Catastrophic Fisher Explosion: Early Phase Fisher Matrix Impacts Generalization

Stanisław Jastrzębski [1 2] **Devansh Arpit** [3] **Oliver Åstrand** [2] **Giancarlo Kerg** [4] **Huan Wang** [3] **Caiming Xiong** [3] **Richard Socher** [3] **Kyunghyun Cho** [* 2 5] **Krzysztof J. Geras** [* 1 2]

Abstract

The early phase of training a deep neural network has a dramatic effect on the local curvature of the loss function. For instance, using a small learning rate does not guarantee stable optimization because the optimization trajectory has a tendency to steer towards regions of the loss surface with increasing local curvature. We ask whether this tendency is connected to the widely observed phenomenon that the choice of the learning rate strongly influences generalization. We first show that stochastic gradient descent (SGD) implicitly penalizes the trace of the Fisher Information Matrix (FIM), a measure of the local curvature, from the start of training. We argue it is an implicit regularizer in SGD by showing that explicitly penalizing the trace of the FIM can significantly improve generalization. We highlight that poor final generalization coincides with the trace of the FIM attaining a large value early in training, to which we refer as catastrophic Fisher explosion. Finally, to gain insight into the regularization effect of penalizing the trace of the FIM, we show that it limits memorization by reducing the learning speed of examples with noisy labels more than that of the examples with clean labels.

1. Introduction

The exact mechanism behind implicit regularization effects in training of deep neural networks (DNNs) remains an extensively debated topic despite being considered a critical component in their empirical success (Neyshabur, 2017; Zhang et al., 2017; Jiang et al., 2020b). For instance, it is commonly observed that using a moderately large learning rate in the early phase of training results in better generalization (LeCun et al., 2012; Salakhutdinov, 2014; Jiang et al., 2020b; Bjorck et al., 2018).

Recent work suggests that the early phase of training of DNNs might hold the key to understanding some of these implicit regularization effects. In particular, the learning rate in the early phase of training has a dramatic effect on the local curvature of the loss function (Jastrzebski et al., 2020; Cohen et al., 2021). Recent work has found that when using a small learning rate, the local curvature of the loss surface increases along the optimization trajectory until optimization is *close to instability*.[1]

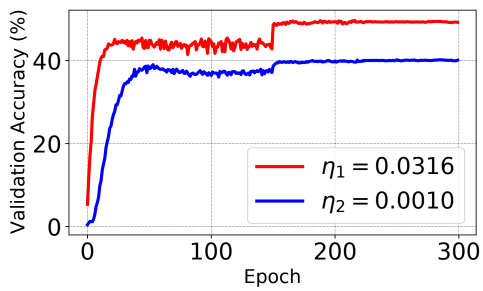
These observations lead to a natural question: does the instability, and the corresponding dramatic change in the local curvature, in the early phase of training influence generalization? We investigate this question through the lens of the Fisher Information Matrix (FIM), a matrix that can be seen as approximating the local curvature of the loss surface (Martens, 2020; Thomas et al., 2020). Achille et al. (2019); Jastrzebski et al. (2019); Golatkar et al. (2019); Lewkowycz et al. (2020) independently suggest that effects of the early phase of training on the local curvature critically influence the final generalization, but did not directly test this proposition.

Our main contribution is to show that *implicit regularization effects due to using a large learning rate can be explained by its impact on the trace of the FIM* ($\mathrm{Tr}(\mathbf{F})$)*, a quantity that reflects the local curvature, from the beginning of training*. Our results suggest that the instability in the early phase is a critical phenomenon for understanding optimization in DNNs. This is in contrast to many prior theoretical works, which generally do not connect implicit regularization effects in SGD to the large instability in the early phase of training (Chaudhari & Soatto, 2018; Smith et al., 2021).

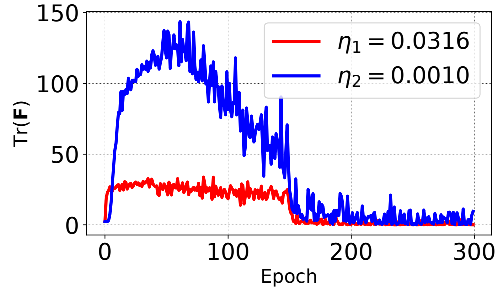
[*]Equal contribution [1]NYU Langone Medical Center, New York University, USA [2]Center of Data Science, New York University, USA [3]Salesforce Research, USA [4]Université de Montréal, Canada [5]CIFAR Azrieli Global Scholar. Correspondence to: Stanisław Jastrzębski <staszek.jastrzebski@gmail.com>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

[1]That is, a small further increase in the local curvature is not possible without divergence. Interestingly, when training using gradient descent with a learning rate η, the largest eigenvalue of the Hessian of the training loss was observed to reach the critical value of $\frac{2}{\eta}$ (Cohen et al., 2021), at which training oscillates along the eigenvector corresponding to the largest eigenvalue of the Hessian.



(a) Validation accuracy



(b) The trace of the Fisher Information Matrix

Figure 1: Catastrophic Fisher explosion phenomenon demonstrated for Wide ResNet trained using stochastic gradient descent on the TinyImageNet dataset. Training with a small learning rate leads to a sharp increase in the trace of the Fisher Information Matrix (FIM) early in training (right), which coincides with strong overfitting (left). The trace of the FIM is a measure of the local curvature of the loss surface. Training is done with either a learning rate optimized using grid search ($\eta_1 = 0.0316$, red), or a small learning rate ($\eta_2 = 0.001$, blue).

We demonstrate on image classification tasks that $\mathrm{Tr}(\mathbf{F})$ early in training correlates with the final generalization performance across settings with different learning rates or batch sizes. We then show evidence that explicitly regularizing $\mathrm{Tr}(\mathbf{F})$, which we call Fisher penalty, recovers generalization degradation due to training with a sub-optimal (small) learning rate, and can significantly improve generalization when training with the optimal learning rate. On the other hand, achieving large $\mathrm{Tr}(\mathbf{F})$ early in training, which may occur in practice when using a relatively small learning rate, or due to bad initialization, coincides with poor generalization. We call this phenomenon catastrophic Fisher explosion. Figure 1 illustrates this effect on the TinyImageNet dataset (Le & Yang, 2015).

