All reviewers & AC Thanks to all the reviewers for their valuable feedback. We would like to begin by clarifying our main contributions:

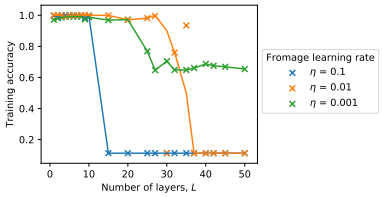
(1) We developed a **new framework for first order optimisation** (Lemma 1). Unlike existing first order optimisation frameworks—such as *mirror descent* and *steepest descent*— where it is not clear which distance metric to use, our framework makes a strong case that the correct distance metric measures the **relative breakdown in gradient**.

(2) We bounded the **relative breakdown in gradient for multilayer perceptrons** (Theorem 1). This provides a principled way to derive deep learning algorithms. Moreover, we believe this application of **perturbation theory to understand deep learning** is highly original and could inspire much followup work.

(3) We derived the Fromage algorithm, which is **admittedly very similar to LARS**. Besides the fact that LARS is susceptible to numerical overflow (Figure 2), Fromage and LARS have very similar performance. Whilst we acknowledge this, we believe that it **does not detract from our main contributions**.

Reviewer 1 Thank you for your valuable comments.

(1) The inset figure shows the performance of Fromage for training MLPs of varying depth using three learning rates. The figure shows that a smaller learning rate is needed to train deeper MLPs, as suggested by Lemma 2. The fact that Fromage with $\eta = 0.01$ trained the off-the-shelf networks in Table 1 was a surprising empirical observation.



(2) In the GAN literature, FID score is a standard proxy for sample quality.

(3) Fromage is used in Figure 1 because we could not get SGD to train such deep multilayer perceptrons.

(4) On *dividing out the gradient scale*—this approach (taken by Adam) requires more learning rate tuning than Fromage. *Gradient clipping*—this approach requires tuning both the learning rate *and* the clipping threshold.

Reviewer 2 These are great questions. We shall address each in turn.

(1) For fixed learning rate η, the iterates do "jitter" around a minimum. This is because Fromage always induces a relative change of η in the weights. Convergence is attained by decaying η when the loss plateaus. [If you are wondering how this fits with our optimisation theory, let us again inspect Equation 2. Close to a minimum of the loss function $\mathcal{L}$, the term $\partial\mathcal{L}/\partial f$ will become the main cause of relative change in the gradient $\nabla_{W_l}\mathcal{L}$, but our deep relative trust model neglects this term $\partial\mathcal{L}/\partial f$.]

(2) Fromage and LARS are similar, but **LARS is empirically motivated and lacks theoretical analysis.**

(3) The large constant factor in Theorem 1 is a result of assuming that the *worst case* happens at every network layer. In practice this is too pessimistic, so we neglect the constant factor in modelling assumption 1. It is also plausible that in modern network architectures, conditioning techniques like BatchNorm and skip connections improve the constant factor. This is a good line of inquiry for further research.

(4) RELU takes an input in $\mathbb{R}$ and projects it on to $\mathbb{R}^+$—thus RELU "transmits half its input domain". For a vector of inputs x, a reasonable model is $\|\text{relu}(x)\|_2 = \|x\|_2/2$. This model may be more realistic when BatchNorm is used, since BatchNorm centers the average input to RELU.

(5) To compare our proposed optimisation framework to steepest descent, note that Lemma 1 can be rewritten as:

$$\mathcal{L}(W + \Delta W) \leq \mathcal{L}(W) + g(W)^T \Delta W + \max_{t\in[0,1]} \|g(W + t\Delta W) - g(W)\|_2 \cdot \|\Delta W\|_2.$$

From this perspective, our framework is analogous to steepest descent with distance: $\max_{t\in[0,1]} \|g(W + t\Delta W) - g(W)\|_2 \cdot \|\Delta W\|_2$. The advantage over steepest descent is that our scheme explicitly tells you how to build your distance metric—you must measure or bound the breakdown in gradient $\max_{t\in[0,1]} \|g(W + t\Delta W) - g(W)\|_2$.

Reviewer 3 Thank you for your review.

(1) We present the analogue of Figure 1 for a 16-layer residual network in the inset figure. As can be seen, this model does violate the Euclidean smoothness assumption—the gradient and loss breakdown are both quasi-exponential. But the breakdown is less severe than for the 16-layer network without skip connections—suggesting that skip connections do improve the loss landscape.