NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

as a manuscript

Maxim Kodryan

TRAINING DYNAMICS AND LOSS LANDSCAPE OF NEURAL NETWORKS WITH SCALE-INVARIANT PARAMETERS

PhD Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow — 2023

The PhD Dissertation was prepared at National Research University Higher School of Economics.

Academic Supervisor: Dmitry P. Vetrov, Candidate of Sciences, National Research University Higher School of Economics.

1 Introduction

Topic of the thesis

Scale invariance is one of the key properties inherent in the parameters of most modern neural network architectures. Provided by the ubiquitous presence of layers of normalization of intermediate activations and/or weights, scale invariance, as the name implies, consists in the invariance of the function implemented by the neural network when its parameters are multiplied by an arbitrary positive scalar. In this work, we investigate the effects of this property on the training dynamics of neural network models, as well as its influence on the intrinsic structure of the loss landscape.

In the first part of the work, we consider and analyze the effect of periodically repeating cycles of convergence and destabilization when neural networks are trained using normalization and weight decay (WD) techniques. As shown by theoretical and empirical analysis, this behavior is a consequence of the competing influence of weight decay and scale invariance on the norm of neural network parameters. Thus, there is a periodic change of the sphere on which the model is trained, which ultimately leads to the observed periodic behavior of the optimization process.

In the second part of the work, we reveal the intrinsic structure of the loss landscape of neural networks with scale-invariant parameters by fixing the sphere on which the model is trained. It has been analytically and experimentally shown that in such a setting, three training regimes can be distinguished: convergence, chaotic equilibrium, and divergence. Each regime is characterized by a number of its specific features and allows to highlight certain properties of the intrinsic loss landscape of scale-invariant neural networks, which are also reflected in the actual practice of training neural network models, for example, when designing a learning rate schedule. The described effects of training scale-invariant models on the sphere are studied in various settings using both the more classical cross-entropy and Mean Squared Error (MSE) loss functions on classification problems, which has shown promise in recent studies [1].

Relevance

Despite the tremendous empirical progress in the field of deep learning in recent decades, the search for a satisfactory justification of the principles of design and inference of deep

neural network models is still extremely relevant [2]. Many questions are still left unanswered. They include both particular issues related to isolated effects of the training process and the properties of final solutions, for example, double descent of the test loss [3, 4] or the so-called grokking [5], and global problems related to the internal structure of the loss landscape and optimization dynamics of neural networks, e.g., the ability of modern neural networks to memorize the entire training set [6], the presence of "minefields" [7], the connectivity of modes [8, 9] in the loss landscape, and similar overparameterized learning phenomena [10]. The interpretability and predictability of training and inference of deep learning models is a necessary condition not only for their wider and safer application, but also for the development of new ways to improve them, which will not only rely on practical intuition and heuristics, but on a rigorous scientific method as well.

Normalization techniques such as Batch Normalization (BN) [11], Layer Normalization [12], or Weight Normalization [13] are commonly used in modern neural network architectures and are shown empirically to often help stabilize the learning process and improve the final quality of models. However, their use further complicates the understanding of the processes occurring in neural networks. Despite some progress in understanding certain properties provided by the use of normalization in neural networks, many questions still remain unsolved [14, 15]. In particular, the role of normalization in determining the effective structure of the loss function surface is not completely clear, as well as how exactly it affects the learning dynamics of modern normalized neural networks. These issues have gained particular relevance in recent years due to the discovered effects of singularity and instability in certain modes of application of normalization techniques [16, 17, 18, 19] despite they are believed to provide stabilization of the learning process of neural networks.

Perhaps the most general, and therefore key, consequence of using arbitrary normalization techniques in a neural network architecture is the scale invariance property of the weights of this network preceding the normalization layers. Due to the fact that normalization is usually applied after almost every hidden layer of the neural network, in practice it turns out that the vast majority of model parameters obtain this property. This circumstance highlights the main difference between normalized neural networks and networks without the use of normalization techniques, so it cannot be ignored when studying the impact of normalization on optimization dynamics and the loss landscape. Thus, the actual research and interpretation of normalization techniques must rely on the

property of scale invariance and its consequences, as demonstrated by recent work in this area [18, 19, 20, 21, 22, 23, 24, 25].

The first part of this work is devoted to discovering, researching and explaining the effect of periodic behavior of neural networks training with normalization and weight decay techniques. Weight decay is a widely used technique for training machine learning models, which consists in scalar multiplication of parameters by a given positive coefficient less than one after each training iteration and acts as a generalized classical L_2 regularizer [22, 26, 27]. Despite the fact that scale-invariant models by definition do not depend on the actual value of the parameters norm, it turns out that weight decay nevertheless significantly affects the training dynamics of such models due to a non-trivial change in the so-called effective learning rate (ELR). Prior work dedicated to investigating this effect has come up with some controversy about how this influence determines the final behavior of the optimization dynamics. Some share the view that training normalized models using weight decay must eventually reach a state of equilibrium, when all observable metrics, including the value of the effective learning rate, the norm of parameters, empirical risk, etc., stabilize in some fixed value, which generally has a beneficial effect on learning [19, 20, 28, 23]. Others, on the contrary, argue that WD after a certain number of training iterations will bring the weight norm too close to zero, which will lead to numerical instabilities and divergence of the optimization process [17, 18, 19]. In this work, the described contradiction is resolved and it is demonstrated that both positions are valid in a certain sense. On the one hand, the learning dynamics of normalized models with WD indeed constantly encounters instabilities for the above reason. On the other hand, such instabilities are consistent, which leads to periodic behavior of the learning dynamics. This periodic behavior has a regular structure, which makes it possible, among other things, to consider it as a kind of generalization of the equilibrium principle. In this work, we provide a detailed experimental and theoretical analysis, describing and substantiating the mechanisms behind such periodic behavior. The main paper on this topic also explores its implications and effects in relation to training modern deep learning models.

In the second part of the work, we study the loss landscape structure of scale-invariant neural networks on their intrinsic domain, i.e., the sphere. Since scale-invariant models inherently do not change when the parameters move along the radial direction from the origin, their natural domain can be considered a sphere instead of the entire parameter space. Accordingly, their training trajectory can also be effectively viewed through the

projection onto the sphere in order to better understand how the optimization dynamics works on the true domain. However, the effective learning rate, which is responsible for the optimization rate on the unit sphere, changes non-trivially during standard training of scale-invariant models, especially with the use of WD, as, in particular, was shown in the previous part of the work. Thus, it is difficult to study the intrinsic loss landscape, since the size of the effective optimization step cannot be controlled even when fixing the standard learning rate (LR). In this work, we solve this problem by switching to the optimization of fully scale-invariant neural networks directly on the sphere using the projected stochastic gradient descent (SGD) method. Such a training procedure eliminates the effect of a dynamically changing effective learning rate and fixes it to a given value by construction, since it eliminates the variability of the parameters norm during training and completely transfers the dynamics to the natural domain. This allows us to study in detail and in a controlled way the intrinsic structure of the loss landscape of scale-invariant neural networks. It turns out that training of scale-invariant neural networks on a sphere can be carried out in three regimes depending on the given ELR value: convergence, chaotic equilibrium, and divergence. Each regime possesses a number of distinctive features and reveals certain properties of the actual loss landscape structure, for example, the presence of a whole spectrum of functionally and geometrically different global minima corresponding to different ELR values of the first regime, high-sharpness zones preventing convergence and separating the first regime from the second, as well as local and global regions of stabilization of the optimization dynamics in the second training regime. Two papers of the author were dedicated to the study of the features of these regimes and their consequences on the training dynamics and the loss landscape of neural networks with scale-invariant parameters: the first one focuses on the study of the classical cross-entropy loss function and for the first time reveals the main properties of the three regimes, the second one considers the case of MSE loss for classification problems [1] and extends the results of the previous work. Among other things, these papers demonstrate how these regimes manifest themselves in standard training of modern deep learning architectures and how they can be used in practice, for example, to find optimal LR schedules.

