Chaotic Dynamics are Intrinsic to Neural Network Training with SGD

Luis M. Herrmann
Center for Digital Health - AG Roland Eils
Berlin Institute of Health
Kapelle-Ufer 2, 10117 Berlin
luis.herrmann@charite.de

Maximilian Granz
FU BioRobotics Lab
Freie Universität Berlin
Arnimallee 7, 14195 Berlin
maximilian.granz@fu-berlin.de

Tim Landgraf
FU BioRobotics Lab
Freie Universität Berlin
Arnimallee 7, 14195 Berlin
tim.landgraf@fu-berlin.de

Abstract

With the advent of deep learning over the last decade, a considerable amount of effort has gone into better understanding and enhancing Stochastic Gradient Descent so as to improve the performance and stability of artificial neural network training. Active research fields in this area include exploiting second order information of the loss landscape and improving the understanding of chaotic dynamics in optimization. This paper exploits the theoretical connection between the curvature of the loss landscape and chaotic dynamics in neural network training to propose a modified SGD ensuring non-chaotic training dynamics to study the importance thereof in NN training. Building on this, we present empirical evidence suggesting that the negative eigenspectrum - and thus directions of local chaos - cannot be removed from SGD without hurting training performance. Extending our empirical analysis to long-term chaos dynamics, we challenge the widespread understanding of convergence against a confined region in parameter space. Our results show that although chaotic network behavior is mostly confined to the initial training phase, models perturbed upon initialization do diverge at a slow pace even after reaching top training performance, and that their divergence can be modelled through a composition of a random walk and a linear divergence. The tools and insights developed as part of our work contribute to improving the understanding of neural network training dynamics and provide a basis for future improvements of optimization methods.

1 Introduction

In the last decade, the advent of Deep Learning has led to an explosive development of powerful machine learning models capable of solving ever more complex problems. However, while numerous different architectures have been developed, the underlying optimization procedure - Stochastic Gradient Descent (SGD) (Robbins, 2007) and its descendants (Duchi et al., 2011; Kingma and Ba, 2017) - has largely remained the same, with many aspects of its nature not being fully understood yet. The role of second order dynamics is one of those aspects, as are chaotic dynamics of Artificial Neural Network (ANN) training, both being notoriously difficult to investigate due to the high computational cost involved. In this paper, we establish a connection between the two aspects and generate new

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

insights by studying the intersection of these two through a series of experiments conducted on small-sized models trained on natural datasets. Our main four contributions are as follows:

1. By modeling ANN training with SGD as a time-discrete dynamical system, we propose a modified SGD algorithm ensuring non-chaotic training dynamics to study the importance of chaos in ANN training.
2. We find empirical evidence suggesting that directions of negative curvature - and thus local chaos - cannot be removed without hurting the training performance of ANNs.
3. We show empirically that the network dynamics start out diverging exponentially at the beginning of the training but transition asymptotically against polynomial behaviour as the model performance converges.
4. Elaborating on the previous aspect, we show that even as the model training converges, the distance between similarly initialized models continues to grow at a small pace, and that this behaviour can be modelled as a sum of a linear divergence and a random walk.

2 Theoretical preliminaries

2.1 Definition of chaos

A popular saying characterizing the notion of chaos states that 'The flapping of a butterfly's wings can cause a storm on the other side of the world.', sometimes also briefly referenced as the butterfly effect. Formally, a dynamical system is chaotic if it is (1) **sensitive to initial conditions**, (2) topologically transitive and (3) has periodic points of the system which are dense in state space (Skokos, 2009, p.3, Def.1). Although a complete analysis of chaos dynamics of a dynamical system would encompass quantifying all three aforementioned properties, property (1) is the most relevant property for model optimization: Given a small deviation in the parameter state of a model at any point of the optimization, one should hope to obtain a similar solution at the end of the optimization procedure, and by extension a solution that performs similarly well both on training and validation data. Therefore, we use chaotic to refer to the sensitivity to initial conditions.[1]

2.2 Recap: Lyapunov Exponents

A common tool for characterizing chaos are Lyapunov exponents, which measure the evolution of displacement vectors over time. Consider a dynamical system $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{f}(\boldsymbol{\theta}^{(t)})$, where $\boldsymbol{\theta}^{(t)}$ is the system state and $\boldsymbol{f}$ the dynamics function which determines how the system evolves over time. Given an initial, infinitesimally small displacement vector $\delta\boldsymbol{\theta}_0$ s.t. $\boldsymbol{\theta}_0' = \boldsymbol{\theta}_0 + \delta\boldsymbol{\theta}_0$, the evolution of the displacement vector over time is given by the so-called tangent map, typically denoted $d_w\boldsymbol{\Phi}^{(t)}$ or $\boldsymbol{Y}^{(t)}$ in matrix notation, satisfying

$$\delta\boldsymbol{\theta}^{(t)} = \boldsymbol{Y}^{(t)}\delta\boldsymbol{\theta}^{(0)} = \frac{\partial\boldsymbol{\theta}^{(t)}}{\partial\boldsymbol{\theta}^{(0)}}\delta\boldsymbol{\theta}^{(0)}, \tag{1}$$

where the matrix $\boldsymbol{Y}^{(t)}$ is the Jacobian of the dynamics function at time t w.r.t. the state at the initial time step. When the state $\boldsymbol{\theta}^{(t)}$ depends only on the state at the previous time point $\boldsymbol{\theta}^{(t-1)}$ the action of the tangent map can be decomposed through application of the chain rule to

$$\boldsymbol{Y}^{(t)} = \boldsymbol{J}_f(\boldsymbol{\theta}^{(t)})\boldsymbol{Y}^{(t-1)} = \frac{\partial\boldsymbol{\theta}^{(t)}}{\partial\boldsymbol{\theta}^{(t-1)}}\boldsymbol{Y}^{(t-1)}, \tag{2}$$

with $\boldsymbol{J}_f$ being the Jacobian of the dynamics function at time step t. The Lyapunov exponents are defined as the eigenvalues (the Lyapunov spectrum) of the matrix

$$\boldsymbol{\Lambda}^{(\infty)} = \lim_{t\to\infty}\frac{1}{2t}\ln\boldsymbol{Y}^{(t)T}\boldsymbol{Y}^{(t)}, \tag{3}$$

which we will refer to as *Lyapunov matrix* from now on. With this definition, we expect initial system perturbations $\Delta\boldsymbol{\theta}_0$ to evolve according to

$$\Delta\boldsymbol{\theta}(t) \propto e^{\lambda_1 t}\Delta\boldsymbol{\theta}_0, \tag{4}$$

[1] A more rigorous definition of chaos would require the system's invariant set to be bounded.

where λ_1 is the maximum eigenvalue of the Lyapunov matrix, also called **maximum Lyapunov Characteristic Exponent** (mLCE). The respective eigenvector gives us the direction of maximum expansion of an initial displacement as the dynamic system evolves. Thus, the mLCE can be used to characterize three different types of system dynamics:

1. $\lambda_1 > 0$: Nearby trajectories **diverge exponentially** (chaotic).
2. $\lambda_1 = 0$: Nearby trajectories **diverge polynomially** (edge-chaotic).
3. $\lambda_1 < 0$: The distance of nearby trajectories is **upper-bounded**.

Although the mLCE is sufficient to characterize the overall system dynamics, a more fine-grained analysis can be obtained in a natural manner by considering the full eigenvalue spectrum of the Lyapunov matrix, with the biggest eigenpair $(\lambda_1, \boldsymbol{v}_1)$ giving us the "most chaotic" direction of the system evolution, $(\lambda_2, \boldsymbol{v}_2)$ the second-most chaotic, and so on assuming $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$. Thus, for a small initial perturbation $\boldsymbol{\theta}'^{(0)} = \boldsymbol{\theta}^{(0)} + \varepsilon \boldsymbol{v}_i$, we expect the perturbation to evolve approximately as dictated by the respective eigenpair as $\Delta\boldsymbol{\theta}(t) \propto e^{\lambda_i t}\varepsilon \boldsymbol{v}_i$. Along any direction, we can expect the bound

$$e^{\lambda_N(t)t}\varepsilon \leq \Delta\boldsymbol{\theta}(t) \leq e^{\lambda_1(t)t}\varepsilon \tag{5}$$

to hold for sufficiently small ε s.t. the first order approximation

$$\boldsymbol{\theta}'^{(t)} \approx \boldsymbol{\theta}^{(t)} + \boldsymbol{Y}^{(t)}(\varepsilon \boldsymbol{w}) \tag{6}$$

of the training process at time step t, where $\boldsymbol{w}$ is a unit-length perturbation axis, is accurate.

2.3 Curvature and chaos

A neural network evolving according to SGD with learning rate γ (without momentum) and trained with data batches $\boldsymbol{z}^{(t)}$ at time step t can be described by the equation

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} - \gamma \boldsymbol{g}(\boldsymbol{\theta}^{(t-1)}; \boldsymbol{z}^{(t-1)}), \tag{7}$$

where $\boldsymbol{g}(\boldsymbol{\theta}^{(t-1)}; \boldsymbol{z}^{(t-1)})$ is the loss gradient w.r.t. to network parameters $\boldsymbol{\theta}^{(t-1)}$. We show in Appendix A.1 and A.3 that the evolution of the tangent map without and with momentum can be described as

$$\boldsymbol{Y}^{(t+1)} = (\boldsymbol{I} - \gamma \boldsymbol{H}^{(t)})\boldsymbol{Y}^{(t)} \quad \text{and} \quad \boldsymbol{Y}^{(t+1)} = \boldsymbol{Y}^{(t)} - \gamma \left(\sum_{s=1}^{t} \beta^{(t-s)} \boldsymbol{H}^{(s)} \boldsymbol{Y}^{(s)} \right), \tag{8}$$

respectively. This result shows that, discounting the effects of random batch sampling in SGD, the chaotic dynamics depend on the Hessian and thus on the curvature of the loss landscape during training.

2.4 Local chaos and negative Hessian eigenvalue spectrum

Given the connections we established above, one could hope to improve the training performance of SGD by using second-order curvature information to avoid/promote a chaotic evolution of the system, assuming chaos to be detrimental/beneficial for neural network training, e.g. by pruning the directions of maximum chaos from the gradient updates. Unfortunately, the Lyapunov matrix is not practical for this purpose because it considers the chaos over the entire training up to the current time point, so it can only inform us of a chaotic evolution of the system a posteriori. To mitigate this problem, one can instead consider a greedy approach where at any given time step t, one looks at the finite-time Lyapunov matrix after a single time step, i.e.

$$\boldsymbol{\Lambda}^{(t,t+1)} = \frac{1}{2}\ln(\boldsymbol{I} - \gamma \boldsymbol{H})^T(\boldsymbol{I} - \gamma \boldsymbol{H}) = \frac{1}{2}\ln(\boldsymbol{I} - \gamma \boldsymbol{H})^2 \tag{9}$$

The eigenvalues of this matrix are called the **Local Lyapunov Exponents (LLE)** at time step t. Analogously to the LCEs, we say the system is *locally chaotic* in the direction of eigenpair $(\lambda_i, \boldsymbol{v}_i)$ iff $\lambda_i > 0$. Using the fact that $\boldsymbol{\Lambda}^{(t,t+1)}$ has the same eigenvectors as $\boldsymbol{H}$, we can relate the eigenspectrum of the Hessian to local chaotic dynamics as follows:

Theorem 2.1. *Given a neural network with N parameters, let $\boldsymbol{H}$ be the Hessian of the loss w.r.t. network parameters, and let $(\lambda_i, \boldsymbol{v}_i)_{i=1}^N$ be the Hessian's eigenpairs. Furthermore, let us assume that the network is trained using SGD with learning rate γ and no momentum. Then, the network's LLEs indicate locally chaotic training behaviour in the direction of $\boldsymbol{v}_i$ if*

$$\lambda_i < 0 \vee \lambda_i > \frac{2}{\gamma} \tag{10}$$

For a proof see Appendix A.5. Our result leads to the interesting observation that sufficiently large positive eigenvalues (as determined by the learning rate) imply locally chaotic training behaviour in SGD. More interestingly however, negative eigenvalues automatically lead to locally chaotic training behaviour.

2.5 Pruning chaotic updates

Since the eigenpairs $(\lambda_i, \boldsymbol{v}_i)_{i=1}^N$ of $\boldsymbol{\Lambda}^{(t,t+1)}$ give us the directions of local chaos ordered by magnitude of the LLEs, we can remove the locally chaotic directions from the parameter updates in SGD by projecting the update vector onto non-locally-chaotic eigenvectors of $\boldsymbol{\Lambda}^{(t,t+1)}$. More specifically:

Theorem 2.2. *Given a neural network with Hessian $\boldsymbol{H}^{(t)}$, eigenpairs $(\lambda_i^{(t)}, \boldsymbol{v}_i^{(t)})_{i=1}^N$ of $\boldsymbol{\Lambda}^{(t,t+1)}$ and gradient updates $\Delta\boldsymbol{\theta}^{(t)}$, let there be $k \leq N$ chaotic eigenpairs, i.e. $\lambda_1 \geq ...\lambda_k > 0$. Suppose all chaotic components are removed through projection, leading to a modified update vector*

$$\Delta\tilde{\boldsymbol{\theta}}^{(t)} = \boldsymbol{V}^{(t)}(\boldsymbol{V}^{(t)})^T\Delta\boldsymbol{\theta}^{(t)}, \qquad \boldsymbol{V}^{(t)} = \textstyle\sum_{l>k} \boldsymbol{v}_l^{(t)}\boldsymbol{v}_l^{(t)^T}. \tag{11}$$

Then, for training with SGD without momentum, the updated training dynamics are given by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \gamma\boldsymbol{V}^{(t)}\boldsymbol{V}^{(t)^T}\boldsymbol{g}(\boldsymbol{\theta}^{(t)}), \qquad \boldsymbol{Y}^{(t+1)} = (\boldsymbol{I} - \gamma\boldsymbol{V}^{(t)}\boldsymbol{V}^{(t)^T})\boldsymbol{Y}^{(t)} \tag{12}$$

and for sufficiently small variations of the initial parameters, the system is guaranteed not to have exponentially diverging orbits.

Proof is provided in Appendix A.7.

3 Related Work

Since the full Lyapunov eigenspectrum is very expensive to calculate, previous work analyzing chaotic dynamics has so far mostly been limited either to small networks or to the largest mLCE, as investigated for instance by Das et al. (2000). Several investigations of chaos in feedforward ANNs such as those by Li (2019); Feng et al. (2019); Zhang et al. (2021) have focused on sensitivity of outputs w.r.t. the inputs rather than w.r.t. the training dynamics, modelling the network function itself as a discrete dynamical system. Recently, Vogt et al. (2020) have calculated approximations of the Lyapunov exponents in RNNs and found correlations between better generalization (through lower validation loss) and a smaller mLCE. To our knowledge, Sasdelli et al. (2021) are the only ones to have investigated chaos dynamics in ANN training by SGD, but interestingly, they model ANN training by SGD as an approximately time-continuous dynamical system and not more accurately as a time-discrete dynamical system, thus arriving at slightly different theoretical results. While their theoretical work also links the biggest Lyapunov exponent to the most negative eigenpair of the network Hessian, positive eigenpairs as indicators of chaos disappear in their considerations, which is not the case for time-discrete SGD without further assumptions. Unfortunately, they do not investigate how the directions of chaotic evolution in general and negative curvature in particular affect the training process.

There is a considerable body of work on the eigenvalue spectra of ANN Hessians that can be tied into our theoretical findings, including an analysis of the full eigenvalue spectrum of a small 784-2-10 MLP for MNIST digit classification by Sagun et al. (2017). Among other findings, the authors remark a two-phase distribution of eigenvalues with a bulk centered around zero and large positive top eigenvalues (although not sufficiently big to satisfy $\lambda > 2/\gamma$), as well as the **presence of negative eigenvalues throughout the entire training**. Further analysis of the negative spectrum by Alain et al. (2019) has indicated that individual negative eigenvalues on average lead to the biggest improvement

of the loss in neural networks, but that current optimization algorithms appear not to be good at optimizing in directions of negative curvature. This finding was extended by Gur-Ari et al. (2018), who found that the gradient of the Hessian mostly resides in the space spanned by the top Hessian eigenvectors and does not mix with the bulk, suggesting a marginal importance of the negative Hessian eigenpairs for the training dynamics.

To the best of our knowledge, we are the first to use time-discrete modelling of training dynamics by SGD to derive predictions of how curvature affects chaotic dynamics, and the first to explore extensively - using the full Hessian spectrum at every training step - how directions of different curvature affect the dynamics of the training process. This extensive investigation provides us with empirical evidence allowing us to establish a novel connection between chaotic training dynamics and the training performance.

4 Experiments

Taking into account all aspects discussed so far, we can make the following claims:

1. Due to the presence of negative eigenvalues in all training steps and assuming the MLP for MNIST is a representative model for Feedforward ANNs, **ANNs can be expected to have positive LLEs at every training step**.
2. Unless the positive top eigenvalues of the Hessian are sufficiently big to satisfy $\lambda > 2/\gamma$ - which was not the case for the MLP for MNIST investigated by Sagun et al. (2017) - the **chaotic behaviour of ANN training by SGD is exclusively determined by negative eigenvalues in the bulk**.

What remains to be verified experimentally is whether the negative eigenspectrum has little or at worst a detrimental effect on the training performance, as suggested by the findings of Alain et al. (2019) and Gur-Ari et al. (2018), or if it is actually benificial to the performance of ANN training.

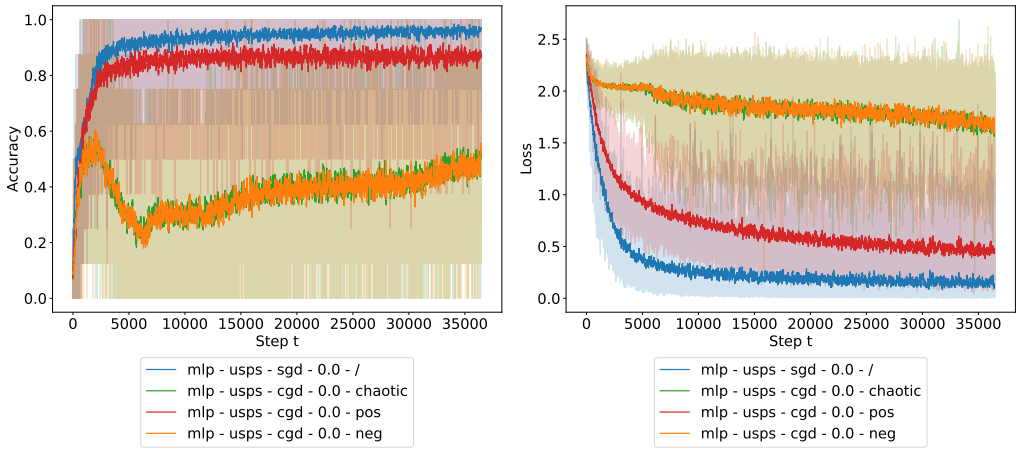
Methods The code for all our experiments is written in Python using the Pytorch framework (Paszke et al., 2017), and available at GitHub [2]. For the calculation of the Hessians, we use O2Grad (Anonymous, 2022), a package on top of Pytorch that enables faster calculation of the Hessian of small ANNs using 2nd order backpropagation. As datasets, we use USPS (Hull, 1994) and FashionMNIST (Xiao et al., 2017) (with images subsampled to 16×16 pixels), since these datasets contain sufficiently low-dimensional, natural data to allow for the calculation of the Hessian and its eigenvalue decomposition at every training step, both for a 784-20-10 MLP and for a small 2D CNN. To modify the parameter updates of the model as proposed in Theorem 2.2, we use the Pytorch implementation of SGD and alter it slightly to implement an own class CGD (Chaos-sensitive Gradient Descent) with the ability to filter parts of the eigenvalue spectrum.

