Width of Minima Reached by Stochastic Gradient Descent is Influenced by Learning Rate to Batch Size Ratio

Width of Minima Reached by Stochastic Gradient Descent is Influenced by Learning Rate to Batch Size Ratio

Stanislaw Jastrzębski[1 2 3] , Zachary Kenton[1 3] , Devansh Arpit[3] , Nicolas Ballas[4] , Asja Fischer[5] , Yoshua Bengio[3 6] , Amos Storkey[7]

Abstract. We show that the dynamics and convergence properties of SGD are set by the ratio of learning rate to batch size. We observe that this ratio is a key determinant of the generalization error, which we suggest is mediated by controlling the width of the final minima found by SGD. We verify our analysis experimentally on a range of deep neural networks and datasets.

1 Introduction

Deep neural networks (DNNs) have demonstrated good generalization ability and achieved state-of-the-art performances in many application domains despite being massively over-parameterized, and despite the fact that modern neural networks are capable of getting an error close to zero on the training data [20]. What is the reason for their good generalization performance, remains an open question.

The standard way of training DNNs involves minimizing a loss function using stochastic gradient descent (SGD) and its variants [3]. Since the loss functions of DNNs are typically non-convex functions of the parameters, with complex structure and potentially multiple minima and saddle points, SGD generally converges to different regions of the parameter space, with different geometric and generalization properties, depending on optimization hyper-parameters and initialization.

Recently, several works [1,2,15] have investigated how SGD impacts the generalization of DNNs. It has been argued that wide minima tend to generalize better than sharp ones [6,15]. One paper [15] empirically showed that a larger batch size correlates with sharper minima and worse generalization performance.

In this paper we find that the critical control parameter for SGD is not the batch size alone, but the ratio of the learning rate (LR) to batch size (BS), i.e. LR/BS. SGD performs similarly for different batch sizes but a constant

[1] First two authors contributed equally
[2] Jagiellonian University, staszek.jastrzebski@gmail.com
[3] MILA, Université de Montréal
[4] Facebook AI Research
[5] Faculty of Mathematics, Ruhr-University Bochum
[6] CIFAR Senior Fellow
[7] School of Informatics, University of Edinburgh

LR/BS. On the other hand higher values for LR/BS result in convergence to wider minima, which indeed seem to result in better generalization.

Our main contributions are as follows:

- We note that any SGD processes with the same LR/BS value is a discretization of the same stochastic differential equation (SDE).
- We derive a relation between LR/BS and the width of the minimum found by SGD.
- We verify experimentally that the SGD dynamics are similar when rescaling the LR and BS by the same amount.
- We demonstrate experimentally that a larger LR/BS correlates with a wider endpoint of SGD and better generalization.

2 Theory

Let us consider a model parameterized by $\boldsymbol{\theta}$ where the components are θ_i for $i \in \{1, \ldots, q\}$. For N training examples $\boldsymbol{x}_n, n \in \{1, ..., N\}$, the loss function, $L(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} l(\boldsymbol{\theta}, \boldsymbol{x}_n)$, and the corresponding gradient $\mathbf{g}(\boldsymbol{\theta}) = \frac{\partial L}{\partial \boldsymbol{\theta}}$, are defined based on the sum over the loss values for *all* training examples.

Stochastic gradients $\mathbf{g}^{(S)}(\boldsymbol{\theta})$ arise when we consider a minibatch $\mathcal{B}$ of size $S < N$ of random indices drawn uniformly from $\{1, ..., N\}$ and form an (unbiased) estimate of the gradient based on the corresponding subset of training examples $\mathbf{g}^{(S)}(\boldsymbol{\theta}) = \frac{1}{S}\sum_{n\in\mathcal{B}} \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}, \boldsymbol{x}_n)$.

We consider SGD with learning rate η, as defined by the update rule

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \boldsymbol{g}^{(S)}(\boldsymbol{\theta}_k) \ , \tag{1}$$

where the index k enumerate the discrete update steps.

2.1 Learning rate to batch size ratio determines SGD dynamics

In this section we derive SGD as a discretization of an SDE in which the learning rate and batch size only enter in their ratio. Other SDEs which discretize to SGD have been considered in earlier work [12,11].

