Explaining Two Strange Learning Curves

Chen, Zhiyi; Loog, Marco; Krijthe, Jesse H.

DOI
10.1007/978-3-031-39144-6_2

Publication date
2023

Document Version
Final published version

Published in
Artificial Intelligence and Machine Learning - 34th Joint Benelux Conference, BNAIC/Benelearn 2022, Revised Selected Papers

Explaining Two Strange Learning Curves

Zhiyi Chen[1,3(✉)], Marco Loog[1,2], and Jesse H. Krijthe[1]

[1] Delft University of Technology, Delft, The Netherlands
zhiychen@student.ethz.ch
[2] Radboud University, Nijmegen, The Netherlands
[3] ETH Zürich, Zürich, Switzerland

Abstract. Learning curves illustrate how generalization performance of a learner evolves with more training data. While this is a useful tool to characterize learners, not all learning curve behavior is well understood. For instance, it is sometimes assumed that the more training data provided, the better the learner performs. However, counter-examples exist for both classical machine learning algorithms and deep neural networks, where errors do not monotonically decrease with training set size. Loog *et al.* [12] describe this monotonicity problem, and present several regression examples where simple empirical risk minimizers display unexpected learning curve behaviors. In this paper, we will study two of these proposed problems in detail and explain what caused the odd learning curves. For the first, we use a bias-variance decomposition to show that the monotonic increase in the learning curve is caused by an increase in the variance, which we explain by a mismatch between the model and the data generating process. For the second problem, we explain the recurring increases in the learning curve by showing only two solutions are attainable by the learner. The probability of obtaining a configuration of training objects that leads to the high risk solution typically decreases as the training set size increases. However, for particular training set sizes, additional configurations that produce the high risk solution become possible. We prove that these additional configurations increase the probability of the high risk solution and therefore explain the unusual learning curve. These examples contribute to a more complete understanding of learning curves and the possibilities and reasons behind their various behaviors.

1 Introduction

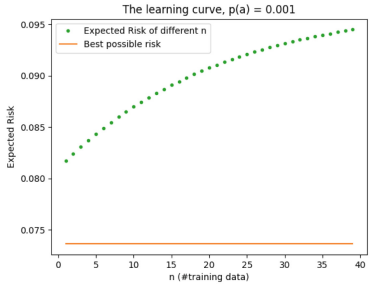
Learning curves are plots demonstrating how the number of training samples influences the generalization performance of learners. They are essential tools to understand and compare the learning behavior of different machine learning models. Among others, they have been used to predict the maximum achievable accuracy, estimate how much data is required for the desired accuracy, predict the generalization performance of learners [5,9] to save computational costs and avoid the usage of the excess training samples [6].

A large quantity of research investigated learning curves for different problems or tried to find a common model for learning curves of various problems

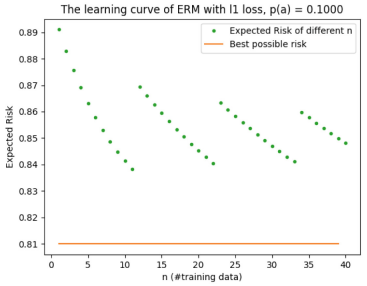
T. Calders et al. (Eds.): BNAIC/Benelearn 2022, CCIS 1805, pp. 16–30, 2023.
https://doi.org/10.1007/978-3-031-39144-6_2

[18]. Different models have been proposed to describe learning curves, such as exponential or logarithmic models [5,6,8]. While generating the learning curves for assorted problems, many unexpected behaviors of learning curves have been observed. Some 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 [6,7,15].

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 [2]. Two other striking examples are presented in Loog *et al.* [12]. In the first, the expected risk increases as more training data are provided (Fig. 1a). In the second problem, the learning curve shows a periodic pattern (Fig. 1b). The reasons for the non-monotonically decreasing learning curves in these examples are poorly understood. Our goal in this work is to explain why these behaviors occur and to contribute to a better understanding of learning curve behavior.



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



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

Fig. 1. The two studied learning curves with unexpected behaviors

The remainder of this paper is structured in the following way. Section 2 will present other works in the field on learning curves 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. Finally, Sect. 5 will discuss how our study answers the research question, discuss its limitations and conclude.

