Mistakes in our work. We've found several oversights in our paper. E.g.: our Cor3 fails to state that the computed loss difference between SGD and ODE is the leading-order difference *due to noise*, i.e., that scales with some higher cumulant such as C. Of course, even without noise, there is also a difference due to time discretization, given by 's embeddings into ODE's $T = kT_0$ grid minus its embeddings into SGD's $T = T_0$ grid. For large k, $(h/k)^2\binom{kT_0}{2} \approx h^2T_0^2/2$, so ODE suffers $(T_0^2/2 - \binom{T_0}{2})$uvalue( ) $= (h^2T_0/2)(GHG)$ more loss than SGD for noiseless loss landscapes.

And Tab1's caption should clarify the same point.

The section introducing resummation should say 'spanning 11 timesteps' instead of 'spanning 12 timesteps' in its example.

ALso, the "proposition A" we mentioned in our author response is false. Luckily, our paper doesn't depend on it.

[**Ba**] relates strongly with our work; we plan to discuss it in (the revision's analogue of) §5.1, 3.2.

Locating [Ba] in our theory. [**Ba**]'s Thm3.1 computes order-η^2 weight displacements $\theta_T - \theta_0$ in the noiseless case $l_x = l$. The relevant diagrams are thus those with ≤ 2 edges and that contain no gray outlines. Indeed, noiseless $\implies$ cumulants vanish $\implies$ any diagram that contains one or more gray outlines has a uvalue (and rvalue) equal to zero. So a sum over diagrams is the same as a sum over gray-free diagrams, i.e., over each diagram whose partition (Pg5Def1) is maximally fine.

Per §A.6, we use 'rootless' diagrams, e.g. , . These diagrams look different from ordinary ones because we are computing weight displacements $\Delta_l \triangleq \mathbb{E}[\theta_T - \theta_0]$, not test losses $\mathbb{E}[l(\theta_T)]$. Of course, in the noiseless case, those expectation symbols are redundant. Likewise, in the noiseless case Δ_l is a function only of η, T (and of the loss landscape l and the initialization θ_0); in particular, we may set E, B as convenient. Let's set $E = B = 1$.

GD's displacement. So, we seek rootless gray-free diagrams width ≤ 2 edges. and are the only such. Let's use their uvalues as in Pg36Thm3 to compute $\Delta_l(T, \eta)$. We read off:

$$\text{uvalue}(\ ) = G_\mu \eta^{\mu\nu} = hG \qquad \text{uvalue}(\ ) = G_\mu \eta^{\mu\sigma} H_{\sigma\rho} \eta^{\rho\nu} = h^2(HG)$$

The RHSs of the above concretize to the case that $\eta^{\mu\sigma}$ (in our directionality-aware theory a symmetric bilinear form that takes two covectors and outputs a scalar) is h times the standard dot product and that G, H are represented in standard ways as matrices. The diagrams embed (into an $E = B = 1$ grid that looks like the rightmost grid on Pg18) in T and in $\binom{T}{2}$ many ways, respectively.[1] The $\binom{T}{2}$ arises due to Pg19's time-ordering condition: has one embedding for every pair $0 \leq t < t' < T$, where t is the red node's column and t' is the green node's column.

These embeddings have trivial Aut groups (Pg28Exm5), so any fixed T has a grand total:

$$\Delta_l(T, h) = -hTGhT + (h^2(T^2 - T)/2)HG + o(\eta^2)$$

[1] An embedding of a rootless diagram (e.g.) assigns *every* node to a grid cell. Pg19 decrees that we assign only *non-root* nodes when computing $\mathbb{E}[l(\theta_T)]$; indeed, the root node represents the test-time factor l and thus corresponds to no training point or training step. By contrast, every factor of every term in $\mathbb{E}[\theta_T - \theta_0]$ corresponds to some training point n and training step t. So we assign *all* nodes to grid cells. We'll expand §A.6 to note as much.

[BA]'S REGULARIZER. Since EulerMethod (EM) (simulation time h, k steps) is just GD with $\eta = h/k$, $T = k$, we can use $\Delta_{\tilde{l}}(k, h/k)$ to predict EM's behavior —and hence ODE's behavior— on a loss $\tilde{l}$. For k huge and η tiny (in a way that depends on k), $\Delta_{\tilde{l}}(k, h/k)$ is close to

$$\star(h) = -h\tilde{G} + (\tilde{H}\tilde{G})h^2/2$$

(I.e., $\tilde{l}$ analytic $\implies \forall\epsilon\ \exists k_0 \forall k > k_0\ \forall A > 0\ \exists h_0 > h_0 \forall h < h_0$: $\|\Delta_{\tilde{l}}(k, h/k) - \star(h)\| < Ah^2 + \epsilon$.)

To match $\star$ with ordinary GD's one-step displacement $\Delta_l(1, h) = -hG$, we just need $hG = h\tilde{G} - (\tilde{H}\tilde{G})h^2/2 + o(h^2)$; it's enough to set $G = \tilde{G} + (\tilde{H}\tilde{G})h/2$. Recognizing the RHS as a total derivative (as $\nabla(\tilde{G} \cdot \tilde{G}) = 2\tilde{H}\tilde{G}$), we see it's enough that $G = \nabla(\tilde{l} - (h/4)(\tilde{G} \cdot \tilde{G}))$ or:

$$\begin{aligned} l &= \tilde{l} - (h/4)(\tilde{G} \cdot \tilde{G}) \\ &= \tilde{l} - (h/4)(G \cdot G) + o(h^2) \end{aligned}$$

This shows how to turn a loss $\tilde{l}$ (on which we plan to run ODE), into a loss l such that running one GD step on l matches ODE on $\tilde{l}$ to leading non-trivial order. Or how to turn l into $\tilde{l}$. In either case, the key term is $(h/4)(G \cdot G)$ with the appropriate sign.