First, we would like to thank the reviewers for their time and effort.

Reviewer 1: From your brief review, we cannot tell why you believe there to be no take-home message. We describe the warm-start problem, which you note is both important and understudied. We then describe a simple solution that remedies the issue. The original concern *is* resolved by the shrink and perturb trick.

Settings described in the introduction are just batch online learning scenarios, which we faithfully simulate in our main experiments (e.g. Figure 7). Please see response 3 under reviewer 4. Final-version plots will be made printer friendly.

Reviewer 2: 1. Please see Section 3.1 for a detailed analysis of both learning rate and batch size. Figure 3 shows that, while there exist some hyperparameter values for which warm-started models perform as well as randomly-initialized models, they take as long to train as randomly-initialized models. Tuning these parameters can close the genearlization gap, but only by sacrificing the computational efficiency we would like to see from warm starting.

Separately, it is true that it is commonplace to train models using SGD with a learning rate schedule. To address this concern, we performed an online learning experiment where, starting from 100 CIFAR-10 samples, batches of 10000 points are supplied to the learner in sequence. Using a fresh, warm, or shrink-perturb initialization, we train the ResNet using SGD for 350 epochs using the standard learning rate schedule. As expected, a generalization gap between warm-started and randomly-initialized models still exists and shrink-perturb is able to remedy the problem (Fig A). Similar to warm-start plots in the Appendix, the final copy will include a detailed analysis of scheduled optimization. Models were all trained for the same number of epochs, but randomly-initialized models take much longer to converge.

2. We use Pearson to study the similarity between warm-start and final-solution weights. Other notions of correlation, like cosine similarity, Spearman, and euclidean distance provide similar plots. We will show these in the final version.

3. It may be difficult to tell from Figure 4, but these models that have not been trained much actually perform significantly worse than fully-trained models. This might be more clear in Table 3, where we apply aggressive regularization to prevent overfitting and still observe a generalization gap. We will make Figure 4 more clear in light of this discussion.

Reviewer 3: 1. In the Appendix we have extensive experiments using shrink perturb with multilayer perceptrons, rather than only ResNets. We also experiment with batch normalization and weight decay, and will add RNNs (see **1** below).

Reviewer 4: 1. Indeed, this article focused on experiments with image datasets using ConvNets and MLPs. In response to this note, we performed an online learning experiment using a very different architecture and dataset: a two-layer, bidirectional RNN on the IMDB movie reviews dataset, where the task is to predict whether a movie review is positive or negative. We iteratively supply batches of 500 samples to the model, and compare randomly-initialized, warm-started, and shrink-perturb models in the figure below. We show that there exists a performance gap between randomly-initialized and warm-started models, which the shrink and perturb trick closes (Fig B). The analysis here is truncated due to time constraints, but the trend is clear; we will include a thorough analysis in the final copy.

2. It is not true that this phenomenon does not happen in vision literature. For example, active learning work in image classification requires models to be trained from scratch at each round [7, 8]. Not all articles say this explicitly, but popular deep active learning github repositories also show that random reinitialization after selection is necessary [9,10]. Fine tuning on ImageNet works because the source dataset is large and the target datasets are, comparatively, significantly smaller. Please see Section 4.2 which discusses the pre-training setup in detail. We show that when the target dataset is large, it is often better to initialize models from scratch than it is to pre-train on a source dataset. Figure 9 shows that shrink-perturb is effective in this setting as well: models initialized with shrink-perturb perform as well or better than either randomly initializing or pre-training, regardless of whether pre-training is helpful.

3. We believe there might be some confusion here. While we do "toy" two-phased experiments for illustrative purposes in Table 2.1 and Figures 3-5, our primary results are for full online learning scenarios. These experiments are like Figure 7, where samples are iteratively provided to the learner in batches of 1000. Each time the model receives new data, it is appended to the current training set and is trained to convergence before receiving the next batch. These experiments are consistent with what is usually done in active and online learning work. Also note that large tabular figures, like figure 8, also show this fully-online setting. Here, to concisely report results, we are showing only the final accuracies and train times, i.e., after dozens of rounds of online learning. That is, the plots in Figure 7 correspond to a single column of Figure 8. This is likewise true for all of the similar appendix figures. We will make this more clear in the final version.

4. The warm-start problem is not induced by overfitting. Figure 4 shows that if we do not train to convergence, we still observe a generalization gap. As mentioned above, Table 3 also shows that when overfitting is avoided, by applying regularization, a generalization gap is still present. Also note that the shrink-perturb trick is able to efficiently resolve the generalization gap present in regularized models as well, as shown in Appendix 6.2.5 and 6.2.6.