Why Do We Need Weight Decay for Overparameterized Deep Networks?

Francesco D'Angelo, Aditya Varre, Maksym Andriushchenko, Nicolas Flammarion
Theory of Machine Learning Lab
EPFL, Lausanne, Switzerland
`{francesco.dangelo,aditya.varre, maksym.andriushchenko, nicolas.flammarion}@epfl.ch`

Abstract

Weight decay is a broadly used technique for training state-of-the-art deep networks. Despite its widespread usage, its role remains poorly understood. In this work, we highlight that the role of weight decay in modern deep learning is different from its regularization effect studied in classical learning theory. For overparameterized deep networks, we show how weight decay modifies the optimization dynamics enhancing the ever-present implicit regularization of SGD via *loss stabilization*.

1 Introduction

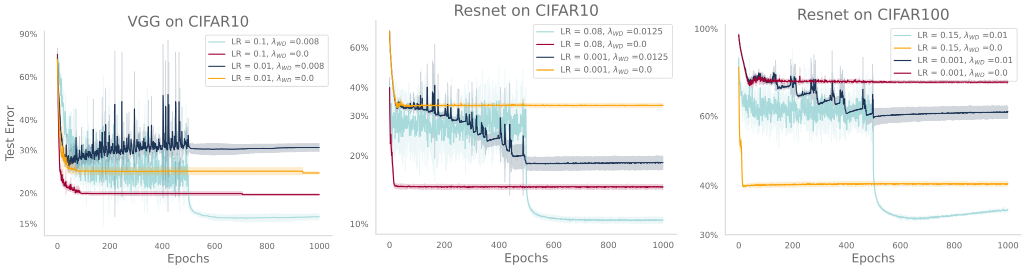
Weight decay (WD) is a widely studied topic in machine learning and numerous hypotheses have been formulated trying to explain its effect. For instance, it serves as a constraint for the network capacity [4], it appears in generalization bounds [17] and it influences the training dynamics via an *effective* learning rate [19]. We argue that despite its widespread usage, its impact on generalization remains poorly understood: in some cases, it acts as a regularizer and in other cases as a tool for better optimization. For example, the difference in test errors in Fig. 1 shows that minimizing the regularized objective alone does not ensure good generalization. Therefore, the regularized objective is insufficient to explain the benefits of WD and a high learning rate appears necessary for optimal performance. This experiment reaffirms the widely acknowledged consensus that *implicit regularization induced by the LR is crucial* [7, 9, 1].

It is natural then to wonder whether weight decay's improvement primarily stems from its ability to control the norm. Fig. 3a clearly illustrates that distinct training trajectories, while resulting in the same final ℓ_2 norm of parameters, can still yield different generalization. Therefore, *the ℓ_2-norm of the learned model's parameters is inconsequential*. This suggests that once the norm is constrained by weight decay, the critical factor influencing the model's generalization is the subsequent choice of LR. Understanding how WD induces these optimization dynamics is crucial for grasping its benefits in generalization. We start by examining the evolution of the norm in Fig. 3a. It rapidly decreases to stabilize within a narrow interval. After the rapid decrease, the optimization resembles the dynamics of SGD projected onto a sphere with a certain radius. We assert that this stage is pivotal for training with weight decay and hypothesize the following key mechanism:

Weight decay maintains parameters norm in a small bounded interval. The resulting projected noise-driven process induces an implicit regularization effect.

The objective of this work is to gain a comprehensive understanding of how the interaction between weight decay and learning rate influences the training dynamics. In particular, we aim to explain the difference in generalization between the yellow and turquoise curves in Fig. 1 by conjecturing and empirically verifying the existence of an implicit regularization mechanism.

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.



(a) (b) (c)

Figure 1: **Training with and w/o weight decay.** We report the test error for VGG (1a) and ResNet (1b, 1c) trained on CIFAR10/100 with and without weight decay and with small and large learning rates. After the first 500 epochs the learning rate is decayed to $\eta = 10^{-4}$ for all the curves.

2 Weight decay for overparameterized deep networks

We start the analysis in a simplified setup that provides foundational insights which help understand the role of weight decay in more general training scenarios.

