EARLY PERIOD OF TRAINING IMPACTS ADAPTATION FOR OUT-OF-DISTRIBUTION GENERALIZATION: AN EMPIRICAL STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Prior research shows that differences in the early period of neural network training significantly impact the performance of in-distribution (ID) data of tasks. Yet, the implications of early learning dynamics on out-of-distribution (OOD) generalization remain poorly understood, primarily due to the complexities and limitations of existing analytical techniques. In this work, we investigate the relationship between learning dynamics, OOD generalization under covariate shift and the early period of neural network training. We utilize the trace of Fisher Information and sharpness, focusing on gradual unfreezing (i.e., progressively unfreezing parameters during training) as our methodology for investigation. Through a series of empirical experiments, we show that 1) changing the number of trainable parameters during the early period of training via gradual unfreezing can significantly improve OOD results; 2) the trace of Fisher Information and sharpness can be used as indicators for the removal of gradual unfreezing during the early period of training for better OOD generalization. Our experiments on both image and text data show that the early period of training is a general phenomenon that can provide Pareto improvements in ID and OOD performance with minimal complexity. Our work represents a first step towards understanding how early learning dynamics affect neural network OOD generalization under covariate shift and suggests a new avenue to improve and study this problem.

1 INTRODUCTION

Deep neural networks have achieved remarkable results on in-distribution (ID) data of a task they trained on but often performed poorly on out-of-distribution (OOD) data under input distribution shifts. OOD performance is critical for real-world applications, such as training on clean images or text but inferencing on noise-corrupted data (Hendrycks & Dietterich, 2019; Michel & Neubig, 2018), data obtained from different time periods (Lazaridou et al., 2021; Yao et al., 2022), across languages or domains (Wang et al., 2021; Talman & Chatzikyriakidis, 2019; Liu et al., 2022; Koh et al., 2021; Gulrajani & Lopez-Paz, 2021). Inadequate generalization to OOD settings is a key issue limiting the robustness and reliability of these models.

Prior research observed that variations in the early period of training have a significant impact on the model's ID performance (Golatkar et al., 2019; Achille et al., 2019; Mosbach et al., 2021; Fort et al., 2020) across scenarios including unimodal and multimodal settings when training from scratch, performing parameter-efficient fine-tuning, or using federated learning. The observation of such a period in diverse applications suggests that the early period of learning is generally important for neural network training (Kleinman et al., 2024), drawing parallels to biological phenomena like the critical learning period in animals (Achille et al., 2019; Kleinman et al., 2023).

In particular, intervening during the early period of training can significantly impact ID generalization at the end of training. Training techniques, such as adjusting optimization hyperparameters (e.g., weight decay, learning rate, or dropout; Golatkar et al. 2019; Jastrzebski et al. 2021; Mosbach et al. 2021; Liu et al. 2023b), using data augmentation (Golatkar et al., 2019; Liu et al., 2023c), or adding noise to weights (Frankle et al., 2020), impact learning dynamics early on and can significantly improve or degrade ID results depending on when they are applied or removed. Despite extensive

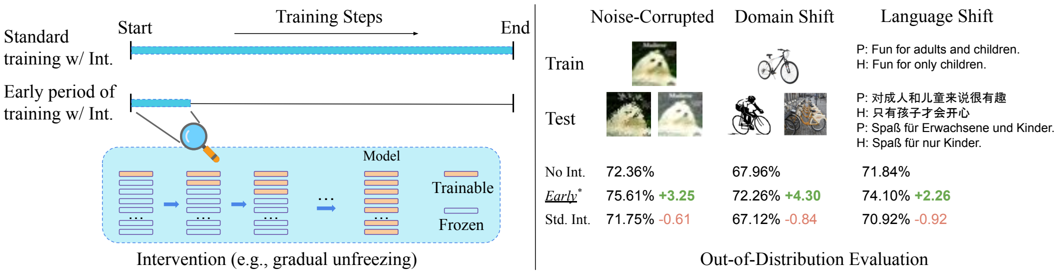


Figure 1: (Left) Interventions during the early period of training are applied for a much shorter time. (Right) Impact of intervention in the early period of training on OOD performance across diverse settings (CIFAR10, Krizhevsky 2009; Hendrycks & Dietterich 2019; Office-Home, Venkateswara et al. 2017; XNLI, Conneau et al. 2018). * indicates optimal OOD results (§5.1).

studies on early learning dynamics and ID generalization, to the best of our knowledge, the impact of the early training period on *OOD* generalization remains unexplored.

In this work, we focus on the impact of the early period of training on OOD generalization, specifically under the common input distribution shift (i.e., covariate shift, encompassing clean to noisy inputs, language, and domain shifts, etc.). We conduct a series of empirical investigations to explore this effect from a previously unexplored perspective – trainable parameters – by using gradual unfreezing (Howard & Ruder, 2018) to intervene in the early period of training (Figure 1). This method is a simple instance of broader training approaches with selective trainable parameters (Kumar et al., 2022; Lee et al., 2023), proven effective in adaptation for OOD generalization. We investigate changes in commonly used metrics to study generalization, namely Fisher Information and loss sharpness (Jastrzebski et al., 2021; Foret et al., 2021; Kwon et al., 2021; Zheng et al., 2021) during the early period of training, exploring their roles in shaping OOD generalization. Through targeted case studies, we demonstrate how leveraging the dynamics of the early period in training can be an effective strategy for "seizing" the moment for generalization across various scenarios.

To the best of our knowledge, we show for the first time that intervening through trainable parameters (i.e., gradual unfreezing) in the early period of training, can significantly enhance OOD generalization under covariate shift in various settings. Our results indicate that sharpness and Fisher Information metrics, though they may not be directly predictive of OOD generalization, can be used as indicators to optimize the timing of intervention removal for better OOD results. We validate this finding in both vision and language tasks, showing its ability to achieve Pareto improvements with minimal complexity. Our analysis and empirical evidence reveal new insights into how early learning dynamics impact neural network generalization, particularly under covariate shift, and suggest new avenues for studying OOD generalization.

2 Related Work

Early period of neural network training. Under the standard usage of the term generalization (in-distribution, where training and testing data are assumed to be from the same distribution), prior work (Golatkar et al., 2019; Achille et al., 2019) shows that the early period of training of neural networks exhibits a "critical learning period" when trained from scratch. Regularization and interventions applied in this critical period affect final task results.

Jastrzebski et al. (2021) indicates that when learning with a lower learning rate, Fisher Information exhibits an "explosion" in the early period of training which impedes ID generalization. Applying regularization to the trace of Fisher Information alleviates the negative impact of the high Fisher Information. Liu et al. (2023c) shows the termination of MixUp (Zhang et al., 2018) early in training and switching to standard empirical risk minimization helps with better ID generalization. You et al. (2020); Frankle et al. (2020) shows that even winning "lottery tickets" emerge in the early period of training with large learning rates. The critical learning period is also found in many other settings, such as in multimodal models (Kleinman et al., 2023), in linear models (Kleinman et al., 2024), in transformers (Mosbach et al., 2021) and federated learning (Yan et al., 2022). However, these works only focus on ID generalization, neglecting the challenges of OOD generalization.

Prior work (Yang et al., 2024; Qiu et al., 2024) investigates into how models adapt to spurious correlations (which is a distinct form of OOD problem compared to the covariate shift examined in our work). These studies focus on the early formation of distinct spurious features during training and the mitigation strategies. However, the impact of trainable parameters (an increasingly important area in light of recent advancements in parameter efficiency and dynamic architectures) remains underexplored.

Kumar et al. (2022); Lee et al. (2023); Liu et al. (2023a) find that training different parts of a model at different times can alter learning dynamics and achieve better OOD results. Encouraged by these findings, we use gradual unfreezing (Howard & Ruder, 2018, a very simple form of training parts of a model at different times) as the main investigative tool in this paper. We focus on two key advancements: 1) more general settings (e.g., training from scratch, fine-tuning), and 2) the characterization of the early period of training and its relationship to OOD generalization.

Fisher Information, sharpness and generalization. Fisher Information has been studied in many prior works such as Chaudhari et al. (2017); Martens & Grosse (2015) to investigate and improve optimization behaviour. Similarly, sharpness is another popular metric used to study optimization behaviour and its relationship to generalization.

Jastrzebski et al. (2017) found a correlation between sharpness and the ratio of learning rate to batch size, which impacts generalization. Jiang et al. (2020); Dziugaite & Roy (2017); Neyshabur et al. (2017) provide theoretical backing for generalization error using sharpness-related measures and empirically show a correlation with generalization. While prior work believes that flatter (less sharp) minima in the loss landscape lead to better generalization in neural networks (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017; Izmailov et al., 2018; Cha et al., 2021), there have been debates on whether sharp minima (such as a high largest eigenvalue of the training Hessian, λ_{max}) imply poor generalization (Dinh et al., 2017) and demonstrate the limits of λ_{max} in explaining ID generalization (Kaur et al., 2023). Andriushchenko et al. (2023) demonstrate that adaptive sharpness is an unreliable metric for OOD generalization in the final solution.

Current research primarily examines the loss landscape at convergence to understand ID generalization. However, the role of Fisher information and sharpness metrics during early training and their relationship to final OOD generalization remains unclear.

3 Preliminaries

We utilize Fisher Information (Fisher, 1925) and sharpness to analyze the training process. Below, we outline the specific metrics used in our experiments.

3.1 Fisher Information Matrix (FIM)

Let x be the inputs and y be the labels of a dataset D. Given a neural network parameterized by w with an output distribution $p_w(\cdot|x)$ for input x, the Fisher Information is defined as:

$$F(w) = \frac{1}{|D|} \sum_{x \in D} \mathbb{E}_{\hat{y} \sim P_w(\cdot|x)} \left[\nabla_w \log p_w(\hat{y}|x) \nabla_w \log p_w(\hat{y}|x)^T \right]. \tag{1}$$

Note that $\hat{y}$ are sampled from $p_w(\cdot|x)$ and not equal to y in general.

Fisher Information reflects the local curvature and measures the amount of information with respect to network parameters, i.e., how sensitive the network predictions are to the small changes in its parameters (Amari & Nagaoka, 2000). A higher value of an element of $F(w)$ indicates that a small change in the corresponding network parameter results in a significant change in the output, which can be interpreted as a "sharper" loss landscape.

Estimating the full $F(w)$ is generally expensive. Prior work shows that the trace of the Fisher Information, $\texttt{tr}(\mathrm{F})$, correlates well with the full Fisher Information when used in real applications to capture signals during the learning process (Achille et al., 2019; Jastrzebski et al., 2021; Sung et al., 2021, inter alia). $\texttt{tr}(\mathrm{F})$ is defined as

$$\texttt{tr}(\mathrm{F}) = \frac{1}{|D|} \sum_{x \in D} \mathbb{E}_{\hat{y} \sim p_w(\cdot|x)} ||\nabla_w \log p_w(\hat{y}|x)||^2. \tag{2}$$

3.2 Sharpness

Let $\mathcal{L}_{\mathcal{D}}(w) = \frac{1}{|D|}\sum_{(x,y)\in D} \log p_w(y|x)$ be the loss over training datasets D, of a neural network parameterized by w, and δ be a small perturbation drawn from a noise distribution, such as a Gaussian distribution $\mathcal{N}(0, \rho^2 diag(c^2))$. The definitions of average and worst-case sharpness are (Foret et al., 2021; Kwon et al., 2021; Andriushchenko et al., 2023; Hochreiter & Schmidhuber, 1997):

$$S^{\rho}_{avg} = \mathbb{E}_{\delta\sim\mathcal{N}(0,\rho^2 diag(c^2))}\mathcal{L}_{\mathcal{D}}(w-\delta) - \mathcal{L}_{\mathcal{D}}(w), \tag{3}$$

$$S^{\rho}_{worst} = \max_{\|\delta\odot c^{-1}\|_p \,\leq\, \rho} \mathcal{L}_{\mathcal{D}}(w-\delta) - \mathcal{L}_{\mathcal{D}}(w), \tag{4}$$

where ρ is a radius parameter of the noise, c is a vector in the parameter space along which sharpness is measured and $\odot c^{-1}$ is element-wise multiplication.

