


We thank reviewers [R1](#), [R2](#), [R3](#) for their substantial time investment and for incisive feedback. We address in turn your concerns over our work’s correctness ([R1](#)), counterintuitiveness ([R3](#)), citations ([R2](#)), and clarity ([R1](#), [R2](#), [R3](#)):

**LIMITS** — We thank [R1](#) for highlighting ways our precision must improve. For instance, Thm 2 gives convergence of *truncated* series (Pg6Par(-1)). Explicitly ([R3](#): note Thm2 defines ‘non-degenerate’ as ‘ $H(\theta_*)$  is positive definite’):

**Thm 2.** *For each  $d$ , each non-deg. local min.  $\theta_*$  has a nbhd  $U$  whose every member  $\theta_0$  induces, via Thm 1, a  $T$ -indexed sequence (of degree- $d$  polynomials  $f_T \in \mathbb{R}[\eta]$ ) that converges (uniformly on any compact set of  $\eta$ s) as  $T \rightarrow \infty$  to some element  $f_1 \in \mathbb{R}[\eta]$ . Here,  $\mathbb{R}[\eta]$  is the formal polynomial ring in  $\dim(\mathcal{M})^2$  many variables.*


So if  $L_{d,T}(\eta)$  is Thm1’s truncation, Thm2 controls  $L_d(\eta) = \lim_{T \rightarrow \infty} L_{d,T}(\eta)$  but not  $L_T(\eta) = \lim_{d \rightarrow \infty} L_{d,T}(\eta)$ , even for  $d, T \gg 1$ . What makes Thm1,2 significant is our **empirical** finding that their formal power series bear on SGD practice (w.r.t. which *any* infinities are idealizations).<sup>1</sup> Formal power series are logically consistent algebraic objects [[Wi](#)], and, as [R2](#) noted, they offer empirically supported intuitive insights into SGD. Since our mathematical analysis was but an inspiring probe and a strong heuristic, we did not examine conditions (such as PrpA’s) under which  $L_d, L_T$  agree:

**Prop A.** *Fix  $U \subseteq \mathcal{M}$  open with compact closure  $\bar{U}$ ,  $\theta_* \in U$  a non-deg. local min. of  $l$ . Assume §B.1 as well as global, prob.-1 bounds  $(\|l_x(\theta)\|, \|\nabla l_x(\theta_*)\|) < C$ . If some  $Q_-, Q_+ \in \text{SPD}$  bound the hessian ( $Q_- < \nabla \nabla l_x(\theta) < Q_+$ ) on  $\bar{U}$ , then for all  $d$  and for any  $\theta_0$  in some nbhd  $V_d$  of  $\theta_*$ : for some  $T_0$  and  $|g|$  in  $\exp(-\text{big}\Omega(T))$ :  $\sup_{T \geq T_0} |L_d(\eta) - L_T(\eta) - g(T)|$  is  $o(\eta^d)$  (and exists on some nbhd in SPSPD of  $\eta = 0$ ). Here,  $SP(S)D$  consists of symmetric positive (semi)definite forms.*

**SHARP MINIMA** — [R3](#) finds Cor 5<sup>2</sup> counterintuitive. We do, too. Compare Fig5  to [[Ke](#)]’s Fig1; note that SGD’s noise consists not of weight *displacements* 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  $\hat{l}(\theta) = l(\theta) + b\theta$  (anonymized code here). 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. In this case, optimization to convergence favors sharp minima ( $\star$ ); but convergence is slow ar flat minima, so flat minima also overfit little ( $\diamond$ ). (Our small- $\eta$  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 Pg12Par5) finds that (contrary to [[Ke](#)]) sharp minima overfit little. Recognizing  $\eta$ ’s role in translating gradients to displacements, our theory accounts for both ( $\star$ ) and ( $\diamond$ ) 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, [FILL IN](#)

**ASSUMPTIONS AND VERIFICATION** — [R3](#), [R1](#) raise concerns about verifiability. [FILL IN](#)

**NOTATION** — [R1](#) notes that we work with vector quantities, not scalars [FILL IN](#). [R2](#), [R3](#) identify ‘Einstein notation’ (though found in tensor statistics literature ([[Cu](#)], [[Am](#)])) as alien to CS at large. Our camera-ready will make all ‘ $\Sigma$ ’s explicit while also referring the reader to [[Cu](#)] (and to as a w pedagogical Einstein-free appendix §D) for tensor manipulation examples. We are also open to translating all diagrams to a more text-friendly representation, e.g.,  $[a][ab : c : d][bcd]$  for  (letters name edges).

**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 pointers 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 PrpA’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) §1.1/2.2/3’s definitions, relegating §2.2/2.1’s Lemma and discussions of terms to §B. §2.4 will include PrpA per [R1](#)’s feedback as well as a cartoon (in Fig5,7’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** — [[Am](#)] S-I. Amari, H. Nagaoka. *Information Geometry*, pg 5. Oxford UP 1993. [[Ba](#)] D.G. Barrett, B. Dherin. *Implicit Gradient Regularization*. ICLR 2021. [[Cu](#)] P. McCullagh. *Tensor Methods in Statistics*, §1.1-1.4, §1.8. Dover 2017. [[Di](#)] L. Dinh, R. Pascanu, S. Bengio, Y. Bengio. *Sharp Minima Can Generalize for Deep Nets*, §1, §5. ICML 2017. [[Ke](#)] N.S. Keskar et alia. *Large-Batch Training for Deep Learning*, §4. ICLR 2017. [[Wi](#)] H. Wilf. *Generatingfunctionology*, §2.1-2.3. Academic Press 1994.

<sup>1</sup>In the classical continuum mechanics (CCM) of thermalized solids, ice cubes have infinite ergy; one manipulates *formal* power series. Yet in experience, CCM well-models many phenomena. [[Qu](#)] D.A. McQuarrie, J.D. Simon. *Physical Chemistry*, §1-1. University Science Books 1997.

<sup>2</sup>i.e.: that overfitting ( $\triangleq l(\theta_T) - l(\theta_0)$  after initializing at a testing minimum)’s second order term is greatest when  $\eta H$  has moderate eigenvalues