Our second contribution is an analysis of why implicitly or explicitly regularizing $\mathrm{Tr}(\mathbf{F})$ impacts generalization. Our key finding is that penalizing $\mathrm{Tr}(\mathbf{F})$ significantly improves generalization in training with noisy labels. We make a theoretical and empirical argument that penalizing $\mathrm{Tr}(\mathbf{F})$ can be seen as penalizing the gradient norm of noisy examples, which slows down their learning. We hypothesize that implicitly or explicitly regularizing $\mathrm{Tr}(\mathbf{F})$ amplifies the implicit bias of SGD to avoid memorization (Arpit et al., 2017; Rahaman et al., 2019). Finally, we also show that small $\mathrm{Tr}(\mathbf{F})$ in the early phase of training biases optimization towards a flat minimum (Keskar et al., 2017).

2. Implicit and explicit regularization of the Fisher Information Matrix

Fisher Information Matrix Consider a probabilistic classification model $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$, where $\boldsymbol{\theta}$ denotes its parameters. Let $\ell(\boldsymbol{x}, y; \boldsymbol{\theta})$ be the cross-entropy loss function calculated for input $\boldsymbol{x}$ and label y. Let $g(\boldsymbol{x}, y; \boldsymbol{\theta}) = \frac{\partial}{\partial \theta}\ell(\boldsymbol{x}, y; \boldsymbol{\theta})$ denote the gradient of the loss computed for an example $(\boldsymbol{x}, y)$.

The central object that we study is the Fisher Information Matrix $\mathbf{F}$ defined as

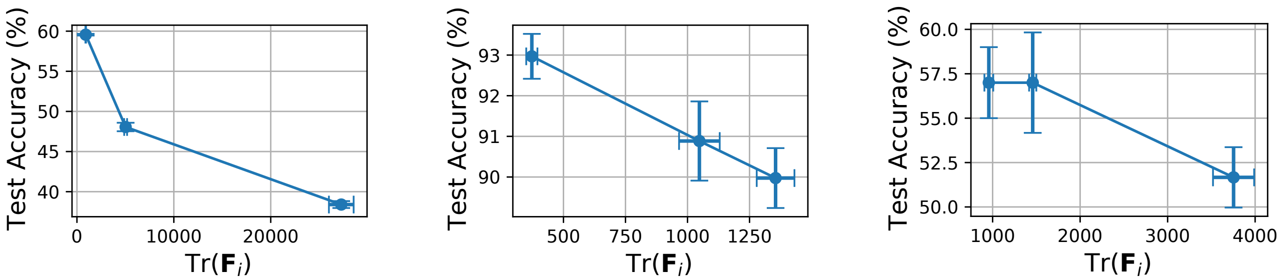
$$\mathbf{F}(\boldsymbol{\theta}) = \mathbb{E}_{x\sim\mathcal{X}, \hat{y}\sim p_\theta(y|\boldsymbol{x})}[g(\boldsymbol{x}, \hat{y})g(\boldsymbol{x}, \hat{y})^T], \qquad (1)$$

where the expectation is often approximated using the empirical distribution $\mathcal{X}$ induced by the training set. Later, we also look into the Hessian $\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta}^2}\ell(\boldsymbol{x}, y; \boldsymbol{\theta})$. We denote the trace of $\mathbf{F}$ and $\mathbf{H}$ matrices by $\mathrm{Tr}(\mathbf{F})$ and $\mathrm{Tr}(\mathbf{H})$.

The FIM can be seen as an approximation to the Hessian (Martens, 2020). In particular, as $p(y|\boldsymbol{x}; \theta) \to \hat{p}(y|\boldsymbol{x})$, where $\hat{p}(y|\boldsymbol{x})$ is the empirical label distribution, the FIM converges to the Hessian. Thomas et al. (2020) showed on image classifications tasks that $\mathrm{Tr}(\mathbf{H}) \approx \mathrm{Tr}(\mathbf{F})$ along the optimization trajectory, which we also demonstrate in Supplement F. Crucially, note that while $\mathrm{Tr}(\mathbf{H})$ uses label information, $\mathrm{Tr}(\mathbf{F})$ does not use any label information, in contrast to the "empirical Fisher" studied for example in Kunstner et al. (2019).

Fisher Penalty The early phase has a drastic effect on the trajectory in terms of the local curvature of the loss surface (Achille et al., 2019; Jastrzebski et al., 2019; Gur-Ari et al., 2018; Lewkowycz et al., 2020; Leclerc & Madry, 2020). In particular, Lewkowycz et al. (2020); Jastrzebski et al. (2019) show that using a large learning rate in stochastic gradient descent biases training towards low curvature regions of the loss surface early in training. For example, using a large learning rate in SGD was shown to result in a rapid decay of $\mathrm{Tr}(\mathbf{H})$ along the optimization trajectory (Jastrzebski et al., 2019).

Our main contribution is to propose and investigate a specific mechanism by which using a large learning rate or a small batch size implicitly influences final generalization. Our first insight is to shift the focus from studying the Hessian, to studying the properties of the FIM. Concretely, we



(a) ImageNet (w/o augmentation)

(b) CIFAR-10 (with augmentation)

(c) CIFAR-100 (w/o augmentation)

Figure 2: Association between $\text{Tr}(\mathbf{F})$ in the initial phase of training ($\text{Tr}(\mathbf{F_i})$) and test accuracy on ImageNet, CIFAR-10 and CIFAR-100 datasets. Each point corresponds to multiple seeds and a specific value of learning rate. $\text{Tr}(\mathbf{F_i})$ is recorded during the early phase of training (2-7 epochs, see the main text for details). The plots show that early $\text{Tr}(\mathbf{F})$ is predictive of final generalization. Analogous results illustrating the influence of batch size are shown in Appendix A.1

.

hypothesize that using a large learning rate or a small batch size improves generalization by implicitly penalizing $\text{Tr}(\mathbf{F})$ from the very beginning of training.

