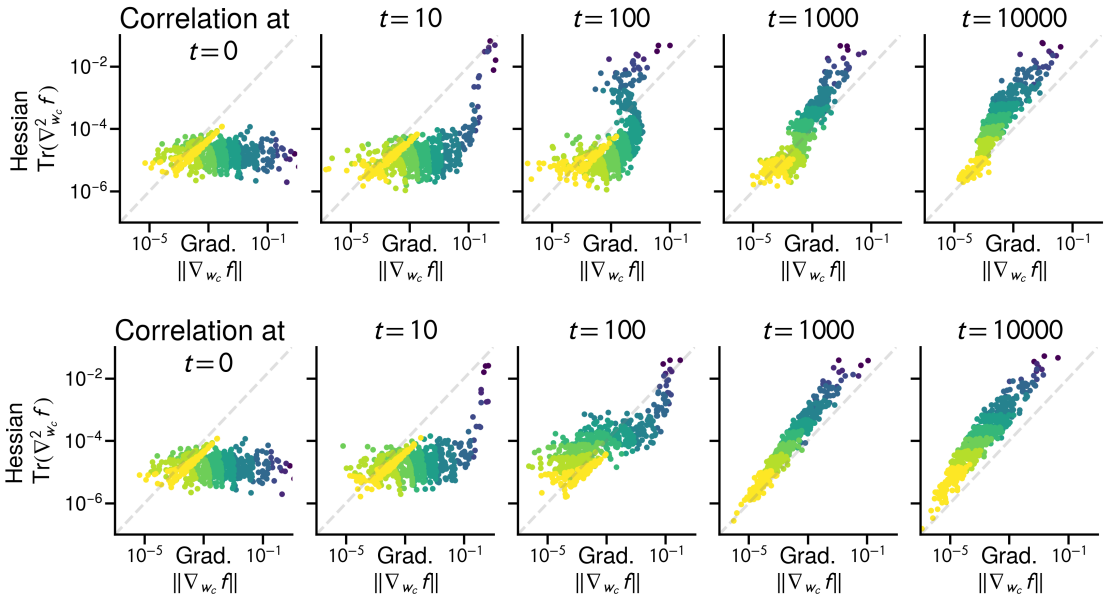

Figure 1: **Correlation between gradients and Hessians on GPT2-Small on WikiText-103.** Evolution of the gradient norm and Hessian trace for each row w_c of the last layer throughout optimization, over the path of GD (top) and Adam (bottom) on the problem of Figure 1.

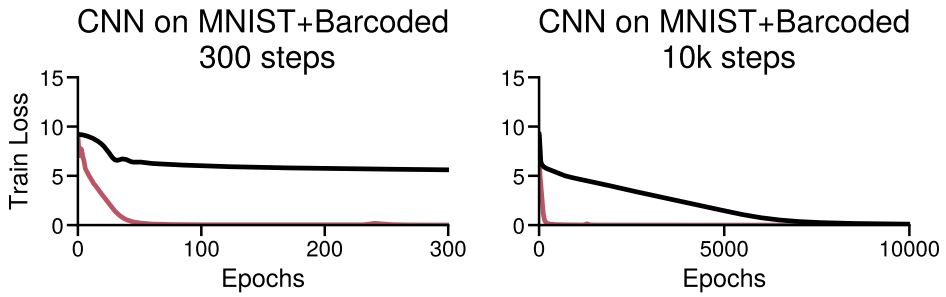

Figure 2: **GD eventually works on Imbalanced MNIST.** We run the same setup as in Figure 2, for longer. Looking at the start of training (left, 300 steps), GD might appear stuck but it can find a better solution if run for longer (right, 10k steps). Adam in red, GD in black, both with momentum.

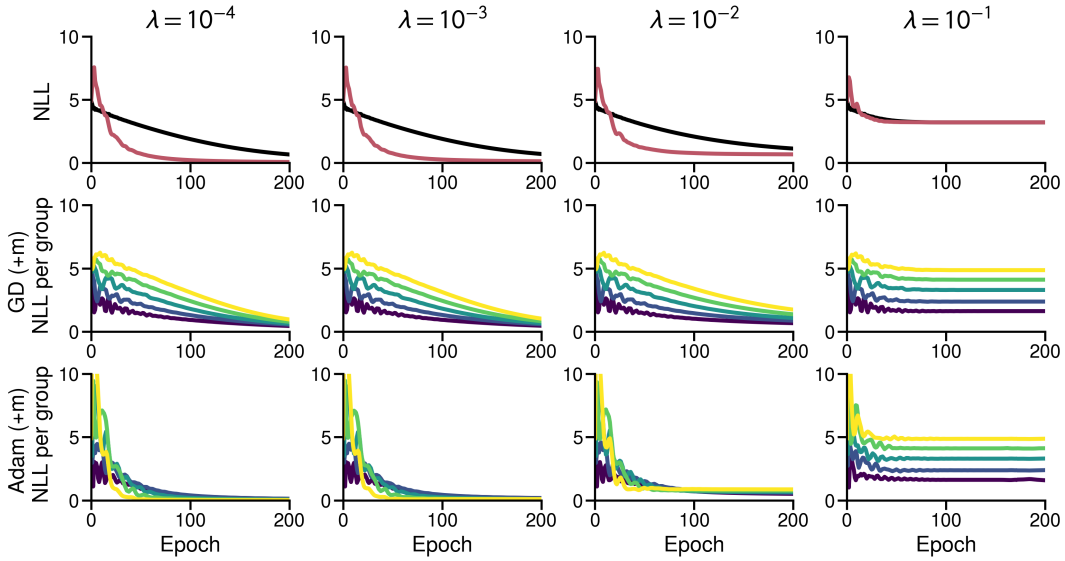

Figure 3: **The separation between GD and Adam still appears when using regularization.** Using varying levels of L_2 regularization λ. The plots show the negative log-likelihood (not including the L_2 penalty).