2 Related Work

There is a large number of studies regarding learning curves in general. Many researchers have tried to find suitable functional models for learning curves [5, 6,10,18]. Duin [3] investigated the learning curves of a variety of algorithms

to find a reasonably well-performing algorithm for small-sample-size problems. Likewise, much research has focussed on understanding the behavior of learning curves, leading to various assumptions about and insights into how learning curves can behave. Haussler *et al.* [7], as an example, developed a theory to find rigorous bounds for learning curves. Provost *et al.* [16] suggested that 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. Others have claimed that the accuracy should increase as more data is provided [6,7,15].

However, while investigating learning curves for various problems, many learning curves not conforming with these assumptions have been discovered. A well-known example is learning curves that exhibit a "double decent" or peaking pattern. This phenomenon was probably first recorded in [17] and relates to the currently equally popular double-descent complexity curves [1,4,13,15]. An fairly recent and complete overview of badly behaving learning curves can be found in [11].

In [12], the authors show that even with a simple distribution and a basic learner, 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 estimate learning curves using real-world datasets with unknown distributions, Loog *et al.* [12] proposed a simple distribution and used it to generate artificial datasets. Since the distribution is known and simple, the expected risk can be calculated exactly, 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.

3 Problem Setting and Methodology

We will first formally describe the two problems and introduce terminology in Sect. 3.1. Then, we will present the two disparate solution strategies we applied to explain the behavior of the learning curves in Sect. 3.2.

3.1 Problem Setting: The Distribution and the Learners

The two problems are originally proposed in [12]. In both problem settings, the following aspects are the same:

The ground truth distribution $\mathcal{D}$ is $(x, y) \in \mathbb{R} \times \mathbb{R} = \mathcal{Z}$, this distribution only has a non-zero probability at two points $a = (1, 1)$ and $b = (\frac{1}{10}, 1)$. Let $P((x, y) = a) = p_a$ and $P((x, y) = b) = 1 - p_a = p_b$.

The hypothesis class $\mathcal{H}$ is all linear functions without intercepts; i.e., $\mathcal{H} = \{h(x) = \beta x \,|\, \beta \in \mathbb{R}\}$.

Both are regression problems and use ERM as the learner. A learner $\mathcal{A}$ maps the set of all possible datasets to elements in the hypothesis class, i.e. $\mathcal{A} : \mathcal{Z} \cup \mathcal{Z}^2 \cup \mathcal{Z}^3 \cup ... \cup \mathcal{Z}^n \rightarrow \mathcal{H}$. Let $\mathcal{L} : \mathcal{H} \rightarrow \mathbb{R}$ denote the loss function,

$\mathcal{R} : \mathcal{H} \to \mathbb{R}$ 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 $\mathcal{R}(h) = \mathbb{E}_{(x,y)}\mathcal{L}(h(x), y)$ and the empirical risk $\hat{\mathcal{R}}(h)$, given a training dataset S^n, is $\hat{\mathcal{R}}(h) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(h(x), y)$. An empirical risk minimizer $\mathcal{A}_{erm}$ is a learner which outputs the hypothesis with minimum empirical risk, given a set of training data.

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

Problem I: $p_a = 0.001$. The loss function is L2 loss: $\mathcal{L}(h) = (h(x) - y)^2$. The empirical risk minimizer $\mathcal{A}_{erm}$ has a closed-form solution $(X^T X)^{-1} X^T Y$, where $X = \begin{bmatrix} x_1, x_2, \ldots, x_n \end{bmatrix}^T$ and $Y = \begin{bmatrix} y_1, y_2, \ldots, y_n \end{bmatrix}^T$.

Problem II: $p_a = 0.1$. The loss function is L1 loss: $\mathcal{L}(h) = |h(x) - y|$. The empirical risk minimizer $\mathcal{A}_{erm}$ does not have a closed-form solution in general. However, we will show (Sect. 4.2) that it has a closed-form solution when $X, Y \in \mathbb{R}$.

3.2 Disparate Methods for Analyzing the Problems

Due to the divergent nature of the two problems, we will approach them using distinct methods. For **Problem I**, we use a bias-variance decomposition to break down the expected risk into *bias* and *variance* terms, and analyze the resulting terms. This method has previously been used in [14] to explain the double descent phenomenon occurring in the learning curves of linear regression (ERM with L2 loss). After observing the curves of these two terms, we focus on the *variance* term and further inspect the cause of its increase.

When interpreting the learning curve of **Problem II**, we will first derive the closed-form solution of $\mathcal{A}_{erm}$, to show only two solutions are possible: the risk optimal solution and a sub-optimal solution. We show that the expected risk depends on the probability of $\mathcal{A}_{erm}$ producing the sub-optimal hypothesis. We then show that this probability can be decomposed into the sum of the probability of different configurations of training objects. The periodic behavior is then explained by the interplay between the decreasing probability of existing configurations and periodic increases in probability due to new configurations becoming possible.

