We would like to thank the reviewers for their positive and constructive feedback. We hope that with our detailed answers below we can initiate a fruitful discussion that resolves all concerns.

— R1.1): Validity of Theorem for large ϵ. Experiment in Sec. 7.16. R1 is correct, the Theorem establishes *exact* equivalence only for ϵ small enough s.t. $\mathcal{B}^p_\epsilon(\mathbf{x}) \subset X(\phi_\mathbf{x})$. However, we experimentally verify that the correspondence holds *approximately* (to a very good degree) in a region much larger than $X(\phi_\mathbf{x})$: see Sec. 5.4 “Validity of linear approx.”, specific. Fig. 4 (left), as well as Sec. 5.5 “Activation Patterns” specific. Fig. 5 and Fig. 9. Note, Fig. 9 shows the *average* (incl. std. errors) over $\mathbf{x}$ and corresp. adv. $\mathbf{x}^*$ from the test set (not just a single data point). These experiments confirm that $\mathbf{J}_f(\mathbf{x})$ is a good approx. (negligible deviation from linearity) and that the activation pattern change is small ($\sim 3\%$ in adv., $\sim 1\%$ in random directions) within $\mathcal{B}^p_{\epsilon^*}(\mathbf{x})$ for *realistic* ϵ^* commonly used during AT.

Further comments: We will clarify the power iteration notation, expand the related work and move the Frobenius norm regularization to the additional ninth page for the camera-ready version. Thank you for your high-quality review.

— R2.1): Still not clear how such regularization affects the training of robust classifiers and how it characterizes the sensitivity of the model to adversarial examples. Our Theorem confirms that a network’s sensitivity to adversarial examples is characterized through its spectral properties: it is the dominant singular vector (resp. the maximizer $\mathbf{v}^*$ in Lemma 1) corresponding to the largest singular value (resp. the (p, q)-operator norm) that determines the optimal adversarial perturbation and hence the sensitivity of the model to adversarial examples. The effect of AT and d.d. ONR on the training of robust classifiers is twofold: (i) they dampen the singular values, see Sec. 5.2 and specific. Fig. 2 and (ii) they give rise to models that are significantly more linear around data than normally trained ones, see Sec. 5.4 and specific. Fig. 4 (left) as well as Sec. 5.5 specific. Fig. 5 and Fig. 9. Note also that our results directly explain why input gradient regularization and FGM based AT do not sufficiently protect against iterative adversarial attacks, see Sec. 5.2 and in particular Sec. 7.8 in the Appendix. We will add a summary at the end of the theory section to emphasize these contributions more thoroughly.

R2.2) ℓ_q-norm loss between logits of the clean and perturbed inputs not consistent with practice. It is not uncommon to use ℓ_q-norm losses for adversarial training. See for instance: [1] Harini, Kurakin, and Goodfellow, "Adversarial logit pairing" or [2] Sabour et al., "Adversarial manipulation of deep representations" (both use an ℓ_2-norm loss on the logits / internal representations). We will add additional pointers to clarify this.

Note also that we have a Sec. 7.7 "Cross-Entropy based Adversarial Loss Function" in the Appendix, where we discuss the effect of the loss function on the directional derivative in the power-method like formulation of AT. In particular, the gradient of the loss w.r.t. the logits of the classifier takes the same “prediction - target” form for both the sum-of-squares error as well as the softmax cross-entropy loss, see "A note on canonical link functions". We will move some of this discussion to the main part in the camera-ready version to emphasize this connection.

R2.3) Theorem requires small ϵ. See reply to **R1.1)**.

R2.4) Plot local linearity and activation patterns during training. Explore whether AT / d.d. ONR can help stabilize activation pattern changes against adversarial examples. It is clear from Figs. 5 & 9 that AT and d.d. ONR do improve the stability of activation patterns against adversarial examples, as both d.d. ONR and AT significantly increase the size of the ReLU cells (in both random and adv. directions) compared to the normally trained model. During training, we see a gradual progression towards the end-state. We will add the plots to the camera-ready version.

R2.5) Generalize Theorem 1 to allow for activation pattern changes. Proving such an extension for the “approximate correspondence” between AT and d.d. ONR is *highly non-trivial* and thus out-of-scope of the current paper: one would have to take into account how much “nearby” Jacobians can change based on the crossings of ReLU boundaries, which is complicated by the fact that the impact of such crossings depends heavily on (i) the specific activation pattern at input $\mathbf{x}$, (ii) the precise values of the weights and biases in the network, and (iii) where in the network the units that change their state are. We will add a note to highlight these challenges.

R2.6) What if we only consider the target logit? Then the Jacobian is a vector, will the theoretical result still hold? How will the spectrum look? This can be formulated as a special case of our analysis and our results would still hold (the optimal perturbation would align with the “Jacobian vector” and there would only be one singular value). Note, however, that in general, the effectiveness of adversarial perturbations depends on the relative increase of one logit over the decrease of another.

— R3.1) Paper only proves correspondence for adversarial training with an l_q-norm loss. See reply to **R2.2)**

R3.2) Further datasets are needed to validate the effectiveness of the theoretical analysis. We have confirmed all our experiments on SVHN and TinyImageNet. The results and conclusions hold as expected. We will add the corresponding plots to the camera-ready version.

R3.3) Formatting issues, e.g. in bibliography. We will standardize the formatting. Thank you for the pointer.