The goal of this work is to reveal and study the features of the training dynamics and the structure of the loss landscape of neural networks with scale-invariant parame-

ters. This will improve the interpretability of modern neural network models that use normalization techniques.

2 Key results and conclusions

Contributions. The main contributions of this work can be summarized as follows:

1. We investigated the training dynamics of normalized neural networks in the entire parameter space with weight decay. We discovered and analyzed both experimentally and theoretically the effect of periodic behavior of such dynamics.

2. We resolved the contradiction that has developed in the literature regarding the result of this optimization dynamics (equilibrium vs. instability) via the described periodic behavior. We derived the generalized equilibrium principle.

3. We investigated the training dynamics of fully scale-invariant neural networks on their natural domain, i.e., the sphere. We discovered and analyzed both experimentally and theoretically three regimes of such training: convergence, chaotic equilibrium, and divergence; we also distinguished their main characteristics.

4. By studying these regimes, we revealed a number of properties of the intrinsic loss landscape of scale-invariant models, including the existence of a spectrum of various global minima, high-sharpness zones, and regions of stabilization of optimization dynamics.

5. Additionally, we studied the three regimes in the case of training with MSE loss function on classification problems.

Theoretical and practical significance. This work continues the general current trend in the field of deep learning to find and develop satisfactory justifications for the mechanisms behind the design and inference of neural network models. The focus of this work is on the principle of scale invariance provided by the use of normalization techniques that are ubiquitous in most modern architectures. The obtained results not only allow us to identify and explain the various properties of the training dynamics and the structure of the loss landscape of normalized models, but also help to generalize previous knowledge and develop more efficient ways to train neural networks. In particular, with the help of the revealed periodic behavior from the first part of the work, it was possible to resolve

the contradiction that has developed in the literature about the learning dynamics of normalized neural networks with weight decay, while the study of the properties of the identified three training regimes on the sphere from the second part served as the basis for interpreting and developing learning rate schedules. The derived theoretical results make it possible to strengthen and formalize the obtained empirical intuition, and in themselves are of interest as a working mathematical model describing scale-invariant dynamics.

Key aspects/ideas to be defended:

1. The discovered periodic behavior of training dynamics of normalized neural networks with weight decay, its experimental and theoretical analysis.

2. The derived principle of generalized equilibrium, resolving the conflict of two contradictory positions regarding the dynamics of such training: equilibrium verses instability.

3. Three discovered regimes of training fully scale-invariant neural networks on the sphere using both cross-entropy and MSE loss functions: convergence, chaotic equilibrium, and divergence; their experimental and theoretical analysis.

4. The revealed properties of the loss landscape of scale-invariant neural networks on the sphere: the spectrum of different global minima, high-sharpness zones, regions of stabilization of optimization dynamics, and others.

Personal contribution. In the first paper, the author formulated and proved all the presented theoretical results. The author made the main contribution to the review of related work, in particular, he established the existence of a contradiction regarding the result of the studied training dynamics and proposed its resolution through the discovered periodic behavior. The author also participated in setting up experiments, analyzing empirical results and writing the text together with Ekaterina Lobacheva and other co-authors.

In the second paper, the author also formulated and proved all the presented theoretical results. The author made the main contribution to the writing of the text and the review of related work. He participated with other co-authors in the analysis and interpretation of empirical results, including establishing the main characteristics of the three regimes of training on the sphere and their implications for the loss landscape. The

author also assisted in setting up experiments, in which the main role was played by Ekaterina Lobacheva and Maksim Nakhodnov.

In the third work, the author was one of the initiators of the study of three training regimes with MSE loss function, and also assisted the main author Maksim Nakhodnov in interpreting and systematizing the results, reviewing the literature, and setting up experiments.

Publications and probation of the work

The author is the main author in two first-tier publications and the second author in one second-tier publication on the dissertation topic.

* — authors with equal contribution.

First-tier publications

1. Ekaterina Lobacheva*, **Maxim Kodryan***, Nadezhda Chirkova, Andrey Malinin, Dmitry Vetrov. On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay. In Advances in Neural Information Processing Systems, 2021 (NeurIPS 2021). Vol. 34, pages 21545-21556. CORE A* conference.

2. **Maxim Kodryan***, Ekaterina Lobacheva*, Maksim Nakhodnov*, Dmitry Vetrov. Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes. In Advances in Neural Information Processing Systems, 2022 (NeurIPS 2022). Vol. 35, pages 14058-14070. CORE A* conference.

Second-tier publications

1. Maksim Nakhodnov, **Maxim Kodryan**, Ekaterina Lobacheva, Dmitry Vetrov. Loss Function Dynamics and Landscape for Deep Neural Networks Trained with Quadratic Loss. Published in Doklady Mathematics in 2022. Vol. 106, issue 1 (supplement), pages 43-62. The journal contains English translations of papers published in Doklady Akademii Nauk (Proceedings of the Russian Academy of Sciences), indexed in Scopus.

Reports at scientific conferences and seminars

1. Conference on Neural Information Processing Systems, December 2021. Topic: “On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay”.

2. Seminar Mathematical Machine Learning MPI MIS + UCLA, December 2021. Topic: "On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay".

3. Machine Learning Summer School by EMINES School of Industrial Management, July 2022. Topic: "On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay".

4. Seminar of the Bayesian methods research group, October 2022. Topic: "Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes".

5. Conference Fall into ML, November 2022. Topic: "Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes".

6. Conference on Neural Information Processing Systems, December 2022. Topic: "Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes".

7. Seminar AIRI AIschnitsa, December 2022. Topic: "Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes".

8. Conference of the Faculty of Computer Science in Voronovo, June 2023. Topic: "Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes".

Volume and structure of the work. The thesis contains an introduction, contents of publications, and a conclusion. The full volume of the thesis is 103 pages.

3 Content of the work

3.1 Periodic behavior of normalized neural networks training with weight decay

Normalized neural networks are those whose architecture uses either normalization layers, such as Batch Normalization [11] or Layer Normalization [12], or directly normalization of weights [13]. The vast majority of modern deep neural networks are normalized, including, for example, the popular ResNet [29] and Transformer [30] architectures. The modern standard for training neural networks has become stochastic methods using the weight

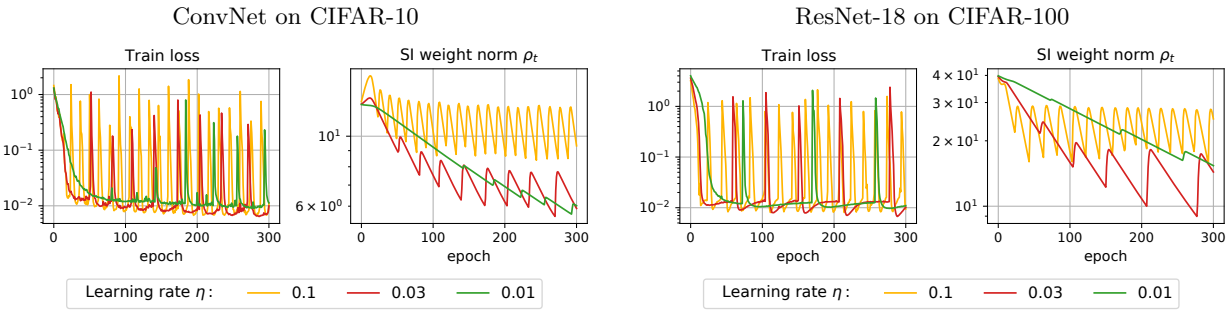

Figure 1: Periodic behavior of ConvNet on CIFAR-10 and ResNet-18 on CIFAR-100 trained using SGD with WD of 0.001 and different learning rates. Each panel on the left shows training loss vs. epoch, on the right — scale-invariant parameters norm vs. epoch.

decay technique, which stabilizes the optimization process and also plays the role of a regularizer, improving generalization of the final solutions [22, 26, 27].