In order to study the effect of implicit regularization of $\text{Tr}(\mathbf{F})$, we introduce a regularizer that explicitly penalizes $\text{Tr}(\mathbf{F})$. First, we note that $\text{Tr}(\mathbf{F})$ can be written as

$$\text{Tr}(\mathbf{F}) = \mathbb{E}_{x\sim\mathcal{X},\hat{y}\sim p_\theta(y|\boldsymbol{x})}\left[\|\frac{\partial}{\partial\theta}\ell(\boldsymbol{x},\hat{y})\|_2^2\right]. \quad (2)$$

Thus, to regularize $\text{Tr}(\mathbf{F})$, we can simply add $\frac{1}{M}\sum_{i=1}^{M}\|g(\boldsymbol{x}_i,\hat{y}_i)\|^2$ term to the loss function, which can be efficiently back-propagated through. This, however, requires a large number of samples to efficiently regularize $\text{Tr}(\mathbf{F})$, as we show in our experiments. Instead, we add the following term to the loss function:

$$\ell'(\boldsymbol{x},y) = \frac{1}{B}\sum_{i=1}^{B}\ell(\boldsymbol{x}_i,y_i) + \alpha\left\|\frac{1}{B}\sum_{i=1}^{B}g(\boldsymbol{x}_i,\hat{y}_i)\right\|^2, \quad (3)$$

where $(\boldsymbol{x},y)$ is a mini-batch of size B, $\hat{y}_i$ is sampled from $p_{\boldsymbol{\theta}}(y|\boldsymbol{x}_i)$, α is a hyperparameter. Importantly, the equation does not involve target labels. Finally, we compute the gradient of the second term only every 10 optimization steps, and in a given iteration use the most recently computed gradient. We refer to this regularizer as Fisher penalty (FP).

In the experiments, we show that this formulation efficiently penalizes $\text{Tr}(\mathbf{F})$. We attribute this largely to the fact that $\left\|\frac{1}{B}\sum_{i=1}^{B}g(\boldsymbol{x}_i,\hat{y}_i)\right\|^2$ and $\text{Tr}(\mathbf{F})$ are strongly correlated during training. We discuss this observation, and ablate the approximations, in more detail in Supplement C.

Catastrophic Fisher Explosion To illustrate the concepts mentioned in this section, we train a Wide ResNet model (depth 44, width 3) (Zagoruyko & Komodakis, 2016) on the TinyImageNet dataset with SGD and two different learning rates. We illustrate in Figure 1 that the small learning rate leads to dramatic overfitting, which coincides with a sharp increase in $\text{Tr}(\mathbf{F})$ in the early phase of training. We also show in Supplement D that these effects cannot be explained by the difference in learning speed between runs with smaller and learning rates. We call this phenomenon catastrophic Fisher explosion.

Why Fisher Information Matrix The benefit of the FIM is that it can be efficiently regularized during training. In contrast, Wen et al. (2018); Foret et al. (2021) had to rely on certain approximations to efficiently regularize curvature. Equally importantly, the FIM is related to the gradient norm, and as such its effect on learning is more interpretable. We will leverage this connection to argue that FP slows down learning on noisy examples in the dataset.

Concurrent work on a different mechanism Barrett & Dherin (2021); Smith et al. (2021) concurrently argue that the implicit regularization effects in SGD can be expressed as a form of a gradient norm penalty. The key difference is that we link the mechanism of implicit regularization to the fact optimization is *close to instability* in the early phase. Supporting this view, we observe that FP empirically performs better than gradient penalty, and that FP works best when applied from the start of training.

3. Early-phase $\text{Tr}(\mathbf{F})$ and final generalization

Using a large learning rate (η) or small batch size (S) in SGD steers optimization to a lower curvature region of the loss surface. However, it remains a hotly debated topic whether such choices explain strong regularization effects (Dinh et al., 2017; Yoshida & Miyato, 2017; He et al., 2019; Tsuzuku et al., 2020). We begin by studying the con-

Table 1: Fisher penalty (FP) effectively models implicit regularization that arises in SGD due to using large learning rates. Using a 10-30x smaller learning rate (Baseline) results in up to 9% degradation in test accuracy on popular image classification benchmarks (c.f. to *optimal* η^*). Adding FP, which explicitly regularizes $\mathrm{Tr}(\mathbf{F})$, substantially improves generalization and closes the gap to η^*. Green cells correspond to runs that finished with, at most, 1% lower test accuracy than when training with $\eta*$.

Setting	η^*	Baseline	$\mathrm{GP_x}$	GP	FP	$\mathrm{GP_r}$
Wide ResNet / TinyImageNet (aug.)	54.67%	52.57%	52.79%	56.44%	**56.73%**	55.41%
DenseNet / CIFAR-100 (w/o aug.)	66.09%	58.51%	62.12%	64.42%	**66.41%**	66.39%
VGG11 / CIFAR-100 (w/o aug.)	45.86%	36.86%	45.26%	47.35%	**49.87%**	48.26%
WResNet / CIFAR-100 (w/o aug.)	53.96%	46.38%	**58.68%**	57.68%	57.05%	58.15%
SimpleCNN / CIFAR-10 (w/o aug.)	76.94%	71.32%	75.68%	75.73%	79.66%	**79.76%**

nection between $\mathrm{Tr}(\mathbf{F})$ and generalization in experiments across which we vary η or S in SGD.

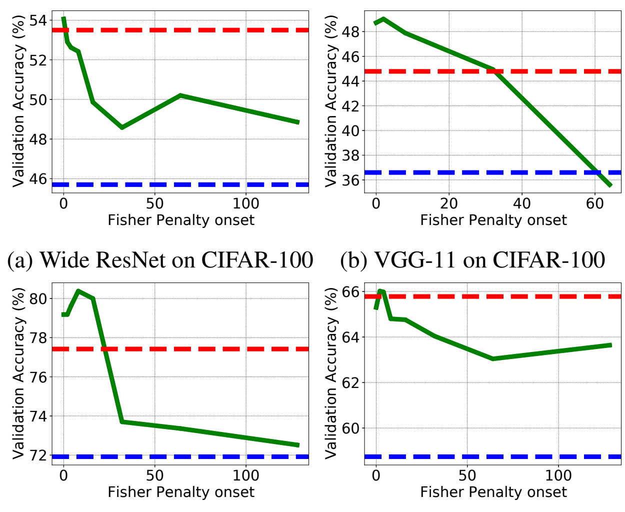
Experimental setup We run experiments in two settings: (1) ResNet-18 with Fixup (He et al., 2016; Zhang et al., 2019) trained on the ImageNet dataset (Deng et al., 2009), (2) ResNet-26 initialized as in (Arpit et al., 2019) and trained on the CIFAR-10 and CIFAR-100 datasets (Krizhevsky, 2009). We train each architecture using SGD, with various values of η, S, and random seed.

We define $\mathrm{Tr}(\mathbf{F_i})$ as $\mathrm{Tr}(\mathbf{F})$ during the initial phase of training. How long we consider the early-phase $\mathrm{Tr}(\mathbf{F})$ to be is determined by measuring when the training loss crosses a task-specific threshold ϵ that roughly corresponds to the moment when $\mathrm{Tr}(\mathbf{F})$ achieves its maximum value. For ImageNet, we use learning rates 0.001, 0.01, 0.1, and $\epsilon = 3.5$. For CIFAR-10, we use learning rates 0.007, 0.01, 0.05, and $\epsilon = 1.2$. For CIFAR-100, we use learning rates 0.001, 0.005, 0.01, and $\epsilon = 3.5$. In all cases, training loss reaches ϵ between 2 and 7 epochs across different hyper-parameter settings. We repeat similar experiments for different batch sizes in Supplement A.1. The remaining training details can be found in Supplement I.1.