Notations. Let $(x_i, y_i)_{i=1}^n$ be the training inputs and labels where $x_i \in \mathcal{D},\ y_i \in \mathbb{R}^c$, and c is number of classes. Let $h : \mathbb{R}^p \times \mathcal{D} \to \mathbb{R}^c$ be the hypothesis class of neural network and for any parameter $\mathbf{w} \in \mathbb{R}^p$ where the function $h(\mathbf{w}, \cdot) : \mathcal{D} \to \mathbb{R}^c$ represents the network predictions. We assume for this section that the network is overparameterized and capable of achieving perfect training accuracy. The training loss $\mathcal{L}$ and the ℓ_2-regularized training loss $\mathcal{L}_\lambda$ are given respectively as: $\mathcal{L}(\mathbf{w}) := \frac{1}{N}\sum_{i=1}^N \ell\left(y_i, h(\mathbf{w}, x_i)\right)$ and $\mathcal{L}_\lambda(\mathbf{w}) := \mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\left\|\mathbf{w}\right\|^2$, where $\ell(\cdot,\cdot) : \mathbb{R}^c \times \mathbb{R}^c \to \mathbb{R}$ denotes the cross-entropy loss function. With $i_t \sim \mathbb{U}([N])$, the SGD algorithm on $\mathcal{L}_\lambda(\mathbf{w})$ (here with batch size 1 and with replacement) with a learning rate (LR) η is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla_{\mathbf{w}}\ell\left(y_{i_t}, h(\mathbf{w}_t, x_{i_t})\right) - \eta\lambda\mathbf{w}_t. \tag{1}$$

Experimental setup. We train VGG [18] without BatchNorm and ResNet [6] models on CIFAR-10/CIFAR-100 using SGD and step-decay [6] as LR schedule. Moreover, we compare different values of ℓ_2-regularization coefficient λ. By decaying the LR we divide the training into two separate phases: (1) **large-LR phase** which exploits the SGD noise, and (2) **fine-tuning phase** which uses a small LR.

2.1 Warmup: optimization on the sphere with scale invariance

To isolate the implicit regularization effect from the large initial drop of the norm, we consider a simplified setting. We train scale-invariant networks [10] with projected SGD on the sphere $\mathbb{S}^{(p-1)}$:

$$\mathbf{w}_{t+1} = \Pi_{\mathbb{S}^{(p-1)}}\left(\mathbf{w}_t - \eta\nabla_{\mathbf{w}}\ell\left(y_{i_t}, h(\mathbf{w}_t, x_{i_t})\right)\right) \text{ where } \quad \Pi_{\mathbb{S}^{(p-1)}} : \mathbf{w} \mapsto {}^{\mathbf{w}}/\left\|\mathbf{w}\right\|_2. \tag{2}$$

The training framework still consists of two phases separated by a LR decay. The primary insight from the sphere analysis is depicted in Fig. 2: the test performance achieved in the fine-tuning phase depends on the LR used in the large-LR phase and moreover, there is an optimal value. Our investigation reveals that the key to understand this behavior and the dependence on the LR lies in the noisy dynamics in the first phase.

The noise driven process. We introduce the key ingredients of SGD noise and subsequently exploit the properties of their approximations to investigate the implicit regularization. Let $g_t = \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}_t) - \nabla_{\mathbf{w}}\ell\left(y_{i_t}, h(\mathbf{w}_t, x_{i_t})\right)$ denote the noise in the gradient.

- **(P1)** Under reasonable approximations (details in Prop. 3) the scale of the noise is proportional to the train cross-entropy loss, i.e., $\mathbb{E}\left[\|g_t\|^2\right] \sim \mathcal{L}(\mathbf{w}_t)$. Hence, a higher training loss implies a larger noise in the stochastic gradients. The experiments in Fig. 5,6 show that in the large LR phase, the training loss remains nearly constant. Based on this, we assume $\mathbb{E}\left[\|g_t\|^2\right] \asymp \sigma_\eta^2$.
- **(P2)** We empirically observe that the covariance of the noise $\Sigma_t = \mathbb{E}\left[g_t g_t^\top\right]$ and the Hessian $\nabla^2_{\mathbf{w}}\mathcal{L}(\mathbf{w}_t)$ have the same shape, see App B.4.