Sharpness metrics measure how the loss changes with respect to small changes to model parameters.[1] While both the Fisher Information and sharpness are used for investigating loss landscapes and generalization, they offer different views (parameter space vs. loss) of the training process.

3.3 Gradual Unfreezing

Gradual unfreezing (Howard & Ruder, 2018) progressively increases the number of trainable parameters (i.e., unfreeze, layer-by-layer) of a neural network from the top to the bottom of a network at a fixed interval of training steps, k (i.e., the unfreezing interval). In this paper, we use a modified formulation of gradual unfreezing (Liu et al., 2023a), where we progressively unfreeze “blocks” of parameters during the early period of training top-down (a block of parameters can range from a single layer to several consecutive layers). In our experiments, we use the namespace of the parameters used in standard implementations to determine blocks. See Appendix B for the algorithm. This method, along with the top-down unfreezing, is chosen as the analysis tool due to its proven effectiveness in achieving state-of-the-art performance across various transfer learning settings (Howard & Ruder, 2018; Kumar et al., 2022; Liu et al., 2023a; Reinhardt et al., 2024).

4 Experimental Setup

We study three experimental settings in this work, covering a diverse set of tasks and scenarios including training from scratch then inference with noise-corrupted images, fine-tuning a pre-trained model for domain generalization, and parameter-efficient fine-tuning for language shift generalization. All experimental results are averaged over 6 runs (for MNIST due to high variances) or 4 runs (all other datasets) and only ID data is used for model selection. See Appendix C and Appendix D for details on the evaluation datasets and hyperparameters.

Training from scratch, noise-corrupted input shift. In this setting, we train a ResNet18 (He et al., 2016) from scratch using the MNIST (LeCun et al., 1998), CIFAR10 (Krizhevsky, 2009), or CIFAR100 (Krizhevsky, 2009). For OOD evaluation, we use the corrupted corresponding evaluation datasets, MNIST-C (Mu & Gilmer, 2019), CIFAR10-C (Hendrycks & Dietterich, 2019) and CIFAR-100-C (Hendrycks & Dietterich 2019, averaging results across corruption types and severities. ID evaluation is done on the original test sets.

Fine-tuning from a pre-trained model, domain shift. Here, we fine-tune an ImageNet (Deng et al., 2009) pre-trained vision transformer (ViT, Wu et al. 2020). We use two popular domain shift datasets, namely Office-Home (Venkateswara et al., 2017) and DomainNet (Peng et al., 2019) for evaluation. We use a single source-domain for training, evaluating all other domains that are not part of the training for Office-Home. For DomainNet, we train on the three domains with the least data (for efficiency reasons) and evaluate on the test sets of all other domains that are not the same as the training domain (i.e., Domain$_{\text{train}}$ $\in$ {Sketch, Infograph, Clipart}, Domain$_{\text{test}}$ $\in$ {Sketch, Infograph, Clipart, Real, Painting, Quickdraw}, see Appendix C).

[1] The sharpness can be negative.

Parameter-Efficient Fine-Tuning (PEFT) with a pre-trained model, language shift. We also conduct experiments using a language transformer. Since pre-training and fine-tuning are common for adapting foundational models, we examine the cross-lingual transfer (train with English data, test with other languages) task using PEFT with the LoRA (Hu et al., 2022) adapters. Here, ID data refers to English task data (for training and validation), while OOD data are in other languages. We train with SQuAD (Rajpurkar et al. 2016, English, question and answering task) and MNLI (Williams et al. 2018, English, natural language inference task), and evaluate on XQuAD (Artetxe et al., 2020), MLQA (Lewis et al., 2020) and XNLI (Conneau et al., 2018). We use XLM-RoBERTa (Conneau et al., 2020) as the pre-trained multilingual transformer backbone.

Learning dynamics metrics. We use $\rho = 0.01$ to calculate the sharpness (both average-case and worst-case) with 15 noise samples (it is computationally expensive to use a larger number of noise samples), and $L2$ norm for the worst-case sharpness. We normalize the $\mathtt{tr}(\mathrm{F})$ by the number of trainable parameters. We use the Auto-PGD algorithm (Croce & Hein, 2020) as implemented in Andriushchenko et al. (2023) (we refer the readers to the original papers for details) for computing worst-case sharpness, as it is a hyperparameter-free estimation method.

5 Impact and time-sentisitivity of early interventions on out-of-distribution generalization

5.1 Timing is critical for removing interventions

Here, we investigate how unfreezing interval k affects ID and OOD results. For all experiments in this section, the smallest k is 1 (unfreeze a block of parameters every one batch update) and the largest k is determined by equally dividing the total training steps among all blocks (Howard & Ruder, 2018; Raffel et al., 2022; Kumar et al., 2022; Lee et al., 2023). We show the relative change in the test results (by subtracting the test results using gradual unfreezing to the results using standard training).

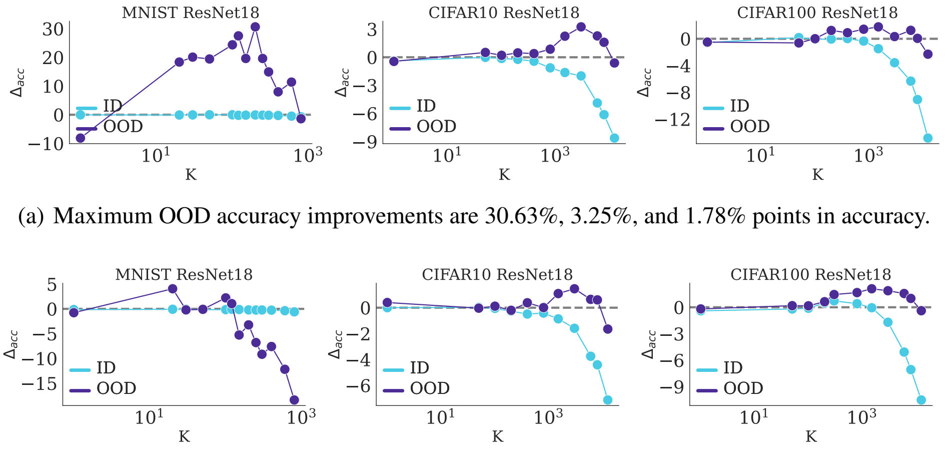
Training from scratch, noise-corrupted input shift. Figure 2 shows the relative change in test results compared to standard training. Gradual unfreezing of trainable parameters significantly impacts OOD generalization (i.e., noise-corrupted images) as early as after a single batch of data, and this effect is particularly pronounced in simpler datasets like MNIST. Extending the unfreezing interval during training initially has minimal impact on ID performance, but later leads to a significant decline, especially at a faster rate with CIFAR. The observed deterioration in ID performance over extended unfreezing intervals mirrors trends from early-stage training interventions and aligns with previous findings (Golatkar et al., 2019; Achille et al., 2019) using other interventions.

The influence of gradual unfreezing on OOD results serves as evidence for the importance of early training periods for OOD generalization. Notably, gradual unfreezing reveals a trade-off between ID and OOD performance for CIFAR datasets, with a brief window where OOD results improve before a sharp decline in ID performance. These results suggest that there may be a critical range of k during early training where ID performance remains stable and OOD performance improves. This time-sensitive observation persists over different learning rates (Figure 2(b)) and model depths (Figure 7 in Appendix F).

Fine-tuning from a pre-trained model, domain shift. Figure 3(a) presents the relative change in domain generalization performance when fine-tuning a pre-trained ViT on single source-domain data. Consistent with previous observations on noise-corrupted input images, results on DomainNet and Office-Home both exhibit a time-sensitive nature in parameter training. [2] Notably, there is also a specific period during training where domain generalization results improve significantly (+2.72% points in accuracy for DomainNet and +4.30% points for Office-Home) with minimal impact on ID evaluation results (DomainNet).

PEFT with a pre-trained model, language shift. We continue to observe a consistent pattern in OOD results (Figure 3(b)) with the prior two scenarios. However, the ID performance for SQuAD also shows improvements. In particular, unfreezing around 1000 and 1600 steps obtain high improvements on average test F1 scores (+2.25% on XNLI and +1.73% on SQUAD). Similar to prior observations, both ID and OOD results are poor when unfreezing occurs later in the training process.

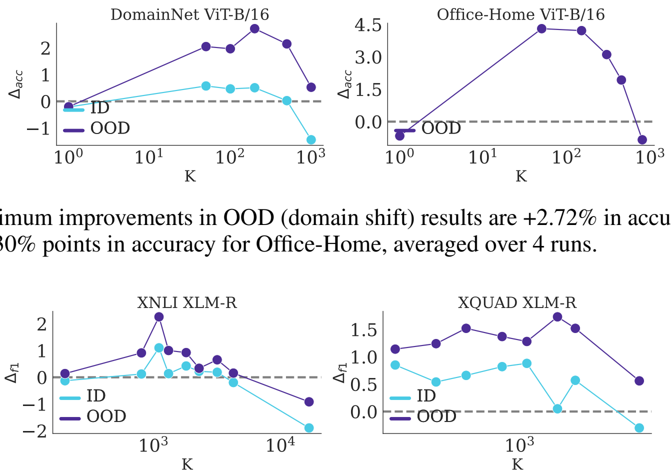
[2] Since there is no official test set for Office-Home, the ID evaluation results are omitted.



(a) Maximum OOD accuracy improvements are 30.63%, 3.25%, and 1.78% points in accuracy.

(b) Maximum OOD accuracy improvements are 4.06%, 1.46%, and 2.10% points in accuracy, using 1/10th of the learning rate as in sub-figure (a).

Figure 2: Changes in ID and OOD (noise-corrupted images) evaluation results when unfreezing parameters at different times (i.e., k) highlight the early training period's impact on OOD generalization. Δ_{acc} is calculated by subtracting gradual unfreezing results from standard training. The x-axis is in the log scale. Each data point on the plot is obtained by averaging over 6 runs for MNIST and 4 runs for CIFAR datasets (a total of 166 experiments per subfigure).



(a) The maximum improvements in OOD (domain shift) results are +2.72% in accuracy for DomainNet and +4.30% points in accuracy for Office-Home, averaged over 4 runs.

(b) The maximum improvements in OOD (language shift) results are +2.26% points on XNLI and +1.73% points on SQUAD in F1, averaged over 4 runs.

Figure 3: Changes in ID and OOD evaluation results when unfreezing parameters at different times (i.e., k) for domain shift (vision) and language shift (text) with pre-trained transformers. Δ is calculated by subtracting the gradual unfreezing results from standard training, averaged over 4 runs (168 experiments in total for subfigure (a) due to single source-domain training and 68 experiments for subfigure (b)). The x-axis is in the log scale.

Summary of findings. The early period of training impacts adaptation to OOD data under covariate shifts. Intervening during this period with gradual unfreezing leads to a time-sensitive trade-off between OOD and ID performance. While factors like data quality or sophisticated learning objectives could influence OOD performance, our results suggest that early, well-timed intervention on trainable parameters also significantly influences models' eventual performance, making this phase a key target for improving OOD results at minimal complexity.

5.2 Discussions

Why does gradual unfreezing help OOD? Kumar et al. (2022, linear-probing then fine-tuning) proposes a method that aligns the classification head (while other parameters kept frozen) with ID data early in training to prevent feature distortion during fine-tuning, leading to better OOD generalization. We suspect that gradual unfreezing could be exploiting a similar mechanism during the early period of training when fine-tuning from a pre-trained model.