Our experiments were run on single GPU nodes of a system featuring an AMD Ryzen Threadripper 1950X processor, 4x Nvidia GeForce RTX 2080 TI GPUs (11GB VRAM) and 64GB RAM, as well as on a second system featuring an Intel(R) Core(TM) i5-8600K, Nvidia GeForce GTX 1080 Ti (11GB VRAM) and a Nvidia Titan XP (12GB VRAM) and 32GB of RAM. Both systems run on Debian Debian GNU/Linux 11 (bullseye).

4.1 Local Chaos Investigation

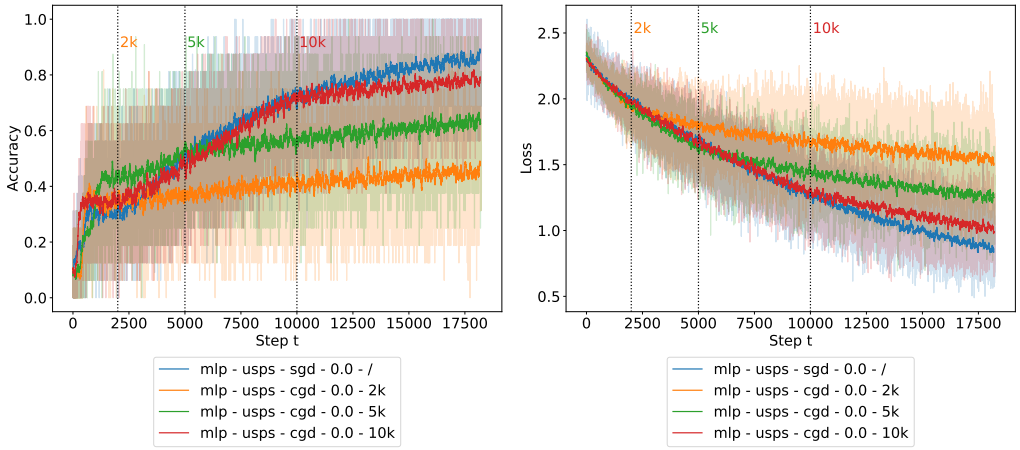
Chaotic direction pruning deteriorates training performance In order to verify the impact of pruning different components on the training performance, we equipped CGD with the ability to filter (1.) chaotic eigenpairs, (2.) only negative eigenpairs and (3.) only positive eigenpairs, and out of those, filter only the k largest by absolute value or prune k at random. As our results for the MLP on USPS in Figure 1 show, models trained without pruning (i.e. regular SGD) reach high/low training accuracies/losses, while the models trained with CGD and pruning of chaotic eigenvalues get stuck at much lower/higher training accuracies/losses. The same phenomenon can be observed for validation losses/accuracies, and on the 2D CNN and FashionMNIST (see Appendix B.1). This is remarkable because it seems to suggest that locally chaotic training behaviour is essential for ANN training in order to quickly achieve reasonable training performance and generalization, and that ANNs trained

[2] `https://github.com/luisherrmann/chaotic_neurips22`



(a) (b)

Figure 1: Line plots of the (a) accuracy and (b) loss curves for an MLP (relu activation) trained on the USPS dataset without momentum. The models trained with CGD use pruning of the full chaotic, negative or positive Hessian spectrum at every time step. The smoothed, solid lines are obtained through local averaging (window size 50) of the respective sequential data. Pruning of the full positive spectrum (red) only has a small effect on training performance, while negative and chaotic pruning strongly limits the network's ability to learn.



(a) (b)

Figure 2: Line plots of the (a) accuracy and (b) loss curves for an MLP trained on the USPS dataset without momentum. We use a sigmoid activation because this slows down convergence and makes the splitting of performance metrics better visible. The models trained with CGD use pruning of the full chaotic Hessian spectrum at every time step after the onset of pruning after 2k, 5k and 10k time steps. The smoothed, solid lines are obtained through local averaging (window size 50) of the respective sequential data). The accuracy and loss curves start diverging from the baseline curve when pruning starts.

by SGD are hardly able to learn in the absence of locally chaotic directions. To make sure this is not an effect occurring only in the early stages of training by SGD, we execute several runs where the pruning of locally chaotic directions starts later in the training. As we can see in Figure 2, training loss and accuracy start improving at a slower rate after the onset of pruning, suggesting that the models learn far more slowly as soon as the pruning of chaotic directions is engaged.

Negative Hessian eigenspectrum drives training by SGD Interestingly, the number of positive chaotic eigenvalues of the Hessian is effectively 0 throughout most of the training (see Appendix

B.2). This means that the chaotic directions pruned during the training come almost exclusively from the negative eigenpairs of the Hessian, which is in line with our previously formulated expectation. However, the extent to which pruning of the negative eigenvalues hinders the training of the network suggests that the negative eigenspectrum of the Hessian is far more important to the training dynamics of neural networks than previously believed by Alain et al. (2019). To better understand the role of the negative and positive eigenspectrum of the Hessian, we perform several runs where we prune (1.) the **entire negative eigenspectrum** and (2.) **the entire positive eigenspectrum** of the Hessian.

As expected, our results in Figure 1 for pruning the full negative eigenspectrum coincide with the results for pruning the full chaotic eigenspectrum of the Hessian: Models trained with pruning of the full negative eigenspectrum of the Hessian reach far lower training performance than models trained with regular SGD. However, **pruning the full positive eigenspectrum of the Hessian only has a small impact on the training performance**: Although the accuracy/loss reaches slightly lower/higher levels than for SGD without pruning, the ANNs trained in this way fare much better than their counterparts with negative pruning. We observe this behaviour across different datasets and models (see Appendix B.1).

These empirical results lead us to the following conclusions:

1. **The negative eigenvalue spectrum of the Hessian contains important information** about the optimization problem since it cannot be discarded without hurting training performance.
2. **The positive eigenvalue spectrum of the Hessian contains less important information** about the optimization since discarding it has less impact on the training.
3. Since the negative eigenspectrum induces locally chaotic training dynamics, **locally chaotic training dynamics are inevitable in SGD** without hurting training performance.

These appear to contradict the findings of Gur-Ari et al. (2018), who observe that the overlap of the gradient is biggest with the positive eigenpairs of the Hessian - an observation that we confirm in our experiments (see Appendix B.2). The apparent contradiction can be resolved through a simplified model of ANN training by SGD, in which **negative curvature precedes positive curvature**: Directions of negative curvature indicate high potential for future loss improvement from moving along those axes, while directions of positive curvature indicate proximity to a local minimum along said axes, and thus reduced potential for loss improvement (see Appendix B.3.1). **Removing the negative components** of the Hessian from the gradient **eliminates possibilities for big future improvements of the loss** and may only allow the optimizer to finish converging in the directions where it was moving anyway, hence why the network is still able to decrease its loss even when the negative eigenspectrum is pruned, though at a smaller pace. This also matches our observation that the count of negative/positive Hessian eigenvalues decreases/increases as the training proceeds (see Appendix B.2).

Note that this model does not explain why the gradient overlap with axes of positive curvature is greater than with axes of negative curvature, only why the former and a high importance of the negative eigenvalues can exist at the same time: Negative curvature is important during an initial discovery phase, and positive curvature during a subsequent exploitation phase.

4.2 Global Chaos Investigation

Our experiments so far have relied on a local quantification of chaos via the LLEs. Next, we extend our investigations to global chaos using finite-time Lyapunov exponents to quantify chaos over longer time spans.

Finite-time Lyapunov exponents In order to better understand long-term chaotic dynamics, we calculate the tangent maps from equation (8) at every time step t and apply eigenvalue decomposition on the resulting finite-time Lyapunov matrix $\mathbf{\Lambda}^{(t)}$ to end up with the corresponding finite-time Lyapunov eigenspectrum. The Lyapunov exponents are above 0 at the beginning of the training and converge against 0 as training progresses (see C.1), suggesting the evolution is initially chaotic, but edge-chaotic in the time limits. Note that since the exponents do not drop below 0, we should expect models with similar initializations to continue diverging even after they reach top performance.

Perturbation analysis The divergence behavior can be further explored by utilizing the Lyapunov eigenspace to perturb models along determined directions. As discussed in section 2.2, this space describes the global divergence directions locally around an initial set of parameters. In our perturbation experiments, we train a model A until convergence at time T_{conv} and calculate the Lyapunov eigenpairs. We then use those eigenpairs to apply initial perturbations onto several models B_1, ..., B_n initialized with the same random seed as A, and observe how their parameters evolve compared to A along the perturbation axes. The latter are derived from the finite-time subspaces

$$V_{\text{chaotic}}(t) := \{\boldsymbol{v} \in \text{Eig}(\boldsymbol{\Lambda}^{(t)}, \lambda) \mid \lambda > 0\}, \qquad V_{\text{non-chaotic}}(t) := \{\boldsymbol{v} \in \text{Eig}(\boldsymbol{\Lambda}^{(t)}, \lambda) \mid \lambda < 0\}$$

according to 5 perturbation strategies: We take (1) the top eigenvector of $\boldsymbol{\Lambda}^{(N)}$ as the **maximally chaotic** direction and (2) the bottom eigenvector as **maximally non-chaotic** direction. Furthermore we consider a linear combination of vectors sampled randomly from (3) $V_{\text{chaotic}}(N)$ and (4) $V_{\text{non-chaotic}}(N)$ that describe random **chaotic** and random **non-chaotic** directions, respectively. Finally, we also sample (5) a completely **random** direction for comparison. The initial parameters are perturbed along the respective directions, where we test different magnitudes $\varepsilon_1, \dots \varepsilon_k$ for each direction to make sure we consider perturbation magnitudes that are in agreement with the locality required by our theoretical preliminaries (see equation (6)).

Ghorbani et al. (2019) finds that the loss landscape near a local minimum is almost flat ($\boldsymbol{H} \approx 0$) and thus using the finite-time Lyapunov matrix at a time point $t = T_{conv}$ where the training is sufficiently converged should provide an acceptable approximation. Therefore, the tangent map $\boldsymbol{Y}^{(t)}$ at the next time step $t > T_{conv} + 1$ and its eigenspace would barely change.

$$\boldsymbol{Y}^{(T_{conv}+1)} = (\boldsymbol{I} - \gamma \boldsymbol{H}^{(T_{conv})})\boldsymbol{Y}^{(T_{conv})} \approx \boldsymbol{Y}^{(T_{conv})} \tag{13}$$

Although we also see a similar distribution of the Hessian eigenspectrum, we do not observe a full convergence of the Hessian to zero (see Appendix B.2). However, the subspaces of chaotic and non-chaotic eigenvectors do stabilise towards the end of the training (see Appendix C.2), making a finite-time analysis with the above perturbation strategies sound.

Distance results Figure 3 (a) shows the distance evolution of an MLP without momentum on USPS and FashionMNIST. For a comparison between model architectures and analysis of momentum, see Appendix C.3. In general, the maximal divergence (empirically ≈ 1 in Figure 3 (a), up to 12 in further experiments in Appendix C.3) lies in range of the expected distance of standard initialized models in Pytorch (LeCun et al., 2012) for the duration of our experiments:

$$\mathbb{E}_{\boldsymbol{\theta}_{1,i}, \boldsymbol{\theta}_{2,i} \sim \mathcal{U}(-\frac{1}{\sqrt{k_i}}, \frac{1}{\sqrt{k_i}})}(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2) \approx 6.38,$$

where k_i are the number of input features of layer i (see Appendix A.8). Still, a surprising result is that the **distances do not seem to saturate** for the duration of our experiments, **which would be expected for a valley of finite size**. In fact, we are still reaching a similar performance with perturbations of magnitudes up to $\varepsilon \leq 7.5$, and only for ≥ 10 there seems to be a significant drop in performance (see Appendix C.3).

Theoretically, the observed distance evolution can be explained as composition of linear divergence and a random walk. Suppose the gradients $\boldsymbol{g}_1(t), \boldsymbol{g}_2(t)$ have distance components $\boldsymbol{w}_1(t), \boldsymbol{w}_2(t)$ with constant distance $\alpha := \|\boldsymbol{w}_1(t) - \boldsymbol{w}_2(t)\|_2^2$

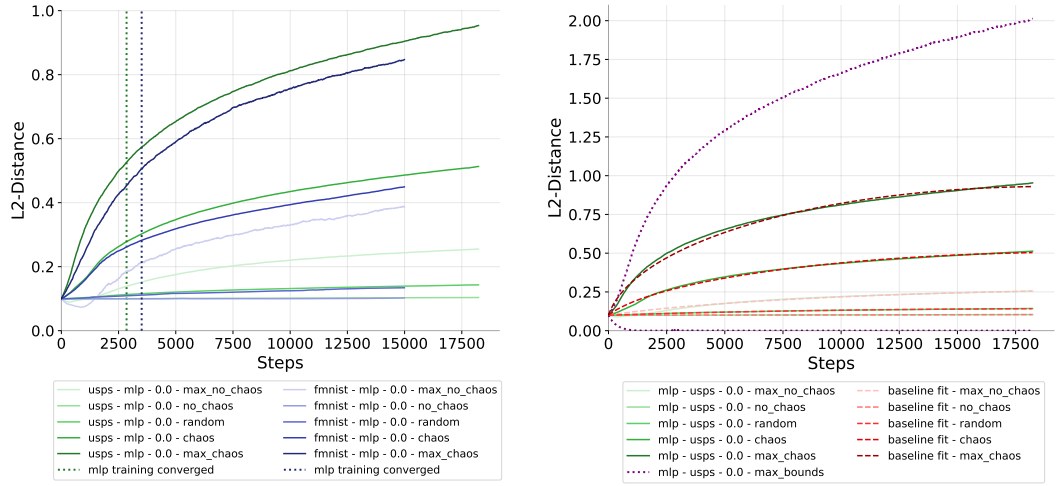
$$\boldsymbol{g}_i(t) \approx \boldsymbol{w}_i(t) + \boldsymbol{\eta}(t) \tag{14}$$

and noise components defined by $\boldsymbol{\eta}_1(t), \boldsymbol{\eta}_2(t) \in \mathcal{N}(0, \sigma^2)$. Then the distance $\Delta\theta(t)$ between two sets of parameters $\boldsymbol{\theta}_1^{(t)}$ and $\boldsymbol{\theta}_2^{(t)}$ evolves as follows:

$$\Delta\theta(t) \approx \mathbb{E}\left[\left\|\sum_{s=0}^{t}(\boldsymbol{w}_1(s) - \boldsymbol{w}_2(s)) + \sum_{s=0}^{t}(\boldsymbol{\eta}_1(s) - \boldsymbol{\eta}_2(s))\right\|_2\right] = \sqrt{\alpha t^2 + \beta t}, \tag{15}$$

with $\beta := 2D\sigma^2$ (see Appendix A.10). For smaller t, this term is still influenced by the random walk term βt whereas for large t the αt^2 term dominates and leads to linear divergence. We give examples of fitting this baseline to the data in Figure 3(b).

Unfortunately, we were unable to explore whether the model distances saturate for a sufficiently long time evolution, which should be the case if the models end up in a bounded loss valley, whereas an



(a) Empirical perturbation evolution

(b) Theoretical modelling

Figure 3: Euclidean distance between a baseline MLP and a perturbed MLP trained with SGD (without momentum). We use 5 different strategies for computing initial perturbations of norm $\varepsilon = 0.1$ using the Lyapunov eigenspace of a model trained beyond convergence of training accuracy. Darker lines represent perturbation directions which we expect to lead to higher model divergence. The observed divergence in (a) on USPS (green) and on FashionMNIST (blue) increases even after convergence of the training accuracy across all perturbation directions. As shown in (b) for the USPS curves, the divergence can be approximated through a function $\sqrt{\alpha t^2 + \beta t}$ (red) within the observed time range. The choice of ε for the trajectories is in agreement with the theoretical bounds (purple) predicted by the maximum/minimum LCE $\lambda_1(t)/\lambda_N(t)$ calculated at every time step.

unbounded loss valley would allow for model divergence to be unbounded as well. A more conclusive analysis on the matter will have to be provided by further studies. In any case, the evolution of training dynamics from chaotic to edge-chaotic in the parameter domain is interesting given that ANNs transition from unchaotic to chaotic in the input domain (Feng et al., 2019), suggesting there may be a complementary relation between the two.

Choice of perturbation length We present the experiments with a perturbation length of $\varepsilon = 0.1$ as we do not find that the approximation holds for larger ε. Surprisingly, we see a similar behavior for smaller perturbation lengths $\varepsilon < 0.1$: the smaller the differences get, the more randomly they behave. We argue that this phenomenon is caused by noise, probably induced by a numerical error. For further analysis see Appendix C.4.

Theoretical bounds In Figure 3 (b), we present the theoretical bounds derived from the maximal and minimal LCE (see equation (5)). Both bounds hold for our choice of $\varepsilon = 0.1$ over the entirety of our experiments where the lower bound converges always to zero (see Appendix C.5). The upper bound, on the other hand, can be utilized as a measure of maximal divergence of a single training run without conducting the full scale of our experiment. Note that for this purpose, it suffices to calculate a Lyapunov-vector-product which can be realized through a Hessian-vector-product for which faster algorithms exist (Pearlmutter, 1994).

5 Discussion

Although we have done our best to perform the pruning experiments on common datasets and model architectures, the experiments are currently difficult to extend to higher-dimensional datasets and models due to the high expense of calculating the Hessian at every training step. For similar reasons, we have not studied the effect of BatchNorm (Ioffe and Szegedy, 2015) and skip connections (He et al., 2015) yet, although we intend to do this at a later time. The high computational cost also limits training time used in our analysis and it could be that some aspects don't apply for longer

training times, e.g. our model for perturbed network divergence (see 3 (b)). In this context, it would also be interesting to investigate possible connections between the edge-chaotic dynamics later in training and overfitting, but we mainly limited the scope of our analysis to training metrics. Another aspect requiring further inquiry is why the model divergence is not always ordered according to our expectations for different perturbation strategies.

