Response to Reviewer 1: Thank you for your detailed review. In turn, we will provide detailed clarifications in the hope that these will persuade you to revisit your verdict and increase your score for our paper.

First, our results are not subsumed by [1] (we use your reference numbering). Their most relevant results show that 2-layer networks have a loss function which, on some convex sets, resembles a convex function in that all local minima are global minima, and all sublevel sets are connected. These two properties do not imply convexity, however; $f(x, y) = x^2y^2$ satisfies these conditions and is nowhere convex. Our results, on the other hand, imply local strong convexity on certain sets, which is a stronger conclusion. They also hold for any network architecture, as opposed to the 2-layer networks of [1]; this is a unique contribution of our paper, since many theoretical results in this area use restricted architectures. Conversely, the results of [1] hold on all of weight space, while our results hold on $U(\lambda, \theta)$, a proper subset which is nonetheless important as it can contain all global minimizers of the problem (see Lemma 4).

The connection between our work and [3] is tenuous; the Gram matrix they study only uses first derivatives of the predictions with respect to the weights and so says nothing about the convexity of the loss function.

Our responses to your technical comments use your enumeration. We have not responded to points we do not dispute.

1. The results of [2] hold for almost all noise realizations, and after selecting a particular realization, small modifications of our results hold. In this setting, only our calculation of the second derivatives of the loss needs to be modified, but roughly the same proof works with a leaky ReLU provided the slope s for negative inputs is bounded.
2. You've misunderstood our derivation of (7). Define $A(R, \lambda, \theta)$ as the set on the right hand side of (7). Then
$$W \in A(R, \lambda, \theta) \quad \text{implies} \quad ||W||_* \leq R \quad \text{and} \quad \ell(W)^{1/2} < \frac{\lambda - \theta}{\sqrt{2}H(H+1)rR^{H-1}}$$
From these two conditions we have $A(R, \lambda, \theta) \subset B(R)$ and
$$\ell(W)^{1/2}||W||_*^{H-1} \leq \ell(W)^{1/2}R^{H-1} < \frac{(\lambda - \theta)R^{H-1}}{\sqrt{2}H(H+1)rR^{H-1}} = \frac{(\lambda - \theta)}{\sqrt{2}H(H+1)r} \Rightarrow A(R, \lambda, \theta) \subset U(\lambda, \theta).$$
3. Here we are referring to the bounded region $A(R, \lambda, \theta)$. We will make this clearer in the paper.
4. If the training data and network architecture is such that zero non-regularized error is possible, then a small enough loss can be obtained at some W regardless of the size of $||W||_*$; see the definition of $U(\lambda, \theta)$ to understand why.
5. Lemmas 5 and 6 do not guarantee a local minimum of ℓ_λ in $U(\lambda, \theta)$. Lemma 4 does, under some conditions.
6. The training process typically halts before the bound is satisfied. Thus, our work may describe the region around a global minimizer which gradient descent does not typically reach. But, see our response to Reviewer 2 for the details of a new experiment we performed in response to their comments which shows the relevance of this bound.

7a. These are distinct but related concepts. Strong convexity is sufficient for strong convexity restricted to gradient paths, but not necessary. Some useful properties of convexity still hold if a function is strongly convex restricted to gradient paths. For example, some guarantees of convergence rate for gradient descent still hold.

7b. Note that we use weight decay, so the loss becomes coercive, and a finite global minimizer is guaranteed.

9. As we mention, $y(a_i, W)$ is locally a polynomial in the weights where the maximum degree of each variable is 1, like $f(x, y, z) = xy + xyz$; differentiating this twice with respect to the same variable always gives 0.

11. You are correct that this doesn't guarantee the existence of a critical point, but this is not necessary for the validity of the lemma. The existence of a critical point could be guaranteed, for example, by Lemma 4.

Response to Reviewer 2: Thank you for your thoughtful review. Based on your comments we have conducted a simple regression experiment with target function $f(x) = 0$ and 100 data points sampled uniformly in $[-1, 1]$. We used an architecture of $H = 1, n_1 = 2$, no biases, no ReLU on the output, and weight decay parameter $\lambda = 1$. Despite having a trivial target function, this problem is non-convex. Experiments show that in 100 independent trials, the bound given in (6) is satisfied for some θ for $51.6\% \pm 24.0\%$ (mean ± standard deviation) of the loss change over training - see the paper for the definition of this metric. Hence, gradient descent does enter the set $U(\lambda, \theta)$, but only after some descent.

This experiment provides a stronger connection between the theory and numerics. It is considerably simpler than the simplest example we tested before, which had $H \geq 2$. We will include the details in the final version of the paper.

Response to Reviewer 3: We are grateful for your kind words about our paper. We studied the deep and non-linear networks used in practice with no unrealistic assumptions, and we are thankful that you appreciate this.

In terms of using our results to explain existing empirical findings, there is one comment we can make. Empirically, we observed that the addition of batch normalization helped the gradient descent trajectory enter the piecewise strongly convex regime much sooner than without it (compare the results for MNIST to those for CIFAR10 and 100). A possible conjecture is that batch normalization convexifies the loss function; this is an interesting avenue for future research.