For regression, the shape of the covariance of the stochastic gradients, when the labels are injected with Gaussian noise, also matches the shape of the Hessian. This crucial observation is used in several works [2, 12, 3] to demonstrate the implicit regularization properties of SGD. Specifically, Damian

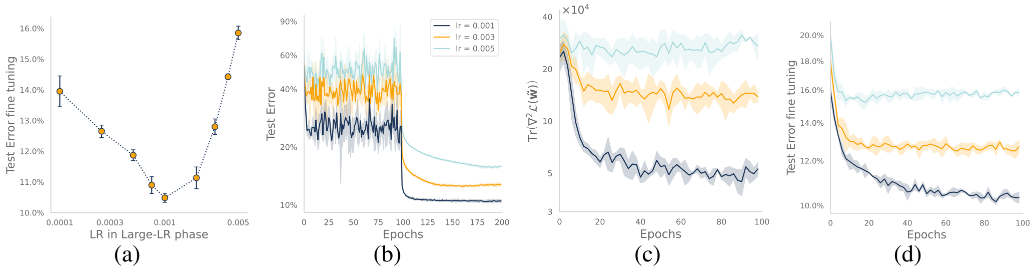


Figure 2: **Training scale-invariant ResNets on the sphere.** We train on CIFAR10 with three different large LR for the first 100 epochs and decay it to $\eta = 10^{-4}$ afterwards. Figure (2a) reports the test error with respect to different LRs in the first phase showing the existence of an optimal value. Figure (2d) reports the test error along the iterations. Figures (2c, 2d) report the decreasing trend of the trace of the Hessian and test error after fine-tuning for 100 epochs with $\eta = 10^{-4}$ every 2 epochs.

et al. [3], Pillaud-Vivien et al. [15] show that the SGD trajectory closely tracks the solution of a regularized problem. Leveraging (**P2**), we conjecture that a similar result should hold in our analysis and that the dynamics of SGD on the sphere for classification tracks closely a regularized process.

Conjecture 1. *Consider the algorithm Eq. 2 with $\mathbf{w}_0$ initialized from a distribution $\mu_0\left(\mathbb{S}^{(p-1)}\right)$. For any input x, let $\mathbf{w}_t, h(\mathbf{w}_t, x)$ be the random variables that denote the iterate at time t and its functional value. The stochastic process $\left(h(\mathbf{w}_t, x)\right)_{t\in\mathbb{N}}$ will converge to a stationary distribution $\mu_\eta^\infty(x)$ with mean $\bar{\mu}_\eta(x)$ for which the following property holds,*

$$\bar{\mu}_\eta(x) = h\left(\mathbf{w}_\eta^*, x\right), \text{ where } \mathbf{w}_\eta^* := \underset{\mathbf{w}\in\mathbb{S}^{(p-1)}}{\arg\min}\ \mathcal{L}(\mathbf{w}) + \eta\sigma_\eta^2 \operatorname{Tr}\left(\nabla^2\mathcal{L}(\mathbf{w})\right). \tag{3}$$

The important difference is that, unlike [2, 3], we do not need to add noise to the labels at each iteration, instead, the large-LR phase induces a label noise-like behaviour similar to [1].

What is the purpose of the fine-tuning phase? Even at stationarity[1], the values of the loss $\mathcal{L}(\mathbf{w}_t)$ and of $\operatorname{Tr}\left(\nabla^2\mathcal{L}(\mathbf{w}_t)\right)$ are still dominated by the noise. This noise obscures any discernible trend along the trajectory, making it challenging to argue convincingly about convergence to the minimum of the regularized loss. While Langevin dynamics suggest LR annealing to approach the mean of the stationary distribution, this technique does not fully resolve the issue. The noise is state-dependent and decreasing the LR might change the stationary distribution and potentially the regularized objective. An alternative approach is to project the iterate $\mathbf{w}_t$ onto a manifold where the loss matches the value evaluated at the mean. Analyzing the evolution of $\operatorname{Tr}\left(\nabla^2\mathcal{L}\right)$ at these projected iterates might reveal evidence of a regularized process. For an illustration, refer to Fig. 4. This projection corresponds to the fine-tuning phase and is accomplished with early-stopped gradient flow (SGD with a small LR).