In this paper, we have shown that chaotic dynamics of ANN training by SGD are linked to negative curvature. We have found strong evidence that using negative curvature information is essential for training, and we have concluded that chaotic behaviour is inherent to training by SGD. Although globally chaotic dynamics are mostly present at the beginning of the training behaviour and fade towards the end of training, we have shown that models with slightly different initializations continue to diverge linearly in parameter space and provided a model for this behavior. Our work provides a theoretical starting point and tools for further investigations aiming to advance the understanding of chaos in ANN training. In this context, we especially recommend investigating the divergence of trajectories at the end of the training and exploring possible connections to generalization capability (see A.1), as well as the relation between chaos in the input and in the parameter domain.

6 Acknowledgements

We are deeply thankful to Julius Upmeier zu Belzen and Roland Eils from the Berlin Institute of Health for providing the main author with the time and support to write the paper. We would also like to thank Leon Sixt from the Biorobotics Lab and Martin Aleksandrov from the Dahlem Center for Machine Learning and Robotics for giving us valuable advice for the preparation and rebuttal of the paper. Maximilian Granz was supported by the Elsa-Neumann-Scholarship by the state of Berlin. We are also grateful to Nvidia for providing us with a Titan Xp and to ZEDAT for granting us access to their HPC system. Finally, we would like to thank the anonymous reviewers of our paper for taking their time to review our work and for supplying interesting points of discussion.

References

Alain, G., Roux, N. L., and Manzagol, P.-A. (2019). Negative eigenvalues of the hessian in deep neural networks.

Anonymous (2022). O2Grad - A Pytorch Extension for 2nd Order Backpropagation. Draft attached in the Supplementary Material.

Das, A., Roy, A., and Das, P. (2000). Chaos in a three dimensional neural network. *Applied Mathematical Modelling*, 24(7):511–522.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. 12:2121–2159.

Feng, L., Zhang, L., and Lai, C. H. (2019). Optimal machine intelligence at the edge of chaos.

Ghorbani, B., Krishnan, S., and Xiao, Y. (2019). An investigation into neural net optimization via hessian eigenvalue density. *CoRR*, abs/1901.10159.

Gur-Ari, G., Roberts, D. A., and Dyer, E. (2018). Gradient descent happens in a tiny subspace.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). *Efficient BackProp*, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg.

Li, H. (2019). Analysis on the nonlinear dynamics of deep neural networks: Topological entropy and chaos.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

Pearlmutter, B. A. (1994). Fast Exact Multiplication by the Hessian. *Neural Computation*, 6(1):147–160.

Robbins, H. E. (2007). A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

Sagun, L., Bottou, L., and LeCun, Y. (2017). Eigenvalues of the hessian in deep learning: Singularity and beyond.

Sasdelli, M., Ajanthan, T., Chin, T.-J., and Carneiro, G. (2021). A chaos theory approach to understand neural network optimization. In *2021 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–10.

Skokos, C. (2009). The lyapunov characteristic exponents and their computation. *Lecture Notes in Physics*, page 63–135.

Vogt, R., Touzel, M. P., Shlizerman, E., and Lajoie, G. (2020). On lyapunov exponents for rnns: Understanding information propagation using dynamical systems tools. *CoRR*, abs/2006.14123.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

Zhang, L., Feng, L., Chen, K., and Lai, C. H. (2021). Edge of chaos as a guiding principle for modern neural network training.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Where to find mentioned contributions of the abstract in the paper:
   - **The paper establishes a theoretical connection between curvature and locally chaotic dynamics**:
     We show this in Theorem 2.1 in section 2.4
   - **Locally chaotic dynamics are mainly determined by the negative eigenvalue spectrum of the Hessian**:
     Applying Theorem 2.1 to the results of logging the maximal Hessian eigenvalue λ_1 (see Appendix B.2) verifies that local chaos is rarely induced by the top eigenspectrum of the Hessian as $\lambda_1 < 2/\gamma$ for most of the training.

- **The directions of local chaos cannot be removed from SGD without hurting the training performance**:
  See subsection 4.1 and Figure 1.
- **We challenge the widespread understanding of convergence to a confined region in parameter space and show models diverge at a slow pace even after reaching top training performance**:
  See subsection 4.2 and Figure 3.
- **Their divergence can be modelled through a composition of a random walk and a linear divergence**:
  See equation 15 and Figure 3.
- **The tools and insights developed as part of our work contribute to improving the understanding of neural network training dynamics and provide a basis for future improvements of optimization**:
  See subsection 4.2 § Theoretical distance evolution as well as section 5.

(b) Did you describe the limitations of your work? [Yes] See discussion section 5

(c) Did you discuss any potential negative societal impacts of your work? [No] No, the scope of this paper is to improve understanding of SGD and we do not expect any immediate negative societal impacts from this.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [Yes]

   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs can be found in the proof section in Appendix A except the equations from the preliminaries section 2 which are taken from (Skokos, 2009).

3. If you ran experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Not yet, these will be included in the final submission including supplementary material by 26th May.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Aside from logging specifics, all training details required to reproduce our experiments given the code base are specified. More detailed explanations for reproducing the experiments will be provided with the code in the supplementary material.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] No, our work does not aim to beat the performance of existing models on certain tasks but merely to improve the understanding of how SGD works. Errors bars would not be purposeful here.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Yes, since the computations we run involving O2Grad require GPUs with state-of-the-art VRAM size and can thus only be reproduced by people with access to comparably poweful systems.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite the creators of the main assets used, i.e. Pytorch, USPS and FashionMNIST.

   (b) Did you mention the license of the assets? [No] No, the datasets (USPS and FashionMNIST) are released under no license and under an open source license (MIT), respectively, and commonly used in the ML community. As such, mentioning the license in the paper is superfluous. We declare the Python packages used for our experiments as part of the code in the supplementary material.

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include a self-developed asset (O2Grad) in the supplementary material.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] No, the datasets used are publicly available under open source licenses, explicit consent by the creators is not necessary.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] No, neither should apply to the USPS or to the FashionMNIST dataset.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs

Tangent map of SGD

Theorem A.1. *Consider a neural network with N parameters $\theta^{(t)} \in \mathbb{R}^N$ at time step t and training with SGD according to*

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{f}(\theta^{(t)}) = \boldsymbol{\theta}^{(t)} - \gamma \boldsymbol{g}(\boldsymbol{\theta}^{(t)}, \boldsymbol{z}^{(t)}), \tag{16}$$

where γ is the learning rate and $\boldsymbol{z}^{(t)}$ is a data batch sampled at time step t. Then, the tangent map of the corresponding dynamical system can be iteratively calculated as

$$\boldsymbol{Y}^{(t+1)} = (\boldsymbol{I} - \gamma \boldsymbol{H}^{(t)})\boldsymbol{Y}^{(t)}. \tag{17}$$

Proof. The tangent map of a dynamical system at time step which only depends on the previous time step can be obtained as

$$\boldsymbol{Y}^{(t+1)} = \boldsymbol{J}_f(\theta^{(t)})\boldsymbol{Y}^{(t)} = \frac{\partial \boldsymbol{\theta}^{(t+1)}}{\partial \boldsymbol{\theta}^{(t)}} \boldsymbol{Y}^{(t)}, \tag{18}$$

where $\boldsymbol{J}_f(\boldsymbol{\theta}^{(t)})$ is the Jacobian of the evolution function $\boldsymbol{f}$ of the dynamical system. In this case, this is simply

$$\frac{\partial \theta_i^{(t+1)}}{\partial \theta_j^{(t)}} = \frac{\partial \theta_i^{(t)}}{\partial \theta_j^{(t)}} - \gamma \frac{\partial}{\partial \theta_j^{(t)}} g_i(\boldsymbol{\theta}^{(t)}) = \delta_{ij} - \gamma H_{ij}(\boldsymbol{\theta}^{(t)}), \tag{19}$$

where δ_{ij} is the Kronecker delta and $\boldsymbol{H}$ denotes the Hessian of the loss w.r.t. network parameters and we are done. □

Tangent map of SGD with momentum

Theorem A.2. *Consider a neural network evolving according to SGD with learning rate γ and momentum β as*

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{v}^{(t)}, \qquad \boldsymbol{v}^{(t)} = -\gamma \boldsymbol{g}(\boldsymbol{\theta}^{(t)}; \boldsymbol{z}^{(t)}) + \beta \boldsymbol{v}^{(t-1)}, \tag{20}$$

The parameters at time step $t+1$ depend on the gradients at all previous time steps and the dependency can be written as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{v}^{(t)}, \qquad \boldsymbol{v}^{(t)} = -\gamma \sum_{s=0}^{t} \beta^{(t-s)} \boldsymbol{g}(\boldsymbol{\theta}^{(s)}, \boldsymbol{z}^{(s)}). \tag{21}$$

Proof. For the sake of convenience, we omit writing down the dependency on the batch $\mathbf{z}(t)$. Induction start for $t = 0$, where we assume $v^{(0)} = 0$:

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} - \gamma \boldsymbol{g}(\boldsymbol{\theta}^{(0)}) = \boldsymbol{\theta}^{(0)} - \gamma \beta^0 \boldsymbol{g}(\theta^{(0)}) \tag{22}$$

Induction step assuming equation holds true for time step t:

$$\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \theta^{(t)} + \boldsymbol{v}^{(t)} && (23)\\
&= \boldsymbol{\theta}^{(t)} - \gamma \boldsymbol{g}(\boldsymbol{\theta}^{(t)}) + \beta \boldsymbol{v}(\boldsymbol{\theta}^{(t-1)}) && (24)\\
&= \boldsymbol{\theta}^{(t)} - \gamma \boldsymbol{g}(\boldsymbol{\theta}^{(t)}) + \beta \cdot (-\gamma) \sum_{s=0}^{t-1} \beta^{(t-1-s)} \boldsymbol{g}(\boldsymbol{\theta}^{(s)}) && (25)\\
&= \boldsymbol{\theta}^{(t)} - \gamma \boldsymbol{g}(\theta^{(t)}) - \gamma \sum_{s=0}^{t-1} \beta^{(t-s)} \boldsymbol{g}(\boldsymbol{\theta}^{(s)}) && (26)\\
&= \boldsymbol{\theta}^{(t)} - \gamma \sum_{s=0}^{t} \beta^{t-s} \boldsymbol{g}(\boldsymbol{\theta}^{(s)}) && (27)
\end{aligned}$$

□

Theorem A.3. *Consider a neural network evolving according to SGD with learning rate γ and momentum β as*

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{v}^{(t)}, \qquad \boldsymbol{v}^{(t)} = -\gamma \boldsymbol{g}(\boldsymbol{\theta}^{(t}; \boldsymbol{z}^{(t)}) + \beta \boldsymbol{v}^{(t-1)}, \tag{28}$$

where $\boldsymbol{g}$ is the gradient of the loss w.r.t. parameters $\theta^{(t)}$ at time point t. Then, the tangent map $\boldsymbol{Y}^{(t+1)}$ of the dynamical system is given by

$$\boldsymbol{Y}^{(t+1)} = \boldsymbol{Y}^{(t)} - \gamma \left(\sum_{s=1}^{t} \beta^{t-s} \boldsymbol{H}^{(s)} \boldsymbol{Y}^{(s)} \right). \tag{29}$$

Proof. At time step t, the update rule for the gradient depends on all previous time steps according to Theorem A.2. Thus, the tangent map can be obtained through application of the chain rule:

$$\boldsymbol{Y}^{(t+1)} = \frac{d\boldsymbol{\theta}^{(t+1)}}{d\boldsymbol{\theta}^{(0)}} = \sum_{s=0}^{t} \frac{\partial \boldsymbol{\theta}^{(t+1)}}{\partial \boldsymbol{\theta}^{(s)}} \frac{d\boldsymbol{\theta}^{(s)}}{d\boldsymbol{\theta}^{(0)}} = \sum_{s=0}^{t} \frac{\partial \boldsymbol{\theta}^{(t+1)}}{\partial \boldsymbol{\theta}^{(s)}} \boldsymbol{Y}^{(s)}. \tag{30}$$

Using Theorem A.2 again, the partial derivatives can be resolved as

$$\frac{\partial \boldsymbol{\theta}^{(t+1)}}{\partial \boldsymbol{\theta}^{(s)}} = \begin{cases} \boldsymbol{I} - \gamma \beta^0 \partial_{\boldsymbol{\theta}^{(t)}} \boldsymbol{g}(\boldsymbol{\theta}^{(t)}) = \boldsymbol{I} - \gamma \boldsymbol{H}^{(t)}, & s = t \\ -\gamma \beta^{t-s} \partial_{\boldsymbol{\theta}^{(s)}} \boldsymbol{g}(\boldsymbol{\theta}^{(s)}) = -\gamma \beta^{t-s} \boldsymbol{H}^{(s)}, & s < t \end{cases} \tag{31}$$

And reinserting into equation (30), we obtain the desired result

$$\boldsymbol{Y}^{(t+1)} = (\boldsymbol{I} - \gamma \boldsymbol{H}^{(t)}) \boldsymbol{Y}^{(t)} - \sum_{s=0}^{t-1} \gamma \beta^{t-s} \boldsymbol{H}^{(s)} \boldsymbol{Y}^{(s)} = \boldsymbol{Y}^{(t)} - \gamma \sum_{s=0}^{t} \beta^{t-s} \boldsymbol{H}^{(s)} \boldsymbol{Y}^{(s)} \tag{32}$$

□

Hessian eigenvalues and Local Chaos

Theorem A.4. *Let $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ be a positive defined matrix, and the natural logarithm of a matrix be defined as the infinite series*

$$\ln \boldsymbol{A} = \sum_{k=1}^{\infty} \frac{(-1)^k}{k} (\boldsymbol{A} - \boldsymbol{I})^k \tag{33}$$

Then, if $(\lambda_i, \boldsymbol{v}_i)_{i=1}^N$ is the set of all eigenpairs of $\boldsymbol{A}$, $(\ln \lambda_i, \boldsymbol{v}_i)_{i=1}^N$ is the set of all eigenpairs of $\ln A$.

Proof. Since $\mathbf{A}$ is positive defined, its eigenvalues λ_i are all positive, so $\ln \lambda_i$ is defined for all eigenvalues of $\mathbf{A}$. Consider an arbitrary eigenvalue of $\mathbf{A}$ satisfying $\mathbf{A}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$. Clearly, $(\mathbf{A} - \boldsymbol{I})\boldsymbol{v}_i = \mathbf{A}\boldsymbol{v}_i - \boldsymbol{I}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i - \boldsymbol{v}_i = (\lambda_i - 1)\boldsymbol{v}_i$. It follows immediately that $(\mathbf{A} - \boldsymbol{I})^k \boldsymbol{v}_i = (\lambda_i - 1)^k \boldsymbol{v}_i$ and thus:

$$\ln \mathbf{A}\boldsymbol{v}_i = \sum_{k=1}^{\infty} \frac{(-1)^k}{k} (\mathbf{A} - \boldsymbol{I})^k \boldsymbol{v}_i = \sum_{k=1}^{\infty} \frac{(-1)^k}{k} (\lambda_i - 1)^k \boldsymbol{v}_i = \ln \lambda_i \boldsymbol{v}_i \tag{34}$$

Since $\ln \mathbf{A} \in \mathbb{R}^{N \times N}$, the matrix can have at most N eigenvalues, so we know that these are all eigenvalues of $\ln \mathbf{A}$ and we are done. □

Theorem A.5. *Given a neural network with N parameters, let $\boldsymbol{H}$ be the Hessian of the loss w.r.t. network parameters, and let $\{(\lambda_i, \boldsymbol{v}_i)\}_{i=1}^N$ be the Hessian's eigenpairs. Furthermore, let us assume that the network is trained using SGD with learning rate γ. Consider the eigenvalues for which $\lambda_i \neq \frac{1}{\gamma}$. Then, the network's LLEs indicate locally chaotic training behaviour in the direction of v_i if*

$$\lambda_i \leq 0 \vee \lambda_i \geq \frac{2}{\gamma} \tag{35}$$

Proof. If $(\lambda_i, \boldsymbol{v}_i)$ is an eigenpair of $\boldsymbol{H}$, $(1-\gamma\lambda_i, \boldsymbol{v}_i)$ is an eigenpair of $\boldsymbol{J}_f = (\boldsymbol{I} - \gamma\boldsymbol{H})$, since $(\boldsymbol{I} - \gamma\boldsymbol{H})\boldsymbol{v} = (1-\gamma\lambda_i)\boldsymbol{v}$. Due to the identity matrix and the Hessian being symmetric, $(\boldsymbol{I} - \gamma\boldsymbol{H})^T(\boldsymbol{I} - \gamma\boldsymbol{H}) = (\boldsymbol{I} - \gamma\boldsymbol{H})^2$. Furthermore, if $\mathbf{A}$ is a positive definite matrix and $(\lambda_i, \boldsymbol{v}_i)$ is an eigenpair of $\mathbf{A}$, then $(\ln\lambda_i, \boldsymbol{v}_i)$ is an eigenpair of $\ln\mathbf{A}$ (see Appendix A.4). If all eigenvalues of $(\boldsymbol{I} - \gamma\boldsymbol{H})^2$ are positive and nonzero due to $\lambda_i \neq \frac{1}{\gamma}$, the matrix is positive definite and it follows that $\frac{1}{2}\ln(1-\gamma\lambda_i)^2$ is an eigenvalue of

$$\boldsymbol{\Lambda}_x^{(1)} = \frac{1}{2}\ln(\boldsymbol{I} - \gamma\boldsymbol{H})^T(\boldsymbol{I} - \gamma\boldsymbol{H})$$

for the same eigenvector $\boldsymbol{v}_i$ and thus an LLE. To indicate locally chaotic training behaviour, the LLE must satisfy

$$\ln(1-\gamma\lambda_i)^2 \geq 0 \quad \Leftrightarrow \quad (1-\gamma\lambda_i)^2 \geq 1 \tag{36}$$

Solving for λ_i gives the desired result and we are done. □

Now, consider the edge case where some eigenpair $(\lambda_i, \boldsymbol{v}_i)$ of the Hessian $\boldsymbol{H}$ satisfies $\lambda_i = \frac{1}{\gamma}$ s.t. for an infinitesimally small displacement $\boldsymbol{\delta v} \parallel \boldsymbol{v}_i$

$$\boldsymbol{J}_f\boldsymbol{\delta v} = (\boldsymbol{I} - \gamma\boldsymbol{H})\boldsymbol{\delta v} = (1 - \gamma \cdot 1/\gamma)\boldsymbol{\delta v} = 0. \tag{37}$$

That is, the dynamical system exhibits a maximal contraction along the axis of $\mathbf{v}_i$, meaning that the system does not evolve along this axis. Therefore, these eigenpairs can be excluded from our considerations on chaotic behaviour. The respective eigenvalues can still be accounted for in the definition of the Local Lyapunov spectrum by either accepting eigenvalues of $-\infty$, or by adapting the definition of the LLEs to be the singular values of

$$\tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^* = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = (\boldsymbol{I} - \gamma\boldsymbol{H})^2, \tag{38}$$

where the first decomposition is the zero-truncated SVD. In this case, the singular pairs $(\sigma_i, \boldsymbol{u}_i)$, where $\boldsymbol{u}_i$ is the singular vector corresponding to singular value σ_i, define Local Lyapunov Exponent (LLE) and corresponding Lyapunov vector, respectively. Note that since $(\boldsymbol{I} - \gamma\boldsymbol{H})^2$ is always positive-semidefinite, the singular value decomposition is identical to the eigenvalue decomposition $(\boldsymbol{I} - \gamma\boldsymbol{H}) = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1}$ provided it exists. Therefore the Lyapunov spectrum can be constrained to exclude Hessian eigenpairs where $\lambda = \frac{1}{\gamma}$ by considering the zero-truncated SVD. This definition is also interesting because it allows us to loosely interpret the top Lyapunov vectors as the directions of maximum trajectory variance if we performed a PCA on trajectory distances in the limit for an infinite number of perturbed trajectories sampled, with the perturbation magnitude and evolution time of the trajectories going to 0.

