Also for k-means
more data does not imply better performance

Loog, Marco; Krijthe, Jesse H.; Bicego, Manuele

DOI
10.1007/s10994-023-06361-6

Publication date
2023

Document Version
Final published version

Published in
Machine Learning

Machine Learning (2023) 112:3033–3050
https://doi.org/10.1007/s10994-023-06361-6

Also for *k*-means: more data does not imply better performance

Marco Loog[1] · **Jesse H. Krijthe**[1,2] · **Manuele Bicego**[3]

Received: 7 December 2022 / Revised: 28 April 2023 / Accepted: 28 June 2023 /
Published online: 25 July 2023

Abstract
Arguably, a desirable feature of a learner is that its performance gets better with an increasing amount of training data, at least in expectation. This issue has received renewed attention in recent years and some curious and surprising findings have been reported on. In essence, these results show that more data does actually not necessarily lead to improved performance—worse even, performance can deteriorate. Clustering, however, has not been subjected to such kind of study up to now. This paper shows that *k*-means clustering, a ubiquitous technique in machine learning and data mining, suffers from the same lack of so-called monotonicity and can display deterioration in expected performance with increasing training set sizes. Our main, theoretical contributions prove that 1-means clustering is monotonic, while 2-means is not even weakly monotonic, i.e., the occurrence of nonmonotonic behavior persists indefinitely, beyond any training sample size. For larger *k*, the question remains open.

Keywords Learning curve · *k*-Means clustering · *k*-Means algorithm · Monotonicity · Smartness · Performance improvement

Editors: Fabio Vitale, Tania Cerquitelli, Marcello Restelli, and Charalampos Tsourakakis.

✉ Marco Loog
marco.loog@ru.nl

Jesse H. Krijthe
j.h.krijthe@tudelft.nl

Manuele Bicego
manuele.bicego@univr.it

[1] Radboud University, Nijmegen, The Netherlands

[2] Delft University of Technology, Delft, The Netherlands

[3] University of Verona, Verona, Italy

1 Introduction

k-Means continues to remain among the most-used techniques both in data mining and in machine learning. Next to being employed for analysis, *k*-means is also actively researched and further developed as such. Among recent work, for instance, we find that Stemmer (2021) studied *k*-means in the differentially private setting, Klochkov et al. (2021) developed a robustified version and report non-asymptotic bounds, Liu (2021) provided improved learning bounds, Ghadiri et al. (2021) came up with a fair *k*-means, while Cohen-Addad et al. (2021) developed and analyzed an online version. The, largely theoretical, study that we provide in this work concerns the learning curve behavior of *k*-means.

Learning curves (or, alternatively, sample complexity curves (Zubek and Plewczynski 2016)) consider, on a specific learning problem, the average performance of a learner against the sample size of the training set (Perlich, 2010; Viering & Loog, 2022). At least since the work by Vallet et al. (1989), it is known that such curves can exhibit counterintuitive behavior and show deteriorating performance with increasing training set sizes, i.e., these learners can display so-called nonmonotonic behavior (Loog et al., 2019).

The investigation of such a, arguably, problematic property (Loog & Viering, 2022) has found renewed attention in the wake of the important work by Belkin et al. (2019). They raise various issues concerning classical learning theory in the light of modern-day overparameterized learners. The phenomenon for which this work seems to be cited primarily, however, was essentially described by Vallet et al. (1989) already (see also Loog et al., 2020). Following Belkin et al. (2019), it is nowadays referred to as double descent, also in the learning curve setting: the curve starts off as expected and improves with increasing numbers of training samples, then its performance starts to deteriorate over a consecutive number of training sizes, following which it again gets to improved performance with more training data.

There is a gamut of surprising and curious learning curve behaviors next to double descent. One of the more peculiar ones is the zigzag curve that least absolute deviation regression can display and which has been explained recently by Chen et al. (2023). Loog and Viering (2022) provide a complete overview of the present state of affairs. Our current work is, more specifically, in line with the findings from Loog et al. (2019). There it is shown that learners that rely on empirical risk minimization (ERM), being at the basis of many learners these days, can act nonmonotonically no matter the training sample size. Stated differently, with some sense of drama, Loog et al. (2019) show that even if, during the training phase, we optimize for the loss that we are also using during the test phase,[1] test-time performance can still deteriorate in expectation. The original work showed the emergence of such behavior for classification, regression, and density estimation.

1.1 Contributions and outline

In this work, we show, in a precise sense, that *k*-means clustering is not devoid of such quirks either. Our contribution is limited in the sense that we merely manage to prove

[1] This may seem like an odd statement, but it is often not realized that, for instance, most classifiers actually optimize an essentially different objective function than what is going to be used during their evaluation. Not surprisingly, this can have unexpected consequences and hamper analysis (Loog & Duin, 2012; Ben-David et al., 2021; Loog et al., 2016).

nonmonotonicity for $k = 2$. Nevertheless, this is significant, firstly, for the fact that we can at all prove that something like this happens for k-means and, secondly, because it demonstrates that nonmonotonicity extends to settings beyond classification, regression, and density estimation.

The next section goes through some preliminaries and provides further related work. It primarily puts k-means in the context of empirical risk minimization and introduces the precise notion of monotonicity. Section 3 formulates our two main results: for $k = 1$ the learning curve of k-means always behaves properly, i.e., it improves with more data, but for $k = 2$ its behavior can be problematic. More precisely, we show that no matter the size of the training set, there are clustering problems for which the 2-means performance still becomes worse with even more training data. While the proof for the former result is provided in the same section, all of Sect. 4 revolves around the proof of the second result. Section 5 discusses and concludes our work.

2 Preliminaries and additional related work

We formulate k-means clustering within the framework of empirical risk minimization (ERM) and touch upon some further relevant literature in this context. In addition, we make precise the notion of monotonicity that we are going to employ. Regarding the latter, we largely follow the notation and definitions as proposed in (Viering et al., 2019; Loog et al., 2019).

2.1 Empirical risk minimization and k-means

Let $S_N = \left(x_1, \ldots, x_N\right) \in \overset{N}{\mathcal{X}}$ be a training set of size N. This is, in our k-means setting, an i.i.d. sample from a distribution D over the standard d-dimensional feature space $\mathcal{X} = \overset{d}{\mathbb{R}}$. In addition, we have as hypothesis class the set of sets of k means

$$\mathcal{M}_k = \left\{m = \{m_1, \ldots, m_k\} \middle| m_i \in \mathbb{R}^d, \forall i \in \{1, \ldots, k\}\right\}. \tag{1}$$

Note that, as every $m \in \underset{k}{\mathcal{M}}$ is an actual set, their cardinality $|m|$ is smaller than k in case $m_i = m_j$ for one or more pairs $i \neq j$. That is, redundant means are discarded.

