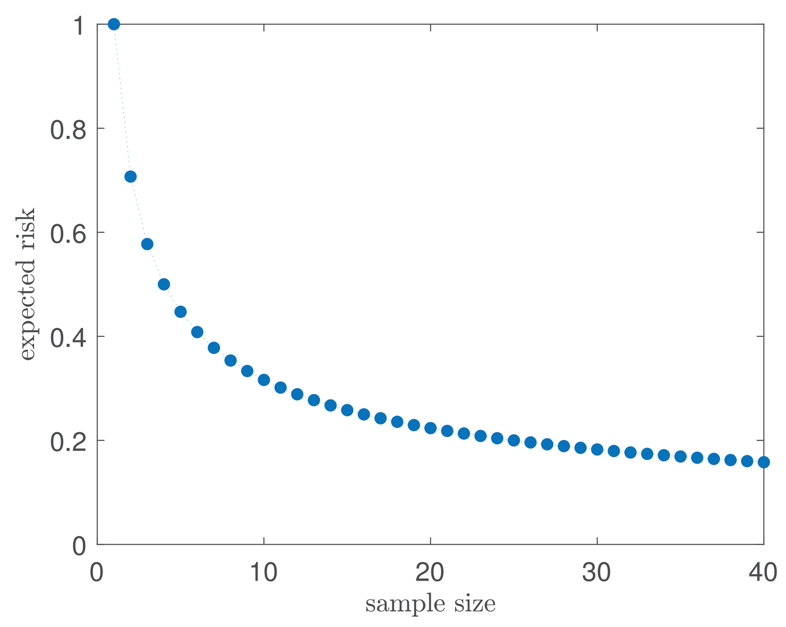
Minimizers of the Empirical Risk and Risk Monotonicity

Marco Loog[+×] **Tom Viering**[+] **Alexander Mey**[+]

[+]Delft University of Technology [×]University of Copenhagen

Introduction. It seems intuitive for learning curves to display monotonic improvement with every added training sample. We show: various well-known empirical risk minimizers can, in fact, behavior non-monotonically, even for arbitrarily large training sample sizes.



Risk monotonicity. $S_n = (z_1, \dots, z_n)$ training set i.i.d. from distribution D over a domain $\mathcal{Z}$. $\mathcal{H}$ hypothesis class and $\ell : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$ a loss function. Objective: minimize

$$R_D(h) := \mathop{\mathbb{E}}_{z \sim D} \ell(z, h). \tag{1}$$

Let $\mathcal{S} := \mathcal{Z} \cup \mathcal{Z}^2 \cup \mathcal{Z}^3 \cup \dots$ and learner $A_{\text{erm}} : \mathcal{S} \to \mathcal{H}$ minimizes the empirical risk R_{S_n} over the training set:

$$A_{\text{erm}}(S_n) := \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, h). \tag{2}$$

Definition. *A learner is weakly monotonic with respect to a loss ℓ if there is an integer $N \in \mathbb{N}$ such that for all $n \geq N$ and for all distributions D on $\mathcal{Z}$,*

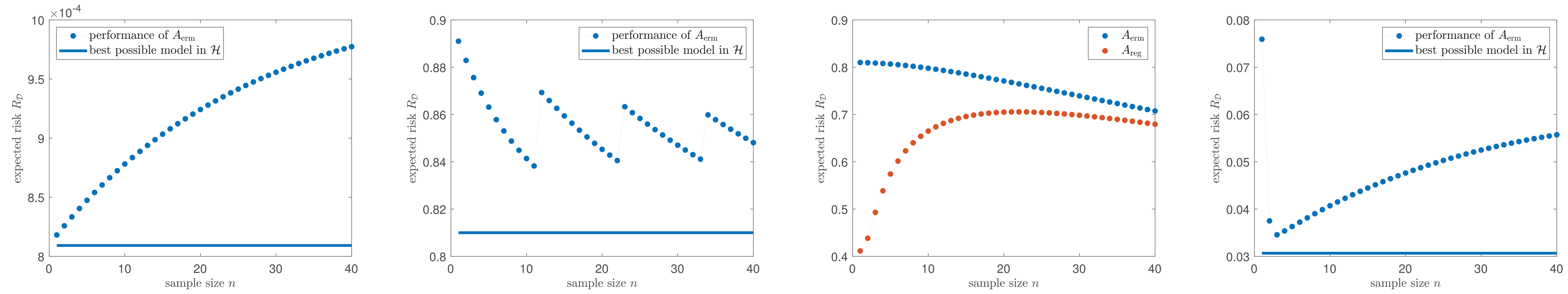
$$\mathop{\mathbb{E}}_{S_{n+1} \sim D^{n+1}} [R_D(A(S_{n+1})) - R_D(A(S_n))] \leq 0. \tag{3}$$

Theorem 0. *Take $\mathcal{H}$ the class of normal distributions with fixed covariance, the mean to be estimated, $\mathcal{Z} \subset \mathbb{R}^d$, and the negative log-likelihood as loss. If $\mathcal{Z}$ is bounded, the learner A_{erm} is monotonic.*

Of more interest are the negative results.

Theorem 1. *If $\exists$ open ball B_0 that contains* 0, *such that $B_0 \subset \mathcal{Z}$, then estimating the variance using negative log-likelihood of a one-dimensional normal density* is not *weakly monotonic.*

Theorem 2. *Consider a linear A_{erm} without intercept and assume it either optimizes the squared, the absolute, or the hinge loss. Assume $\mathcal{Y}$ contains at least one nonzero element. If $\exists$ open ball B_0 that contains* 0, *such that $B_0 \subset \mathcal{X}$, then this risk minimizer* is not *weakly monotonic.*



(a) Squared loss. $P(a) = 0.00001$

(b) Absolute loss. $P(a) = 0.1$

(c) Squared and regularized loss. $P(a) = 0.01$

(d) Squared loss and bias term.

Experiments. Subfigures a, b, c consider distributions with two points: $a = (1, 1)$ and $b = (\frac{1}{10}, 1)$ (first coordinate input, second output). Subfigure d's distribution is supported on three points: $a = (1, 1)$, $b = (\frac{1}{10}, -1)$, and $c = (-1, 1)$ (input, output) with $P(a) = 0.01$, $P(b) = 0.01$, and $P(c) = 0.98$.