Theorem A.6. *Given a neural network with N parameters, let $\boldsymbol{H}^{(t)}$ be the Hessian of the loss w.r.t. network parameters at time step t. Furthermore, let $\lambda_1^{(s)}$ be the maximum LLE of the training dynamics at time step s for $0 < s \leq t$. Then, the mLCE of the system evolution can be upper-bounded as*

$$\lambda_1 \leq \lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} \lambda_1^{(s)} \tag{39}$$

Proof. As previously established, the mLCE λ_1 is defined as the maximum eigenvalue of $\boldsymbol{\Lambda}^{(\infty)}$, which is precisely the limit of the time-normalized spectral norm of the tangent map:

$$\lambda_1 = \lambda_{\max} \lim_{t\to\infty} \frac{1}{2t}\ln \boldsymbol{Y}^{(t)T}\boldsymbol{Y}^{(t)} = \lim_{t\to\infty} \frac{1}{t}\ln \lambda_{\max}\sqrt{\boldsymbol{Y}^{(t)T}\boldsymbol{Y}^{(t)}} = \lim_{t\to\infty} \frac{1}{t}\ln\|\boldsymbol{Y}^{(t)}\| \tag{40}$$

For SGD without momentum, this can be further rewritten using the chain rule and the rules for the logarithm:

$$\lambda_1 = \lim_{t\to\infty} \frac{1}{t}\ln\left\|\prod_{s=1}^{t} \frac{\partial\boldsymbol{\theta}^{(s)}}{\partial\boldsymbol{\theta}^{(s-1)}}\right\| \leq \lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t}\ln\left\|\frac{\partial\boldsymbol{\theta}^{(s)}}{\partial\boldsymbol{\theta}^{(s-1)}}\right\| \tag{41}$$

Using the definition of the spectral norm again, we realize the eigenvalues in the sum are the maximum LLEs and we end up with

$$\lambda_1 \leq \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \ln \lambda_{max} \sqrt{\boldsymbol{J}_f(\boldsymbol{\theta}^{(s+1)})^T \boldsymbol{J}_f(\boldsymbol{\theta}^{(s+1)})} \tag{42}$$

$$\leq \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \lambda_{max} \ln \sqrt{\boldsymbol{J}_f(\boldsymbol{\theta}^{(s+1)})^T \boldsymbol{J}_f(\boldsymbol{\theta}^{(s+1)})} = \lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t} \lambda_1^{(s)}. \tag{43}$$

□

Theorem A.7. *Given a neural network with Hessian $\boldsymbol{H}^{(t)}$, eigenpairs $(\lambda_i^{(t)}, \boldsymbol{v}_i^{(t)})_{i=1}^n$ of $\Lambda^{t,t+1}$ and gradient updates $\Delta\theta^{(t)}$, let there be $k \leq n$ chaotic eigenpairs, i.e. $\lambda_1 \geq ...\lambda_k > 0$. Suppose all chaotic components are pruned through projection, leading to a modified update vector*

$$\Delta\tilde{\boldsymbol{\theta}}^{(t)} = \boldsymbol{v}^{(t)}(\boldsymbol{v}^{(t)})^T \Delta\boldsymbol{\theta}^{(t)}, \qquad \boldsymbol{v}^{(t)} = \sum_{l>k} \boldsymbol{v}_l^{(t)} \boldsymbol{v}_l^{(t)^T}. \tag{44}$$

Then, for SGD training with learning rate of γ and without momentum, the updated training dynamics are given by

$$\theta^{(t+1)} = \boldsymbol{\theta}^{(t)} - \gamma \boldsymbol{v}^{(t)} \boldsymbol{v}^{(t)^T} \boldsymbol{g}(\boldsymbol{\theta}^{(t)}), \qquad \boldsymbol{Y}^{(t+1)} = (\boldsymbol{I} - \gamma \boldsymbol{v}^{(t)} \boldsymbol{v}^{(t)^T}) \boldsymbol{Y}^{(t)} \tag{45}$$

and for sufficiently small variations of the initial parameters, the system is guaranteed not to have exponentially diverging orbits.

Proof. The updated training dynamics are given by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \Delta\tilde{\boldsymbol{\theta}}^{(t)} = \boldsymbol{\theta}^{(t)} - \boldsymbol{v}^{(t)} \boldsymbol{v}^{(t)^T} \Delta\theta^{(t)} = \boldsymbol{\theta}^{(t)} - \gamma \boldsymbol{v}^{(t)} \boldsymbol{v}^{(t)^T} \boldsymbol{g}(\boldsymbol{\theta}^{(t)}, \mathbf{x}^{(t)}, \boldsymbol{Y}^{(t)}). \tag{46}$$

Thus, we can derivate w.r.t. $\theta^{(t)}$ to obtain the dynamics Jacobian and get

$$\frac{\partial \theta_i^{(t+1)}}{\partial \theta_j^{(t)}} = \frac{\partial}{\partial \theta_j^{(t)}} \left(\theta_i^{(t)} - \gamma \sum_k \left\{ \boldsymbol{v}^{(t)} \boldsymbol{v}^{(t)^T} \right\}_{ik} g_k^{(t)}(\boldsymbol{\theta}^{(t)}, \mathbf{x}^{(t)}, \boldsymbol{Y}^{(t)}) \right) \tag{47}$$

$$= \frac{\partial \theta_i^{(t)}}{\partial \theta_j^{(t)}} - \gamma \sum_k \left\{ \boldsymbol{v}^{(t)} \boldsymbol{v}^{(t)^T} \right\}_{ik} \frac{\partial g_k^{(t)}(\boldsymbol{\theta}^{(t)}, \mathbf{x}^{(t)}, \boldsymbol{Y}^{(t)})}{\partial \theta_j^{(t)}} \tag{48}$$

$$= \delta_{ij} - \gamma \sum_k \left\{ \boldsymbol{v}^{(t)} \boldsymbol{v}^{(t)^T} \right\}_{ik} H_{kj}, \tag{49}$$

or more succinctly in matrix notation

$$\frac{\partial \theta^{(t+1)}}{\partial \theta^{(t)}} = \boldsymbol{I} - \gamma \boldsymbol{v}^{(t)} \boldsymbol{v}^{(t)^{(T)}} \boldsymbol{H}^{(t)}. \tag{50}$$

Now, we still need to show that the system is guaranteed not to have exponentially diverging orbits. If $\boldsymbol{v}_i^{(t)}$ is a pruned Lyapunov vector (i.e. $i > k$), its eigenvalue for the modified dynamics Jacobian becomes 1.

$$(\boldsymbol{I} - \gamma \boldsymbol{v}^{(t)} \boldsymbol{v}^{(t)^T} \boldsymbol{H}^{(t)}) \boldsymbol{v}_i^{(t)} = \boldsymbol{v}_i^{(t)} - \gamma \boldsymbol{v}^{(t)} \boldsymbol{v}^{(t)^T} \lambda_i \boldsymbol{v}_i^{(t)} \tag{51}$$

$$= \boldsymbol{v}_i^{(t)} - \gamma \lambda_i \underbrace{\boldsymbol{v}^{(t)} \boldsymbol{v}^{(t)^T} \boldsymbol{v}_i}_{=0} \tag{52}$$

$$= \boldsymbol{v}_i^{(t)} \tag{53}$$

It follows immediately that the eigenvalue of the local Oseledet matrix (the LLE) is $\ln(1) = 0$. Otherwise, if $\boldsymbol{v}_i^{(t)}$ was a preserved Lyapunov vector, then it satisfies

$$\boldsymbol{v}^{(t)} \boldsymbol{v}^{(t)^T} \boldsymbol{v}_i^{(t)} = \boldsymbol{v}_i^{(t)} \tag{54}$$

by design and thus its LLE is preserved. Summing up, the eigenspectrum of the modified dynamics Jacobian satisfies the eigenvalue equation

$$\frac{\partial \boldsymbol{\theta}^{(t+1)}}{\partial \boldsymbol{\theta}^{(t)}} \boldsymbol{v}_i = \begin{cases} 1\boldsymbol{v}_i, & k \leq i \\ (1-\lambda_i)\boldsymbol{v}_i, & 0 \leq i < k \end{cases} \tag{55}$$

such that all eigenvalues translate to non-chaotic or edge-chaotic LLEs smaller than (or equal to) 0. Applying Theorem A.6 and using the fact that the LLEs, in particular the maximum LLE at any time step, are all smaller than 0 yields

$$\lambda_1 \leq \lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t} \lambda_1^{(s)} \leq \lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t} 0 = 0. \tag{56}$$

Hence, the trajectories diverge at most polynomially and we are done. □

Distance Statements

Theorem A.8. *Two standard initialized 784-20-10 MLPs in Pytorch have a expected L2-distance of* ≈ 6.38.

Proof. We calculate the distance between single layers first and differentiate between weight and bias. Pytorch initializes weights $\boldsymbol{w}$ and biases $\boldsymbol{b}$ uniformly with

$$\boldsymbol{w}_i, \boldsymbol{b}_i \sim \mathcal{U}(-\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{k}}), \quad k := \text{\# in_features} \quad \text{and} \quad m := \text{\# out_features}$$

<u>1. case: weight</u>

So let $\boldsymbol{w}_i^{(1)}, \boldsymbol{w}_i^{(2)} \sim \mathcal{U}(-\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{k}})$ for all i.

$$\begin{aligned} \mathop{\mathbb{E}}_{\boldsymbol{w}_i^{(1)}, \boldsymbol{w}_i^{(2)} \sim \mathcal{U}(-\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{k}})} [\|\boldsymbol{w}^{(1)} - \boldsymbol{w}^{(2)}\|_2^2] &= \mathop{\mathbb{E}}_{\boldsymbol{w}_i \sim \mathcal{U}(-\frac{2}{\sqrt{k}}, \frac{2}{\sqrt{k}})} [\|\boldsymbol{w}\|_2^2] \\ &= \sum_{i=1}^{km} \mathop{\mathbb{E}}_{\boldsymbol{w}_i \sim \mathcal{U}(-\frac{2}{\sqrt{k}}, \frac{2}{\sqrt{k}})} [w_i^2] = \sum_{i=1}^{km} \mathrm{Var}(\boldsymbol{w}_i) = km \frac{4^2}{12k} = \frac{4m}{3} \end{aligned}$$

Thus $\mathbb{E}[\|\boldsymbol{w}^{(1)} - \boldsymbol{w}^{(2)}\|_2] = \sqrt{\frac{4m}{3}}$.

<u>2. case: bias</u>

Analogously for the bias where the dimension of the bias is not km but m, we end up with $\mathbb{E}[\|\boldsymbol{b}^{(1)} - \boldsymbol{b}^{(2)}\|_2] = \sqrt{\frac{4m}{3k}}$.

We now calculate the expected distance over all layers of the two 784-20-10 MLPs $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}$:

$$\mathbb{E}[\|\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}\|_2] = \sqrt{\frac{4(m_1+m_2)}{3} + \frac{4m_1}{k_1} + \frac{4m_2}{k_2}} = \sqrt{\frac{4(30)}{3} + \frac{4(20)}{3(784)} + \frac{4(10)}{3(20)}} \approx 6.38$$

□

Theorem A.9. *Two random Gaussian walks* $\boldsymbol{x_1}(t), \boldsymbol{x_2}(t) = \boldsymbol{x_i}(t-1) + \boldsymbol{\eta_i}(t)$ *with* $\boldsymbol{x_1}(0), \boldsymbol{x_2}(0) = \boldsymbol{0}$ *and* $\boldsymbol{\eta_1}(t), \boldsymbol{\eta_2}(t) \in \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ *have an expected distance of* $\sqrt{2TD}\sigma$ *until time step* T.

Proof.

$$\begin{aligned} \mathbb{E}\left[\|\boldsymbol{x}_1(t) - \boldsymbol{x}_2(t)\|_2\right] &= \mathop{\mathbb{E}}_{\boldsymbol{\eta}_1(t), \boldsymbol{\eta}_2(t) \in \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})} \left[\|\sum_{t=1}^{T} \boldsymbol{\eta}_1(t) - \boldsymbol{\eta}_2(t)\|_2\right] \\ &\overset{(1)}{=} \mathop{\mathbb{E}}_{\boldsymbol{\eta}'(t) \in \mathcal{N}(0, \sigma^2 \boldsymbol{I})} \left[\|\sum_{t=1}^{2T} \boldsymbol{\eta}'(t)\|_2\right] \overset{(2)}{=} \mathop{\mathbb{E}}_{\boldsymbol{\eta}'(t) \in \mathcal{N}(0, 2T\sigma^2 \boldsymbol{I})} [\|\tilde{\boldsymbol{\eta}}(t)\|_2] \\ &\overset{(3)}{=} \sqrt{2TD}\sigma. \end{aligned} \tag{57}$$

Equation (1) is valid, because $\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) = -\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ and (2), because variances of the sum two independent variables add up. For (3), observe that

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \in \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})}[\|\boldsymbol{x}\|_2^2] = \sum_{i=1}^{D} \mathop{\mathbb{E}}_{x_i \in \mathcal{N}(0, \sigma^2)}[x_i^2] = \mathop{\mathbb{E}}_{x' \in \mathcal{N}(0, D\sigma^2)}[x'^2] = D\sigma^2 \tag{58}$$

□

Theorem A.10. *Let the gradients of two models $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mathbb{R}^D$ be defined as*

$$g_i(t) = \boldsymbol{w}_i(t) + \boldsymbol{\eta}_i$$

where $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ are independent and identically distributed, isotropic noise terms and $\| \sum_{s=1}^{t} (\boldsymbol{w}_1(s) - \boldsymbol{w}_2(s)\|_2 = \alpha t$ diverges linearly over time. Assume both models are initialized with the same set of parameters $\theta_1^{(0)} = \theta_2^{(0)}$. Then distance of the two models evolves as

$$\mathop{\mathbb{E}}_{\eta_1, \eta_2 \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})}[\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2] = \sqrt{\alpha^2 t^2 + \beta t},$$

where $\beta = 2D\sigma^2$.

Proof.

$$\mathop{\mathbb{E}}_{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})} \left[\left\| \textstyle\sum_{s=0}^{t} (\boldsymbol{w}_1(s) - \boldsymbol{w}_2(s)) + \sum_{s=0}^{t} (\boldsymbol{\eta}_1(s) - \boldsymbol{\eta}_2(s)) \right\|_2 \right] \tag{59}$$

We want to calculate the term (59) where $\sum_{s=0}^{t} (\boldsymbol{w}_1(s) - \boldsymbol{w}_2(s))$ defines the linear divergence term and $\sum_{s=0}^{t} (\boldsymbol{\eta}_1(s) - \boldsymbol{\eta}_2(s))$ is the random Gaussian walk as in equation (15). For simplification, assume that we take the expected value over $\boldsymbol{\eta}_1(s), \boldsymbol{\eta}_2(s) \in \mathcal{N}(\mathbf{0}, \sigma^2 I)$ in all of the following expression

$$\begin{aligned}
&\mathbb{E}\left[\left\| \sum_{s=0}^{t} (\boldsymbol{w}_1(s) - \boldsymbol{w}_2(s)) + \sum_{s=0}^{t} (\boldsymbol{\eta}_1(s) - \boldsymbol{\eta}_2(s)) \right\|_2^2 \right] \\
&= \mathbb{E}\left[\left\| \sum_{s=0}^{t} (\boldsymbol{w}_1(s) - \boldsymbol{w}_2(s)) \right\|_2^2 + \left\| \sum_{s=0}^{t} (\boldsymbol{\eta}_1(s) - \boldsymbol{\eta}_2(s)) \right\|_2^2 \right. \\
&\qquad \left. + \sum_{s,s'=0}^{t} \langle (\boldsymbol{w}_1(s') - \boldsymbol{w}_2(s')), \boldsymbol{\eta}_1(s) - \boldsymbol{\eta}_2(s)) \rangle \right] \\
&= \alpha^2 t^2 + \beta t + \mathbb{E}\left[\sum_{s,s'=0}^{t} \langle (\boldsymbol{w}_1(s') - \boldsymbol{w}_2(s')), \boldsymbol{\eta}_1(s) - \boldsymbol{\eta}_2(s)) \rangle \right]
\end{aligned}$$

where $\beta := 2D\sigma^2$, which follows from Theorem A.9. Now this leaves to show that the last summand is zero. In fact, observe that for all $v \in \mathbb{R}^D$

$$\mathop{\mathbb{E}}_{\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, 2\sigma^2 \boldsymbol{I})}[\langle \boldsymbol{v}, \boldsymbol{\omega} \rangle] = \sum_{i=1}^{D} v_i \mathop{\mathbb{E}}_{\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, 2\sigma^2 \boldsymbol{I})}[\boldsymbol{\omega}_i] = 0,$$

which can be applied to $\boldsymbol{v} = \boldsymbol{w}_1(s) - \boldsymbol{w}_2(s)$ and $\boldsymbol{\omega} = \boldsymbol{\eta}_1(s) - \boldsymbol{\eta}_2(s)$ for all s. □

Theorem A.11. *The modified Subspace Similarity defined as*

$$\textit{overlap}(V, W) = \frac{\textit{Tr}(P_V, P_W)}{\sqrt{\textit{Tr}(P_V)\textit{Tr}(P_W)}} \sqrt{\frac{d_{\textit{max}}}{d_{\textit{min}}}} = \frac{1}{d_{\textit{min}}} \|\textit{flatten}(V^T W)\|_2^2$$

where P_V, P_W are the orthogonal profectors onto the subspaces V, W fulfills the following properties:

1. *If $V \perp W$, then overlap$(V, W) = 0$*
2. *If $V \subseteq W$ or $W \subseteq V$, then overlap$(V, W) = 1$*
3. *For arbitrary subspaces V, W, overlap$(V, W) \in [0, 1]$*

Proof. We first show the equality of the definition. The orthogonal projector is defined as $P_V := \boldsymbol{V}\boldsymbol{V}^T$ using $\boldsymbol{V}$ as a matrix of orthonormal basis vectors (in the columns) and as the space V to simplify the following equations. Now the numerator equals

$$\text{Tr}(P_V, P_W) = \text{Tr}(\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{W}\boldsymbol{W}^T) \overset{*}{=} \text{Tr}((\boldsymbol{W}^T\boldsymbol{V})(\boldsymbol{V}^T\boldsymbol{W})) = \|\text{flatten}(\boldsymbol{V}^T\boldsymbol{W})\|_2^2$$

where for * we used the cyclic property of the trace. We utilize this property again to calculate

$$\text{Tr}(P_V) = \text{Tr}(\boldsymbol{V}\boldsymbol{V}^T) = \text{Tr}(\boldsymbol{V}^T\boldsymbol{V}) = \text{Tr}(\text{Id}_V) = \dim(V)$$

The rest of the equation follows by applying the two results.

