# The Dipping Phenomenon

Marco Loog and Robert P.W. Duin

Pattern Recognition Laboratory
Delft University of Technology
Delft, The Netherlands
prlab.tudelft.nl

**Abstract.** One typically expects classifiers to demonstrate improved performance with increasing training set sizes or at least to obtain their best performance in case one has an infinite number of training samples at ones's disposal. We demonstrate, however, that there are classification problems on which particular classifiers attain their optimum performance at a training set size which is finite. Whether or not this phenomenon, which we term dipping, can be observed depends on the choice of classifier in relation to the underlying class distributions. We give some simple examples, for a few classifiers, that illustrate how the dipping phenomenon can occur. Additionally, we speculate about what generally is needed for dipping to emerge. What is clear is that this kind of learning curve behavior does not emerge due to mere chance and that the pattern recognition practitioner ought to take note of it.

## 1 On Learning Curves and Peaking

The analysis of learning curves, which describe how a classifier's error rate behaves under different training set sizes, is an integral part of almost any proper investigation into novel classification techniques or unexplored classification problems [7]. Though sometimes interest goes only to its asymptotics [9], the learning curve is especially informative in the comparison of two or more classifiers when considering the whole range of training set sizes. It indicates at what samples sizes the one classifier may be preferable over the other for a particular type of problem. Also, by means of extrapolation, the curve may give us some clue on how many additional samples may be needed in a real-world problem to reach a particular error rate. Such analyses are readily impossible on the basis of a point estimate as, for example, obtained by means of leave one out cross-validation on the whole data set at hand.

The learning curve one typically expects to observe falls off monotonically with increasing training set size (see Figure 1). The rate of decrease depends on the particular problem considered and the complexity of the classifier employed. Such behavior can indeed be demonstrated in certain settings in which the classifier selected typically fits the underlying data assumptions well, see for instance [1,10]. In a similar spirit, various bounds on learning curves also show monotonic decrease for the expected true error rate with increasing training set sizes [5,16].

G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 310–317, 2012.

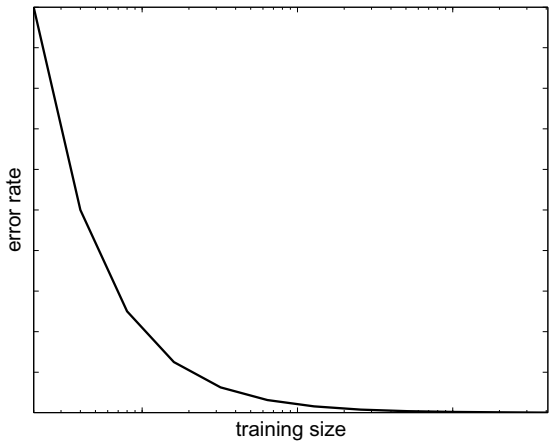


**Fig. 1.** An idealized learning curve in which the error rate drops monotonically with an increasing training set size

That such monotonic behavior can, however, not always be guaranteed has already been known at least since the mid nineties. Both Opper and Kinzel [12] and Duin [4] describe what is nowadays referred to in pattern recognition as the peaking phenomenon for learning curves: the error rate attains a local maximum that does not coincide with the smallest training sample size considered. This phenomenon has been described and investigated, for instance, for the Fisher discriminant classifier [4,13,14], for particular perceptron rules [12,11], and for lasso regression [8]. The naming of this phenomenon alludes to the peaking phenomenon for increasing feature sizes (as opposed to increasing training set sizes, which this paper is concerned with) as originally identified by Hughes [6] in the 1960s. Hughes' phenomenon for such feature curves shows that, for a fixed training sample size, the error initially drops but beyond a certain dimensionality typically starts to rise again.

On the basis of what we know about peaking, we may adjust our expectation about learning curves and speculate classifiers to at least obtain their best performance when an infinite number of training samples is used. But also this turns out to be false hope as this work demonstrates. It appears there are classification problems on which particular classifiers attain their optimal performance at a training set size which is finite. In contrast with peaking, we term this phenomenon dipping as it concerns a minimum in the learning curve, in fact, a non-asymptotic, global minimum.

The next three section of the paper, Sections 2, 3, and 4, give some simple examples, for three artificial classification problems in combination with specific classifiers, which demonstrate how the dipping phenomenon emerges. Though artificial, the examples clearly illustrate that this kind of learning curve behavior does not merely emerge due to chance, e.g. due some unfortunate draw of training data, but that it is an issue structurally present in particular problem-classifier combinations. The final section, Section 5, speculates on what

generally is needed for dipping. It also offers some further discussions and concludes this contribution.