Stochastic Gradient Descent: We focus on SGD in the context of large datasets. Consider the loss gradient for a randomly chosen data point,

$$\boldsymbol{g}_n(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}, \boldsymbol{x}_n). \tag{2}$$

Viewed as a random variable induced by the random sampling of the data items, $\boldsymbol{g}_n(\boldsymbol{\theta})$ is an unbiased estimator of the gradient $\boldsymbol{g}(\boldsymbol{\theta})$. For typical loss functions this estimator has finite covariance which we denote by $\mathbf{C}(\boldsymbol{\theta})$.

The batch estimate $\boldsymbol{g}^{(S)}(\boldsymbol{\theta})$ is the arithmetic mean of the components $\boldsymbol{g}_n(\boldsymbol{\theta})$. By the central limit theorem, for sufficient large batch size $\boldsymbol{g}^{(S)}(\boldsymbol{\theta})$ is approximately Gaussian distributed with mean $g(\boldsymbol{\theta})$ and variance $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = (1/S)\mathbf{C}(\boldsymbol{\theta})$.

Stochastic gradient descent (1) can be written as

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \boldsymbol{g}(\boldsymbol{\theta}_k) + \eta(\boldsymbol{g}^{(\boldsymbol{S})}(\boldsymbol{\theta}_k) - \boldsymbol{g}(\boldsymbol{\theta}_k)) \ , \tag{3}$$

where we have established that $(\boldsymbol{g}^{(\boldsymbol{S})}(\boldsymbol{\theta}_k) - \boldsymbol{g}(\boldsymbol{\theta}_k))$ is an additive zero mean Gaussian random noise with variance $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = (1/S)\mathbf{C}(\boldsymbol{\theta})$. Hence we can rewrite (3) as

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \boldsymbol{g}(\boldsymbol{\theta}_k) + \frac{\eta}{\sqrt{S}}\boldsymbol{\epsilon} \ , \tag{4}$$

where $\boldsymbol{\epsilon}$ is a zero mean Gaussian random variable with covariance $\mathbf{C}(\boldsymbol{\theta})$.

Stochastic Differential Equation: Consider now a stochastic differential equation[1] of the form

$$d\boldsymbol{\theta} = -\mathbf{g}(\boldsymbol{\theta})dt + \sqrt{\frac{\eta}{S}}\mathbf{R}(\boldsymbol{\theta})d\boldsymbol{W}(t) \ , \tag{5}$$

where $\mathbf{R}(\boldsymbol{\theta})\mathbf{R}(\boldsymbol{\theta})^T = \mathbf{C}(\boldsymbol{\theta})$, $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{U}(\boldsymbol{\theta})\boldsymbol{\Lambda}(\boldsymbol{\theta})^{\frac{1}{2}}$, and the eigendecomposition of $\mathbf{C}(\boldsymbol{\theta})$ is given by $\mathbf{C}(\boldsymbol{\theta}) = \mathbf{U}(\boldsymbol{\theta})\boldsymbol{\Lambda}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})^T$, with diagonal matrix $\boldsymbol{\Lambda}(\boldsymbol{\theta})$ containing the eigenvalues and orthonormal matrix $\mathbf{U}(\boldsymbol{\theta})$ containing the eigenvectors of $\mathbf{C}(\boldsymbol{\theta})$.

This SDE can be discretized using the Euler-Maruyama (EuM) method[2] with stepsize η to obtain precisely the same equation as (4).

Hence we can say that SGD implements an EuM approximation[3] to the SDE (5). As much as the discretized approximation is valid, the SGD optimization process must inherit all the properties[4] of the underlying SDE. Specifically we note that in the underlying SDE the learning rate and batch size only appear in the ratio η/S, which we also refer to as the stochastic noise. This implies that these are not independent variables in SGD. Rather it is only their ratio that affects the path properties of the optimization process. The only independent effect of the learning rate η is to control the stepsize of the EuM method approximation, affecting only the per batch speed at which the discrete process follows the dynamics of the SDE. There are, however, more batches in an epoch for smaller batch sizes, so the per data-point speed is the same.

2.2 LR/BS ratio controls trace of Hessian at a minimum.

We argue in this paper that there is a theoretical relationship between the expected loss value, the level of stochastic noise η/S in SGD, and the width of the minimum explored at this final stage of training. We derive that relationship in

[1] See [12] for a different SDE which also has a discretization equivalent to SGD.
[2] See e.g. [9].
[3] For a more formal analysis, not requiring central limit arguments, see an alternative approach [11] which also considers SGD as a discretization of an SDE. Note that the batch size is not present there.
[4] Including the paths of the dynamics, the equilibria, the shape of the learning curves.

this section. In the following section we will then continue by investigating our theoretical findings empirically.

