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 $d$th order truncation $L_d(\eta, T)$, a degree-$d$ polynomial in $\eta$ 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 points out that, though Thm 2 controls $\mathrm{LHS}(\eta) \triangleq \lim_{d\to\infty} \lim_{T\to\infty} L_d(\eta, T)$, it is $\mathrm{RHS}(\eta) \triangleq \lim_{T\to\infty} \lim_{d\to\infty} L_d(\eta, T)$ that more interests us. So how do LHS and RHS relate?

**Proposition.** *Assume Apx B.1's setting and suppose that $\nabla l(\theta_\star) = 0$ and that on some 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 $\theta$.* Then *for any initialization $\theta_0 \in V$ in some neighborhood $V$ of $\theta_\star$ and for any homogeneous polynomial $p(\eta)$ (of $\eta$'s dim × dim many components) with exactly one root (at $\eta = 0$): $\lim_{\eta\to 0}(LHS(\eta) - RHS(\eta))/p(\eta) = 0$.*

TODO: prove, discuss, and discuss irrelevance!

**Sharp Minima** — R2 notes that Cor 5 statement — that the amount of 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 $\eta$, greatest when the hessian has moderate curvature with respect to $\eta$ — is counterintutive. Comparing Fig 5 to [**Ke**]'s Fig 1, we note that SGD's natural noise structure is *not* that of *displacements* in weight space; rather, it is that of additive error terms $\nabla l_x(\theta) - \nabla l(x)$ in the *gradient* estimate. Consider a one-dimensional $\theta$ and imagine a quadratic 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 $b^2/(2a)$. Thus, for fixed $b$, sharper minima (larger $a$) overfit less. The gradient covariance $C$ controls $b^2$, explaining Cor 5's $C/2H$ factor. This example suggests that if we optimize to convergence, sharp minima overfit least. But it is near flat minima that convergence is slowest, so for fixed $\eta, T$, we expect overfitting to vanish as the Hessian shrinks. By making explicit $\eta$'s role in translating gradients into displacements, our theory accounts for both effects, thus reconciling [**Ke**]'s pro-flat intuitions with [Di]'s pro-sharp intuitions. We view it as a merit that our formalism makes such counterintuitive phenomena visible.

**Citations** — Thank you for pointing us to Barrett et al.

**Clarity** — Our work uses a convention standard in physics and in high-dimensional statistics.


[**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.