We are grateful to the reviewers for their time and comments. For the reviewers' convenience, we briefly state below the novel contributions of our work, as summarized by R4 (whom we thank for the expert summary):

> "*This paper shows that under mild conditions, SGD converges to a critical point of general non-convex functions and avoids all strict saddle points, with probability* 1*. It also presents a convergence rate analysis of SGD once it enters the neighborhood of a local minimum.*"

In what follows, we address the reviewers' comments in order, tagging the reviewers concerned in each as **#RX**.

#R1: Relation to Lee et al [17,18]. As we explain in Lines 66-68 of the introduction, **[17,18] do not study stochastic gradient descent,** but gradient descent with full, *perfect* gradients – i.e., a ***deterministic*** algorithm. Specifically, [17,18] show that **deterministic** gradient descent avoids strict saddles from almost every initial condition. The reviewer is therefore not correct in interpreting this statement as an "in probability" result for SGD: *the results of [17,18] provide* ***no*** *guarantees for SGD, from* ***any*** *initial condition.*

Additional comments: The stochasticity in SGD makes for a drastically different, much more difficult setting. In the full gradient case, there is a well-defined drift that drives GD away from saddle points. This persistent push is no longer present in SGD: this is a crucial difference which we feel may be at the source of this misunderstanding.

#R1#R2: Relation to Pemantle [27]. Pemantle's work only applies to **isolated, linearly unstable saddle points,** it does not cover saddle points with a **non-trivial center manifold.** In the deterministic case, strict saddles can indeed be excluded thanks to the existence of local diffeomorphism results based on the center manifold theorem. However, in the stochastic case, the presence of a non-trivial center manifold requires fundamentally different techniques from differential geometry, as we explain in detail in Appendices D.2 and D.3. The reason for this is that there is no longer a persistent drift away from the center stable manifold (in technical terms, there is no "shadowing"). **This major difficulty is not present in Pemantle's work** (which, again, cannot deal with non-trivial center manifolds); the only relation with [27] is two technical lemmas on random numerical sequences (Lemmas D.1 and D.2).

Additional comments: The reviewers may have thought that we are making a significantly more restrictive Morse-Smale assumption for the problem's objective – we emphasize here that *this is not the case.*

#R1#R2: On the rates of Jin et al [14]. First, as can be seen from (E.3) and (E.42), Thm. 4 gives the precise bound

$$\mathbb{E}[f(X_n) - f(x^*) \,|\, X_1 \in \mathcal{U}_1] \leq \tfrac{2}{\beta}\tfrac{2\gamma^2}{1-\delta}\tfrac{G^2+\sigma^2}{2\alpha\gamma-1}\tfrac{1}{n} + o\big(\tfrac{1}{n}\big). \tag{1}$$

We will put this expression for $p = 1$ in the main text. Beyond this, there are two key factual misunderstandings:

1. **The statements for SGD in [14] and related papers are also asymptotic** because they involve an unknown, probabilistic constant hidden in the $\mathcal{O}(\cdot)$ notation; see Theorem 3, Corollary 4 and Theorem 5 in [14], as well as the corresponding statements in the papers mentioned by R1.
2. The asymptotic value convergence rate of [14] and related papers is $\mathcal{O}(1/\sqrt{n})$; by contrast, **the value convergence guarantee that we provide is $\boldsymbol{\mathcal{O}(1/n)}$.** The reviewers are therefore incorrect in stating that our rates are similar to those of [14] and related works.

#R1: From high probability to probability 1 via Borel-Cantelli. This is not possible for (at least) two reasons:

1. The target probability threshold ζ of Ge et al. is hard-coded in the algorithm's step-size. Therefore, getting results for different probability thresholds (in order to apply Borel-Cantelli) would necessitate running different algorithms, destroying in this way the validity of the results of Ge et al.
2. Even if this vital obstacle were to be somehow overcome, the logarithmic dependence of the step-size of Ge et al. on ζ implies that the induced step-size policy would have to vanish at an exponential rate in order to apply Borel-Cantelli. However, it is well known from standard results in stochastic approximation that SGD with summable step-size policies *does not converge* (Kushner and Yin, 1997, Chap. 4), so this approach would fail.

#R2: On Bonnabel (2013). We thank R2 for bringing this paper to our attention, we will definitely discuss it! At the same time, we should point out that **Bonnabel's paper makes the explicit assumption that SGD remains in a compact set** (cf. Theorems 1 and 2). Boundedness assumptions of this kind are prevalent in the literature, and **this is precisely one of the key gaps that our paper closes:** convergence of SGD ***without*** implicit, unverifiable boundedness assumptions. This was the main weakness identified by R2, so we hope that the above clarifies the merits of our work.

#R3: On the rates of escape. Deriving rates of escape that hold with probability 1 is a whole new paper in itself.

#R3: On the size of $\mathcal{U}_1$. The size of $\mathcal{U}_1$ only depends on the landscape of f around x^*, *not* δ; see (E.17) and (E.18).

#R3: On Dashamand et al. Dashamand et al. refine the analysis of Ge et al. and provide positive probability results for second-order stationary points. There is no overlap with our techniques or results; we will cite it to make this clear.

#R4: On the hitting time to $\mathcal{U}_1$. This is a very difficult global-to-local estimate. To the best of our knowledge, no one has succeeded in making progress on similar questions in general non-convex settings, so we do not address this here.

#R4: On the dependence on d. Great question! The rate does not explicitly depend on d, see (1) above.