We will define the width of a minimum in terms of the trace $Tr(\mathbf{H}(\boldsymbol{\theta}))$ of the Hessian at the minimum: the lower the $Tr(\mathbf{H}(\boldsymbol{\theta}))$, the wider the minima. For notational convenience, in the rest of this section we drop the dependence of $\mathbf{H}(\boldsymbol{\theta})$ and $\mathbf{C}(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$.

In order to derive the required relationship, we will make the following assumptions about the final phase of training:

Assumption 1 As we expect the training to have arrived in a local minima, the loss surface can be approximated by a quadratic bowl, with minimum at zero loss (reflecting the ability of networks to fully fit the training data). Given this the training can be approximated by an Ornstein-Unhlenbeck process. This is a similar assumption as made by previous papers [12,13].

Assumption 2 The covariance of the gradients and the Hessian of the loss approximation are approximately equal, i.e. we can assume $\mathbf{C} = \mathbf{H}$. A closeness of the Hessian and the covariance of the gradients in practical training of DNNs has been reportedbefore [14,21].

Based on Assumptions 1 and 2, the Hessian is positive definite, and matches the covariance $\mathbf{C}$. Hence its eigendecomposition is $\mathbf{H} = \mathbf{C} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$, with $\boldsymbol{\Lambda}$ being the diagonal matrix of positive eigenvalues, and $\mathbf{V}$ an orthonormal matrix. We can reparameterize the model in terms of a new variable $\boldsymbol{z}$ defined by $\boldsymbol{z} \equiv \mathbf{V}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_*)$ where $\boldsymbol{\theta}_*$ are the parameters at the minimum.

Starting from the SDE (5), and making the quadratic approximation of the loss $l(\boldsymbol{\theta}) \approx (\boldsymbol{\theta} - \boldsymbol{\theta}_*)^T\mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_*)$ results in an Ornstein-Unhlenbeck (OU) process for $\boldsymbol{z}$, that is

$$d\boldsymbol{z} = -\boldsymbol{\Lambda}\boldsymbol{z}dt + \sqrt{\frac{\eta}{S}}\boldsymbol{\Lambda}^{1/2}d\mathbf{W}(t) \ . \tag{6}$$

It is a standard result that the stationary distribution of an OU process of the form of (6) is a Gaussian with zero mean and covariance $\mathrm{cov}(\boldsymbol{z}) = \mathbb{E}(\boldsymbol{z}\boldsymbol{z}^T) = \frac{\eta}{2S}\mathbf{I}$. Moreover, in terms of the new parameters $\boldsymbol{z}$, the expected loss can be written as

$$\mathbb{E}(l) = \frac{1}{2}\sum_{i=1}^{q}\lambda_i\mathbb{E}(z_i^2) = \frac{\eta}{4S}\mathrm{Tr}(\boldsymbol{\Lambda}) = \frac{\eta}{4S}\mathrm{Tr}(\mathbf{H}) \tag{7}$$

where the second equality follows from the expression for the OU covariance.

We see from Eq. (7) that the learning rate to batch size ratio controls the trade-off between width and expected loss associated with the SGD dynamics at a minimum centred at a point of zero loss, more specifically, it follows $\frac{\mathbb{E}(l)}{Tr(\mathbf{H})} \propto \frac{\eta}{S}$. If two runs of SGD, with different LR/BS ratios have the same expected final loss $\mathbb{E}(l)$, the SGD process with the higher η/S ratio, must be associated to a smaller $\mathrm{Tr}(\mathbf{H})$ and hence must have found a different minimum, and indeed one that is wider. In the experiments which are described in the following section we compare geometrical properties of minima with the same loss value (but different generalization properties) to empirically analyze this relationship between $Tr(\mathbf{H})$ and $\frac{\eta}{S}$.

3 Experiments

We now present an empirical analysis motivated by the theory discussed in the previous section.

Learning dynamics of SGD depend on LR/BS. In this section we experimentally investigate the approximation of SGD as an discretization of the SDE given in Eq. (5), analysing how the learning dynamics are affected by the learning rate to batch size ratio.