As mentioned above, normalization induces scale invariance of neural network weights that precede normalization layers or that are directly normalized. Due to the ubiquity of the use of normalization techniques in modern architectures, such weights are in the majority, so studying the effect of scale invariance on optimization dynamics turns out to be an urgent problem. This is of particular importance due to the non-trivial and even unexpected interplay between scale invariance and weight decay, as will be shown below.

In this section, we present a study of the periodic behavior of the dynamics of normalized models training with weight decay (Fig. 1), which, among other things, can be considered as a generalization of two conflicting points of view about the outcome of such training: equilibrium versus instability. To simplify the presentation and to exhibit the most representative case, here we consider experiments with convolutional neural networks using BN and trained using the SGD algorithm with a constant learning rate; however, in the main paper, we show that the above results are also valid when using other architectures, normalization techniques, optimization algorithms, including classical gradient descent or the Adam [31] optimizer, and even more general scale-invariant models.

Background and formulation of the problem

To clarify the formulation of the problem, we describe the main consequences of scale invariance on the dynamics of training and its interaction with weight decay. Consider

an arbitrary scale-invariant function $f(x)$ such that

$$f(\alpha x) = f(x),\ \forall x, \forall \alpha > 0. \tag{1}$$

The equation (1) is essentially the definition of scale invariance. By differentiating both parts of the equality (1) with respect to x and with respect to α, one can obtain the following fundamental properties of the gradient of arbitrary scale-invariant functions (see Lemma 1.3 in Li and Arora [18]):

$$\begin{cases} \langle \nabla f(x), x \rangle = 0,\ \forall x & \text{(2a)} \\ \nabla f(\alpha x) = \dfrac{1}{\alpha} \nabla f(x),\ \forall x, \forall \alpha > 0. & \text{(2b)} \end{cases}$$

Consider the gradient descent optimization of $f(x)$ with learning rate η and weight decay λ:

$$x_{t+1} = (1 - \eta\lambda)x_t - \eta \nabla f(x_t). \tag{3}$$

The above properties lead to two important implications regarding the dynamics of the optimization process. First, according to the property (2a), the shift of x in the direction of $-\nabla f(x)$, i.e., gradient descent step, always increases $\|x\|$, while weight decay, on the contrary, decreases $\|x\|$ (see Fig. 2 for an illustration). The interaction of these "centripetal" and "centrifugal" forces can lead to a non-trivial change in $\|x\|$ during optimization. Secondly, according to the property (2b), despite the fact that the value of the function $f(x)$ itself is invariant under the multiplication of x by α, the dynamics changes significantly when optimization is performed on different scales of the parameters norm. For smaller norms, the optimization takes larger steps, which can lead to instability, while for larger norms, the steps are smaller and the optimization process may be slow to converge.

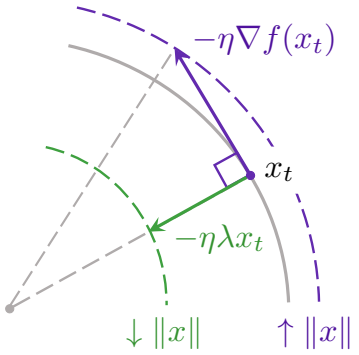

Figure 2: Illustration of "centripetal force" of weight decay and "centrifugal force" of gradient when optimizing scale-invariant functions.

Since the scale-invariant $f(x)$ can be considered as a function on the sphere, its optimization dynamics is often analyzed on the unit sphere, that is, for $\|x\| = 1$. It can be shown that the training dynamics in the entire parameter space can be equivalently represented on the unit sphere using the notions of effective gradient and effective learning rate instead of similar concepts in optimization [18, 19, 20, 21, 22, 23, 24]. The effective

gradient is defined as the gradient at a point projected onto the unit sphere and can be obtained from (2b) as $\nabla f(x/\|x\|) = \nabla f(x)\|x\|$. The effective learning rate is defined as $\tilde{\eta} = \eta/\|x\|^2$ [21, 24]. A change in $\|x\|$ does not affect the effective gradient by definition and is reflected only in the effective learning rate: the lower the norm, the higher the ELR and, in fact, the greater the actual optimization steps. In what follows, the following notation will be used for the norm of parameters, gradient, effective gradient, and the ELR at iteration t, respectively: $\rho_t \equiv \|x_t\|$, $g_t \equiv \|\nabla f(x_t)\|$, $\tilde{g}_t \equiv \|\nabla f(x_t/\|x_t\|)\| = \rho_t g_t$, and $\tilde{\eta}_t \equiv \eta/\rho_t^2$.

Existing contradiction in the literature

Let us briefly outline the essence of the controversy that has formed in the community regarding the training of normalized models with scale-invariant parameters using the weight decay technique (3). As shown earlier, the dynamics of such training non-trivially changes the norm of parameters due to the interaction of "centrifugal force" of the gradient and "centripetal force" of weight decay, which in turn affects the step size of the optimization process. Thus, in the literature, there are two contradictory points of view regarding the result of the corresponding dynamics.

On the one hand, works like Li et al. [19] or Wan et al. [23] claim that such training leads to the *equilibrium* state, where the "centripetal force" is compensated by the "centrifugal force" and, ultimately, the norm of scale-invariant weights (together with other training statistics) will tend to some constant value. Several other works take a similar view [20, 28].

On the other hand, a number of works emphasize that due to the weakening of the role of the gradient with training progress, the use of weight decay can bring the parameters too close to the zero point, which leads to *instability* due to an excessive increase in the effective learning rate. In particular, the work of Li et al. [17] shows that approaching the zero point in normalized neural networks leads to numerical errors after the optimization step and subsequent training failure. Li and Arora [18] also emphasize that scale-invariant functions are ill-conditioned near the origin, and in a simplified way prove that convergence is impossible if both normalization and weight decay are used (however guaranteed if either of them is absent). Moreover, despite their equilibrium presumption, Li et al. [19] empirically demonstrate that the loss function constantly fluctuates between low and high values when gradient descent with weight decay is used.

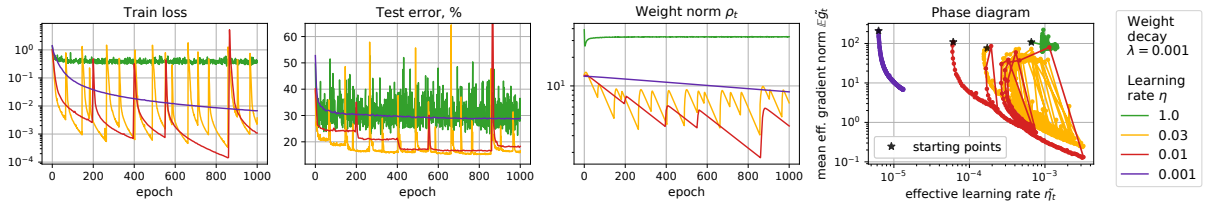

Figure 3: Periodic behavior of training a scale-invariant neural network ConvNet on CIFAR-10.

