We thank the reviewers for their comments and time.

Reviewer #1. *(Relationship to prior work, particularly (Lyu and Li, 2019).)* We will expand our Related Work section to explicitly separate our technical contribution from prior work, expanding comments found throughout this response. As the reviewer mentions, prior work often *assumes* directional convergence and alignment, but neither indicates a possible proof, nor even provides conclusive evidence. Regarding (Lyu and Li, 2019), they sidestepped directional convergence by using subsequences (cf. their Theorem 4.4), and in fact they provided a pathological example where directional convergence *fails* (cf. their Theorem J.1). Our key technical tool for directional convergence, the notion of o-minimal definability, was only used in (Lyu and Li, 2019) to ensure a nonsmooth chain rule. Alignment, in our exact form, does not appear in prior work, which excites us as it implies both existing and new margin maximization results.

(Overly technical presentation.) We agree, with hindsight, that our presentation was encumbered with too much focus on technical details, such as how to handle nonsmooth models as mentioned by the reviewer. We will expand intuition and machine learning connections, and move material to the appendices, as the reviewer suggests.

(Minor comments.) Thank you! We will address these comments in our revisions.

Reviewer #2. *(Discrete-time analysis.)* We agree that a discrete-time analysis is essential. We touched upon this in our "Concluding Remarks", but will expand the material; for instance, one can easily adapt our analysis to handle extremely small step sizes, but handling a practical choice is much more challenging. Getting convergence rates is another important open problem, and might be hard even for the gradient flow, despite our present work.

(Non-homogeneous networks.) We agree this is important, and provide illustrative support, in the form of Figure 2b on page 3 (DenseNet) and Figure 3 in the appendix (ResNet), which does seem to suggest directional convergence holds.

("One might argue that the directional convergence part is relatively easier".) We will highlight in our revised Related Work section that this question is tricky, and has stymied many mathematicians. As discussed at the end of Section 1.1 in our submission, to prove the related gradient conjecture of René Thom, mathematicians had to develop the whole area of o-minimal structures. Even with this powerful technique, as far as we know the existing results on directional convergence all roughly assume piecewise polynomials or real-analytic functions, and therefore cannot analyze a nonsmooth function composed with the exp/logistic loss, which is more relevant to the deep learning community.

("Sub-differential and continuous dynamics are a big field".) We surveyed and cited recent work, e.g., (Davis et al., 2020); we had to prove many new results (e.g., the unbounded nonsmooth Kurdyka-Łojasiewicz inequalities) to show directional convergence and alignment.

Reviewer #3. *(Generalization.)* We agree that generalization is essential, and will include expanded discussion in our revisions. Briefly, on the empirical side, we point the reviewer to the large-scale experiment we cited, by Shallue et al. (2018), which finds that even uncommonly large amounts of training do not seem to hurt generalization. On the theoretical side, we will mention various margin-based generalization bounds, and tie them to our alignment property.

(Relationship to (Lyu and Li, 2019).) Our analysis of directional convergence is not relevant to this prior work: they did not prove directional convergence but instead must use subsequences. Please refer to lines 2-9 above for more details.

(Non-homogeneity and discrete time.) We agree these are important; please refer to lines 14-19 above for more details.

Reviewer #5. *(Regarding "(1)".)* Our analysis can be extended to many other decreasing losses with an exp tail, but we chose to focus on the exp/logistic loss to highlight the key ideas, and moreover because the logistic loss is one of the most widely-used losses in machine learning. Since our core technical work is on solutions at infinity, losses like the squared loss, which imply a (finite) minimizer, require a different analysis, though some of our lemmas will still apply.

(Regarding "(2)".) This initialization assumption was first introduced in prior work (Lyu and Li, 2019), and allows us to focus on the late-training regime; we can handle random initialization by first invoking a standard overparameterized analysis as a lemma (these results only hold near random initialization). Regarding "it seems not hard to show the initialization is close to an optimal classifier", firstly we stress this is an orthogonal concern to our directional convergence result, which ensures gradient flow eventually stabilizes. Secondly, "optimal" in our setting is typically "maximum margin", where it seems the late-training phase is essential, and our results handle a few such cases.

(Regarding "(3)".) The prior work (Ji and Telgarsky, 2018a) only considered deep linear networks, and the proofs there share almost no techniques with the proofs of our main results, in Theorems 3.1 and 4.1. If this was a typo and the reviewer intended (Ji and Telgarsky, 2019), then our potential function $\mathcal{J}$ in Section 4 is indeed based on their dual potential, but their setting is linear and convex, and our nonlinear nonconvex analysis is significantly different. Regarding the relationship to math literature, such as (Kurdyka 2000a, 2006), we had to overcome many obstacles missing from these prior works, such as how to handle the exp/logistic loss with nonsmooth models.