We thank all the Reviewers for a careful reading of our paper and for providing useful suggestions for improvements, which we will be happy to implement in the camera-ready version.

Reviewer #1

As we state at the beginning of Sec. 2, Theorems 1 and 3 hold for any activation function, including $\tanh$. Theorem 1 states that in the limit $n \to \infty$ of infinite length of the bit string with the number of layers L kept fixed the Hamming distance scales as $\sqrt{n/(F'(1)\ln n)}$, where $F'(1)$ depends on L but not on n. As we state in Remark 2, for ReLU we always have $F'(1) \leq 1$, while for $\tanh$ depending on the variances of weights and biases $F'(1)$ may grow exponentially with L. Therefore, for finite values of L and n with the $\tanh$ activation function, $F'(1)$ may become comparable with $n/\ln n$ and significantly affect the Hamming distance. We will clarify this point in the camera-ready version.

Reviewer #2

Exploiting our results to understand the stability of trained neural networks under adversarial perturbations is an extremely interesting line of research which we are currently pursuing.

We have performed additional experiments on the MNIST dataset to explore the correlation between the Hamming distance of a training or test picture from the closest classification boundary and the correctness of its classification. Figure 1 shows that incorrectly classified pictures are significantly closer to the boundary than correctly classified ones, thus implying an empirical correlation between Hamming distance and generalization. We will include a discussion in the camera-ready version.

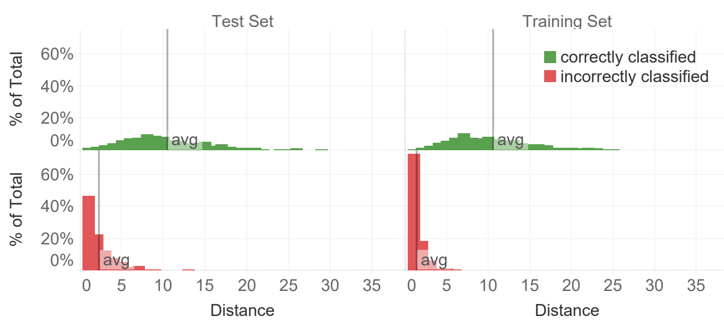

Figure 1: Histogram of correctly and incorrectly classified pictures shows that trained neural networks are far more likely to misclassify points closer to a classification boundary for both the training and test sets. Results are aggregated across 20 different trained neural networks. All neural networks are trained to classify whether digits are even or odd and are trained for 10 epochs using the adam optimizer on the MNIST dataset.

We will move the MNIST results to the main paper swapping them with the detailed proofs and modify Sec. 1.1 as suggested.

The kernel entry associated to two inputs lying on the sphere is a function of their squared Euclidean distance, which coincides with the Hamming distance in the case of bit strings. We are currently working on extending our results to continuous inputs replacing the Hamming distance with the squared Euclidean distance.

We will add in the camera-ready version a discussion on the convergence rate to the Gaussian probability distribution.

We conjecture that the training process keeps the classification boundaries as far as possible from the training pictures, and this results in having most of the pictures that represent a digit still far from the boundaries. Therefore, the distance to the closest boundary is larger for a training or test picture than for a random picture. We will add a comment on this in the camera-ready version.

F is the function that provides the entries of the kernel of the Gaussian process associated to the neural network in terms of the scalar product of the inputs, as defined in eq. (2). We will make the definition more clear in the camera-ready version.

Reviewer #3

We will implement all the suggestions: we will relabel Theorem 3 as Theorem 2, move the MNIST analysis to the main paper swapping it with the detailed proofs, modify Sec. 1.1 as suggested and add error bars to the plots. The reference to Sec. 6.1 is a typo due to a previous version of the paper, the correct reference is Sec. 4.1.