Explain Strange Learning Curves in Machine Learning

Zhiyi Chen
Supervisor(s): Tom Viering, Marco Loog
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

The learning curve illustrates how the generalization performance of the learner evolves with more training data. It can predict the amount of data needed for decent accuracy and the highest achievable accuracy. However, the behavior of learning curves is not well understood. Many assume that the more training data provided, the better the learner performs. However, many counter-examples exist for both classical machine learning algorithms and deep neural networks. As presented in previous works, even the learning curves for simple problems using classical machine learning algorithms have unexpected behaviors. In this paper, we will explain what caused the odd learning curves generated while using ERM to solve two regression problems. Loog *et al.* [1] first proposed these two problems. As a result of our study, we conclude that the unexpected behaviors of the learning curves under these two problem settings are caused by incorrect modeling or the correlation between the expected risk and the output of the learner.

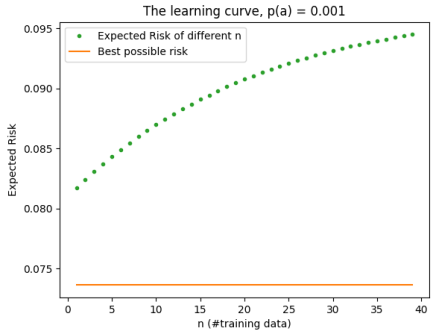
1 Introduction

Learning curves are plots demonstrating how the number of training samples influences the generalization performance of learners. Thus, the learning curves can display how well the learner solves the problem. They have been an essential tool for researchers to predict the maximum achievable accuracy, estimate how much data is required for the desired accuracy, and evaluate the generalization performance of learners [2, 3]. Moreover, their ability to predict the amount of data needed can be used in large learning problems to save computational costs and avoid the usage of the excess training samples [4].

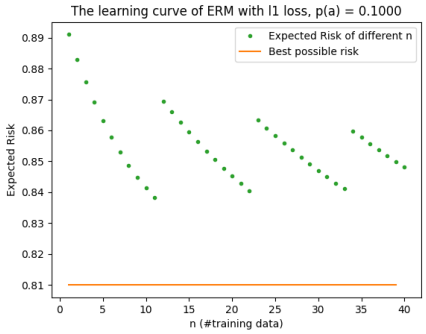
A large quantity of research investigated the learning curve for different problems or tried to find a common model for learning curves of various problems [2, 4]. Many learning curve models have been proposed, such as the exponential or logarithmic models [2, 4, 5]. While generating the learning curves for assorted problems, many unexpected behaviors of learning curves have been observed. Some of the learning curves exhibit non-monotonic behaviors. This phenomenon is opposed to the common assumption that "The more training data, the better the performance of the learner", as proposed in [4, 6, 7]. One well-known example is the sample-wise double descent learning curve, which exists not only in simple models such as linear regression, but also appears in deep neural networks [8].

In our study, we mainly investigate the two learning curves with unexpected behavior introduced by Loog *et al.* [1]. Under the two proposed problem settings, the learning curves are not decreasing monotonically, which means the generalization performance of the learner is not always improving with more data. As shown in Figure 1a, the expected risk for the first problem increases as more training data is provided. On the other hand, the learning curve of the second problem shows a periodic pattern, as presented in Figure 1b. Instead of decreasing monotonically, its decrease is followed by a jump upwards periodically. Our target is to explain these learning curves. **The main question is** why these learning curves appear? More specifically, what caused the non-monotonic behavior of the learning curve? Is such behavior caused by distribution, loss of function, learner, or combined?

The remainder of this paper is structured in the following way. Section 2 will present other works in the field of the learning curve and discuss how our work is related to them. Section 3 will introduce the problem settings of the investigated learning curves and our analysis methodology. Section 4 will display the results of the analysis. Section 5 will discuss how our study answers the research question and its limitations. Section 6 will go



(a) How the size of the training data influence the performance of $\mathcal{A}_{erm}$ with L2 loss and linear functions without intercept.



(b) How the size of the training data influence the performance of $\mathcal{A}_{erm}$ with L1 loss and linear functions without intercept.

Figure 1: Comparison of the variance and bias terms for ERM and ridge regression

through the integrity and reproducibility of our results. Eventually, section 7 will provide the conclusions and recommendations for future works.

2 Related Work

There is a large number of studies regarding learning curves in general. Many researchers have tried to find the best function model for the learning curves [2, 4, 9]. Duin [10] investigated the learning curves of a variety of algorithms to find a reasonably well-performing algorithm for small-sample-size problems. Likewise, many efforts have been paid to understand the behavior of the learning curves, and there are various assumptions about how the learning curves should behave. Haussler *et al.* [7] developed a theory to find a rigorous bound for the learning curves. Provost *et al.* [11] suggested that the learning curves should exhibit a steep decrease in error at the early stage, a more gentle decrease in the middle stage, and a plateau afterward. Some claimed that the accuracy should increase as more data is provided [4, 6, 7].

However, while investigating learning curves for various problems, many learning curves not conforming with the previous assumptions have been discovered. Such unexpected behaviors of learning curves occur in both classical machine learning algorithms [6] and in deep neural networks [12]. The learning curves of many problems exhibit the "double decent" pattern, as presented in Figure 2. In [1], it is shown that even with a simple distribution and a basic learner, the learning curves can be ill-behaved. These problems and their learning curves are the focus of this paper. Unlike most of the previously mentioned studies, which used real-life datasets with unknown distributions, Loog *et al.* [1] proposed a simple distribution and used it to generate artificial datasets. Since the distribution is known and simple, the expected risk can be calculated, instead of estimated using test sets. Thus, the possibility that the odd learning curves are caused by non-representative test sets, is safely ruled out. We will try to explain why these learning curves have unexpected behaviors.

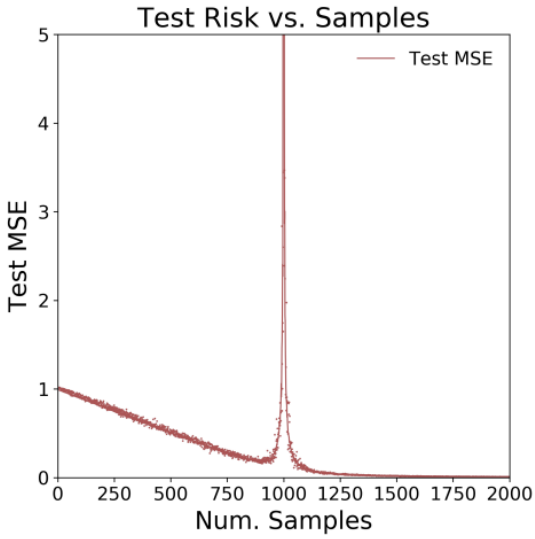


Figure 2: The double descent pattern [6]

3 Problem Setting and Methodology

We will first introduce the settings of the two problems and explain terminologies in section 3.1. Then, we will present the methodology we applied to solve the questions in section 3.2. We have adopted two disparate methods for analyzing the two problems, which will be introduced in detail separately.

