Dear Reviewers, we would like to take this opportunity to thank you for your precise and constructive feedback. Below the paragraphs in italics are extracts from the reviews. Citations refer to the bibliography of the paper.

Reviewer #1. *[...] the paper mentions "some basic vision or sound recognition tasks" (line 33) - I'd like to ask some examples of such tasks.*

Authors: Our noiseless assumption approximately holds for supervised learning tasks with little ambiguity of the output given the input - but the rule giving the output given the input can be complex. An example from [18, Section 6] is the classification of images of cats versus dogs. For typical images, the output is unambiguous; humans indeed achieve a near-zero error. In sound recognition, one could think of the recovery of the melody from a tune, an unambiguous (but tremendously complex!) task. We will add these examples to the final version. Note that in Appendix D, we generalize our results to the case where some ambiguity (i.e., some additive noise) is present.

R1: *Looking at the statement of the Theorem 1, seems that it should be applicable in finite-dimensional spaces with invertible covariance matrices. [...] In particular, for X distributed with a finite support and has identity covariance matrix, the conditions (a) and (b) hold for arbitrarily large positive α, however the theorem statement implies that the estimates will go to zero at an arbitrarily large polynomial rate, which is not true.*

A.: Theorem 1 does apply in finite-dimensional spaces. In the example described by the reviewer, SGD converges exponentially; this is a surprising effect of the noiseless model. Indeed,

$$\mathbb{E}[\|\theta_n - \theta_*\|^2] = \mathbb{E}[\|(I - \gamma X_n \otimes X_n)(\theta_{n-1} - \theta_*)\|^2]$$
$$\leq \mathbb{E}[\|\theta_{n-1} - \theta_*\|^2] - 2\gamma\mathbb{E}[\langle \theta_{n-1} - \theta_*, (X_n \otimes X_n)(\theta_{n-1} - \theta_*)\rangle] + \gamma^2 R_0 \mathbb{E}[\langle \theta_{n-1} - \theta_*, (X_n \otimes X_n)(\theta_{n-1} - \theta_*)\rangle] .$$

As $\gamma R_0 \leq 1$ and by assumption of the reviewer, $\mathbb{E}[X_n \otimes X_n] = I$, we obtain $\mathbb{E}[\|\theta_n - \theta_*\|^2] \leq (1-\gamma)\mathbb{E}[\|\theta_{n-1} - \theta_*\|^2]$. Note that in finite dimensional spaces, the non-asymptotic polynomial bounds of Theorem 1 can be better than the exponential rates, for small number of iterations n. This is detailed in the paper for the gossip process (lines 288-298). We will add this remark on the application to finite-dimensional spaces to the final version.

R1: *The paper does not give any theoretical argument to the 'tightness' of the proposed bounds, [...].*

A.: We prove both upper bounds (Theorems 1 & 3) and lower bounds (Theorems 2 & 4) on the performance of SGD that almost match: they have the same asymptotic in n. Thus the bounds describe the actual behavior of SGD, and this is confirmed by simulations: **the bounds are "tight" in this sense**. Note that we do not mean "optimal" here.

R1: *I would suggest to add the results with additive noise assumption for infinite-dimensional spaces, to put the proposed model into a perspective.*

A.: We will give the non-parametric optimal rates with additive noise from Caponnetto & De Vito [9], reached by ridge regression. The perspective with our work is that these rates are slower than n^{-1}, while we prove rates faster than n^{-1} because of our noiseless assumption. This is stated lines 65-68.

A.: We also thank **R1** for pointing out a typo and a notation without definition.

Reviewer #2. *I think the claim about the Sobolev smoothness is overstatement because it depends on a somewhat strong condition (line 215-216) which strongly restricts the class of kernels.*

A.: We respectfully disagree. It is true that this condition does not cover C^∞ kernels, including the Gaussian kernel. However, **this condition is relevant for less regular kernels**, that have a power decay in Fourier. Line 215-216 defines the rate of decay in Fourier. In theory, one could imagine that the lower and upper bounds hold for different s, in which case one could have a theory by adapting lines 215-232. However, the point of the section is only to illustrate our theory, and for all "less regular" kernels that we know and cite, the condition 215-216 holds, so we kept things simple. We will add this discussion to the final version.

R2: *How does the averaging technique work in the noiseless setting?*

A.: Averaging does not seem to accelerate the averaging process (Section 3.2). Extrapolating to all SGDs, we expect that averaging is useful only for reducing additive noise, and thus would not accelerate in the noiseless setting. However, this question deserves a rigorous study that we wish to conduct in future work.

Reviewer #2 and Reviewer #4 both asked for a deeper comparison with [18]. [18] indeed analyses the zero (or low) noise setting, and allow for the optimal function to lie in the kernel space. However, they do not exploit when the function is more regular than being in the kernel space, i.e., when $\alpha_1 > 0$ with our notation, $\beta > 1/2$ with theirs. In fact, they leave this case as an open problem in their Section 6. Thus, a fair comparison can only be made when $\alpha_1 = 0$, $\beta = 1/2$. In this case, SGD and [18] both achieve the same rate $O(n^{-1})$. We will add this discussion to lines 71-73.

Reviewer #4. *Line 35: Citation for when it is called multiplicative noise would be great.* **A.:** We will cite [13]. Thanks.

Reviewer #5 did not express any concern; we only thank her/him for the encouraging review.