We thank the reviewers and are happy that they found the solution simple (R1,R4) and the empirical results promising (R1,R2,R3,R4) in a field of high relevance to the community(R3).

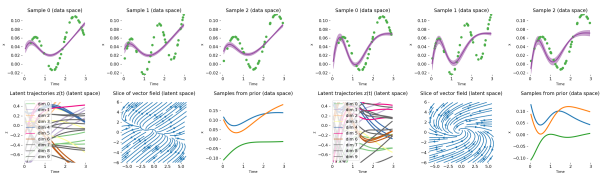
The questions raised by R3 are valid and we thank R3 for raising them since it would aid in making the text better. We thank R3 for pointing us to the references of stiff ODEs in the field of numerical analysis. We would definitely add these as references. The experimental setting for section 4.4 was similar to the one used in the examples/ode_demo.py file of the official github repository of neural ode. Instead of using a 2-D ODE we used a 1-D ODE to be able to better visualize and analyze the behavior. We used Eq 12 in the paper as mentioned in the numerical methods book by Chapra et al on page 752. It was surprising that a Neural ODE was able to model a 2 dimensional model with ease while the stiff equation was extremely hard for the Neural ODE to model. We agree that we have barely scratched the surface in terms of stiff ODEs. We had mentioned in lines 287-291 explaining that it was an empirical observation regarding the failure of neural ODE based models where STEER was able to somehow work better. The text wanted to highlight the problems of stiff ODEs and demonstrate that it could become a real challenge moving forwards as the area of Neural ODE matures. We would be open to removing the section until we understand more about the reason why STEER performs better in those circumstances. As R3 points out the method still would have merits without the stiff ODE section.

The concerns raised regarding the proof is also valid on closer inspection. The proof, though valid does not add much in the discussion of the proposed STEER method since T is only drawn after a theta-parameter update has been conducted. Our rationale was based on the fact that Picard's iteration enables us to prove existence of a unique solution for Neural ODEs and the modified Picard's iteration enables us to show the existence of a unique solution for one with changes in the end time. However, modifications to the Picard's iterates on closer inspection are not the exact same modifications introduced by STEER due to the random sampling involved. We would be open to removing the discussion of the proof from the paper.

As R4 pointed out, the eigenvectors can be complex or real. The ratio of the magnitudes of the real parts of the eigenvalues defines the stiffness ratio. We will mention dataset details as pointed out for Tables 1 and 5. The units of time of table 1 is hours. The sentence can be better phrased as RNNs have also been used for irregularly sampled timeseries models with good results.

R4 and R3 have pointed out the comparison to fixed depth solvers such as resnets. Extensive comparisons between resnets and ODEs with similar number of parameters were made in Neural ODE Chen et al. (2018) and Augmented Neural ODE Dupont et al. (2019). STEER doesn't provide a huge performance improvement in terms of accuracy over these 2 techniques. We would however include the results for a fair comparison. With respect to the optimum values of b as R4 and R3 have pointed out. We have shown some experiments in the supplementary regarding the values of b for optimal performance. The general observation has been that the greater the value of b the better the performance. In case of feedforward models with fixed limits of integration $[0, 1]$ the best performance was obtained when $b = 0.99$. In the case of generative models with continuous normalizing flows, however there were some numerical issues when we tried values of b as high as feedforward models. The best results while also ensuring numerical stability was obtained with $b = 0.5$ with fixed limits of integration $[0, 1]$. In case of irregularly sampled timeseries models, since the points are irregularly sampled, hence the limits of integration change every time. In such a case rather than explicitly tuning b we tune the ϵ parameter as mentioned in line 207 the main text of the paper. Similar to the observations in the preceding cases, the best results are obtained using b as high as possible. In this case it implies ϵ as low as possible without numerical instability. The best results were obtained with $\epsilon = 0.05$.

As R1 has rightly pointed out, the integration times are arbitrary in the case of Neural ODEs. The dynamics of these models rearrange to similar qualitative behaviors even when the integration times are varied by instance. We observe that STEER changes the behavior of these models in qualitatively meaningful ways as shown in Figure 1. The principled work by Finlay et al. (2020) is based on concepts of optimal transport. We have shown via experiments on generative modeling (Table 1) that our technique can be used to augment their technique to obtain even faster convergence times.



Neural ODE STEER

Figure 1: A simple addition of t to $sin(t)$ changes the qualitative performance of Neural ODE vs STEER.

R2 points out the rationale behind the increase in the performance in case of timeseries models and feedforward models. Previous studies Dupont et al. (2019) have corroborated that simpler dynamics leads to better solutions. The temporal regularization is intended to improve the dynamics for neural ODE flows as demonstrated via multiple experiments. It could include unintended improvements in accuracy that we observe in Tables 2,4 and 5 and Figure 1. We intend to explore these aspects in future work.