3.1 Problem Setting: the Distribution and the Learners

These two problems are originally proposed in [1]. In both problem settings, the following aspects are the same:

- **The ground truth distribution** $\mathcal{D}$ is $(x, y) \in \mathbb{R} \times \mathbb{R}$, where $P(x = 1) = p_a, P(x = \frac{1}{10}) = 1 - p_a = p_b$ and $y = 1$. Simply put, the domain $\mathcal{Z}$ of this distribution is consisted of only two points $a = (1, 1)$ and $b = (\frac{1}{10}, 1)$, while the probability of $P((x, y) = a) = p_a, P((x, y) = b) = 1 - p_a = p_b$.
- **The hypothesis class** $\mathcal{H}$ is all linear functions without intercepts; i.e., $\{h(x) = \beta x | \beta \in \mathbb{R}\}$.
- **Both of them are regression problems and use ERM (empirical risk minimizer) as the learner.** A learner $\mathcal{A}$ maps the set of all possible datasets to elements in the hypothesis class, i.e. $\mathcal{A} : \mathcal{Z} \cup \mathcal{Z}^2 \cup \mathcal{Z}^3 ... \rightarrow \mathcal{H}$. A ERM $\mathcal{A}_{erm}$ is a learner which outputs the hypothesis with minimum empirical risk, given a set of training data. Let $\mathcal{L} : \mathcal{H} \rightarrow \mathbb{R}$ denote the loss function, $\mathcal{R} : \mathcal{H} \rightarrow \mathbb{R}$ denote the risk function, and $S^n = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ denote a set of samples with size n. The risk $\mathcal{R}(h)$ for a hypothesis $h \in \mathcal{H}$ is $\mathcal{R}(h) = \mathbb{E}_{(x,y)}\mathcal{L}(h)$ and the empirical risk $\hat{\mathcal{R}}(h)$, given a training dataset S^n, is $\hat{\mathcal{R}}(h) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(h)$.

The main differences between these two problem settings lie in the value of p_a and the loss function.

- **Problem I:** $p_a = 0.001$. The loss function is L2 loss. The loss of a hypothesis h for a sample (x, y) is $\mathcal{L}(h) = (h(x)-y)^2$. The empirical risk on a given training dataset with n samples is $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(h(x_i)-y_i)^2$. The expected risk is $R(h) = \mathbb{E}_{(x,y)}(h(x)-y)^2$. The empirical risk minimizer $\mathcal{A}_{erm}$ has a closed-form solution $(X^TX)^{-1}X^TY$, where $X = \begin{bmatrix} x_1, x_2, \ldots, x_n \end{bmatrix}^T$ and $Y = \begin{bmatrix} y_1, y_2, \ldots, y_n \end{bmatrix}^T$.
- **Problem II:** $p_a = 0.1$. The loss function is L1 loss. The loss of a hypothesis h for a sample (x, y) is $\mathcal{L}(h) = |h(x) - y|$. The empirical risk on a given training dataset with n samples is $\hat{R}(h) = \sum_{i=1}^{n}|h(x_i) - y_i|$. The expected risk is $R(h) = \mathbb{E}_{(x,y)}|h(x) - y|$. The empirical risk minimizer $\mathcal{A}_{erm}$ does not have a closed-form solution in general. However, it indeed has a closed-form solution when $X, Y \in \mathbb{R}$, which will be derived in section 4.2.

3.2 Disparate Methods of Analyzing the Problems

Due to the divergent nature of the two problems, we have tackled them using distinct methods. For **Problem I**, we used the bias-variance decomposition to break down the expected risk into *bias* and *variance* terms, and analyzed the resulting terms. This method has been used in [13] to explain the double descent phenomenon occurring in the learning curves of linear regression (ERM with L2 loss). Since we have a similar setting, we decided to adopt the method. We used the bias-variance decomposition here and observed the change of *bias* and *variance* terms with respect to the number of training samples. After observing the curves of these two terms, we focused on the *variance* term and further inspected what is the cause of its increase.

When interpreting the learning curve of **Problem II**, we started by finding the closed-form solution of $\mathcal{A}_{erm}$. Then we calculated the best possible solution of β, i.e, $\beta = \arg\min_{\beta\in\mathbb{R}} \mathbb{E}_{(x,y)}|\beta x - y|$. While figuring out the solutions, we discovered that the expected risk is correlated with the probability of $\mathcal{A}_{\upharpoonright\nabla\Updownarrow}$ producing a distinct hypothesis. Then, we drew our attention to that probability.

4 Analysis

In this section, we will present the analysis of the two problems using the method we introduced in section 3.2. In section 4.1, we will investigate **Problem I**. The derivation of the bias-variance decomposition will be provided, along with the analysis of the two terms. **Problem II** will be explored in section 4.2. We will show how we derive the closed-form solution of $\mathcal{A}_{erm}$ and the best possible β. Furthermore, we will explain why the probability of $\mathcal{A}_{erm}$ producing one specific hypothesis leads to the strange behavior of the learning curve.

4.1 Problem I

We applied bias-variance decomposition to the expected risk for **Problem I** and identified the cause of the increasing learning curve by observing how the resulting terms change with respect to the number of training samples. We proposed using ridge regression to solve the problem and then analyzed why the problem occurs. The bias-variance decomposition, the

use of ridge regression, and the causes of the problem will be presented in detail in the following sections.

4.1.1 Bias-Variance Decomposition

In the setting of **Problem I**, the loss function of a given hypothesis h on a sample $(x, y)$is $\mathcal{L}(h) = (h(x) - y)^2$. The risk is $R(h) = \mathbb{E}_{(x,y)}(h(x) - y)^2$. Therefore, the true risk of a fixed sample size n $\mathbb{E}_{S^n} R(A_{erm}(S^n)) = \mathbb{E}_{S^n}\mathbb{E}_{(x,y)}(\hat{\beta}x - y)^2$. This expression can be decomposed in the following way.

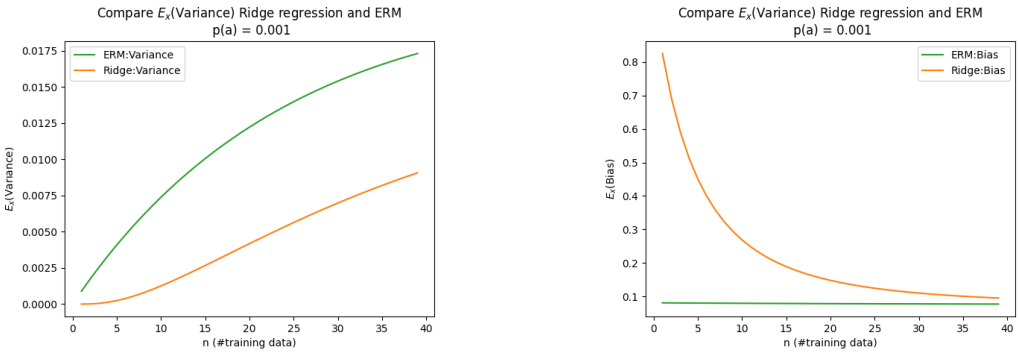
$$
\begin{aligned}
\mathbb{E}_{S^n} R(A_{erm}(S^n)) &= \mathbb{E}_{S^n}\mathbb{E}_{(x,y)}(\hat{\beta}x - y)^2, \qquad (\hat{\beta} = A_{erm}(S^n)) \\
&= \mathbb{E}_{S^n}\mathbb{E}_x(\hat{\beta}x - y)^2 \\
&= \mathbb{E}_x\mathbb{E}_{S^n}(\hat{\beta}x - \mathbb{E}_{S^n}\hat{\beta}x + \mathbb{E}_{S^n}\hat{\beta}x - y)^2 \\
&= \mathbb{E}_x\mathbb{E}_{S^n}\left\{(\hat{\beta}x - \mathbb{E}_{S^n}\hat{\beta}x)^2 + (\mathbb{E}_{S^n}\hat{\beta}x - y)^2 + 2(\hat{\beta}x - \mathbb{E}_{S^n}\hat{\beta}x)(\mathbb{E}_{S^n}\hat{\beta}x - y)\right\} \\
&= \mathbb{E}_x\left\{Var_{S^n}(\hat{\beta})x^2 + (\mathbb{E}_{S^n}\hat{\beta}x - y)^2 + \mathbb{E}_{S^n}G\right\} \\
\mathbb{E}_{S^n}G &= \mathbb{E}_{S^n}\left\{2(\hat{\beta}x - \mathbb{E}_{S^n}\hat{\beta}x)(\mathbb{E}_{S^n}\hat{\beta}x - y)\right\} \\
&= 2\mathbb{E}_{S^n}\left\{x^2\hat{\beta}\mathbb{E}_{S^n}\hat{\beta} - \hat{\beta}xy - x^2\mathbb{E}_{S^n}\hat{\beta}\mathbb{E}_{S^n}\hat{\beta} + \mathbb{E}_{S^n}\hat{\beta}xy\right\} \\
&= 2\left\{x^2\mathbb{E}_{S^n}\hat{\beta}\mathbb{E}_{S^n}\hat{\beta} - \mathbb{E}_{S^n}\hat{\beta}xy - x^2\mathbb{E}_{S^n}\hat{\beta}\mathbb{E}_{S^n}\hat{\beta} + \mathbb{E}_{S^n}\hat{\beta}xy\right\} \\
&= 0 \\
Thus,\ \mathbb{E}_{S^n} R(A_{erm}(S^n)) &= \mathbb{E}_x\left\{Var_{S^n}(\hat{\beta})x^2 + (\mathbb{E}_{S^n}\hat{\beta}x - y)^2\right\} \\
&= \mathbb{E}_x x^2 Var_{S^n}(\hat{\beta}) + \mathbb{E}_x(\mathbb{E}_{S^n}\hat{\beta}x - y)^2
\end{aligned}
$$