Now statement 1. follows directly from the fact that if $V \perp W$ then $\boldsymbol{V}^T\boldsymbol{W} = \boldsymbol{0}$. For statement 2. observe that if $W \subseteq V$ then the orthogonal projection of W onto V is the identity: $P_V \mid_W = \text{Id}_W$. So the similarity is

$$\text{overlap}(V, W) = \frac{1}{\dim(W)}\text{Tr}(\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{W}\boldsymbol{W}^T) = \frac{\text{Tr}(\boldsymbol{W}\boldsymbol{W}^T)}{\dim(W)} = \frac{\text{Tr}(\boldsymbol{W}^T\boldsymbol{W})}{\dim(W)} = 1$$

Finally for statement 3., we see that $\|\text{flatten}(\boldsymbol{V}^T\boldsymbol{W})\|_2^2 \geq 0$ is clearly positive and what is left to prove is that it is bounded by 1. Let $n := \dim(V)$, $m := \dim(W)$ and $\boldsymbol{v}_1, ..., \boldsymbol{v}_n$ and $\boldsymbol{w}_1, ..., \boldsymbol{w}_m$ be an orthonormal basis of V, W, respectively. Assume with out loss of generality $m < n$

$$\|\text{flatten}(\boldsymbol{V}^T\boldsymbol{W})\|_2^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\langle \boldsymbol{v}_i, \boldsymbol{w}_j\rangle^2$$

Now extend the orthonormal basis of V to a orthonormal basis $B := v_1, ..., v_n, ..., v_k$ of the whole space $V' \supseteq V, W$. And write the orthonormal basis of W in terms of B: $w_j = \sum_{i=1}^{k}\lambda_{ji}v_i$. Then

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\langle \boldsymbol{v}_i, \boldsymbol{w}_j\rangle^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_{ji}^2\langle \boldsymbol{v}_i, \boldsymbol{v}_i\rangle^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_{ji}^2$$

Now $\sum_{i=1}^{n}\lambda_{ji}^2 \leq 1$, since

$$1 = \|\boldsymbol{w}_j\|_2^2 = \left\langle \sum_{i=1}^{k}\lambda_{ji}\boldsymbol{v}_i, \sum_{i=1}^{k}\lambda_{ji}\boldsymbol{v}_i \right\rangle = \sum_{i=1}^{k}\lambda_{ji}^2 \geq \sum_{i=1}^{n}\lambda_{ji}^2$$

And we apply this inequality to what we previously ended up with

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_{ji}^2 \leq \sum_{j=1}^{m}1 = m$$

to arive at what we wanted to show: $\text{overlap}(V, W) = \frac{1}{m}\|\text{flatten}(V^TW)\|_2^2 \leq \frac{m}{m} = 1$. □

A.1 Generalization and long-term chaos

In this subsection, we provide a mathematical argument for why we can assume chaotic dynamics (in the sense of sensitivity to initial conditions) to generally have an impact on generalization properties of ANNs trained with SGD. Consider an ANN trained with some procedure M to yield a parametrized function f_θ^M by the end of the training, where $\theta := \boldsymbol{\theta}^{(T)}$ are the parameters at the end of the training. Let S be the set of all data points that could potentially be used to train a network for its respective task, and let $S^{train} \subset S$ be the entirety of data points used to actually train the network. Furthermore, let $\mathcal{L}(f_\theta^M; x)$ be the loss from calculating the network outputs, given a sample $x \in S$. Let us assume the input data to be structured s.t. $\forall x \in S \exists \epsilon > 0$ s.t. $\forall y \in \mathcal{B}_\epsilon(s) : y \in S$. A training procedure that

leads to good generalization should be one that minimizes the variance of the loss of the learned function with respect to data points $x \in S$. Thus, the generalization capacity can be metricized by

$$\text{Var}[\mathcal{L}(f_\theta^M)] = \mathbb{E}_{\boldsymbol{x}\in S}\left[\left(\mathcal{L}(f_\theta^M;\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\in S}\left[\mathcal{L}(f_\theta^M;\boldsymbol{x})\right]\right)^2\right] = \mathbb{E}_{\boldsymbol{x}\in S}\left[\left(\mathcal{L}(f_\theta^M;\boldsymbol{x}) - \bar{\mathcal{L}}\right)^2\right] \quad (60)$$

In particular, the generalization properties should be good within any finite region $\tilde{S} \subset S$. Assuming $\mathcal{L}(f_\theta^M;\boldsymbol{x})$ is a piecewise-smooth function with respect to x and assuming x is not a border point at a derivative discontinuity, the loss function may be expressed using a Taylor polynomial:

$$\mathcal{L}(f_\theta^M;\boldsymbol{x}_s + \boldsymbol{\delta x}) = \mathcal{L}(f_\theta^M;\boldsymbol{x}_s) + \boldsymbol{\delta x}^T \nabla_{\mathcal{L}} f_\theta^M + \frac{1}{2}\boldsymbol{\delta x}^T \boldsymbol{H}_{\mathcal{L}}(f_\theta^M;\boldsymbol{x}_s)\delta\boldsymbol{x} + ... \quad (61)$$

Let us consider the ability of the function f_θ to generalize on points in a vicinity $\tilde{S}$ of a point x_s which was part of the training data, and where all other points $x \in \tilde{S}\backslash\{x_S\}$ were not seen during the training procedure M. For the sake of simplicity, let us assume that x_s was seen exactly once during the training, namely at time step s. Then from the chain rule, it follows that

$$\mathcal{L}'(f_\theta^M;\boldsymbol{x}_s) = \frac{d\mathcal{L}}{d\boldsymbol{x}}|\boldsymbol{x} = \boldsymbol{x}_s = \frac{d\mathcal{L}}{d\boldsymbol{\theta}^{(T)}}\frac{d\boldsymbol{\theta}^{(T)}}{d\boldsymbol{\theta}^{(s+1)}}\frac{d\boldsymbol{\theta}^{(s+1)}}{d\boldsymbol{x}_s} = \frac{d\mathcal{L}}{d\boldsymbol{\theta}^{(T)}}\boldsymbol{Y}^{(s+1,T)}\frac{d\boldsymbol{\theta}^{(s+1)}}{d\boldsymbol{x}_s} \quad (62)$$

When the training procedure M is SGD (without momentum), the update rule gives us that

$$\frac{d\boldsymbol{\theta}^{(s+1)}}{d\boldsymbol{x}_s} = \frac{d}{d\boldsymbol{\theta}^{(s)}}(\boldsymbol{\theta}^{(s)} - \gamma\boldsymbol{g}(\boldsymbol{\theta}^{(s)};\boldsymbol{x}_s)) = \boldsymbol{I} - \gamma\frac{\partial^2\mathcal{L}(\boldsymbol{\theta}^{(s)};\boldsymbol{x}_s)}{\partial\boldsymbol{x}_s\partial\boldsymbol{\theta}^{(s)}} = \boldsymbol{I} - \gamma\boldsymbol{H}_{x\theta}^{(s)}, \quad (63)$$

where the tangent map $\boldsymbol{Y}^{(s+1,T)}$ gives us the sensitivity of the parameters at the end of the training procedure to the parameters at a previous time point $\boldsymbol{\theta}^{(s+1)}$. Due to the product rule, the higher-order terms of the Taylor expansion will also contain some dependency on $\boldsymbol{Y}^{(s+1,T)}$. However, for more clarity, we restrict ourselves to first-order terms to end up with the expression

$$\text{Var}_{\tilde{S}}[\mathcal{L}(f_\theta^{SGD})] \approx \mathbb{E}_{\boldsymbol{x}\in\tilde{S}}\left[\left(\mathcal{L}(f_\theta^{SGD};\boldsymbol{x}_s) + \frac{d\mathcal{L}}{d\boldsymbol{\theta}^{(T)}}\boldsymbol{Y}^{(s+1,T)}(\boldsymbol{I} - \gamma\boldsymbol{H}_{x\theta}^{(s)})\boldsymbol{\delta x} - \bar{\mathcal{L}}\right)^2\right] \quad (64)$$

What this argument tells us is that, in order for the generalization capacity of a trained function f_θ^{SGD} to be good at least for unseen samples within some vicinity of previously seen samples $\boldsymbol{x}_s$, training dynamics should not be too sensitive on initial conditions. The argument can be extended to vicinities of arbitrary sets of points seen during training by considering the expected values over respective regions.

B Further Analysis on Locally Chaotic Dynamics

B.1 Pruning Experiments: Performance Metrics

USPS

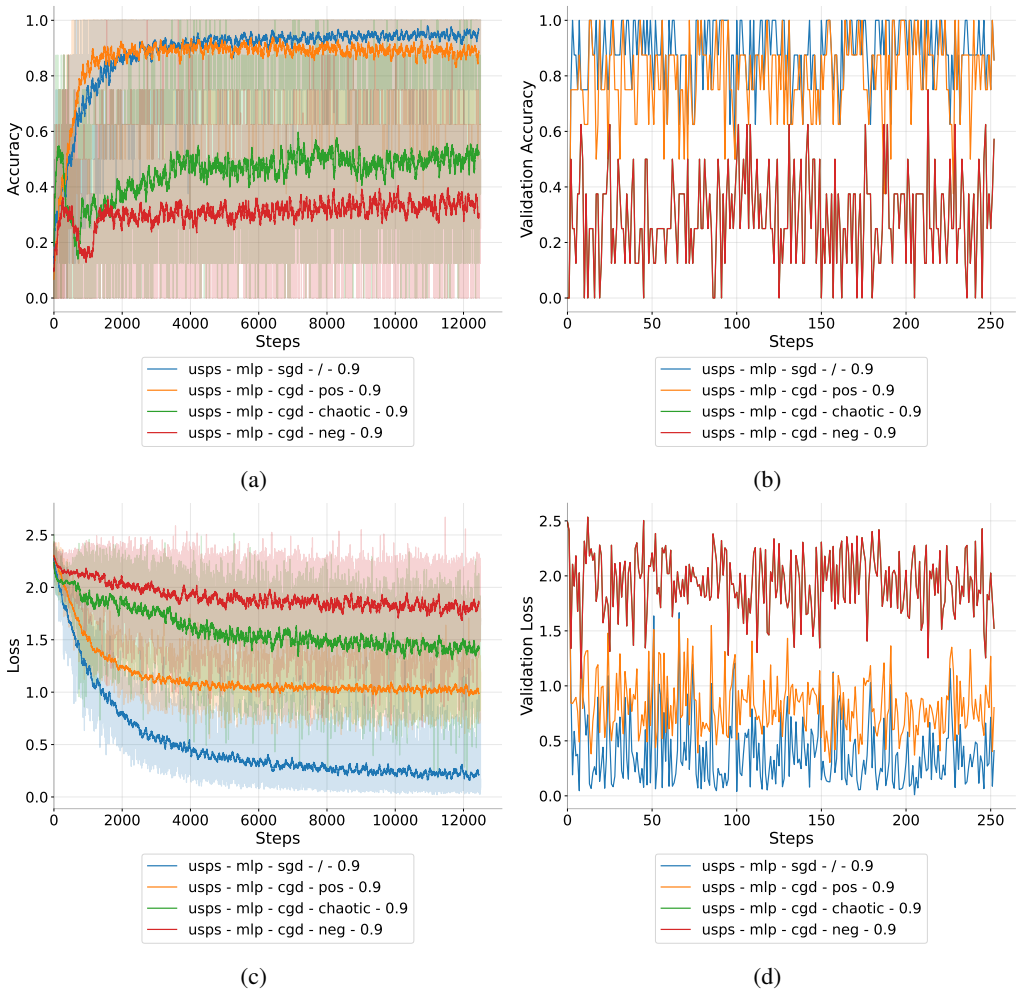


Figure B.1.1: Line plots of the (a) training accuracy and (b) validation accuracy, as well as (c) training loss and (d) validation loss for an MLP (relu activation) trained on the USPS dataset with momentum (0.9). The models trained with CGD use pruning of the full chaotic Hessian spectrum at every time step. The smoothed, solid lines are obtained through local averaging (window size 50) of the respective sequential data. Models trained with pruning of the full chaotic/negative spectrum show much slower convergence than the model trained with full positive spectrum pruning and reach lower performance in the observed time range. All models are initialized with the same random seed.

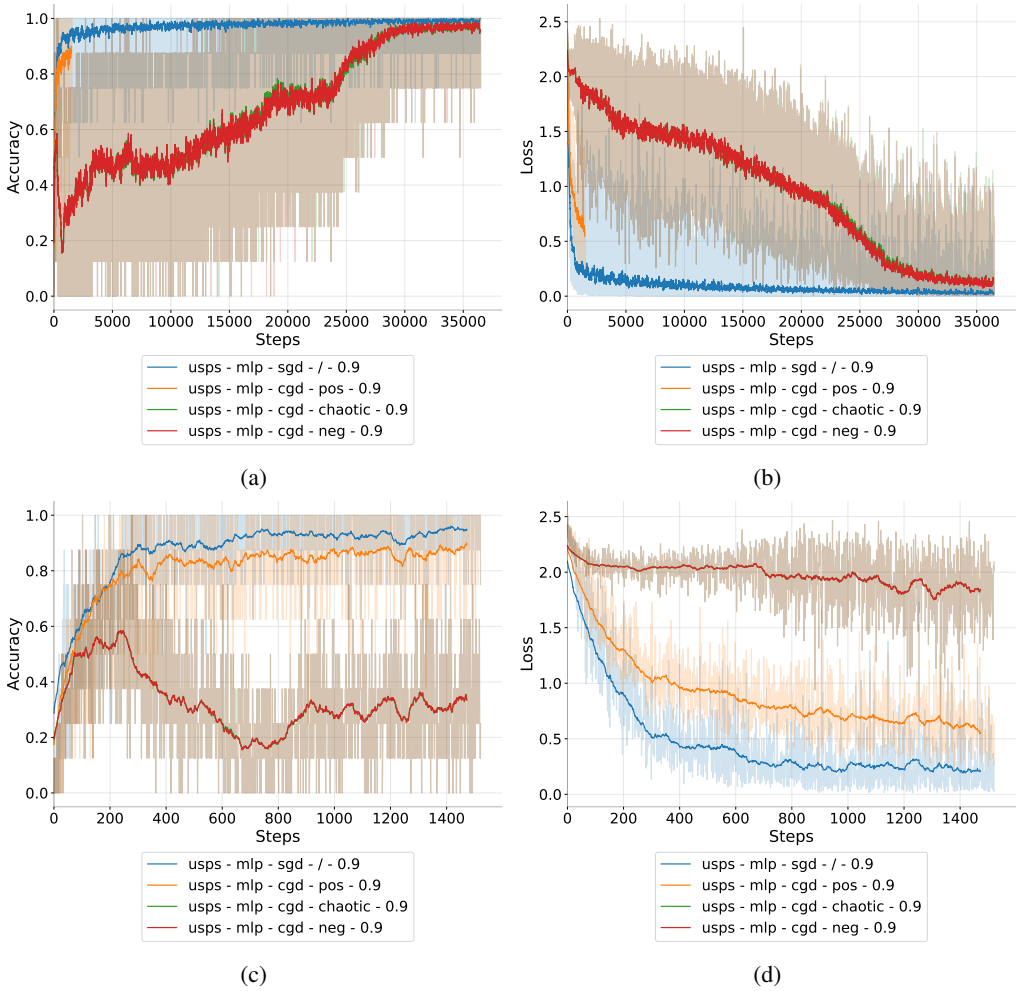


Figure B.1.2: Line plots of the (a)/(c) accuracy and (b)/(d) loss curves for an MLP (relu activation) trained on the USPS dataset with momentum (0.9). The models trained with CGD use pruning of the full chaotic, negative or positive Hessian spectrum at every time step. For greater clarity, the plots in the bottom row show a smaller time window. The smoothed, solid lines are obtained through local averaging (window size 50) of the respective sequential data. Using pruning of the full positive spectrum (yellow), the loss becomes singular, crashing the run. This suggests some information about positive curvature is required to avoid an excessively chaotic evolution of the system. However, the run with positive pruning still reaches high performance on both metrics before crashing. Meanwhile, the runs with full negative/chaotic pruning take much longer to converge, suggesting the negative Hessian eigenpairs are essential to enable fast training by SGD.