Let $\mathrm{g_b}$ denote the gradient of a mini-batch in the training set $b \in \mathcal{B}$. Let $\hat{\mathrm{g}}$ denote the full-batch gradient for the training set. Define the (average) *gradient similarity* at each training step by $\mathrm{GS} \triangleq \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} \frac{\mathbf{g_b}\cdot\hat{\mathbf{g}}}{\|\mathbf{g_b}\|\|\hat{\mathbf{g}}\|}$ where $(\cdot)$ denotes vector multiplication. Tracking GS during training, we find that when training from scratch, GS is higher when using gradual unfreezing than standard training during the early period of training. The difference in GS disappears after the early period (Figure 4 shows GS for the classification head, additional layers in Appendix H.1). This suggests that gradual unfreezing could help to better align early mini-batch gradients to the full-batch gradient. Improving alignment could prevent overfitting to specific mini-batches and reduce learning spurious features, especially early in training, where such features can have lasting deleterious effects. To verify, we conduct additional experiments using the WaterBirds dataset (Sagawa et al., 2020, commonly used for spurious correlation study). We find that gradual unfreezing indeed improves worst-group accuracy over standard training (see Appendix F.4).

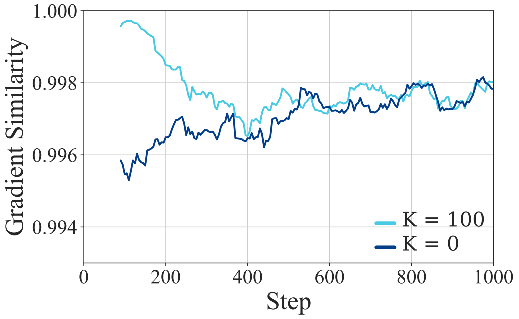


Figure 4: Gradient similarity (mini-batch vs full-data) for the classification head of a ResNet18 trained with CIFAR10. The mini-batch gradient is more similar to the full-data gradient in the early period of training when gradual unfreezing is applied (K=100) compared to standard training (K=0).

Other interventions during the early period of training. Beyond gradual unfreezing, we found that other interventions also exhibit the time-critical nature of training for OOD generalization. Currently, the list of interventions includes learning rate warm-up and delaying the application of a regularizer that minimizes $\mathrm{tr}(\mathrm{F})$ (Jastrzebski et al., 2021). We refer the reader to Appendix F.2 and F.3 for details and additional results. While the gain in OOD generalization for other interventions is less significant than restraining trainable parameters (i.e., through gradual unfreezing), these additional cases indicate that the time-critical nature of removing/applying intervention for OOD generalization is a general phenomenon. We will investigate them in detail in the future.

6 Can we use learning dynamics to “seize” the early period of training for OOD Generalization?

6.1 Learning dynamics analysis

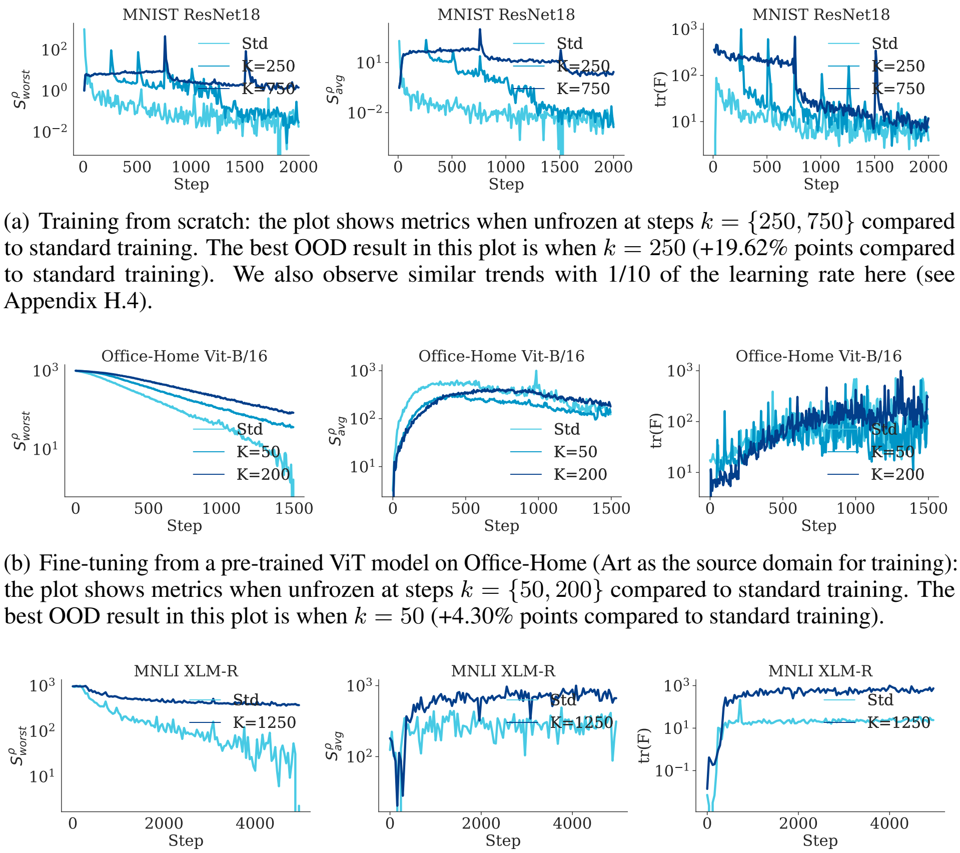
To analyze the characteristics of the early period of training with gradual unfreezing, we examine the learning dynamics using the three metrics described in §3.

Figure 5(a) and Figure 5(c) show the learning dynamics for the early period of training from scratch or with PEFT (in both cases, only randomly initialized parameters are updated). We see that by initially freezing and subsequently gradually unfreezing the trainable parameters, we induce higher Fisher Information and S^{ρ}_{avg}, S^{ρ}_{worst} at the beginning of training compared to standard training. In general, the longer we withhold parameters, the higher the level of sharpness and $\mathrm{tr}(\mathrm{F})$ we can sustain. Unfreezing parameters reduce these metrics.

Figure 5(b) (domain shift, fine-tuning a pre-trained backbone) shows some inconsistency in the metrics initially. However, as parameters are withheld longer, sharpness and $\mathrm{tr}(\mathrm{F})$ sustain. Note that, unlike the previous cases, this experiment directly fine-tunes a pre-trained model without adding randomly initialized parameters.

While S^{ρ}_{avg}, S^{ρ}_{worst}, and $\mathtt{tr}(\mathrm{F})$ differ in definition, they are all sensitive to the early period of training.[3] We identify a pattern consisting of two phases: 1) an initial phase of rapid change (e.g., before the first 50-100, 1000 or 2000 steps in the three subfigures in Figure 5 respectively), and 2) a subsequent stabilization phase where the rate of change of the metric decreases.

Summary of findings. *1)* Gradual unfreezing alters learning dynamics during the early period of training, as measured by the metrics in §3. *2)* The inconsistent learning dynamics across metrics when fine-tuning a pre-trained backbone suggest that sharpness alone does not reliably predict OOD generalization in the modern transfer learning setup. This points to the need to develop new theoretical metrics in different OOD scenarios. Empirically, low sharpness during *early training period* does not guarantee optimal OOD results (e.g., GU increases sharpness early on when training from scratch, yet yields better OOD results empirically), despite the recent success of many sharpness minimization methods for ID. *3)* In standard training (without interventions), metrics exhibit two phases: an initial phase of rapid change, followed by a stabilization phase.



(a) Training from scratch: the plot shows metrics when unfrozen at steps $k = \{250, 750\}$ compared to standard training. The best OOD result in this plot is when $k = 250$ (+19.62% points compared to standard training). We also observe similar trends with 1/10 of the learning rate here (see Appendix H.4).

(b) Fine-tuning from a pre-trained ViT model on Office-Home (Art as the source domain for training): the plot shows metrics when unfrozen at steps $k = \{50, 200\}$ compared to standard training. The best OOD result in this plot is when $k = 50$ (+4.30% points compared to standard training).

(c) Fine-tuning of a text Transformers with LoRA adapters: the plot shows metrics when unfrozen at steps $k = \{1250\}$ compared to standard training.

Figure 5: Learning dynamics with three metrics: $\mathtt{tr}(\mathrm{F})$, $S^{\rho}avg$, and $S^{\rho}worst$. Unfreezing parameters at different steps impact early learning dynamics. The y-axis is log-scaled and normalized between 0 and 1000 for clarity.

[3] See Appendix H for additional learning dynamics with similar trends.

6.2 Do learning dynamics signal the right time for intervention removal?

Learning dynamics criteria for improved OOD generalization. Here, we investigate the learning dynamics of ResNet18 on MNIST when training from scratch. We combined our findings in §5.1 and §6.1 to arrive at the hypothesis that the optimal range of k for achieving the best overall ID and OOD performance is after the initial rapid change of sharpness and $\mathtt{tr}(\mathrm{F})$ but not too long afterward. For instance, the ID results deteriorated rapidly after 800-1000 training steps for the CIFAR datasets in Figure 2, while OOD results were still improving.

This observation suggests that the optimal time to remove intervention (i.e., unfreeze parameters in our case) while maintaining ID results (less than 0.5 points decrease in accuracy) and achieving better OOD results should meet two specific criteria: **1)** after the initial rapid change of sharpness or $\mathtt{tr}(\mathrm{F})$, and **2)** before the stabilization phase progresses too far.

Criterion **2)** is evident across all figures in our prior experiments in §5, as larger values of k consistently degrade both ID and OOD performance. To assess the criterion **1)**, we focus on the MNIST dataset and identify $\hat{k}$ as the earliest ending step of the initial rapid-changing phase among the three metrics (S^{ρ}_{worst}, S^{ρ}_{avg} and $\mathtt{tr}(\mathrm{F})$). We then experiment with 10 different k values, spaced 10 steps apart, both less and greater than $\hat{k}$. For $k < \hat{k}$, we obtained median OOD (ID) accuracies of 52.72 (98.93). For $\hat{k} < k$, we obtained median OOD (ID) accuracies of 53.54 (98.91). This result helps to validate the first criterion since the median OOD accuracy is lower for $k < \hat{k}$ with minimal change in ID accuracy. Together, this analysis suggests that the stabilization of metrics after the initial phase could be a useful signal to determine the optimal ***time*** to introduce new trainable parameters.

Hypothesis validation with a heuristic algorithm. To further validates our hypothesis, we use a heuristic algorithm that satisfies the above-mentioned criteria to determine the stabilization time of the three metrics (we first detect a significant change in metrics, then detect the stabilization point of the metrics, the algorithm is given in Appendix E). The OOD results are then compared with ten random sampled k values per dataset to determine the winning rate (i.e., the percentage of times when the value picked by the algorithm is better than a randomly sampled value).

Training from scratch, noise-corrupted input shift. As shown in Table 1, using a heuristic algorithm is better than performing a random hyperparameter search the majority of the time. In most cases, the degradation of ID accuracy is within 0.5 percentage points. This further validates that the stabilization of S^{ρ}_{worst}, S^{ρ}_{avg} and $\mathtt{tr}(\mathrm{F})$ could signify the removal of interventions (in our case gradual unfreezing) to trade-off a small amount of ID performance for better OOD results. While $\mathtt{tr}(\mathrm{F})$ shows better results, there isn't a clear winning metric for intervention removal due to: 1) the metrics exhibit high noise during training, and 2) the stabilization points determined by different metrics either match or are very close to each other. We defer the exploration of more sophisticated algorithms to future work. Nevertheless, our experiments show that an optimal intervention window exists that can effectively balance good ID and ODD results and the stabilization of sharpness and $\mathtt{tr}(\mathrm{F})$ could signal the right time to remove interventions.

Table 1: Results using the heuristic algorithm to find $\hat{k}$ for gradual unfreezing (GU). Best OOD results are bolded. The algorithm can determine the same value of $\hat{k}$ in different metrics in multiple cases (hence the same results). WR stands for winning rate (OOD). See Appendix E for visualization of $\hat{k}$ overlay on the learning dynamics.