We call the term $\mathbb{E}_x x^2 \cdot Var_{S^n}(\hat{\beta})$ *variance* and the term $(\mathbb{E}_{S^n}\hat{\beta}x - y)^2$ *bias*. The following Figure 3 shows how $\mathbb{E}_x(bias)$ and *variance* are changing with respect to the training size n. As shown in Figure 3, the increase rate of variance terms surpasses the decrease rate of the bias term, which leads to the result of an increasing expected risk.

Figure 3: How are the Bias and Variance terms evolving with respect to the growth of n

4.1.2 Ridge Regression

We have applied ridge regression $\mathcal{A}_{ridge} : \mathcal{Z} \cup \mathcal{Z}^2 \cup \mathcal{Z}^3 ... \rightarrow \mathcal{H}$ instead, in order to decrease the variance and its increase rate. Assume $S^n = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, $\mathcal{A}_{ridge}(S^n) = \arg\min_{\beta \in \mathbb{R}} \lambda ||\beta|| + \sum_{i=1}^{n} (\beta x_i - y_i)^2$. $\lambda \in [0, +\infty)$ is a hyper-parameter controlling the strength of regularization effect. The larger the λ, the stronger the regularization effect is. The closed-form solution of $\mathcal{A}_{ridge}$ is $(X^T X + \lambda \mathbf{I})^{-1} X^T Y$, where $X = \begin{bmatrix} x_1, x_2, \ldots, x_n \end{bmatrix}^T$ and $Y = \begin{bmatrix} y_1, y_2, \ldots, y_n \end{bmatrix}^T$. Due to the regularization term in ridge regression, the algorithm is more stable, which means "a small change of the input does not change the output much." [14, Chapter 13] Therefore, the variance of the output is lower, compared to $\mathcal{A}_{erm}$, while the bias is higher. As shown in Figure 4a, with $\lambda = 0.1$, the variance is lowered and grows at a slower rate. Whereas, the bias starts at a higher value and decreased faster, as shown in Figure 4b.



(a) Compare $\mathbb{E}_x(x^2) \cdot variance$

(b) Compare $\mathbb{E}_x(bias)$

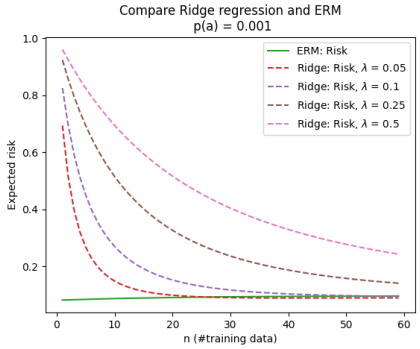
Figure 4: Comparison of the variance and bias terms for ERM and ridge regression

We have experimented further with $\lambda = \{0.05, 0.1, 0.25, 0.5\}$. All learning curves of $\mathcal{A}_{ridge}$ with various λ values are decreasing monotonically for $n = 1, 2, ..., 40$, as depicted in Figure 5. Figure 5b shows that with a relatively small λ, $\mathcal{A}_{ridge}$ can achieve a lower expected risk compared to $\mathcal{A}_{erm}$, when n is large enough. We can also observe that for $\lambda = 0.05$, the learning curve starts to increase again as n increases. This phenomenon can be attributed to the weak regularization effect posed the small λ value, as the other learning curves are still decreasing.

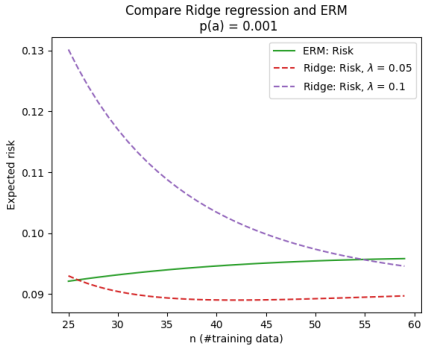
4.1.3 Explain *Variance* Increase

After decomposing the true risk and observing the behavior of the bias and variance terms, we have further investigated what causes the increase in variance, which is contrary to the intuition that a larger number of training samples should lead to a lower variance. We proposed that the increase in variance is caused by the fact that this distribution does not fit the linear model $Y = \beta X + \epsilon$, where $\mathbb{E}\epsilon = 0$, and X and ϵ are independent with each other. Therefore cannot be learned by the $\mathcal{A}_{erm}$ here, which is performing the ordinary least squares method for linear regression.

To prove our claim, we have constructed a similar distribution with four points $a_1 = [1, \frac{3}{2}], a_2 = [1, \frac{1}{2}], b_1 = [\frac{3}{4}, \frac{5}{4}], b_2 = [\frac{3}{4}, \frac{1}{4}]$, each with probability $\frac{1}{2}p_a$, $\frac{1}{2}p_a$, $\frac{1}{2}p_b$, $\frac{1}{2}p_b$. This



(a) Overview of performance with different lambda



(b) Zoom in on $\lambda = \{0.05, 0.1\}$ and $n = [25, 60]$

Figure 5: How the size of the training data influence the performance of $\mathcal{A}_{ridge}$ with different λ and linear functions without intercept.

distribution can be reformulated as $Y = \beta X + \epsilon$. $\beta = 1$, X is a discrete random variable with two possible outcomes: 1 and $\frac{3}{4}$; $P(X = 1) = p_a$ and $P(X = \frac{3}{4}) = 1 - p_a = p_b$. ϵ is also a random variable with two possible outcomes: $\frac{1}{2}$ and $-\frac{1}{2}$. Each outcome has a probability of $\frac{1}{2}$ and $\mathbb{E}\epsilon = \frac{1}{2} \cdot \frac{1}{2} + (-\frac{1}{2}) \cdot \frac{1}{2} = 0$. Therefore, this distribution with four points fits the linear model. As shown in Figure 6, the $Var_{S^n}(\hat{\beta})$ decreases as the number of training samples increases.

