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

Reviewer 1

Novelty We agree there is a lot of prior classical work on numerical analysis ODEs and all references that are relevant to us are cited in the paper. Reviewer 3 agrees on the importance of introducing these ideas to the machine learning community. Besides, our contribution is in the *application* to optimization: in this regard, the results derived in Thm. 3,4 and C2 are novel. As suggested by the reviewer, we will revise the text to emphasize where our contribution lies.

Context not treated appropriately While designing this paper we spent several months researching into the dynamical systems and numerical analysis literature. After speaking directly with leading researchers in these fields, we condensed our knowledge in a readable yet exhaustive overview of the theory of shadowing and error analysis (Sec. 2). The same ideas are then presented again in a self-contained yet less abstract way in the main sections, directly applied to optimization methods. Given the amount of effort from our side in trying to collect the most relevant ideas from the literature to make them easily accessible to the community (most papers we cite work in more abstract settings, relying on manifolds or using a somewhat outdated notation), *we would kindly ask the reviewer if he/she could be more precise in his/her claim that we do not "treat the context appropriately", ideally by giving concrete examples*. To further convince the reviewer that our discussion and understanding of the literature is complete, we invite him/her to check [5]: a seminal paper in numerical analysis which shows how central notions in the analysis of Euler's method cannot be applied to get useful global bounds (their Sec.1) hence one has to rely on perturbed hyperbolic splittings (their Sec.3).

The ODE method On the same line, we also remind the reviewer that the "ODE method" results only holds asymptotically and under stepsizes decreasing to zero. Instead, crucially, we consider fixed stepsizes and our bounds hold from the very first iteration: this a completely different setting which *cannot* be captured by the ODE method.

"it is not clear how to use these results other than in the negative" Please note that Thm. 3 is tight: the result precisely describes the behavior of GD and we therefore believe this result is valuable for the community, even if that could appear as a "negative" result (i.e. *not possible to improve*). We will add a comment on this, see note below.

Tightness of formula for ϵ: first we note that the bound for δ in Prop. 3.1. cannot be improved; indeed it coincides with the well-known *local* truncation error of Euler's method. Next, pick $f(x) = \frac{1}{2}x^2$, $x_0 = 1$ and $h = \frac{1}{L} = 1$. For $k \in \mathbb{N}$, gradients are smaller than 1 for both GD-ODE and GD, hence $\ell = L = \mu = 1$. Our formula for the *global* shadowing radius gives $\epsilon = \frac{hL\ell}{2\mu} = 0.5$, equal to the local error $\delta = \ell L h^2/2$ — i.e. as tight the well-established local result. In fact, GD jumps to 0 in one iteration, while $y(t) = e^{-t}$; hence $y(1) - x_1 = 1/e \approx 0.37 < 0.5$. For smaller steps like $h = 0.1 < \frac{1}{L}$, our formula predicts $\epsilon = 0.05 = 10\delta$. In simulation, we have maximum deviation at $k = 10$ and is $\approx 0.02 = 4\delta$— which is only 2.5 times smaller than our prediction. $\square$

Reviewer 2

h too small Actually, our results hold for *any stepsize* $h \leq 1/L$: as can be seen from Thm. 3, each choice of h will determine a shadowing radius ϵ such that $h = 2\mu\epsilon/(L\ell)$. We agree that ϵ can be large in some cases, but this bound in nonetheless tight (see answer to R1) and it's not possible to improve it. Indeed, *it captures a fundamental property of the ODE approximation* — which is what we are after. We will update our conclusion section making a remark on this.

Thm.4 only holds for quadratics We kindly point the reviewer to line 234 and App C4, where the mentioned result is *generalized to non-quadratic saddles*. The extension to more complex landscapes is discussed qualitatively (see also reply to R3) at line 238 and tested in the experimental section. We will emphasize these results in the revised version.

Lipschitzness We addressed this in footnote 6: "[...] we can pick $\ell = L\|x_0 - x^*\|$ even tho quadratics are not Lip."

Reviewer 3

More generality In full generality, the theory we present can be applied to a cost which, restricted to a (fixed) subspace, is strongly convex and, restricted to the remaining directions, is strongly concave. This cost can be then also perturbed as in Thm. C2. After an extensive literature search, we feel confident in claiming that the described setting is aligned with what the classical literature on long time numerical error approximation (i.e. shadowing) can tackle. To convince the reader of this fact, we invite him/her to read the introduction of [5]: a seminal paper in numerical analysis which studies shadowing (i.e. long time approximation) *near hyperbolic saddles* (i.e. when there exists an approximate hyperbolic splitting of $\mathbb{R}^d$). The tools in our paper, specifically the proof of Thm. C.2., are inspired by this work. We will add a comment making clear what we said in the first sentence of this paragraph: i.e. *our results can be applied outside the quadratic regime* (see also gluing below) and are in line with what is known in numerical analysis.

Gluing We thank R3 for the interest in this idea, which we indeed only sketched and would like to discuss quantitatively in the potential additional page. In short, the outlined simple gluing procedure is successful *if the number of unstable directions is non-increasing*. In numerical analysis, this was explored in [9] (see 2nd paragraph of their intro). We will add this result in the form of a theorem and make clear what objectives can be studied with this approach.