Minimizers of the Empirical Risk and Risk Monotonicity

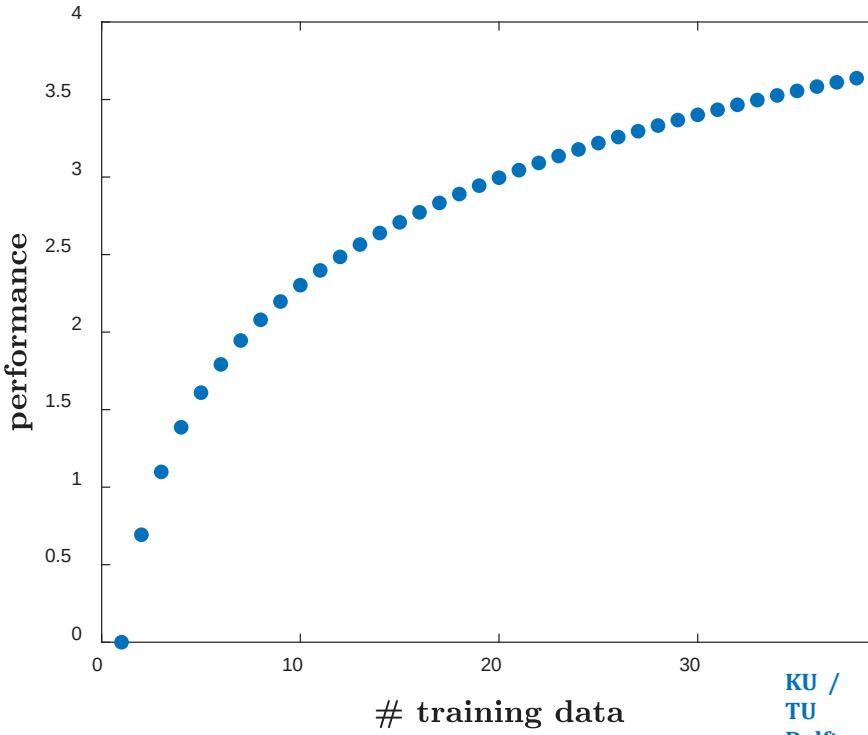
Loog • Viering • Mey

Delft University of Technology • University of Copenhagen

KU / TU Delft

Minimizers of the Empirical Risk and Risk Monotonicity

Popularly formulated,
our works asks :
Can one expect improved
generalization performance
with more training data?



Minimizers of the Empirical Risk and Risk Monotonicity

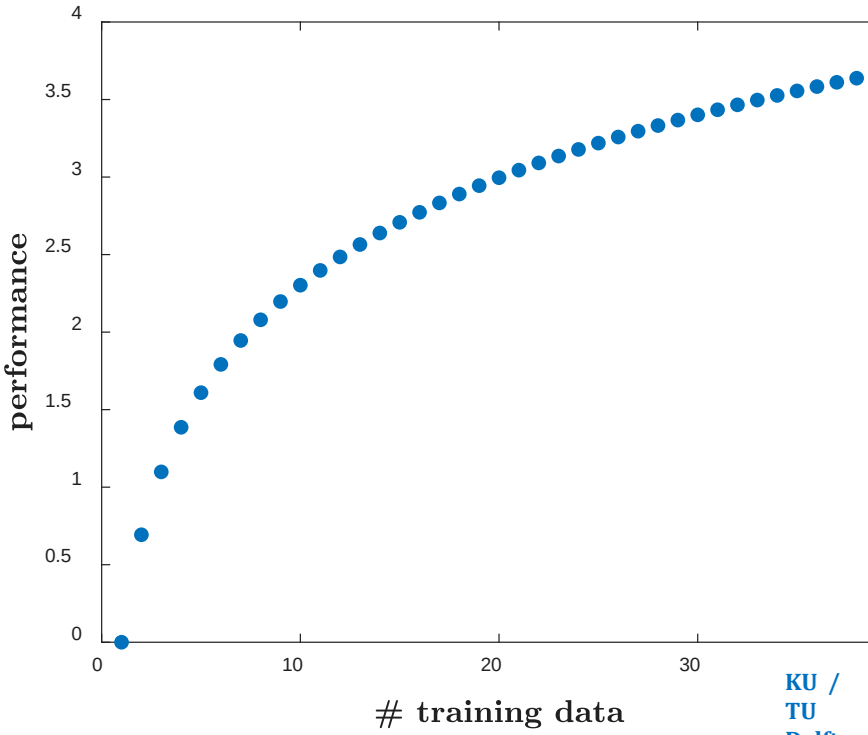
Popularly formulated,
our works asks :
Can one expect improved
generalization performance
with more training data?

