
Multirate Training of Neural Networks

Tiffany Vlaar¹ Benedict Leimkuhler¹

Abstract

We propose multirate training of neural networks: partitioning neural network parameters into “fast” and “slow” parts which are trained on different time scales, where slow parts are updated less frequently. By choosing appropriate partitionings we can obtain substantial computational speed-up for transfer learning tasks. We show for applications in vision and NLP that we can fine-tune deep neural networks in almost half the time, without reducing the generalization performance of the resulting models. We analyze the convergence properties of our multirate scheme and draw a comparison with vanilla SGD. We also discuss splitting choices for the neural network parameters which could enhance generalization performance when neural networks are trained from scratch. A multirate approach can be used to learn different features present in the data and as a form of regularization. Our paper unlocks the potential of using multirate techniques for neural network training and provides several starting points for future work in this area.

1. Introduction

Multirate techniques have been widely used for efficient simulation of multiscale ordinary differential equations (ODEs) and partial differential equations (PDEs) (Rice, 1960; Gear, 1974; Gear & Wells, 1984; Günther & Rentrop, 1993; Engstler & Lubich, 1997; Constantinescu & Sandu, 2013). Motivations for using multirate techniques are the presence of fast and slow time scales in the system dynamics and to simulate systems which are computationally infeasible to evolve with a single stepsize.

In their most general formulation the multirate methods we consider in this work involve separating the model param-

¹Department of Mathematics, University of Edinburgh, Edinburgh, United Kingdom. Correspondence to: Tiffany Vlaar <Tiffany.Vlaar@ed.ac.uk>.

eters Θ into multiple components $\Theta_1, \dots, \Theta_N$ corresponding to different time scales. Slow parameters are updated less frequently than their fast counterparts but with larger step-sizes. Synchronization of the parts occurs every slow time step. This is illustrated for two time scales (and accompanying fast Θ_F and slow Θ_S parameters) in Figure 1.

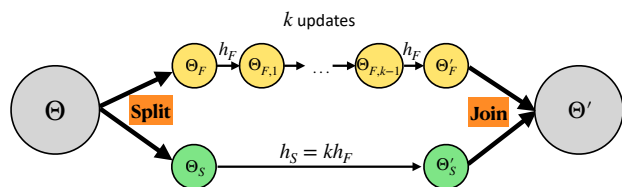


Figure 1. The basic principle of the multirate techniques considered in this paper is illustrated for two time scales in this figure. We first split our model parameters Θ into fast and slow components, Θ_F and Θ_S , respectively. The fast components are then updated every step with stepsize h_F , whereas the slow components are updated every k steps with stepsize $h_S = k \cdot h_F$.

The idea of using fast and slow weights in a machine learning context has been around for a long time (Feldman, 1982; Hinton & Plaut, 1987; Ba et al., 2016), originally inspired by neuroscience as synapses in the brain have dynamics at different time scales. However, the use of multirate methods has so far been largely overlooked for this area. In this work we seek to change this. We propose a novel multirate training scheme and show its use in various neural network training settings. We describe connections with the current machine learning literature in Section 6.

To demonstrate how multirate methods may be applicable in deep learning applications, consider a WideResNet-16 architecture trained on the patch-augmented CIFAR-10 dataset (Li et al., 2019) using SGD with momentum and weight decay and different learning rates (Figure 2). In this dataset a noisy patch of 7×7 pixels is added to the center of some CIFAR-10 images. Some images contain both the patch and CIFAR-10 data, while other images only contain the patch or are patch-free. When training using a large learning rate, the network is unable to memorize the patch, but achieves high accuracy on patch-free data. Meanwhile, when training using a small learning rate the network can memorize the patch quickly, but the accuracy on clean data is lower. We demonstrate that a multirate approach trained on two

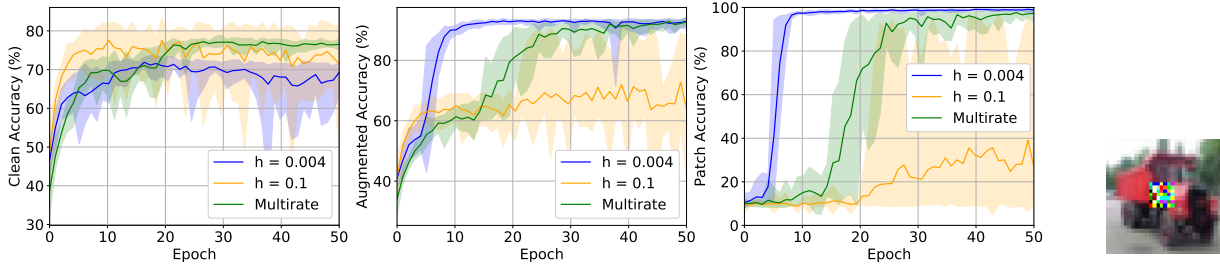


Figure 2. WideResNet-16 architecture trained on patch-augmented CIFAR-10 data (Li et al., 2019). An example of a CIFAR-10 image with a patch is given on the right. Of the training data: 20% is patch-free, 16% has only the patch, and the rest has both data and patch. More details are provided in Appendix C.1. Left: clean validation set. Middle: augmented data with patches. Right: patch-only data. A network trained using a small learning rate (blue) learns the patch quickly, whereas a large learning rate (orange) gives higher accuracy on clean data. A multirate scheme (green) trained on both time scales ($h_F = 0.004$, $h_S = 0.1$, see Section 3 and Appendix C.1) is able to memorize the patches and to simultaneously obtain high accuracy on the clean data.