One of the key results of this thesis is the discovered periodic behavior of training normalized models with weight decay, which, as shown below, allows us to resolve the described fundamental contradiction through the principle of generalized equilibrium.

Periodic behavior and the underlying mechanisms

For the sake of clarity, the experimental results presented here and below cover the case of fully scale-invariant convolutional neural networks (ConvNet, ResNet-18) trained using stochastic gradient descent with weight decay on natural image classification problems CIFAR [32]. To ensure full scale invariance, after each convolutional layer, a BN layer with non-trainable affine parameters is added, and the last linear layer is set fixed. This guarantees scale invariance of all trainable parameters of the model, while practically does not worsen its quality [33, 18]. The main work additionally investigates the case of standard architectures with the presence of non-scale-invariant parameters trained in more conventional settings using momentum, learning rate schedule and data augmentation; in short, all results, including periodic behavior, hold true as long as training is long enough and the LR schedule is not too aggressive (see also Fig. 1).

Figure 3 demonstrates the periodic behavior of training a scale-invariant neural network ConvNet on the CIFAR-10 dataset for different learning rates. In the optimization process, instabilities are clearly encountered, which, however, do not lead to complete divergence, but cause a new training cycle; the observed periodicity of destabilizations, as well as the behavior of the learning dynamics within each period, is regular and obeys some generalized equilibrium law. Thus, one can empirically conclude that the presumptions of equilibrium and instability turn out to be simultaneously valid in one way or another. More rigorous theoretical results supporting this conclusion are given below.

The observed periodic behavior arises from the interplay between normalization and weight decay, namely, due to their competing influence on the norm of the scale-invariant

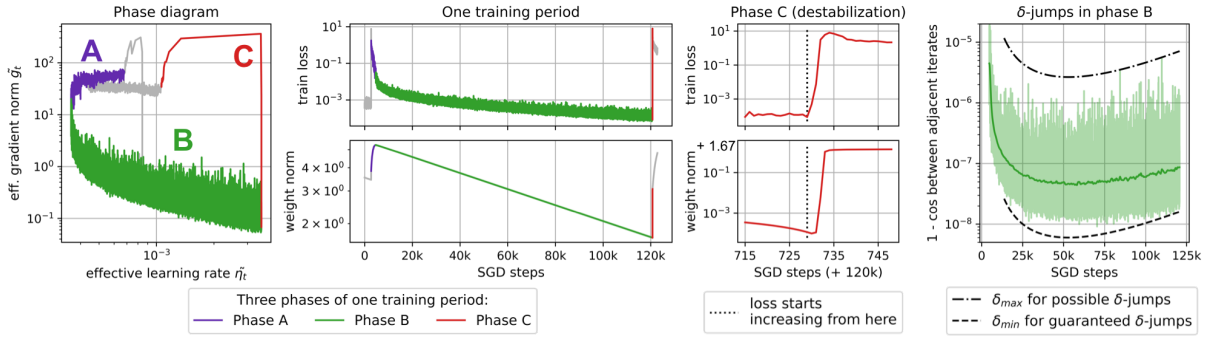

Figure 4: One training period of the scale-invariant ConvNet on CIFAR-10: three phases can be distinguished. The rightmost plot compares the empirically observed cosine distance between weights at adjacent iterations with the theoretically derived bounds.

weights. As discussed earlier, WD tends to decrease the norm of the parameters, while gradients from the loss function tend to increase it (see Fig. 2). These two forces alternately outweigh each other over fairly long training periods, resulting in periodic behavior.

To clarify the details, let us take a closer look at one training period by analyzing Figure 4. At the beginning of the period, high values of the training loss are observed, and therefore, large gradients outweigh the effect of weight decay. This leads to a rapid decrease in training loss, an increase in the parameters norm, and a decrease in the ELR value. The described phase is indicated by the letter A in the plots. Further, as the training loss decreases, there comes a point where the gradients become small enough to be outweighed by WD. As a result, the weight norm begins to decrease, and the effective learning rate increases, which is indicated as the B phase in the plots. Finally, when the weight norm becomes too small and the ELR, on the contrary, too high, the optimization takes several large steps and leaves the vicinity of the minimum point. Effective gradients sharply increase with the value of the loss function and, multiplied by a high ELR, lead to a rapid increase in the weight norm (phase C). The detailed plots for the C phase (the third panel from the left in Figure 4) confirm that the training loss begins to increase earlier than the norm of the parameters. Eventually, when the weight norm becomes large, the effective learning rate decreases and stops the divergence process, leading to a new period of training.

The main paper also provides additional ablation studies demonstrating that periodic behavior can indeed be eliminated by fixing the norm of the parameters (see also the next section of this thesis). This reconfirms the proposed substantiation of the periodic

behavior mechanisms via the dynamics of the weight norm and, consequently, the effective learning rate.

Theoretical analysis

This subsection presents the main theoretical results concerning the optimization dynamics of scale-invariant functions with weight decay (3), which supplement and reinforce the above empirical observations. Mainly, concepts and results are formulated to explain the causes of destabilization between phases B and C of the period, and a theorem is presented that formally generalizes the equilibrium presumption [19, 23]. The exact formulations together with the proofs are given in the main work.

First of all, the dynamics (3) is reformulated in terms of the dynamics of the norm of parameters and gradients:

$$\rho_{t+1}^2 = (1 - \eta\lambda)^2\rho_t^2 + \eta^2\tilde{g}_t^2/\rho_t^2. \tag{4}$$

Next, the notion of a δ-*jump* is introduced, which reflects the necessary condition for the occurrence of destabilization in the analysis of dynamics (3) using (4).

Definition 1. *The dynamics* (3) *performs a* ***δ-jump*** *if the cosine distance between adjacent iterations exceeds a given threshold $\delta > 0$:*

$$1 - \cos(x_t, x_{t+1}) > \delta.$$

The first result on necessary and sufficient conditions for the occurrence of a δ-jump is formulated as follows.

Proposition 1. *Let the effective gradient norm be locally bounded: $\ell \leq \tilde{g}_t \leq L$. Then the following approximate conditions for a δ-jump hold:*

$$\begin{cases} \rho_t^2 \lessapprox \dfrac{\eta L}{\sqrt{2\delta}} \implies \delta\text{-jump is possible}, & \text{(5a)} \\ \rho_t^2 \lessapprox \dfrac{\eta \ell}{\sqrt{2\delta}} \implies \delta\text{-jump is guaranteed}. & \text{(5b)} \end{cases}$$

The fulfillment of these conditions in practice are depicted in the right panel of Figure 4. It can be seen that the actual dynamics of the cosine distance between adjacent iterations remains within the δ-jump bounds derived from Proposition 1. It can also be seen that the general trend of this dynamics demonstrates an increase in δ-jumps towards the end of the B phase, which indicates accumulating instability.

The following proposition helps to predict the time of the occurrence of a δ-jump, and hence destabilization, depending on the choice of the learning rate η and the weight decay factor λ.