The particular loss function that we consider in our case could be termed the cluster-wise or group-wise squared loss. We, however, go with within-group squared (WGS) loss, inspired by the term within-group sum of squares that Hartigan (1978) considers in the context of k-means:

$$\begin{aligned} \ell_{\text{WGS}} : \mathcal{X} \times \mathcal{M}_k &\to \mathbb{R} \\ (x, m) &\mapsto \min_{i \in \{1, \ldots, |m|\}} \|x - m_i\|^2, \end{aligned} \tag{2}$$

The ultimate objective is to minimize the expected loss, i.e., the risk:

$$R_D(m) := \mathop{\mathbb{E}}_{x \sim D} \ell_{\text{WGS}}(x, m). \tag{3}$$

As we do not know the actual underlying distribution D, the principle of ERM (Vapnik, 1982) suggests the learner to rely on its empirical distribution, defined by the training sample of size N, and to consider the loss on this distribution:

$$R_{S_N}(m) := \frac{1}{N}\sum_{j=1}^{N} \ell_{\mathrm{WGS}}(x_j, m) = \frac{1}{N}\sum_{j=1}^{N} \min_{i\in\{1,\ldots,|m|\}} \|x_j - m_i\|^2. \tag{4}$$

This empirical risk is then equivalent to Hartigan (1978)'s within-group sum of squares.

We can now define the following:

Definition 1 (k-means clustering) k-means clustering is the learner A_k that maps from the set of all samples $\mathcal{S} := \bigcup_{i=1}^{\infty} \overset{i}{\mathcal{X}}$ to the hypothesis class $\mathcal{M}_k$, i.e., $A_k : \mathcal{S} \to \mathcal{M}_k$, according to the optimality assignment

$$\begin{aligned} \mathcal{A}(S_N) &:= \operatorname*{argmin}_{m\in\mathcal{M}_k} R_{S_N}(m), \\ A_k(S_N) &:= \operatorname*{argmin}_{m\in\mathcal{A}(S_N)} |m|. \end{aligned} \tag{5}$$

The first minimizers in our definition picks out all mean sets that minimize the empirical risk. $\mathcal{A}(S_N)$ can indeed be a set of solutions, in particular when $N < k$. The second minimizer then makes sure that we obtain a solution set of minimal cardinality.

The ERM view on k-means that we consider is equivalent to the formulation that Pollard (1981) provides. Different formulations are possible, for instance, where ones does not consider the squared distance to the closest means, but where one looks for a partitioning of the space in optimal regions. The latter can be found, for example, in the works by Dalenius (1950), MacQueen (1967), Ben-David et al. (2006). The former "center-based" formulation can also be found in Rakhlin (2005), Rakhlin and Caponnetto (2006) and Ben-David (2004) (under the name vector quantization problem). Buhmann (1998) provides a slightly different ERM setting and normalizes the influence of every cluster with its size. Bock (2007) relates different center-based and partitioning based approaches.

2.2 Idealized and practical *k*-means

Note that the specific k-means that we consider is, in some sense, idealized, because we actually assume that it minimizes the empirical risk globally, which is known to be an NP-hard problem (Dasgupta, 2008; Aloise et al., 2009).

For this reason, in practice, one often needs to resort to suboptimal approaches when carrying out the optimization in Eq. (5). This is where the well-known alternating optimization method of assigned points to means, then updating the means, and repeating this process comes in (Steinhaus, 1956; Jancey 1966; Bock 2007). Often, k-means is actually associated with exactly this algorithm, which we expressly do not do in this paper.

The successful k-means++ algorithm (Arthur & Vassilvitskii, 2007) has been shown to provide a reasonable solution to the k-means problem, despite it being NP-hard. More specifically, it is guaranteed to achieve a WGS risk in polynomial time that is, in expectation, no worse than $\log k + 2$ times the optimal. It should be noted that k-means++ considers the WGS risk of the training set, while we are primarily interested in the expected loss of k-means on the full problem distribution.

2.3 Monotonic or smart?

We note that more than 20 years before Loog et al. (2019), Devroye et al. (1996) already suggested the notion of smart rules, which refers to classifiers that show non-increasing expected error rates with increasing training set sizes. The terms smart and smartness, may be better choices than the terms monotonic and monotonicity. The latter term, as it is used originally, does of course not distinguish between increasing or decreasing curves. What we are after, in particular, is a nonincreasing curve in terms of the expected risk. As we evaluate in terms of the same loss as the one that is being optimized during training, this risk is not the error rate in our setting, but equals the expected within-group squared loss from Eq. (3).

2.4 Monotonicity

The behavior we are interested in is that we get better, or at least not worse, test performance when having more data to train on. In particular, we want k-means to not perform worse with increasing N in terms of the expected within-group squared loss as given by Eq. (3). As adding a single very bad sample can always ruin performance, it is reasonable to merely ask for such performance non-deterioration in expectation, i.e., over all possible samples S_N and S_{N+1} of size N and $N+1$, respectively. A basic initial definition is therefore the following.

Definition 2 (local monotonicity) A_k is $(D, \ell_{\text{WGS}}, N)$-monotonic with respect to a distribution D and an integer $N \in \mathbb{N} := \{1, 2, \dots\}$ if

$$\Delta_N^{N+1} := \mathop{\mathbb{E}}_{S_N \sim D^N}[R_D(A_k(S_N))] - \mathop{\mathbb{E}}_{S_{N+1} \sim D^{N+1}}[R_D(A_k(S_{N+1}))] \geq 0. \tag{6}$$

The two entities we would like to get rid of in Definition 2 are N and D. The former, because we would like our learner to act monotonically irrespective of the sample size. The latter, because we typically do not know the underlying distribution. What we do know, however, is in which domain we are operating, which is $\mathbb{R}^d$ for our k-means. Therefore, employing the difference Δ_N^{N+1} as defined in Eq. (6), the following is appropriate.

Definition 3 (local $\mathbb{R}^d$-monotonicity) A_k is (locally) $(\mathbb{R}^d, \ell_{\text{WGS}}, N)$-monotonic with respect to an integer $N \in \mathbb{N}$ if, for all distributions D on $\mathbb{R}^d$ for which Δ_N^{N+1} exists, $\Delta_N^{N+1} \geq 0$.

Note that the above definition is a refinement of the original from (Viering et al., 2019; Loog et al., 2019). In particular, we added the statement on the existence of the difference in expected risks to make sure that the learner is $\mathbb{R}^d$-monotonic even though on some distributions the necessary integrals may not exist or the difference in itself is problematic. Such issues arise, for instance, when both expectations in the difference evaluate to infinity and Δ_N^{N+1} would become $\infty - \infty$.