## 2 Basic Dipping for Linear Classifiers

Consider a two-class classification problem consisting of one Gaussian distribution and one mixture of two Gaussian distributions (Figure 2). The Gaussians of the second class appear on either side of the Gaussian of the first class. A perfectly symmetric situation is considered here: there is symmetry in the overall distribution and the class priors are equal. It should be stressed, however, that this perfect symmetry is definitely not needed to observe a dipping behavior, just like there is no need to stick to Gaussian distributions. This configuration, however, enables us to easily explain why dipping occurs.

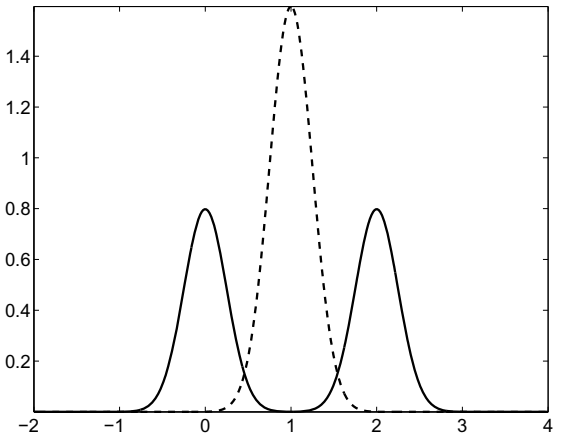


**Fig. 2.** Distribution of two-class data used to illustrate basic dipping

Let us consider what happens when we would make an expected learning curve for the nearest mean classifier (NMC, [3]). In the case of large total training set sizes, both means will be virtually on top of each other and the expected classification error will reach a worst case performance of 0.5. If, however, we go to smaller and smaller sample sizes, these means will in expectation be further and further apart due to their difference in variance. In the extreme case in which we have one observation from both classes, the one mean will be around the mode of class one and the other will be near one of the two modes of class two. Though one will still have means that lead to an error rate of about 0.5, chances are very slim. There will, however, be many configurations that both classify the first class and one lobe of the second class more or less correctly, which gives an expected error of around 0.25 as only the second lobe of the second class gets misclassified.

In conclusion, the smaller the sample size is the higher the probability is that the NMC delivers a performance considerably better than chance. Figure 3 gives

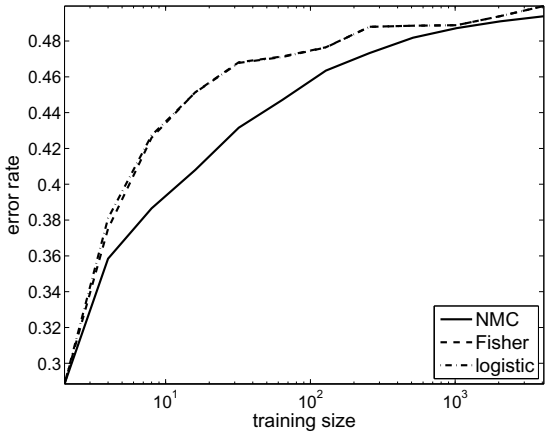


**Fig. 3.** Dipping learning curves for three linear classifiers, viz., NMC, Fisher discriminant, and logistic regression, based on the one-dimensional distribution presented in Figure 2

the expected learning curve (an average over 1000 repetitions) for training set sizes ranging from $2^1$ to $2^{12}$ (compare to Figure 1). The same figure displays learning curves for the Fisher discriminant and logistic regression as well. Both linear classifiers also suffer from dipping and an explanation for this goes along the same lines as for the NMC.

## 3 Delayed Dipping

The following example demonstrates that the occurrence of the dip can be at any point along the learning curve. Let us again consider the NMC but now the classification problem changes to the one illustrated in Figure 4. The first class is a Gaussian distribution and the second class is a noisy ring positioned around the first class with a variable radius. Again the priors are taken equal.

When the radius of the ring is small, we are basically back in a situation similar to the one in Section 2 and one would observe dipping as in Figure 3. The more training samples one would have, the closer the two means would get. Though this is bad in case the ring is near the center class, when the ring grows larger and larger, while the noise level stays the same, more observations in fact lead to improved performance up to a certain level. Having one observation per class would mean that the larger part of the ring is going to be misclassified to the center class. Increasing the total training set size, however, moves the mean of the second class closer and closer to the mean of the first class. As long as the second class mean does not move into the region where the first class becomes dense, moving closer to the center will lead to a better classification of class two and therefore a better overall performance. When the ring grows infinitely large, the two means can be virtually the same (relative to the size of the ring) while the first class is classified nearly perfect and as good as half of the ring is correctly classified. This happens when the training set grows infinite as well.