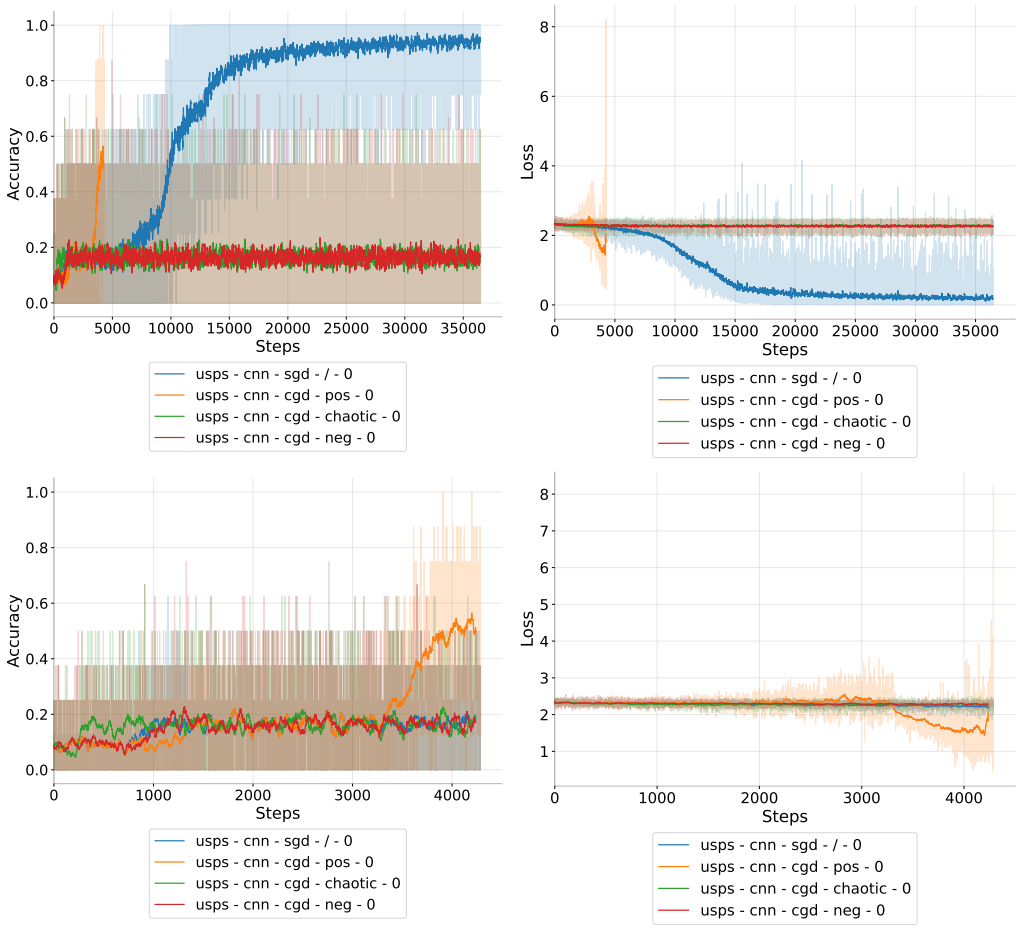


Figure B.1.3: Line plots of the (a)/(c) accuracy and (b)/(d) loss curves for a CNN trained on the USPS dataset without momentum. The models trained with CGD use pruning of the full chaotic, negative or positive Hessian spectrum at every time step. For greater clarity, the plots in the bottom row show a smaller time window. The smoothed, solid lines are obtained through local averaging (window size 50) of the respective sequential data. Using pruning of the full positive spectrum (yellow), the loss becomes singular, crashing the run. This suggests some information about positive curvature is required to avoid an excessively chaotic evolution of the system. However, the run with positive pruning still reaches high performance on both metrics before crashing. Meanwhile, the runs with full negative/chaotic pruning are unable to achieve significant improvements on both metrics.

FashionMNIST

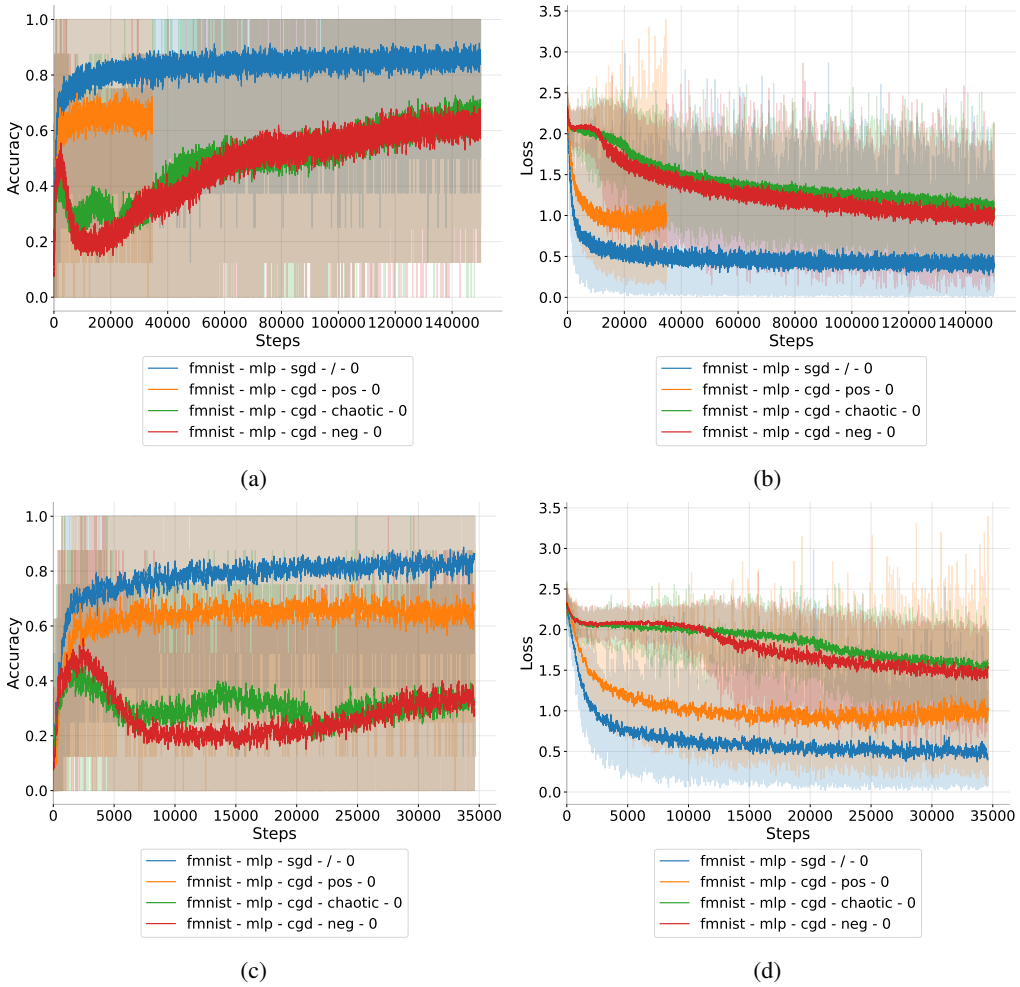


Figure B.1.4: Line plots of the (a)/(c) accuracy and (b)/(d) loss curves for an MLP trained on the FashionMNIST dataset without momentum. The models trained with CGD use pruning of the full chaotic, negative or positive Hessian spectrum at every time step. For greater clarity, the plots in the bottom row show a smaller time window. The smoothed, solid lines are obtained through local averaging (window size 50) of the respective sequential data. Using pruning of the full positive spectrum (yellow), the loss becomes singular, crashing the run. The run with positive pruning quickly reaches high performance on both metrics before crashing. The runs with full negative/chaotic pruning need much longer to reach a comparable performance on the training metrics.

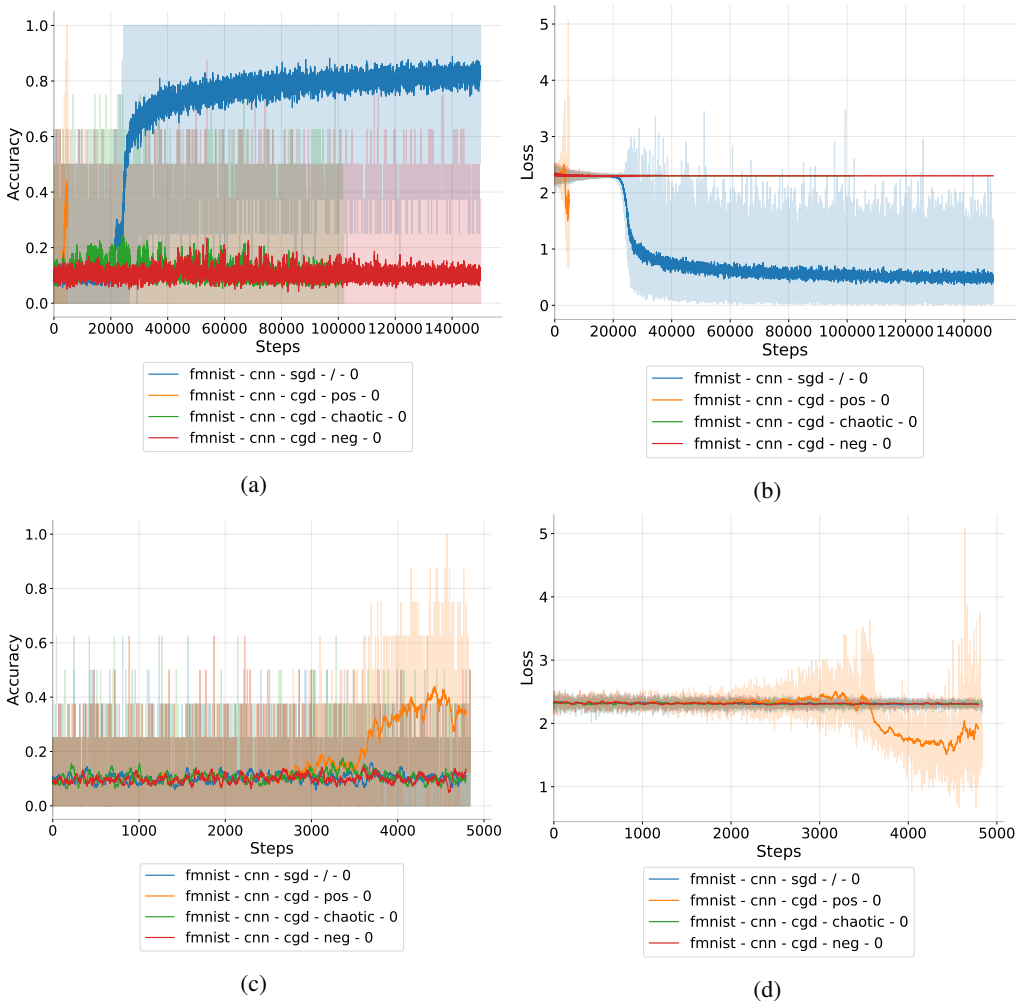


Figure B.1.5: Line plots of the (a)/(c) accuracy and (b)/(d) loss curves for a 2D CNN trained on the FashionMNIST dataset without momentum. The models trained with CGD use pruning of the full chaotic, negative or positive Hessian spectrum at every time step. For greater clarity, the plots in the bottom row show a smaller time window. The smoothed, solid lines are obtained through local averaging (window size 50) of the respective sequential data. Using pruning of the full positive spectrum (yellow), the loss becomes singular, crashing the run. The run with positive pruning quickly reaches high performance on both metrics before crashing. The runs with full negative/chaotic pruning need much longer to reach a comparable performance on the training metrics.

B.2 Pruning Experiments: Eigenspaces

USPS

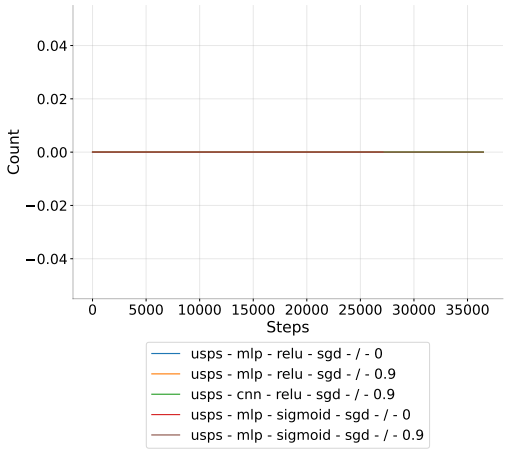


Figure B.2.1: Count of positive eigenvalues of the Hessian that satisfy the chaos criterion $\lambda_i > \frac{2}{\gamma}$ throughout the training, for several runs. The count is effectively zero.

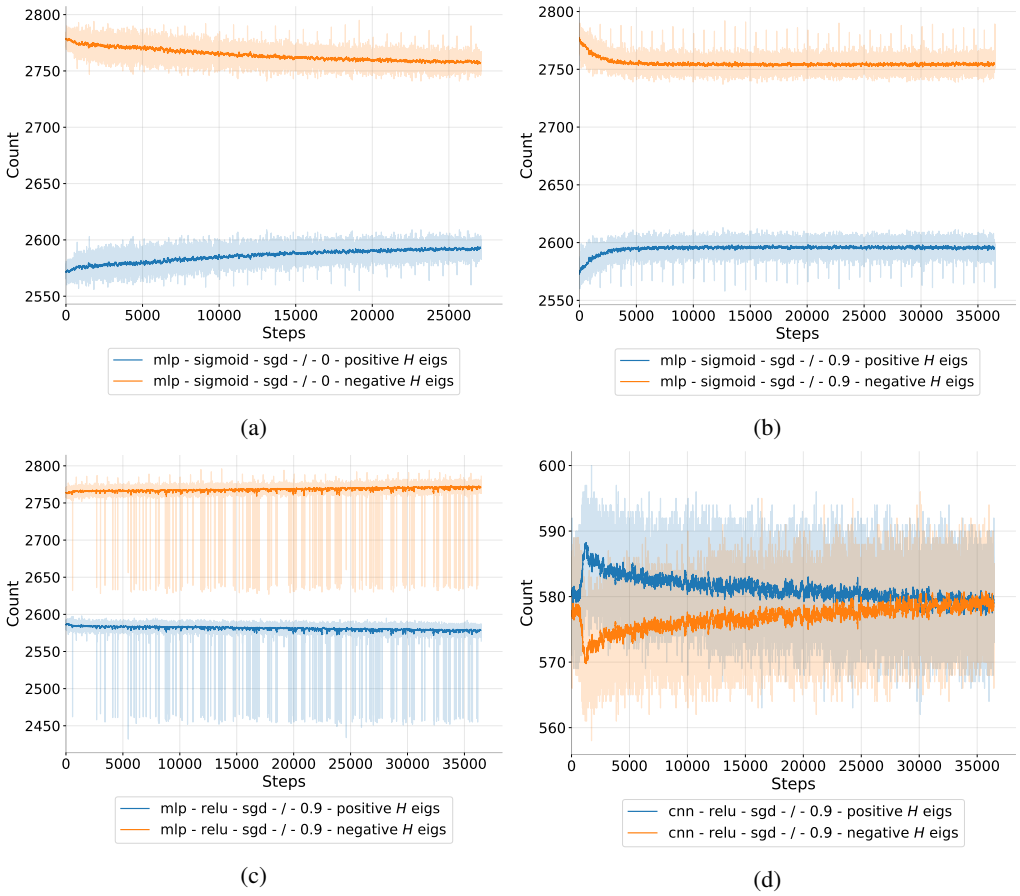
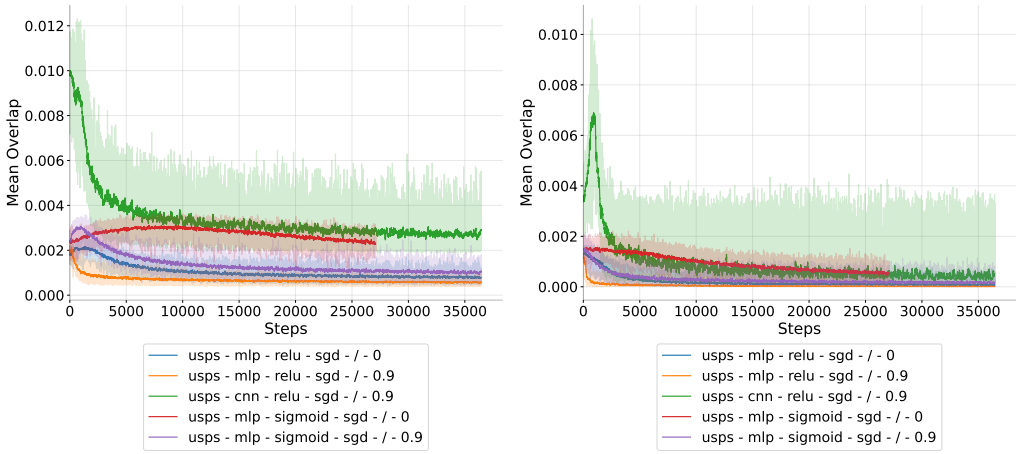


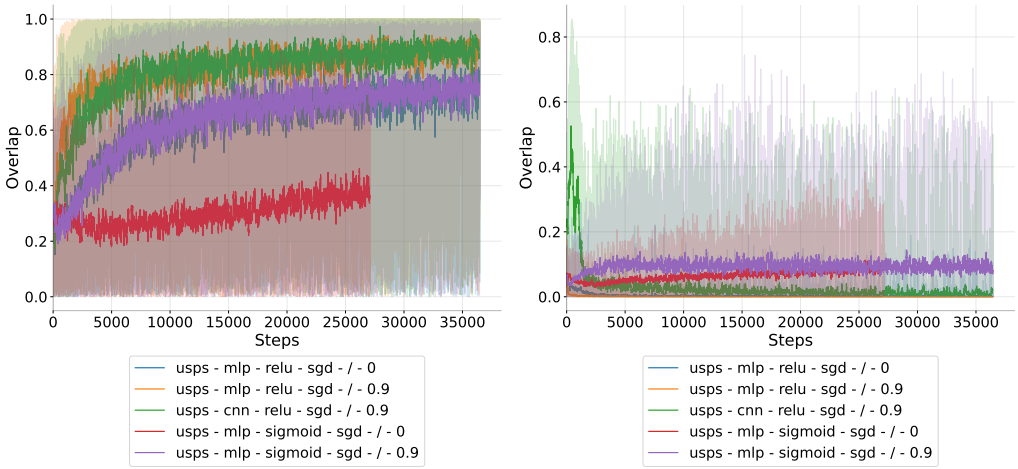
Figure B.2.2: **Number of positive** ($\lambda_i > 0$) and **negative** ($\lambda_i < 0$) eigenvalues of the Hessian for different runs on the USPS dataset. For the MLP models with sigmoid activation (a) and (b), the number of positive eigenvalues increases as training proceeds, while the number of negative eigenvalues decreases (regardless of momentum). With ReLU activation (c), the number of positive and negative eigenvalues is subject to strong fluctuations, although the positive eigenvalue count seems to decrease very slightly, opposed to the sigmoid runs. For a CNN (d), the count of positive eigenvalues increases very quickly, suggesting a very fast exploration phase in agreement with our model, but eventually relaxes back to the initial levels. The evolution of the negative eigenvalue count is complementary.



