We thank reviewers R1, R2, R3, R4 for a wealth of constructive comments. The reviewers expressed interest in our results but had concerns over our exposition's clarity. In response to these concerns, we will thoroughly revise the paper as detailed below, and in addition to the NeurIPS submission process we will try to submit to a journal. As R1 noted, conferences have the advantage of visibility for ideas we are proud of.

PLANNED RE-FACTORING OF THE NARRATIVE —— Per R1, R4, we will

PLANNED ECONOMIZING OF SPACE —— R1: we will use the full 8 pages and remove distracting content such as the Chladni plate image and archaic references from the paper body.

PLANNED CLARIFICATION OF CLAIMS —— We will expand Sect 1.2 to less tersely define the test loss $l(\theta)$, the generalization gap, and the tensors $M_1^1, M_1^2, \cdots$, where we adopt R1's suggested M notation. —— —— —— Per R1, we will render hand-drawn figures cleanly, and rework Fig 4 entirely by plotting an SGD trajectory through three cross-sectional slices of the landscape, in each slice indicating gradient noise with contour lines and expected loss with a colored heatmap. Only two slices are necessary, but to emphasize that the effect happens for all time, we'll show the trajectory for three slices. We will also orient Fig 4 to align with Fig 1a.

PLANNED DEFENSE OF SIGNIFICANCE ——

TECHNICAL CLARIFICATIONS —— R1 Ln 51**: How is the expression in an eigenbasis of ηH? It looks like only H.** H is rank $(0, 2)$ (no upper idxs, two lower idxs), so it maps vectors to covectors and without further structure we can't speak of its eigenvalues. That's why we use $(\eta H)^\mu_\nu$, a linear map that maps vectors to vectors, to get an eigenbasis. Then $\eta(H_{\mu\mu} + H_{\nu\nu} + \cdots)$ is short for "the μth eigval of ηH, plus the νth eigval ...". On Ln 51$\frac{1}{2}$ we are explicit instead of using summation convention; accordingly, we disobey the usual syntax for upper/lower idxs. This is confusing, so we will use uniform notation throughout the revision. —— R1 **I am skeptical of the authors' claim that this work reconciles conflicting results on sharp vs flat minima.** FILL IN —— R2 Sect 3**: What was being plotted, architectures, etc?** Fig 3.a shows the test losses (y axis) attained after fixed-time SGD runs with different learning rates (x axis), with one random initialization. We'll expand all figures' captions and discussion. Appx C.2.1 lists architectures. —— R2 Sect 2.5, R4 Sect 3.2**: Do corrections proposed satisfy these scaling relationships? That higher-order approximations outperform lower-order approximations feels tautological. Very artificial example does not motivate third order dynamics.** Some but not all of our corrections obey SDE noise-scaling laws in that they are functions of η/B. We view our experiments as verifying that we forgot no factors of 2 etc. As with many NeurIPS papers, our contribution is theoretical, and we suggest but do not demonstrate that this theory may one day improve training of modern neural nets. Our artificial examples are typical in that the third order contributions that they isolate are all present in generic loss landscapes. We show how to interpret third order terms, yielding insight when they are non-negligible. These terms may be negligible in practice, but experiments on real data (Fig 3 green lines) suggest they are sometimes substantial. —— R3 Dfn 1**: What do diagrams stand for? Are they valid math?** Formally, diagrams represents (sets of) terms in a Taylor expansion. Appx A.6 gives visual intuition. Appx B gives defns and proofs. —— R3 Ln 119, R4 Cor 3**: Is this ERM? What does *test* refer to? Formally define $l(\theta)$ and generalization gap** We study SGD as an approximate method for ERM. The test loss is the expectation $l(\theta) \triangleq \mathbb{E}_{x\sim\mathcal{D}}[l_x(\theta)]$ over fresh samples x from the underlying distribution $\mathcal{D}$, as suggested by Ln 65's word "unbiased". The generalization gap on a training set $\mathcal{S} \sim \mathcal{D}^N$ is $\mathbb{E}_{x\sim\mathcal{D}}[l_x(\theta)] - \mathbb{E}_{x\sim\mathcal{S}}[l_x(\theta)]$. Like prior work [e.g. Chaudari], our predictions depend on the underlying (and in practice unknown) distribution $\mathcal{D}$; one may obtain qualitative insight (e.g. Sect 2.3) and unbiased estimates (Appx C.6) with just training data. —— R4 Cor 3**: Can one find a term l_c that works globally? Can it be computed at less cost than running SGD?** Yes, Appx C.6 gives estimates for expressions of arbitrary order with only constant factor time overhead. E.g. $2l_x(\theta) \cdot \nabla(l_x(\theta) - l_y(\theta))$ is for any fixed θ an unbiased estimate of ∇C, for $x, y \sim \mathcal{D}$. This local estimate may thus be computed at each step as θ_t evolves.