time scales can both memorize the patch and obtain a high accuracy on the patch-free data. Multirate methods thus show potential for simultaneously gathering information on different features of the data, for settings where fixed learning rate approaches fail.

In this work we illustrate the benefit of using multirate techniques for a variety of neural network training applications. As main application we use a multirate approach to obtain computational speed-up for transfer learning tasks by evaluating the gradients associated with the computationally expensive (slow) part of the system less frequently (Section 4). PyTorch code supporting this work, including a ready-to-use torch.optimizer, has been made available at <https://github.com/TiffanyVlaar/MultirateTrainingOfNNs>.

The contributions of this paper are as follows:

- We propose multirate training of neural networks, which requires partitioning neural network parameters into fast and slow parts. We illustrate the versatility of this approach by demonstrating the benefits of different partitioning choices for different training applications.
- (Section 3) We describe a novel multirate scheme that uses linear drift of the slow parameters during the fast parameter update and show that the use of linear drift enhances performance. We compare its convergence properties to vanilla SGD.
- (Section 4) We use our multirate method to train deep neural networks for transfer learning applications in vision and NLP in almost half the time, without reducing the generalization performance of the resulting model.
- (Section 5) We show that a multirate approach can be used to provide some regularization when training neural networks from scratch. The technique randomly selects new subsets of the neural network to form the slow parameters using an iterative process.

We conclude that multirate methods can enhance neural network training and provide a promising direction for future theoretical and experimental work.

2. Background

Multirate methods use different stepsizes for different parts of the system. Faster parts are integrated with smaller stepsizes, while slow components are integrated using larger stepsizes, which are integer multiples of the fast stepsize. Multirate methods have been used for more than 60 years (Rice, 1960) in a wide variety of areas (Engstler & Lubich, 1997; Günther & Rentrop, 1993). Gear (1974) analyzed the accuracy and stability of Euler-based multirate methods applied to a system of ODEs with slow and fast components.

The system of ODEs that forms the starting point for most neural network training schemes is $d\theta = G(\theta)dt$, where $\theta \in \mathbb{R}^n$ are the neural network parameters and G represents the negative gradient of the loss of the entire dataset. As a starting point for our multirate approach we partition the parameters as $\theta = (\theta_F, \theta_S)$, with $\theta_F \in \mathbb{R}^{n_F}$, $\theta_S \in \mathbb{R}^{n_S}$, $n = n_F + n_S$, and obtain system of ODEs:

$$d\theta_F = G_F(\theta)dt, \quad d\theta_S = G_S(\theta)dt, \quad (1)$$

where G_F and G_S are the gradients with respect to θ_F and θ_S , respectively.

For neural network training the loss gradient is typically evaluated on a randomly selected subset of the training data and the pure gradient in Eq. (1) is subsequently replaced by a noisy gradient which we denote $\tilde{G}(\theta)$. Further, most training procedures incorporate momentum (Polyak, 1964; Sutskever et al., 2013). In the stochastic gradient Langevin dynamics method of Welling & Teh (2011), the system is further driven by constant variance additive noise. As a somewhat general model, one may consider a partitioned underdamped Langevin dynamics system of stochastic dif-

ferential equations of the form:

$$\begin{aligned} d\theta_\alpha &= p_\alpha dt, \quad \text{where } \alpha = F, S \\ dp_\alpha &= \tilde{G}_\alpha(\theta)dt - \gamma_\alpha p_\alpha dt + \sqrt{2\gamma_\alpha\tau_\alpha} dW_\alpha, \end{aligned} \quad (2)$$

with momentum $p = (p_F, p_S) \in \mathbb{R}^n$ and hyperparameters $\gamma_\alpha, \tau_\alpha > 0$. When evaluating the gradient on the full dataset, Langevin dynamics is provably ergodic (Mattingly et al., 2002), under mild assumptions, and samples from a known distribution. In this paper we will focus on the case $\tau_\alpha = 0$, which corresponds to standard stochastic gradient descent (SGD) with momentum under re-scaling of the hyperparameters, however, our multirate approach can easily be extended to the more general case. We have also opted to use the same momentum hyperparameter (γ_α in Eq. (2)) for both subsystems to provide a fair comparison with standard SGD with momentum. Using different optimizer hyperparameters, as well as exploration of methods which combine different optimizers for different components, is left for future study (see Section 7 and Appendix A). Algorithms can easily be designed based on partitioning into multiple independent components (not just two) evolving at different rates, as we illustrate in Section 3.1.