4 Analysis

4.1 Problem I

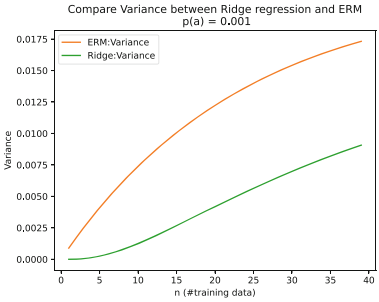
We apply a bias-variance decomposition to the expected risk for **Problem I** and identify the cause of the increase in the learning curve by observing how the resulting terms change with respect to the number of training samples. We show that ridge regression can mitigate the problem and then analyze why the problem occurs.

Bias-Variance Decomposition. In the setting of **Problem I**, given the L2 loss and linear hypotheses without intercept, the true risk for a fixed sample size n is $\mathbb{E}_{S^n}\mathcal{R}(\mathcal{A}_{erm}(S^n)) = \mathbb{E}_{S^n}\mathbb{E}_{(x,y)}(\hat{\beta}x - y)^2$, $\hat{\beta} = \mathcal{A}_{erm}(S^n)$. This expression can be decomposed in the following way:

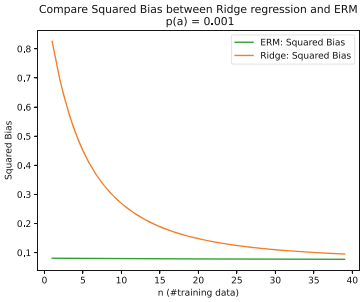
$$\begin{aligned}
\mathbb{E}_{S^n}\mathcal{R}(\mathcal{A}_{erm}(S^n)) &= \mathbb{E}_{S^n}\mathbb{E}_{(x,y)}(\hat{\beta}x - y)^2 \\
&= \mathbb{E}_{(x,y)}\mathbb{E}_{S^n}(\hat{\beta}x - \mathbb{E}_{S^n}\hat{\beta}x + \mathbb{E}_{S^n}\hat{\beta}x - y)^2 \\
&= \mathbb{E}_{(x,y)}\mathbb{E}_{S^n}\Big\{(\hat{\beta}x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

We will call the term $\mathbb{E}_x x^2 \cdot Var_{S^n}(\hat{\beta})$ *variance* and the term $\mathbb{E}_{(x,y)}(\mathbb{E}_{S^n}\hat{\beta}x - y)^2$ *squared bias*. Figure 2 shows how squared bias and variance are changing with respect to the training size n. This shows that the increase of the variance term surpasses the decrease of the squared bias term, leading to an increasing expected risk.

Fig. 2. Squared bias and Variance terms with respect to the growth of n in Problem I



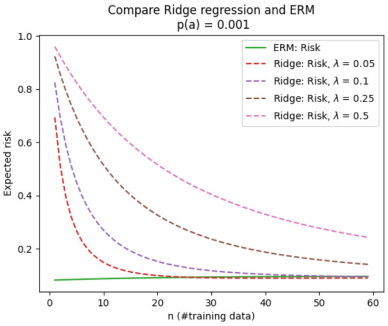
(a) Variance



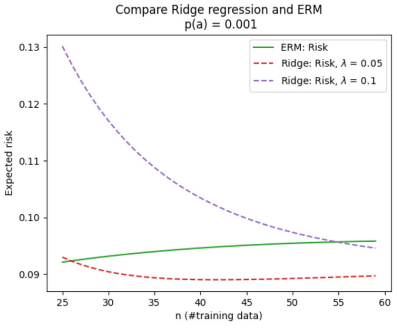
(b) Squared Bias

Fig. 3. Comparison of the variance and squared bias terms for ERM and ridge regression

Mitigating the Variance Increase. We consider ridge regression $\mathcal{A}_{ridge}$, to see whether regularization can mitigate the increase in variance and lead to a decreasing learning curve. $\mathcal{A}_{ridge}(S^n) = \arg\min_{\beta\in\mathbb{R}} \lambda||\beta||^2 + \sum_{i=1}^{n}(\beta x_i - y_i)^2$, where $\lambda \in [0, +\infty)$ is a hyper-parameter controlling the strength of the regularization effect. The larger the λ, the stronger the regularization effect is. The closed-form solution of $\mathcal{A}_{ridge}$ is $(X^T X + \lambda \mathbf{I})^{-1} X^T Y$. As shown in Fig. 3a, with $\lambda = 0.1$, the variance is lower compared to $\mathcal{A}_{erm}$ and grows at a slower rate, while the squared bias starts at a higher value and decreases faster, as shown in Fig. 3b. Looking at different values for lambda, $\lambda = \{0.05, 0.1, 0.25, 0.5\}$, we find that learning curves of $\mathcal{A}_{ridge}$ decrease monotonically for $n = 1, 2, ..., 40$, (Fig. 4a), except when lambda is small. When zoomed in (Fig. 4b), the figure also 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.