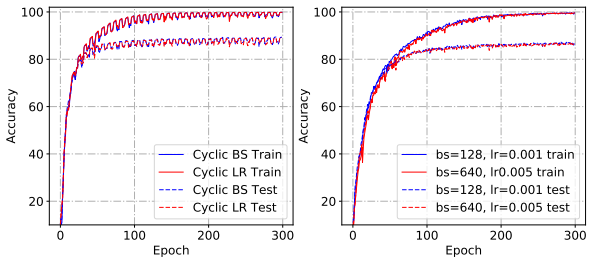


Fig. 1: SGD training of VGG11 on CIFAR10. Left: cyclic learning rate or batch size schedules. Right: constant η and S. The approximate match between red and blue curves implies that dynamics are set by the ratio of learning rate to batch size.

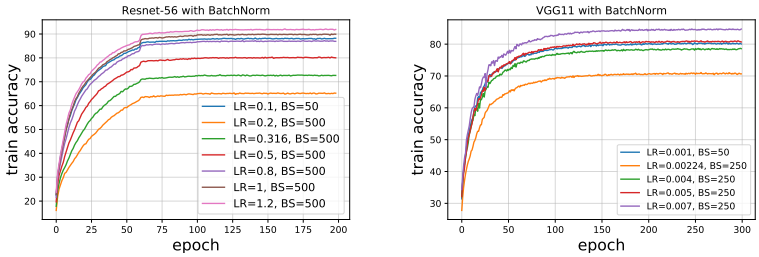


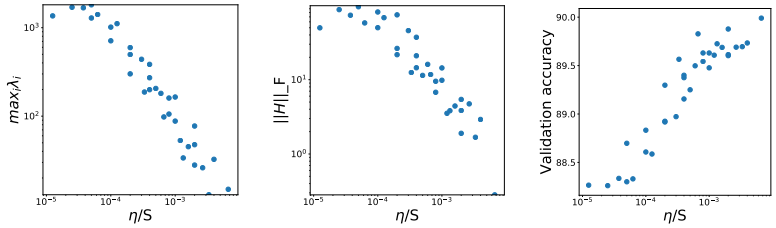
Fig. 2: Training (ResNet (left) and VGG11 (right) on CIFAR10) with a small batch size in comparison to training with a larger batch size in combination with different learning rates. Rescaling learning rate exactly by the same amount as the batch size (left, brown; right red) gives closest match to small-batch curves.

We first look at the results of four experiments in which we trained the VGG11 architecture[5] [16] on the CIFAR10 dataset, shown in Fig 1[6]. The left plot compares two experimental settings: a cyclic batch size (CBS) schedule (blue) oscillating between 128 and 640 while using a fixed learning rate of $\eta = 0.005$, compared to a cyclic learning rate (CLR) schedule (red) oscillating between 0.001 and 0.005 with a fixed batch size of $S = 128$. The right plot compares the results

[5] We have adapted the final layers to be compatible with the CIFAR10 dataset.
[6] Each experiment was repeated for 5 different random initializations.

for two other experimental settings: constant learning rate to batch size ratios of $\frac{\eta}{S} = \frac{0.001}{128}$ (blue) versus $\frac{\eta}{S} = \frac{0.005}{640}$ (red). We emphasize the similarity of the curves for each pair of experiments, demonstrating that the learning dynamics are approximately invariant under changes in learning rate or batch size that keep the ratio η/S constant.

We next ran experiments in which we rescaled the learning rate with different values when going from a small batch size to a large one, to yield a comparison to rescaling the learning rate with exactly the same value as the batch size. In Fig. 2 we show the results from two experiments on ResNet56 and VGG11, both trained with SGD and batch normalization on CIFAR10. In both settings the blue line corresponds to training with a small batch size of 50 and a small starting learning rate[7]. The other lines correspond to models trained with different learning rates and a larger batch size. It becomes visible that when rescaling η by the same amount as S (brown curve for ResNet, red for VGG11) the learning curve matches fairly closely the blue curve. Other rescaling strategies such as keeping the ratio $\eta/\sqrt{S}$ constant, as suggested by [7], (green curve for ResNet, orange for VGG) lead to larger differences in the learning curves.



