We thank reviewers R1, R2, R3, R4 for a wealth of constructive comments. Reviewers expressed interest in our results but had concerns over our exposition's clarity. In response to these concerns, we will revise the paper as detailed below, and also 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 R4 and R1, we will highlight and frontload the stories of how and why SGD tends toward small C regions and C-flat minima. —— ★ Synthesizing R4's and R1's suggestions, we will disentangle the body and appendices by relegating diagrams to the appendices. We leave but three traces in the body: (A) we will note that our main technical contribution is a diagram-based formalism that streamlines use of our Key Lemma and that yields physical intuition; (B) after writing the relevant Taylor series (with lots of $\cdots$s) in the paper body and interpreting the corollary-relevant terms, we will present (as a foretaste) one diagram while broadly sketching the Taylor terms' combinatorics; (C) though we will move the Key Lemma (which does not mention diagrams) to the body and state all Cor.s without reference to diagrams, we will state (but not emphasize) our two diagram-related Thm.s.

Planned economizing of space —— R1: we will move Broader Impacts and remove distractions (e.g. Chladni plate image) from the body, freeing ≈ 1.2 pages. —— Moving diagrams to the appendices, we free up ≈ 1.2 pages (but ≈ 0.8 consumed again by ★). —— The net 1.6 freed pages we spend on clarifying our discussion: 0.2 pages each to: Sect 1 (frame problem (SGD dynamics with known loss landscape statistics) and solution (Taylor series), highlighting potential for insight (Cor 1,3); Sect 1.2 (what are l, l_x?); Sect 1.3 and Cor 5 (comparison to SDE work per R4/R2's questions); Cor 1 (expand discussion of mechanism and interpretation); Cor 3 (supplement with cartoon trajectory akin to revised Fig 4); Sect 3 (explain which measurements are plotted); Sect 3.3 (discuss Cor 2 as quantifying implict regularization and compare to previously published intuitions per R1's comments); Fig 4 (new figure takes space).

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. We will briefly review tensor 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.

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 were explicit instead of using summation convention; accordingly, we disobeyed the usual syntax for upper/lower idxs. This was 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.** We erred in stating our contribution so strongly. Instead of reconciling conflicting proved results, our work explains why two existing intuitions may co-exist, by observing that the two intuitive stories ("flat minima are robust to displacement in weights" / "sharp minima are robust to gradient noise") use different noise models. Our theory, by quantifying how gradient noise leads eventually to weight displacements, then allows us to calculate how curvature affects overfitting (Fig 3 bottom right). —— 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?** 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 $\boldsymbol{l(\theta)}$ 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.