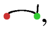
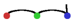
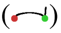
This is a brief note on [**Ba**], the paper suggested by R2. [**Ba**] relates strongly with our work; we plan to discuss it in (the revision's analogue of) §5.1, 3.2.

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 fuzzy outlines. Indeed, noiseless $\implies$ cumulants vanish $\implies$ any diagram that contains one or more fuzzy outlines has a uvalue (and rvalue) equal to zero. So a sum over diagrams is the same as a sum over fuzzless diagrams, i.e., over each diagram whose partition (Pg5Def1) is a union of size-one sets.

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 $E = B = 1$.

Deriving [Ba]'s regularizer. So, we seek rootless fuzzless 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:

uvalue( ) $= G_\mu \eta^{\mu\nu} = hG$ 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:

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

How does Δ_l relate to [**Ba**]'s Thm3.1? We can use it to predict ODE's behavior on a loss $\tilde{l}$. More precisely, we can use it to predict the EulerMethod (EM)'s behavior for any Euler mesh size. Specifically, if the EM divides the integration domain into k chunks, then using EM to integrate an ODE over duration h is the same as running GD over k timesteps with learning rate h/k. In other words, EM's displacement is $\Delta_{\tilde{l}}(k, h/k)$; which for k huge and η tiny (in a way that depends on k) is

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

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 is enough to set $G = \tilde{G} + (\tilde{H}\tilde{G})h/2$. Recognizing the right hand side as a total derivative (as $\nabla(\tilde{G} \cdot \tilde{G}) = 2\tilde{H}\tilde{G}$), we may set

$$G = \nabla(\tilde{l} - (h/4)(\tilde{G} \cdot \tilde{G})) \qquad l = \tilde{l} - (h/4)(\tilde{G} \cdot \tilde{G}) = \tilde{l} - (h/4)(G \cdot G) + o(h^2)$$

This shows us how to turn a loss $\tilde{l}$ (on which we plan to run ODE), and from there obtain a loss l such that running one GD step on l matches ODE on $\tilde{l}$ to leading non-trivial order. [**Ba**]'s result is

[1] Note that an embedding of a rootless diagram like is an assignment of *all* its nodes to grid cells. Pg19's condition is that we assign non-root nodes when calculating $\mathbb{E}[l(\theta_T)]$; intuitively, this is because the root node represents a test-time measurement l and the latter doesn't correspond to any training point or training timestep. Here we compute $\mathbb{E}[\theta_T - \theta_0]$, every factor of every term of which corresponds to some training point n and training timestep t. Thus, we assign *all* nodes to grid cells. We will expand §A.6 to note as much (and analogously for §A.6's other two variants).

actually the mirror image of this — how to start with l and obtain $\tilde{l}$ — and the argument is likewise parallel.

Mistakes in our work.

MISTAKE

References. [**Ba**] D.G.Barrett, B.Dherin. *Implicit Gradient Regularization*. ICLR 2021.