We thank the reviewers for their efforts. Below we address the main comments.

Reviewer #1

We thank the reviewer for his/her positive feedback! Experiments do show that the phenomena described in our results extends to architectures beyond fully connected networks, and deep networks.

We will add details about these experiments in the final version, and leave the extension of our results to more architectures as a future direction.

Reviewer #2

We thank the reviewer for his/her very positive feedback!

We will address his comments in the final version of the paper.

Reviewer #3

We thank the reviewer for his/her overall positive feedback.

We will address his comments in the final version of the paper.

Reviewer #4

We thank the reviewer for his/her feedback. The reviewer is concerned mostly about readability, and we will make every effort to clarify the points that he/she has raised. Yet, given that (1) the reviewer thinks that "The results are strong and very interesting", (2) the positive feedback of the other three reviewers, and (3) the fact that the other reviewers wrote that "the paper is well-rewritten and easy to follow", that "The theoretical claims are sound and the proofs seem correct", and that "(correctness) Seems fine" we ask him/her to reconsider his/her score.

We next address specific concerns:

1. "I do not understand theorem 3.1, what is the length of the gradient flow? What does it mean to fool the network with a gradient flow starting at x_0? I admit the result is not clear at all. In all the paper, I am a bit disturbed by $d_{k+1} = o(d_k)$, it is not clear for me what it means in this context."

   The length of gradient flow is the length of the trajectory of gradient flow, until the network sign is flipped.

   "to fool the network with a gradient flow starting at x_0" means to start gradient flow from x_0 and to reach a point in which the sign of the network is different. That new point is the adversarial example, as it is so close in terms of the Euclidean distance.

   "$d_{k+1} = o(d_k)$" means that we assume that the dimension decreases in every layer. To illustrate that, this holds for instance if $\sqrt{d_k} \leq d_{k+1} \leq d_k/\log(k)$ for every k
2. "The footnote making equivalence between lower bound on singular values and c-surjectivity does not appear immediate or even false"

   We will add a proof to the final version
3. "There are no experiments backing the theoretical claims: it would be good to illustrate them."

   Please see our response to reviewer # 1