Proposition 2. *Denote $\kappa = \sqrt{\frac{\eta}{2\lambda}}$. Under the assumptions of Proposition 1:*

- $\rho_0^2 > \kappa\ell \wedge \delta < \eta\lambda\frac{L^2}{\ell^2} \Rightarrow$ ***minimal*** *δ-jump time is $t_{\min} = \mathcal{O}(1/4\eta\lambda)$;*
- $\rho_0^2 > \kappa L \wedge \delta < \eta\lambda\frac{\ell^2}{L^2} \Rightarrow$ ***maximal*** *δ-jump time is $t_{\max} = \mathcal{O}(1/2\eta\lambda)$.*

Corollary 1. *Thus, we can conclude that instabilities, and hence periods, occur with a frequency directly proportional to the learning rate $\times$ weight decay product $\eta\lambda$.*

Finally, the final and main theoretical result — the *generalized equilibrium principle* — generalizes the previous equilibrium presumption [19, 23] and thus resolves the contradiction with the position of instability: in the course of training, periodic behavior must stabilize within certain limits.

Theorem 1. *Under the assumptions of Proposition 2, if $2\eta\lambda L \leq \ell$, then the following bounds on parameters norm hold:*

$$\kappa\ell \leq \rho_t^2 \leq \kappa L,\ t \gg 1.$$

If $\rho_0^2 > \kappa L$, then ρ_t^2 converges linearly to the interval $[\kappa\ell, \kappa L]$ in $\mathcal{O}\left(1/\eta\lambda\right)$ time.

Empirical confirmation of the above statements is given in detail in the main work, but it can also be seen in Figure 3. One can notice that as the hyperparameter η increases at fixed λ, the periods become more frequent in accordance with Proposition 2 and Corollary 1. One can also notice that, despite the periodic behavior, from a certain training epoch, the norm of parameters (as well as other metrics) clearly lies within certain boundaries, confirming the statement formulated in Theorem 1.

Conclusion and other results

In this part of the thesis, the phenomenon of periodic behavior of the dynamics of training normalized neural networks with weight decay is analyzed in detail. An explanation of the mechanisms behind this periodic behavior is given, as well as theoretical results that reinforce empirical intuition. Finally, through the principle of generalized equilibrium of periodic behavior, the contradiction was resolved regarding such dynamics: equilibrium or instability.

The main work also provides many additional results, omitted here, including empirical analysis of the consequences of periodic behavior, such as the warm-up stage and minima achieved at different training periods, and ablation studies of periodic behavior in different settings of conventional neural network training.

3.2 Three regimes of training scale-invariant neural networks on the sphere

In the previous section, it was mentioned that scale-invariant functions (1) are inherently defined on a sphere, which is their natural domain. As a rule, the unit sphere is implied, for which the notions of effective gradient and effective learning rate are introduced. However, in this work, a broader definition is considered and the natural, or intrinsic, domain of scale-invariant functions will be understood as a sphere of arbitrary fixed radius in the parameter space. When studying neural networks with scale-invariant weights, therefore, the question arises about the structure of the loss landscape on the natural domain in order not to take into account symmetries that do not essentially affect the function implemented by the model.

Usually, studies of loss landscape structure and/or training dynamics of neural networks are conducted in controlled experiments, when all hyperparameters, including the learning rate, are set to fixed or at least controlled values [34, 35, 36, 37, 38]. As was discussed in detail earlier, in the case of training neural networks with scale-invariant parameters in the entire space, the effective learning rate turns out to be a non-trivially changing value, even if all learning hyperparameters are fixed. This complicates and even distorts the understanding of the intrinsic structure of the loss landscape of such models.

In order to correct this shortcoming, in this section we consider the training of completely scale-invariant models on the sphere of fixed radius using projected (stochastic) gradient descent. In this case, the ELR value turns out to be completely controlled at the stage of setting up the experiment; in particular, it can be fixed to a given constant value. It turns out that, depending on the ELR value, such optimization on the sphere can be carried out in three regimes: convergence, chaotic equilibrium, and divergence (see Fig. 5). The first regime (low ELR values) can be considered as a typical case of convergence to a minimum with a monotonically decreasing value of the training loss. The second regime (medium ELR values) demonstrates a consistent oscillatory behavior of the loss function around some value, separated from both the global minimum and random guessing. This

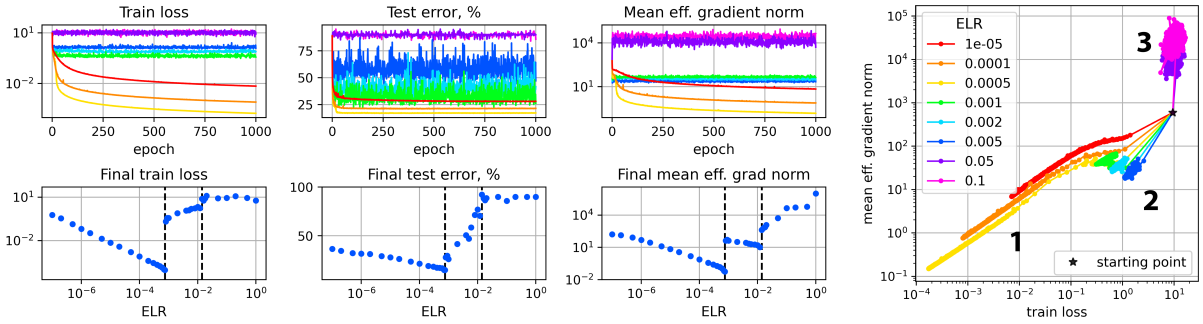

Figure 5: Three regimes of training a scale-invariant neural network on the sphere: (1) convergence for low ELRs, (2) chaotic equilibrium for medium ELRs, and (3) divergence for high ELRs. ConvNet on CIFAR-10. Dashed lines indicate boundaries between regimes.

regime is called *chaotic equilibrium* as it resembles the equilibrium state described in some previous works [19, 23], as mentioned before. The last third regime (high ELR values) is a destabilized, divergent training mode associated with an excessively large optimization step size. Each regime allows revealing certain features of the loss landscape structure on the sphere, which expand and deepen the previous results.

Further in this section, we adopt the specified training protocol using fully scale-invariant neural networks, which are obtained from standard architectures by the method described in the previous section. In the two main papers on the topic of this section, the case of training scale-invariant models in the entire space, as well as conventional training of neural networks, including learning rate schedules, is additionally investigated from the perspective of three regimes.

Theoretical analysis

The emergence of the above regimes of training models with scale-invariant parameters on the sphere by the gradient projection method can be analytically confirmed. To substantiate it, we derived the following theoretical results, clarifying the general properties of such optimization dynamics and analyzing in detail the example of a concrete scale-invariant function. Specific formulations and proofs of the above statements are described in the second of the three mentioned works of the author, related to this thesis.

Consider the function $F(\boldsymbol{\theta})$ of the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^P$, which can be divided into n groups: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$, where each $\boldsymbol{\theta}_i \in \mathbb{R}^{p_i}$ and $\sum_{i=1}^{n} p_i = P$. We will assume that each of these groups is scale-invariant, i.e., multiplication of any of $\boldsymbol{\theta}_i$ by a positive scalar, with the others fixed, does not change the value of the function F. Note that this is a

typical case for neural networks with several normalized layers, since each of these layers individually is scale-invariant. Naturally, if a function is scale-invariant with respect to several groups of parameters, then it is also scale-invariant with respect to their union, so the entire parameter vector $\boldsymbol{\theta}$ is also scale-invariant.

