Open Problem: Monotonicity of Learning

Tom Viering T.J.VIERING@GMAIL.COM
Alexander Mey A.MEY@TUDELFT.NL
Delft University of Technology, The Netherlands

Marco Loog M.LOOG@TUDELFT.NL
University of Copenhagen, Denmark
Delft University of Technology, The Netherlands

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We pose the question to what extent a learning algorithm behaves monotonically in the following sense: does it perform better, in expectation, when adding one instance to the training set? We focus on empirical risk minimization and illustrate this property with several examples, two where it does hold and two where it does not. We also relate it to the notion of PAC-learnability.

Keywords: Finite sample behavior, Learnability, Monotonic learning

Introduction. Recently, there has been an increasing amount of attention on machine learning algorithms that are presently referred to as robust or safe, meaning that even when assumptions are violated, performance will not degrade significantly (Loog, 2010). The focus is mostly on settings that are slightly different from supervised learning such as online learning (Koolen et al., 2016), domain adaptation (Liu et al., 2015) and semi-supervised learning (Krijthe and Loog, 2017). The open problem presented here makes the point that such robustness and safety properties are not even fully understood for standard supervised learning and density estimation.

We focus on what we will refer to as the *monotonicity* of a learner's performance: given one additional training instance, to what extent can we expect a learner to improve? Or, equivalently, when is the so-called learning curve monotone (Shalev-Shwartz and Ben-David, 2014)? While this property is undoubtedly desirable, and most of us expect such behavior, there are surprising counterexamples. This open problem asks to unravel this behavior.

Understanding theoretical properties of learning curves can set expectations for practitioners. For example, if we know that a learner is monotone, but we observe the opposite in practice, we know that this behaviour must have another explanation, such as a finite sampling effect.

Preliminaries and Related Work. Let $S_n = (z_1, \ldots, z_n)$ be a training set of size n, sampled i.i.d. from an (unknown) distribution D over a domain $\mathcal{Z}$. The learner A we consider performs *empirical risk minimization* (ERM). Its output is $A(S_n)$, i.e., a hypothesis h from a prespecified set $\mathcal{H}$ that minimizes the empirical risk over S_n based on a loss function $\mathcal{L} : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$. In statistical learning, performance is measured through this loss and the aim is to minimize the true risk $L_D(h) = \mathbb{E}_{z \sim D}\, \mathcal{L}(h, z)$. One can define classification problems, regression, and density estimation in such terms.

Before we formally introduce the concept of monotonicity, we mention related works that already report on non-monotone learning behavior. Duin (1995) and Opper and Kinzel (1996) describe the so-called peaking phenomenon for classification: when the dimensionality is approximately equal to the size of the training set, the risk in terms of the zero-one loss and mean squared error has a maximum (it peaks). This happens for models that require estimates of the (pseudo-)inverse of the covariance matrix (Raudys and Duin, 1998), such as linear regression.

Loog and Duin (2012) describe what they call dipping: the evaluation risk attains a global minimum for some finite n. Even for $n \to \infty$ the risk never recovers. This phenomenon can occur when there is a mismatch between target (e.g. zero-one) and surrogate loss (e.g. hinge). Ben-David et al. (2012) analyze this mismatch between surrogate and zero-one loss in more detail.

We focus on the setting where the loss the learner optimizes matches the loss it is evaluated with. Thus the observed behaviour in our examples cannot be explained through the dipping phenomenon. This makes our findings more unexpected and the open problem more appealing. Note, indeed, that our learner A (performing ERM) is implicitly associated with a specific loss $\mathcal{L}$ and set $\mathcal{H}$.

The Monotonicity Property. The idea is that with an additional instance a learner should improve its performance in expectation over the training set. We need the following building block.

Definition 1 (local monotonicity) *A learner A is locally or (D, n)-monotone with respect to a distribution D and an $n \in \mathbb{N}$ if*

$$\mathop{\mathbb{E}}_{S_{n+1} \sim D^{n+1}} L_D(A(S_{n+1})) \leq \mathop{\mathbb{E}}_{S_n \sim D^n} L_D(A(S_n)).$$

We may want to construct stronger properties from this, e.g. monotonicity for all n. Also, since the distribution D is unknown, we may want monotonicity to hold for any D on the domain $\mathcal{Z}$.

