceil$, in this case 11.

Can the learning curve change behavior?

- When $\beta = \frac{y_b}{x_b}$, $P(a)x_a - P(b)x_b < 0$



- When $\lceil x_a/x_b + 1 \rceil = 21$: