We thank the reviewers (**R1**, **R2**, **R3**, **R4**) for their time and expertise. We are grateful to the reviewers for letting us know which definitions and explanations were unclear, we will improve them in the revised version. We also thank them for pointing out the typos, that we will correct in the revised version.

R1, R2: Significance of the work – We underline that our work is methodological in nature. Indeed, the introduction of the finite batch size provides a conceptual step forward the understanding of the dynamics of mini-batch gradient descent with DMFT. Moreover, in contrast to [25], considering structured data allows to investigate supervised learning and questions on generalization properties. The theory we build opens way to many detailed investigations about the nature of the difference between GD and SGD in simple models, and we are certain it will be followed by numerous works (for examples of ongoing follow-up investigations, see below).

R1, R2: Insights from theory and future works implications – (i) DMFT provides a natural tool to characterize the noise introduced by the finite batch size: at variance with [25], in Eq. (15) we have a Gaussian noise but also a stochastic variable $s(t)$ which encodes the batch size noise. Recent literature (*arXiv:1912.00018*) investigated the behavior of SGD gradient as a function of time and measured the corresponding statistical properties. Due to the dimensional reduction performed by DMFT, comparing these experimental results with simple models by sampling the effective stochastic process $h(t)$ is much easier than by direct simulations. In addition, we have direct access to two-point correlation functions of the stochastic gradient which we expect to be present and important. Those are numerically much more demanding to get in simulations than say the test error. The DMFT approach can instead be used to compute the full time dependent correlations of the SGD gradient. Furthermore, from the response and correlation functions we can extract an effective temperature (*arXiv:cond-mat/9611044*) which is a direct and quantitative measure of the noise that governs the dynamics. In contrast, computing two-point response functions from numerical simulations is much more challenging. **(ii)** Optimal stopping time: we are working on deriving an analytic formula for it as a function of the model parameters.

R1: Extensions to more realistic data – We are working on generalizing the DMFT analysis of SGD to: **(i)** Models of structured data with low intrinsic dimension embedded in large dimension, such as the Hidden Manifold Model [9], or the Random Features Model (*arXiv:1908.05355*); **(ii)** Generalised Linear Models. DMFT can be used to study both the recovery transition of Gradient Flow as well as how it changes when the finite batch size is employed.

R1: Clarity of $R(t,t')$ – Since we focus on the linear response regime, we couple an infinitesimal local field $H_i(t)$ to each variable w_i, changing the loss as follows: $\mathcal{H}(\mathbf{w}) \to \mathcal{H}(\mathbf{w}) - \sum_{i=1}^{d} H_i \mathrm{w}_i$, and hence adding and extra term $H_i(t)$ to the right hand side of Eq. (10). We will add a more detailed explanation in the revised version.

R1: Other mean-field methods (NTK, mean field theory of batch normalization) – We will extend the discussion about mean-field analysis of infinitely wide networks in the introduction.

R2, R4: Learning rate – In our approach the learning rate is - strictly speaking - zero (we are in the stochastic gradient flow regime) and enters only in the discretization of the DMFT equations for the numerical solutions.

R2: "DMFT equations (20) have some local optima" – This does not happen. The DMFT fixed point equations are deterministic, hence at given initial condition the solution is unique. This can be seen as follows. The kernels computed by DMFT are causal and a simple integration scheme of the equations is just extending those kernels progressively in time starting just from their initial value which is completely deterministic given by the initial condition for the stochastic process. Furthermore, we check this independently by simulations and we find a very good agreement with the theoretical results even when the problem is non-convex. We will add this discussion.

R2, R4: Optimal batch size – We thank for this pointer, our methodology is indeed well fitted to address this question. An important parameter in our setting is the persistence time, in Fig. 3 right we see that the smaller the persistence time the better the early stopping error. We will investigate whether in other cases an optimum can be found.

R3: The batch size is $O(n)$ and its noise averages out– Even though we consider extensive batch size, the randomness due to that does not disappear in the large n limit. Indeed in Eq. (15) we have a Gaussian noise but also a stochastic variable $s(t)$ encoding the batch size noise. Therefore even if the batch size is extensive, the resulting noise does not average out. This is apparent e.g. in Fig. 1 left where the early stopping error depends clearly on the batch size.

R3, R4: Extensions to two-layer architectures – We are working on generalizing the DMFT to a committee machine with finite number (K) of hidden units (see for instance *arXiv:1806.05451*). In this case, instead of one effective stochastic process for the typical gap as in Eq. (15), we will have (K) coupled such processes.

R4: Non-standard nonlinearity – In relation to the point above, we can consider a committee machine, and the hidden units together with the nonlinear activation will indeed introduce non-convexity. However, the fact that our analysis allows to consider generic non-convex loss functions is interesting per se, as it is not the case for other methods in existing literature.

R4: Non monotonicity of generalization minimum in experiments – In Figure 3, the exact location of the generalization minimum at finite dimension is not precise due to large fluctuations (see the errorbars).

R4: Persistent-SGD vs SGD – We compare the two algorithms in Figure 4.

R4: Solving DMFT numerically – We will add more details on that in the revised version.