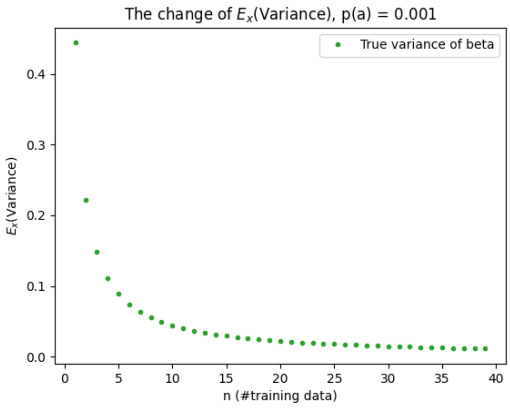


Figure 6: How the size of the training data influence $Var_{S^n}\hat{\beta}$

We further inspected the term $Var_{S^n}(\hat{\beta})$ to verify that the linear model property of this distribution leads to the decrease of variance. In order to investigate the variance, we first

calculate $\mathbb{E}_{S^n}\hat{\beta}$.

$$\begin{aligned}\mathbb{E}_{S^n}\hat{\beta} &= \mathbb{E}_{S^n}(X_n^T X_n)^{-1}X_n^T Y \\ &= \mathbb{E}_{S^n}(X_n^T X_n)^{-1}X_n^T(X_n\beta+\epsilon_n) \\ &= \mathbb{E}_{S^n}(X_n^T X_n)^{-1}(X_n^T X_n)\beta + (X_n^T X_n)^{-1}X_n^T\mathbb{E}_{S^n}\epsilon_n \\ \text{Since } \mathbb{E}_{S^n}\epsilon_n = 0, &\quad \mathbb{E}_{S^n}\hat{\beta} = \beta\end{aligned}$$

Then we investigated $Var_{S^n}(\hat{\beta})$:

$$\begin{aligned}Var_{S^n}(\hat{\beta}) &= \mathbb{E}_{S^n}(\hat{\beta}-\beta)^2 \\ &= \mathbb{E}_{X^n}\mathbb{E}_{\epsilon_n}(\hat{\beta}-\beta)^2 \\ &= \mathbb{E}_{X^n}\mathbb{E}_{\epsilon_n}((X_n^T X_n)^{-1}X_n^T(X_n\beta+\epsilon_n)-\beta)^2 \\ &= \mathbb{E}_{X^n}\mathbb{E}_{\epsilon_n}((X_n^T X_n)^{-1}X_n^T X_n\beta + (X_n^T X_n)^{-1}X_n^T\epsilon_n-\beta)^2 \\ &= \mathbb{E}_{X^n}\mathbb{E}_{\epsilon_n}((X_n^T X_n)^{-1}X_n^T\epsilon_n - 0)^2 \\ &= \mathbb{E}_{X^n}Var_{\epsilon_n}((X_n^T X_n)^{-1}X_n^T\epsilon_n) \\ &= \mathbb{E}_{X^n}\left[(X_n^T X_n)^{-1}X_n^T Var_{\epsilon_n}\epsilon_n X_n(X_n^T X_n)^{-1}\right] \\ \text{Since all training samples are i.i.d },Var_{\epsilon_n}\epsilon_n &= Var(\epsilon)\cdot {}_n\mathbb{I}_n \\ &= \mathbb{E}_{X^n}\left[Var(\epsilon)(X_n^T X_n)^{-1}(X_n^T X_n)(X_n^T X_n)^{-1}\right] \\ &= Var(\epsilon)\mathbb{E}_{X^n}(X_n^T X_n)^{-1} \\ &= Var(\epsilon)\mathbb{E}_{X^n}(\sum_{i=1}^{n}x_i^2)^{-1}\end{aligned}$$

Since $\mathbb{E}_{X^n}(\sum_{i=1}^{n}x_i^2)^{-1}$ decreases as n increases, the variance also decreases as n increases.

4.2 Problem II

To explain the unexpected periodic behavior of the learning curve for **Problem II**, we first deduced the closed-form solution for both $\mathcal{A}_{erm}$ and the optimal β. Then, we discovered that expected risk is positively correlated to the probability of $\mathcal{A}_{erm}$ outputting one specific hypothesis, whose curve also has a periodic pattern. We then further investigated why the curve that probability has periodic behavior. The details will be provided in the following sections.

4.2.1 Closed-form Solution of $\mathcal{A}_{erm}$ and the Optimal β

In the setting of **Problem II**, the loss of a hypothesis h on a sample (x, y) is $\mathcal{L}(h) = |h(x) - y|$.The empirical risk on a given training dataset with n samples is $\hat{R}(h) = \sum_{i=1}^{n}|h(x_i) - y_i|$. Therefore, $\mathcal{A}_{erm} = \arg\min_{h\in\mathcal{H}}\hat{R}(h)$, where $\mathcal{H} = \{h(x) = \beta x|\beta \in \mathbb{R}\}$. So $\mathcal{A}_{erm}$ can also be expressed as $\mathcal{A}_{erm} = \arg\min_{\beta\in\mathbb{R}}\sum_{i=1}^{n}|\beta x_i - y_i|$. We first derive the closed form solution for $\mathcal{A}_{erm}$. Let n denote the size of the training dataset, n_a denote the number of point

$a = (x_a, y_a)$ in the training dataset, and n_b denote that of point $b = (x_b, y_b)$.

$$\sum_{i=1}^{n} |\beta x_i - y_i| = n_a|\beta x_a - y_a| + n_b|\beta x_b - y_b|$$

(1) For $\beta \in [0, \frac{y_a}{x_a})$

$$\begin{aligned}\frac{d}{d\beta}(n_a|\beta x_a - y_a| + n_b|\beta x_b - y_b|) &= \frac{d}{d\beta}(n_a(y_a - x_a\beta) + n_b(y_b - x_b\beta)) \\ &= -n_a x_a - n_b x_b \\ &< 0\end{aligned}$$

(2) For $\beta \in [\frac{y_a}{x_a}, \frac{y_b}{x_b}]$

$$\begin{aligned}\frac{d}{d\beta}(n_a|\beta x_a - y_a| + n_b|\beta x_b - y_b|) &= \frac{d}{d\beta}(n_a(x_a\beta - y_a) + n_b(y_b - x_b\beta)) \\ &= n_a x_a - n_b x_b\end{aligned}$$

(3) For $\beta \in (\frac{y_b}{x_b}, +\infty)$

$$\begin{aligned}\frac{d}{d\beta}(n_a|\beta x_a - y_a| + n_b|\beta x_b - y_b|) &= \frac{d}{d\beta}(n_a(x_a\beta - y_a) + n_b(x_b\beta - y_b)) \\ &= n_a x_a + n_b x_b \\ &> 0\end{aligned}$$

If $n_a x_a - n_b x_b \geq 0$, then the derivative is only negative when $\beta \in [0, \frac{y_a}{x_a})$, which means the function stops decreasing when $\beta \geq \frac{y_a}{x_a}$. Therefore, the minimum of this function is reached at the point $\beta = \frac{y_a}{x_a}$. In the other case, when $n_a x_a - n_b x_b < 0$, the the derivative is negative when $\beta \in [0, \frac{y_b}{x_b})$, which means the function stops decreasing when $\beta \geq \frac{y_b}{x_b}$. Therefore, the minimum of this function is reached at the point $\beta = \frac{y_b}{x_b}$. The closed form solution of $\mathcal{A}_{erm}$ is thus the following.

