We thank all the reviewers for their valuable feedback. We address their individual concerns below.

Reviewer 1

We are grateful for the very positive comments on our work and for the outlined valuable directions for future research. We totally agree on the need to establish a more rigorous connection between the continuous- and discrete-time perspectives in optimization — this is exactly the focus of our current research: we see this paper as our first step in this direction. We will include a discussion of the references you indicate in your review, including the mean field limit work you mentioned which is related to our interest for future research directions.

Reviewer 2

Models are easy extensions of existing ones As correctly stated by the reviewer (and actually also by us in the paper), SDE approximations of SGD (with fixed learning rate) were already presented in previous literature (see e.g. line 151). In this paper we extend these models to include the effect of decaying learning rates, increasing batch-sizes and variance reduction. We agree that the construction of such models, to the eyes of an expert reader, might not be technically very challenging — but this is precisely why *we dedicate only one page to this part* before switching our focus to the convergence analysis. Moreover, we would like to point out that

1. studying such extensions is of chief importance in machine learning, since modern SGD algorithms (see e.g. [41], [7], [29]) rely on decaying learning rates, increasing batch-sizes and variance reduction;
2. the construction of the model is *just one of the 5 contributions* listed in our paper. As stated in the introduction, our main goal is to show how such model can guide the analysis of commonly used stochastic methods.

Possible extension to continuous-time models for momentum-based accelerated method Some continuous-time stochastic momentum SDEs have been studied in [32] — a truly beautiful work from which we took a lot of inspiration. However, this work is of a different nature: it does not focus on providing a correspondence between continuous models and discrete algorithms; but instead analyses a very general yet interpretable SDE in the *convex* setting. That said, we agree on the need to extend our methodology to stochastic momentum methods: we have in fact already derived similar results, but decided not to include it in our submission to focus more on our other contributions. Indeed, as noted by R1, the paper is already quite dense and the study of momentum methods *deserves to be explored in a separate work.*

Reviewer 3

We apologize for the typos/wrong reference numbers, and we thank the reviewer for pointing them out: [36] at line 101 should be [40], [37] at line 127 should be [36].

Assumption (Hσ) restrictive (comment 2) We would first like to point out that such assumption is commonly used in the continuous-time literature (see e.g. (H_3) in [40] and Eq.(8) in [32,NeurIPS Proceedings]). Nevertheless, we thank the reviewer for the constructive comment and *will try to update our proofs* using the suggested localization argument.

Generality of landscape stretching result (comment 4) We thank the reviewer for the interesting comment. We believe that the landscape stretching phenomenon is actually quite general and would also hold e.g. asymptotically under strong convexity[1]: indeed it is well known that, by Taylor's theorem, in the neighborhood of the solution to a strongly convex problem the cost *behaves as its quadratic approximation*. In dynamical systems, this linearization argument can be made very precise and goes under the name of Hartman-Grobman theorem. Since the process we study is memoryless (no momentum), at some point it will necessarily enter a neighborhood of the solution where the dynamics is described by the landscape stretching result. We will add a comment on this in the updated version of this paper. We are also thankful for the comment about multiplicative noise in high dimensions; yet, at least in the simple case we considered, the volatility is not actually a function of the state. We will nevertheless add a short note on this.

Weak approximation of SGD (comment 1 and 3) *We clearly stated at line 129* that [26] and [36] do not assume Gaussianity of Z_k. We are very fund of these works (including [B]), yet we find that practical implications are limited since the approximation bound explodes (as the standard Euler global error does) as a function of the final time point. That is why we explicitly said (line 135) that *our approach in this paper is different*: we explore the connection between continuous and discrete exclusively by providing matching rates (and the algebraic equivalence in Sec 3.2 and App. A.2.). We are very sorry for the confusion, but we felt this point was clear given that it is also stressed many times in Sec. 3.2 as well as in the list of contributions (line 59) and in the abstract (line 6). Nonetheless, we understand that this is a crucial point and *we will do our best to update line 127-137* to make sure there is absolutely no confusion in where our contribution lies with respect to prior work. We also thank the reviewer for pointing out [C] to us (very interesting work), which we will also include in the discussion.

In summary we would benefit from the additional ninth page to extend and update the discussion as outlined above. In particular, to improve transparency, we would *transfer details from App. D* to the main paper as the reviewer suggests.

[1] and also in the neighborhood of any hyperbolic fixed point, with implications about saddle point evasion.