	MNIST RN18	**CIFAR10 RN18**	**CIFAR100 RN18**	**WR**
Method	ID / OOD	ID / OOD	ID / OOD	-
Standard	99.06$_{\pm 0.08}$/33.36$_{\pm 10.81}$	93.32$_{\pm 0.23}$/72.36$_{\pm 0.63}$	71.07$_{\pm 0.36}$/45.10$_{\pm 0.39}$	-
GU$_{S^{\rho}_{worst}}$	98.78$_{\pm 0.15}$/52.48$_{\pm 7.70}$	93.06$_{\pm 0.06}$/72.75$_{\pm 0.84}$	70.68$_{\pm 0.18}$/45.19$_{\pm 0.62}$	60%
GU$_{S^{\rho}_{avg}}$	98.78$_{\pm 0.15}$/52.48$_{\pm 7.70}$	93.02$_{\pm 0.05}$/72.58$_{\pm 0.49}$	70.67$_{\pm 0.20}$/45.35$_{\pm 0.60}$	60%
GU$_{\mathtt{tr}(\mathrm{F})}$	98.91$_{\pm 0.26}$/**54.12**$_{\pm 10.23}$	93.02$_{\pm 0.10}$/**73.56**$_{\pm 0.45}$	70.78$_{\pm 0.31}$/**45.82**$_{\pm 0.56}$	83%

PEFT with a pre-trained model, language shift. Results using $\mathtt{tr}(\mathrm{F})$ to determine $\hat{k}$ for gradual unfreezing are shown in Table 2 ($\hat{k}$ values are in Appendix E, results with S^{ρ}_{worst} / S^{ρ}_{avg} are in Table 6 in the Appendix) and the winning rate is 80%. The ID results are not sacrificed in this experimental setting, hence further pointing towards that the stabilization of sharpness and $\mathtt{tr}(\mathrm{F})$ could signify 'when' to remove intervention in the early period of training for better OOD generalization.

Table 2: Cross-lingual transfer results of standard training and using $\mathtt{tr}(\mathrm{F})$ to determine unfreezing interval $\hat{k}$ for gradual unfreezing (GU), best OOD results are bolded. WR stands for winning rate, averaged over 10 randomly sampled k per training dataset. EM is the exact match score.

	XQuAD		**MLQA**		**XNLI**	**WR**
Method	F1- En/X-ling	EM- En/X-ling	F1- X-ling	EM- X-ling	Acc- En/X-ling	
Standard	$82.96_{\pm 0.49}$/$68.72_{\pm 0.85}$	$71.39_{\pm 0.25}$/$52.64_{\pm 0.66}$	$56.27_{\pm 0.80}$	$40.93_{\pm 0.55}$	$83.17_{\pm 0.29}$/$71.84_{\pm 0.52}$	-
$\mathrm{GU}_{\mathtt{tr}(\mathrm{F})}$	$83.77_{\pm 0.57}$/**70.70**$_{\pm 0.27}$	$72.33_{\pm 0.69}$/**54.40**$_{\pm 0.27}$	**58.47**$_{\pm 0.21}$	**42.31**$_{\pm 0.17}$	$83.36_{\pm 0.13}$/**72.49**$_{\pm 0.42}$	80%

Fine-tuning from a pre-trained model, domain shift. The domain generalization results on DomainNet, using $\mathtt{tr}(\mathrm{F})$ to determine $\hat{k}$ for gradual unfreezing, achieve an accuracy of 37.86%, compared to 35.34% with standard training, with a winning rate of 90%. The complete results are in Table 7 in appendix. Once again, our results align with the trends observed in the previous two cases.

Summary of findings. Our case studies show that the learning dynamics can be effective indicators for determining the optimal timing for intervention removal, with $\mathtt{tr}(\mathrm{F})$ being a slightly better metric overall (when trainable parameters are randomly initialized). While the improvement in OOD performance may be modest, this higlights the connection between learning dynamics and OOD generalization, as well as its potential applications. Since our understanding of the relationship between learning dynamics and OOD generalization is nascent, we hope this initial work will encourage further exploration in this area.

7 Conclusions

In this work, we investigate the early period of training and its impact on OOD generalization under covariate shift. We show that interventions by altering trainable parameters (i.e., progressively changing the number of trainable parameters through gradual unfreezing) during the early period of training improve OOD generalization. This is validated across various vision and language tasks, achieving Pareto improvements with minimal complexity. We emphasize the overlooked role of trainable parameters during the early period of training. Unlike prior work on ID generalization, we empirically observed that sharpness and $\mathtt{tr}(\mathrm{F})$ during the early period of training may not be indicative of the OOD generalization, but can be indicative of "when" to remove interventions.

In light of these findings, it is also essential to consider the broader context of the training strategy studied in this work. The significance of methods that modify only parts of the final model, along with the growing focus on efficient training and fine-tuning — such as freezing parameters (e.g., Adapters, Houlsby et al. 2019; Pfeiffer et al. 2020; Hu et al. 2022) or dynamic architectures (Yoon et al., 2018; Evci et al., 2022; Gu et al., 2021) — cannot be overstated. Our findings contribute to a deeper understanding of the early period of training and OOD generalization, and suggest new research directions, including the development of theoretical metrics to better predict OOD generalization.

References

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL `https://openreview.net/forum?id=BkeStsCcKQ`.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In *International*

Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, pp. 840–902. PMLR, 2023. URL https://proceedings.mlr.press/v202/andriushchenko23a.html.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://aclanthology.org/2020.acl-main.421.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 22405–22418, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcb41ccdc4363c6848a1d760f26c28a0-Abstract.html.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1YfAfcgl.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.747. URL https://doi.org/10.18653/v1/2020.acl-main.747.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020. URL http://proceedings.mlr.press/v119/croce20b.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 2017. URL http://proceedings.mlr.press/v70/dinh17b.html.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017. URL http://auai.org/uai2017/proceedings/papers/173.pdf.

Utku Evci, Bart van Merrienboer, Thomas Unterthiner, Fabian Pedregosa, and Max Vladymyrov. GradMax: Growing neural networks using gradient information. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=qjN4h_wwUO.

Rory A. Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22:700 – 725, 1925.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/405075699f065e43581f27d67bb68478-Abstract.html.

Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The early phase of neural network training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Hkl1iRNFwS.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 10677–10687. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/87784eca6b0dea1dff92478fb786b401-Abstract.html.

Xiaotao Gu, Liyuan Liu, Hongkun Yu, Jing Li, Chen Chen, and Jiawei Han. On the transformer growth for progressive BERT training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5174–5180, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.406. URL https://aclanthology.org/2021.naacl-main.406.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pp. 876–885. AUAI Press, 2018. URL http://auai.org/uai2018/proceedings/papers/313.pdf.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/fb64a552feda3d981dbe43527a80a07e-Abstract-Conference.html.

Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos J. Storkey. Three factors influencing minima in SGD. *ArXiv*, abs/1711.04623, 2017. URL https://api.semanticscholar.org/CorpusID:7311295.

Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4772–4784. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jastrzebski21a.html.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Simran Kaur, Jeremy Cohen, and Zachary Chase Lipton. On the maximum hessian eigenvalue and generalization. In *Proceedings on "I Can't Believe It's Not Better! - Understanding Deep Learning Through Empirical Falsification" at NeurIPS 2022 Workshops*, volume 187 of *Proceedings of Machine Learning Research*, pp. 51–65. PMLR, 03 Dec 2023. URL https://proceedings.mlr.press/v187/kaur23a.html.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical learning periods for multisensory integration in deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 24296–24305. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02327. URL https://doi.org/10.1109/CVPR52729.2023.02327.

Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical learning periods emerge even in deep linear networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Aq35gl2c1k.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang.

WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL http://proceedings.mlr.press/v139/koh21a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *10th International Conference on Learning Representations, ICLR 2019, Online, Apr 25-29, 2022*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5905–5914. PMLR, 2021. URL http://proceedings.mlr.press/v139/kwon21b.html.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomás Kociský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 29348–29363, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f5bf0ba0a17ef18f9607774722f5698c-Abstract.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Yoonho Lee, Annie S. Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=APuPRxjHvZ.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7315–7330, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.653. URL https://aclanthology.org/2020.acl-main.653.

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 7069–7086. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.476. URL https://doi.org/10.18653/v1/2022.emnlp-main.476.

Chen Cecilia Liu, Jonas Pfeiffer, Ivan Vulic, and Iryna Gurevych. Improving generalization of adapter-based cross-lingual transfer with scheduled unfreezing. *CoRR*, abs/2301.05487, 2023a. doi: 10.48550/ARXIV.2301.05487. URL https://doi.org/10.48550/arXiv.2301.05487.

Zhuang Liu, Zhiqiu Xu, Joseph Jin, Zhiqiang Shen, and Trevor Darrell. Dropout reduces underfitting. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22233–22248. PMLR, 2023b. URL https://proceedings.mlr.press/v202/liu23aq.html.

Zixuan Liu, Ziqiao Wang, Hongyu Guo, and Yongyi Mao. Over-training with mixup may hurt generalization. In *The Eleventh International Conference on Learning Representations, ICLR*

2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023c. URL https://openreview.net/pdf?id=JmkjrlVE-DG.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. PEFT: state-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

James Martens and Roger B. Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 2408–2417. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/martens15.html.

Paul Michel and Graham Neubig. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 543–553, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1050. URL https://aclanthology.org/D18-1050.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=nzpLWnVAyah.

Norman Mu and Justin Gilmer. MNIST-C: A robustness benchmark for computer vision. *CoRR*, abs/1906.02337, 2019. URL http://arxiv.org/abs/1906.02337.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5947–5956, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/10ce03a1ed01077e3e289f3e53c72813-Abstract.html.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 1406–1415. IEEE, 2019. doi: 10.1109/ICCV.2019.00149. URL https://doi.org/10.1109/ICCV.2019.00149.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pp. 46–54. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-DEMOS.7. URL https://doi.org/10.18653/v1/2020.emnlp-demos.7.

Guanwen Qiu, Da Kuang, and Surbhi Goel. Complexity matters: Dynamics of feature learning in the presence of spurious correlations. *CoRR*, abs/2403.03375, 2024. doi: 10.48550/ARXIV.2403.03375. URL https://doi.org/10.48550/arXiv.2403.03375.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1), jun 2022. ISSN 1532-4435. URL https://dl.acm.org/doi/abs/10.5555/3455716.3455856.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Max Reinhardt, Gregor Geigle, Radu Timofte, and Goran Glavaš. Improving vision-language cross-lingual transfer with scheduled unfreezing. In Jing Gu, Tsu-Jui (Ray) Fu, Drew Hudson, Asli

Celikyilmaz, and William Wang (eds.), *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pp. 155–166, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.alvr-1.13. URL https://aclanthology.org/2024.alvr-1.13.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Yi-Lin Sung, Varun Nair, and Colin Raffel. Training neural networks with fixed sparse masks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 24193–24205, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cb2653f548f8709598e8b5156738cc51-Abstract.html.

Aarne Talman and Stergios Chatzikyriakidis. Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 85–94, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4810. URL https://aclanthology.org/W19-4810.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5385–5394. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.572. URL https://doi.org/10.1109/CVPR.2017.572.

Kexin Wang, Nils Reimers, and Iryna Gurevych. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 671–688, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.59. URL https://aclanthology.org/2021.findings-emnlp.59.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. doi: 10.18653/V1/N18-1101. URL https://doi.org/10.18653/v1/n18-1101.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *CoRR*, abs/2006.03677, 2020. URL https://arxiv.org/abs/2006.03677.

Gang Yan, Hao Wang, and Jian Li. Seizing critical learning periods in federated learning. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 8788–8796. AAAI Press, 2022. doi: 10.1609/AAAI.V36I8.20859. URL https://doi.org/10.1609/aaai.v36i8.20859.

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2953–2961. PMLR, 2024. URL https://proceedings.mlr.press/v238/yang24c.html.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei Koh, and Chelsea Finn. Wild-Time: A benchmark of in-the-wild distribution shift over time. In *Advances in*

Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/43119db5d59f07cc08fca7ba6820179a-Abstract-Datasets_and_Benchmarks.html.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Sk7KsfW0-.

Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Yingyan Lin, Zhangyang Wang, and Richard G. Baraniuk. Drawing early-bird tickets: Toward more efficient training of deep networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJxsrgStvr.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Yaowei Zheng, Richong Zhang, and Yongyi Mao. Regularizing neural networks via adversarial model perturbation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 8156–8165. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zheng_Regularizing_Neural_Networks_via_Adversarial_Model_Perturbation_CVPR_2021_paper.html.

A LIMITATIONS

There are several limitations to our work. We empirically evaluate three model architectures, and three types of covariate shift, and our selection may not encompass all possibilities and real-world applications. However, we believe our insights have generalizable value. Our empirical observations reveal correlations between changes in training dynamics and OOD generalization. Future research could explore causal interventions to better understand and enhance this relationship.

In our work, we utilize a hand-crafted algorithm to determine metric stabilization time, aiming to show that sharpness stabilization could indicate the time needed for the removal of interventions. However, leveraging training dynamics during the early period of training requires additional computation compared to standard training. While this is not a concern for our study, as our focus is on uncovering insights, more efficient, theory-driven algorithms and metrics should be explored for future practical applications.

B GRADUAL UNFREEZING

Following the notations and algorithm in Liu et al. (2023a), let FORWARD($*$) be the standard forward pass, and BACKWARD($*$) calculates gradients and performs updates for trainable parameters. The modified gradual unfreezing algorithm is in Algorithm 1.

In our experiments, we partition the blocks by their natural namespaces as follows:

ResNet18: The definition block follows the standard implementation of ResNet, with an input convolution layer and a batch norm group together as the additional block. The model parameters are partitioned into 5 blocks, and a classification head.

VGG11: The definition block follows the standard implementation of VGG, with 8 blocks in total. The classification head consists of 3 linear layers with a ReLU function in between. The results are in the Appendix.

Algorithm 1 Gradual Unfreezing

Require: A model's eventual trainable parameters are partitioned into blocks $j \in \{0, \ldots, L-1\}$ parameterized by θ_j, with a task-specific classification head C, and an unfreezing interval k. A set $\mathcal{S}$ of the indices of parameter blocks to unfreeze.

```
Initialize $C, \theta_j$ for all $j$
$\mathcal{S} \leftarrow \{C\}$
$j \leftarrow L - 1$
for $i = 1 \ldots \mathrm{N}$ do
    Sample a data batch $b \sim D$
    if $i \mod k == 0$ and $i \leq kL$ then
        $\mathcal{S} \leftarrow \mathcal{S} \cup \{\theta_j\}$
        $j \leftarrow j - 1$
    end if
    FORWARD($*$)
    BACKWARD($\mathcal{S}$)
end for
```

XLM-RoBERTa + LoRA: The experiment follows Liu et al. (2023a). Each parameter block consists of 2 sets of LoRA adapters added to the query and value of the backbone transformer from the same layer. The LoRA parameters are partitioned into 12 blocks, and a classification head, where the classification head and the last layer of LoRA adapters are trainable initially.

C DATASETS

We provide additional information on the datasets used for evaluation in our experiments.

MNIST-C (Mu & Gilmer, 2019), **CIFAR10-C/CIFAR100-C** (Hendrycks & Dietterich, 2019): This is the noise-corrupted version of the classic image classification datasets MNIST/CIFAR10/CIFAR-100. There are 15 different corruptions in the evaluation dataset, namely frost, fog, gaussian blur, gaussian noise, glass blur, impulse noise, jpeg compression, motion blur, pixelate, saturate, shot noise, snow, spatter, speckle noise, and zoom blur, across 5 severity levels. There are a total of 10 classes each for MNIST-C/CIFAR10-C, and 100 classes for CIFAR100-C.

Office-Home (Venkateswara et al., 2017): This is an image classification task where the images are organized into four different domains: Clipart, Art, Photo and Real. There are a total of 65 classes for classification in this dataset. We considered four domain transfer settings: from Clipart(C) to Art(A)/Photo(P)/Real(R); from A to C/P/R; from P to A/C/R; and from R to A/P/C.

DomainNet (Peng et al., 2019): This is an image classification task where the images are organized into six different domains: Infograph, Sketch, Real, Quickdraw, Painting, and Clipart. There are a total of 345 classes for classification in this dataset. Due to resource constraints and efficiency, we considered three transfer settings (with the least amount of training data): from Infograph(I) to Sketch(S)/Real(R)/Quickdraw(Q)/Painting(P)/Clipart(C); from C to I/S/R/Q/P; and from S to I/R/Q/P/C.

XQuAD (Artetxe et al., 2020): This is a parallel dataset for evaluating cross-lingual question answering, with an evaluation set covering 11 languages (excluding English): Arabic, German, Greek, Spanish, Hindi, Russian, Thai, Turkish, Vietnamese, Chinese, Romanian. The task is the classify the start and end of the answer given a question and a context.

MLQA (Lewis et al., 2020): This is a highly parallel dataset for evaluating cross-lingual question answering. The dataset consists of an evaluation set covering 6 languages (excluding English): Arabic, German, Spanish, Hindi, Vietnamese and Simplified Chinese. The task is the classify the start and end of the answer given a question and a context.

XNLI (Conneau et al., 2018): This is a multilingual natural language inference dataset covering 14 languages (excluding English): French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu. The task is to classify a pair of sentences as having either an entailment, contradiction or neutral relationship.

D HYPERPARAMETERS

The hyperparameters are listed in Table 3 for our experiments. We use the default hyperparameters for the AdamW optimizer, except for the learning rate. All other hyperparameters for the transformer experiments follow Liu et al. (2023a), and we use the HuggingFace PEFT (Mangrulkar et al., 2022) implementations of LoRA. We report results over 6 random seeds for MNIST (due to the high variance in OOD results), and we use 4 random seeds for all other experiments. Standard data augmentation techniques are applied across all experiments. For the CIFAR datasets, we use random cropping and horizontal flipping. For the domain shift datasets, we apply resizing and cropping, horizontal flipping, colour jittering, and grayscaling, following the approach in Gulrajani & Lopez-Paz (2021). The experiments use a single NVIDIA P100, A6000 or A100 GPU depending on the availability.

For our domain shift experiments, we used the total training steps and conducted evaluations across various settings with a single source-domain training (hyperparameter determined following Gulrajani & Lopez-Paz 2021).

For calculating S^{ρ}_{worst} and S^{ρ}_{avg}, we use $L2$ norm and $\rho = 0.01$ with 15 examples. We follow the setup in Andriushchenko et al. (2023) and use the implementation with 2048 data points from the training data (un-augmented when calculating sharpness metrics) for all experiments. We use a batch size of 256, except for SQuAD (the batch size is 32) for calculating all the metrics. The sharpness and $\mathtt{tr}(\mathrm{F})$ are recorded every 10 batches (steps) for all datasets.

Table 3: Hyperparameters used in our experiments.

	MNIST	CIFAR10	CIFAR10	CIFAR100	SQuAD	MNLI	Office-Home	DomainNet
	RN18	RN18	VGG11	RN18	XLM-R	XLM-R	Vit-B/16	Vit-B/16
optimizer	AdamW	SGD	SGD	SGD	AdamW	AdamW	AdamW	AdamW
lr scheduler	const.	const.	const.	const.	linear	linear	const.	const.
lr_d	0.01	0.1	0.15	0.01	0.0005	0.0005	0.00005	0.00005
batch size	128	128	128	128	32	128	128	128
training epochs	10	200	200	200	15	15	-	-
training steps	-	-	-	-	-	-	5000	15000
weight decay	0.01	0	0	0.0005	0.01	0.01	0.01	0.01
momentum	0.9	0	0	0.9	-	-	-	-
LoRA r	-	-	-	-	8	8	-	-
LoRA alpha	-	-	-	-	8	8	-	-
LoRA dropout	-	-	-	-	0.2	0.2	-	-

E Algorithm to Determine the Unfreeze Time

To verify our hypothesis, we use a simple algorithm with a heuristic to determine the unfreezing interval $\hat{k}$, Algorithm 2 presents the flow, τ is 3 or 8 and ϵ is 0.02 (i.e., the percentage of change in the signal is within 2%). The algorithm takes $t_{\Delta_{\hat{S}}}$ as the input, which is the index marking the end of the rapid increase of the signal using a similar logic.

Algorithm 2 Find Stabilization

```
procedure FIND_STABILIZATION_BY_MEAN($\hat{S}$, $t_{\Delta_{\hat{S}}}$, $\tau$, $\epsilon$) ▷ $\hat{S}$ is an array of normalized signal when only the head is trainable, $t_{\Delta_{\hat{S}}}$ is the index marking the end of the rapid increasing of the signal, $\tau$ is the window for smoothing the signals, $\epsilon$ is the threshold in changes of the signal for stabilization.
    if $t_{\Delta_{\hat{S}}} > 0$ then
        $\hat{S} = \hat{S}[t_{\Delta_{\hat{S}}}:]$
    end if
    $\mu_{\hat{S}}$ = moving_average($\hat{S}$, $\tau$)
    $\Delta_{\mu_{\hat{S}}}$ = np.abs(np.diff($\mu_{\hat{S}}$))
    for i, $\delta$ in enumerate($\mu_{\hat{S}}$) do
        if $\delta \leq \epsilon$ then
            index = i                    ▷ The first time where the change is smaller than $\tau$.
            break
        end if
    end for
    if $t_{\Delta_{\hat{S}}} > 0$ then
        index = index + $t_{\Delta_{\hat{S}}}$
    end if
    return index
end procedure
```

Using the heuristic algorithm, we determine the value $\hat{k}$ for experiments in Table 4, where we observe the determined $\hat{k}$ are very close to each other except for VGG with CIFAR10 and XLM-R with SQuAD. All the $\hat{k}$ values are shown visually in Figure 6, overlaying on top of the learning dynamics.

Table 4: Different k determined by Algorithm 2.

Metric	**MNIST RN18**	**CIFAR10 RN18**	**CIFAR100 RN18**	**CIFAR10 VGG11**	**SQuAD XLM-R**	**MNLI XLM-R**
S^{ρ}_{worst}	270	230	260	960	810	780
S^{ρ}_{avg}	270	270	250	1010	1090	720
$\texttt{tr}(\mathrm{F})$	210	260	230	250	1310	790

Table 5: Results using the heuristic algorithm (Appendix E) to find $\hat{k}$ for gradual unfreezing (GU), best OOD results are bolded. The algorithm can determine the same value of $\hat{k}$ in different metrics in multiple cases (hence the same results). WR stands for winning rate (OOD).