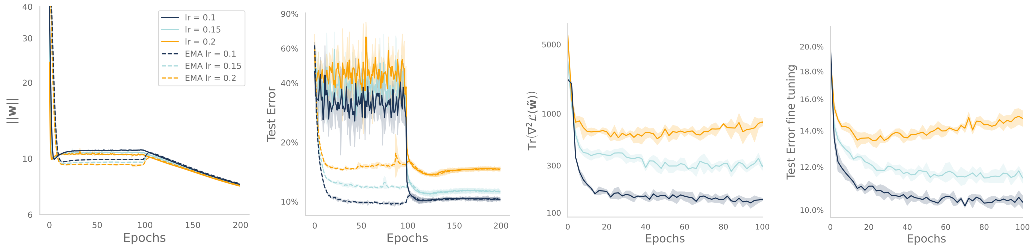
Interpretation of the conjecture and links to generalization. The empirical observations in Fig. 2 show that when two different LRs η_l (large) and η_s (small) are used in the Large-LR phase, models with different generalization properties are obtained after the fine-tuning phase. Our conjecture explains this gap as two solutions $\bar{\mu}_{\eta_l}$ and $\bar{\mu}_{\eta_s}$ of the regularized problem having different strength of regularization ($\eta_l\sigma_{\eta_l}^2$ vs $\eta_s\sigma_{\eta_s}^2$). The conjecture further explains the U-shape generalization curve in Fig. 2a where optimal regularization results in good test performance, and models beyond that level are over-regularized. The regularization is implicit and is solely due to the noisy dynamics.

Revealing the implicit regularization mechanism. Directly measuring the loss $\mathcal{L}$ or $\operatorname{Tr}\left(\nabla^2 L\right)$ at $\mathbf{w}_t$ fails to reveal any decreasing trend due to the noise. Thus, we use the fine-tuning process to exhibit such trend. During fine-tuning, the iterate $\mathbf{w}_t$ is projected to a nearby point, denoted as $\tilde{\mathbf{w}}_t$, such that $\mathcal{L}(\tilde{\mathbf{w}}_t) \sim \mathcal{L}(\mathbf{w}_\eta^*)$. Since their loss values are similar, we compare $\operatorname{Tr}\left(\nabla^2\mathcal{L}(.)\right)$ at $\mathbf{w}_\eta^*$ and $\tilde{\mathbf{w}}_t$. In the experiments in Fig. 2c, we report $\operatorname{Tr}\left(\nabla^2\mathcal{L}(.)\right)$ along the fine-tuned iterates $\tilde{\mathbf{w}}_t$ and observe a decreasing trend. The trajectory of the iterates $(\mathbf{w}_t)_{t\geq 0}$ closely follows the trajectory of the fine-tuned iterates $(\tilde{\mathbf{w}}_t)_{t\geq 0}$ which converge to $\mathbf{w}_\eta^*$. This mechanism elucidates how the trajectory of SGD implicitly biases the model towards a regularized solution that leads to enhanced generalization.

2.2 A unifying theme: beyond scale invariance and spherical optimization

The spherical case studied in the previous subsection paints a clear picture. When isolated from the evolution of the norm, the stochastic dynamics induced by SGD and large LRs provide better control

[1] Assuming the existence of a stationary distribution, the iterates $\mathbf{w}_t$ are eventually realizations from it.



(a) (b) (c) (d)

Figure 3: **Training standard ResNets with weight decay.** We train on CIFAR-10 with $\lambda_{WD} = 0.015$, three different large LRs for the first 100 epochs and decay them to $\eta = 10^{-3}$ afterwards. The norm in Fig. 3a converges to the same value after the LR decay while the test error in Fig. 3b is different. Fig. (3c, 3d) report the decreasing trend of $\mathrm{Tr}(\nabla^2)$ and test error after fine-tuning for 100 epochs with $\eta = 10^{-3}$ every 2 epochs.

over the trace of the Hessian of the model and thus enforce a useful regularization which translates into good generalization properties. In this section, we demonstrate that a similar picture holds in the case of standard training with weight decay. We extend the Conjecture 1, to hold beyond spherical optimization and for networks which are not scale invariant.

Conjecture 2. *Consider the algorithm in Eq. 1 with $\mathbf{w}_0$ initialized from a distribution $\mu_0\left(\mathbb{R}^{(p)}\right)$. For any input x, let $\mathbf{w}_t, h(\mathbf{w}_t, x)$ be the random variables that denote the iterate at time t and its functional value. The stochastic process $\left(h(\mathbf{w}_t, x)\right)_{t\in\mathbb{N}}$ converges to the stationary distribution $\mu^{\infty}_{\eta,\lambda}(x)$ with mean $\bar{\mu}_{\eta,\lambda}(x)$ for which the following property holds,*

$$\bar{\mu}_{\eta,\lambda}(x) = h\left(\mathbf{w}^*_{\eta,\lambda}, x\right), \text{ where } \quad \mathbf{w}^*_{\eta,\lambda} := \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min}\ \mathcal{L}_\lambda(\mathbf{w}) + \eta\sigma^2_{\eta,\lambda}\,\mathrm{Tr}\left(\nabla^2\mathcal{L}\right). \tag{4}$$