Results Figure 2 shows the association between $\mathrm{Tr}(\mathbf{F_i})$ and test accuracy across runs with different learning rates. We show results for CIFAR-10 and CIFAR-100 when varying the batch size in Figure 7 in the Supplement. We find that $\mathrm{Tr}(\mathbf{F_i})$ correlates well with the final generalization in our setting, which provides initial evidence for the importance of $\mathrm{Tr}(\mathbf{F})$. It also serves as a stepping stone towards developing a more granular understanding of the role of implicit regularization of $\mathrm{Tr}(\mathbf{F})$ in the following sections.

4. Fisher Penalty

To understand the significance of the identified correlation between $\mathrm{Tr}(\mathbf{F_i})$ and generalization, we now run experiments in which we directly penalize $\mathrm{Tr}(\mathbf{F})$. We focus our attention on the identified effect that using a high learning rate has on $\mathrm{Tr}(\mathbf{F})$, especially early in training.



(a) Wide ResNet on CIFAR-100 (b) VGG-11 on CIFAR-100

(c) Simple CNN on CIFAR-10 (d) DenseNet on CIFAR-100

Figure 3: Fisher Penalty has to be applied early in training to close the generalization gap to the optimal learning rate (c.f. the red horizontal line to the blue horizontal line). Each subplot summarizes an experiment in which we apply Fisher Penalty starting from a certain epoch (x axis) and measure the final test accuracy (y axis).

Experimental setting We use a similar setting as in the previous section, but we include larger models. We run experiments using Wide ResNet (Zagoruyko & Komodakis, 2016) (depth 44 and width 3, with or without BN layers), SimpleCNN (without BN layers), DenseNet (L=40, K=12) (Huang et al., 2017) and VGG-11 (Simonyan & Zisserman, 2015). We train these models on either the CIFAR-10 or the CIFAR-100 datasets. Due to larger computational cost, we replace ImageNet with the TinyImageNet dataset (Le & Yang, 2015) (with images scaled to 32×32 resolution) in these experiments.

To investigate if the correlation between $\mathrm{Tr}(\mathbf{F_i})$ and the final generalization holds more generally, we apply Fisher penalty in two settings. First, we use a learning rate 10-30x smaller than the optimal one, which both incur up to 9% degradation in test accuracy and results in large value of $\mathrm{Tr}(\mathbf{F_i})$. We also remove data augmentation from the CIFAR-10 and the CIFAR-100 datasets to ensure that training a with small learning rate does not result in underfitting. In the second setting, we add Fisher penalty in training with an optimized learning rate using grid search (η^*) and train with data augmentation.

Fisher penalty penalizes the gradient norm computed using labels sampled from $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$. We hypothesize that a similar, but weaker, effect can be introduced by other gradient norm regularizers. We compare FP to: (1) penalizing the input gradient norm $\|\boldsymbol{g}_x\|^2 = \|\frac{\partial}{\partial \boldsymbol{x}}\ell(\boldsymbol{x}, y)\|^2$, which we denote by $\mathrm{GP_x}$ (Varga et al., 2018; Rifai et al., 2011; Drucker & Le Cun, 1992); (2) penalizing the vanilla mini-batch gradient (Gulrajani et al., 2017), which we denote by GP; and penalizing the mini-batch gradient computed with random labels $\|\boldsymbol{g_r}\|^2 = \|\frac{\partial}{\partial \boldsymbol{x}}\ell(\boldsymbol{x}, \hat{y})\|^2$ where $\hat{y}$ is sampled from a uniform distribution over the label set ($\mathrm{GP_r}$). We are not aware of any prior work using GP or $\mathrm{GP_r}$ in supervised training, with the exception of Alizadeh et al. (2020) where the authors penalized ℓ_1 norm of gradients to compress the network towards the end of training, and concurrent work of Barrett & Dherin (2021). We note that regularizing $\mathrm{GP_x}$ is related to regularizing the Jacobian input-output of the network (Hoffman et al., 2019; Chan et al., 2020).

We tune the hyperparameters on the validation set. More specifically for α, we test 10 different values spaced uniformly between $10^{-1} \times v$ and $10^{1} \times v$ on a logarithmic scale with $v \in \mathbb{R}_+$. For TinyImageNet, we evaluate 5 values spaced equally on a logarithmic scale. We include the remaining experimental details in the Supplement I.2.

Fisher Penalty improves generalization Table 1 summarizes the results of the main experiment. First, we observe that a suboptimal learning rate (10-30x lower than the optimal) leads to dramatic overfitting. We observe a degradation of up to 9% in test accuracy, while achieving approximately 100% training accuracy (see Table 6 in the Supplement).

Fisher penalty closes the gap in test accuracy between the small and optimal learning rate, and even achieves better performance than the optimal learning rate. A similar performance was observed when minimizing $\|g_r\|^2$. We will come back to this observation in the next section.

GP and $\mathrm{GP_x}$ reduce the early value of $\mathrm{Tr}(\mathbf{F})$ (see Table 4 in the Supplement). They, however, generally perform worse than $\mathrm{Tr}(\mathbf{F})$ or $\mathrm{GP_r}$ and do not fully close the gap between small and optimal learning rate. We hypothesize they improve generalization by a similar but less direct mechanism than $\mathrm{Tr}(\mathbf{F})$ and $\mathrm{GP_r}$, which we make more precise in Section G in the Supplement. We note that GP was proposed in Barrett & Dherin (2021); Smith et al. (2021) as the term that SGD implicitly regularizes (see Related Work for more details).

Table 2: Fisher penalty (FP) improves generalization in 4 out of 5 settings when applied with the optimal learning rate η^* and trained using standard data augmentation. In 3 out of 5 settings the difference between FP and η^* is relatively small (below 2%), which is expected given that FP is aimed at reproducing the regularization effect of large η, and we compare to training with the optimal η^*.