	MNIST RN18	**CIFAR10 RN18**	**CIFAR100 RN18**	**WR**	**CIFAR10 VGG11**	**WR**
Method	ID / OOD	ID / OOD	ID / OOD	-	ID / OOD	
Standard	99.06/33.36	93.32/72.36	71.07/45.10	-	88.62/71.63	-
$\mathrm{GU}_{S^{\rho}_{worst}}$	98.78/52.48	93.06/72.75	70.68/45.19	60%	87.69/71.47	40%
$\mathrm{GU}_{S^{\rho}_{avg}}$	98.78/52.48	93.02/72.58	70.67/45.35	60%	87.71/**72.37**	100%
$\mathrm{GU}_{\texttt{tr}(\mathrm{F})}$	98.91/**54.12**	93.02/**73.56**	70.78/**45.82**	83%	88.40/71.86	60%

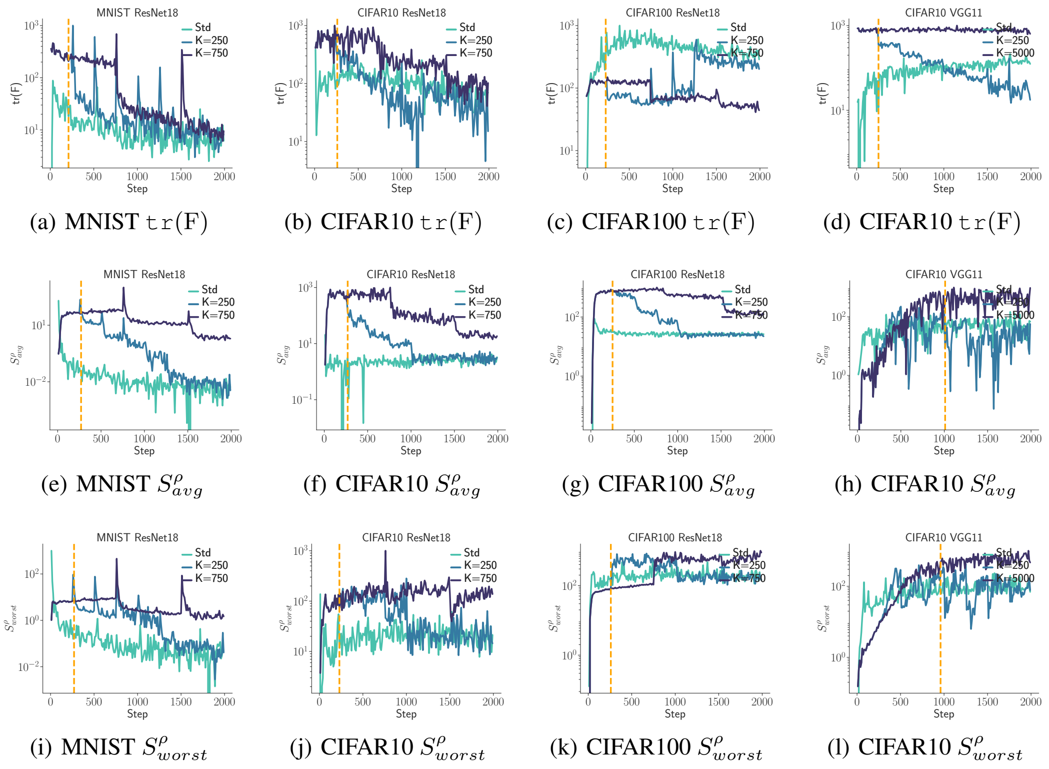
Table 5, Table 6 shows the complete results for 1) training from scratch evaluating on noise-corrupted inputs, and 2) PEFT tuning for cross-lingual transfer. While all results are better than the standard training, empirically, $\texttt{tr}(\mathrm{F})$ is a metric that gives a better winning rate compared to a random hyperparameter search.

Table 6: Cross-lingual transfer results of standard training and using all 3 metrics to determine the unfreezing interval $\hat{k}$ for gradual unfreezing (GU), best OOD results are bolded. WR stands for winning rate, averaged over 10 randomly sampled k per training dataset.

	XQuAD		**MLQA**		**XNLI**	**WR**
Method	F1- En/X-ling	EM- En/X-ling	F1- X-ling	EM- X-ling	Avg- En/X-ling	
Standard	82.96/68.72	71.39/52.64	56.27	40.93	83.17/71.84	-
GU$_{S^{\rho}_{worst}}$	83.78/70.09	72.10/54.17	57.86	42.02	82.83/72.13	45%
GU$_{S^{\rho}_{avg}}$	83.84/70.00	72.12/53.69	58.10	42.03	83.03/72.27	40%
GU$_{\mathtt{tr}(\mathrm{F})}$	83.77/**70.70**	72.33/**54.40**	**58.47**	**42.31**	83.36/**72.49**	80%

Table 7: Domain generalization results of standard training and using $\mathtt{tr}(\mathrm{F})$ to determine unfreezing interval $\hat{k}$ for gradual unfreezing (GU), best OOD results are bolded. WR stands for winning rate, averaged over 10 randomly sampled k per training dataset.

	DomainNet	**WR**
Method	OOD	
Standard	35.34	-
GU$_{S^{\rho}_{worst}}$	**37.95**	90%
GU$_{S^{\rho}_{avg}}$	37.80	90%
GU$_{\mathtt{tr}(\mathrm{F})}$	37.86	90%



(a) MNIST $\mathtt{tr}(\mathrm{F})$ (b) CIFAR10 $\mathtt{tr}(\mathrm{F})$ (c) CIFAR100 $\mathtt{tr}(\mathrm{F})$ (d) CIFAR10 $\mathtt{tr}(\mathrm{F})$

(e) MNIST S^{ρ}_{avg} (f) CIFAR10 S^{ρ}_{avg} (g) CIFAR100 S^{ρ}_{avg} (h) CIFAR10 S^{ρ}_{avg}

(i) MNIST S^{ρ}_{worst} (j) CIFAR10 S^{ρ}_{worst} (k) CIFAR100 S^{ρ}_{worst} (l) CIFAR10 S^{ρ}_{worst}

Figure 6: Learning dynamics with $\hat{k}$ given by Algorithm 2 (vertical line).

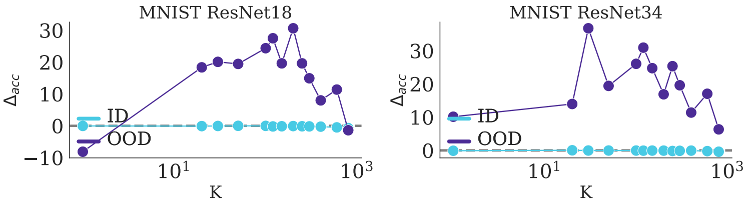


Figure 7: Changes in ID and OOD evaluation results when unfreezing parameters at different times (i.e., k) for ResNet18 and ResNet34. The Δ is calculated by subtracting the gradual unfreezing results from standard training, averaging 6 runs. The x-axis is on a log scale.

Table 8: ID and OOD evaluation results of ResNet18 and ResNet34 on MNIST. k^* is the optimal k that produces the best average OOD results.

	ResNet18	**ResNet34**
	ID/OOD	ID/OOD
Std	99.06/33.36	99.24/10.04
k^*	98.98/63.99	99.20/46.80

F Impact and Time-sensitivity of interventions on out-of-distribution results

F.1 Model Size

We further experimented with a larger ResNet (ResNet34) on the MNIST dataset (chosen for its efficiency as we needed to conduct 86 experiments to generate the subfigure). The time-sensitive nature remains consistent across ResNet models of different depths. Figure 7 shows the results and the numerical results are in Table 8, indicating that this time sensitivity in OOD generalization persists across models with different depths. Interestingly, the larger model (ResNet34) shows no significant differences in ID results while exhibiting greater variation and degradation in OOD results compared to the smaller model (ResNet18).

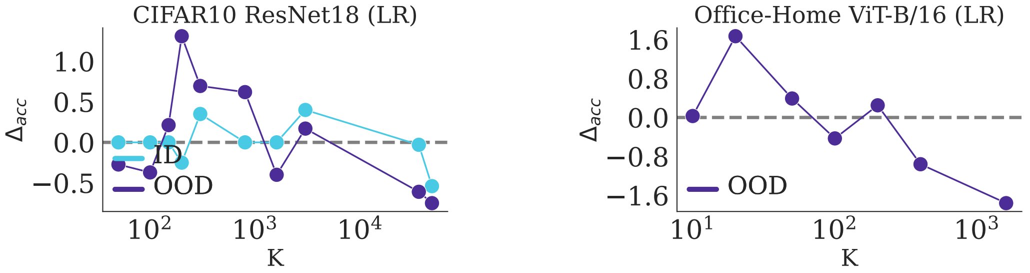
F.2 Learning Rate Warm-up

We also experiment with a simple learning rate warm-up (step function, single step), starting from a reduced learning rate (1/10 or 1/5 of the target learning rate) and switching at a specific time k. We evaluate this approach with: 1) ResNet18 trained from scratch, test on noise-corrupted CIFAR10, starting at 1/10 of the target learning rate, and 2) fine-tuning the pre-trained ViT, test on the Office-Home dataset (domain shift) from 1/5 of the target rate. The results are in Figure 8.

Although the improvements are smaller compared to gradual unfreezing, adjusting the learning rate switch timing increases OOD accuracy by up to +1.31% with ResNet18, with minimal impact on ID performance. The maximum improvements on the Office-Home dataset is +1.67%. This again highlights the importance of timing in applying or removing interventions for better OOD generalization.

F.3 Fisher Penalty

Prior work shows that regularizing $\mathrm{tr}(\mathrm{F})$ can help with ID generalization (Jastrzebski et al., 2021) (training from scratch). Let $\mathcal{J}$ be the original loss, the total loss with Fisher penalty is in Eqn. 5. Following the simple CNN setting in (Jastrzebski et al., 2021, Appendix I.2), we train a simple 4-layer CNN (with one MaxPooling layer, no dropout) and a final fully connected layer of 128 hidden units on the CIFAR10 dataset from scratch with data augmentations. The model is trained for 300 epochs using an SGD optimizer with batch size 128, momentum 0.9, and a learning rate decay of 0.1 after epochs 150 and 225. We use a starting learning rate of 0.001 (a smaller learning rate than the default) and apply the Fisher penalty (FP) with a strength of 0.01 (α) every 10 steps. The model is evaluated with noise-corrupted inputs (i.e., CIFAR10-C) during the test.



(a) The maximum improvements in OOD results (noise-corrupted inputs) are +1.31%.

(b) The maximum improvements in OOD results (domain shift) are +1.67%.

Figure 8: Changes in ID and OOD evaluation results when unfreezing parameters at different times (i.e., k) highlight the early training period's influence on OOD generalization in different settings with learning rate warm-up (step function, single step). The Δ is calculated by subtracting the gradual unfreezing results from standard training, averaging 4 runs. The x-axis is on a log scale.

$$\mathcal{J}_{total} = \mathcal{J} + \alpha * \mathtt{tr}(\mathrm{F})\,. \tag{5}$$

First, in this small learning rate setting, the ID results improved from 84.45 to 85.39 on average over 4 random seeds (compared to no FP) when we applied FP to the training. Then, we experiment with delaying the application of the FP regularizer by k steps, and Figure 9 shows the results compared to applying FP from the beginning of the training. Delaying the application of the FP decreases ID results by a small fraction, but increases OOD results compared to no delay. The best k appears to be between 1000 to 3000 steps (largest OOD increase, smallest ID decrease). Figure 10 shows the learning dynamics with no FP penalty (Std) and with the FP applied with no delay (k=0) or a delay of 2000 steps (i.e., k=2000). The sharpness profile of the k=2000 curve follows a high-to-low trend.

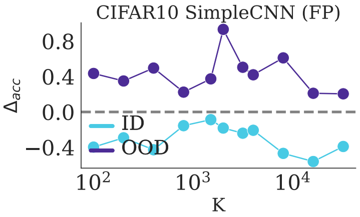


Figure 9: Change in ID and OOD evaluation results when applying Fisher penalty at different times.

The stabilization of sharpness and $\mathtt{tr}(\mathrm{F})$ once again coincides with the period of improved OOD results in Figure 9. This supports our hypothesis that the point at which sharpness and $\mathtt{tr}(\mathrm{F})$ stabilize marks the optimal time to apply a regularizer that reduces sharpness during early training.

We also use the same algorithm (Appendix E) to determine the time for application of FP (i.e., find a time k when the sharpness metrics or FIM stabilizes). The ID/OOD results are 85.04/66.25, 85.22/66.65 and 84.98/65.97, determined using S^{ρ}_{worst}, S^{ρ}_{avg} and $\mathtt{tr}(\mathrm{F})$, respectively (k=1980/1930/2410). As expected, the OOD results are better than applying the FP from the beginning of the training.

F.4 Spurious Correlation Experiment

We hypothesize that the effectiveness of gradual unfreezing on OOD generalization stems from implicitly regularizing learning from possible spurious features present in the dataset. To test this hypothesis, we experiment with a pre-trained ResNet18 (on ImageNet) and the WaterBirds dataset (Sagawa et al., 2020) which the spurious features are known. Training hyperparameters are determined based on Izmailov et al. (2022) (learning rate=3e-3, epochs=100, weight decay=1e-4, batch size=32, SGD optimizer, 4 runs).

