We thank reviewers R1, R2, R3 for their feedback. Reviewers had concerns over our work's correctness (R1), counterintuitiveness (R3), citations (R2), and clarity (R1,R2,R3). We address these concerns in sqeuence.

Limits — view the expected testing loss as a function $L(\eta, T)$. For each d is a dth order truncation $L_d(\eta, T)$, a degree-d polynomial in η whose coefficients depend on T. Thm 2 gives a sufficient condition for $L_{d,\infty}(\eta) \triangleq \lim_{T\to\infty} L_d(\eta, T)$ to exist as well as a formula for $L_{d,\infty}$. R1 observes that, though Thm 2 controls $\text{LHS}(\eta) \triangleq \lim_{d\to\infty} \lim_{T\to\infty} L_d(\eta, T)$, it is $\text{RHS}(\eta) \triangleq \lim_{T\to\infty} \lim_{d\to\infty} L_d(\eta, T)$ that more interests us. How do LHS and RHS relate?

Proposition A. *Assume §B.1's boundedness and analyticity properties.* Supposing *that* $\nabla l(\theta_\star) = 0$ *and that on some open neighborhood* U *of* $\theta_\star$ *the hessian* $\nabla\nabla l_x(\theta)$ *is lower-bounded by some strictly positive definite form* $Q(\theta)$ *continuous in* θ, then *for any initialization* $\theta_0 \in V$ *in some open neighborhood* V *of* $\theta_\star$ *and for any homogeneous polynomial* $p(\eta)$ *(of* η*'s dim* $\times$ *dim many components) with no roots (besides* $\eta = 0$*):* $\lim_{\eta\to 0}(LHS(\eta) - RHS(\eta))/p(\eta) = 0$.

Proof idea. Gradient and hessian bounds give for all ϵ some δ, δ' so that for all T, all $\eta < \delta$, and all θ_0 with $|\theta_0 - \theta_\star| < \delta'$: $|\theta_T - \theta_0| < \epsilon$ with probability 1. In fact, □

TODO: prove, discuss, and discuss irrelevance! PropA is a straightforward extension of the proof of Thm 2. But from our viewpoint, PropA is inessential: in practice (e.g., with CIFAR conv-nets) we have observed none of the pathologies that PropA seeks to control (Page6, last par). We the authors prefer to admit both mathematical analysis and scientific measurement as means of discovering.

Sharp Minima — R2 finds Cor 5's statement (that overfitting (defined as the increase in testing loss l upon initializing at a local minimum of l and then training) is, to second order in η, greatest when the ηH has moderate eigenvalues) counterintutive. We do, too. Compare Fig 5 to [**Ke**]'s Fig 1 and note that SGD's noise structure is *not* that of *displacements* in weight space; rather, it is that of error terms $\nabla l_x(\theta) - \nabla l(x)$ in the *gradient* estimate. Say dim = 1 and imagine a testing loss $l(\theta) = a\theta^2/2$ and a training loss $\hat{l}(\theta) = l(\theta) + b\theta$. At the training minimum $\theta = -b/a$, the testing loss is $b^2/(2a)$. So for fixed b, sharp minima ($a \gg 1$) overfit less (we invite R2 to run the gist at gist.github.com/anonymous-taylor-series to see this 'in person'). This example suggests that (**A**) if we optimize to convergence, sharp minima overfit less; that (**B**) convergence is slow near flat minima explains why theory and measurement find that flat minima also overfit little. (Our small-η assumption rules out the possibility that H is so sharp that SGD diverges: in the regime $1/T \ll \eta H \ll 1$, sharper minima overfit less). Prior work (see Page12, par 5) finds that (contrary to [**Ke**]) sharp minima overfit little. By explicating η's role in translating gradients into displacements, our theory accounts for both (**A**) and (**B**), thus unifying existing pro-flat and pro-sharp intuitions (e.g., [**Ke**] and [**Di**]). We view it as a merit that our formalism makes such counterintuitive phenomena visible.

Implicit Regularization —

Assumptions — Reviewers

Notation — Our paper used a tensor-index convention found in tensor statistics ([**Mc**], [**Dy**]) but not in CS at large. Our camera-ready will make all '$\sum$s explicit. Based on R2,R3's feedback, we believe this small change will substantially improve a reader's experience.

Organization — Reviewers R2,R3 struggled with our paper's organization. We believe that decomposing the paper into three tracks (to be selected between based on a reader's goals) will help us tell our story without causing the burden of too many frontloaded concepts or the confusion of too many forward references to backloaded concepts. **Track A** [pgs 1-4], for the reader who wants a 'free sample', will eschew diagrams, general theorems, and §1.1, 2.2's heavy notations. It will illustrate Taylor series via §2.1's proof, identify the concrete terms relevant to §3.3, state (with §B.1's assumptions explicit and with PropA's level of precision) Cor 4, and conclude with §4.2's verification of SGD's sensitivity to curl. **Track B** [pgs 1-4, 5-12], for she who seeks physical intuition for our corollaries, will take Track A as motivation for the formalism of §1.1 and §2.2, which we will illustrate as in §A.4's. §2.3 will discuss of re-summation physically, in the style of Fig 5, 7. §2.4 will include PropA per R1's feedback. For space, we'll move §4 to §C; but each of §3.1, 3.2, 3.4 will briefly summarize the relevant empirical confirmation. **Track C** [pgs 4-12, 15-42], for she who wishes to use and extend our formalism, will . We will re-organize the paper accordingly.

Surprise — We are glad that some of our results surprised R2, R3. We believe these results, and more importantly the physical-geometric viewpoint that led to them, are worth sharing. Recalling that many of our favorite papers are those offering a disorientingly fresh viewpoint, we hope that the reviewers can feel the same about our work.

[**Ba**] D.G.Barrett, B.Dherin. Implicit Gradient Regularization. ICLR 2021.

[**Di**] L.Dinh, R.Pascanu, S.Bengio, Y.Bengio. Sharp Minima Can Generalize for Deep Nets. ICML 2017.

[**Dy**] E.Dyer, G.Gur-Ari. Asymptotics of Wide Networks from Feynman Diagrams. ICLR 2020.

[**Ke**] N.S.Keskar et alia. On Large-Batch Training for Deep Learning. ICLR 2017.

[**Mc**] P.McCullagh. Tensor Methods in Statistics. Dover Books 2017.