Let us write the algorithm for minimizing $F(\boldsymbol{\theta})$ on the sphere of radius ρ with a fixed learning rate η:

$$\begin{cases} \hat{\boldsymbol{\theta}}^{(t)} \leftarrow \boldsymbol{\theta}^{(t)} - \eta \nabla F(\boldsymbol{\theta}^{(t)}), \\ \boldsymbol{\theta}^{(t+1)} \leftarrow \hat{\boldsymbol{\theta}}^{(t)} \cdot \frac{\rho}{\left\|\hat{\boldsymbol{\theta}}^{(t)}\right\|}. \end{cases} \tag{6}$$

By analogy with the previously considered case of a single scale-invariant group, the notions of an individual effective gradient and an individual effective learning rate for each group $\boldsymbol{\theta}_i$ are introduced as analogs of the regular gradient and learning rate, but calculated at the point $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_i/\rho_i, \ldots, \boldsymbol{\theta}_n)$, where $\rho_i \equiv \|\boldsymbol{\theta}_i\|$. Denote the effective gradient norm for the group $\boldsymbol{\theta}_i$ as $\tilde{g}_i = g_i \rho_i$, where $g_i \equiv \|\nabla_{\boldsymbol{\theta}_i} F(\boldsymbol{\theta})\|$, and the corresponding effective learning rate as $\tilde{\eta}_i = \eta/\rho_i^2$. We also denote the norm of the total effective gradient and the total (fixed) effective learning rate as $\tilde{g} = g\rho$, where $g \equiv \|\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})\|$, and $\tilde{\eta} = \eta/\rho^2$.

From the parameters norm relation $\sum_{i=1}^n \rho_i^2 = \rho^2$, the following fundamental relation is derived, which connects individual effective learning rates with the total one:

$$\sum_{i=1}^{n} \frac{1}{\tilde{\eta}_i} = \frac{1}{\tilde{\eta}}. \tag{7}$$

Further, the notion of effective step size (ESS) is introduced, which is a complete analogue of the regular optimization step size, but calculated on the unit sphere. By definition, the effective step size is equal to the product of the effective learning rate and the effective gradient norm: $\tilde{\eta}_i \tilde{g}_i$ for an individual group $\boldsymbol{\theta}_i$ and $\tilde{\eta}\tilde{g}$ for the full parameter vector $\boldsymbol{\theta}$. This value indicates how much the parameters actually change after the optimization step, taking into account their scale invariance. It turns out that the total ESS can be represented as a convex combination of the individual ones after squaring:

$$(\tilde{\eta}\tilde{g})^2 = \sum_{i=1}^{n} \omega_i (\tilde{\eta}_i \tilde{g}_i)^2, \ \sum_{i=1}^{n} \omega_i = 1, \ \omega_i \propto \frac{1}{\tilde{\eta}_i}. \tag{8}$$

Thanks to the above equations, it is possible to derive the dynamics of individual ELRs update during the process (6):

$$\tilde{\eta}_i^{(t+1)} \leftarrow \tilde{\eta}_i^{(t)} \frac{1 + (\tilde{\eta}\tilde{g}^{(t)})^2}{1 + (\tilde{\eta}_i^{(t)} \tilde{g}_i^{(t)})^2}. \tag{9}$$

This central result of the analysis of the training dynamics on the sphere allows us to conclude that with a high/low value of the individual ESS at a given iteration, the ELR value at the next iteration must decrease/increase. Thus, since the values of the effective step size and the effective learning rate are closely related by definition, the *negative feedback* principle arises, when large ELRs should become smaller, and small ones should become larger. This principle is general and key to the analysis of the training regimes.

To visually explain the differences between the three training regimes, we present the following example of a function with several scale-invariant parameter groups, for which optimization properties were studied depending on the chosen total ELR value:

$$F(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \alpha_i f(x_i, y_i) = \sum_{i=1}^{n} \alpha_i \frac{x_i^2}{x_i^2 + y_i^2}, \tag{10}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\alpha_i > 0$. Thus, each of the n pairs (x_i, y_i) forms a separate scale-invariant group of parameters.

For simplicity, let $\rho = 1$, that is, the optimization (6) is carried out on the unit sphere, and hence $\tilde{\eta} = \eta$. To facilitate the analysis of the optimization dynamics of the general function (10), we derive the following result on the convergence of each of its constituent subfunctions $\alpha_i f(x_i, y_i) = \alpha_i \frac{x_i^2}{x_i^2 + y_i^2}$.

Proposition 3. *For a function $f_\alpha(x, y) = \alpha \frac{x^2}{x^2+y^2}$ optimized on a sphere with an effective learning rate $\tilde{\eta}$:*

1. *if $\tilde{\eta} < \frac{1}{\alpha}$, there is a linear convergence to a minimum;*
2. *if $\tilde{\eta} > \frac{1}{\alpha}$, there is a stabilization at level $\frac{1}{2}\left(\alpha - \frac{1}{\tilde{\eta}}\right)$.*

Now, knowing the conditions for the convergence of the F subfunctions, namely, when the individual effective learning rate is less than $1/\alpha_i$, by taking into account the relation (7), we can conclude that the first regime, i.e., convergence, is observed under the following condition:

$$\tilde{\eta} < \frac{1}{\sum_{i=1}^{n} \alpha_i}. \tag{11}$$

Otherwise, the situation in which all individual ELRs are below the convergence threshold is impossible, and therefore the dynamics of their update (9) turns out to be undamped and must converge to some equilibrium according to the negative feedback principle. Naturally, the state of equilibrium of such dynamics is the situation in which all individual ESS values become equal, which is equivalent to reaching the following levels for individual

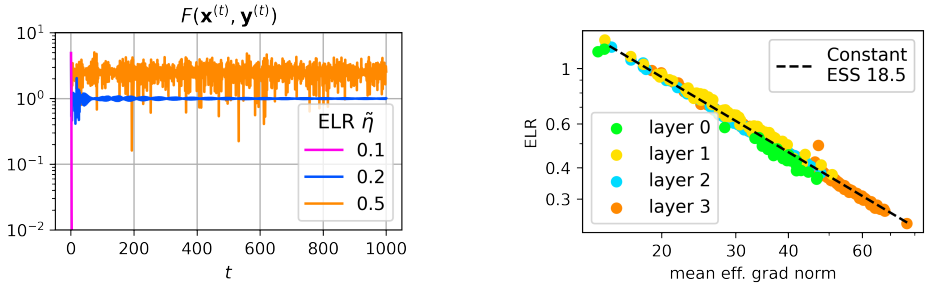

Figure 6: Verification of theoretical results. Left: optimization dynamics of the function (10) using different ELRs corresponding to the three regimes. Right: equilibration of individual ESS values in the second regime when training a scale-invariant neural network on the sphere.

ELR values:

$$\tilde{\eta}_i^* \equiv \frac{\tilde{\eta} \sum_{j=1}^{n} \alpha_j}{\alpha_i}. \tag{12}$$

Thus, the system has to enter a state of chaotic equilibrium, that is, the second training regime. Finally, if the value of $\tilde{\eta}$ is too high, chaos will dominate over the equilibrium and the optimization will switch to the third regime of divergence.

Figure 6 depicts some results of experimental verification of the obtained theoretical derivations. On the left is the behavior of the function (10) when optimizing on the unit sphere with three different effective learning rates corresponding to the three regimes according to the threshold rule (11). It can be seen that the smallest value makes the function converge quickly, the medium one causes the function to stabilize at a certain level, and the largest one leads to the most chaotic behavior. On the right is the predicted equilibration of effective step sizes in the second regime for individual scale-invariant groups of a real ConvNet neural network trained on the sphere. It can be seen that the effective step sizes, which are the product of the effective learning rates and the effective gradients norm, do actually concentrate around a certain value indicated by the dotted black line on the plot.