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

The same procedure is applied to find $\arg\min_{h \in \mathcal{H}} R(h)$, or equivalently $\beta = \arg\min_{\beta \in \mathbb{R}} \mathbb{E}_{(x,y)}|\beta x - y|$.

$$\mathbb{E}_{(x,y)}|\beta x - y| = p_a|\beta x_a - y_a| + p_b|\beta x_b - y_b|$$

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

Under this problem setting, $p_a x_a - p_b x_b = \frac{1}{10} \cdot 1 - \frac{1}{10} \cdot \frac{9}{10} > 0$, $\beta = \frac{y_a}{x_a}$.

4.2.2 Analysis of the Expected Risk

We then analyzed the expected risk for a given n. Let $P_{S^n}(\hat{\beta} = \rho)$ denote the probability of $\mathcal{A}_{erm}$ outputting ρ when the size of the training dataset is n, $\frac{y_a}{x_a} = \beta$ as $\hat{\beta} = \hat{\beta}_1$, $\frac{y_b}{x_b} \neq \beta$ as

$\hat{\beta}_2$, $P_{S^n}(\hat{\beta} = \hat{\beta}_1)$ as P_1^n, and $P_{S^n}(\hat{\beta} = \hat{\beta}_2)$ as $P_2^n = 1 - P_1^n$

$$\begin{aligned}
\mathbb{E}_{S^n} R(A_{erm}(S^n)) &= \mathbb{E}_{S^n} \mathbb{E}_{(x,y)} |\hat{\beta} x - y| \\
&= \mathbb{E}_{(x,y)} \mathbb{E}_{S^n} |\hat{\beta} x - y| \\
&= \mathbb{E}_{(x,y)} (P_1^n |\hat{\beta}_1 x - y| + P_2^n |\hat{\beta}_2 x - y|) \\
&= P_1^n \cdot \mathbb{E}_{(x,y)} |\hat{\beta}_1 x - y| + P_2^n \cdot \mathbb{E}_{(x,y)} |\hat{\beta}_2 x - y|
\end{aligned}$$

As $\hat{\beta}_1 = \arg\min_{\beta \in \mathbb{R}} \mathbb{E}_{(x,y)} |\beta x - y|$, the smaller P_2^n is, the larger P_1^n and the smaller the true risk for n. Therefore, we further investigate how P_2^n changes with respect to the number of training samples. As shown in Figure 7, P_2^n follows the same decreasing followed by sudden increase periodic pattern. If we change the value of p_a such that $p_a x_a - p_b x_b < 0$, then

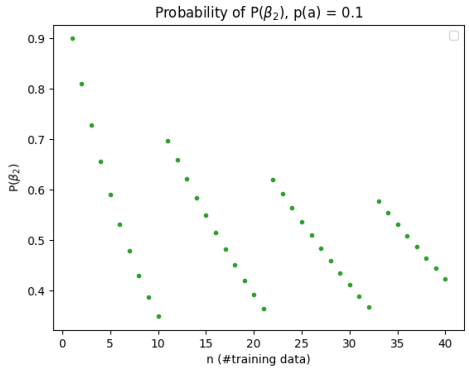


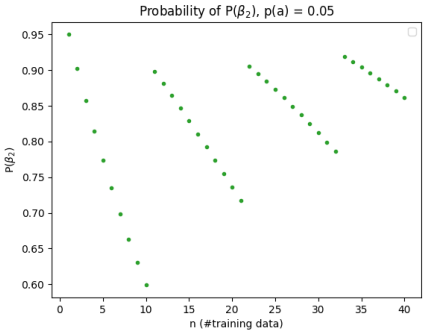
Figure 7: The change of P_2^n with respect to n

$\beta = \frac{y_b}{x_b}$. In this case, $\hat{\beta}_2 = \arg\min_{\beta \in \mathbb{R}} \mathbb{E}_{(x,y)} |\beta x - y|$ and the smaller P_2^n is, the larger the true risk. This claim is supported by the following setting. We set $p_a = 0.05$ and all the other values remain the same, then $p_a x_a - p_b x_b = 0.05 \cdot 1 - 0.95 \cdot \frac{1}{10} = -0.045 < 0$. The learning curve and the curve for P_2^n under this setting is displayed in Figure 8. The curve of P_2^n still has the "decrease then increase " periodic behavior. Whereas, the learning curve, instead of following the same trend, exhibits the complete opposite behavior. It increases, while the curve of P_2^n decreases.

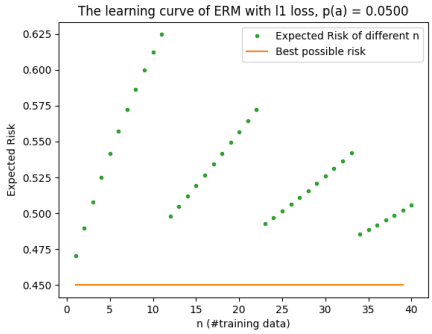
4.2.3 Explaining the Periodic Pattern of P_2^n

The question of why the learning curve behaves as in Figure 1b can be reduced to the question why the curve of P_2^n has the behavior shown in Figure 7. In order to investigate P_2^n, we need to understand when will $\mathcal{A}_{erm}$ output $\hat{\beta}_2$.

$$\hat{\beta} = \frac{y_b}{x_b} = \hat{\beta}_2 \quad \text{if } n_a x_a - n_b x_b < 0$$



(a) The learning curve when $p_a = 0.05$



(b) The change of P_2^n when $p_a = 0.05$

Figure 8: The curve of P_2^n and the learning curve when $\hat{\beta}_2 = \beta$

Thus, for a given n, $P_2^n = P_{S^n}(n_a x_a - n_b x_b < 0)$. n_b can be substituted by $n - n_a$.

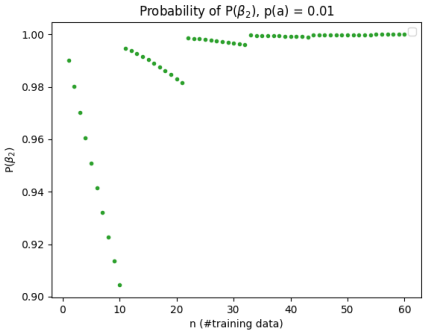
$$\begin{aligned} n_a x_a - n_b x_b &< 0 \\ n_a x_a - (n - n_a) x_b &< 0 \\ n_a (x_a + x_b) &< n x_b \\ n_a &< \frac{x_b}{x_a + x_b} n \\ n_a &< \frac{n}{\frac{x_a}{x_b} + 1} \end{aligned}$$

$P_{S^n}(n_a x_a - n_b x_b < 0) = P_{S^n}(n_a < \frac{n}{\frac{x_a}{x_b} + 1}) = \sum_{i \in N_A} P_{S^n}(n_a = i)$, where $N_A = \{i \in \mathbb{N} |\ i < \frac{n}{\frac{x_a}{x_b} + 1}\}$. Since $n, i \in \mathbb{N}$, $|N_A|$ increases by 1, when n increases by $\lceil \frac{x_a}{x_b} + 1 \rceil$. In this problem setting $\lceil \frac{x_a}{x_b} + 1 \rceil = 11$ and as shown in the Figures 1b and 7, the curves have a sudden increase when $11|n$. Therefore, we claim that the increase of $|N_A|$ is the cause of the sudden increase. In order to prove this, we observe $\sum_{i \in N_A} P_{S^n}(n_a = i)$ before and after $|N_A|$ increase by 1. The proof of this claim is provided in Appendix A. We also claim that P_2^n decreases while n increases and $|N_A|$ stays the same, the proof of which is provided in Appendix B.