The double-descent phenomenon that we covered earlier shows that learning curves can have some difficulties being monotonous from the start. The second best thing to hope for is that a learner becomes monotonic after some sample size, which leads to a weak form of monotonicity.

Definition 4 (weak $\mathbb{R}^d$-monotonicity) A_k is weakly $(\mathbb{R}^d, \ell_{\text{WGS}}, n)$-monotonic if there is an integer $n \in \mathbb{N}$ such that for all $N \geq n$, the learner is locally $(\mathbb{R}^d, \ell_{\text{WGS}}, N)$-monotonic.

If the n from Definition 4 can be set to 1, the learner is called globally $\mathbb{R}^d$-monotonic.

Definition 5 (global $\mathbb{R}^d$-monotonicity) A_k is globally $(\mathbb{R}^d, \ell_{\text{WGS}})$-monotonic if for every integer $N \in \mathbb{N}$, the learner is locally $(\mathbb{R}^d, \ell_{\text{WGS}}, N)$-monotonic.

2.5 Learning curves and bounds

Meek et al. (2002) were possibly the first to consider learning curves for clustering techniques in an application setting. In particular, they studied mixtures of Gaussians, but the notion can be transferred readily to the setting of k-means as seen in what follows. Interestingly, apart from Meek et al. (2002), there seems little additional learning curve work.

There is more work available on theoretical bounds for k-means. These report non-asymptotic bounds on the excess risk—or excess distortion as it is also called in this setting. Some of the more recent works are (Levrard, 2015; Maurer, 2016; Chichignoud & Loustau, 2014; Levrard, 2013; Biau et al., 2008). Earlier mentioned (Klochkov et al., 2021) is one of the recent additions, which, alternatively, studies a robust version of k-means. Their bounds show the typical $1/N$ or $\sqrt{1/N}$ power-law behavior in terms of the training sample size N. Clearly, as these are bounds, they do not necessarily say a lot about the actual (local) behavior of the learning curve.

3 Main theoretical results

Having laid the groundwork in the previous section, we can now come to our main results.

Theorem 1 *A_1 is globally $(\overset{d}{\mathbb{R}}, \ell_{\text{WGS}})$-monotonic.*

The proof follows in Sect. 3.1. It is fairly uninvolved—certainly compared to the proof of the next theorem. Nevertheless, the result is there both for completeness and to contrast it with what happens when $k = 2$, in which case the behavior changes notably.

Theorem 2 *For A_2 in combination with any integer $N \geq 14$, there exists a distribution D for which $\Delta_N^{N+1} < 0$ and, therefore, A_2 is not weakly $(\mathbb{R}^d, \ell_{\text{WGS}}, n)$-monotonic for any $n \in \mathbb{N}$.*

The proof of this result is relatively involved and could be considered a bit cumbersome. We therefore dedicate a separate section to it, which is Sect. 4.

3.1 Proof of Theorem 1

Proof It is easy to check that the minimizing hypothesis of the empirical risk $A_1(S_N)$ is attained for the mean: $m_N := \frac{1}{N}\sum_{i=1}^{N} x_i$, as we are dealing with only one cluster.

Let $\mathbb{E}$ now denote the expectation both over the training sample $S_N \sim D^N$ and over the test sample $x \sim D$, where this latter expectation comes from the risk R_D expressed by Eq. 3.

With μ the true mean of the distribution D, we can write the following for the expected within-cluster loss for sample size N:

$$\mathbb{E}\left[\|x - m_N\|^2\right] = \mathbb{E}\left[\|x - \mu + \mu - m_N\|^2\right] = \mathbb{E}\left[\|x - \mu\|^2\right] + \mathbb{E}\left[\|m_N - \mu\|^2\right]. \tag{7}$$

Only the last term matters as it is the only part depending on the training set size. We have

$$\begin{aligned}
&\mathbb{E}\left[\|m_N - \mu\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{n} x_i - \mu\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{n} x_i\right\|^2\right] - \|\mu\|^2 \\
&= \mathbb{E}\left[\frac{1}{N^2}\sum_{i=1}^{n}\|x_i\|^2 + \frac{1}{N^2}\sum_{j\neq k} x_j^T x_k\right] - \|\mu\|^2 \\
&= \mathbb{E}\left[\frac{N}{N^2}\|x\|^2 + \frac{N^2-N}{N^2}\|\mu\|^2\right] - \|\mu\|^2 = \mathbb{E}\left[\frac{1}{N}\|x\|^2\right] - \frac{1}{N}\|\mu\|^2 \\
&= \frac{1}{N}\mathbb{E}\left[\|x - \mu\|^2\right].
\end{aligned} \tag{8}$$

Now, for Δ_N^{N+1} (as defined in Eq. (6)) to exist, $\mathbb{E}\left[\|x - \mu\|^2\right]$ needs to be finite. In that case, as $\mathbb{E}\left[\|x - \mu\|^2\right]$ is nonnegative, we have that

$$\Delta_N^{N+1} = \frac{1}{N}\mathbb{E}\left[\|x - \mu\|^2\right] - \frac{1}{N+1}\mathbb{E}\left[\|x - \mu\|^2\right] \geq 0 \tag{9}$$

for any N and so A_1 is globally monotonic. □

4 Proof of Theorem 2

At a high level, the proof is fairly straightforward. We explicitly construct a class of distributions that has two free parameters and we show that by having those parameters depend on N in the right way, we can always make the learning curve go up when going from a sample of size N to one of size $N + 1$ on that same distribution, i.e., $\Delta_N^{N+1} < 0$. As we can construct such a distribution for any $N \geq 14$ (so the distribution is allowed to be different for every $(N, N + 1)$-pair), A_2 is not locally $\mathbb{R}^d$-monotonic for any $N \geq 14$. As such, A_2 is also not weakly monotonic for any $n \in \mathbb{N}$, as there is always an $N \geq n$ and a corresponding distribution for which $\Delta_N^{N+1} < 0$.

The distributions that we construct for this consist simply of three point masses in one-dimensional feature space $\mathbb{R}$. Because we can always embed this one-dimensional problem in $\mathbb{R}^d$ for any $d \in \mathbb{N}$, we can limit ourselves to this specific problem in our proof.

The complication in proving Theorem 2 stems mainly from the fact that we cannot explicitly evaluate the difference Δ_N^{N+1} between two consecutive points on the learning curve in full. We therefore demonstrate that the curve goes up by showing that the difference is strictly negative by bounding it away from zero.

We start the preparations for the proof in the next subsection where we introduce our parameterized three-point problem. In Sect. 4.2, we then formulate and prove six lemmas, which give us different handles on parts of the behavior of the true expected risks. Sect. 4.3 brings it all together and finalizes the proof of Theorem 2.