(a) Largest eigenvalue. (b) Frobenius norm. (c) Validation accuracy.

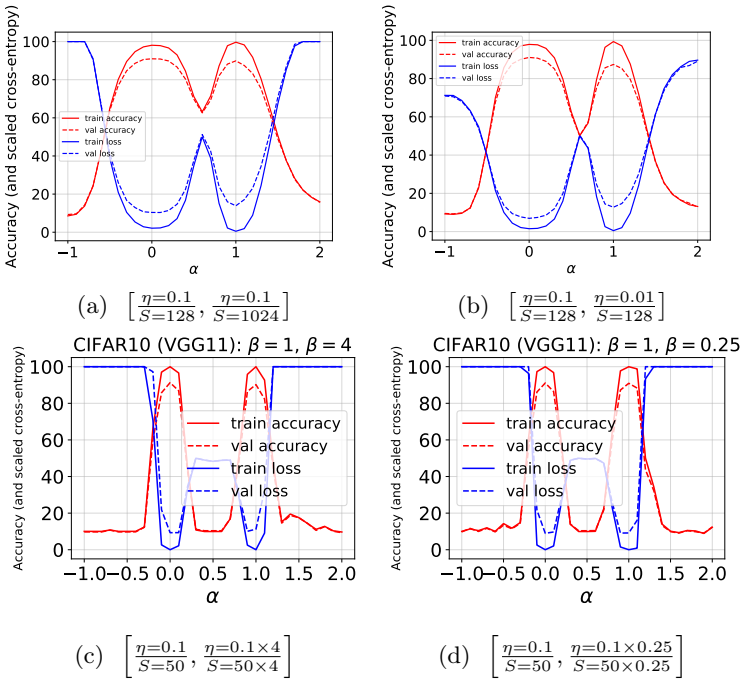
Fig. 3: Ratio of learning rate to batch size, η/S, for a grid of η and S for a 4 layer ReLU MLP on FashionMNIST. Higher η/S correlates with lower Hessian maximum eigenvalue and lower Hessian Frobenius norm, i.e. wider minima, and better generalization. The validation accuracy is consistent for different batch sizes, and different learning rates, as long as the ratio is constant.

Geometry and generalization depend on LR/BS. In this section we investigate experimentally the impact of learning rate to batch size ratio on the geometry of the region that SGD ends in. We trained a series of 4-layer batch-normalized ReLU MLPs on Fashion-MNIST [19] with different η and S[8]. To access the loss curvature at the end of training, we computed the largest eigenvalue and we approximated the Frobenius norm of the Hessian (higher values imply a sharper minimum) using the finite difference method. Fig. 3a and

[7] We used a adaptive learning rate schedule with η dropping by a factor of 10 on epochs 60, 100, 140, 180 for ResNet56 and by a factor of 2 every 25 epochs for VGG11.

[8] Each experiment was run for 200 epochs in which most models reached an accuracy of almost 100% on the training set.

Fig. 3b show the values of these quantities for minima obtained by SGD for different $\frac{\eta}{S}$, with $\eta \in [5e-3, 1e-1]$ and $S \in [25, 1000]$. As $\frac{\eta}{S}$ grows, the norm of the Hessian at the minimum decreases, suggesting that higher values of $\frac{\eta}{S}$ push the optimization towards flatter regions. Figure 3c shows the results from exploring the impact of $\frac{\eta}{S}$ on the final validation performance, which confirms that better generalization correlates with higher values of $\frac{\eta}{S}$. Taken together, Fig. 3a, Fig. 3b, and Fig. 3c imply that as $\frac{\eta}{S}$ increases SGD finds wider regions which correlate well with better generalization[9].



(a) $\left[\frac{\eta=0.1}{S=128}, \frac{\eta=0.1}{S=1024}\right]$ (b) $\left[\frac{\eta=0.1}{S=128}, \frac{\eta=0.01}{S=128}\right]$