In this experiment, the accuracy and worst-group accuracy (WGA) for standard training (Empirical Risk Minimization) were 96.88% and 56.93%, respectively, compared to 96.88% and 58.29% with gradual unfreezing. This result confirms that gradual unfreezing achieves better OOD results by implicitly regularizing spurious correlations.

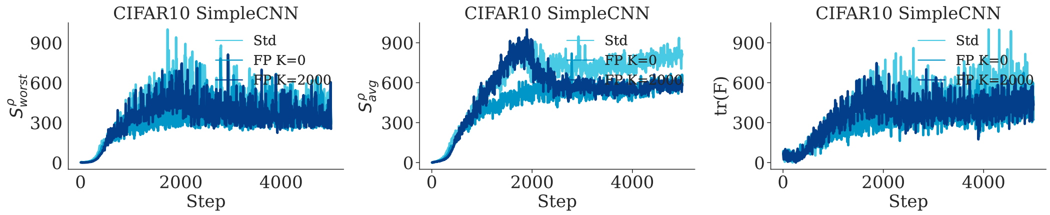


Figure 10: Learning dynamics of a simple CNN model on CIFAR10 with and without the application of the Fisher penalty from Jastrzebski et al. (2021). Std means standard training without using the Fisher penalty. k=0 means the Fisher penalty is applied at the start of training. k=2000 means the Fisher penalty is applied with a delay of 2000 steps. The y-axis is normalized and results are smoothed using a rolling window for better visualization.

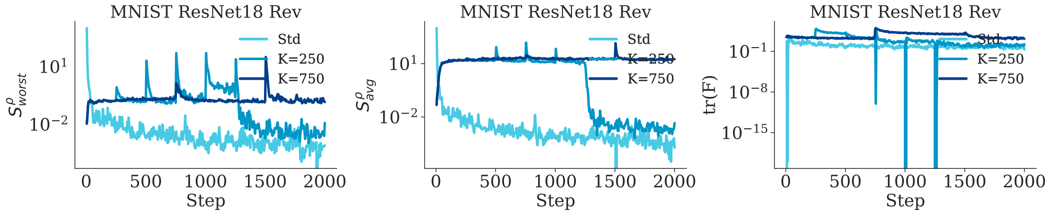


Figure 11: Learning dynamics of a ResNet 18 model trained with MNIST. *Rev* indicates that the unfreezing order progresses from the bottom layers to the top layers. Similar to the trends using top-down order, higher sharpness and $\mathtt{tr}(\mathrm{F})$ are observed.

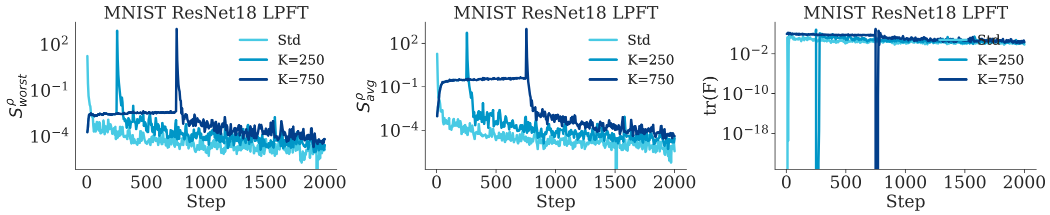


Figure 12: Learning dynamics of a ResNet 18 model trained with MNIST using LP-FT as in Kumar et al. (2022). Similar to the trends using top-down order, higher sharpness and $\mathtt{tr}(\mathrm{F})$ are observed.

G Properties of the Final Solutions and OOD Results

Changing the learning dynamics in the early period of training inevitably results in different final solutions. We plot the final solution's λ_{max} (largest eigenvalue of training data feature), S^{ρ}_{worst} and S^{ρ}_{avg} against the OOD test results in Figure 13 respectively.

While in general the sharpness measures and OOD have negative correlations (i.e., the smaller the sharpness values the better, especially S^{ρ}_{worst} has a consistent negative correlation), they are not always statistically significant (e.g., for MNIST). The learning rate has a big impact on the final solutions' sharpness. Furthermore, such as in Figure 13 (c), we can even attain slightly positive correlations. Our results complement the findings in Andriushchenko et al. (2023), which serve as evidence pointing towards the need for developing robust new metrics and thorough investigation for OOD generalization.

H Additional Learning Dynamics

H.1 Gradient Similarity

Figure 14 illustrates additional gradient similarity (between mini-batch gradients and the full gradient, §5.2) during the early period of training for ResNet18. On average, gradient similarity increases when trainable parameters are constrained. Additionally, higher layers exhibit greater similarity to the full gradient at the beginning of training.

H.2 Feature Rank

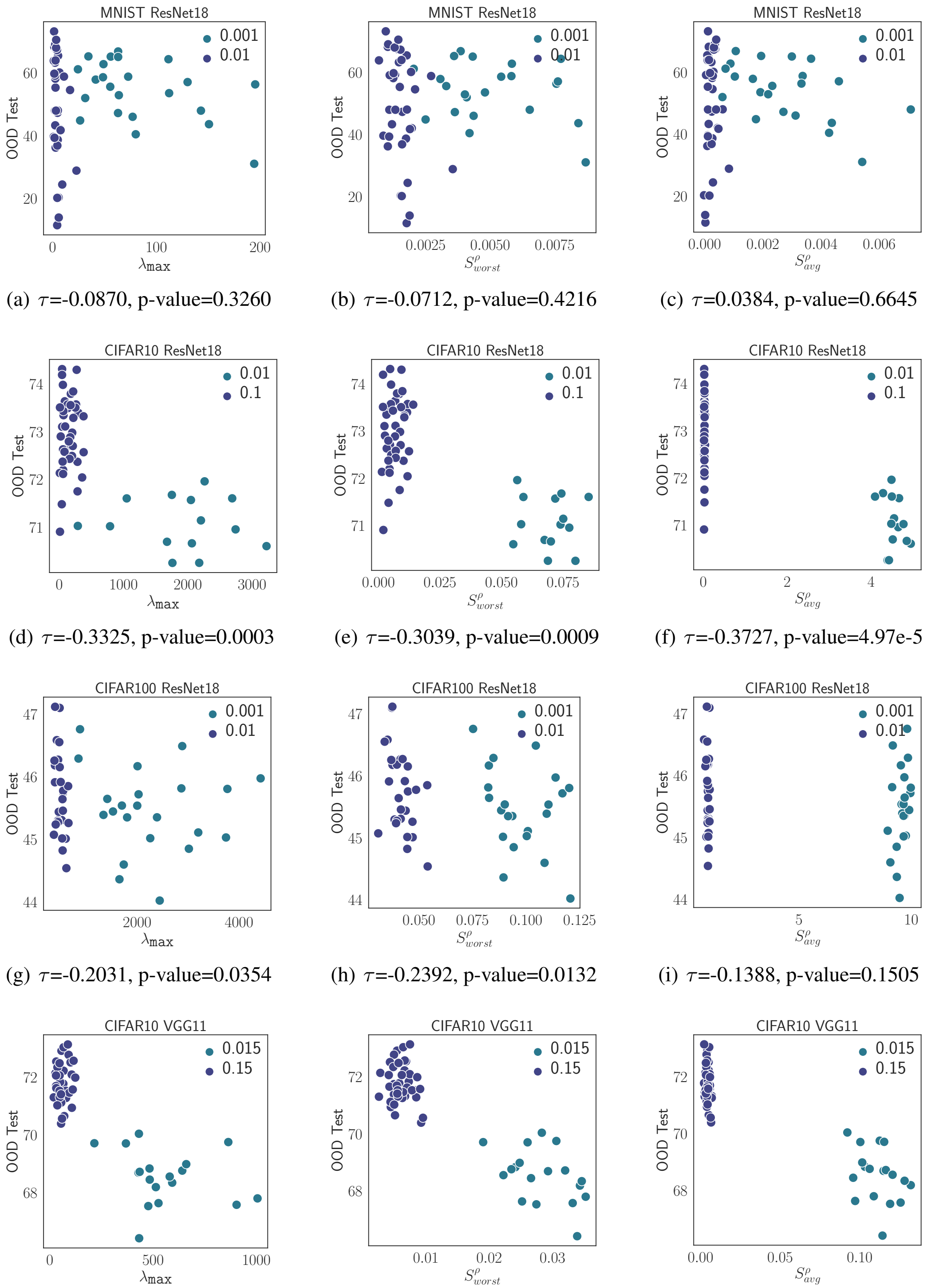
Figure 15 shows the evolution of feature ranks before the classification head for the first 2000 training steps. We observe that standard training typically starts with a lower feature rank, and as training progresses, the feature rank gradually increases. When withholding parameters from training, the feature ranks are high at the beginning of the learning period. As parameters are gradually released, the feature ranks decrease compared to their initial values.

H.3 SQuAD

In Figure 16, we present the learning dynamics for XLM-R with SQuAD in the early period of learning. The learning dynamics show a similar trend as the SQuAD dataset, the S^{ρ}_{worst} value is also negative, and withholding trainable parameters increases the S^{ρ}_{worst} during training based on Eqn. 4 in our main paper.

H.4 Training from Scratch

Figure 17 shows all the learning dynamics in the early period of training using the same learning rate in §D with gradual unfreezing. Figure 18 shows the learning dynamics in the early period of training using 1/10th of the learning rate specified in §D with gradual unfreezing. We observe consistent trends.



(a) τ=-0.0870, p-value=0.3260

(b) τ=-0.0712, p-value=0.4216

(c) τ=0.0384, p-value=0.6645

(d) τ=-0.3325, p-value=0.0003

(e) τ=-0.3039, p-value=0.0009

(f) τ=-0.3727, p-value=4.97e-5

(g) τ=-0.2031, p-value=0.0354

(h) τ=-0.2392, p-value=0.0132

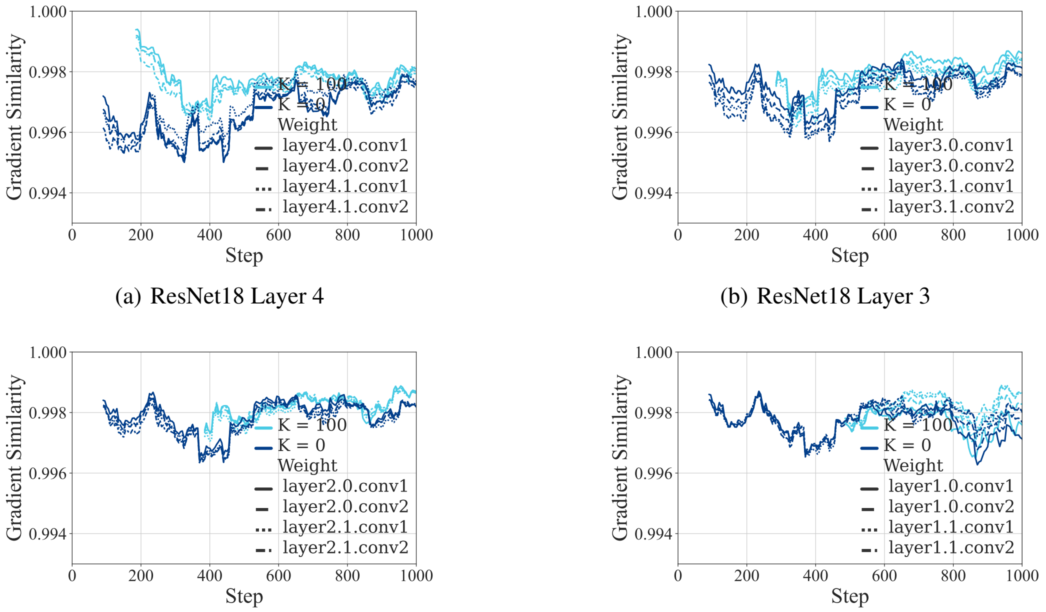
(i) τ=-0.1388, p-value=0.1505

(j) τ=-0.4167, p-value=3.12e-6

(k) τ=-0.4670, p-value=1.74e-7

(l) τ=-0.5254, p-value=4.13e-7

Figure 13: Final feature λ_{max}, S^{ρ}_{worst}, and S^{ρ}_{avg} versus the OOD test results (coloured by learning rate), labelled with Kendall's τ and p-value.



