We sincerely thank all the reviewers for their meticulous work and their helpful and detailed comments. Specifically, thank you for your suggestions on expanding upon our current figure/simulations as well as adding more intuition and high level ideas so as to help the reader through the admittedly complex proof. We will gladly incorporate them!

[To Reviewer #1:] Thank you very much for your hard work and your supportive comments for our paper!
1."perhaps between section 3 and 4 . . . what would be great." Thank you for this suggestion. We will update the revised version analogously.
2."Regardless,... supplementary article." In section 14.4.2 titled "Variable Step Size", the step-size $\eta_t = \frac{B}{\rho\sqrt{t}}$ is used in SGD. For the Stable manifold theorem in line 30, we refer to Chapter 5 Theorem III.7 in "Global Stability of Dynamical Systems" by M. Shub and in line 130, we refer to Section 2.7 in "Differential Equations and Dynamical Systems" by L. Perko.
3."Might be interesting to add a note... Is this correct?" Yes, for the constant step-size α, we need α to be less than the inverse of largest eigenvalue ([10], Lee et al, 2019). In the case of vanishing step-sizes the upper bound is unnecessary from an asymptotic perspective, but it is practically better to start with $\alpha_0 < 1/L$.
4."Does the paper... stuck in a saddle point" The last parts of the statements of Theorem 4.1 and Corollary 4.2 assert that almost all points in a neighborhood of a saddle point will be transported out of the neighborhood in a finite number of steps. The convergence criterion "$f(x_{k+1}) \leq f(x_k) - \alpha_k(1 - \frac{\alpha_k L}{2})||\nabla f(x_k)||^2$ for all k" for convex functions can be used in neighborhoods of local minima (e.g.p126, *Convex Optimization Algorithms*, by Bertsekas). It asserts that once a point enters a neighborhood of local minimum, it converges. So convergence can be attained by combining these results.
5."This is of course... that area of study." Recent progress (e.g.[6] and Jin et al 2017) show that perturbed gradient descent can escape saddles efficiently and then find approximate local minima in polynomial time. The efficiency of deterministic methods relies heavily on the specific structure of functions and initialization. In "GD can take exponential time...", Du et al constructed a function for which GD is significantly slowed down by saddles. But GD works well for the functions of matrix factorization(Jain et al 2017), phase retrieval, dictionary learning and so on (pointed in [10]). Generic efficiency results for first order methods without noise is an interesting direction for future research. Especially for variable step-sizes, very little is known. Methods using information of Hessian and curvature are studied in recent references (e.g."A geometric analysis of phase retrieval" Sun et al and "On noisy negative curvature descent" Liu and Yang, "A generic approach..." by Reddi et al, 2017). We will expand upon these connections in the revision.

[To Reviewer #2:] Thank you very much for your hard work and your supportive comments for our paper!
"If it was a space issue..." The case of coordinate descent is left open due to technical difficulties. Our main result (Theorem 1.1) relies on the fact that the differential of update rules can be written as $I - \alpha_k H$ where H has no complex eigenvalues. The differential of coordinate descent can lead to complex eigenvalues, and thus our stable manifold theorem (Theorem 4.1) cannot be applied as is. Actually the issue is even more complicated. Suppose we had a generalization of Theorem 4.1 that addresses complex eigenvalues, it is not clear if it would imply the desired result for coordinate descent, because the differential of coordinate descent update rule cannot be written in the form of $I - \alpha_k H$ with H diagonalizable. New ideas will probably be required to address these intricate issues. We will make a formal remark about this to hopefully stimulate future research on this important direction.

[To Reviewer #3:] Thank you for your work, comments and suggestion on improvements.
1."Since, the latter condition is no longer required, ... unclear." The condition of $\sum \alpha_k^2 < \infty$ is unnecessary in the deterministic case. It can be understood in the following sense:
Intuition: 1. The reason to introduce $\sum \alpha_k^2 < \infty$ is to make the total variance of the increments finite (e.g. p699 of [16] R.Pemantle,1990). But deterministic methods can be seen as stochastic methods with noise identically 0, so the "total variance" is identically 0 (which is finite) for any step-sizes. 2. Consider the case when the step-size is a constant α (then $\sum \alpha^2 = \infty$). We already know that our result is true in this case since this is exactly the result of [10] Lee et al, 2019. So the condition $\sum \alpha_k^2 = \infty$ should not cause any obstruction in our deterministic setting.
Technical sketch: Formally, the proofs of Theorem 4.1 and Corollary 4.2 do not require $\sum \alpha_k^2 < \infty$. Briefly the proof is to use the Banach Fixed Point Theorem to show the existence and uniqueness of the stable manifold. The place where property of α_k is actually used is in the proofs of Lemma C.1 and Lemma C.2, where we show that the sums $R_k = \sum_{i=0}^{\infty} \alpha_{k+1+i} \prod_{j=k+1}^{k+1+i} (1 - \alpha_j \lambda)^{-1}$ with $\lambda < 0$, and $S_k = \alpha_k + \sum_{i=0}^{k-1} \alpha_i \prod_{j=i+1}^{k} (1 - \alpha_j \lambda)$ with $\lambda > 0$ are bounded as $k \to \infty$. To see that the restriction of $\sum \alpha_k^2 < \infty$ is unnecessary, consider the case of constant step-size (which trivially implies $\sum \alpha_k^2 = \infty$), then R_k and S_k become geometric series with ratios less than 1, so boundedness immediately follows.
2. "Do the additional conditions (lines 167-168) lead to elimination of the condition?" That is correct. These conditions are used to obtain the expression in Line 54 in the supplementary material.