(a) Overview of performance with different lambda



(b) When zooming in, we can observe that lower expected risk is achieved when n is large enough with lambda 0.1 and 0.05.

Fig. 4. Compare the performance of different lambda values

Explaining the *Variance* Increase. The increase in the variance term runs contrary to the intuition that a larger number of training samples should lead to lower variance. One explanation for the effect is that the distribution does not fit the linear model $Y = \beta X + \epsilon$, where $\mathbb{E}\epsilon = 0$, and X and ϵ are independent. We will show that if the data does fit this model, we can expect a decreasing learning curve. In order to investigate the variance, we first calculate $\mathbb{E}_{S^n}\hat{\beta}$:

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

Since $\mathbb{E}_{S^n}\epsilon_n = 0$, we have $\mathbb{E}_{S^n}\hat{\beta} = \beta$. Next consider $Var_{S^n}(\hat{\beta})$:

$$\begin{aligned}Var_{S^n}(\hat{\beta}) &= \mathbb{E}_{S^n}(\hat{\beta} - \beta)^2 \\ &= \mathbb{E}_{X_n}\mathbb{E}_{\epsilon_n}((X_n^T X_n)^{-1} X_n^T (X_n\beta + \epsilon_n) - \beta)^2 \\ &= \mathbb{E}_{X_n}\mathbb{E}_{\epsilon_n}((X_n^T X_n)^{-1} X_n^T X_n\beta + (X_n^T X_n)^{-1} X_n^T \epsilon_n - \beta)^2 \\ &= \mathbb{E}_{X_n}\mathbb{E}_{\epsilon_n}((X_n^T X_n)^{-1} X_n^T \epsilon_n)^2 \\ &= \mathbb{E}_{X_n} Var_{\epsilon_n}((X_n^T X_n)^{-1} X_n^T \epsilon_n) \\ &= \mathbb{E}_{X_n}\left[(X_n^T X_n)^{-1} X_n^T Cov_{\epsilon_n}\epsilon_n X_n (X_n^T X_n)^{-1}\right]\end{aligned}$$

Since all training samples are i.i.d., we have $Cov_{\epsilon_n}\epsilon_n = Var(\epsilon)\cdot\mathbb{I}_n$. Hence

$$\begin{aligned}Var_{S^n}(\hat{\beta}) &= \mathbb{E}_{X_n}\left[Var(\epsilon)(X_n^T X_n)^{-1}(X_n^T X_n)(X_n^T X_n)^{-1}\right] \\ &= Var(\epsilon)\mathbb{E}_{X_n}(X_n^T X_n)^{-1} \\ &= Var(\epsilon)\mathbb{E}_{X_n}(\sum_{i=1}^{n} x_i^2)^{-1}\end{aligned}$$

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

Under the same assumptions, the squared bias term $\mathbb{E}_{(x,y)}(\mathbb{E}_{S^n}\hat{\beta}x - y)^2 = \mathbb{E}_\epsilon\mathbb{E}_x(\beta x - (\beta x + \epsilon))^2 = \mathbb{E}_\epsilon\epsilon^2$, which is a constant. Therefore, when the distribution fits the linear model, the learning curve will always decrease. Even though the distribution in **Problem I** can be modelled as $Y = \beta X + \epsilon$, where $\beta = \arg\min_{\beta\in\mathbb{R}}\mathbb{E}_(x,y)(\beta x - y)^2$, ϵ is dependent on X and $\mathbb{E}\epsilon \neq 0$. We can therefore conclude that one possible reason for the increasing learning curve is that these two essential conditions are missing.

4.2 Problem II

To explain the unexpected periodic behavior of the learning curve for **Problem II**, we first derive the closed-form solution for both $\mathcal{A}_{erm}$ and the optimal β.

Then, we show that the expected risk depends on the probability of $\mathcal{A}_{erm}$ outputting a specific hypothesis. We then investigate why the curve of this probability has periodic behavior.