There are two differences compared to Conjecture 1: (a) the loss term in the regularized objective is replaced by a ℓ_2-regularized loss and (b) most importantly the strength of the regularization $\sigma_{\eta,\lambda}$, now depends on both the LR and the WD parameter λ. Our experiments in Fig. 3, provide empirical validation for this conjecture. When trained with different LRs and then fine-tuned, the training converges to models with different test performances. This difference is primarily attributed to the varying regularization strengths $\sigma_{\eta,\lambda}$. When fine-tuning every two epochs along the trajectory as reported in Fig. 3c, the quantity $\mathrm{Tr}(\nabla^2)$ is decreasing closely following a regularized process. A similar trend can be observed in Fig. 3d for the test performance when fine-tuning along the trajectory. These observations strongly indicate the benefits of generalization arising from implicit regularization.

Exponential moving average. As discussed in the spherical case, the iterates are noisy realizations and measuring either $\mathcal{L}$ or $\mathrm{Tr}(\nabla^2)$ at the iterates is not informative. However, we can reduce the noise by averaging.[2] Intuitively the average should be close to $\mathbf{w}^*_{\eta,\lambda}$ and the experiment in Fig. 3b confirms this intuition. We consider an exponential moving average (EMA) of the SGD iterates with parameter $\beta = 0.999$ and show that the test error is lower for large LR which enjoy better regularization. This provides further justification for our conjecture and also highlights the practical advantage of obtaining the best model by a simple exponential moving average instead of fine tuning.

Effective learning rate vs. high training loss. Existing works [21] have explored the relationship between LR and WD, introducing the concept of effective LR. These works primarily emphasize that training with WD results in a higher effective LR, without clarifying how this high LR contributes to improved generalization. We address this gap by proposing that a higher LR leads to an increase in $\sigma_{\eta,\lambda}$, consequently enhancing generalization. We claim that examining the high training loss, which approximates the scale of $\sigma_{\eta,\lambda}$, offers a more insightful explanation for the enhanced generalization ability. Analyzing the training curve in Figure 5, the training loss remains consistently high during training with weight decay, entering a phase termed "loss stabilization," [1]. We assert that WD contributes to achieving this *loss stabilization* phase in classification tasks, leveraging the implicit regularization induced by stochastic dynamics.

Conclusion. We illustrated how WD, constraining the norm of the parameters in a small interval, keeps the noise of SGD at a certain scale given a sufficiently large LR. Moreover, this induces an implicit regularization effect whose strength depends on both the LR η and the WD parameter λ. Crucially, different strengths of this regularization lead to different generalization of the model.

[2] Note that on the sphere, we need to compute the mean on a manifold which is a harder problem

References

[1] M. Andriushchenko, A. Varre, L. Pillaud-Vivien, and N. Flammarion. SGD with large step sizes learns sparse features. In *International Conference on Machine Learning*, 2023.

[2] G. Blanc, N. Gupta, G. Valiant, and P. Valiant. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In *Conference on Learning Theory*, 2020.

[3] A. Damian, T. Ma, and J. D. Lee. Label noise sgd provably prefers flat global minimizers. In *Advances in Neural Information Processing Systems*, volume 34, pages 27449–27461, 2021.

[4] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *ICCV*, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[7] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2016.

[8] M. Kodryan, E. Lobacheva, M. Nakhodnov, and D. P. Vetrov. Training scale-invariant neural networks on the sphere can happen in three regimes. *Advances in Neural Information Processing Systems*, 35:14058–14070, 2022.

[9] Y. Li, C. Wei, and T. Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, 2019.

[10] Z. Li and S. Arora. An exponential learning rate schedule for deep learning. *arXiv preprint arXiv:1910.07454*, 2019.

[11] Z. Li, K. Lyu, and S. Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. In *Advances in Neural Information Processing Systems*, volume 33, pages 14544–14555, 2020.

[12] Z. Li, T. Wang, and S. Arora. What happens after sgd reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*, 2021.

