We thank the reviewers for their insightful comments and suggestions. We respond to the major concerns below and will incorporate all comments in the next revision.

Summary of contributions We presented a novel theoretical and empirical study of the gradient dynamics of overparameterized shallow ReLU networks trained with a least-squares loss. Our results are valid both in the finite and infinite width functional settings. We distinguish two extremal regimes in terms of generalization behavior: "adaptive" and "kernel". The effect of each regime can be quantified in terms of a conserved quantity which depends on the initialization and on the scaling as the number of neurons grows large. In the kernel regime, the training problem converges to a kernel regression over a Sobolev space $\mathcal{H}^{2,2}$ as the number of neurons approaches infinity. Furthermore, in 1D, under mild technical assumptions, the kernel case reduces to cubic spline interpolation. In the mean-field limit, the adaptive regime with regularization converges to regression in $\mathcal{H}^{1,2}$, yielding linear splines with knots at the samples. For a finite number of neurons, our presentation of the adaptive regime is qualitative. We note that dynamics are fully determined by the residuals and the velocity field induced by gradient flow always pushes neurons towards the samples. For finite neurons, we observe solutions which adapt to the input data with knots converging at samples.

Analysis of the adaptive regime (R2, R3) Our results on the adaptive regime reinterpret those appearing in *[Maennel et al.]* and *[Savarese et al.]* in the framework of mean-field analysis. The functional representation in terms of linear splines is established in the limit of infinite width (by combining *[Saverese et al.]* with *[Chizat and Bach NeurIPS'18]*) under appropriate initial conditions and using TV regularisation. Our analysis in the adaptive regime for finite neurons is thus qualititative, but we believe it clarifies the role of initialization and parametrization. Rigorously quantifying the effect of having a finite number of neurons is an important next step, as is the extension to other neural architectures. We highlight however that our main technical contribution in this work is to rigorously establish the implicit bias in the kernel regime in terms of cubic splines, for generic parameter initializations. We therefore provide one of the first instances of an explicit distinction between the "adaptive" and the "kernel" regimes in terms of generalization: formally, we can show that kernel training converges to a kernel regression in $\mathcal{H}^{2,2}$ and, following *[Chizat and Bach NeurIPS'18]*, that adaptive training in the mean-field limit converges to a regularised regression in a Sobolev space $\mathcal{H}^{1,2}$. If the paper is accepted, we will emphasize these technical contributions.

Insights into higher dimensional inputs (R1) The statements and formulation in the paper generalize to higher dimensional inputs, however they do not paint a complete picture of the dynamics in this setting. For higher dimensional full parameters ($\boldsymbol{a} \in \mathbb{R}^{m\times p}, \boldsymbol{b} \in \mathbb{R}^m, \boldsymbol{c} \in \mathbb{R}^m$) representing $f_{\boldsymbol{z}} : \mathbb{R}^p \to \mathbb{R}$, the associated reduced parameters can be viewed as spherical coordinates identifying each neuron with a unit-norm vector $||d(\boldsymbol{\theta}_i)|| = 1$ in $\mathbb{R}^{p+1}$ and a radius $r_i = c_i||(\boldsymbol{a}_i, b_i)||_2$.

In higher dimensions, the samples correspond to hyperplanes in phase space, and the possible configurations of attractors and repulsors become more complex. For example, when reduced neurons lie on one of the attractor hyperplanes, they follow dynamics in the lower dimensional subspace. The difficulty with the analysis in higher dimensions is that it involves the combinatorics of arrangements of hyperplanes corresponding to the sample points. We leave full categorization of these dynamics to future work, however we were able to verify experimentally that the dynamics in higher dimensions are qualitatively very similar to the 1D case, leading to concentration of neurons in the adaptive regime and smooth interpolants in the kernel regime. If the paper is accepted, we will include these experimental results.

Definition of linear splines (R2, R3) We will give a formal definition of adaptive linear splines in the next revision of the paper: A linear spline is a piecewise linear function $\varphi : \mathbb{R} \to \mathbb{R}$ whose knots $e_i \in \mathbb{R}, i = 1 \ldots m$ are the boundaries between pieces. We say that the spline is adaptive if the knots are also variable, i.e., if the function can be written as $\varphi(x', e_1, \ldots, e_m) : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$. Alternatively, we can view adaptive linear splines in the funtional setting as functions $\varphi : \mathbb{R} \to \mathbb{R}$ which interpolate the data points and minimize $\|\varphi\|_{\mathcal{H}^{1,2}} := \int |\varphi''(u)|du$.

Kernel learning for polynomially wide networks (R1) We were not aware of these results for polynomially wide networks, and will cite them in revised version of the paper. These, in fact, seem complementary to our results which demonstrate that for increasing width, the dynamics rapidly approach the kernel regime. In the revision, we will include an experiment demonstrating results for varying finite widths.

Missing citations (R3) The missing citations pointed out by the reviewer are relevant and will be addressed in the next revision. In particular we believe our work is complementary to "A Convergence Theory for Deep Learning via Over-Parameterization" since we can quantify for both a finite and infinite number of neurons how much the dynamics behave like the kernel regime versus the adaptive regime by considering $\boldsymbol{\delta}$ and m in Equation (21).

Presentation of the results (R3) We have prepared a revised version of the paper where our results are presented more rigorously. In particular, we have improved Proposition 4 and formalized our discussion relating the RKHS norm with linearized curvature. In the adaptive setting, in addition to the infinte width analysis, we have clarified the qualitative description of the dynamics with finite neurons and added experiments illustrating the role of attractive samples throughout the training process (in 1D and in higher dimensions).