Setting	η^*	FP
WResNet / TinyImageNet	54.70±0.0%	**60.00±0.1%**
DenseNet / C100	**74.41±0.5%**	74.19±0.5%
VGG11 / C100	59.82±1.2%	**65.08±0.5%**
WResNet / C100	69.48±0.3%	**71.53±1.2%**
SimpleCNN / C10	87.16±0.2%	**87.52±0.5%**

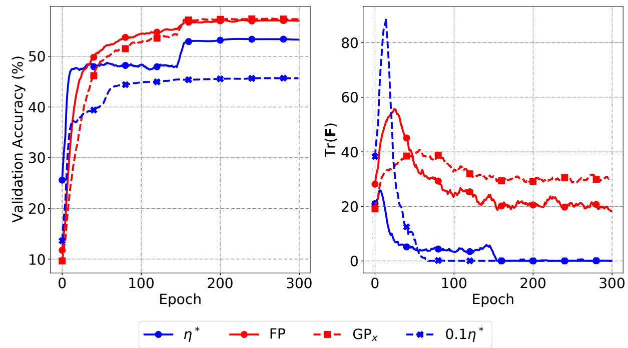
We also investigate whether similar conclusions hold in large batch size training. In experiments with CIFAR-10 and SimpleCNN, we find that we can close the generalization gap due to training with a large batch size by using Fisher Penalty. We provide further details in Supplement E.

In the second experimental setting, we apply FP to a network trained with the optimal learning rate η^*. According to Table 2 (see Table 7 for training accuracies), Fisher Penalty improves generalization in 4 out of 5 settings. The gap between the baseline and FP is relatively small in 3 out of 5 settings (below 2%), which is natural given that we already regularize training implicitly by using the optimal η..

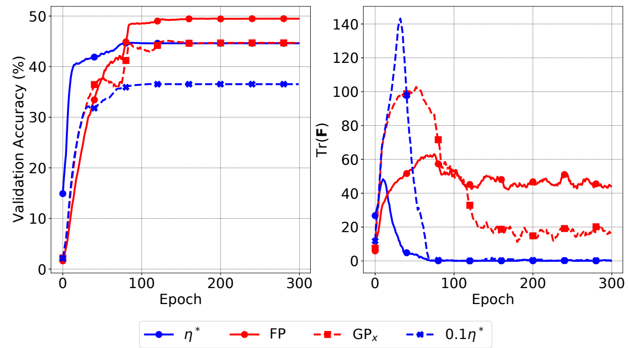
Geometry and generalization in the early phase of training Here, we investigate if penalizing $\mathrm{Tr}(\mathbf{F})$ early in training matters for the final generalization. A positive answer would further strengthen the link between the early phase of training and implicit regularization effects in SGD. We run experiments on CIFAR-10 and CIFAR-100.

First, we observe that all gradient-norm regularizers reduce the early value of $\mathrm{Tr}(\mathbf{F})$ closer to $\mathrm{Tr}(\mathbf{F})$ achieved when trained with the optimal learning rate η^*. We show this effect with Wide ResNet and VGG-11 on CIFAR-100 in Figure 4, and for other experimental settings in the Supplement. We also tabulate the maximum achieved values of $\mathrm{Tr}(\mathbf{F})$ over the optimization trajectory in Supplement A.2.

To test the importance of explicitly penalizing $\mathrm{Tr}(\mathbf{F})$ early in training, we start applying it after a certain number of



(a) Wide ResNet on CIFAR-100 (w/o aug.)



(b) VGG-11 on CIFAR-100 (w/o aug.)

Figure 4: Using Fisher Penalty (FP) in training with a sub-optimal (small) learning rate drastically reduces the tendency to reach highly curved regions of the loss surface that arises when training with a small learning rate (compare the peak values of $\mathrm{Tr}(\mathbf{F})$). At the same time, FP also significantly improves generalization. Penalizing the input gradient norm ($\mathrm{GP_x}$) also impacts $\mathrm{Tr}(\mathbf{F})$ but achieves worse generalization. Each subfigure shows validation accuracy (left) and $\mathrm{Tr}(\mathbf{F})$ (right). Curves were smoothed for clarity.

epochs $E \in \{1, 2, 4, 8, 16, 32, 64, 128\}$. We use the best hyperparameter set from the previous experiments. Figure 3 summarizes the results. For both datasets, we observe a consistent pattern. When FP is applied starting from a later epoch, final generalization is significantly worse, and the generalization gap arising from a suboptimal learning rate is not closed. Interestingly, there seems to be a benefit in applying FP after some warm-up period, which might be related to the widely used trick to gradually increase the learning rate in the early phase (Gotmare et al., 2019).

4.1. Fisher Penalty Reduces Memorization

It is not self-evident why regularizing $\mathrm{Tr}(\mathbf{F})$ should influence generalization. In this section, we show that explicit penalization of $\mathrm{Tr}(\mathbf{F})$ improves learning with datasets with noisy labels. To study this, we replace labels of the examples in the CIFAR-100 dataset (25% or 50% of the training set) with labels sampled uniformly. We refer to these examples as *noisy examples*. While label noise in real datasets is not uniform, methods that perform well with uniform label noise generally are more robust to label noise in real datasets (Jiang et al., 2020a). We also know that datasets such as CIFAR-100 contain many labeling errors (Song et al., 2020). Hence, examining if $\mathrm{Tr}(\mathbf{F})$ reduces memorization of synthetic label noise provides an insight into why it improves generalization in our prior experiments.

We argue FP should reduce memorization. Under certain assumptions, $\mathrm{Tr}(\mathbf{F})$ is equivalent to the norm of the noisy examples' gradient. As such, we would expect $\mathrm{Tr}(\mathbf{F})$ to decrease the norm of the gradient of the noisy examples compared to the gradient of the clean examples. Assuming that the gradient norm of a given group of examples is related to its learning speed (Chatterjee, 2020; Fort et al., 2020b), FP should promote learning first clean examples, before learning noisy examples. We make this argument more precise in the Supplement H.

To study whether the above happens in practice, we compare FP to $\mathrm{GP_x}$, $\mathrm{GP_r}$, and mixup (Zhang et al., 2018). While mixup is not the state-of-the-art approach for learning with noisy labels, it is competitive among approaches that do not require additional data nor multiple stages of training. In particular, it is a component in several state-of-the-art methods (Li et al., 2020; Song et al., 2020). For gradient norm based regularizers, we evaluate 6 different hyperparameter values spaced uniformly on a logarithmic scale, and for mixup we evaluate $\beta \in \{0.2, 0.4, 0.8, 1.6, 3.2, 6.4\}$. We experiment with the Wide ResNet and VGG-11 models. We describe remaining experimental details in Supplement I.3.