4.1 A parameterized three-point problem and its risk

Definition 6 (three-point problem) This clustering problem, which depends on two parameters c and p, both in the open interval (0, 1), considers three locations in one-dimensional Euclidean space with associated probability mass function P. The specific points are A at -1, B at the origin, and C at c, while the associated probability masses are $P(A) = p$ and $P(B) = P(C) = \frac{1}{2}(1-p)$.

In addition, let us now introduce the following notation and definitions. Firstly, let the number of training samples from A, B, and C, equal i, j, and k, respectively. Secondly, let $\ell_X(i,j,k)$ equal the true loss incurred at point $X \in \{A,B,C\}$. It is important to note that these three losses are, of course, dependent on the precise counts for i, j, and k, as those determine the hypothesis for that training set. We denote its associated hypothesis from $\mathcal{M}_k$ by $m(i,j,k)$. A further definition we use is

$$R(i,j,k) := p\ell_A(i,j,k) + \tfrac{1}{2}(1-p)\big(\ell_B(i,j,k) + \ell_C(i,j,k)\big), \tag{10}$$

which denotes the true risk given the counts (i,j,k) for the three points A, B, and C.

Finally, for a training set of size N, the expected risk for the three-point clustering problem, which we simply denote by $E(N)$ from now on, is equal to

$$E(N) := \sum_{i=0}^{N} \sum_{j=0}^{N-i} \frac{N!}{i!j!(N-i-j)!} p^i \left(\frac{1-p}{2}\right)^{N-i} R(i,j,N-i-j), \tag{11}$$

where the count for point C equals $k = N-i-j$.

4.2 Six preparatory lemmas

The six lemmas presented in this section provide four different types of results. To start with, Lemma 1 describes a specific situation in which a minimizer for our three-point problem can be identified easily and uniquely. Lemmas 2 and 3 provide simplifications of some of the expressions involving binomials an risks that we will encounter. Lemmas 4 and 5 provide bounds for some of the expressions that appear in the proof of Theorem 2 when considering the difference in expected risk Δ_N^{N+1} at training sample size N and $N+1$ under the three-point distribution parameterized by the same c and p. Our ultimate lemma shows a specific one-dimensional function to be negative beyond a certain point. This is merely a technical result, that will ultimately be used to lower-bound the increase in the risk.

Lemma 1 *With $i \geq 1$ and $j + k \geq 1$,*

$$m(i,j,k) = \left\{-1, c\frac{k}{j+k}\right\} \tag{12}$$

is the unique minimizer of the empirical risk for the three-point problem if and only if

$$c^2 k(i+j) < i(j+k). \tag{13}$$

Proof First observe that using two rather than one mean, will always improve the empirical risk in the setting considered. Secondly, assigning points observed at the same location to two different means can never reduce the minimum risk. Thirdly, associating one mean with the middle location B and the other with the observations on both sides at A and C cannot be optimal. All in all, the optimal solution should either associate the one mean with A and the other with B and C or the one with A and B and the other with C. The mean associated with one point should, of course, be exactly that point. The other mean, in order to minimize the squared loss, is the weighted average of the two observed locations. As a result, either the hypothesis as specified in Eq. (12) is the optimizer or $\left\{-\frac{i}{i+j}, c\right\}$ is.

Now, with $N = i + j + k$, the respective empirical risks for these two hypotheses are

$$\frac{j}{N}\left(c\frac{k}{j+k}\right)^2 + \frac{k}{N}\left(c - c\frac{k}{j+k}\right)^2 = \frac{c^2jk}{N(j+k)} \tag{14}$$

and

$$\frac{i}{N}\left(-1 + \frac{i}{i+j}\right)^2 + \frac{j}{N}\left(\frac{i}{i+j}\right)^2 = \frac{ij}{N(i+j)}. \tag{15}$$

One therefore uniquely chooses the first hypothesis in case the value in Eq. (14) is (strictly) smaller than that in Eq. (15), as these are the only two hypotheses that we need to consider. It is easy to see that this holds if and only if $c^2k(i+j) < i(j+k)$. □

Lemma 2 *For the three-point problem from Definition 6 and any $N \geq 1$,*

$$\sum_{j=0}^{N} \frac{N!}{j!(N-j)!}\left(\frac{1-p}{2}\right)^N R(0, j, N-j) = \left(\frac{c(c+2p)}{2^N} + p\right)(1-p)^N, \tag{16}$$

which is the contribution to the expected risk of all training sets that do not contain samples at A.

Proof When $j = 0$ or $j = N$, the hypotheses minimizing the empirical risk are in fact single means at c and 0, respectively. In these cases, we have

$$R(0, N, 0) = p + \tfrac{1}{2}(1-p)c^2 \tag{17}$$

$$R(0, 0, N) = p(1+c)^2 + \tfrac{1}{2}(1-p)c^2. \tag{18}$$

For $1 \leq j \leq N-1$, both B and C are in the training set and the minimizer equals $m(0, j, k) = \{0, c\}$. The associated true risk is therefore $R(0, j, k) = p$. Working out the summation now leads to

$$\sum_{j=0}^{N} \frac{N!}{j!(N-j)!}\left(\frac{1-p}{2}\right)^N R(0,j,N-j) =$$
$$\left(\frac{1-p}{2}\right)^N \left(p + \tfrac{1}{2}(1-p)c^2 + p(1+c)^2 + \tfrac{1}{2}(1-p)c^2 + \sum_{j=1}^{N-1} \frac{N!}{j!(N-j)!}p\right) = \tag{19}$$
$$\left(\frac{1-p}{2}\right)^N \left(p + p(1+c)^2 + (1-p)c^2 + (2^N-2)p\right) =$$
$$\left(\frac{c(c+2p)}{2^N} + p\right)(1-p)^N. \qquad \square$$

Lemma 3 *Consider the three-point problem from Definition 6 and assume* $c^2 < \frac{4(N-1)}{N^2}$. *If* $N \geq 2$ *and* $1 \leq i \leq N-1$, *then*

$$\sum_{j=0}^{N-i} \frac{N!}{i!j!(N-i-j)!}p^i\left(\frac{1-p}{2}\right)^{N-i} R(i,j,N-i-j) =$$
$$\binom{N}{i}\frac{c^2(N-i+1)(1-p)^{N-i+1}p^i}{4(N-i)}. \tag{20}$$