[13] Z. Li, T. Wang, and D. Yu. Fast mixing of stochastic gradient descent with normalization and weight decay. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=sof8l4cki9.

[14] V. Papyan. The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size. *arXiv preprint arXiv:1811.07062*, 2018.

[15] L. Pillaud-Vivien, J. Reygner, and N. Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In *Conference on Learning Theory*, 2022.

[16] L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, and L. Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[17] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *NeurIPS*, 2014.

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

[20] S. Wojtowytsch. Stochastic gradient descent with noise of machine learning type. Part I: Discrete time analysis. *arXiv preprint arXiv:2105.01650*, 2021.

[21] G. Zhang, C. Wang, B. Xu, and R. Grosse. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2018.

Appendix

A Training details

CIFAR-10/100 experiments. We train a VGG network without BatchNorm and preactivation ResNet-18 on CIFAR-10 and ResNet-34 on CIFAR-100 without data augmentations. We use standard SGD *without momentum* for all experiments. We note that ℓ_2 regularization and weight decay are exactly the same in this case. We use the standard He initialization [5] for all parameters. To make ResNets scale-invariant, we follow the approach of Li et al. [11] consisting of fixing the last layer, removing the learnable parameters of the normalization layers and adding a normalization layer in the skip connection. For the experiments in Fig.1, VGG is trained with LR = 0.1 and LR = 0.01 and weight decay parameter is fixed to be either $\lambda_{WD} = 0.0$ or $\lambda_{WD} = 0.008$. The ResNet-18 is trained with LR = 0.08 and LR = 0.001 and $\lambda_{WD} = 0.0$ or $\lambda_{WD} = 0.0125$. The ResNet-34 is trained with LR = 0.15 and LR = 0.001 and weight decay parameter $\lambda_{WD} = 0.0$ or $\lambda_{WD} = 0.01$. The total number of epochs is 1000 in all experiments in Fig.1 and all the LR are decayed at epoch 500 to 0.0001. For the experiments in Fig. 2 we use scale-invariant ResNet-18 and project the SGD iterates on the unitary sphere. We test the following LRs in the large-LR phase $(0.0001, 0.0005, 0.00075, 0.001, 0.002, 0.003, 0.004, 0.005)$ to show different generalization performance. After 100 epochs all the learning rates are decayed to the same value 0.0001. In Fig. 2c and Fig. 2d we finetune every 2 epochs for 100 additional epochs with LR=0.0001. For the experiments in Fig. 3 we test three different LRs $(0.1, 0.15, 0, 2)$ and decay all of them to 0.001 after the first 100 epochs. To obtain Fig. 3c and Fig. 3d we finetune every 2 epochs for 100 additional epochs with LR=0.001. All the experiments are conducted for 5 different random seeds.

B Weight decay for overparametrized deep networks: additional details and figures

B.1 A graphical illustration of the fine-tuning phase

Here, we plot an illustrative graphic in Figure 4 to give an idea of what happens during the fine-tuning phase.

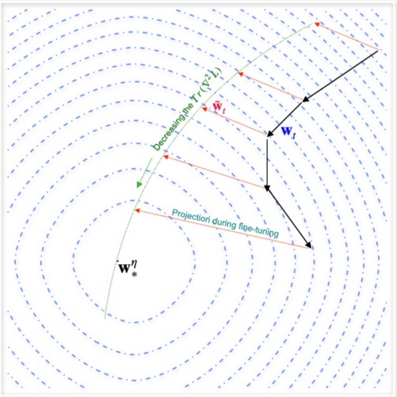

Decreasing the $Tr(\nabla^2 L)$
$\tilde{\mathbf{w}}_t$
$\mathbf{w}_t$
Projection during fine-tuning
$\mathbf{w}_*^\eta$

Figure 4: **A graphical illustration of the fine-tuning phase**.

B.2 Supporting material

Here we prove that the scale of noise is well approximated by training loss in the case of binary classification instead of classification in the case of multiple classes. The proof follows the lines of Wojtowytsch [20].

