We thank all the four reviewers for their constructive and positive feedback. Please find our answers to major questions raised. Other points will be dealt with in the revised version.

Limited setting (Reviewers $\#1$ and $\#4$). We acknowledge that our current setting is limited (two-layer neural networks with fixed second layer) and that considering more general architectures is a natural follow up of our work. Preliminary results indicate that our approach can be extended to the full two-layer neural network setting. However, note that the derivation of these results is based on the submitted paper. We think that a manuscript gathering the two settings would be too long and that is why we decided to first consider the simpler case and leave the extensions for future work. In addition, we emphasize that the two-layer neural network setting is common when studying the effect of overparameterization in neural networks. Finally, note that even though this assumption is restrictive, recent works (Suzuki [2020]) have pointed out that ResNets can be rigorously approached by a sum of two-layer neural networks.

Relevance of the parameter α (Reviewer $\#1$). First recall that α corresponds to the decay power of the stepsize. Our motivation to consider decreasing stepsizes comes from the fact that they are commonly used in practice. Besides, another reason follows from a higher order analysis that we briefly explain. Whereas the convergence rates we derive only depend on $\beta \in [0,1]$, if we now turn to a Central Limit Theorem (CLT), as established in Sirignano and Spiliopoulos [2020] for $\beta = \alpha = 0$, preliminary computations suggest that we obtain convergence rates of the form $N^{(1-\beta)/(2-2\alpha)}$. Hence, we expect an interplay between α and β to appear in this weak expansion. The rigorous derivation of such results is left for future work. We will add a discussion on this matter.

Form of the learning rate (Reviewers $\#1$ and $\#2$). In the revised version of our manuscript, we will explain in more details the dependency with respect to β. More precisely we will highlight that even though the learning rate in SGD scales as $\mathcal{O}(N^\beta)$ the learning rate in the mean-field dynamics scales as $\mathcal{O}(N^{\beta-1})$. In addition, we will emphasize that the term $(n + \gamma_{\alpha,\beta}(N)(N)^{-1})^{-\alpha}$ can be replaced by $(n+c)^{-\alpha}$ with $c > 0$ at the cost of modifying the limiting SDE.

Comparison with previous works and difference between the two regimes (Reviewer $\#3$). The formulation we consider in the paper has already been studied in several works to better understand the behaviour of overparameterized neural networks and their optimization using SGD or simply gradient descent. However, the main difference between our work and previous studies is the use of a stepsize depending on the width of the neural network which leads to a different mean-field limit. In contrast to the one obtained previously, this limit has a diffusion term. This illustrates that SGD has a potential regularization effect. In future work, we plan to rigorously investigate this phenomenon by establishing generalization bounds for the two regimes and compare them.

Comparison with other formulations (Reviewer $\#1$). In what follows we try to clarify the following remark of Reviewer 1: *"Depending on how one thinks about it, the learning rate in previous papers on infinite width SGD depends on the number of hidden units."*. Set $a_N = \sum_{k=1}^N F(w^{k,N}, x)$ and consider the two functionals $\mathscr{R}^N(w^{1:N}) = \int_{\mathsf{X}\times\mathsf{Y}} \ell(N^{-1} a_N, y)\mathrm{d}\pi(x,y)$, $\bar{\mathscr{R}}^N(w^{1:N}) = \int_{\mathsf{X}\times\mathsf{Y}} \ell(a_N, y)\mathrm{d}\pi(x,y)$. Let $(W_n^{1:N})_{n\in\mathbb{N}}$ be the SGD scheme associated with the minimization of $\mathscr{R}^N$ and $(\bar{W}_n^{1:N})_{n\in\mathbb{N}}$ the one associated with the minimization of $\bar{\mathscr{R}}^N$, defined by the recursions (1) $W_{n+1}^{k,N} = W_n^{k,N} - \gamma_1 N^{-1} \int_{\mathsf{X}\times\mathsf{Y}} \partial_1 \ell(N^{-1}\sum_{k=1}^N F(W_n^{k,N}, x), y)\nabla F(W_n^{k,N}, x)\mathrm{d}\pi(x,y)$, (2) $\bar{W}_{n+1}^{k,N} = \bar{W}_n^{k,N} - \gamma_2 \int_{\mathsf{X}\times\mathsf{Y}} \partial_1 \ell(\sum_{k=1}^N F(\bar{W}_n^{k,N}, x), y)\nabla F(\bar{W}_n^{k,N}, x)\mathrm{d}\pi(x,y)$. From these definitions, we get that no choice for the stepsizes γ_1 or γ_2 implies that $(W_n^{1:N})_{n\in\mathbb{N}} = (\bar{W}_n^{1:N})_{n\in\mathbb{N}}$ because of the different scalings in the loss function, *i.e.* $\sum_{k=1}^N F(W_n^{k,N}, x)$ is multiplied by $1/N$ in (1) and not in (2). As a result, there is no immediate link between the setting we consider with a stepsize which depends on the number of hidden units and the classical setting for SGD. However, in the specific case where $\partial_1 \ell$ is positively homogeneous, further conclusions can be drawn. This is currently under investigation. We will mention this observation in the paper.

Differentiability of the features (Reviewer $\#2$). In our paper, we assume some high order differentiability conditions for the feature function in order to derive our results. We suspect that our regularity assumptions can be relaxed but at the expense of significant technical complications. That is why we have decided to limit our theoretical study to the smooth case. Nevertheless, in our experimental analysis we used ReLU activation functions which are not differentiable to illustrate that our findings also hold empirically in the non-smooth setting.

Comparison with other quantitative results (Reviewer $\#3$). We agree with reviewer 3 that previous works have established quantitative propagation of chaos results, see Mei et al. [2018, 2019]. However, we found our quantitative results to be of interest since the propagation of chaos results in [Mei et al., 2018, Theorem 3] hold in high probability for different criteria (Fortet-Mourier metric and risk evaluation). In that respect we believe that our results in the case where $\beta = 0$ complement the ones of Mei et al. [2018, 2019] and extend them in the case where $\beta \neq 0$. We will add this remark in the revised version of our paper to better acknowledge the results of Mei et al. [2018, 2019].

Assumptions in the main document (Reviewer $\#4$). We tried to simplify A1 in the main document but the gain of space was negligeable and that is why we think that it is better to keep the general formulation we have for the moment. However, we acknowledge that a discussion on this set of assumptions is in order and we plan to add it in the revision of our manuscript.