Closed-form Solution of $\mathcal{A}_{erm}$ and the Optimal β. In the setting of **Problem II**, the loss is L1 loss. Therefore, $\mathcal{A}_{erm}$ can be 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 points $a = (x_a, y_a)$ in the training dataset, and n_b denote the number of points $b = (x_b, y_b)$. The empirical risk for all hypotheses β is therefore

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

Consider the sign of the gradient with respect to β, we discard the positive $\frac{1}{n}$,

(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 $n_a x_a - n_b x_b > 0$, 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 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 $\beta = \frac{y_b}{x_b}$. Finally, when $n_a x_a - n_b x_b = 0$, both $\frac{y_a}{x_a}$ and $\frac{y_b}{x_b}$ are minima, among other solutions. In that case we pick $\beta = \frac{y_a}{x_a}$ as the minimizer. Note that this choice will not affect the qualitative behavior: it will

merely shift the curve by one sample. 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 $\arg\min_{h \in \mathcal{H}} \mathcal{R}(h) = \mathbb{E}_{(x,y)} |\beta x - y| = p_a |\beta x_a - y_a| + p_b |\beta x_b - y_b|$ to give

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

Under the setting of **Problem II**, $p_a x_a - p_b x_b = \frac{1}{10} \cdot 1 - \frac{1}{10} \cdot \frac{9}{10} > 0$, $\beta = \frac{y_a}{x_a}$.

Analysis of the Expected Risk. We now analyze 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, $\hat{\beta}_1 = \frac{y_a}{x_a}$, $\hat{\beta}_2 = \frac{y_b}{x_b}$ the two possible solutions, and $P_1^n = P_{S^n}(\hat{\beta} = \hat{\beta}_1)$ and $P_2^n = P_{S^n}(\hat{\beta} = \hat{\beta}_2) = 1 - P_1^n$ the probability of attaining these solutions. The expected risk can be written as the risk of each solution multiplied by the probability of attaining that solution:

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

As $\hat{\beta}_1 = \arg\min_{\beta \in \mathbb{R}} \mathbb{E}_{(x,y)} |\beta x - y|$, the risk of $\hat{\beta}_1$ is smaller than the risk of $\hat{\beta}_2$. Hence, the smaller P_2^n is, the larger P_1^n and the smaller the expected risk for n. Therefore, to explain the periodic behavior (shown in Fig. 5), we must investigate how P_2^n changes with respect to the number of training samples.

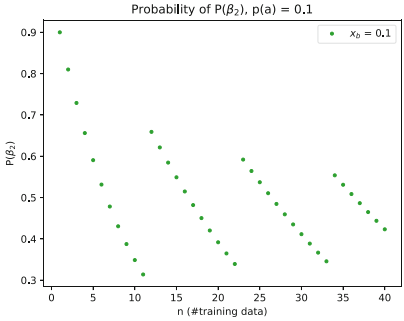
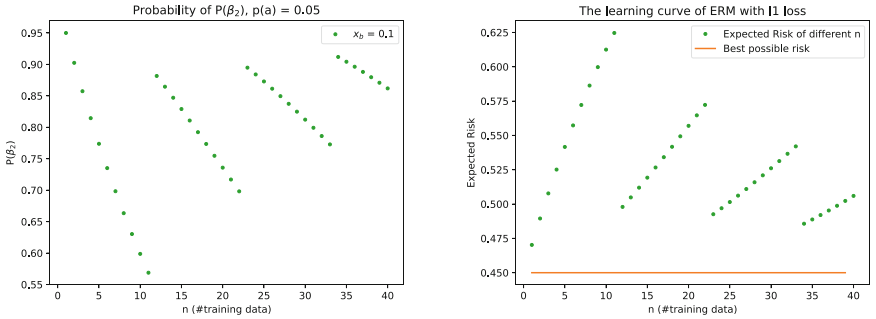


Fig. 5. The change of P_2^n with respect to n

Note that if we change the value of p_a such that $p_a x_a - p_b x_b < 0$, then $\beta = \frac{y_b}{x_b}$. 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 expected risk will be. An example is shown in Fig. 6 where we set $p_a = 0.05$ and all the other values remain the same.