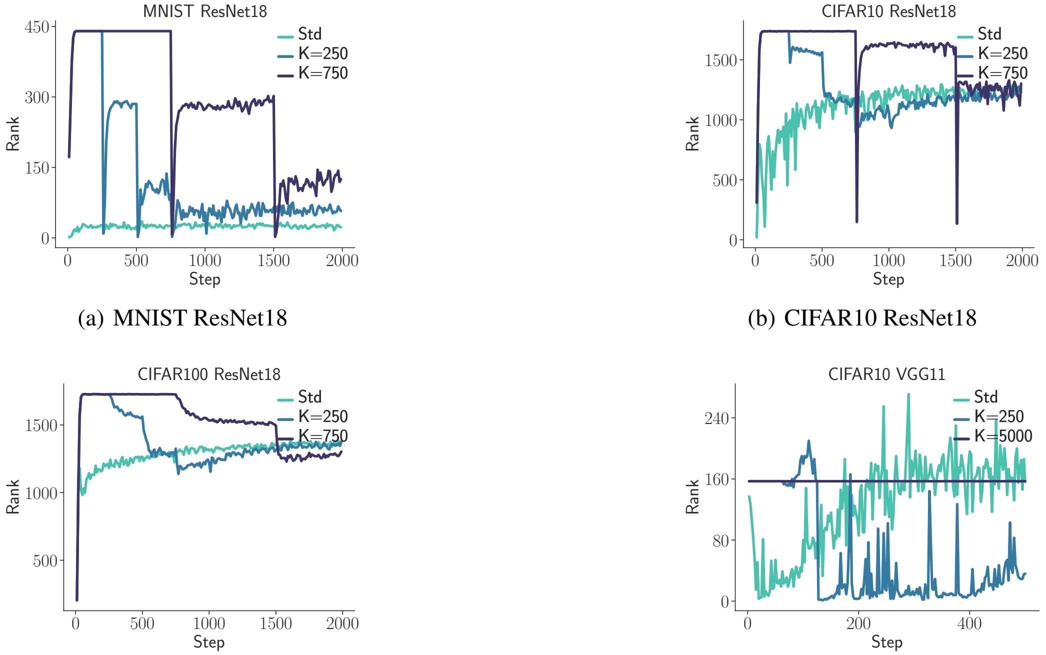
(a) ResNet18 Layer 4

(b) ResNet18 Layer 3

(c) ResNet18 Layer 2

(d) ResNet18 Layer 1

Figure 14: Gradient similarity (mini-batch vs full-data) for the convolution layers in ResNet18 during training with CIFAR10. The mini-batch gradient is, on average, more similar to the full-data gradient when gradual unfreezing is applied (K=100) compared to standard training (K=0). This effect is more pronounced in the higher layers.



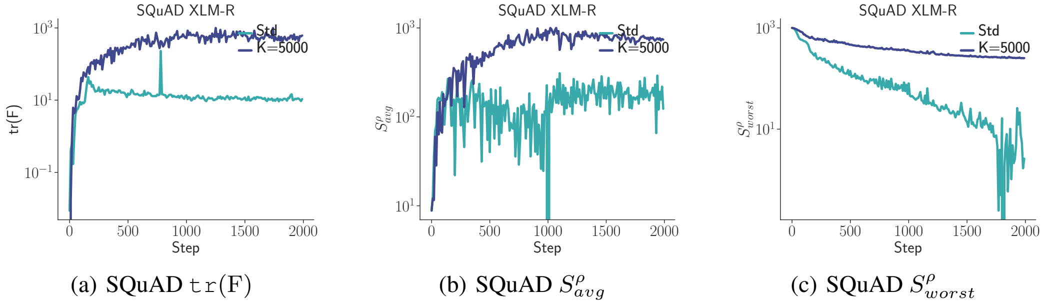
(a) MNIST ResNet18

(b) CIFAR10 ResNet18

(c) CIFAR100 ResNet18

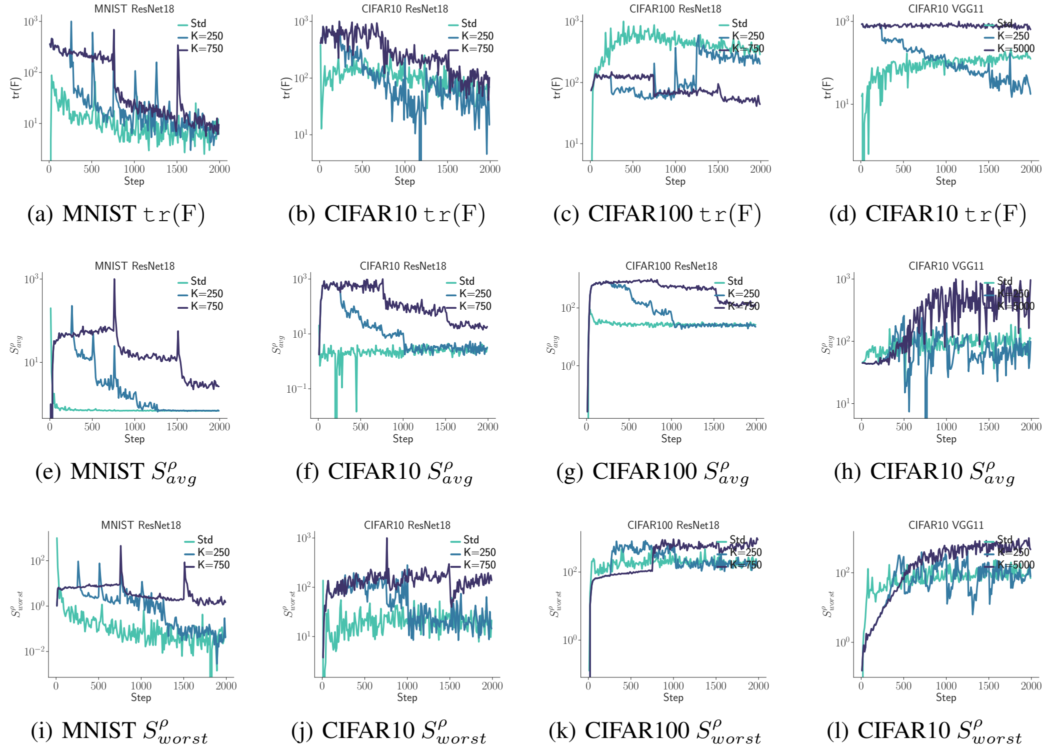
(d) CIFAR10 VGG11

Figure 15: Change of feature ranks before the classification head. The sudden decrease in feature ranks is due to unfreezing the trainable parameters.



(a) SQuAD $\mathtt{tr}(\mathrm{F})$ (b) SQuAD S^{ρ}_{avg} (c) SQuAD S^{ρ}_{worst}

Figure 16: Learning dynamics of XLM-R with LoRA training with SQuAD, y-axis for figures are in the log scale, the original value sharpness value in sub-figure (c) is negative where we take the absolute value before visualization. All values are normalized between 0 and 1000 for visualization.

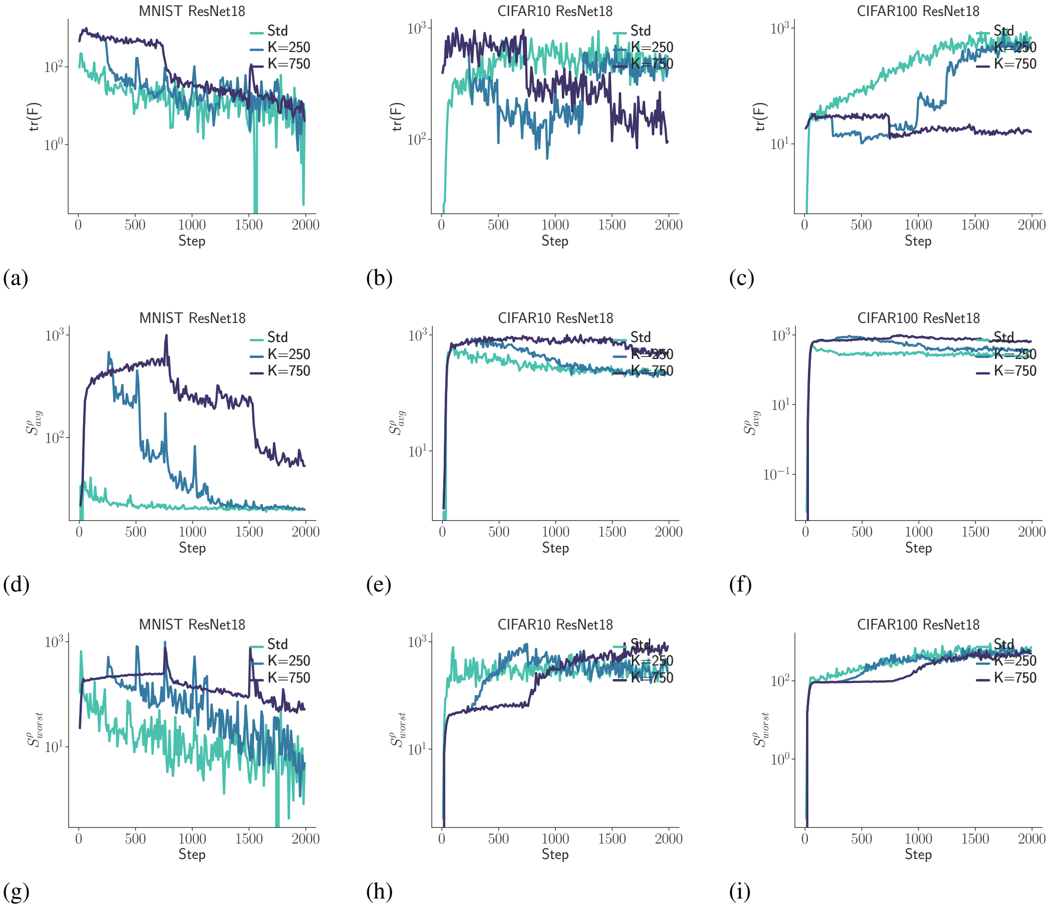


(a) MNIST $\mathtt{tr}(\mathrm{F})$ (b) CIFAR10 $\mathtt{tr}(\mathrm{F})$ (c) CIFAR100 $\mathtt{tr}(\mathrm{F})$ (d) CIFAR10 $\mathtt{tr}(\mathrm{F})$

(e) MNIST S^{ρ}_{avg} (f) CIFAR10 S^{ρ}_{avg} (g) CIFAR100 S^{ρ}_{avg} (h) CIFAR10 S^{ρ}_{avg}

(i) MNIST S^{ρ}_{worst} (j) CIFAR10 S^{ρ}_{worst} (k) CIFAR100 S^{ρ}_{worst} (l) CIFAR10 S^{ρ}_{worst}

Figure 17: Unfreezing parameters at different times affects the learning dynamics in the early period of training (with lr_d). We show $\mathtt{tr}(\mathrm{F})$, S^{ρ}_{avg} and S^{ρ}_{worst} when parameters are unfrozen at steps $k = \{250, 750\}$ for ResNet and $k = \{250, 5000\}$ for VGG, versus standard training. The y-axis uses a log scale and is normalized between 0 and 1000 for visualization.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Figure 18: Unfreezing parameters at different times affect the learning dynamics in the early period of training. We show $\mathtt{tr}(\mathrm{F})$, S^{ρ}_{avg} and S^{ρ}_{worst} when parameters are unfrozen at steps $k = \{250, 750\}$ for ResNet, versus standard training. The y-axis uses a log scale and is normalized between 0 and 1000 for visualization. We use 1/10th of the learning rate specified in §D.