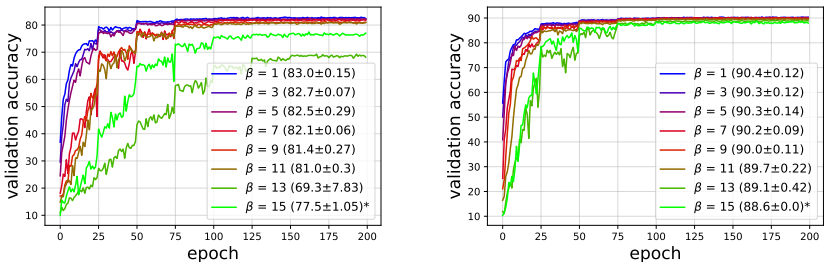
(c) $\left[\frac{\eta=0.1}{S=50}, \frac{\eta=0.1\times4}{S=50\times4}\right]$ (d) $\left[\frac{\eta=0.1}{S=50}, \frac{\eta=0.1\times0.25}{S=50\times0.25}\right]$

Fig. 4: Interpolations between models with interpolation coefficient α. $\alpha = 0$ corresponds to one trained model (1st element of sub-caption), $\alpha = 1$ to another (2nd element of subcaption). (a), (b): Resnet56 with different ratio $\frac{\eta}{S}$. (c), (d): VGG11 with the same ratio, but different η, S. Higher ratios result in wider minima (a,b) as seen by the width of the basin around $\alpha = 0$, whilst the same ratio gives the same width minima (c,d), despite differences in batch size and learning rate.

In Fig. 4 we qualitatively illustrate the behavior of SGD with different $\frac{\eta}{S}$. We follow [15] by investigating the loss on the line interpolating between the parameters of two models with interpolation coefficicent α. In Fig. 4(a,b) we consider Resnet56 models on CIFAR10 resulting from optimization with different $\frac{\eta}{S}$. We see sharper regions on the right for the lower $\frac{\eta}{S}$. In Fig. 4(c,d) we consider

[9] Assuming the network has enough capacity

VGG-11 models trained on CIFAR10 with the same ratio $\frac{\eta=0.1\times\beta}{S=50\times\beta}$ but different β. We see the same sharpness for the same ratio. Experiments were repeated several times with different random initializations and qualitatively similar plots were achieved.



(a) Train dataset size 12000 (b) Train dataset size 45000

Fig. 5: Validation accuracy for different dataset sizes and different β values for fixed ratio $\frac{\beta\times(\eta=0.1)}{\beta\times(S=50)}$. The curves diverging from the blue shows the approximation of the SDE discretized to SGD breaking down for large β, which is magnified for smaller dataset size.

Breakdown of η/S scaling. We expect discretization errors to become important when the learning rate gets large. We also expect the assumptions based on the central limit theorem to break down for a large batch size and smaller dataset size.

We show this experimentally in Fig. 5, where similar learning dynamics and final performance can be observed when simultaneously multiplying the learning rate and batch size by a factor β up to a certain limit[10]. This is done for a smaller training set size in Fig. 5 (a) than in (b). The curves don't match when β gets too large as expected from our approximations.

4 Related work

The analysis of SGD as an SDE is well established in the stochastic approximation literature, see e.g. [10]. It was shown by [11] that SGD can be approximated by an SDE in an order-one weak approximation. However, batch size does not enter their analysis. In contrast, our analysis makes the role of batch size evident and shows the dynamics are set by the ratio of learning rate to batch size. The work of [8] reproduce the SDE result of [11] and further show that the covari-

[10] Experiments are repeated 5 times with different random seeds. The graphs denote the mean validation accuracies and the numbers in the brackets denote the mean and standard deviation of the maximum validation accuracy across different runs. The * denotes at least one seed diverged.

ance matrix of the minibatch-gradient scales inversely with the batch size[11] and proportionally to the sample covariance matrix over all examples in the training set. The authors of [12] approximate SGD by a different SDE and show that SGD can be used as an approximate Bayesian posterior inference algorithm. In contrast, we show the ratio of learning rate over batch influences the width of the minima found by SGD. We then explore each of these experimentally linking also to generalization.

Many works have used stochastic gradients to sample from a posterior, see e.g. [18], using a decreasing learning rate to correctly sample from the actual posterior. In contrast, we consider SGD with a fixed learning rate and our focus is not on applying SGD to sample from the actual posterior.

Our work is closely related to the ongoing discussion about how batch size affects sharpness and generalization. Our work extends this by investigating the impact of both batch size and learning rate on sharpness and generalization. In [15] it's shown empirically that SGD ends up in a sharp minimum when using a large batch size. In [7] the learning rate is rescaled with the square root of the batch size, and more epochs are trained for to reach the same generalization with a large batch size. The empirical analysis of [5] demonstrated that rescaling the learning rate linearly with batch size can result in same generalization. Our work theoretically explains this empirical finding, and extends the experimental results on this.