(a) Positive-gradient overlap

(b) Negative-gradient overlap

Figure B.2.3: Average overlaps between the normalized (a) positive Hessian eigenvectors and the gradient, and (b) negative Hessian eigenvectors and the gradient, where each curve corresponds to a different run configuration for the USPS dataset. The overlap between two vectors is measured as absolute cosine similarity. For all models, the positive spectra have on average a greater overlap with the gradient than the negative spectra, and the overlaps decrease as training progresses.

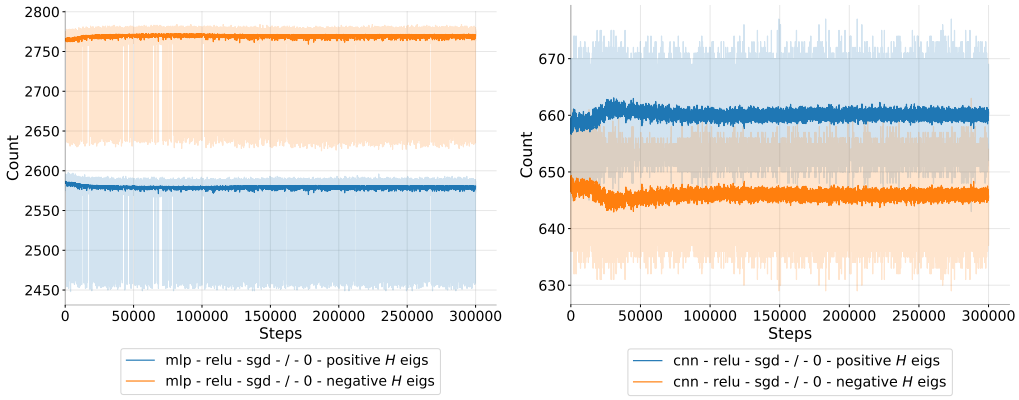


(a) Top-gradient overlap

(b) Bottom-gradient overlap

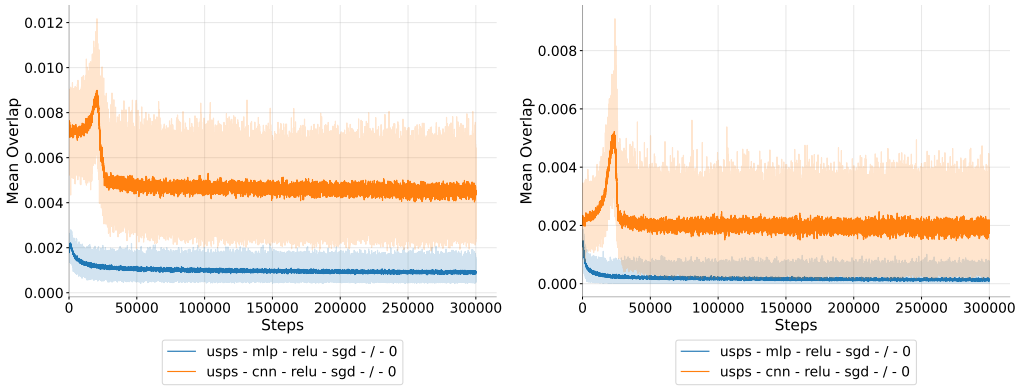
Figure B.2.4: Overlaps between the (normalized) (a) top Hessian eigenvector and the gradient, and (b) the minimum (i.e. most negative) Hessian eigenvector and the gradient, where each curve corresponds to a different run configuration. The overlap between two vectors is measured as absolute cosine similarity. The overlap with the top eigenvector is on average bigger than with the minimum eigenvector, in agreement with the findings of Gur-Ari et al. (2018). Although subject to wild fluctations, the overlap with the top eigenvector increases on the long run during training, while the average overlap with the minimum eigenvectors decreases quickly at the beginning of the training (after an initial increase for the MLP), suggesting the negative spectrum of the Hessian is more important during an initial **discovery phase**, while the positive spectrum becomes more important during a subsequent **discovery phase**.

FashionMNIST



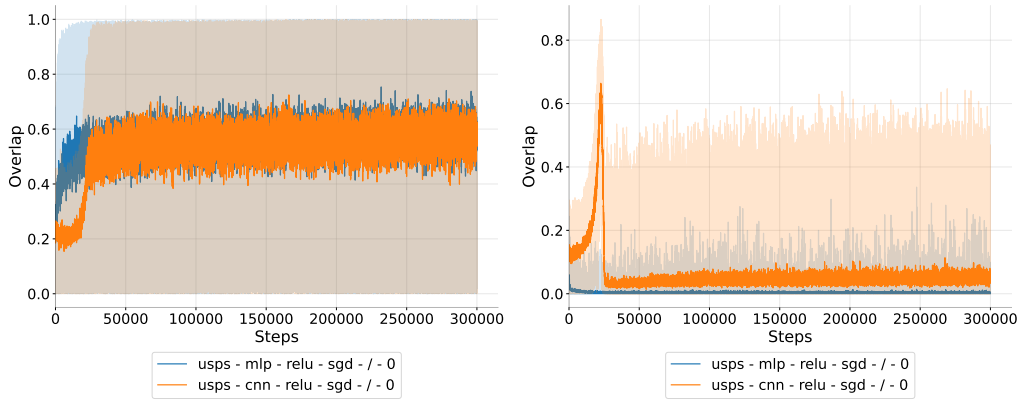
(a) (b)

Figure B.2.5: Number of positive ($\lambda_i > 0$) and negative ($\lambda_i < 0$) eigenvalues of the Hessian for different runs on the FashionMNIST. The eigenvalues behave analogously as for the USPS dataset. In the MLP with ReLU, the positive eigenvalue count slightly decreases at the beginning of the training, while for the CNN, the positive eigenvalue count initially increases an then relaxes to lower levels. The evolution of the negative eigenvalue count is complementary.



(a) Positive-gradient overlap (b) Negative-gradient overlap

Figure B.2.6: Average overlaps between the (normalized) (a) positive Hessian eigenvectors and the gradient, and (b) negative Hessian eigenvectors and the gradient, where each curve corresponds to a different run configuration for the USPS dataset. The overlap between two vectors is measured as absolute cosine similarity. The situation here is similar to the USPS dataset: For all models, the positive spectra have on average a greater overlap with the gradient than the negative spectra, and the overlaps decrease as training progresses.



(a) Top-gradient overlap

(b) Bottom-gradient overlap

Figure B.2.7: Overlaps between the (normalized) (a) top Hessian eigenvector and the gradient, and (b) the minimum (i.e. most negative) Hessian eigenvector and the gradient, where each curve corresponds to a different run configuration. The overlap between two vectors is measured as absolute cosine similarity. The situation here is similar to the USPS dataset: The overlap with the top eigenvector is on average bigger than with the minimum eigenvector, in agreement with the findings of Gur-Ari et al. (2018). Although subject to wild fluctuations, the overlap with the top eigenvector increases on the long run during training, while the average overlap with the minimum eigenvectors decreases quickly at the beginning of the training for most runs.

B.3 Additional

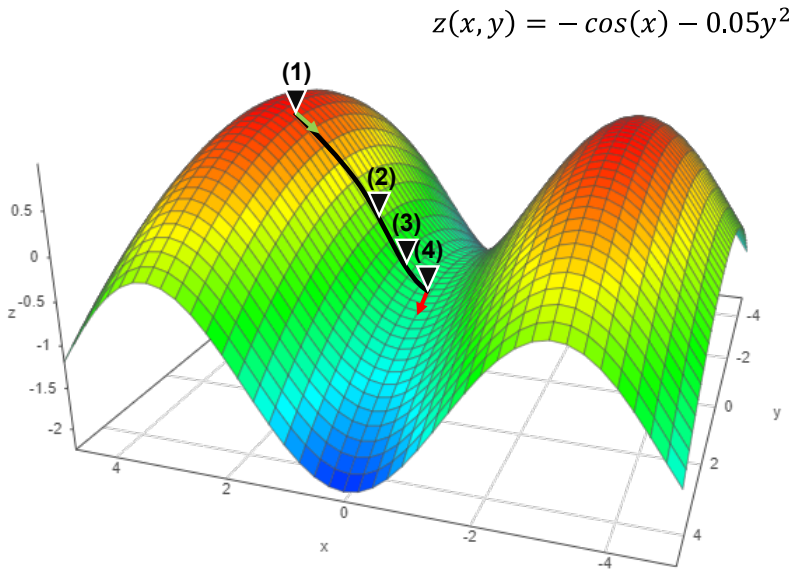
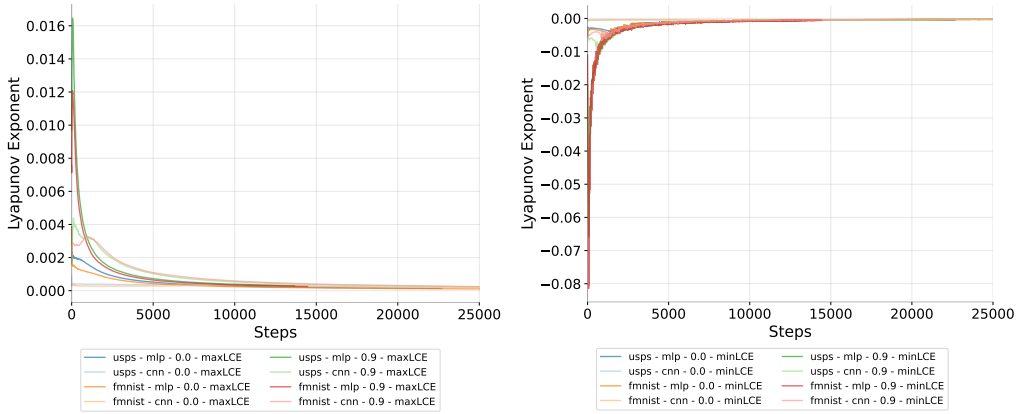


Figure B.3.1: A simplified model of training by SGD illustrating how negative curvature precedes positive curvature: At the beginning of the training, the optimizer is near a high-dimensional local maximum or saddle-point (1), from which the loss can be reduced by updating the parameters in directions of high negative curvature. As the optimizer passes the inflection point (2) along that given axis, the curvature along that axis becomes positive (3), but the potential for loss improvements in this direction is smaller at this point. In the vicinity of the local minimum (4) along this axis (actually a saddlepoint), the optimizer encounters new axes of high negative curvature and start moving in those directions. Image generated using `https://academo.org/demos/3d-surface-plotter`.

C Further Analysis on Globally Chaotic Dynamics

C.1 Lyapunov Exponents



(a) Maximum LCE (mLCE)

(b) Minimum LCE

Figure C.1.1: Maximum (a) and minimum (b) Lyapunov characteristic exponents (LCE) of the finite-time Lyapunov matrix at every time step for different runs with SGD. Across all run configurations, we observe a quick drop of the maximum LCEs to 0, suggesting the dynamics start out chaotic, but converge against edge-chaotic as the training advances. Interestingly, the minimum LCEs increase, suggesting an overall transition of the system to edge-chaotic behaviour. Although not visible in this plot, the minimum LCE is less stable than the maximum LCE, since the minimum eigenvalues of the Lyapunov matrix (before applying the natural logarithm) occasionally become 0, meaning the Lyapunov exponent would be $-\infty$. The respective points are not visible in the plot.

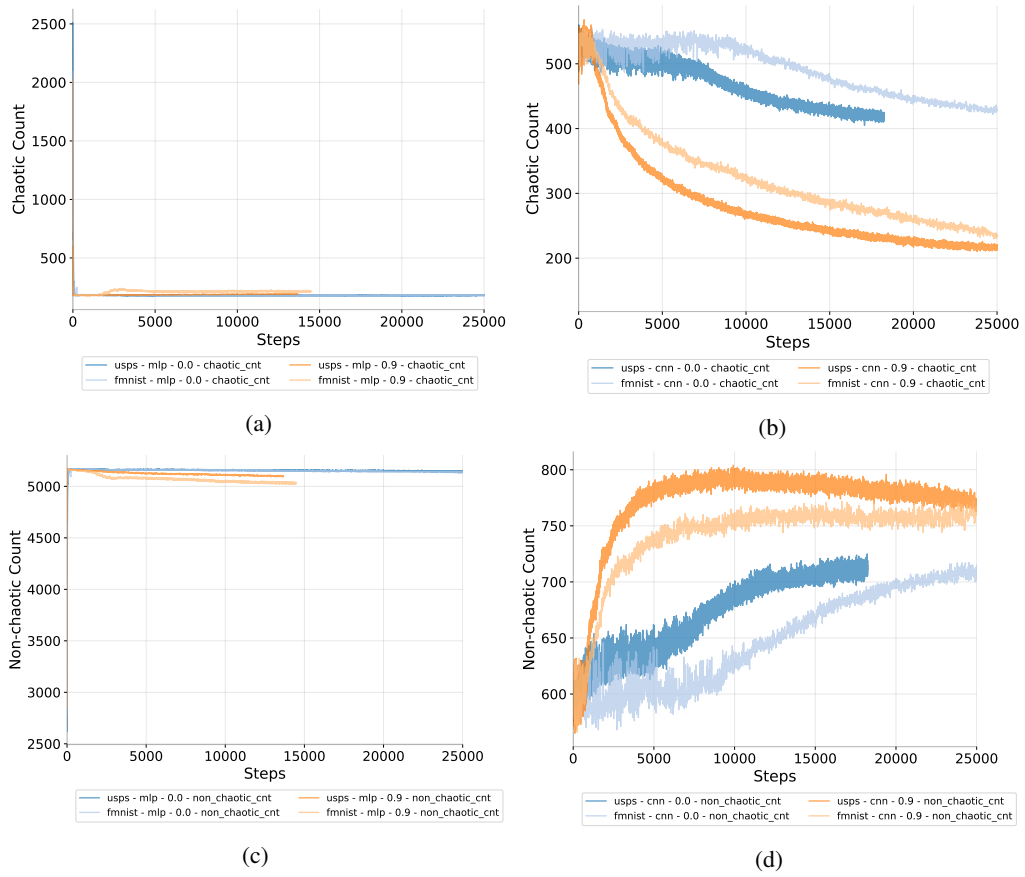


Figure C.1.2: Chaotic (a)/(b) and non-chaotic count (c)/(d) for MLP (a)/(c) and CNN (b)/(d). The counts are defined as $\dim(V_{\text{chaotic}}(t))$ and $\dim(V_{\text{non-chaotic}}(t))$, respectively. The chaotic count for the MLP drops quickly from a high value to a low value and stays almost constant afterwards. For the CNN the chaotic count drops more slowly. All of this can be said for the non-chaotic count vice-versa as the non-chaotic count is $D -$ chaotic_count where D is the total number of parameters of the model. All different plots back up the thesis that training is initially chaotic but quickly/slowly stabilizes in a less chaotic regime over time.

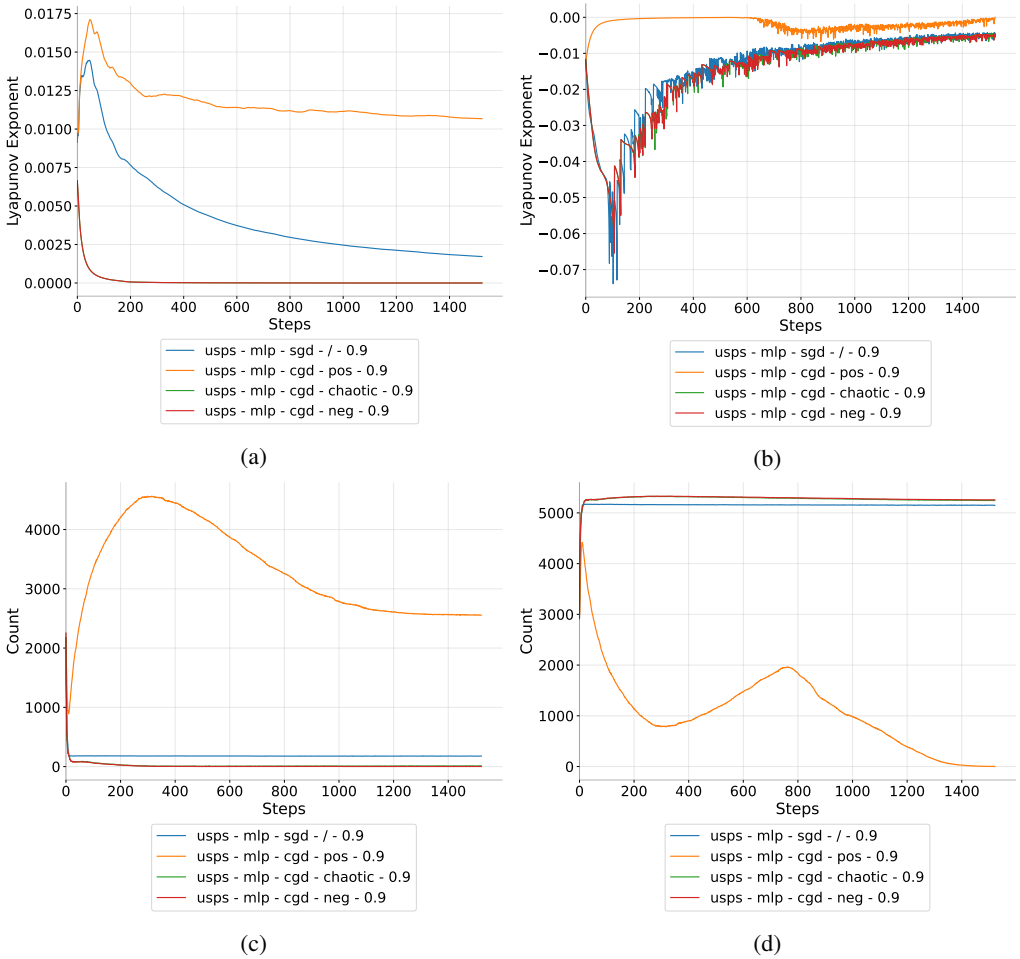


Figure C.1.3: Maximum (a) and minimum (b) Lyapunov exponent for an MLP trained on the USPS dataset with momentum (0.9). Even with momentum, CGD is effective at quickly dampening the maximum Lyapunov exponent to values close to 0 for chaotic and negative pruning. Without pruning, the Lyapunov exponent also converges to 0, but at a slower pace. The minimum Lyapunov exponents converge against values close to 0, indicating the chaotic dynamics overall converge against edge-chaotic behaviour. This is also reflected by count of (c) chaotic and (d) unchaotic eigenvalues. The number of chaotic eigenvalues quickly drops in regular SGD and then stabilizes at a low (but nonzero) level. For positive pruning, the number of nonchaotic eigenvalues drops to 0, causing the training to become unstable.

C.2 Subspace Similarity of Chaotic spaces

In this section we provide an argument for why it is sound to talk about global chaos when we approximate the Lyapunov eigenspace with the eigenspaces $V_{\text{chaotic}}(T_{\text{conv}})$ and $V_{\text{non-chaotic}}(T_{\text{conv}})$ until training convergence T_{conv}.