Results To test whether FP reduces the speed with which the noisy examples are learned, we track the training and validation accuracy, and the gradient norm. We compute these metrics separately on the noisy and clean examples. Figure 5 summarizes the results for VGG-11, and we show results for ResNet-32 in Figure 11 in the Supplement.

We observe that FP limits the ability of the model to memorize data more strongly than it limits its ability to learn from clean data. Figure 5 confirms applying FP results in training accuracy on noisy examples being lower for the same accuracy on clean examples, compared to the baseline.

We can further confirm our interpretation of the effect $\mathrm{Tr}(\mathbf{F})$ has on training by studying the gradient norm. As visible in the top panel of Figure 5, the gradient norm evaluated on noisy examples is larger than on clean examples, and the ratio is closer to 1 when FP is applied with a larger coefficient. Interestingly, we also observe that the angle between

Table 3: Fisher Penalty (FP) and GP_r both reduce memorization competitively to mixup. We measure test accuracy at the best validation point in training with either 25% or 50% examples with noisy labels in the CIFAR-100 dataset.

Label Noise	Setting	Baseline	Mixup	GP_x	FP	GP_r
25%	VGG-11 / CIFAR-100	41.74%	52.31%	45.94%	**60.18%**	58.46%
	ResNet-52 / CIFAR-100	53.30%	**61.61%**	52.70%	58.31%	57.60%
50%	VGG-11 / CIFAR-100	30.05%	39.15%	34.26%	**51.33%**	50.33%
	ResNet-52 / CIFAR-100	43.35%	**51.71%**	42.99%	47.99%	50.08%

the noisy and clean examples' gradients is negative early in training, which we plot in Figure 12 in the Supplement.

Finally, we summarize test accuracies (at the best validation epoch) in Table 3. Penalizing $\mathrm{Tr}(\mathbf{F})$ reduces memorization competitively to mixup, improving test accuracy by up to 21.28%. Furthermore, FP strongly outperforms penalizing GP_x, consistent with with trends in the prior Sections. We also again observe FP to perform comparably to penalizing GP_r. Under the assumption that the model predictive distribution is close to random, FP is equivalent to penalizing GP_r. Assuming this holds in the early phase of training, we interpret this as a further corroboration of the importance of applying FP in the early phase (c.f. Section 4). In the Supplement, we also report training accuracy separately on the clean and noisy examples for the experiment with 25% noisy examples.

Taken together, the results suggest implicit or explicit regularization of FP improves generalization at least in part by strengthening the bias of SGD to learn clean examples before learning noisy examples (Arpit et al., 2017). We also note that related conclusions about the effect using large learning rates were reached by Jastrzebski et al. (2017); Li et al. (2019).

5. Early $\mathrm{Tr}(\mathbf{F})$ influences final curvature

To provide further insight into why it is important to regularize $\mathrm{Tr}(\mathbf{F})$ during the early phase of training, we establish a connection between the early phase of training and the wide minima hypothesis (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017) which states that *flat* minima *typically* correspond to better generalization. Here, we use $\mathrm{Tr}(\mathbf{H})$ as a measure of flatness.

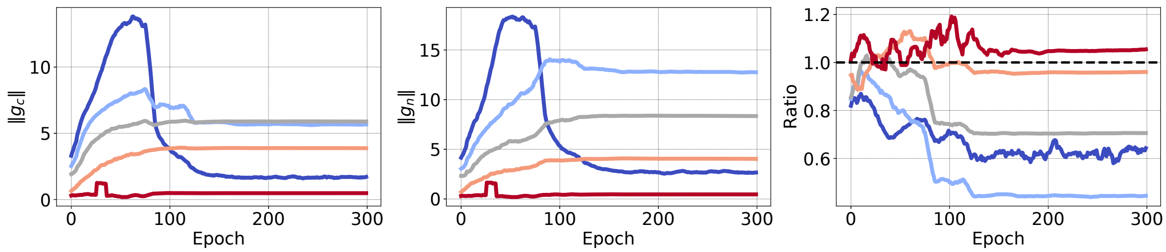
Experimental setting We investigate how likely it is for an optimization trajectory to end up in a wide minimum in two scenarios: 1) when optimization exhibits small $\mathrm{Tr}(\mathbf{F})$ early on, and 2) when optimization exhibits large $\mathrm{Tr}(\mathbf{F})$ early on. We train two ResNet-26 models for 20 epochs using high and low regularization configurations. At epoch 20 we record $\mathrm{Tr}(\mathbf{F})$ for each model. We then use these two models as initialization for 8 separate models each, and continue training using the low regularization configuration with different random seeds. The motivation behind this experiment is to investigate if the degree of regularization in the early phase biases the model towards minima with certain flatness ($\mathrm{Tr}(\mathbf{H})$) even though no further high regularization configurations are used during the rest of the training. For all these runs, we record the best test accuracy along the optimization trajectory along with $\mathrm{Tr}(\mathbf{H})$ at the point corresponding to the best test accuracy. We describe the remaining experimental details in Supplement I.4.

Results We present the result in Figure 6 for the CIFAR-100 datasets, and for CIFAR-10 in Supplement A.4. A training run with a lower $\mathrm{Tr}(\mathbf{F})$ during the early phase is more likely to end up in a wider minimum as opposed to one that reaches large $\mathrm{Tr}(\mathbf{F})$ during the early phase. This happens despite that the late phases of both sets of models use the low regularization configuration. The latter runs always end up in sharper minima. In Supplement I.4 we also show evolution of $\mathrm{Tr}(\mathbf{H})$ throughout training, which suggests that this behavior can be attributed to curvature stabilization happening early during training.

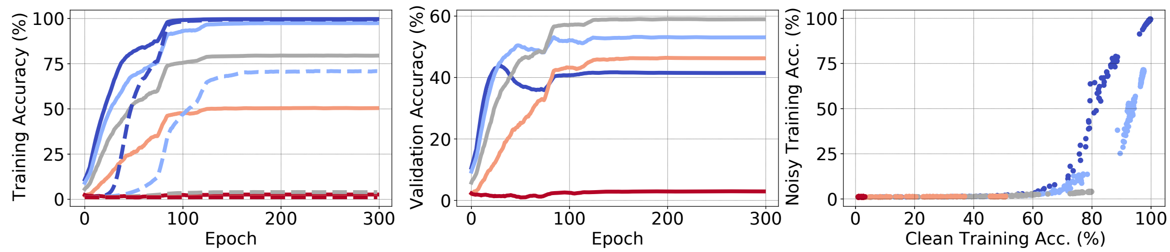
6. Related Work

Implicit regularization effects are critical to the empirical success of DNNs (Neyshabur, 2017; Zhang et al., 2017). Much of it is attributed to the choice of hyperparameters in SGD (Keskar et al., 2017; Smith & Le, 2018; Li et al., 2019), low complexity bias induced by gradient descent (Xu, 2018; Jacot et al., 2018; Arora et al., 2019; Hu et al., 2020), the cross-entropy loss function (Poggio et al., 2018; Soudry et al., 2018), or the importance of the early phase of training (Achille et al., 2019; Jastrzebski et al., 2020; Fort et al., 2020a; Frankle et al., 2020; Golatkar et al., 2019; Lewkowycz et al., 2020). However, developing a mechanistic understanding of how SGD implicitly regularizes DNNs remains a largely unsolved problem.

