We thank all the reviewers for their time and effort to evaluate our paper. We appreciate that the reviewers find our paper to be original, theoretically deep, well-organized, clearly written, complete, and our contributions to have a big potential to be influential. We believe that we addressed all the raised issues. The detailed responses are given below.

R1. We thank the reviewer for the positive and insightful comments. As suggested by the reviewer, we will fix all the minor issues: **(1)** We will replace $\bar{\varepsilon}$ with ξ to prevent confusion. **(2)** We will define the concept of metastability in more detail and provide more references. **(3)** To increase clarity, we will make A6 more concise by representing the constants with Big-O notation and providing their explicit definitions in the supplementary document. **(4)** We agree that our most important contribution is the fact that Theorem 2 enables the use of the metastability results for Lévy-driven SDEs for their discretized counterpart. We will highlight this fact more clearly in a remark. Thank you for this suggestion. **(5)** In our paper, we mainly aim at providing more understanding on the connections between SGD and wide minima. In the current literature, the implication of better generalization by wide minima is still in an hypothesis phase and is a very active research field on its own. Nevertheless, several empirical results have conformed with this phenomenon. We will add a new paragraph on the connection between our results and generalization by discussing these points.

R3. We are grateful to the reviewer for the positive and encouraging comments. We will fix the typo as requested.

R4. We thank the reviewer for the detailed comments. We suspect that a simple misunderstanding about the scope and the contributions of our paper might have influenced the opinions of the reviewer. We will now clarify this misunderstanding and we hope this would help the reviewer to reconsider their score.

As acknowledged by Reviewers 1 and 3, modeling SGD as an SDE driven by Lévy motion has been already proposed in a recent paper, Simsekli et al., ICML 2019 [6]. In that paper, the α-stable noise assumption was indeed validated empirically in various deep learning settings. Therefore, the empirical validation of the assumption has been in a way validated in the community.

A clear limitation of [6] (also mentioned in their paper) is that the authors did not develop new theory and used existing metastability properties of the **continuous-time** Lévy-driven SDEs (which we summarized in Section 2) as a proxy for the **discrete-time** dynamics of SGD. Approximating SGD as a continuous-time SDE has already raised several theoretical questions (as we mentioned in the beginning of page 3), since the behaviors of these two systems might be significantly different. In our paper, we provided explicit conditions (Theorem 2) such that *the discretized process inherits the metastability properties of its continuous-time limit*. Hence, our main contribution is to establish this technically challenging theoretical result, which required us to first build a more general result about Lévy-driven SDEs (Theorem 3) and use this result in a non-trivial way to relate the exit-times of the two processes (Theorem 2).

Quality-Clarity: We thank the reviewer for this insightful question. First, we would like to clarify that *our results only hold for connected neighborhoods of a local minimum*. This can be verified by checking the assumption in Theorem 1 which explicitly requires the interval to be a neighborhood around a local minimum. Since we are relating the exit time of the discretized process to the conclusions of Theorem 1, we automatically inherit this condition (to see this clearly, Eqs 8-10 are explicitly defined using a neighborhood of local minimum $\bar{w}$). For exit times in d-dimensions, we still require the set of interest to be a neighborhood of a local minimum (see A1-5 in [18]). Now, for simplicity assume that we are in $\mathbb{R}$, and consider two local minima and define two intervals such that these intervals contain exactly the basins around the local minima (similar to Figure 1 right in the paper). Then, if the exit time from Basin 1 is longer than Basin 2, it immediately implies that Basin 1 has a larger diameter. If we combine this fact with A3-5 (or the assumption in Theorem 1), which make sure that the function behaves globally regularly (gradient Hölder and dissipative), we can directly deduce that Basin 1 will be more flat. The same argumentation can be made for $\mathbb{R}^d$ with the careful construction of [18]. We agree that the connection with this notion and the other notions of flatness (e.g. spectrum of the Hessian) is not immediate, yet we believe that there is an explicit link and it definitely deserves further investigation. We also agree that this is a subtle point and we will clarify it by stating it explicitly.

Improvements: **(1)** We will define metastability in more detail as suggested. **(2)** The Brownian systems need exponential time in the height of the basin (line 143, see also [17] Sec 3.1). We will explicitly define the exit times as suggested. **(3)** We underline that our result is already for $\mathbb{R}^d$ with an explicit dependency on d. The multi-dimensional version of Theorem 1 is available in [18] and makes a non-trivial connection between exit-times and the dimension. However, that theorem would require us to introduce several technical constructs, which we could not simply accommodate in our paper due to space limitations. We will summarize [18] in the supp. doc. for completeness. **(4-5)** That is correct, as in [16], x is the initial point of the process, taken uniformly in the interval $[-a, a]$. In the multi-dimensional setting, the initial point $W(0)$ is in the neighbourhood of radius a, centered at the local minimum $\bar{w}$. We will clarify these notations, thank you for pointing out. **(6)** The composite noise exhibits the same metastability behavior as pure α-stable noise, is more general, and is mathematically more convenient for our analysis (see line 232). **(7)** In general, the law of the processes (6) and (7) are not the same (for any $\eta > 0$). However, one can show that (7) converges (in law) to (6) when η goes to zero (see [39]).