Proposition 3. *Assume $\|\mathbf{w}\| \in [a,b]$, for any $x \in \mathcal{D}$, $\|\nabla h(\mathbf{w}, x)\| \in [m, M]$ holds. For n sufficiently large, there exists constants c_1, c_2 such that*

$$c_1 \mathcal{L}(\mathbf{w}) \le \mathbb{E}\left[\|g(\mathbf{w})\|^2\right] \le c_2 \mathcal{L}(\mathbf{w})$$

Proof. The noise in the case when the gradient is computed at (x_i, y_i) is

$$g(\mathbf{w}) = \ell^{'}(y_i, h(\mathbf{w}, x_i))\nabla h(\mathbf{w}, x_i) - \frac{1}{n}\sum_i \nabla \ell^{'}(y_i, h(\mathbf{w}, x_i))\nabla h(\mathbf{w}, x_i),$$

Taking the expectation over uniform sampling over i, we have,

$$\mathbb{E}\|g\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\ell^{'}(y_i, h(\mathbf{w}, x_i))\right)^2 \|\nabla h(\mathbf{w}, x_i)\|^2 - \frac{1}{n^2}\|\sum_i \nabla \ell^{'}(y_i, h(\mathbf{w}, x_i))\nabla h(\mathbf{w}, x_i)\|^2 \quad (5)$$

<u>**Upper bound**</u>: Using the self-bounding property of the binary cross entropy, i.e., $\left(\ell'^2\right) \le l$ and $\|\nabla h(\mathbf{w}, x)\|^2 \le M^2$.

$$\mathbb{E}\|g\|^2 \le M^2 \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h(\mathbf{w}, x_i)) = M^2 \mathcal{L}(\mathbf{w}).$$

<u>**Lower bound**</u>: Again since the iterates are bound, we can assume there exists a constant c such that $\left(\ell'^2\right) \ge cl$. as the second term in 5 is decreasing with $O(n^{-2})$, we can assume that the first term is dominating and relevant and can lower bound the first term as,

$$\mathbb{E}\|g\|^2 \ge cm^2 \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h(\mathbf{w}, x_i)) = cm^2 \mathcal{L}(\mathbf{w}).$$

This proves the proposition. □

B.3 Comparison with the related works

Our focus is on an empirical illustration of the implicit regularization phenomenon, hence we refrain from attempting to prove this general conjecture, which we believe is a challenging task. The existing theoretical works [2, 12, 3] present two major weaknesses; they are essentially limiting analysis and as such fail at capturing the entire trajectory and they primarily target regression tasks. The powerful mathematical framework for scale-invariant networks developed by Li and Arora [10], Li et al. [11] allows them to study in detail the benefits of normalization and its interplay with weight decay. By means of this framework, they state a fast equilibrium conjecture, which gives qualitative guarantees for the speed of convergence of the stochastic process to the stationary distribution in function space. They disentangle the evolution of the norm and the direction of the parameters and show how the evolution of the direction only depends on the intrinsic LR $\lambda_i = \eta\lambda$. However, a qualitative description of the stationary distribution, its dependence on this intrinsic LR and the relationship with generalization is missing [11, Figure 3(d)]. We attempt to fill this gap by providing a qualitative depiction of the stationary distribution and its dependence on the intrinsic LR shading some light towards understanding the relationship with generalization. The work of Kodryan et al. [8] reports a similar observation, where the best test loss is achieved at a LR where the loss neither converges nor diverges but does not provide any explanation.

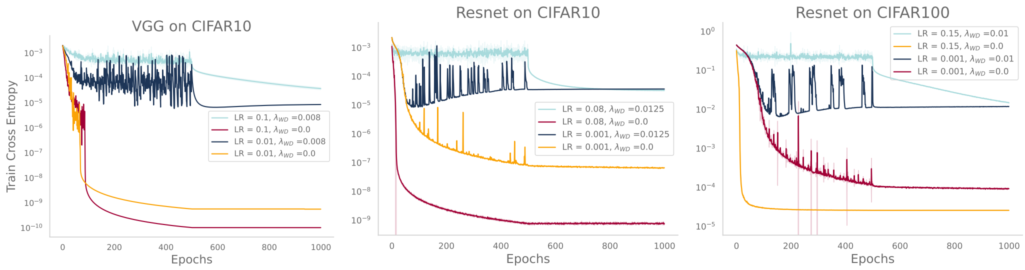
On the benefit of normalization. Our conjecture characterizes the mixing distribution but does not delve into the speed of the mixing process. In our experiments, we observe that normalization plays a pivotal role in the speed of mixing. Li et al. [11] observes a similar phenomenon in the case of scale-invariant networks, specifically the fast equilibrium conjecture, which is addressed by Li et al. [13]. We note that this phenomenon persists even when the models are not scale-invariant.