(a) The change of P_2^n (b) The learning curve

Fig. 6. The curve of P_2^n and the learning curve when $p_a = 0.05$ leading to $\beta = \hat{\beta}_2$

Explaining the Periodic Pattern of P_2^n. The question of why the learning curve behaves as in Fig. 1b can be reduced to the question why the curve of P_2^n has the behavior shown in Fig. 5. To investigate P_2^n, we need to understand when $\mathcal{A}_{erm}$ will output $\hat{\beta}_2$. This happens when

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

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 to

$\lceil \frac{x_a}{x_b} + 1 \rceil k + 1$, where $k \in \mathbb{N}$. In this problem setting $\lceil \frac{x_a}{x_b} + 1 \rceil = 11$ and as shown in the Figs. 1b and 5, the curves have an increase when $n \in \{x \in \mathbb{N} | x = 11k+1, k \in \mathbb{N}\}$. Therefore, we claim that the increase of $|N_A|$ is the cause of the sudden increase. In order to prove this, we consider $\sum_{i \in N_A} P_{S^n}(n_a = i)$ before and after $|N_A|$ increases by 1.

Let $M = \lceil \frac{x_a}{x_b} + 1 \rceil$ be the length of the periodicity and consider the difference in probability when the sample is increased by one, such that $|N_A|$ increases from k to $k+1$: $P_2^{kM+1} - P_2^{kM}$ for any integer $k > 0$. Two sources contribute to this difference: the change in probability of the existing k configurations and the addition of the probability of the new configuration. The change in the former is given by

$$\sum_{j=0}^{k-1} \binom{kM+1}{j} (p_a)^j (1-p_a)^{kM+1-j} - \sum_{j=0}^{k-1} \binom{kM}{j} (p_a)^j (1-p_a)^{kM-j}.$$

Note that this is the difference between the conditional distribution functions (CDF) of two binomial distributions. These CDFs are equal to the regularized incomplete beta function (indicated by I_x). We therefore have:

$$I_{(1-p_a)}((M-1)k+2, k) - I_{(1-p_a)}((M-1)k+1, k) = \\ - \binom{Mk+1}{k} \frac{k}{Mk+1} p_a^k (1-p_a)^{(M-1)k+1}, \quad (1)$$

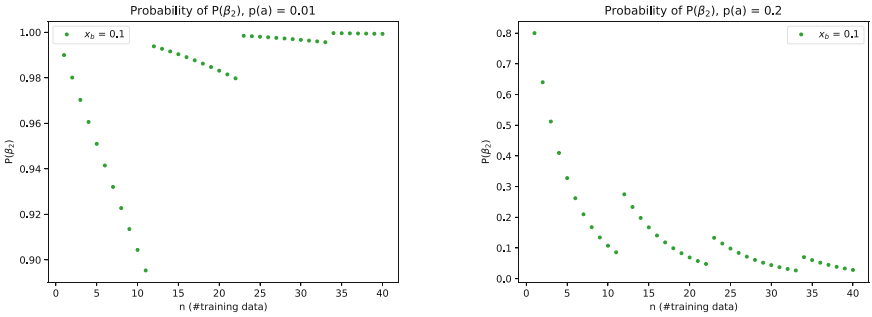
using the identity $I_x(a+1, b) = I_x(a, b) - \binom{a+b}{a} \frac{b}{a+b} x^a (1-x)^b$. Note this difference is negative for any probability p_a.

Next consider the increase in P_2^n caused by the additional configuration. This is given by

$$\binom{Mk+1}{k} (p_a)^k (1-p_a)^{(M-1)k+1}.$$

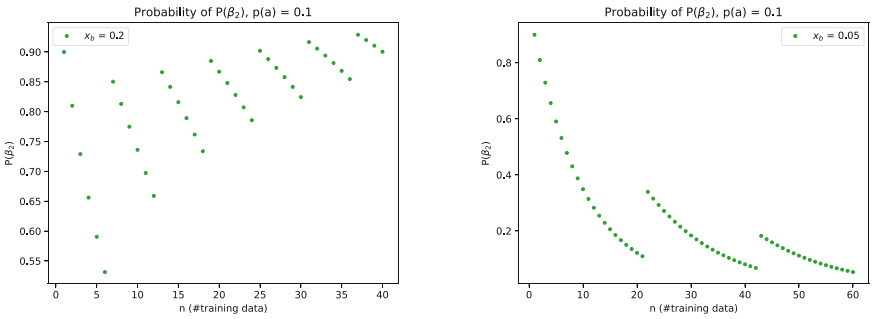
Comparing this to (1), we find that the decrease in probability is always $\frac{k}{Mk+1}$ times the increase caused by the additional configuration, hence the increase is always bigger than the decrease. So, whenever $|N_A|$ does not increase when n increases, P_2^n decreases. However, when $|N_A|$ increases as well (which happens every M training objects), P_2^n is guaranteed to increase. This also directly implies that the non-monotonic behavior for this learning curve keeps occurring for arbitrarily large n.