Empirical results

This subsection contains the main experimental results regarding the loss landscape structure of scale-invariant neural networks on the sphere, obtained by analyzing three training regimes. The details of the experiments, as well as additional results, are given in the second and third works of the attached list of articles by the author of this thesis. Here, to simplify the presentation, we will consider a scale-invariant convolutional neural network

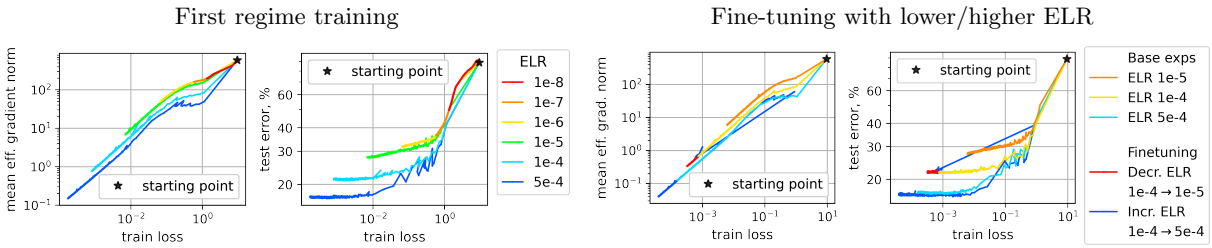

Figure 7: First training regime. Left: training with fixed ELR converges to areas with different sharpness and generalization. Right: fine-tuning with a smaller ELR remains on the same trajectory, but with an increased one jumps out and converges to a more optimal basin.

ConvNet, trained on the task of classifying natural images CIFAR-10 using the specified projected SGD method of the form (6). The main displayed metrics are: training loss, test error, and average norm of the effective stochastic gradient of the model, which simultaneously plays the role of a measure of the optimization progress and the sharpness of the found minimum, as justified in the main text of the papers. All training launches were carried out from the same initialization point and with the same data order to ensure complete fixation of the experiment setting and leveling the effect of training stochasticity when comparing the results. First, we consider the case of the classical cross-entropy loss function.

Training with small values of the effective learning rate leads to the first regime, called *convergence*. The optimization shows typical convergence behavior (see Fig. 5): after a few epochs, the model is able to reach an area with very low training loss and continues to converge to the minimum point. The rate of convergence directly depends on the value of the ELR. In addition, training with different ELR values leads to solutions with different sharpness (mean effective gradients norm) and generalization (test error): higher values lead to a solution with lower sharpness and better generalization. Additional results from the main work of the author also confirm that the optima achieved after training with different ELR values not only differ in the described characteristics, but also are geometrically located in distinct linearly disconnected basins of the loss landscape.

The chosen ELR value affects not only the rate of convergence and the properties of the final solution: the entire optimization trajectory differs significantly for runs with different values. To analyze training trajectories regardless of optimization speed, consider the evolution of sharpness and generalization versus training loss. The corresponding plots are shown for various ELR values in Figure 7, left. For the lowest values, the

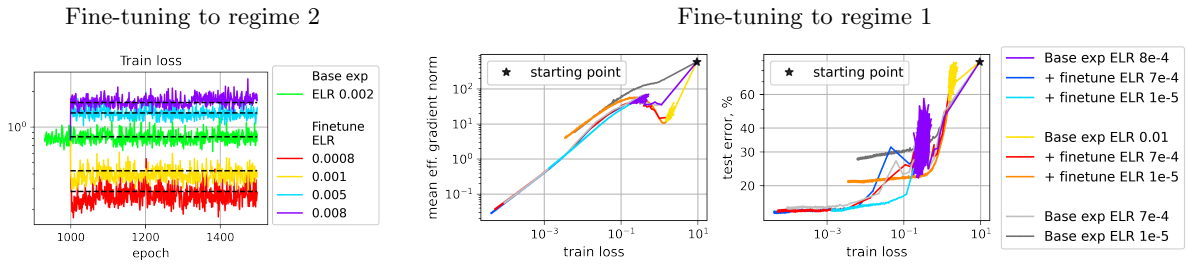

Figure 8: Fine-tuning from the second regime with a different ELR value. Left: to the second regime, leads to stabilization at a new level. Right: to the first regime, leads to convergence to different optima when starting from a high ELR and to the same type of minimum when starting from a low ELR.

trajectories coincide, while training is too slow to converge to the optimum. For other values, as the effective learning rate increases, the trajectories begin to shift down and to the left on the diagrams, while in the region of low loss values, they appear parallel. Such differences demonstrate that training at a higher ELR not only leads to a better endpoint in terms of sharpness and generalizing, but also better conditions the entire optimization trajectory. We also conducted an experiment with fine-tuning from the end of the trajectory attributed to a given ELR value with a different ELR (see Fig. 7, right). Fine-tuning with a lower value remains in the same basin and continues to move along the same trajectory without finding other minima. Fine-tuning with a higher value after a short time goes beyond the neighborhood of the current optimum and converges into a new basin, the characteristics of which correspond to the new effective learning rate.

Thus, the analysis of the first regime allows us to conclude that the intrinsic structure of the loss landscape of scale-invariant neural networks contains a variety of optima that differ in the characteristics of sharpness and generalization, which form a one-to-one correspondence with the chosen value of the effective learning rate.

In the second regime, the optimization noisily stabilizes around a certain value for various training metrics, for example, for the training loss (see Fig. 5), which is called *chaotic equilibrium*. In this regime, the value of the effective learning rate is too high to converge to the optimum, but still not enough to diverge completely. Thus, the optimization process stabilizes in a certain region of the loss landscape with a practically fixed level of training metrics, which is determined by the chosen ELR value. Additional experiments from the main works show that this region is locally convex for moderate values of the effective learning rate in the second regime.

To demonstrate that the value of the effective learning rate uniquely determines the loss stabilization level, we conducted an experiment with fine-tuning with a different ELR value of the second regime (see Fig. 8, left). Changing the value of the effective learning rate appropriately changes the optimization dynamics, namely, brings it to a level corresponding to the new value. We also conducted experiments with fine-tuning from different starting ELR values of the second regime with a new value of the first regime. As can be seen from the right plots of Figure 8, the resulting optima are highly dependent on the starting value. Fine-tuning of models pretrained with low ELR values of the second regime always converges to points with the same sharpness/generalization characteristics. For large starting values, fine-tuning leads to a variety of trajectories depending on the chosen effective learning rate from the first regime.

Thus, after analyzing the second training regime, we can conclude that in the loss landscape on the sphere, stabilization regions can be distinguished, in which the optimization dynamics is fixing at a certain level depending on the chosen value of the effective learning rate. Such regions can be either local (convex, allow convergence into a single type of minimum) or global (non-convex, contain many different minima).

In addition to the above, it can be assumed that the main difference between the first and second optimization regimes is associated with the presence of zones of increased sharpness in the optimization trajectories of neural networks. In the effective gradient norm versus training loss diagram in Figure 5, right, one can observe that sharpness reaches its peak exactly at the transition point between the first two regimes. This allows us to conclude that training with only sufficiently small ELR values makes it possible to pass this bottleneck and enter the convergence regime. In the main papers, this reasoning is additionally substantiated with the help of a detailed analysis of transitions between regimes and the relationship between these transitions and the epoch-wise double descent [4].