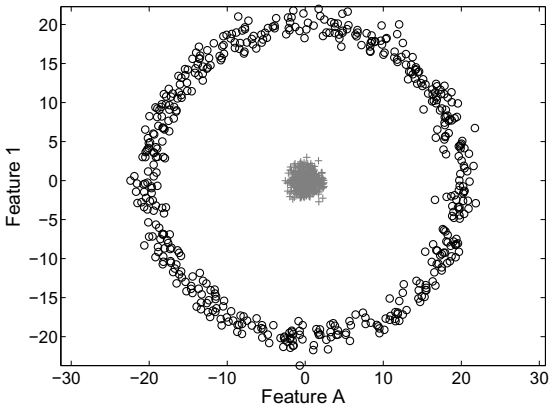


**Fig. 4.** Single instantiation of a distribution of two-class data used to illustrate early and late dipping for the NMC. The outer ring can vary in diameter based on which the time of dipping can be controlled.

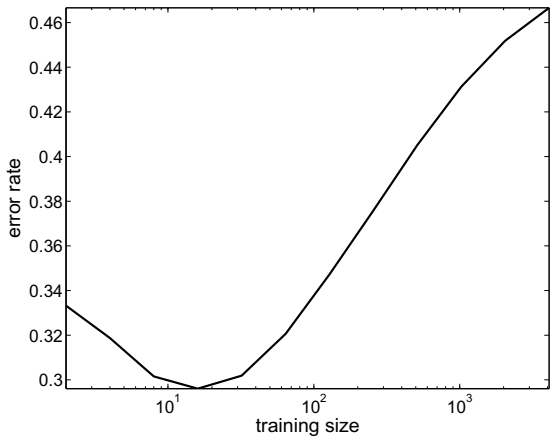


**Fig. 5.** Learning curve for the NMC based on the two-class distribution in Figure 4. It shows the dipping phenomenon to occur away from the smallest sample sizes.

In conclusion, by means of the variable ring diameter, one can tune the occurrence of the dip for the NMC to an arbitrary position along the learning curve. Figure 5 gives a learning curve that dips at a training sample of 16, which is obtained for a radius of 20 with a Gaussian standard deviation and a ring noise level standard deviation of 1.

## 4 Dipping of QDA

Our final example shows that dipping is not limited to linear discriminants but may also be encountered when employing more flexible classifiers. Here we

consider classical quadratic discriminant analysis (QDA, [10]). Figure 6 shows the class configurations used—a variation to the one from Section 2. Figure 7 displays the learning curve obtained by QDA[1]. A reason rather similar to the one given in Section 2 can be given for the observed dipping, though it is slightly more involved because of the more complex classifier considered. Here we merely note that in case of large sample sizes the decision boundary is close to the middle and the error rate gets close to the worst case solution, which is slightly less than 0.5. For smaller sample sizes the decision boundary shifts away from the middle, which on average leads to an improvement in classification error as can be observed in the learning curve from Figure 7.

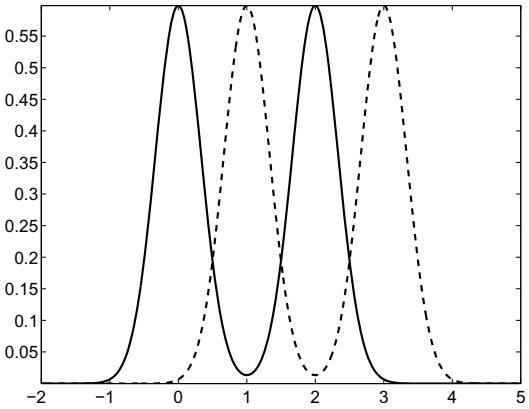


**Fig. 6.** Distribution of two class data used to let QDA dip

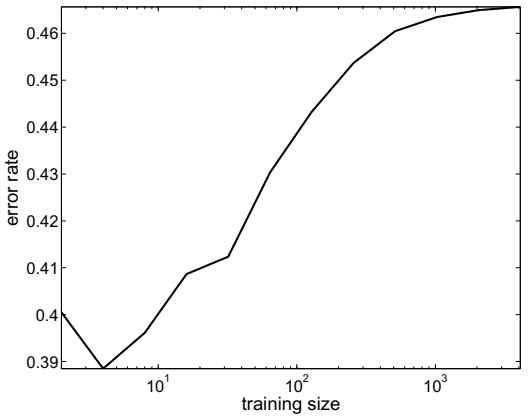


**Fig. 7.** Learning curve for QDA related to the two-class distribution in Figure 6, illustrating that dipping is not limited to the simplest of classifiers.