Mixing in the function space. A simpler conjecture could have been that the iterates $(\mathbf{w}_t)_{t\ge 0}$ mix towards a solution of the regularized objective $\mathbf{w}^*_\eta$. However, Li et al. [11] argues against mixing in the parameter space, emphasizing the necessity of considering the function space. Hence, our conjecture is formulated to capture stationarity in function space.

B.4 Additional figures

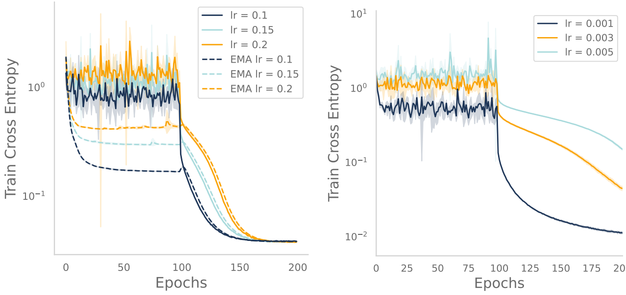
In this section, we report additional experimental results related to Section 2 in the main text.

Training curves for VGG and ResNets. In Fig. 5 we report the train cross entropy for VGG and ResNet18 on CIFAR10 and ResNet34 trained on CIFAR100. We can observe how when weight decay is used in combination with large LR, the train cross entropy stabilizes at some approximately constant level. In Fig. 6 we report the train cross entropy for scale-invariant ResNet on the sphere in Fig. 6b and for standard ResNet trained with weight decay and different large LRs in Fig. 6a. In both cases we can observe different levels of stabilization for the cross entropy depending on the LR deployed in the large-LR phase.



(a) (b) (c)

Figure 5: **Training with and w/o weight decay.** We report the train cross entropy for VGG (5a) and ResNet (5b, 5c) trained on CIFAR10/100 with and without weight decay and with small and large learning rates. After the first 500 epochs the learning rate is decayed to $\eta = 10^{-4}$ for all the curves.



(a) (b)

Figure 6: **Cross-entropy of standard and scale-invariant ResNets** we train on CIFAR10 with three different large LR for the first 100 epochs and decay it to $\eta = 10^{-3}$ for the standard ResNets with $\lambda_{WD} = 0.015$ Figure 6a and to $\eta = 10^{-3}$ for the scale-invariant ones 6b.

Connection between SGD covariance and Hessian. Much of the literature related to implicit bias relies on the assumption that the covariance of the noise of SGD is strictly related to the hessian of the loss function as discussed in Sec 2. Denoting the Hessian $\mathrm{H}(\mathbf{w}) := \nabla^2 \mathcal{L}(\mathbf{w})$ we can write it as the so-called Gauss-Newton decomposition [16, 14] $\mathrm{H}(\mathbf{w}) = \mathrm{G}(\mathbf{w}) + \mathrm{E}(\mathbf{w})$. To measure the cosine similarity (CS) between $\mathrm{w}(\mathbf{w})$ and the covariance Σ_t we compute

$$CS = \mathbb{E}\left[\cos\left(\mathrm{H}(\mathbf{w})v, \Sigma_t v\right)\right]$$

where v is sampled from the Gaussian distribution in $\mathbb{R}^p$ and $\cos(u, v) = {}^{\langle u,v\rangle}/\left\|u\right\|\left\|v\right\|$. The results are reported in Fig. 7

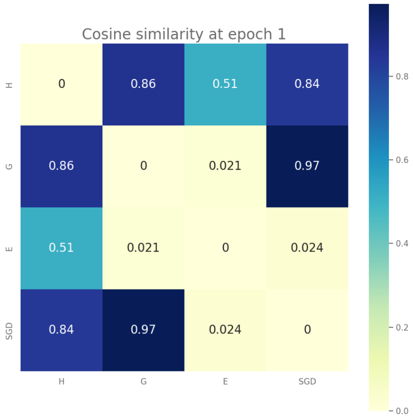


Figure 7: **Cosine similarity between hessian and Noise covariance:** we compute the cosine similarity between the hessian and the covariance of the SGD noise for a scale-invariant ResNet after one epoch with large lr $\eta = 0.005$. The results show how the two matrices are correlated and in particular how the SGD noise covariance is highly correlated with $\mathrm{G}(\mathbf{w})$.