Many prior works have proposed explicit regularizers aimed at finding low curvature solutions (Hochreiter & Schmidhuber, 1997). Chaudhari et al. (2017) proposed a Langevin dynamics based algorithm. Wen et al. (2018); Izmailov et al. (2018); Foret et al. (2021) propose finding wide minima



(a) Gradient norm on clean examples (left), noisy examples (middle), and their ratio (right); evaluated on the training set.



(b) Training accuracy on clean and noisy examples (solid/dashed lined, left), validation accuracy (middle), and a scatter plot of training accuracy on clean vs noisy examples (right).

Figure 5: Fisher penalty reduces memorization by slowing down training on data with noisy labels more strongly than does training on clean data, as indicated by relative differences in gradient norm (top) and training accuracy (bottom) between the two groups of examples. Experiment run with VGG-11 on the CIFAR-100 dataset. Blue to red color represents increasing regularization coefficient (from 10^{-2} to 10^{1}).

through approximations involving averaging gradients or parameters at the neighborhood of the current parameter state. In contrast, our focus is on elucidating a mechanistic link behind the implicit regularization effects in SGD and the early phase of training. To corroborate our hypothesis, we develop Fisher Penalty, an efficient and novel explicit regularizer that aims to reproduce the regularization effect of training with a large learning rate.

Our work contributes to a better understanding of the role of the FIM in training and generalization of deep neural networks. The FIM was also used to define complexity measures such as the Fisher-Rao norm (Karakida et al., 2019; Liang et al., 2019), and can be seen as approximating the local curvature of the loss surface (Martens, 2020; Thomas et al., 2020). Most notably, the FIM defines the distance metric used in natural gradient descent (Amari, 1998).

Penalizing $\mathrm{Tr}(\mathbf{F})$ is related to penalizing the input gradient norm or the input-output Jacobian of the network, which were shown to be effective regularizers for deep neural networks (Drucker & Le Cun, 1992; Varga et al., 2018; Hoffman et al., 2019; Chan et al., 2020; Moosavi-Dezfooli et al., 2019). Most closely related is Moosavi-Dezfooli et al. (2019) who regularize the curvature in the input space using a stochastic approximation scheme. Our results suggest that penalizing FP is a more effective regularizer than other tested variants of gradient norm penalities. However, we did not run an extensive comparison. In particular, we did not compare directly to Moosavi-Dezfooli et al. (2019).

Chatterjee (2020); Fort et al. (2020b) show that SGD avoids memorization by extracting commonalities between examples due to following gradient descent directions shared between examples. Other works have also connected the implicit regularization effects of using large learning rates with preventing memorization (Arpit et al., 2017; Li et al., 2019; Jastrzebski et al., 2017). Liu et al. (2020) was first to note the difference in gradient norms between noisy and clean examples that emerges in the early phase of training. Our work is complementary to these findings. We make a direct connection between memorization, the instability in the early phase of training, and implicit regularization effects arising from using large learning rates in SGD.

Concurrent works have also proposed that implicit regularization effects in SGD can be understood as a form of gradient penalty. Barrett & Dherin (2021); Smith et al. (2021) show that SGD implicitly penalizes the mini-batch gradient norm, with the strength controlled by the learning rate, and study GP as an explicit regularizer. Analogously, we argue that SGD implicitly regularizes $\mathrm{Tr}(\mathbf{F})$, a closely related quantity, which can be expressed as the squared gradient norm under labels sampled from $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$. We found that penalizing GP did not always close the generalization gap due to using a small learning rate. Overall, our results suggest that the fact that optimization is *close to instability* (Jastrzebski et al., 2020) in the early phase is critical for

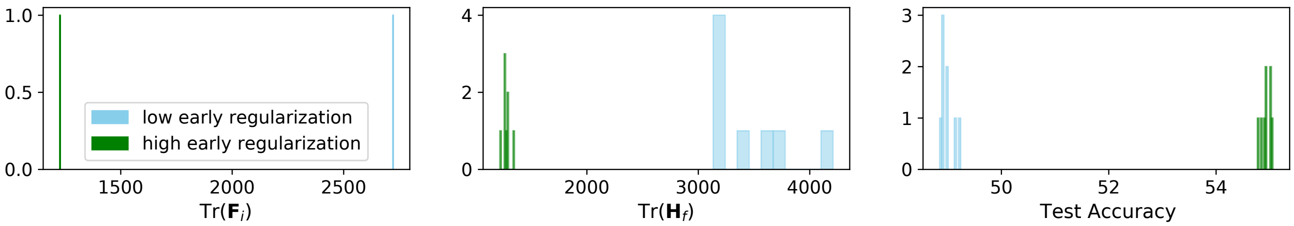


Figure 6: Optimization trajectories passing through regions with low $\text{Tr}(\mathbf{F})$ ($\text{Tr}(\mathbf{F_i})$) during the early phase of training reach wider minima. Two ResNet-56 models are trained with two different levels of regularization for 20 epochs on CIFAR-100. Each model is then continued trained using the low regularization configuration with different random seeds. Left: $\text{Tr}(\mathbf{F})$ at the end of the 20 epochs ($\text{Tr}(\mathbf{F_i})$). Middle: A histogram of $\text{Tr}(\mathbf{H})$ at the best test accuracy epoch along the trajectory ($\text{Tr}(\mathbf{H_f})$). Right: a histogram of test accuracy.

understanding implicit regularization effects in SGD.

7. Conclusion

The dramatic instability and changes in the curvature that happen in the early phase of training motivated us to probe its importance for generalization (Jastrzebski et al., 2020; Cohen et al., 2021). We investigated if these effects might explain some of the implicit regularization effects attributed to SGD such as the poorly understood generalization benefit of using large learning rates.