Definition 2 ($\mathcal{Z}$-monotonicity) *A learner A is $\mathcal{Z}$-monotone if, for all $n \in \mathbb{N}$ and distributions D on $\mathcal{Z}$, it is (D, n)-monotone.*

Examples. We now turn to some illustrations and consider to what extent they are $\mathcal{Z}$-monotone. In the remainder, we refer to $\mathcal{Z}$-monotone as monotone. It will be clear from the context what $\mathcal{Z}$ is.

Example I: mean estimation of a normal distribution (monotone). We perform density estimation with a normal distribution with *fixed variance* $\sigma^2 > 0$ and *unknown mean*. The hypothesis class is $\mathcal{H}_\sigma = \left\{ h : z \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2} \right) \mid \mu \in \mathbb{R} \right\}$. We choose the domain $\mathcal{Z} \subset [-1, 1]$. This choice ensures that any distribution D has a finite mean and finite variance. We use negative log-likelihood as loss. Thus ERM is equivalent to maximum likelihood (ML) estimation for this setting. The optimum that ERM finds is $\mu = \frac{1}{n} \sum_i z_i$. The expected risk equals

$$\mathop{\mathbb{E}}_{S_n \sim D} L_D(A(S)) = \frac{1}{2} \log(2\pi\sigma^2) + \frac{\sigma_D^2}{2\sigma^2} \left(1 + \frac{1}{n} \right),$$

where σ_D^2 is the true variance of D. So the expected risk decreases monotonically in n.

Example II: variance estimation of a normal distribution (not monotone). We take the same domain and loss function as in Example I, but now estimate the variance, while keeping the mean fixed to 0. The hypothesis set becomes $\mathcal{H}_{\mu=0} = \left\{ h : z \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{z^2}{2\sigma^2} \right) \mid \sigma > 0 \right\}$ and the ML estimate equals $\sigma = \frac{1}{n} \sum_i z_i^2$. This example does not obey the monotone principle. Consider a distribution D that only has support on $\{1, \frac{1}{10}\}$. Let D be given by the probability mass function $p(1) = \alpha$ and $p(\frac{1}{10}) = 1 - \alpha$. For $0 < \alpha < 0.0235$ one can easily check numerically that $\mathbb{E}_{S_1} L_D(A(S_1)) < \mathbb{E}_{S_2} L_D(A(S_2))$, showing that the monotonocity property does not hold.

Example III: linear regression (not monotone). Take $\mathcal{H} = \{h \mapsto wx | w \in \mathbb{R}\}$ as hypothesis set and use the mean squared error as loss function. We choose the domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with $\mathcal{X} \subset [-1, 1]$ and $\mathcal{Y} \subset [0, 1]$. We define D through a probability mass function $p(x, y)$. Take $p(\frac{1}{10}, 1) = 1 - \alpha$ and $p(1, 1) = \alpha$, and $p(x, y) = 0$ otherwise. Again, one can find numerically that $\mathbb{E}_{S_1} L_D(A(S_1)) < \mathbb{E}_{S_2} L_D(A(S_2))$ for $0 < \alpha < 0.0047$. This shows this learner is not monotone.

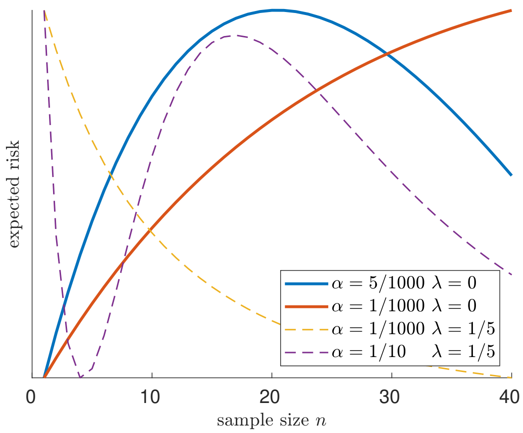

Figure 1: Non-monotone behavior as observed in Example III.

Figure 1 plots a rescaled version of the expected risk against the sample size n for several settings. The thick lines correspond to ERM. First of all, observe that by changing α, we can shift the peak. This shows that the behaviour is unrelated to the peaking behaviour (Duin, 1995), since peaking would occur at $n \approx d = 1$. Second, if we add λI to the empirical covariance matrix, which corresponds to L_2-regularization of w, we still observe non-monotone behavior, now even for larger values of α (see the dashed lines in Figure 1).

