We thank reviewers R1, R2, R3 for substantial time investment, incisive feedback, and [**Ba**].

Limits. R1 highlights ways our precision must improve. Thm2 states: ⦃*For each d, every non-deg. local min.* $\theta_\star$ *has a nbhd* U *whose every member* θ_0 *induces, via Thm1, a sequence* $(L_{d,T} : T \in \mathbb{N})$ *of truncations, each a degree-d polynomial in* η*, that converges ptwise as* $T \to \infty$ *to some polynomial* L_d.⦄ So if $L_{d,T}(\eta)$ is Thm1's truncation, Thm2 controls $L_d(\eta) = \lim_{\tilde{T}\to\infty} L_{d,\tilde{T}}(\eta)$ but not $L_T(\eta) = \lim_{\tilde{d}\to\infty} L_{\tilde{d},T}(\eta)$, even for $d, T \gg 1$. **Empirically**, we find that Thm1/2's *formal* power series [**Wi**] predict SGD-in-practice (w.r.t. which *any* infinities are idealizations). Regarding our mathematics as but a strong heuristic,[1] we didn't examine when L_d, L_T agree. Still: **Prop A**. ⦃*Fix* $U \subseteq \mathcal{M}$ *open,* $\theta_\star \in U$ *a non-deg. local min. of l. Assume §B.1 and global, prob.*-1 *bounds* $(|l_x(\theta)|, \|\nabla l_x(\theta_\star)\|) < C$. ***If*** $\exists Q_-, Q_+ \in$ *SPD bounding the hessian* $(Q_- < \nabla\nabla l_x(\theta) < Q_+)$ *on* U, ***then*** $\forall d$ *and* $\forall \theta_0$ *in some nbhd* V_d *of* $\theta_\star$*:* $\exists T_0, A, B > 0$ *so that* $\sup_{T\geq T_0} ReLU(|L_d(\eta) - L_T(\eta)| - \exp(A - BT))$ *exists on some nbhd in SPSD of* $\eta = 0$ *and is* $o(\eta^d)$.⦄ **SP(S)D** consists of symmetric positive (semi)definites.

Sharp Minima. Like us, R3 finds Cor5 counterintutive.[2] SGD's noise consists not of weight **displacements** but of error terms $\nabla l_x - \nabla l$ in the **gradient** estimate; compare Fig5 to [**Ke**]'s Fig1. Say θ is 1D with $l(\theta) = a\theta^2/2$ and training loss $\hat{l}(\theta) = l(\theta) + b\theta$. At $\hat{l}$'s min. $\theta = -b/a$, $l(\theta) = b^2/(2a)$. So for fixed b, sharp min'a ($a \gg 1$) overfit less (demo here). C controls b^2, hence Cor5's $C/2H$ factor. Here, opt'z'n *to convergence* favors sharp min'a ($\star$); cnv'gnce is slow at flat min'a, so flat min'a also overfit little ($\diamond$). (Our small-η assumption precludes H from being so sharp that SGD diverges: we treat $\eta H \ll 1$). Prior work (Pg12Par5, e.g. [**Ke**] and [**Di**]) supports both pro-flat and pro-sharp intuitions. Recognizing η's role in translating gradients to displacements, we account for both ($\star$,$\diamond$), unifying existing intuitions (§4.3). It is a merit that our theory makes such counterintuitive phenomena visible.

ODE. [**Ba**]'s LemA.3 specializes LemKey. In our terms, [**Ba**]'s Thm3.1 computes η^2 weight displacements using *fuzzless* diagrams (noiseless ≡ cumulants vanish ≡ fuzzy diagrams vanish); see Tab1 for the leading corrections to [**Ba**] due to noise. Per §A.6 (fix E=B=1), GD displaces θ by $\Delta^l_{GD}(\eta, T) = -T + \binom{T}{2} + o(\eta^2)$. Now, $2 = \nabla^\mu$, whence arises [**Ba**]Pg2's $\lambda R = \eta G^2/4 = /(2 ! \cdot 2)$. (Note: our analysis applies because, via Euler, ODE 'is' k steps of rate-η/k GD ($k \gg 1$)).

Notation. R3 recognizes our expectands as tensor expressions; they are often fully contracted (so scalar) and are always random variables in some $\mathbb{R}^k$. Per R2,R3, we'll disemploy 'Einstein notation' and cite [**Cu**] (+ a new §D) for tensor examples. If advised, we'll also forgo diagrams: e.g. $[a][ab : c : d][bcd]$ for (letters name edges). R3, Pg6Thm2 defines '**non-degenerate**' as '$H \in$ SPD'.

Organization. R2,R3 stress the paper's narrative challenge. We'll arrange the paper into 3 self-contained tracks, each pertinent to a different goal: TrkA [pgs 1-4], for casual readers, will eschew diagrams, theorems, and §1.1/§2.2's heavy notations; illustrate Taylor series via §2.1's proof; identify §3.3's terms; state Cor4 (w/ §B.1's assumptions explicit, w/ PrpA's precision); explain §4.2's curl effect. TrkB [pgs 1-4, 5-12], for seekers of physical intuition, will use TrkA to motivate (and §A.4 to illustrate) §1.1/§2's definitions; relegate §2.2/2.1's LemKey/discussion to §B; add to §2.3.1 a resummation cartoon à la Fig5,7. For space, §C will absorb §4. TrkC [pgs 5-12, 15-45], for our theory's extenders, will include PrpA (per R1) and more explicit statements and arguments throughout.

References. [**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.

---

1. By cl'cal.mech. (CM) of thermal continua, ice cubes have energy=∞ [McQuarrie '97, §1-1]. Still, CM gives insight.
2. i.e.: *that **overfitting*** ($\triangleq l(\theta_T) - l(\theta_0)$ where θ_0 is a min. of l) *has an* η^2 *term greatest when* ηH *has moderate eigenvalues.*