[1] As the per class sample sizes sometimes equals one, the covariance matrices in QDA were moderately regularized in this experiment.

## 5 Discussion and Conclusion

For four different classifiers we have demonstrated that the dipping phenomenon can be observed. We explained why it emerges in a basic setting using linear classifiers, sketched how the dipping point can attain an arbitrary location along the learning curve, and illustrated the possibility that also discriminants more complicated than linear can show dipping behavior. What seems to be a essential requirement is that the model underlying the classifier does not suit the classification problem considered very well. Curves similar to those on Figure 2 can be generated for linear discriminant analysis (LDA), the perceptron, or the linear support vector machine. All in all, it raises the question at what complexity classifiers will not suffer from dipping any longer. More specifically: can we find problems for which $k$ nearest neighbors or the Parzen classifier show this type of behavior? Or are nonparametric techniques immune to dipping? Certainly for the Parzen classifier, when one would keep the kernel's bandwidth fixed, we would not be surprised if particular data configurations will even make this classifier dip. To date, however, we have been unsuccessful in finding an illustration of such behavior.

It may even be that still less is needed for dipping to potentially happen. Even if the type of decision boundaries that can be modeled by a particular classifier is in principle rich enough to include the Bayes decision boundary for the problem at hand, the learning routine or estimation procedure might be unable to find the correct fit. An example is the Fisher discriminant, which is not always able to separate linearly separable classes. The underlying problem is that we want to minimize the expected classification error but in reality we always have to settle for a surrogate loss that is all but a bad approximation to the 0-1 loss. Maybe due to this discrepancy, "anything" can happen: for any classifier one might be able to find a, potentially rather pathological, data set for which the classifier dips. That this state of affairs may not be completely accurate is, however, demonstrated by the existence of so-called universally consistent classifiers (see, for instance, [15]). Though such results on universality should, in turn, also be interpreted with care [2].

A completely different question this work also raises is whether one should treat the training set size just like any other free parameter a classifier has. Should one, for example, also cross-validate over the number of training samples to be used for training? Another issue of interest is whether the phenomenon can be observed in any real-world problem and how it affects such setting.

Irrespective of the previous questions, we think dipping is a phenomenon that one should keep in mind when studying learning curves. When observed, it may not be ascribed blindly to chance or a bad training sample. It might just be inherent in the combination of problem at hand and classifier employed.

## References


1. Amari, S., Fujita, N., Shinomoto, S.: Four types of learning curves. Neural Computation 4(4), 605–618 (1992)
2. Ben-David, S., Srebro, N., Urner, R.: Universal learning vs. no free lunch results. In: Philosophy and Machine Learning Workshop NIPS 2011 (December 2011), http://www.dsi.unive.it/PhiMaLe2011/
3. Duda, R., Hart, P.: Pattern classification and scene analysis. John Wiley & Sons (1973)
4. Duin, R.: Small sample size generalization. In: Proceedings of the Scandinavian Conference on Image Analysis, vol. 2, pp. 957–964 (1995)
5. Haussler, D., Kearns, M., Seung, H., Tishby, N.: Rigorous learning curve bounds from statistical mechanics. Machine Learning 25(2), 195–236 (1996)
6. Hughes, G.: On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory 14(1), 55–63 (1968)
7. Jain, A., Duin, R., Mao, J.: Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(1), 4–37 (2000)
8. Krämer, N.: On the peaking phenomenon of the lasso in model selection. Arxiv preprint arXiv:0904.4416 (2009)
9. Langley, P.: Machine learning as an experimental science. Machine Learning 3(1), 5–8 (1988)
10. McLachlan, G.: Discriminant Analysis and Statistical Pattern Recognition. John Wiley & Sons (1992)
11. Opper, M.: Learning to generalize. In: Frontiers of Life, vol. 3(part 2), pp. 763–775. Academic Press (2001)
12. Opper, M., Kinzel, W.: Statistical mechanics of generalization. In: Models of Neural Networks III, ch. 5. Springer (1995)
13. Raudys, S., Duin, R.: Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. Pattern Recognition Letters 19(5), 385–392 (1998)
14. Skurichina, M., Duin, R.: Stabilizing classifiers for very small sample sizes. In: Proceedings of the 13th International Conference on Pattern Recognition, vol. 2, pp. 891–896. IEEE (1996)
15. Steinwart, I.: Consistency of support vector machines and other regularized kernel classifiers. IEEE Transactions on Information Theory 51(1), 128–142 (2005)
16. Vapnik, V.: Estimation of dependences based on empirical data. Springer (1982)