We appreciate the valuable comments, which urged us to embody explicit connections to practices of learning. Apology that not all comments are replied here and our replies have to be short due to space, but they'll be fully addressed in a revision. We plead a reconsideration based on the improvement, as our contribution is truly innovative and nontrivial.

Re: connection to learning, and when Cond.1&2 hold. Here is an example (simplified and only briefly explained for length) in which the loss will be multiscale as considered in our paper: train a 2-layer neural network to fit data $\{x^k, y^k\}_k$, where the output $y^k = y_0^k + y_1^k + \xi^k$ admits a decomposition into large scale behavior $y_0^k = g_0(x^k)$, microscopic detail $y_1^k = \epsilon g_1(\epsilon x^k)$, and i.i.d. noise ξ_k. Assume g_0 and g_1 are regular enough so that universal approximation (UA) works and they can be approximated by wide enough neural networks with $\mathcal{O}(1)$ weights. Consider MSE loss $\sum_k \|y^k - \sum_i a_i\sigma(W_i x^k + b_i)\|^2$ with σ being the periodic activation in a recent progress [Implicit Neural Representations with Periodic Activation Functions, 2020]. Then there exists a minimizer and in its neighborhood the loss satisfies Cond.1& 2: omit k WLOG, absorb bias into weight, and rewrite the loss as (denote by $\theta = [a_i, W_i]_i$)

$$f(\theta) = \Big\| y_0 - \textstyle\sum_{i\in I} a_i\sigma(W_i x) + \epsilon y_1 - \sum_{j\notin I} a_j\sigma(W_j x)\Big\|^2 = \Big\| g_0(x) - \sum_{i\in I} a_i\sigma(W_i x)\Big\|^2$$

$$+ 2\epsilon\Big\langle g_0(x) - \textstyle\sum_{i\in I} a_i\sigma(W_i x), g_1(\epsilon x) - \sum_{j\notin I} a_j\sigma(W_j x)\Big\rangle + \epsilon^2\Big\| g_1(\epsilon x) - \sum_{j\notin I} a_j\sigma(W_j x)\Big\|^2$$

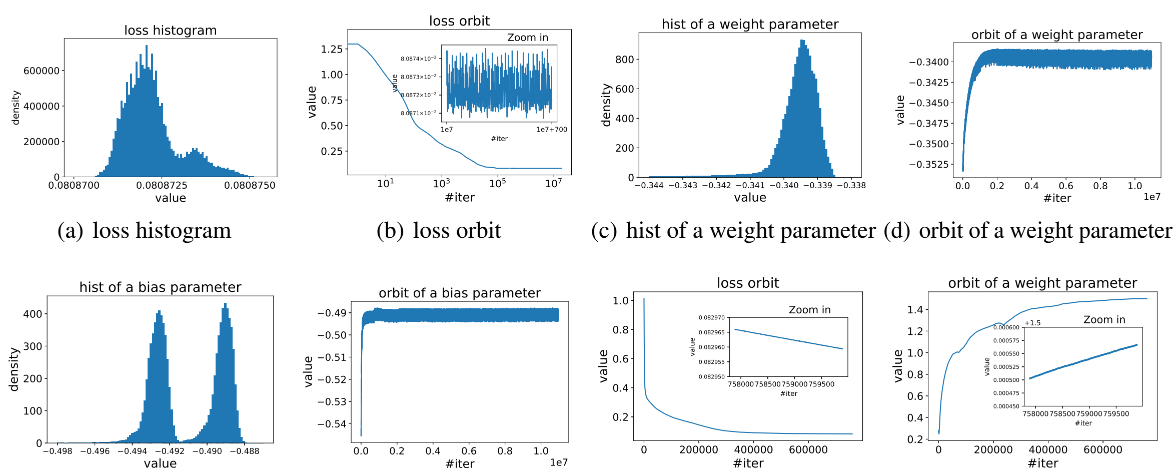
where I and I^c are sets of nodes, each large enough for UA to ensure vanishing loss. Renormalize by letting $\hat{x} = \epsilon x$ so that UA works for $g_1(\cdot)$, then the 2nd term rewrites as

$$2\epsilon\Big\langle g_0(x) - \textstyle\sum_{i\in I} a_i\sigma(W_i x), g_1(\hat{x}) - \sum_{j\notin I} a_j\sigma\Big(\frac{W_j}{\epsilon}\hat{x}\Big)\Big\rangle.$$

This is in the form of $\epsilon\hat{f}_1(\theta/\epsilon, \theta)$ for some $\hat{f}_1(\phi, \varphi)$ that is quasiperiodic in ϕ (quasiperiodic because $\hat{x}$ is multi-dim). The 3rd term rewrites similarly. Thus, we see $f(\theta) = f_0(\theta) + f_{1,\epsilon}(\theta)$ where f_0 is the 1st term and $f_{1,\epsilon}(\theta) = \epsilon\hat{f}_1(\theta/\epsilon, \theta) + \epsilon^2\hat{f}_2(\theta/\epsilon, \theta)$ for some $\hat{f}_1$, $\hat{f}_2$ quasiperiodic in the 1st argument. Such $f_{1,\epsilon}$ satisfies Cond.1&2 due to its quasiperiodic small scale. $\square$

Like most theory papers, we also present numerical experiments in which our conclusions still hold although conditions for our theorems no longer apply. Thanks to the reviews the following will be added (and expanded):

Neural network training. We use fully connected 5-16-2 MLP to regress UCI Airfoil Self-Noise Data Set, with leaky ReLU, MSE as loss, and batch gradient. Fig.1 shows large LR again produces stochasticity as our paper studies.



(a) loss histogram (b) loss orbit (c) hist of a weight parameter (d) orbit of a weight parameter

(e) hist of a bias parameter (f) orbit of a bias parameter (g) loss orbit (small LR) (h) parameter orbit (small LR)

Figure 1: (a)-(f) use LR=0.0165 (large) and demonstrate stochasticity originated from chaos as GD converges to a statistical distribution rather than a local min. (g,h) use LR=0.001 (small) and GD converges to a local min.

Re: $f_{1,\epsilon}$ satisfying Cond.1&2 is like a random variable; tautology? $f_{1,\epsilon}$ does contribute like a r.v., but this needs to be proved, which is one of our main contributions – note both x and $f_{1,\epsilon}(x)$ are deterministic even under Cond.1&2! Cond.1&2 use auxiliary random variables to define the needed $f_{1,\epsilon}$, but $f_{1,\epsilon}$ is a deterministic function.

Re: weaken isotropic noise assumption? We don't require isotropic 'noise'. Kindly see e.g., Thm.2, which contains 2 statements: (i) convergence to stochastic behavior for general covariance; (ii) explicit characterization of the limiting statistics when covariance is isotropic (note the same thing holds for SGD).

Re: valid in multi-dim? Apology that multi-dim. and nonconvex demonstrations were left in Appendix C.2, C.3.3, & C.5. This rebuttal also adds a neural network example, which is high-dim. & nonconvex, and our conclusion still holds.