We thank all three reviewers for their considerate appraisal and constructive comments. All agree that the idea presented in the paper is interesting. Reviewer #1 notes, in addition, that the work is fundamental in nature and that the discussed phenomenon is non-intuitive, intriguing, and warrants careful study. Reviewer #2 states that the analysis of the example distributions is interesting and clear. Reviewer #3 considers our results surprising and deems the originality and the quality of the work to be high. In what follows, we respond to the most important issues from the particular reviews.

Reviewer #1

We cannot do otherwise than agree with the sentiment expressed that more intuition into the when and why of (non)monotonicity in empirical risk minimization would be appreciated. This should also be apparent from the discussion that we provide in the paper. We do hope that our work spurs others to delve into what we think is an interesting topic. Indeed, the primary insight obtained in this submission is that nonmonotonicity at all happens, even in fairly standard settings. This is of course of interest in itself. At this point, we are simply unable, however, to offer any deeper intuition or to comment on the potential occurrence of nonmonotonicity in practice.

The reviewer's remark about identifying properties that increase the risk on nonmonotonicity is interesting. We will include it in the discussion as another direction for future research and relate it to our suggestion of $\mathcal{D}$-monotonicity, which aims to ensure monotonicity by making assumptions about the distributions considered. For further clarification, in advance of the discussion that comes with Equation (8) in Section 6, we will add one or two sentences directly after Lemma 1 and Remark 2. These will state that, to some extent, the results from the lemma and the remark show that if the learning of the single point b does not happen fast enough, local monotonicity cannot be guaranteed. The necessary space for all additional text can be realized through the compression or even removal of (parts of) Section 3 and the categorical distribution example in Section 4 (as suggested by the reviewer).

We thank the reviewer for the minor comments provided, which we will use to further clarify our paper.

Reviewer #2

We agree that research into the exact conditions that allow for monotonicity would be most welcome. Similarly, it is of interest to further study the relationship between PAC-learnability and monotonicity. The discussion in our paper states some related directions of interest. Currently, however, we are simply not able to provide any further results into such directions. What we should emphasize more clearly in our paper though, is that PAC-learnability is an essentially different concept, independent of monotonicity. That is, learning problems of all four possible combinations exist: not PAC-learnable and monotonic, PAC-learnable and not monotonic, etc. For instance, the memorize algorithm (line 213 in our paper) is monotone, while it has infinite VC-dimension and so is not PAC-learnable. As such, combinatorial measures like the VC-dimension may not be useful in the analysis of monotonicity.

Incidentally, the reviewer could definitely be right in believing that there are quite a few researchers that would doubt that learning curves necessarily show improved performance with more data. We of course admit that our own believe is based on anecdotal evidence and personal experience—which should be clear from our submission as well. Indeed, up to now, we have met very few people who doubted the monotonicity statement and who were not surprised by our findings. Also, we do think the quote from the book by Shalev-Shwartz and Ben-David, both leading researchers in the field of learning theory, is telling. It states that the learning curve "must start decreasing once the training set size is larger than the VC-dimension" (see page 153 in the free online copy of the book; cited on page 1 of our paper).

Finally, we agree that monotonicity, especially its global form, is rather strict and it is of interest to think about relaxations. One option, making assumptions on the class of distributions, which we call $\mathcal{D}$-monotonicity, is already mentioned in the last section of our submission. The convergence of the risk (as suggested by the reviewer), however, seems a rather different notion that is already fulfilled by any learner that is consistent. We believe that our definition of local monotonicity is of interest exactly because it is not a limit property, but a finite sample size characteristic.

Reviewer #3

We could see that our work lacks some appeal for the practitioner and that it may be primarily the more theoretically inclined NeurIPS attendee that finds our results of interest. Still, our finding is for many so counterintuitive that it can even spark curiosity in the most ardent practitioner. Moreover, we believe that it is important that practitioners understand what type of learning curve behavior is possible, for which our work provides some first results.

Certainly, the two-point discrete distributions are merely constructs to show that nonmonotonic behavior can at all arise. They are definitely not practical. To throw light on more practical settings, we need to further characterize risk monotonicity. A full characterization of this phenomenon—something we also state in the paper's discussion—is of course an ultimate goal, since it would bring clarity to whether such behavior could occur in more realistic settings.