Proof According to Lemma 1, with $k = N-i-j$, the hypothesis from Eq. (12) is the unique optimum, if $c^2(N-i-j)(i+j) \leq i(N-i)$. The left-hand side is quadratic in i and j and takes its maximum for any $i+j = \frac{N}{2}$. The right-hand side is quadratic and takes its minimum in $i = 1$ and $i = N-1$. Therefore, under the assumptions that $c^2 < \frac{4(N-1)}{N^2}$, we indeed have that

$$c^2(N-i-j)(i+j) \leq c^2\frac{N}{2}\frac{N}{2} < (N-1) \leq i(N-i). \tag{21}$$

Now, the optimal hypothesis $\left\{-1, c\frac{k}{j+k}\right\}$ has corresponding true risk

$$R(i,j,k) = \tfrac{1}{2}(1-p)\left(\left(c\frac{k}{j+k}\right)^2 + \left(c - c\frac{k}{j+k}\right)^2\right) = \tfrac{1}{2}c^2(1-p)\frac{j^2+k^2}{(j+k)^2} \tag{22}$$

and we find the desired identity by working the sum on the left-hand side of Eq. (20).

$$\sum_{j=0}^{N-i} \frac{N!}{i!j!(N-i-j)!}p^i\left(\frac{1-p}{2}\right)^{N-i} R(i,j,N-i-j) =$$
$$\sum_{j=0}^{N-i} \frac{N!}{i!j!(N-i-j)!}p^i\left(\frac{1-p}{2}\right)^{N-i} \tfrac{1}{2}c^2(1-p)\frac{j^2+(N-i-j)^2}{(N-i)^2} =$$
$$\frac{N!}{i!(N-i)!}\frac{c^2p^i(1-p)^{N-i+1}}{2^{N-i+1}(N-i)^2}\times \tag{23}$$
$$\sum_{j=0}^{N-i} \frac{(N-i)!}{j!(N-i-j)!}\left(2j^2 - 2(N-i)j + (N-i)^2\right) = (\star)$$

The next equality is established with the use of the identities $\sum_{k=0}^{n}\binom{n}{k}=2^n$, $\sum_{k=0}^{n}\binom{n}{k}k=n2^{n-1}$, and $\sum_{k=0}^{n}\binom{n}{k}k^2=(n^2+n)2^{n-2}$. All of these can be readily derived from the binomial theorem: the first sum is standard, the other two identities are reported in Gould (2010) as Eq.s (1.70) and (1.67), respectively.

$$\begin{aligned}(\star)&=\binom{N}{i}\frac{c^2p^i(1-p)^{N-i+1}}{2^{N-i+1}(N-i)^2}\times\\ &\left(((N-i)^2+N-i)2^{N-i-1}-(N-i)^2 2^{N-i}+(N-i)^2 2^{N-i}\right)\\ &=\binom{N}{i}\frac{c^2(N-i+1)p^i(1-p)^{N-i+1}}{4(N-i)}.\end{aligned}\tag{24}$$

□

Lemma 4 *Given $N\geq 2$, if $p<\frac{1}{N+7}$, then*

$$\begin{aligned}&\binom{N}{i}\frac{c^2(N-i+1)(1-p)^{N-i+1}p^i}{4(N-i)}+\\ &-\binom{N+1}{i}\frac{c^2(N-i+2)(1-p)^{N-i+2}p^i}{4(N-i+1)}\leq\\ &\left(1-\frac{N+7}{N+6}(1-p)\right)\binom{N}{i}\frac{c^2(N-i+1)(1-p)^{N-i+1}p^i}{4(N-i)}<0\end{aligned}\tag{25}$$

for all $1\leq i\leq N-1$.

Proof To start with, observe that the term $1-\frac{N+7}{N+6}(1-p)$ equals 0 when p takes on its supremum $\frac{1}{N+7}$. As soon as p becomes smaller than $\frac{1}{N+7}$, the term becomes strictly negative and we have the second, strict inequality.

To show the first inequality, note initially that we can rewrite it into the equivalent requirement that

$$\begin{aligned}\frac{\binom{N+1}{i}\frac{c^2(N-i+2)(1-p)^{N-i+2}p^i}{4(N-i+1)}}{\binom{N}{i}\frac{c^2(N-i+1)(1-p)^{N-i+1}p^i}{4(N-i)}}&=\frac{(N-i+2)(N+1)(N-i)(1-p)}{(N-i+1)^3}\geq\\ &\frac{N+7}{N+6}(1-p).\end{aligned}\tag{26}$$

Demonstrating this for $i=1$ comes down to showing that

$$\frac{(N+1)^2(N-1)}{N^3}\geq\frac{N+7}{N+6}.\tag{27}$$

To see that this holds, multiply left and right by $N^3(N+6)$ and reorganize terms to come to $5N^2-7N-6\geq 0$. Equality is attained for $N=-\frac{3}{5}$ and $N=2$, and so the inequality indeed holds for all $N\geq 2$, as we are dealing with a convex quadratic equation.

To show now that the inequality holds for any $1 \leq i \leq N-1$, consider the derivative to i:

$$\frac{d}{di}\frac{(N-i+2)(N+1)(N-i)}{(N-i+1)^3} = \frac{(N+1)(i^2-2(N+1)i+N^2+2N-2)}{(N-i+1)^4}.$$

This only becomes zero, when $i^2-2(N+1)i+N^2+2N-2=0$, which only happens when $i = N+1\pm\sqrt{3}$. Both these solutions are larger than $N-1$. Therefore, on the whole interval $[1, N-1]$, the derivative is positive, $\frac{(N-i+2)(N+1)(N-i)}{(N-i+1)^3}$ is strictly increasing over that same domain, and we have that

$$\frac{(N-i+2)(N+1)(N-i)}{(N-i+1)^3} \geq \frac{(N+1)^2(N-1)}{N^3} \geq \frac{N+7}{N+6} \tag{28}$$

for all $i \in [1, N-1]$ in general and for $1 \leq i \leq N-1$ in particular. □

Lemma 5 *For the three-point problem from Definition 6, any $N \geq 2$, and any $p \leq \frac{1}{2}$:*

$$\begin{aligned} &p^N R(N,0,0)+\left(\frac{c(c+2p)}{2^N}+p\right)(1-p)^N \\ -&p^{N+1}R(N+1,0,0)-\left(\frac{c(c+2p)}{2^{N+1}}+p\right)(1-p)^{N+1} \\ &\leq\left(\frac{c(c+2p)}{2^N}+2p^2\right)(1-p)^N. \end{aligned} \tag{29}$$

Proof The term $R(N, 0, 0)$ gives the risk in case the training set only contains samples from location A, which means we have one mean at -1 and therefore

$$R(N,0,0) = \tfrac{1}{2}(1-p)\big(1+(1+c)^2\big). \tag{30}$$

Using this term together with Eq. (16) from Lemma 2, we can rewrite the part of Eq. (29) on the left-hand side of the inequality, as

$$\begin{aligned} &p^N\tfrac{1}{2}(1-p)\big(1+(1+c)^2\big)+\left(\frac{c(c+2p)}{2^N}+p\right)(1-p)^N \\ -&p^{N+1}\tfrac{1}{2}(1-p)\big(1+(1+c)^2\big)-\left(\frac{c(c+2p)}{2^{N+1}}+p\right)(1-p)^{N+1} = \\ &\frac{c(p+1)(c+2p)(1-p)^N}{2^{N+1}}+p^2(1-p)^N+\frac{1}{2}(c^2-2c+2)(1-p)^2p^N. \end{aligned} \tag{31}$$