Moreover, we can conclude that the shape of the curve showing how P_2^n changes with respect to n will always demonstrate such periodic patterns regardless of the value of p_a. As shown in Fig. 7, 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 Fig. 8, the duration of one period is always equal to $\lceil \frac{x_a}{x_b} + 1 \rceil$.



(a) When $p_a = 0.01$ (b) When $p_a = 0.2$

Fig. 7. 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 $x_b = \frac{1}{20}, \lceil \frac{x_a}{x_b} + 1 \rceil = 21$

Fig. 8. The behavior of P_2^n with different values of $\lceil \frac{x_a}{x_b} + 1 \rceil$

5 Discussion and Conclusion

Our goal was to explain why the learning curves generated under two problem settings proposed by Loog et al. [12] exhibit non-monotonic behavior. For **Problem I**, we adopted the bias-variance decomposition to split the expected risk into bias and variance terms. 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 demonstrated that ridge regression can suppress the rapid increase in variance. It is unexpected for variances to increase with more training samples. We suggested that this increase is caused by the fact that the distribution of this problem does not fit certain modeling assumptions (linearity, independence, and zero expected error). We supported this insight by showing that if these assumptions do hold, the variance decreases with more training samples. While this shows one way of adapting the problem to guarantee monotonicity, alternative assumptions about the problem are possible (for instance, letting go of linearity) that we have not explored here. In addition, these sufficient assumptions also do not offer a direct intuitive account of why the variance increases.

For **Problem II**, we showed that the change in the probability of A_{erm} outputting one of two possible hypotheses leads to the periodic pattern of the learning curve. In this way, the problem is reduced to explaining why the curve of this probability has periodic behavior. For this probability, we prove that the factors that cause a decrease in the probability will periodically be negated by an additional configuration increasing the probability.

While our goal was to shed some light on learning curves for two specific problems, both using artificial distributions, an important question is how these insights generalize to similar problems, and how they can inform our understanding of strange learning curves that are observed empirically. Nevertheless, the study of specific cases, as presented in this work, helps to better understand learning behavior, also in general. In fact, we would be glad to see the study of artificial data sets to regain prominence within machine learning research in the large.

A Appendix

We use two known but non-trivial identities in our proof. For completeness we prove these identities here.

Claim. Given a binomial distribution $f(k,n,p) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, the CDF $F(k,n,p) = P(X \leq k) = I_{1-p}(n-k, 1+k)$

Proof. Denote $1-p$ as q,

$$
\begin{aligned}
I_q(n-k,1+k) &= \frac{\mathbf{\Gamma}(n+1)}{\mathbf{\Gamma}(n-k)\mathbf{\Gamma}(k+1)} \int_0^q t^{n-k-1}(1-t)^k dt \\
&= \frac{n!}{(n-k-1)!k!} \left[\frac{1}{n-k} t^{n-k}(1-t)^k \right] \Bigg|_0^q \\
&+ \frac{n!}{(n-k-1)!k!} \underbrace{\left[\frac{k}{n-k} \int_0^q t^{n-k}(1-t)^{k-1} dt \right]}_{\text{(Integration by parts)}} \\
&= \binom{n}{k} q^{n-k} p^k + \underbrace{\frac{n!}{(n-k)!(k-1)!} \left[\int_0^q t^{n-k}(1-t)^{k-1} dt \right]}_{\text{Integration by parts}} \\
&= \binom{n}{k} q^{n-k} p^k + \binom{n}{k-1} q^{n-(k-1)} p^{k-1} \\
&+ \underbrace{\frac{n!}{(n-k+1)!(k-2)!} \left[\int_0^q t^{n-k+1}(1-t)^{k-2} dt \right]}_{\text{Repeated application of integration by parts}} \\
&= \sum_{i=0}^{k} \binom{n}{i} q^{n-i} p^i \\
&= P(X \leq k)
\end{aligned}
$$

Claim. $I_x(a+1,b) = I_x(a,b) - \binom{a+b}{a}\frac{b}{a+b}x^a(1-x)^b$

Proof.

