Author Response ("Margin-Based Generalization Lower Bounds for Boosted Classifiers")

We thank the reviewers for the time and expertise invested in these reviews.

Response to the First Reviewer (Reviewer #3). Regarding the intuition behind the "first part" of the distribution (lines 164-179): Thanks for pointing out the confusion, we will try to make the presentation more precise. The intuition we wanted to get across was the following: Assume we could assign a probability of $\frac{1}{10m}$ to every point in $\mathcal{X}$ (which as you say require $|\mathcal{X}| = 10m$). Then most data points would not be sampled. This is great for proving a high generalization error: On a sample of m points, one would only see a small constant fraction of $\mathcal{X}$ and the error would be about $1/2$. Now the issue with the above is that the sample S will consist of m distinct points with a randomly chosen label. This makes it impossible to construct a small hypothesis set that can guarantee a good margin on the sample (point 1. of Theorem 1). Thus instead we create only d points with a probability of being sampled of $1/10m$. Of course the distribution is not proper now (as you also remark), hence we need to add one point with large mass. Since a sample will miss a constant fraction of the d points, the generalization error will be proportional to d/m. The last part of the proof is choosing the largest d for which we can still guarantee good margins on the sample. It turns out we can choose d (and hence $|\mathcal{X}|$) as $\theta^{-2} \ln |\mathcal{H}|$. We will make sure to comment on $|\mathcal{X}|$ in the final version.

We will also try to add more intuition about the role of the two distributions in the final version of the paper. Intuitively, the second distribution with the slightly biased labels is preferable in terms of proving large generalization error because it ensures that an algorithm often will be wrong also on points it has already seen in the sample. But we cannot use that distribution all the time, as it also means that many margins will become negative (when the same point has been sampled with two different labels, the margin will be negative on one sample). So the proof "uses" the second distribution as much as the threshold τ allows, and uses the first distribution the remaining time (which yields better margins but smaller error).

Response to the Second Reviewer (Reviewer #4). Regarding the statement of Theorem 1, the change of classifiers is essential to give the theorem meaning. And in fact, it makes the theorem even *stronger*. In particular, since Theorem 1 applies *for all algorithms*, it *simultaneously* applies to *every* algorithm that maximizes margins. One could e.g. choose the algorithm that given a sample spends arbitrarily much time in order to find the voting classifier with the best margins possible on the sample, and still that algorithm would have a large error. Theorem 1 also applies to all other algorithms, even those not trying to maximize margins but maybe something completely different. Since we want Theorem 1 to state something about all possible algorithms, we cannot hope to show that the classifier constructed by the algorithm performs well on the training set, that is, it has a good empirical margin. Hence we need the change of classifier. On the other hand, if we discard the first condition altogether, then the claim becomes almost trivial (at least for smaller values of τ) as a uniform distribution over $\mathcal{X} \times \{-1, 1\}$ can then fail any classifier. We will make sure to emphasize this further in a final version.

Regarding the constant in the exponent in the definition of m for Theorem 2, the thing is that any constant exponent bounded away from 1 will do here. To be concrete the proof requires $\frac{m}{\ln m} \geq \left(\frac{\ln N}{\theta^2}\right)^{1+1/10}$, which in turn is used in the proof of Claim 10 right after line 461 to get that $\ln(u/d) \geq (1/10) \ln u$. However the choice of $1/10$ is arbitrary. This will be explicitly stated in the final version of the paper.

The main consideration for not including the more technical parts of the proofs was to allow us to provide the reader with an intuitive high-level description. As can be seen in the full version of the paper, which was submitted as supplementary material, all the proofs are provided in detail. At the reviewers' discretion, we can include more of the details in the final version of the paper. In any case, the full version of the paper will be published on arXiv.

Naturally, the typos pointed out by the reviewer will be rectified.

Response to the Third Reviewer (Reviewer #5). We believe there is a slight confusion as to the strength of Theorem 1. In particular, Theorem 1 holds for *every algorithm*. This means that one could e.g. take the algorithm $\mathcal{A}$ that outputs the voting classifier obtaining the best possible margins on the sample, and still that algorithm has large generalization error. And by the first point in Theorem 1, that algorithm $\mathcal{A}$ can (and thus will) actually have good margins. Thus Theorem 1 is *even stronger* than if it had said that the any algorithm $\mathcal{A}$ which produces good margins also has large generalization error. Please also see the response to the Second Reviewer where this is also discussed.

Regarding the statement of Theorem 2, the existence of a classifier that satisfies both properties of the theorem shows that one cannot rely on high margins on the training set in order to attain performance better than the upper bounds provide. Thus showing that the known upper bounds are almost tight, if one relies only on the margin distribution.

Once again, we thank all the reviewers for their time and effort invested in this paper, and for valuable remarks.