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 (and do they exist)?

Proposition A. *Assume §B.1's boundedness and analyticity properties;* ***if*** $\nabla l(\theta_\star) = 0$ *and 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 × dim many components; and with no roots besides* $\eta = 0$*):* $\lim_{\eta\to 0}(LHS(\eta) - RHS(\eta))/p(\eta) = 0$ *is well-defined and vanishes.*

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, FILL IN □

PropA is a straightforward extension of Thm 2's proof. But in some sense, PropA is inessential: in practice we have observed none of the pathologies that PropA seeks to control (Page6, last par). Ultimately, we view our experiments (e.g., with CIFAR conv-nets) as verifying that our theory has substance.

Sharp Minima — R3 finds Cor 5 (that overfitting (i.e., the excess testing loss l incurred after initializing at a testing minimum 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; note that SGD's noise consists not of *displacements* in weight space but of error terms $\nabla l_x(\theta) - \nabla l(x)$ in the *gradient* estimate. Say dim = 1 with testing loss $l(\theta) = a\theta^2/2$ 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. The covariance C controls b^2, explaining Cor 5's $C/2H$ factor. (Run this toy example at gist.github.com/anonymous-taylor-series.) In this case, optimization to convergence favors sharp minima (**A**); but convergence is slow near flat minima, so theory and measurement agree that flat minima also overfit little (**B**). (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**) and unifies 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 and ODE — We thank R2 for the highly relevant article. In brief,

Verification — R3,R1 raise concerns about verifiability.

Notation — R2,R3 note that our use of 'Einstein notation' (though found in tensor statistics ([**Mc**], [**Dy**])) is alien to CS at large. Our camera-ready will make all '$\sum$'s explicit while also referring the reader to [**Mc**], as well as a new Einstein-free appendix §D, for examples of tensor manipulations.

Organization — R2,R3 stress the challenge of organizing our paper so as to avoid both the burden of too many frontloaded concepts and the confusion of too many forward references to concepts delayed. We'll address this challenge by segmenting the paper into three tracks (to be selected by a reader *based on her goals*), each with a self-contained subset of concepts: **Track A** [pgs 1-4], for readers who want 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 Cor 4 (with §B.1's assumptions explicit and with PropA's level of precision), and conclude with §4.2's verification of SGD's sensitivity to curl. **Track B** [pgs 1-4, 5-12], for she who seeks our results and their physical intuitions, will use Track A to motivate (and §A.4 to illustrate) §1.1/2.2/2.3's definitions, relegating §2.2/2.1's Lemma and discussions of terms to §B. §2.4 will include PropA per R1's feedback as well as a cartoon (in Fig 5,7's style) of resummation's 'physics'. 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 5-12, 15-42], for she who wishes to use and extend our formalism, will FILL IN.

References — [**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.