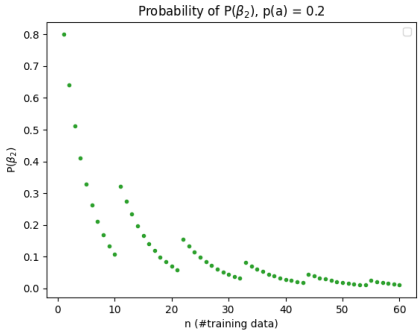
Therefore, we concluded that the shape of the curve showing how P_2^n changes with respect to n will always demonstrate such periodic pattern regardless of the value of p_a. As shown in Figure 9, with either larger or smaller values of p_a the curve of P_2^n still displays the same periodic pattern, which is sudden increase after a fixed period of decrease. Moreover, the duration of one period is dependent on $\lceil \frac{x_a}{x_b} + 1 \rceil$. As illustrated in Figure 10, the duration of one period is always equal to $\lceil \frac{x_a}{x_b} + 1 \rceil$.

5 Discussion

We aimed to explain why the learning curves generated under the two problem settings have unexpected behaviors. For **Problem I**, the study demonstrates a correlation between the increasing learning curve and the distribution not fitting the linear model. We adopted the bias-variance decomposition to split the expected risk into bias and variance terms.

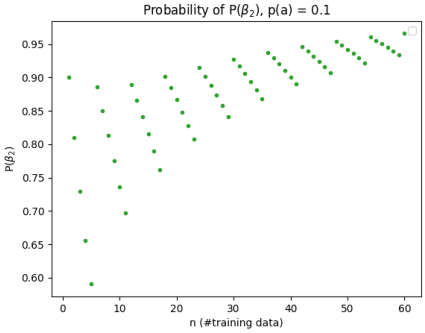


(a) When $p_a = 0.01$

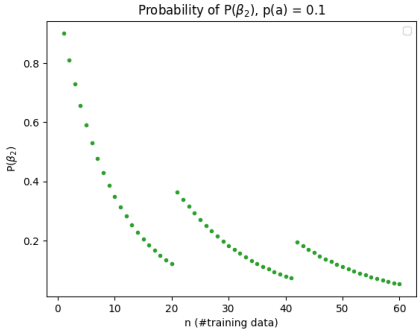


(b) When $p_a = 0.2$

Figure 9: The behavior of P_2^n with different values of p_a



(a) When $x_b = \frac{1}{5}, \lceil \frac{x_a}{x_b} + 1 \rceil = 6$



(b) When $x_b = \frac{1}{20}, \lceil \frac{x_a}{x_b} + 1 \rceil = 21$

Figure 10: The behavior of P_2^n with different values of $\lceil \frac{x_a}{x_b} + 1 \rceil$

The visualization and analysis of these two terms have shown that the rapid increase in variance and the, in contrast, slower decrease in bias leads to the ascending learning curve. We switched to using Ridge Regression to suppress the rapid increase in variance, which results in learning curve decreasing in the same range. It is unusual for variances to increase with more training samples. We suggested that this increase is caused by the fact that this distribution does not fit the linear model. We supported our claim by proposing a similar distribution, conforming to the linear model, and proved that its variance decreases with more training samples.

For **Problem II**, we discovered that the probability of A_{erm} outputting one hypothesis leads to the periodic pattern of the learning curve. Through the analysis of $\hat{\beta}$ and the optimal β, we discovered that A_{erm} only has two possible outcomes, with one equaling to the optimal β denoted as β_1 and the other denoted as β_2. For the sake of simplicity, we use p_1 to refer the probability of A_{erm} outputting β_1 and p_2 for that of β_2. It is shown, with further derivation, that the expected risk is positively correlated with p_2. After plotting how p_2 changes with respect to the number of training samples, we discovered the same periodic pattern as observed in the learning curve. Therefore, the problem was reduced to why the

curve of p_2 has the periodic behavior, which was answered in section 4.2.3.

Even though we have provided detailed explanations for the unexpected behavior of learning curves, there are still some limitations in this study. During the analysis for **Problem I**, to back our claim that the distribution not conforming with the linear model caused the increase in variance, we provided a concrete example. However, presenting only one example is not adequate for rigorously proving our claim. Regarding **Problem II**, we have explained why the learning curve has the periodic pattern shown in Figure 1b and 8b. Whereas, we have not answered whether, with more and more training data, the learning curve will converge to the lowest possible risk. Besides, our results are limited to these two specific problem settings and are not yet generalized to other problems.

6 Responsible Research

Our study used an artificial distribution to generate all the datasets and did not involve any real-life data, thus preventing data breaches from happening. All results and figures are authentic. Throughout our experiments, we have never altered the results to support our claims. We have given credits to and added references in the paper whenever we adopt ideas from other works. Moreover, our study is easily reproducible. The derivations can be reproduced following the steps we provided in the paper. The experiments are written in `Python` using `numpy`, `sympy` and `matplotlib` libraries. Note that we used the `sympy` library to increase the precision for floats and `matplotlib` to generate the figures. As the repository is not public, if anyone is interested in viewing the original code, please contact the author.

7 Conclusions and Future Work

Our study focused on the strange learning curves introduced in Loog *et al.* [1]. The learning curves are generated from two similar regression problems while using ERM as the learner. The two problems are only different in loss function and one parameter value. We adopted disparate methods for the two problem settings. For **Problem I**, we used bias-variance decomposition to inspect the expected risk. For **Problem II**, we analyzed the closed-form solution of the ERM and the optimal hypothesis β.

We have answered our research question for each problem. Regarding **Problem I**, we discovered that the rapid increase in *variance* and the relatively slow decrease in *bias*, with respect to the number of training samples, lead to the ascending learning curve. In addition, we proposed that the increase in variance is caused by the distribution not conforming with the linear model. To strengthen our argument, we constructed a similar distribution yet conforming with the linear model and proved that in this case, the variance decreases. As for **Problem II**, our analysis showed that the ERM only output two possible hypotheses: β_1 equals the optimal β and β_2. The expected risk is positively correlated with the probability of ERM outputting β_2. The curve of this probability also presents the periodic pattern, which produced the strange learning curve. As proved in Appendix A and B, this curve will display the same periodic pattern regardless of certain changes in the problem setting.

Our study has left room for several follow-up questions. One possible research question for future study is to analyze whether the learning curves of both questions will converge to the optimal risk. If they will, how fast will they converge? Is there any upper bound on the expected risk given the size of the training dataset. The learning problems discussed here use artificial distributions. Whereas, the odd learning curves generated for real-life learning

problems also seem of interest to us. Whether some of the unexpected learning curves generated from real-life problems are also caused by similar reasons is another question worth studying.

References

[1] M. Loog, T. Viering, and A. Mey, "Minimizers of the empirical risk and risk monotonicity," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 7478–7487.

[2] L. J. Frey and D. H. Fisher, "Modeling decision tree performance with the power law," in *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. R2, 03–06 Jan 1999. [Online]. Available: https://proceedings.mlr.press/r2/frey99a.html

[3] P. Kolachina, N. Cancedda, M. Dymetman, and S. Venkatapathy, "Prediction of learning curves in machine translation," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, july 2012, pp. 22–30. [Online]. Available: https://aclanthology.org/P12-1003

[4] B. Gu, F. Hu, and H. Liu, "Modelling classification performance for large data sets," in *Advances in Web-Age Information Management*, 2001, pp. 317–328.

[5] G. H. John and P. Langley, "Static versus dynamic sampling for data mining," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, pp. 367–370.