Figure C.2.1 shows that at time point T_{conv} the chaotic subspace and thus due to orthogonality $V_{\text{chaotic}}(t) \perp V_{\text{non-chaotic}}(t)$ the space $V_{\text{non-chaotic}}(t)$ stabilize over time. Where for fixing early time steps t_0, the similarity stabilizes in a lower regime instead of dropping towards randomness, for later time steps $t_0 = T_{\text{conv}}$ the similarity stays close to one. The rate of change should be lower for even later time steps, which is when we saved the eigenspaces $V_{\text{chaotic}}(T_{\text{end}}), V_{\text{non-chaotic}}(T_{\text{end}})$ for the different perturbation strategies. Although for some experiments, e.g. MLP with momentum on FashionMNIST, one can argue that longer training is needed as the subspace similarity still drops significantly.

For comparison, we also estimated the overlap of two random subspaces of the same size as $V_{\text{chaotic}}(t_0)$ and $V_{\text{chaotic}}(t)$ and averaged it over all different runs: Let $d_{(0,i)} = \dim(V_{\text{chaotic}}(t_0, \theta_i))$ and $d_{(t,i)} = \dim(V_{\text{chaotic}}(t, \theta_i))$ for different experiments $i = 1, ..., 8 =: N$. The baseline is defined as

$$\frac{1}{N} \sum_{i=1}^{N} \mathop{\mathbb{E}}_{V_{(0,i)}, V_{(t,i)})} [\text{overlap}(V_{(0,i)}, V_{(t,i)})] \tag{65}$$

where $\dim(V_{(0,i)}) = d_{(0,i)}$ and $\dim(V_{(t,i)})) = d_{(t,i)}$.

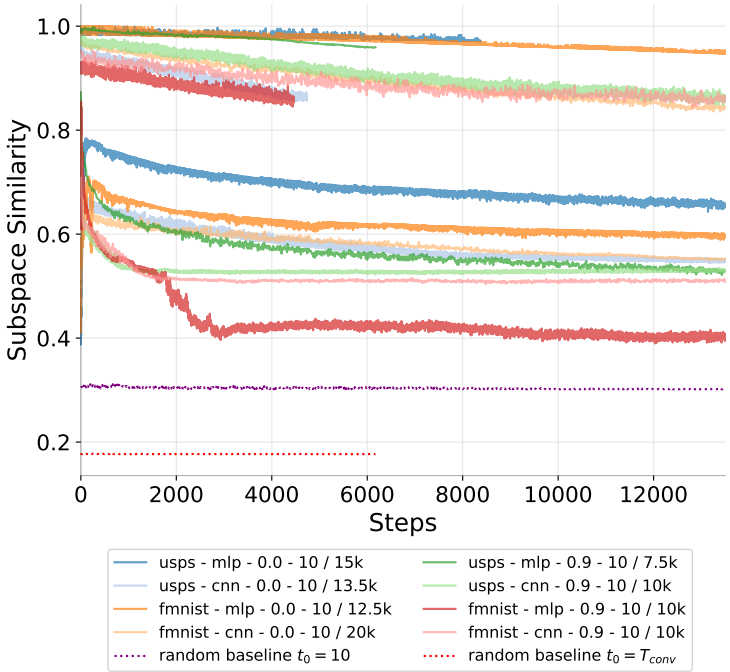
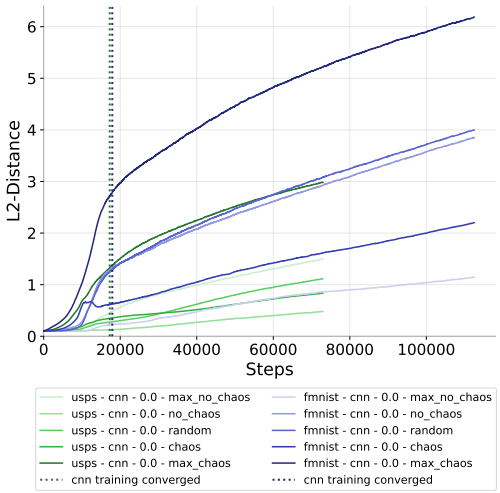
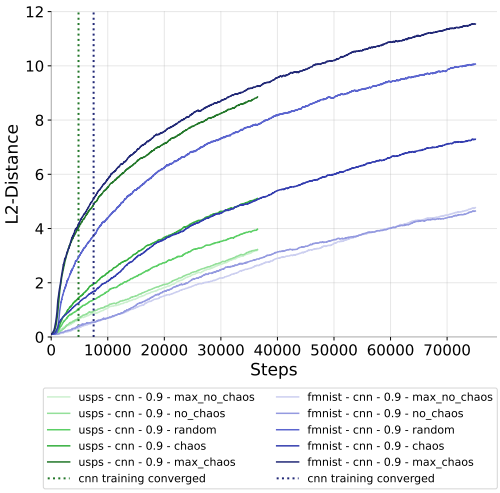


Figure C.2.1: Subspace similarities measured for the same configuration of models (CNN, MLP) and datasets (USPS, FashionMNIST) with and without momentum as we used in our other experiments. We fix a chaotic subspace $V_{\text{chaotic}}(t_0)$ (see section 4.2) at the start $t_0 = 10$ and the end of trainng $t_0 = T_{\text{conv}}$, e.g. indicated with 10 / 10k, and measure the overlap between the fixed subspace and following chaotic subspaces $V_{\text{chaotic}}(t)$ for $t > t_0$. The two plots are visualized with the same color for each model/dataset configuration. We modify the calculation of the metric in order to calculate the similarities between subspaces of different dimensionalities (see A.11). We also plot a random baseline for comparison which is defined as the mean of the expected overlap between random subspaces of the same dimensionality (see equation (65).

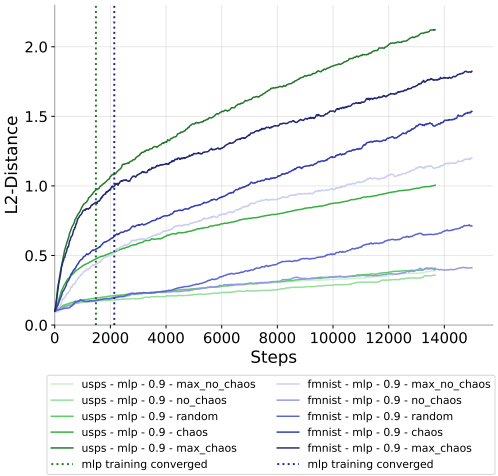
C.3 Distance Experiments



(a) CNN without momentum

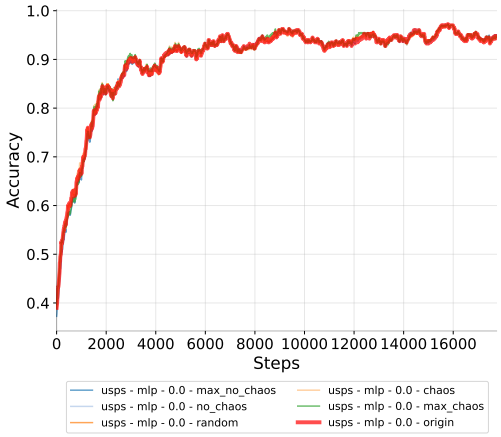


(b) CNN with momentum

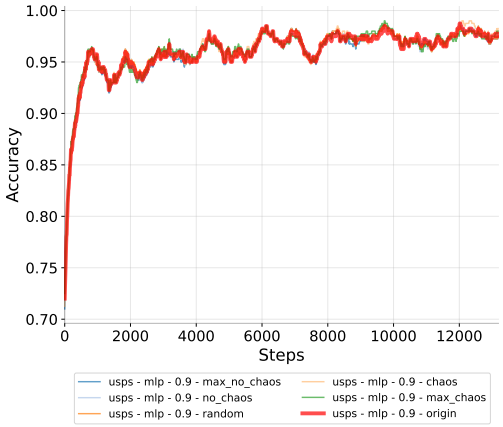


(c) MLP with momentum

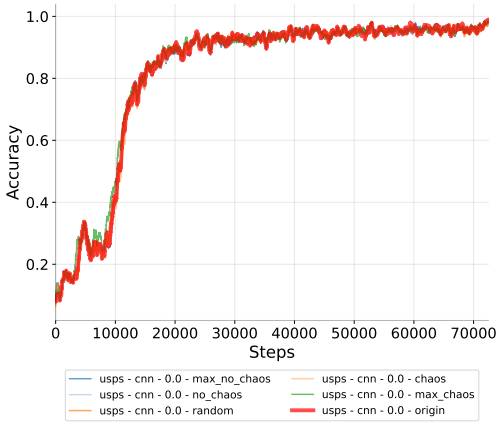
Figure C.3.1: Further distance experiments of Figure 3 (a) with a perturbation magnitude of 0.1. We only see minor differences between datasets so it seems that chaotic behavior is more affected by the model choice and general training parameters. In fact, the CNN model exhibits more global chaotic behavior while having less parameters than the MLP, but more experiments have to be conducted to gather a larger sample size to verify this observation. The models trained with momentum have a greater distance divergence than without momentum which could be explained by the process carrying along energy. This could lead to bigger gradients when entering the valley of the local minimum.



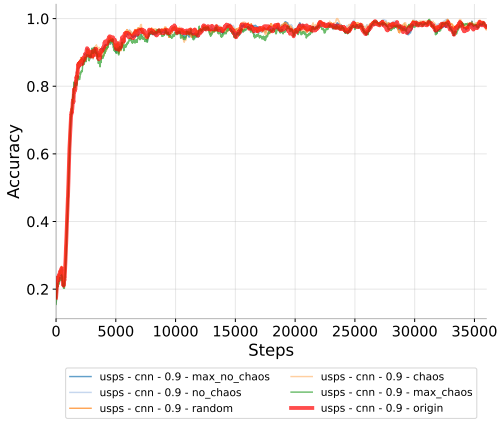
(a) MLP without momentum



(b) MLP with momentum

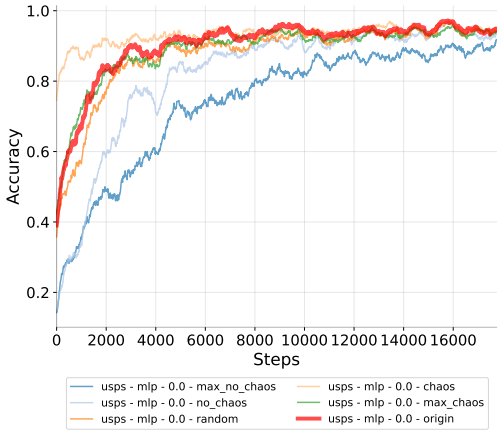


(c) CNN without momentum

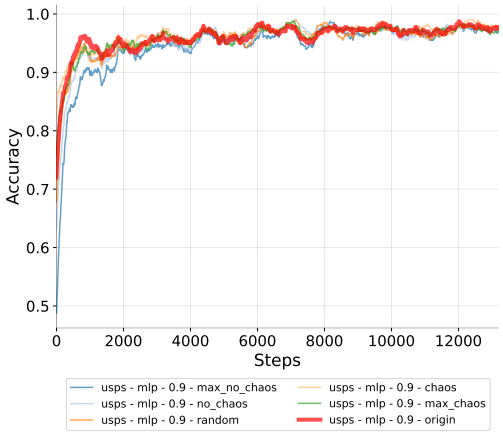


(d) CNN with momentum

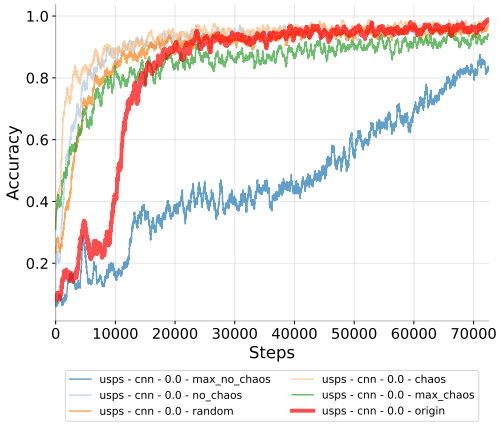
Figure C.3.2: Averaged accuracy (moving average over the last 500 steps) of runs with a perturbation magnitude of $\varepsilon = 0.1$ over the different strategies. We see no significant difference between the models with different perturbation strategies compared to the original model.



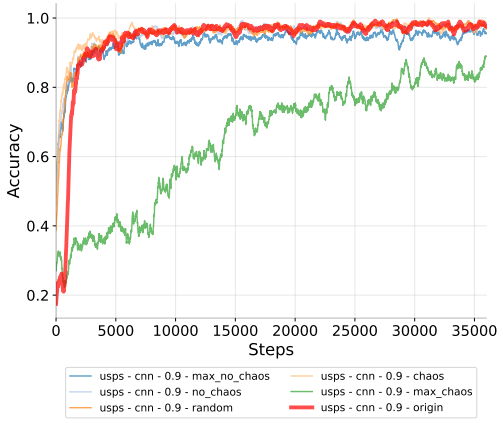
(a) MLP without momentum



(b) MLP with momentum



(c) CNN without momentum



(d) CNN with momentum

Figure C.3.3: Averaged accuracy (moving average over the last 500 steps) of runs with a perturbation magnitude of $\varepsilon = 10$ over the different strategies. They generally perform slightly worse with the exception of the chaotically perturbed model. Surprisingly, we see that this model consistently outperforms the original model, particularly in the beginning of training. This could be connected to what we find for the pruning experiments: Chaos seems to be beneficial especially at the start of the training. Also, the maximal non chaotic perturbation always seems to lead to a worse accuracy than all other strategies with the exception of (d). A further analysis has to be provided to better understand both phenomenons.

C.4 Choice of Perturbation Magnitude

We chose to discuss the experiments with magnitude $\varepsilon = 0.1$, because it shows the expected behavior over all runs we tracked. So this can be said to be cherry picked. Generally, we expect the maximal chaotic perturbation to have the largest distance, and then chaotic, random, non-chaotic and maximal non-chaotic to have smaller distances in that order. For bigger epsilon, we argue that locality breaks and the error of approximation with the Lyapunov matrix gets too high. Surprisingly, we do not find to have the expected order of the perturbation strategies for smaller values of ε, as well. For these cases, we see that the trajectories become more noisy. In fact, this noise seems to be larger than the perturbation magnitudes such that the trajectories end up in random locations in the range of what we see in Figure 3. We calculated the absolute mean positional divergence as a measure of order for every distance over all experiments and show that smaller and higher choices of ε become more random the farther away they are from 0.1 (see Figure C.4.1). What causes the noise is still not clear

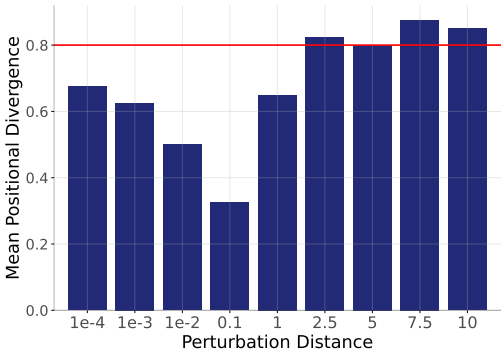
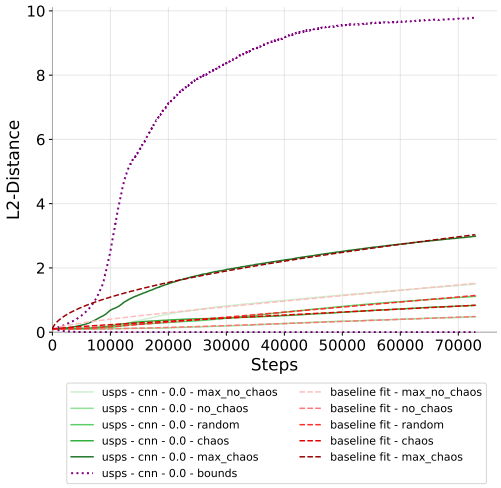


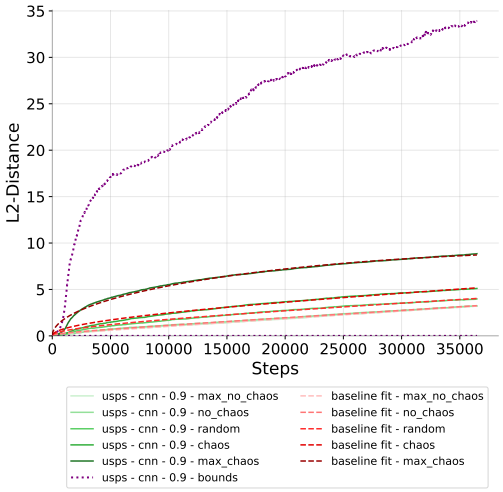
Figure C.4.1: The mean positional divergence: $1/(2N)\sum_{i=1}^{N}|\text{pos}_{\text{strat}(i)} - i|$ over all $N = 5$ different strategies. We expect the strategies to be ordered in the following way: maximal chaotic, chaotic, random, non-chaotic, maximal non-chaotic (from highest divergence to lowest). The red line marks the expected value for a randomly chosen order.

to us as it cannot be caused by SGD due to the fixed random seed. It is probably a numerical error and to understand this phenomenon additional research is needed.

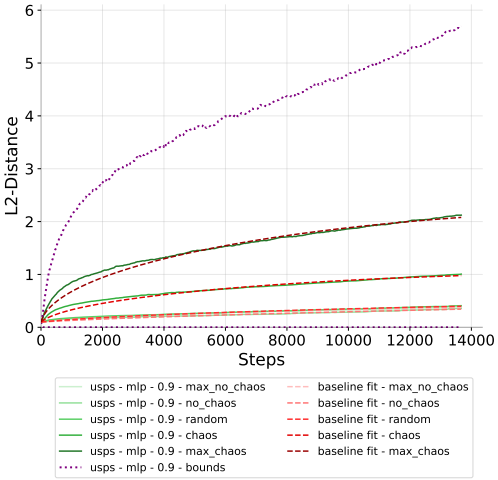
C.5 Theoretical Distance Analysis



(a) CNN without momentum



(b) CNN with momentum



(c) MLP with momentum

Figure C.5.1: Further theoretical distance experiments of Figure 3 (b) with a perturbation magnitude of $\varepsilon = 0.1$. This experiment shows that the bounds hold in practise even when they are not tight which leaves two possible options: Either the first order approximation that defines the bound (see equation 5) is not precise for $\varepsilon = 0.1$ or we were not able do determine the exact direction of maximal/minimal global chaos due to a numerical error (see section C.4)