We thank reviewers for their time and effort!

Reviewer 1

Miscellaneous *(∗)* Thank you for the positive feedback! *(∗)* If one is willing to assume convergence of the unobserved entry, then our analysis admits a finite version. Namely, by Theorem 2 (*cf.* Subsection 3.4 and Appendix D), if the bottom right observation in the basic setting is increased from 0 to $\epsilon > 0$, then the limit of the unobserved entry is at least ϵ^{-1}. *(∗)* We are not aware of a simple approach for adapting our analysis to symmetric matrix factorization; it is a direction we intend to investigate. *(∗)* We will modify the text to put more emphasis on matrix factorization.

Reviewer 2

Miscellaneous *(∗)* Thank you for the feedback and support! *(∗)* Extending our analysis to the regime of large learning rate is an interesting direction we intend to pursue. *(∗)* Our construction and theory indeed extend to higher dimensions — see Appendix C. *(∗)* In the high-dimensional analogue of Theorem 1 (not included in the paper) constants indeed depend on the dimension. *(∗)* We will modify the text to put more emphasis on matrix factorization.

Reviewer 3

Significance of our contribution The ability of norms to explain implicit regularization in matrix factorization is an important open question in the theory of deep learning (both supporting and opposing conjectures were made, with multiple recent works attempting to address them). Our main contribution — settling this open question (negatively) — does not follow from any prior work, in particular Dauber *et al.* (2020) or Suggala *et al.* (2018). These papers carefully construct highly specialized convex objectives on which gradient descent (or variants thereof) does not implicitly minimize certain norms. By this they refute the prospect of norms being implicitly minimized on **every** convex objective. To our knowledge, very few have endorsed this far-reaching prospect. Indeed, implicit regularization is conventionally viewed as stemming from a combination of optimizer and model (objective), not an optimizer on its own.

Relation to prior work Prior work dealing with implicit regularization in matrix factorization and various other models is surveyed in Appendix B. We will add an account for "model-free" analyses such as the aforementioned.

Miscellaneous *(∗)* Intuitively, what drives the unobserved entry towards infinity is the persistent positivity of the determinant — see proof sketch of Theorem 1. *(∗)* Definition of effective rank is provided in Appendix H.

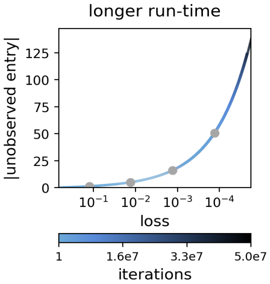
Reviewer 4

Theoretical result for convergence to zero loss There seems to be a misunderstanding — Proposition 4 treats **scaled** identity initialization (α times identity, where $\alpha \in (0, 1]$ is allowed to be arbitrarily small), and thus does not deviate from the regime of near-zero initialization.

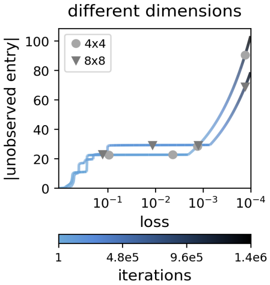
Empirical support for convergence to zero loss In our experiments, the loss value 10^{-4} does **not** represent asymptotic convergence — we merely used it as a stopping criterion for maintaining reasonable run-times. Lower loss values are obtained if one is willing to accommodate longer runs. For example, top plot herein depicts a run with 50M iterations.

Experimentation with additional settings We fear that portions of the paper may have gone unnoticed — beyond the basic setting defined in Subsection 3.1, various additional settings were treated, not only theoretically (Subsection 3.4, Appendixes C and D), but also empirically (Figure 4 in Appendix G). We will add more experimental figures to Appendix G, including ones obtained with different matrix dimensions (see example in bottom plot herein).

Unobserved entry converging to finite value The fact that in some of our experiments the unobserved entry converges to finite value when theory predicts it should diverge to infinity, results from a non-perfect match between: *(i)* the theoretical scheme of gradient flow with balanced initialization; and *(ii)* its practical realization via gradient descent with small learning rate and near-zero initialization. (i) is a standard model for analyzing (ii), and is the subject of the formal conjectures on implicit regularization in matrix factorization. However, despite the fact that lower learning rate and smaller (or more balanced) initialization bring (ii) closer to (i) (as demonstrated in Figure 1), some degree of mismatch will always be present. As reviewer states, even when this leads unobserved entry to converge to finite value, a clear growth is always exhibited, in compliance with our theory, and in contrast to Conjecture 1.



Extension to non-linear models We believe there is a misunderstanding — the whole point of our experiments with tensor factorization was to extend our conclusions beyond linear neural networks. Tensor factorization corresponds to a class of non-linear (polynomial) neural networks for which we can easily define the rank of input-output mappings, and accordingly examine whether that is implicitly minimized by gradient descent. As stated in the paper, in order to apply our conclusions to more conventional non-linear models (*e.g.* ReLU networks), formalizing a notion of rank for their input-output mappings is needed, and that may be a key step towards explaining generalization in deep learning.


different dimensions
● 4x4
▼ 8x8
|unobserved entry|
100
80
60
40
20
0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
loss
1
4.8e5
9.6e5
1.4e6
iterations