[6] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: where bigger models and more data hurt," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124003, dec 2021. [Online]. Available: https://doi.org/10.1088/1742-5468/ac3a74

[7] D. Haussler, M. Kearns, H. S. Seung, and N. Tishby, "Rigorous learning curve bounds from statistical mechanics," *Machine Learning*, vol. 25, pp. 195–236, 1996. [Online]. Available: https://doi.org/10.1007/BF00114010

[8] S. d'Ascoli, L. Sagun, and G. Biroli, "Triple descent and the two kinds of overfitting: where and why do they appear?" *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124002, dec 2021. [Online]. Available: https://doi.org/10.1088/1742-5468/ac3909

[9] M. Last, "Predicting and optimizing classifier utility with the power law," in *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, Oct 2007, pp. 219–224.

[10] R. Duin, "Small sample size generalization," in *9th Scandinavian Conference on Image Analysis*, June 1995, pp. 957–964.

[11] F. Provost, D. Jensen, and T. Oates, "Efficient progressive sampling," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '99, 1999, pp. 23–32. [Online]. Available: https://doi.org/10.1145/312129.312188

[12] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, july 2019. [Online]. Available: https://doi.org/10.1073%2Fpnas.1903070116

[13] P. Nakkiran, "More data can hurt for linear regression: Sample-wise double descent," 2019. [Online]. Available: https://arxiv.org/abs/1912.07242

[14] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning - From Theory to Algorithms.* USA: Cambridge University Press, 2014.

A Appendix

Claim A.1. *whenever n increases by* $\lceil \frac{x_a}{x_b} + 1 \rceil$*, which lead to the increase of* $|N_A|$ *by 1,* P_2^n *increases.*

Proof. Assume when $n = m$, $N_A = 0, 1, ..., k$ and when $n = m + 1$, $N_A = 0, 1, ..., k + 1$

$$\begin{aligned}
\sum_{i=0}^{k} P_{S^m}(n_a = i) &= \binom{m}{0}(1-p_a)^m + \binom{m}{1}(1-p_a)^{m-1}p_a^1 \\
&\quad + \binom{m}{2}(1-p_a)^{m-2}p_a^2 + ... + \binom{m}{k}(1-p_a)^{m-k}p_a^k \\
&= (1-p_a)^{m-k}\Bigg\{\binom{m}{0}(1-p_a)^k + \binom{m}{1}(1-p_a)^{k-1}p_a^1 \\
&\qquad \underbrace{+ \binom{m}{2}(1-p_a)^{k-2}p_a^2 + ... + \binom{m}{k}p_a^k\Bigg\}}_{(1)} \\
&= (1-p_a)^{m-k}\underbrace{\sum_{i=0}^{k}\binom{m}{i}(1-p_a)^{k-i}p_a^i}_{(1)}
\end{aligned}$$

We can apply the same procedure when $n = m + 1$.

$$\begin{aligned}
\sum_{i=0}^{k+1} P_{S^{m+1}}(n_a = i) &= \binom{m+1}{0}(1-p_a)^{m+1} + \binom{m+1}{1}(1-p_a)^m p_a^1 \\
&\quad + \binom{m+1}{2}(1-p_a)^{m-1}p_a^2 + ... + \binom{m+1}{k+1}(1-p_a)^{m-k}p_a^{k+1} \\
&= (1-p_a)^{m-k}\Big\{\binom{m+1}{0}(1-p_a)^{k+1} + \binom{m+1}{1}(1-p_a)^k p_a^1 \\
&\qquad \underbrace{+ \binom{m+1}{2}(1-p_a)^{k-1}p_a^2 + ... + \binom{m+1}{k+1}p_a^{k+1}\Big\}}_{(2)} \\
&= (1-p_a)^{m-k}\underbrace{\sum_{i=0}^{k+1}\binom{m+1}{i}(1-p_a)^{k+1-i}p_a^i}_{(2)}
\end{aligned}$$

We first ignore the common factor $(1-p_a)^{m-k}$ and focus on the parts (1) and (2), which causes the differences. Take arbitrary $0 \le j \le k$ and assume j is even. We examine the coefficients of p_a^j in (1) and (2), denoted as c_1^j and c_2^j.

$$\begin{aligned}
c_1^j &= \binom{m}{0}\cdot\binom{k}{j} - \binom{m}{1}\cdot\binom{k-1}{j-1} + \binom{m}{2}\cdot\binom{k-2}{j-2} - ... + \binom{m}{j} \\
&= \sum_{i=0}^{j}(-1)^{j-i}\binom{m}{i}\cdot\binom{k-i}{j-i} \\
c_2^j &= \binom{m+1}{0}\cdot\binom{k+1}{j} - \binom{m+1}{1}\cdot\binom{k}{j-1} + \binom{m+1}{2}\cdot\binom{k-1}{j-2} - ... + \binom{m+1}{j} \\
&= \sum_{i=0}^{j}(-1)^{j-i}\binom{m+1}{i}\cdot\binom{k+1-i}{j-i}
\end{aligned}$$

We use the theorem $\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$ to decompose terms in c_2^j. Note $\binom{n}{k} = 0$, when $k \in \mathbb{Z}_{<0}$

$$\begin{aligned}
c_2^j &= \sum_{i=0}^{j}(-1)^{j-i}\binom{m+1}{i}\cdot\binom{k+1}{j} \\
&= \sum_{i=0}^{j}(-1)^{j-i}\Bigg[\underbrace{\binom{m}{i}\cdot\binom{k+1-i}{j-i}}_{\alpha} + \underbrace{\binom{m}{i-1}\cdot\binom{k+1-i}{j-i}}_{\gamma}\Bigg]
\end{aligned}$$

After the decomposition, we calculate $c_2^j - c_1^j$ by subtracting the similar terms in c_1^j from the terms labeled with α in c_2^j. The resulting terms of each such subtraction are labeled

with r_α

$$c_2^j - c_1^j = \sum_{i=0}^{j} (-1)^{j-i} \left[\underbrace{\binom{m}{i} \cdot \binom{k-i}{j-i-1}}_{r_\alpha} + \underbrace{\binom{m}{i-1} \cdot \binom{k-(i-1)}{j-i}}_{\gamma} \right]$$

Denote $\binom{m}{i} \cdot \binom{k-i}{j-i-1}$ as a_i, then $\binom{m}{i-1} \cdot \binom{k-(i-1)}{j-i}$ is a_{i-1}

$$\begin{aligned} c_2^j - c_1^j &= \sum_{i=0}^{j} (-1)^i (a_i + a_{i-1}) \\ &= (a_{(-1)} + a_0) - (a_0 + a_1) + ... + (a_{j-1} + a_j) \\ &= a_{(-1)} + a_j \\ &= \binom{m}{-1} \cdot \binom{k+1}{j} + \binom{m}{j} \cdot \binom{k-j}{-1} \\ &= 0 \end{aligned}$$

As shown in the result each term in $c_j^2 - c_j^1$ labeled with r_α will be cancelled by the next term labeled with γ. Thus, only the last r_α term will be left, which is 0. Following a similar procedure, we can derive that $c_j^2 - c_j^1$ is also 0 when j is odd. Let us now consider $j = k+1$.