Example IV: the memorize algorithm (monotone). Ben-David et al. (2011) introduced this binary classifier. When evaluated on a test object x that is also present in the training set, this learner returns the label of that training object. In case multiple training examples share the same x, the majority vote is returned. In case the test object is not present in the training set, a default label is returned. This learner is monotone for any distribution under the zero-one loss as it only updates its decision on points that it observes.

Relation to Learnability. From learning theory we know that if the hypothesis class has finite VC-dimension (or other appropriate complexity), the excess risk of ERM is bounded. This bound will be tighter given a larger training set size n. PAC bounds hold with a particular probability, while we are concerned with the risk in expectation over the sample. However, even bounds that hold in expectation over the training sample will not rule out non-monotone behaviour. The expected risk can go up as long as the expected risk stays below the upper bound. Thus high probability or expected risk bounds are insufficient to guarantee monotonicity.

This is illustrated by our examples: Example VI is monotone but is not learnable (Shalev-Shwartz and Ben-David, 2014). Example III is learnable if a regularizer is added to the objective of ERM or if the hypothesis space $\mathcal{H}$ is restricted such that the norm of w is bounded. However, as we have seen in Figure 1, we still can observe non-monotone behaviour in that case.

Open problem(s). First and foremost, we are interested to identify, especially for commonly employed learners, on which domains $\mathcal{Z}$ they will or may not act monotonically. In view of the peaking behaviour, $\mathcal{Z}$-monotonicity for all n may be too strong for some settings. Perhaps monotonicity is only possible if n is larger than some N that may depend on $\mathcal{Z}$ and A. For Examples II and III it is an open problem whether they satisfy this weaker notion, and for which (smallest) N this notion is satisfied. Other related notions of monotonicity may also be of interest. For example, instead of demanding a lower loss, we may require that the loss does not degrade too much. Or we can demand the property to hold with high probability with respect to both samples.

More generally, we may ask: why and how does this behaviour occur? And maybe more importantly: how can we provably avoid non-monotone behaviour? What conditions does a learner need to satisfy to be monotone? Perhaps particular loss functions lead to monotone learners? What if we allow for learning under regularization or other strategies deviating from strict ERM, for example improper learners or randomized decision rules?

Perhaps it is always possible to find a D for a given $\mathcal{Z}$ on which learners are non-monotone. In that case, is it possible to avoid non-monotone behaviour under some assumptions on D? Realizeability or well-specification could be good candidate-assumptions on D. In fact, this raises the issue to what extent well-specified statistical models can actually be proven to behave monotonically. For instance, is Example II monotone if the problem is well-specified?

All in all, we believe the question of monotonicity of learning offers various tantalizing questions to study, some of which may yet have to be formulated.

References

Shai Ben-David, Nathan Srebro, and Ruth Urner. Universal learning vs. no free lunch results. In *Philosophy and Machine Learning Workshop NIPS*, 2011.

Shai Ben-David, David Loker, Nathan Srebro, and Karthik Sridharan. Minimizing the misclassification error rate using a surrogate convex loss. In *Proceedings ICML*, pages 83–90, 2012.

Robert P W Duin. Small sample size generalization. In *Proceedings of the Scandinavian Conference on Image Analysis*, volume 2, pages 957–964, 1995.

Wouter M Koolen, Peter Grünwald, and Tim van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. In *NIPS*, pages 4457–4465, 2016.

Jesse H Krijthe and Marco Loog. Projected estimators for robust semi-supervised classification. *Machine Learning*, 106(7):993–1008, 2017.

Anqi Liu, Lev Reyzin, and Brian D Ziebart. Shift-pessimistic Active Learning Using Robust Bias-aware Prediction. In *Proceedings of AAAI-15*, pages 2764–2770, 2015.

Marco Loog. Constrained parameter estimation for semi-supervised learning: the case of the nearest mean classifier. In *ECML PKDD 2010*, pages 291–304. Springer, 2010.

Marco Loog and Robert P W Duin. The dipping phenomenon. In *Proceedings of the IAPR S+SSPR*, pages 310–317. Springer, 2012.

Manfred Opper and Wolfgang Kinzel. Statistical mechanics of generalization. In *Models of neural networks III*, pages 151–209. Springer, 1996.

Sarunas Raudys and Robert P W Duin. Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix. *Pattern recognition letters*, 19(5-6):385–392, 1998.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.