We thank all the reviewers for their valuable comments. We first highlight our technical contributions.

Main technical contributions: We emphasize that our main contribution is the first result showing that PSGD and SCRN escape saddle-points and converge to local minimizers faster under Strong Growth Condition (SGC) (which is a consequence of interpolation phenomenon). Prior works (e.g., [VBS18]) considered only convergence to critical points under SGC in the first-order setting. We provide our results in both the zeroth and higher order settings. In the zeroth-order setting, we prove a novel concentration inequality for the zeroth-order gradient estimator which is non-trivial and was not know before. We emphasize that this concentration result does not assume the function is bounded; see also Remark 3. For completeness, we also analyzed the complexity for zeroth-order PSGD without the SGC assumption for unbounded functions, which was not done before in the literature. Furthermore, the analysis of SCRN is also significantly involved under SGC (especially in zeroth-order setup); see also Remark 6 and 7.

Rev #1: NTK regime: NTK viewpoint provides an alternative explanation on efficiency of optimizing algorithms for DNN training. However, a majority of the results based on NTK approach are for polynomially (in depth and sample-size) large-width networks ([1] mention that their polynomial degrees are impractical). Landscape analysis of DNN training (which, roughly speaking is: all local min are global min and so escaping saddle-points and converging to local-min is needed) provides an alternative view for finite-width multilayer neural networks (see e.g., **P2** and its references). Our contributions in this paper are geared towards the later angle – as DNNs are also interpolators in practice, we show that under SGC we can converge faster to local-min. **Other examples:** Another concrete example satisfying SGC condition is online matrix completion (see **P1** for details). We will add this example in detail in our revision. **Contributions:** Please see Lines 2-10 above. **[AZL18]:** At a high-level, the suggested approach would also work. However, the method in [AL18] is a theoretical computer science style reduction approach. It involves a wrapper algorithm on top of PSGD which increases overall runtime. Our result directly analyzes PSGD iterates, more in line with optimization and machine learning type results. Furthermore, a main drawback of the approach in [AZL18] is that it is not directly applicable for the 0th-order setting due to their bounded variance (of stochastic gradient) assumption. Please also see the discussion in Remark 5 for more details. **Variance reduction (VR):** A motivation for the SGC assumption is that it provides *automatic VR*. As in Table 1, for stochastic setting under SGC, PSGD already achieves the corresponding complexity of its deterministic counterpart (without acceleration). It is interesting future work to examine if similar results hold for finite-sum setting. However, most VR methods invariably involve double-loop algorithms and their empirical performance in deep learning has come under close scrutiny in the recent past; see [DB19] and [Sch20].

Rev #2: Intuitions: SGC at a high-level could intuitively be described as an assumption satisfied in interpolation models, providing *automatic variance reduction*. Please also see Lemma 3.1 and Remark 1 for more intuition. We will clarify this more in our revision. **Detailed distribution of stochastic grad:** We clarify that *we do not make any distributional assumption* on the stochastic gradient. SGC is only a variance/moment based assumption. **Finite-sum opt:** Finite-sum opt is a special case of stochastic setting and hence the assumptions are satisfied; see also [MBB18, VBS18]. **Proof Sketch:** Thanks for this suggestion. We will add a proof sketch in our revision. **High-order method:** Making higher-order methods practical is an interesting future work. We will clarify this point and reorganize.

Rev #6: Theoretical contributions: Please see Lines 2-10 above. **Approximate SGC:** Thanks for raising this extremely interesting question. Considering 1st-order setting, departures from SGC can be modeled, for example, by: $E(\|\nabla F(x_t,\xi)\|^2) \leq \rho\|\nabla f(x_t)\|^2 + e_t^2$ where $e_t^2 > 0$ is an additive non-vanishing iteration-dependent noise variance; see also [VBS18]. This assumption could be used for example to model certain specific types of label-noise. When $e_t^2 = \sigma^2, \forall t$, the oracle complexity is $\tilde{O}(\epsilon^{-4})$ since this case is essentially equivalent to the standard stochastic gradient setting and SGC has no effect. But one could obtain rates approaching $\tilde{O}(\epsilon^{-2})$ depending on the decay of the term $R_t = \sum_{i=1}^{t} e_i^2$. If we assume $R_t \approx t^\alpha$ for $\alpha \in (-\infty, 1]$, then we can show that the complexity is $\tilde{O}(\epsilon^{-\max(2,3.5+0.5\alpha)})$. Hence, when $\alpha = -2$, the complexity is $\tilde{O}(\epsilon^{-2.5})$ and when $\alpha \leq -3$, it is $\tilde{O}(\epsilon^{-2})$. In this setup, it may be possible to connect label noise (for specific noise models) to α. This would provide a possible way of incorporating label noise in this setup. It is intriguing to examine this problem rigorously as future work. We will clarify this in our revision. **Experiments:** We will be happy to add simulations and real-world experiments on DNN and Online matrix completion in our revision. We clarify that our main goal in this work is to provide a plausible explanation for the question *why do optimization algorithms for deep-learning models work efficiently in practice despite the associated non-convexity ?*. As an attempt, [MBB18], [VBS18] and related works proposed the SGC condition which is satisfied due to the interpolating nature of DNNs. However, prior works fell short of providing a complete explanation as they only analyze convergence to critical points. Recently, it has been shown in several works that for DNNs (in the finite-width regime), all approximate local minima are also global minima (see reference **P2**) due to which converging to local-min and escaping saddle-points are important. Hence, based on these motivations, we show that PSGD and SCRN converge to local-min faster with SGC condition. We emphasize that this line of work is only one plausible explanation for the above question and there are other directions (e.g., NTK) attempting alternative explanations in the literature.

References: P1 - *Provable Efficient Online Matrix Completion via Non-convex Stochastic Gradient Descent*, Jin et al., NeurIPS 2016. **P2** - Elimination of All Bad Local Minima in Deep Learning, Kawaguchi et al., AISTATS, 2020.