The answer?
Majority of people : yes [of course!]
Our paper : nope, unfortunately not

Making It a Bit More Precise…

That learners become better with more training data seems intuitive

Majority indeed takes it for granted that **learning curves** show improved performance with more data



Making It a Bit More Precise…

The learning setting we consider :

$S_n \in \{z_1, \dots, z_n\}$ sampled i.i.d. from D over a domain $\mathcal{Z}$

Hypothesis class $\mathcal{H}$ and loss $\ell: \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$

Learner A maps from set of all samples $\mathcal{S}$ to hypothesis class $\mathcal{H}$

$R_D(h) = \mathbb{E}_{z \sim D} \ell(z, h)$ the expected true loss

We want to study learning curves :

How does $\mathbb{E}_{S_n \sim D^n}[R_D(A(S_n))]$ with increasing n behave?

Monotonicity

We want to study learning curves :

How does $\mathop{\mathbb{E}}_{S_n \sim D^n}[R_D\big(A(S_n)\big)]$ with increasing n behave?

In particular, we ask to what extent we have

$$\mathop{\mathbb{E}}_{S_{n+1} \sim D^{n+1}}[R_D\big(A(S_{n+1})\big) - R_D\big(A(S_n)\big)] \leq 0$$

for all distributions D on domain $\mathcal{Z}$

We call this (rather basic) property **locally monotonic** in n

That is, the learning curves does not increase

Previous Monotonicity Results

Micchelli (1979) : lower bound learning curves of Gaussian processes

Many studies within context of neural networks were done at end of 1980s, beginning of 1990s; Tishby, Haussler, and others

End of 1990s, beginning of 2000s, focus shifts from neural nets to Gaussian processes; Opper, Sollich, and others

Previous Nonmonotonicity Results

Duin (1995) shows underdetermined setting, least squares methods can behave nonmonotonic in terms of error rate [a]

Devroye (1996) conjectures that consistent classifiers that perform better with increasing training sizes do not exist [b]

Grünwald (2011) : unfortunate choice of prior gives nonmonotonicity

Loog (2012) : best expected 0-1 loss can be attained for finite samples [c]

[a] Directly related results have been presented by Opper (1996, 2001), Duin (2000), Krämer (2009), Belkin (2018), and Spigler (2018)

[b] The conjecture is formulated following some provably nonmonotonic consistent classifiers

[c] Devroye (1996) already showed that likelihood models can converge to suboptimal classifiers; incidentally, Ben-David (2012) provides another dipping example

Our Nonmonotonicity Results

Results by Duin, Devroy, and others hinge on

- Fact that the model is underdetermined and/or
- Discrepancy between loss optimized and loss evaluated with
- [e.g. surrogate loss is optimized, evaluation in error rates]

Our results : nonmonotonicity can occur at all sample sizes even when evaluation loss matches loss optimized (ERM!)

- Moreover, we show such behavior for classification, regression, and density estimation tasks

But Let's Start Off Positively…

Let $\mathcal{H}$ be the class of normal distributions on $\mathbb{R}^d$ with given covariance matrix for which the mean is to be estimated

Take $\mathcal{Z} \subset \mathbb{R}^d$ and as loss the negative log-likelihood

Theorem · If $\mathcal{Z}$ is bounded, the learner A is **globally monotonic** (i.e., locally monotonic for all n)

Our First Nonmonotonicity Result

Consider linear models without intercept :
take $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \{-1, +1\}$ and $\mathcal{H} = \mathbb{R}^d$

Let A be the minimizer of the empirical risk

Theorem · Assume A either optimizes the squared, the absolute, or the hinge loss. Assume $\mathcal{Y}$ contains at least one element. If there exists an open ball $B_0 \subset \mathcal{X}$, then this risk minimizer is **not** locally monotonic for any $n \in \mathbb{N}$