$$
\begin{aligned}
I_x(a+1,b) &= \sum_{i=0}^{b-1}\binom{a+b}{i}x^{(a+b)-i}(1-x)^i \\
&= \sum_{i=0}^{b-1}\left[\binom{a+b-1}{i}+\binom{a+b-1}{i-1}\right]x^{(a+b)-i}(1-x)^i \\
&= x\sum_{i=0}^{b-1}\binom{a+b-1}{i}x^{(a+b-1)-i}(1-x)^i \\
&\quad + \sum_{i=0}^{b-1}\binom{a+b-1}{i-1}x^{(a+b-1)-(i-1)}(1-x)^i \\
&= xI_x(a,b) + \sum_{i=0}^{b}\binom{a+b-1}{i-1}x^{(a+b-1)-(i-1)}(1-x)^i \\
&\quad - \binom{a+b-1}{b-1}x^a(1-x)^b \\
&= xI_x(a,b) + (1-x)I_x(a,b) - \frac{(a+b-1)!}{(b-1)!a!}x^a(1-x)^b \\
&= I_x(a,b) - \frac{(a+b)!}{a!b!}\frac{b}{a+b}x^a(1-x)^b \\
&= I_x(a,b) - \binom{a+b}{a}\frac{b}{a+b}x^a(1-x)^b
\end{aligned}
$$

References

1. Belkin, M., Hsu, D., Ma, S., Mandal, S.: Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proc. Natl. Acad. Sci. **116**(32), 15849–15854 (2019). https://doi.org/10.1073/pnas.1903070116
2. d'Ascoli, S., Sagun, L., Biroli, G.: Triple descent and the two kinds of overfitting: where and why do they appear? J. Stat. Mech. Theory Exp. **2021**(12), 124002 (2021). https://doi.org/10.1088/1742-5468/ac3909
3. Duin, R.: Small sample size generalization. In: 9th Scandinavian Conference on Image Analysis, pp. 957–964 (1995)
4. Duin, R.: Classifiers in almost empty spaces. In: Proceedings 15th International Conference on Pattern Recognition, ICPR-2000, vol. 2, pp. 1–7 (2000). https://doi.org/10.1109/ICPR.2000.906006
5. Frey, L.J., Fisher, D.H.: Modeling decision tree performance with the power law. In: Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. R2 (1999). https://proceedings.mlr.press/r2/frey99a.html

6. Gu, B., Hu, F., Liu, H.: Modelling classification performance for large data sets. In: Advances in Web-Age Information Management, pp. 317–328 (2001)
7. Haussler, D., Kearns, M., Seung, H.S., Tishby, N.: Rigorous learning curve bounds from statistical mechanics. Mach. Learn. **25**, 195–236 (1996). https://doi.org/10.1007/BF00114010
8. John, G.H., Langley, P.: Static versus dynamic sampling for data mining. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD 1996, pp. 367–370. AAAI Press (1996)
9. Kolachina, P., Cancedda, N., Dymetman, M., Venkatapathy, S.: Prediction of learning curves in machine translation. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 22–30 (2012). https://aclanthology.org/P12-1003
10. Last, M.: Predicting and optimizing classifier utility with the power law. In: Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), pp. 219–224 (2007)
11. Loog, M., Viering, T.: A survey of learning curves with bad behavior: or how more data need not lead to better performance (2022). https://arxiv.org/abs/2211.14061
12. Loog, M., Viering, T., Mey, A.: Minimizers of the empirical risk and risk monotonicity. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 7478–7487 (2019)
13. Loog, M., Viering, T., Mey, A., Krijthe, J.H., Tax, D.M.: A brief prehistory of double descent. Proc. Natl. Acad. Sci. **117**(20), 10625–10626 (2020)
14. Nakkiran, P.: More data can hurt for linear regression: sample-wise double descent. arXiv (2019). https://arxiv.org/abs/1912.07242
15. Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., Sutskever, I.: Deep double descent: where bigger models and more data hurt. J. Stat. Mech. Theory Exp. **2021**(12), 124003 (2021). https://doi.org/10.1088/1742-5468/ac3a74
16. Provost, F., Jensen, D., Oates, T.: Efficient progressive sampling. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 1999, pp. 23–32 (1999). https://doi.org/10.1145/312129.312188
17. Vallet, F., Cailton, J.G., Refregier, P.: Linear and nonlinear extension of the pseudo-inverse solution for learning boolean functions. Europhys. Lett. (EPL) **9**(4), 315–320 (1989). https://doi.org/10.1209/0295-5075/9/4/003
18. Viering, T.J., Loog, M.: The shape of learning curves: a review. CoRR abs/2103.10948 (2021). https://arxiv.org/abs/2103.10948