For the first term in the sum of the right-hand side, as $\frac{p+1}{2} \leq 1$, we have the following upper bound.

$$\frac{c(p+1)(c+2p)(1-p)^N}{2^{N+1}} \leq \frac{c(c+2p)(1-p)^N}{2^N}. \tag{32}$$

For the third term, we have

$$\frac{1}{2}(c^2-2c+2) \leq \frac{1}{2}2 = 1 \tag{33}$$

because $c \in (0,1)$ and the quadratic function takes on its supremum at $c = 0$. In addition, as $\left(\frac{p}{1-p}\right)^N \leq (2p)^N$ for all $N \geq 1$ if $p < \frac{1}{2}$, we have

$$(1-p)^2 p^N = p^2 \frac{p^{N-2}}{(1-p)^{N-2}}(1-p)^N \leq p^2(2p)^{N-2}(1-p)^N \leq p^2(1-p)^N. \tag{34}$$

□

Lemma 6 *For $N \geq 14$,*

$$(32N^2 + 8N)2^{-N} + 1 - \frac{4N^2 - N - 7}{2(N+6)(N-1)} < 0. \tag{35}$$

Proof For $N > 1$, the denominator $2(N+6)(N-1)$ is positive. Multiplying all terms with this term and simplifying leads to

$$\begin{aligned} &2(N+6)(N-1)((32N^2 + 8N)2^{-N} + 1) - (4N^2 - N - 7) \\ =&16N(N-1)(N+6)(4N+1)2^{-N} - 2N^2 + 11N - 5. \end{aligned} \tag{36}$$

The logarithm of the expression $16N(N-1)(N+6)(4N+1)2^{-N}$ is concave for $N > 1$ and so we can upper-bound that log-expression by a linear function. In particular, based on the first Taylor polynomial (or tangent) at $N = 14$, we may write:

$$\begin{aligned} &\log(16N(N-1)(N+6)(4N+1)2^{-N}) \\ \leq &\log\left(\frac{25935}{128}\right) + \left(\frac{27857}{103740} - \log(2)\right)(N-14). \end{aligned} \tag{37}$$

As the coefficient $\frac{45299}{157092} - \log(2)$ for the linear term in the latter part of the inequality is negative, we can fill in $N = 14$ to obtain a value that upper-bounds the expression in Eq. (37) for all $N \geq 13$. The exponent of this value, which equals $\frac{25935}{128}$, we can then use to upper-bound Eq. (36) for $N \geq 14$:

$$\begin{aligned} &2(N+6)(N-1)((32N^2 + 8N)2^{-N} + 1) - (4N^2 - N - 7) \\ &= 16N(N-1)(N+6)(4N+1)2^{-N} - 2N^2 + 11N - 5 \\ &\leq \frac{25935}{128} - 2N^2 + 11N - 5. \end{aligned} \tag{38}$$

This last quadratic upper bound is concave and takes on its maximum value at $N = \frac{11}{4}$, which is smaller than 14. Therefore, the value of the upper bound in Eq. (38) at 14, provides an upper bound for $\frac{25935}{128} - 2N^2 + 11N - 5$ for all $N \geq 14$. Its value is $\frac{25935}{128} - 2 \cdot 4^2 + 11 \cdot 4 - 5 = -\frac{5169}{128} < 0$ and, as $2(N+6)(N-1)$ is positive in that case, we also have that the left-hand side of Eq. (35) is strictly smaller than 0 for all $N \geq 14$. □

4.3 Finalizing the proof

Proof of Theorem 2 To start with, that A_2 is not weakly monotonic for any $n \in \mathbb{N}$ follows readily from the first part of the theorem. If for every integer $N \geq 14$, there exists a distribution D such that $\Delta_N^{N+1} < 0$ then there is always a distribution and an $N \geq n$ for which

$\Delta_N^{N+1} < 0$, which means A_2 cannot be weakly $(\overset{d}{\mathbb{R}}, \ell_{\text{WGS}}, n)$-monotonic. The remainder of the proof therefore focuses on demonstrating the first claim from the theorem.

Without loss of generality, we can limit our attention to one-dimensional $\mathbb{R}$, as we can always embed a problem from that space in $\mathbb{R}^d$. As such, nonmonotonicity of A_2 in $\mathbb{R}$ carries over to $\mathbb{R}^d$. Let us therefore merely consider the one-dimensional problem from Definition 6. Moreover, take $N \geq 14$ and take $c = \frac{2}{N}$, in which case c^2 is strictly smaller than $\frac{4(N-1)}{N^2}$. Additionally, take $p = \frac{1}{4N^2}$, which is strictly smaller than $\frac{1}{N+7}$ for $N \geq 14$. These choices make sure that all six lemmas hold and, in addition, determines a specific three-point distribution D for every N. We now show that the learning curve increases on this D when going from N to $N+1$ training samples.

Following Lemma 2 and Lemma 3, we can write

$$\begin{aligned} E(N) = p^N R(N,0,0) + \left(\frac{c(c+2p)}{2^N} + p\right)(1-p)^N \\ + \sum_{i=1}^{N-1} \binom{N}{i} \frac{c^2(N-i+1)(1-p)^{N-i+1}p^i}{4(N-i)}. \end{aligned} \tag{39}$$

Subsequently, let us consider the difference $\Delta_N^{N+1} = E(N) - E(N+1)$, where both expectations are taken with respect to the same underlying distribution D. This difference needs to be smaller than 0 to show that A_2 is not monotonic on the problem defined by D when going from N to $N+1$.

Using Lemmas 4 and 5, we find that

$$\begin{aligned} \Delta_N^{N+1} \leq & \left(\frac{c(c+2p)}{2^N} + 2p^2\right)(1-p)^N \\ & + \sum_{i=1}^{N-1} \left(1 - \frac{N+7}{N+6}(1-p)\right)\binom{N}{i} \frac{c^2(N-i+1)(1-p)^{N-i+1}p^i}{4(N-i)} \\ \leq & \left(\frac{c(c+2p)}{2^N} + 2p^2\right)(1-p)^N + \left(1 - \frac{N+7}{N+6}(1-p)\right)\frac{c^2N^2(1-p)^N p}{4(N-1)}, \end{aligned} \tag{40}$$

where the last inequality holds because all terms in the summation are smaller than 0, so removing those for $i \in \{2, \ldots, N\}$ only increases the value.