For the highest values of the effective learning rate, the most unstable optimization behavior is observed, corresponding to random guessing (see Fig. 5). This is how the third training regime, called *divergence*, manifests itself. In the main works of the author, this regime is directly compared with the random walk and the gradient ascent method: in all three cases, adjacent iterations turn out to be uncorrelated, but the random walk gives a lower bound for the training loss value, and gradient ascent gives an upper bound. Thus,

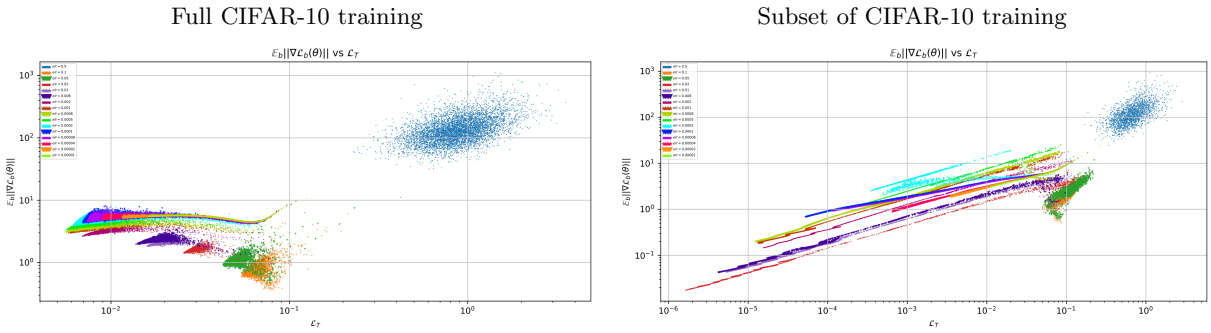

Figure 9: Training loss versus mean effective gradient norm diagrams. Training with an MSE loss function on the sphere. Left: training on a full dataset does not allow optimization to converge and detect the first training regime. Right: training on a subsample of 5000 objects allows all three training regimes to be distinguished.

the behavior of the model in the third regime can be considered as something between random guessing and divergence.

MSE loss function case

In conclusion of this section, we briefly touch the main results obtained when training on the sphere using an MSE loss function.

As can be seen from Figure 9, training the scale-invariant ConvNet model on the sphere using MSE loss did not allow optimization to converge to the optimum for the given iteration budget. Thus, only the second and third optimization regimes can be clearly identified. However, after reducing the sample size to 5000 objects, the model was able to converge for certain values of the effective learning rate and demonstrate the presence of the first regime as well, in full accordance with the case of the cross-entropy loss function. In the second work of the author on the topic of this section, additional experiments are presented based on the analysis of cosine distances between adjacent iterations, which allow one to more subtly distinguish between the regimes and show the presence of an analogue of the first regime even without convergence.

Conclusion and other results

In this section, we analyzed the case of training fully scale-invariant models on the sphere using the projected SGD method with a fixed effective learning rate value. We identified three regimes of such training depending on the specific ELR value: convergence,

chaotic equilibrium, and divergence. We analyzed these regimes both theoretically and experimentally, identified their main characteristics, and obtained information on the intrinsic structure of the loss landscape of scale-invariant neural networks. Among them is the presence of a whole spectrum of various global optima, high-sharpness zones, and regions of stabilization of optimization dynamics, both local and global. In addition to the classical cross-entropy function, the case of using an MSE loss function on classification problems was also studied from the point of view of three training regimes.

The author's two main papers on this topic (second and third in order in the list) provide many additional results omitted in this section, including the manifestation of three regimes in conventional training of neural networks, the interpretation and search for optimal learning rate schedules with their help, linear connectivity in different regimes, analysis of transitions between regimes, ELR schedules, etc.

4 Conclusion

The final section summarizes the main results of the work.

1. We investigated the dynamics of standard training of neural networks using normalization and weight decay techniques. We discovered, explained and analyzed both experimentally and theoretically periodic behavior of such training dynamics.

2. We proposed to resolve the contradiction that has developed in the literature about the result of such training dynamics (equilibrium versus instability) via the principle of generalized equilibrium in periodic behavior. This argument is substantiated both experimentally and theoretically.

3. We investigated the dynamics of training fully scale-invariant neural networks on the sphere of fixed radius using both cross-entropy and MSE loss functions. We discovered and analyzed both experimentally and theoretically three regimes of such training: convergence, chaotic equilibrium, and divergence.

4. Thanks to the study of these regimes, we revealed in more detail the loss landscape structure of scale-invariant models on the natural domain. In particular, we showed the presence of a whole spectrum of global minima, different in their properties, high-sharpness zones, local and global regions of optimization dynamics stabilization.

References

[1] Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations*, 2021.

[2] Leo Breiman. Reflections after refereeing papers for nips. In *The Mathematics of Generalization*, pages 11–15. CRC Press, 2018.

[3] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Science*, 116(32), 2019.

[4] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020.

[5] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

[6] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[7] W Ronny Huang, Zeyad Ali Sami Emam, Micah Goldblum, Liam H Fowl, Justin K Terry, Furong Huang, and Tom Goldstein. Understanding generalization through visualizations. In *"I Can't Believe It's Not Better!" NeurIPS 2020 workshop*, 2020.

[8] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning*. PMLR, 2018.

[9] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, 2018.

[10] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[12] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[13] Tim Salimans and Diederik P Kingma. Weight normalization: a simple reparameterization to accelerate training of deep neural networks. *Advances in Neural Information Processing Systems*, 29, 2016.

[14] Nils Bjorck, Carla P Gomes, Bart Selman, and Kilian Q Weinberger. Understanding batch normalization. *Advances in Neural Information Processing Systems*, 31, 2018.

[15] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in Neural Information Processing Systems*, 31, 2018.

[16] Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. A mean field theory of batch normalization. In *International Conference on Learning Representations*, 2019.

[17] Xiang Li, Shuo Chen, and Jian Yang. Understanding the disharmony between weight normalization family and weight decay. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4715–4722, 2020.

[18] Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2020.

[19] Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33, 2020.

[20] Twan Van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

[21] Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: efficient and accurate normalization schemes in deep networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[22] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2019.

[23] Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of normalized neural network using sgd and weight decay. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[24] Simon Roburin, Yann de Mont-Marin, Andrei Bursuc, Renaud Marlet, Patrick Pérez, and Mathieu Aubry. A spherical analysis of adam with batch normalization. *arXiv preprint arXiv:2006.13382*, 2020.

[25] Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[27] Aitor Lewkowycz and Guy Gur-Ari. On the training dynamics of deep networks with l_2 regularization. *Advances in Neural Information Processing Systems*, 33, 2020.

[28] Vitaliy Chiley, Ilya Sharapov, Atli Kosson, Urs Koster, Ryan Reece, Sofia Samaniego de la Fuente, Vishal Subbiah, and Michael James. Online normalization for training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Conference on Computer Vision and Pattern Recognition*, 2015.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[33] Elad Hoffer, Itay Hubara, and Daniel Soudry. Fix your classifier: the marginal value of training the last weight layer. In *International Conference on Learning Representations*, 2018.

[34] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[35] Nikhil Iyer, V Thejas, Nipun Kwatra, Ramachandran Ramjee, and Muthian Sivathanu. Wide-minima density hypothesis and the explore-exploit learning rate schedule. *arXiv preprint arXiv:2003.03977*, 2020.

[36] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

[37] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

[38] Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Edward Dahl, Zachary Nado, and Orhan Firat. A loss curvature perspective on training instabilities of deep learning models. In *International Conference on Learning Representations*, 2022.