Our Second Nonmonotonicity Result

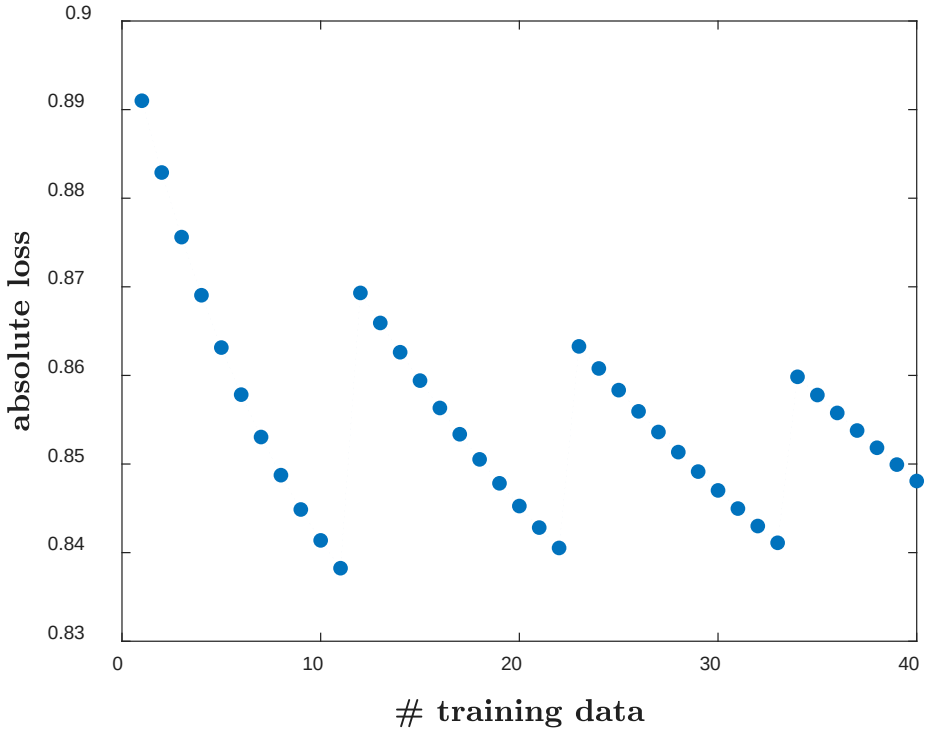
Let $\mathcal{H}$ be the class of normal distributions on $\mathbb{R}$ with given mean for which the variance is to be estimated

With $\mathcal{Z} = \mathbb{R}$, similar to the first results, we have

Theorem · Assume A optimizes the negative log-likelihood. If there exists an open ball $B_0 \subset \mathcal{Z}$, then this risk minimizer is **not** locally monotonic for any $n \in \mathbb{N}$

Some Empirical Results

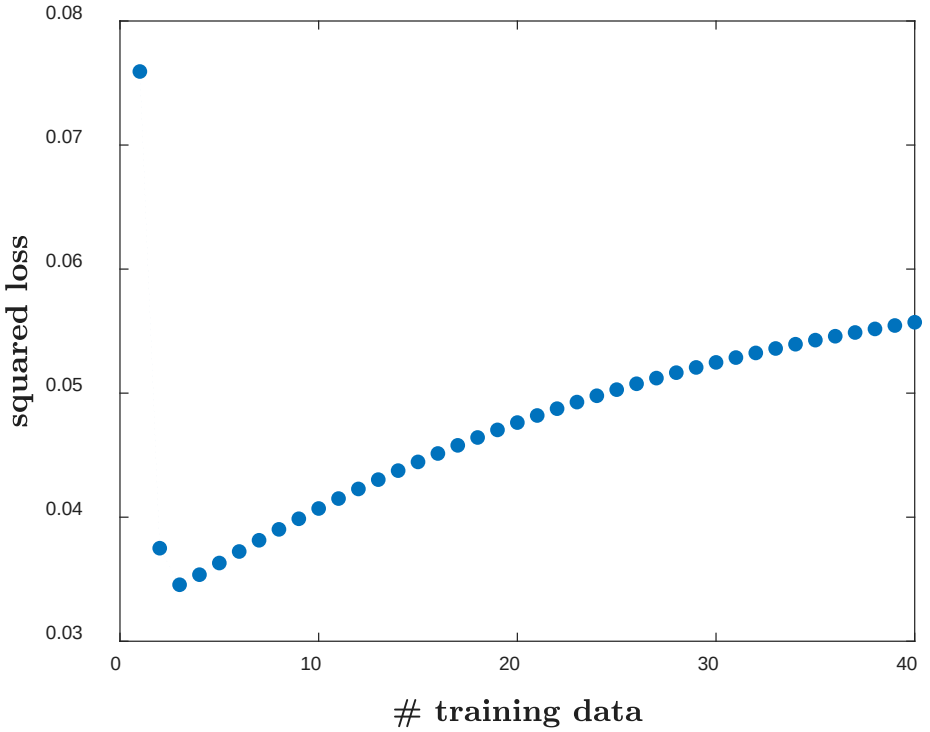
E.g. to illustrate the funky shapes learning curves can take on [a]



[a] Code available through paper's supplement

Some Empirical Results

E.g. to illustrate that, also with intercept, stuff goes wrong [a]



[a] Code available through paper's supplement

Conclusions

Learning curves can display unexpected behavior

Our study :

- Nonmonotonic behavior possible in well-determined ERM
- Also indicates learning curves are not well understood

We raise open issues and possible research directions

- All of which should lead to further insight into such curves