In our next step, we fill in our choices for c and p in the above inequality and simplifying the expression. This gives us the following bound on the change in expected risk on D:

$$\begin{aligned} \Delta_N^{N+1} \leq & \left(\frac{\frac{2}{N}\left(\frac{2}{N} + 2\frac{1}{4N^2}\right)}{2^N} + 2\left(\frac{1}{4N^2}\right)^2\right)\left(1 - \frac{1}{4N^2}\right)^N \\ & + \left(1 - \frac{N+7}{N+6}\left(1 - \frac{1}{4N^2}\right)\right)\frac{\left(\frac{2}{N}\right)^2 N^2\left(1 - \frac{1}{4N^2}\right)^N \frac{1}{4N^2}}{4(N-1)} \\ = & \frac{1}{8N^4}\left((32N^2 + 8N)2^{-N} + 1 - \frac{4N^2 - N - 7}{2(N+6)(N-1)}\right)\left(1 - \frac{1}{4N^2}\right)^N. \end{aligned} \tag{41}$$

In this expression, both $\frac{1}{8N^4}$ and $\left(1-\frac{1}{4N^2}\right)^N$ are positive for all $N \geq 14$. The middle term is strictly negative according to Lemma 6 and, therefore, we have that Eq. (41) is upper-bounded by 0. In particular, we have shown that for every $N \geq 14$ there is a problem distribution D such that $\Delta_N^{N+1} < 0$, which was to be demonstrated. □

5 Discussion and conclusion

Having shown that 1-means is monotonic, while 2-means is not even weakly monotonic, the obvious question that comes up is what we can say about $k > 2$. Our current conviction is that we can probably design problematic cases, similar to the one for 2-means, for $k > 2$ as well. We have, however, not been able to do so up until now. A first step could be to empirically show that, say, for 3-means there are at all distributions where nonmonotonicity occurs. It may only then be sensible to look for a proof showing that it is not weakly monotonic. Incidentally, note that for 3-means, we need at least a four-point problem, which may point at an even more involved proof for this case.

The current proof seems to strongly hinge on the discreteness of the chosen three-point distributions from Definition 6. We conjecture, however, that a similar proof can be constructed on the basis of a class of continuous distributions, rather than discrete ones. Our current idea is that every one of the three discrete locations A, B, and C can probably be replaced by a narrow enough uniform distribution around that point such that the crucial steps in our proof still go through. We do admit, however, that it may be rather nontrivial to precisely reformulate all arguments. Nonetheless, we do believe that the distribution's discreteness is not essential.

Two further research direction that, we think, are of interest are, firstly, whether k-means can be turned into a monotonic learner and, secondly, what we can say about the learning curve behavior of closely related Gaussian mixture models (Welling & Kurihara, 2006; McLachlan et al., 2019; Lücke & Forster, 2019). The latter are closely linked to k-means in particular when the covariance matrices of the mixture components are assumed to be the identity and the mixture priors are all equal. What we wonder especially is whether moving from the within-group squared loss to the (negative) log-likelihood and/or going from hard to soft assignments of points to clusters provides any benefits when it comes to monotonic behavior.

As for the former research question, there are some wrapper techniques to turn learners monotonic in expectation. Viering et al. (2020) and Bousquet et al. (2022) rely on the 0-1 loss specifically and cannot be applied directly to our setting. On the other hand, Mhammedi (2021)'s wrapper only assumes the loss to be bounded, which is readily fulfilled for k-means in case the support of the distribution is bounded as well. Unfortunately, the proof in (Mhammedi, 2021) turned out to be defective and the original result was updated in a correction mentioned in (Footnote 6 (Mhammedi, 2022)). It now only ensures monotonicity up to an additive term with an N^{-1} rate. Still, assuming bounded distributional support, Mhammedi (2021) at least gives us some possibility to control the monotonicity of k-means.

In the context of potential wrappers, what is important to note, however, is that these algorithms essentially change the base algorithm. Are we still dealing with k-means now or is it a different learner altogether? More directly related to the idealized version that we consider are the k-means algorithms in practical use (Sect. 2.2 provides some pointers).

From a monotonicity point of view, the fact that these typically do actually not provide a globally optimal clustering could result in a smoother, maybe even monotonous learning curve. In other words, the regularizing effect from the suboptimality of practical k-means optimization could actually promote monotonicity. Regularization can, however, both fix and create nonmonotonicity (Loog et al., 2019; Nakkiran et al., 2020; Viering & Loog, 2022) and it could be interesting to see how different k-means++ possibly behaves compared to other lesser-optimal algorithms.

All in all, our findings add clustering to the list of potentially nonmonotonic behaving learners, next to some classifiers, regression techniques, and density estimators (cf. Loog et al, 2019). From a theoretical point of view, this is crucial as it provides us, for instance, with a deeper understanding of how learning curves can at all behave. The practitioner may object that "this does not happen on real-world problems." Though we are, in principle, willing to believe this, what proof do we really have? If anything, some recent study by Mohr et al. (2022), on a large number of data sets in combination with various classifiers, showed that nonmonotonic learning curve behavior does occur. The least any practitioner should be is aware that nonmonotonicity can happen.

Acknowledgements We like to sincerely thank the two anonymous reviewers for their critical and insightful comments and the lead guest editor for his considerate and decisive handling of our initial submission. Our gratitude also goes out to Tom J. Viering and Alexander Mey for advising us on the precise interpretations of the works by Bousquet et al. (2022) and Mhammedi (2021). In particular, we thank them for pointing out Footnote 6 in an updated version of the latter work (see Mhammedi, 2022).

Author contributions ML and MB conceived the initial research idea. ML and JHK designed the proof strategy. ML provided the main proof. ML and JHK provided proofs for the lemmas. ML, JHK, and MB performed literature research and wrote the related work section. ML drafted the initial paper and took care of writing the revision. ML, JHK, and MB planned the revision and made critical adjustments to the initial and the revised manuscript.

Funding No financial support was received for this blue-sky research.

Availability of data and material Not applicable.

Code available Not applicable.

Declarations

Conflict of interest Not applicable.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Aloise, D., Deshpande, A., Hansen, P., et al. (2009). Np-hardness of euclidean sum-of-squares clustering. *Machine learning, 75*(2), 245–248.

Arthur, D., Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms, pp 1027–1035.

Belkin, M., Hsu, D., Ma, S. et al (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences 116(32):15,849–15,854.

Ben-David, S. (2004). A framework for statistical clustering with a constant time approximation algorithms for k-median clustering. In: International Conference on Computational Learning Theory, Springer, pp 415–426.