We showed evidence that using a large learning rate in SGD influences generalization by implicitly penalizing the trace of the Fisher Information Matrix ($\text{Tr}(\mathbf{F})$), a measure of the local curvature, from the beginning of training. We argued that (1) the value of early $\text{Tr}(\mathbf{F})$ correlates with final generalization, and (2) explicitly regularizing $\text{Tr}(\mathbf{F})$ can substantially improve generalization, and showed similar results for training with a small batch size. In the absence of implicit or explicit regularization, $\text{Tr}(\mathbf{F})$ can attain large values early in training, which we referred to as catastrophic Fisher explosion.

To better understand the mechanism by which penalizing $\text{Tr}(\mathbf{F})$ improves generalization, we investigated training on a dataset with incorrectly labeled examples. Our key finding is that penalizing $\text{Tr}(\mathbf{F})$ significantly reduces memorization by slowing down learning examples with incorrect labels. We hypothesize that penalizing the local curvature (by using large learning rates or penalizing $\text{Tr}(\mathbf{F})$) early in training improves generalization at least in part by strengthening the implicit bias of SGD to avoid learning noisy labels (Arpit et al., 2017; Li et al., 2019; Liu et al., 2020).

Developing theory that is fully consistent with our findings is an interesting topic for the future. Notably, existing theoretical works generally do not connect implicit regularization effects in SGD to the fact optimization is *close to instability* in the early phase of training (Li et al., 2017; Chaudhari & Soatto, 2018; Smith et al., 2021).

Another exciting topic for the future is connecting these findings to shortcut learning (Geirhos et al., 2020). The tendency of SGD to learn the simplest patterns in the datasets can be detrimental to the broader generalization of the model (Nam et al., 2020). We hope that by better understanding implicit regularization effects in SGD, our work will contribute to developing optimization methods that better optimize for both in and out of distribution generalization.

Limitations

Our work has several limitations. We used mini-batch gradients to approximate $\text{Tr}(\mathbf{F})$ in Fisher Penalty. While experiments suggest it is inconsequential for the main conclusions, more work would be needed to fully establish it. We also experimented only with vision tasks. Finally, we observed that FP slows down learning on clean examples. This feature makes it challenging to apply with very small learning rates, at which underfitting might start to be an issue.

Acknowledgments

GK acknowledges support in part by the FRQNT Strategic Clusters Program (2020-RS4-265502 - Centre UNIQUE - Union Neurosciences & Artificial Intelligence - Quebec). KC was supported by Samsung Advanced Institute of Technology (Next Generation Deep Learning: from pattern recognition to AI) and Samsung Research (Improving Deep Learning using Latent Structure). We thank Catriona C. Geras for proofreading the paper.

References

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Milad Alizadeh, Arash Behboodi, Mart van Baalen, Christos Louizos, Tijmen Blankevoort, and Max Welling. Gradient ℓ_1 regularization for quantization robustness. In *International Conference on Learning Representations*, 2020.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 1998.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, 2017.

Devansh Arpit, Víctor Campos, and Yoshua Bengio. How to initialize your network? robust initialization for weightnorm & resnets. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

David Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021.

Johan Bjorck, Carla P. Gomes, Bart Selman, and Kilian Q. Weinberger. Understanding batch normalization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 2018.

Alvin Chan, Yi Tay, Yew-Soon Ong, and Jie Fu. Jacobian adversarially regularized networks for robustness. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Satrajit Chatterjee. Coherent gradients: An approach to understanding generalization in gradient descent-based optimization. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

François Chollet and others. *Keras*. 2015.

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.

Soham De and Samuel L. Smith. Batch normalization biases residual blocks towards the identity function in deep networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 2009.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, 2017.

H. Drucker and Y. Le Cun. Improving generalization performance using double backpropagation. *IEEE Transactionsf on Neural Networks*, 1992.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

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

Stanislav Fort, Paweł Krzysztof Nowak, Stanislaw Jastrzebski, and Srini Narayanan. Stiffness: A new perspective on generalization in neural networks, 2020b.

Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The early phase of neural network training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace, 2018.

Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 1997.

Judy Hoffman, Daniel A. Roberts, and Sho Yaida. Robust learning with jacobian regularization. 2019.

Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 1990.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, 2018.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 2018.

Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos J. Storkey. Three Factors Influencing Minima in SGD. 2017.

Stanislaw Jastrzebski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos J. Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, 2020a.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020b.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, Proceedings of Machine Learning Research, 2019.

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

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

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

Y. Le and X. Yang. Tiny imagenet visual recognition challenge. 2015.

Guillaume Leclerc and Aleksander Madry. The two regimes of deep network training, 2020.

Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient BackProp. In *Neural Networks: Tricks of the Trade (2nd ed.)*. 2012.

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

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, 2017.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

Tengyuan Liang, Tomaso A. Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, Proceedings of Machine Learning Research, 2019.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in Neural Information Processing Systems*, 2020.

James Martens. New insights and perspectives on the natural gradient method, 2020.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020.

Behnam Neyshabur. Implicit regularization in deep learning. 2017.

Tomaso Poggio, Kenji Kawaguchi, Qianli Liao, Brando Miranda, Lorenzo Rosasco, Xavier Boix, Jack Hidary, and Hrushikesh Mhaskar. Theory of deep learning iii: explaining the non-overfitting puzzle, 2018.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron C. Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, 2019.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, 2011.

Ruslan Salakhutdinov. Deep learning. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Samuel L. Smith and Quoc V. Le. A bayesian perspective on generalization and stochastic gradient descent. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Samuel L Smith, Benoit Dherin, David Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

Jiaming Song, Yann Dauphin, Michael Auli, and Tengyu Ma. Robust and on-the-fly dataset denoising for image classification. In *Computer Vision – ECCV 2020*, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, and Nathan Srebro. The implicit bias of gradient descent on separable data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Valentin Thomas, Fabian Pedregosa, Bart van Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, Proceedings of Machine Learning Research, 2020.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, 2020.

Dániel Varga, Adrián Csiszárik, and Zsolt Zombori. Gradient regularization improves accuracy of discriminative models, 2018.

Wei Wen, Yandan Wang, Feng Yan, Cong Xu, Chunpeng Wu, Yiran Chen, and Hai Li. Smoothout: Smoothing out sharp minima to improve generalization in deep learning, 2018.

Zhiqin John Xu. Understanding training and generalization in deep learning by fourier analysis, 2018.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Residual learning without normalization via better initialization. In *International Conference on Learning Representations*, 2019.