We thank all reviewers for their helpful reviews. Please see our response below.

Correctness of Theorem 1 proof Thank you R3 for pointing out the mistake in the current proof. The mistake is in the very last step of the proof, where we tried to show $W_1 - W_2^\top = 0$ (lines 593-594). Fortunately, we have the following fix that asserts the correctness of Theorem 1. We hope R3 will update their score based on this revised proof.

From Lemma 2 (line 583), $C = \frac{1}{n}(I - W_2 W_1)XX^\top \in \mathbb{R}^{m\times m}$ is positive semi-definite, and $\Lambda^2 = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ is positive definite. Define $A = W_1 - W_2^\top \in \mathbb{R}^{k\times m}$. We prove below that $A = 0$ follows from line 592 which states

$$\forall v \in \mathbb{R}^k, \quad 0 = v^\top ACA^\top v + v^\top \Lambda^2 AA^\top v \tag{1}$$

Proof. Since $ACA^\top \succeq 0$, we have $\forall v,\ v^\top ACA^\top v \geq 0$. Hence, from (1), $\forall v, v^\top \Lambda^2 AA^\top v \leq 0$. Consider setting $v = e_i$, the i^{th} coordinate vector in $\mathbb{R}^k$ (i^{th} entry is 1, and all others are 0). We must have $e_i^\top \Lambda^2 AA^\top e_i = \lambda_i^2 \|A_i\|_2^2 \leq 0$ (A_i denotes the i^{th} row of A). Since $\lambda_i > 0$, we have $A_i = 0$. Since this holds for all $i = 1, \ldots, k$, we have $A = 0$. $\square$

How to choose the non-uniform regularization parameters (R4) This is a great question. It's indeed difficult to choose an "optimal" set of λ values for the non-uniform ℓ_2 regularization (see below for justifications). However, note that our goal is the analysis, and that the difficulty in choosing the λ values a priori contributes to the argument of weak symmetry breaking by the non-uniform ℓ_2 regularization.

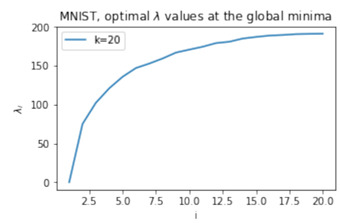

Figure 1: Optimal λ values.

As for why it is difficult to choose an "optimal" set of λ values, note that optimal values at global minima are not optimal in general. At global minima, the optimal λ values can be obtained by solving the $\min\max$ optimization in line 509 (an example of such optimized λ values for MNIST, $k = 20$ is shown in Figure 1). However, this set of λ values concentrate on the larger side, and significantly slow down learning (suboptimal *away from* the global minima). In our experiments, we first make a (rough) estimate of the k^{th} largest data singular value, and linearly space the λ values from 0 to the estimated σ_k value (see Appendix G for details).

Practical utility of the analyzed algorithms (R1, R3) A big part of our motivation was to understand the slowness of learning neural net representations, as opposed to reducing loss — a topic of immense practical importance, as evidenced by the recent flurry of interest on early training effects. Linear autoencoders are one of the few examples where we can determine exactly what representation *ought* to be learned, which makes them a particularly useful model system for understanding convergence of representations.

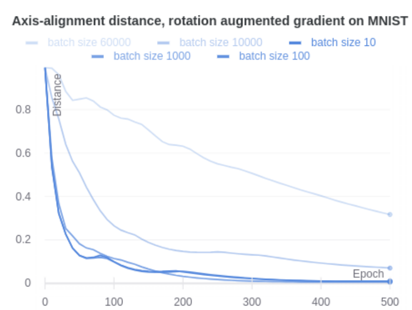

Figure 2: Mini-batch experiment.

Probabilistic interpretations of the non-uniform ℓ_2 regularization and the nested dropout? As pointed out by R2, the probabilistic interpretation of non-uniform ℓ_2 regularization is a straightforward generalization of that studied in Kunin et al. In particular, we can assign a diagonal Gaussian prior to the weights whose precision parameters equal the λ_i value for the corresponding latent dimension. There is a well-known Bayesian interpretation of dropout [Gal and Ghahramani, 2016], and extending this to nested dropout is an interesting direction for future work.

Relation to additional prior work Compared to Wager et al. [2013] (R1), which discusses the connection of dropout and adaptive ℓ_2 regularization in generalized linear models, we work with a different class of models (linear autoencoders) and study a different type of dropout (nested dropout in the latent units). Nevertheless, it is an interesting future direction to investigate the connection between the deterministic nested dropout and common types of regularizers.

Concurrent work [Oftadeh et al., 2020] (pointed out by R3) addresses the identifiability issue in linear autoencoders by proposing a new loss function. The loss function proposed in Oftadeh et al. is a special case of deterministic nested dropout (section 5 of our paper) where the prior $p_B(\cdot)$ is a uniform distribution. Oftadeh et al. show that the local minima correspond to ordered, axis-aligned representations, but do not analyze the speed of convergence under the new loss. Hence, our analysis provides additional insight into their method. We will include the above discussions, as well as the additional citations regarding the connections between linear VAEs and pPCA (R3) in the revised paper.

Mini-batch training for the rotation augmented gradient? (R1) We did additional experiments on MNIST using the rotation augmented gradient, with various batch sizes and $k = 20$ (Figure 2). The results show that the rotation augmented gradient works well with mini-batches. Larger batches improve per-epoch convergence up to a point of diminishing returns, similarly to standard models and algorithms (e.g. Shallue et al., JMLR 2019).

Writing & clarity We thank R2 and R4 for the writing & clarity suggestions, and will address them in the revised paper. Note that the preconditioning of the Adam optimizer is not compatible with the rotation augmented gradient, so only the SGD is relevant for this algorithm (see Appendix G for more experimental details).