We thank all reviewers for the detailed and encouraging comments.

General comments

- *Additional experiments:* We would like to emphasize that the focus of this work is theoretical. Therefore, large-scale experiments and applications are beyond the scope of the paper. Our experiments serve as the first proof of concept, which indeed successfully validate our theoretical findings. More extensive empirical testing is left for future work.
- *Additional background/ proof details:* We agree that it would be instructive to include more detailed proof sketches and additional background on tree-embeddings and hyperbolic methods for ML in the main text. If accepted, we will use the additional page to provide more details on proof ideas and extend the related work and background sections.
- *Training vs. test error in Figure 3:* Here, our aim is to study the learning/optimization aspect of the underlying problem. Therefore, in Figure 3, we focus on the training objective and the margin realized on the training set. That said, since the underlying dataset is well separable, in our experiments with adversarial GD, we do observe that the obtained models achieve zero test error with higher α resulting in faster convergence to zero error. Similarly, we briefly discuss in Appendix F.1 that the hyperbolic perceptron achieves zero test error on well separable data. We will clarify this in the revised version.

  On the other hand, while validating our dimension distortion trade-off analysis (line 319-329 and Appendix F.3), where we are interested in the end-to-end performance based on the distortion introduced by the choice of the embedding space, we do report test errors in Table 2 (Appendix F.3).

R1

- *Saturation of robust loss vs. margin loss:* Please note that the margin in Figure 3(c) shows a worst case quantity (minimum margin over all data points), whereas the α-robust loss in Figure 3(b) depicts an average quantity (mean of the adversarial loss at all data points). We will add a clarifying remark in the revised version.
- *Choice of α in practice:* Assumption 1(2) imposes a norm constraint on the adversarial examples, relative to the maximal norm of the training points. Given the constant R_x, one can estimate an upper bound on α. In addition, an upper bound on α depends on how separable the data set is, i.e., the maximal possible margin. Within these constraints, the choice of α is guided by a trade-off between better robustness and longer training time. We will add a clarifying remark in a revised version.

R2

- *When is a hyperbolic approach beneficial:* As discussed in the paper (e.g., Section 2.2 and Section 5), our approach is suitable for truly hierarchical data, which embeds well into low-dimensional hyperbolic space, but requires a high-dimensional Euclidean embedding space for accurate representation. Many data sets that we encounter in practice are hierarchical (e.g., language, social networks, biological networks etc.), which suggests that hyperbolic ML methods could be beneficial for applications in these areas. A growing body of literature studies this empirically, see, e.g., (Monath et al., KDD'19), (Chami et al., NeurIPS'19), (Klimovskaia et al., Nature Communications'20).

R3

- *Related work:* Thank you for the suggestion. We will add a reference to Ganea et al. in the related work section.
- *Figure 3:* Please note that we characterize robust classifiers by the (worst case) margin they achieve. In Figure 3(c) we see that $\alpha = 1$ indeed achieves the best margin.

R4

- *Hyperbolic vs. Euclidean margin:* Thank you for raising this issue. Yes, it is a notation overload. Up to Section 5, γ denotes the hyperbolic margin. In Section 5 we compare hyperbolic and Euclidean margins, which are denoted as γ_H and γ_E to avoid confusion. We will revise Table 1 accordingly and add a clarifying remark in a revised version.
- *Eq. (4.2):* Thank you for your comment. Currently, we explain the notation in Eq. (4.2) in the paragraph following the equation, with a more detailed discussion of the robust loss deferred to the appendix. We will add a discussion regarding our objective and the process that lead to Eq. (4.2) in the revised version.
- *Additional data sets:* As we discussed above, large-scale experiments are beyond the scope of the paper. For the first proof of concept, we chose a data set that fulfills the assumptions in our theory and allows for validation of the guarantees for both the adversarial approach and the dimension-distortion trade-off. Our chosen data set is significantly larger than the data sets used in Cho et al. (1200 vs 50 data points).