Lemma A.1. $\binom{n}{i} - \binom{n}{i-1} > 0$, *when* $1 \leq i < \dfrac{n+1}{2}$

Proof.

$$\begin{aligned} \binom{n}{i-1} &= \frac{n!}{(n-i+1)!(i-1)!} \\ \binom{n}{i} &= \frac{n!}{(n-i)!(i)!} \\ \binom{n}{i-1} &= \binom{n}{i} \cdot \frac{i}{n-i+1} \\ \binom{n}{i} - \binom{n}{i-1} &= \binom{n}{i} \left(1 - \frac{i}{n-i+1}\right) \end{aligned}$$

For $1 \leq i < \dfrac{n+1}{2}$, $\qquad \dfrac{i}{n-i+1} < \dfrac{\frac{n+1}{2}}{n - \frac{n+1}{2} + 1} = 1$

Since $\binom{n}{i} > 0, \left(1 - \dfrac{i}{n-i+1}\right) > 0 \implies \binom{n}{i} - \binom{n}{i-1} > 0$

□

When k is odd, $k+1$ is even

$$c_1^j = 0, \text{since the highest order of } p_a \text{ is } k$$
$$c_2^j = \binom{m+1}{0} - \binom{m+1}{1} + \binom{m+1}{2} - ... + \binom{m+1}{k+1}$$
$$= \sum_{i=1}^{\frac{(k+1)}{2}} \left\{ \binom{m+1}{2i} - \binom{m+1}{2i-1} \right\} + \binom{m+1}{0}$$

$$k+1 \leq \frac{m+1}{\frac{x_a}{x_b}+1} < \frac{m+1}{2},\ 2i \leq k+1 < \frac{m+1}{2} \implies \sum_{i=1}^{\frac{(k+1)}{2}} \left\{ \binom{m+1}{2i} - \binom{m+1}{2i-1} \right\} \geq 0$$

$$c_2^j \geq \binom{m+1}{0} > 0$$

When k is even, $k+1$ is odd

$$c_1^j = 0, \text{since the highest order of } p_a \text{ is } k$$
$$c_2^j = -\binom{m+1}{0} + \binom{m+1}{1} - \binom{m+1}{2} - ... + \binom{m+1}{k+1}$$
$$= \sum_{i=0}^{\frac{k}{2}} \left\{ \binom{m+1}{2i+1} - \binom{m+1}{2i} \right\}$$

$$k \leq \frac{m+1}{\frac{x_a}{x_b}+1} < \frac{m+1}{2},\ 2i \leq k < \frac{m+1}{2} \implies \sum_{i=0}^{\frac{k}{2}} \left\{ \binom{m+1}{2i} - \binom{m+1}{2i-1} \right\} > 0$$

$$c_2^j > 0$$

Therefore, in (2) - (1) the coefficients of all p_a^j, $0 \leq j \leq k$ are zero and the coefficient of p_a^{k+1} is positive. Therefore, (2) - (1) > 0. Since $(1-p_a)^{m-k}$ is also positive,

$$\sum_{i=0}^{k} P_{S^{m+1}}(n_a = i) > \sum_{i=0}^{k+1} P_{S^m}(n_a = i)$$

Therefore, there will always be an increase in p_2^n whenever n increases by $\lceil \frac{x_a}{x_b} + 1 \rceil$, which leads to the increase of $|N_A|$ by 1. □

B Appendix

Claim B.1. *whenever n increases without increasing $|N_A|$, P_2^n decreases.*

Proof. Since $|N_A|$ stays the same, assume when $n = m$, $N_A = 0, 1, ..., k$ and when $n = m+1$,

$N_A = 0, 1, ..., k$

$$\sum_{i=0}^{k} P_{S^m}(n_a = i) = \binom{m}{0}(1-p_a)^m + \binom{m}{1}(1-p_a)^{m-1}p_a^1$$
$$+ \binom{m}{2}(1-p_a)^{m-2}p_a^2 + ... + \binom{m}{k}(1-p_a)^{m-k}p_a^k$$
$$= (1-p_a)^{m-k}\Bigg\{\binom{m}{0}(1-p_a)^k + \binom{m}{1}(1-p_a)^{k-1}p_a^1$$
$$\underbrace{+ \binom{m}{2}(1-p_a)^{k-2}p_a^2 + ... + \binom{m}{k}p_a^k\Bigg\}}_{(1)}$$
$$= (1-p_a)^{m-k}\underbrace{\sum_{i=0}^{k}\binom{m}{i}(1-p_a)^{k-i}p_a^i}_{(1)}$$

We can apply the same procedure when $n = m + 1$.

$$\sum_{i=0}^{k} P_{S^{m+1}}(n_a = i) = \binom{m+1}{0}(1-p_a)^{m+1} + \binom{m+1}{1}(1-p_a)^m p_a^1$$
$$+ \binom{m+1}{2}(1-p_a)^{m-1}p_a^2 + ... + \binom{m+1}{k}(1-p_a)^{m+1-k}p_a^k$$
$$= (1-p_a)^{m-k}\Bigg\{\binom{m+1}{0}(1-p_a)^{k+1} + \binom{m+1}{1}(1-p_a)^k p_a^1$$
$$\underbrace{+ \binom{m+1}{2}(1-p_a)^{k-1}p_a^2 + ... + \binom{m+1}{k}(1-p_a)p_a^k\Bigg\}}_{(2)}$$
$$= (1-p_a)^{m-k}\underbrace{\sum_{i=0}^{k}\binom{m+1}{i}(1-p_a)^{k+1-i}p_a^i}_{(2)}$$

Following the same procedure of proving **Claim** A.1, we can conclude that the coefficient of all terms p_a^j is zero, when $0 \leq j \leq k$. Let us now consider $j = k + 1$, using **Lemma A.1** as in proof of **Claim A.1** .

When k is odd

$$c_1^j = 0, \text{since the highest order of } p_a \text{ is } k$$
$$c_2^j = \binom{m+1}{0} - \binom{m+1}{1} + \binom{m+1}{2} - ... - \binom{m+1}{k}$$
$$= \sum_{i=0}^{\frac{(k-1)}{2}} \left\{ \binom{m+1}{2i} - \binom{m+1}{2i+1} \right\}$$

$$k-1 < \frac{m+1}{\frac{x_a}{x_b}+1} < \frac{m+1}{2},\ 2i \le k-1 < \frac{m+1}{2} \implies \sum_{i=0}^{\frac{(k-1)}{2}} \left\{ \binom{m+1}{2i} - \binom{m+1}{2i+1} \right\} < 0$$
$$c_2^j < 0$$

When k is even

$$c_1^j = 0, \text{since the highest order of } p_a \text{ is } k$$
$$c_2^j = -\binom{m+1}{0} + \binom{m+1}{1} - \binom{m+1}{2} - ... - \binom{m+1}{k}$$
$$= \sum_{i=1}^{\frac{k}{2}} \left\{ \binom{m+1}{2i-1} - \binom{m+1}{2i} \right\} - \binom{m+1}{0}$$

$$k \le \frac{m+1}{\frac{x_a}{x_b}+1} < \frac{m+1}{2},\ 2i \le k < \frac{m+1}{2} \implies \sum_{i=1}^{\frac{k}{2}} \left\{ \binom{m+1}{2i-1} - \binom{m+1}{2i} \right\} \le 0$$
$$c_2^j \le -\binom{m+1}{0} < 0$$

Therefore, in (2) - (1) the coefficients of all p_a^j, $0 \le j \le k$ are zero and the coefficient of p_a^{k+1} is negative. Therefore, (2) - (1) < 0. Since $(1-p_a)^{m-k}$ is also positive,

$$\sum_{i=0}^{k} P_{S^{m+1}}(n_a = i) < \sum_{i=0}^{k} P_{S^m}(n_a = i)$$

Therefore, whenever n increases while $|N_A|$ stays the same, p_2^n decreases. □