Ben-David, S., von Luxburg, U., Pál, D. (2006). A sober look at clustering stability. In: International Conference on Computational Learning Theory, Springer, pp 5–19.

Ben-David, S., Loker, D, Srebro, N., et al (2012). Minimizing the misclassification error rate using a surrogate convex loss. In: Proceedings of the 29th International Conference on Machine Learning, pp 83–90.

Biau, G., Devroye, L., & Lugosi, G. (2008). On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory, 54*(2), 781–790.

Bock, H. H. (2007). Clustering methods: a history of k-means algorithms. Selected contributions in data analysis and classification pp 161–172.

Bousquet, O. J., Daniely, A., Kaplan, H., et al (2022). Monotone learning. In: Conference on Learning Theory, PMLR, pp 842–866.

Buhmann, J. (1998). Empirical risk approximation: An induction principle for unsupervised learning. Tech. Rep. IAI-TR98-3, Rheinische Friedrich-Wilhelms-Universität.

Chen, Z., Loog, M., Krijthe, J.H. (2023). Explaining two strange learning curves. In: BNAIC/BeNeLearn Post-Proceedings, p accepted.

Chichignoud, M., & Loustau, S. (2014). Adaptive noisy clustering. *IEEE Transactions on Information Theory, 60*(11), 7279–7292.

Cohen-Addad, V., Guedj, B., Kanade, V., et al (2021). Online k-means clustering. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp 1126–1134.

Dalenius, T. (1950). The problem of optimum stratification. *Scandinavian Actuarial Journal, 1950*(3–4), 203–213.

Dasgupta, S. (2008). The hardness of k-means clustering. Technical Report CS2008-0916, Computer Science and Engineering Department, University of California at San Diego.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

Ghadiri, M., Samadi, S., Vempala, S. (2021). Socially fair k-means clustering. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp 438–448.

Gould, H.W. (2010). Fundamentals of series: Table II: Examples of series which appear in calculus, edited and Compiled by Jocelyn Quaintance.

Hartigan, J. (1978). Asymptotic distributions for clustering criteria. The Annals of Statistics pp 117–131.

Jancey, R. C. (1966). Multidimensional group analysis. *Australian Journal of Botany, 14*(1), 127–130.

Klochkov, Y., Kroshnin, A., & Zhivotovskiy, N. (2021). Robust k-means clustering for distributions with two moments. *The Annals of Statistics, 49*(4), 2206–2230.

Levrard, C. (2013). Fast rates for empirical vector quantization. *Electronic Journal of Statistics, 7*, 1716–1746.

Levrard, C. (2015). Nonasymptotic bounds for vector quantization in hilbert spaces. *The Annals of Statistics, 43*(2), 592–619.

Liu, Y. (2021). Refined learning bounds for kernel and approximate *k*-means. *Advances in Neural Information Processing Systems, 34*, 6142–6154.

Loog, M., Duin, R.P. (2012). The dipping phenomenon. In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, pp 310–317.

Loog, M., Viering, T. (2022). A survey of learning curves with bad behavior: or how more data need not lead to better performance. arXiv preprint arXiv:2211.14061.

Loog, M., Krijthe, J.H., Jensen, A.C. (2016). On measuring and quantifying performance: error rates, surrogate loss, and an example in semi-supervised learning. In: Handbook of Pattern Recognition and Computer Vision. World Scientific, p 53–68.

Loog, M., Viering, T., Mey, A. (2019). Minimizers of the empirical risk and risk monotonicity. Advances in Neural Information Processing Systems 32.

Loog, M., Viering, T., Mey, A., et al. (2020). A brief prehistory of double descent. *Proceedings of the National Academy of Sciences, 117*(20), 10625–10626.
Lücke, J., & Forster, D. (2019). k-means as a variational EM approximation of Gaussian mixture models. *Pattern Recognition Letters, 125*, 349–356.
MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, University of California Press, pp 281–298.
Maurer, A. (2016). A vector-contraction inequality for rademacher complexities. In: International Conference on Algorithmic Learning Theory, Springer, pp 3–17.
McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application, 6*, 355–378.
Meek, C., Thiesson, B., & Heckerman, D. (2002). The learning-curve sampling method applied to model-based clustering. *Journal of Machine Learning Research, 2*, 397–418.
Mhammedi, Z. (2021). Risk monotonicity in statistical learning. Advances in Neural Information Processing Systems 34.
Mhammedi, Z. (2022). Risk-monotonicity in statistical learning. arXiv preprint arXiv:2011.14126.
Mohr, F., Viering, T.J., Loog, M., et al (2022). LCDB 1.0: An extensive learning curves database for classification tasks. In: Machine Learning and Knowledge Discovery in Databases, ECMLPKDD. Springer, Lecture Notes in Computer Science, p accepted.
Nakkiran, P., Venkat, P., Kakade, S., et al (2020). Optimal regularization can mitigate double descent. arXiv preprint arXiv:2003.01897.
Perlich, C. (2010). Learning curves in machine learning. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 577–488). Springer.
Pollard, D. (1981). Strong consistency of k-means clustering. The Annals of Statistics pp 135–140.
Rakhlin, A. (2005). Stability of clustering methods. In: NeurIPS Workshop on Theoretical Foundations of Clustering.
Rakhlin, A., Caponnetto, A. (2006). Stability of *k*-means clustering. Advances in neural information processing systems 19.
Steinhaus, H. (1956). Sur la division des corps matériels en parties. Bulletin de l'Académie Polonaise des Sciences, Classe III IV(12):801–804.
Stemmer, U. (2021). Locally private k-means clustering. *The Journal of Machine Learning Research, 22*(1), 7964–7993.
Vallet, F., Cailton, J. G., & Refregier, P. (1989). Linear and nonlinear extension of the pseudo-inverse solution for learning boolean functions. *EPL (Europhysics Letters), 9*(4), 315.
Vapnik, V. (1982). Estimation of Dependences Based on Empirical Data. Springer Series in Statistics, Springer-Verlag.
Viering T, Loog M (2022). The shape of learning curves: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence.
Viering, T., Mey, A., Loog, M. (2019). Open problem: Monotonicity of learning. In: Conference on Learning Theory, PMLR, pp 3198–3201.
Viering, T.J., Mey, A., Loog, M. (2020). Making learners (more) monotone. In: International Symposium on Intelligent Data Analysis, Springer, pp 535–547.
Welling, M., Kurihara, K. (2006). Bayesian k-means as a "maximization-expectation" algorithm. In: Proceedings of the 2006 SIAM international conference on data mining, SIAM, pp 474–478.
Zubek, J., & Plewczynski, D. M. (2016). Complexity curve: a graphical measure of data complexity and classifier performance. *PeerJ Computer Science, 2*, e76.