Anisotropic noise in SGD was studied in [21]. It was found that the gradient covariance matrix is approximately the same as the Hessian, late on in training. In the work of [14], the Hessian is also related to the gradient covariance matrix, and both are found to be highly anisotropic. In contrast, our focus is on the importance of the scale of the noise, set by the learning rate to batch size ratio.

Concurrent with this work, [17] derive an analytical expression for the stochastic noise scale and – based on the trade-off between depth and width in the Bayesian evidence – find an optimal noise scale for optimizing the test accuracy. The work of [4] explores the stationary non-equilibrium solution for the SDE for non-isotropic gradient noise.

5 Conclusion

By approximating SGD as an SDE, we found that the learning rate to batch size ratio controls the optimization dynamics. Furthermore, this ratio is a key determinant of the width of the minima found by SGD. We experimentally investigated our theoretical findings using a range of DNN models and datasets. The results show an approximate invariance of the learning curves under rescaling of learning rate and batch size by the same amount. Moreover they confirm that the ratio of learning rate to batch size correlates with the width of the final minima and the generalization of the associated model, with a higher ratio leading to wider minima and better generalization.

[11] This holds approximately, in the limit of small batch size compared to training set size.

Acknowledgements We thank NSERC, Canada Research Chairs, IVADO and CIFAR for funding. SJ was in part supported by Grant No. DI 2014/016644 and ETIUDA stipend No. 2017/24/T/ST6/00487. This project has received funding from the European Union's Horizon 2020 programme under grant agreement No 732204 and Swiss State Secretariat for Education,Research and Innovation under contract No. 16.0159.

References

1. Advani, M.S., Saxe, A.M.: High-dimensional dynamics of generalization error in neural networks. arXiv preprint arXiv:1710.03667 (2017)
2. Arpit, D., et al.: A closer look at memorization in deep networks. In: ICML (2017)
3. Bottou, L.: Online learning and stochastic approximations. On-line learning in neural networks 17(9), 142 (1998)
4. Chaudhari, P., Soatto, S.: Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. arXiv:1710.11029 (2017)
5. Goyal, P., et al.: Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. ArXiv e-prints (2017)
6. Hochreiter, S., Schmidhuber, J.: Flat minima. Neural Computation 9(1), 1–42 (1997)
7. Hoffer, E., et al.: Train longer, generalize better: closing the generalization gap in large batch training of neural networks. ArXiv e-prints, arxiv:1705.08741
8. Junchi Li, C., et al.: Batch Size Matters: A Diffusion Approximation Framework on Nonconvex Stochastic Gradient Descent. ArXiv e-prints (2017)
9. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer (1992)
10. Kushner, H., Yin, G.: Stochastic Approximation and Recursive Algorithms and Applications. Stochastic Modelling and Applied Probability
11. Li, Q., Tai, C., E., W.: Stochastic modified equations and adaptive stochastic gradient algorithms. In: Proceedings of the 34th ICML (2017)
12. Mandt, S., Hoffman, M.D., Blei, D.M.: Stochastic gradient descent as approximate Bayesian inference. Journal of Machine Learning Research 18, 134:1–134:35 (2017)
13. Poggio, T., et al.: Theory of Deep Learning III: explaining the non-overfitting puzzle. ArXiv e-prints, arxive 1801.00173 (2018)
14. Sagun, L., Evci, U., Ugur Guney, V., Dauphin, Y., Bottou, L.: Empirical Analysis of the Hessian of Over-Parametrized Neural Networks. ArXiv e-prints (2017)
15. Shirish Keskar, N., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. ArXiv e-prints (2016)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. Smith, S., Le, Q.: Understanding generalization and stochastic gradient descent. arXiv preprint arXiv:1710.06451 (2017)
18. Welling, M., Teh, Y.W.: Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the 28th ICML. pp. 681–688 (2011)
19. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. ArXiv e-prints (2017)
20. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530 (2016)
21. Zhu, Z., Wu, J., Yu, B., Wu, L., Ma, J.: The Regularization Effects of Anisotropic Noise in Stochastic Gradient Descent. ArXiv e-prints (2018)