3. Multirate Training of Neural Networks

In Section 3.1 we propose a novel multirate technique that can be directly applied to the training of neural networks and discuss application-specific appropriate choices for the fast and slow parameters. In Section 3.2 we study the convergence properties of the scheme.

3.1. A Partition-based Multirate Approach

The type of multirate algorithms we consider in this work take the following approach for two time scales:

1. Separate model parameters into a fast and slow part.
2. At every step, compute the gradients with respect to the fast variables. Update the fast variables using the optimizer of your choice with fast stepsize h_F .
3. Every $k \in \mathbb{Z}_+$ steps: Compute gradients with respect to the slow variables. Update slow variables using the optimizer of your choice with slow stepsize $h_S = kh_F$.

This multirate approach can be combined with different optimization schemes, such as of the form in Eq. (2). In this work, for our analysis and numerical experiments we shall focus on using as base algorithm stochastic gradient descent (SGD), where the gradients are computed for every mini-batch of m training examples. We will compare our multirate approach with PyTorch’s standard SGD with momentum implementation (Paszke et al., 2017) and hence

for consistency we present our method in the same notation and manner as used in the PyTorch code. Our multirate scheme is described by Algorithm 1. We refer to the model parameters and momenta associated with the slow system as θ_S and p_S , respectively, and for the fast system as θ_F and p_F . We denote by $\mathcal{L}(\theta_S, \theta_F)$ the neural network loss as evaluated on a minibatch of training examples. We use the cross-entropy loss for classification tasks. We use μ to denote the momentum hyperparameter, which we typically set to $\mu = 0.9$.

We discuss variations of Algorithm 1 such as combining this multirate approach with other optimizers, the use of weight decay, or using different initializations for the fast and slow systems in Appendix A.

Linear drift. In Algorithm 1 we continuously push the slow parameters along a linear path defined by their corresponding momenta. This means that although the gradients for the slow parameters are only computed every k steps, the slow neural network parameters do get updated every step in the direction of the previous gradient. This is a novel technique for multirate training, where approaches similar to that in Algorithm 2 are more prevalent. We compare these approaches in ablation studies in Section 4.3 and show that the use of linear drift enhances performance.

Algorithm 1 Multirate SGD with linear drift

```

 $p_S := \mu p_S + \nabla_{\theta_S} \mathcal{L}(\theta_S, \theta_F)$ 
for  $i = 1, 2, \dots, k$  do
   $p_F := \mu p_F + \nabla_{\theta_F} \mathcal{L}(\theta_S, \theta_F)$ 
   $\theta_F := \theta_F - \frac{h}{k} p_F$ 
   $\theta_S := \theta_S - \frac{h}{k} p_S$ 
end for

```

Algorithm 2 Multirate SGD no linear drift

```

 $p_S := \mu p_S + \nabla_{\theta_S} \mathcal{L}(\theta_S, \theta_F)$ 
 $\theta_S := \theta_S - h p_S$ 
for  $i = 1, 2, \dots, k$  do
   $p_F := \mu p_F + \nabla_{\theta_F} \mathcal{L}(\theta_S, \theta_F)$ 
   $\theta_F := \theta_F - \frac{h}{k} p_F$ 
end for

```

Choice of Partitioning. Examples of possible separations of the model parameters into fast and slow components are layer-wise, weights vs. biases, or by selecting (random) subgroups. The appropriate separation is application-specific and will be discussed in more detail in upcoming sections. In Section 4 we explore obtaining computational speed-up using Algorithm 1 through layer-wise partitioning, where our fast parameters are chosen such that the gradients corresponding to the fast system are quick to compute, while gradients of the full net are only computed every k steps. In

Appendix D.2 we study the effect of putting the biases of a neural network on the slow time scale. Finally, in Section 5.1 we use partitioning using random subgroups to develop a regularization technique for neural network training.

Extension to more scales. Although we have presented the algorithm for two time scales, the scheme can easily be extended to more scales. To extend our framework to multiple components operating at r scales, one can use stepsizes $h_i = h_{i-1}/K_{i-1}$, $i = 1, 2, \dots, r$, $K_i \in \mathbb{Z}_+$, recursively dividing the step sequences in Algorithm 1 and 2 into finer ones at each successive level of the parameter hierarchy.

Uncoupled learning rates. In Algorithm 1 and 2 the fast and slow learning rates are coupled. Alternatively, one could introduce an uncoupled learning rate for the slow parameters. This may lead to further performance enhancements, but introduces an extra hyperparameter and thus additional tuning. We provide some ablation studies in Appendix E.

3.2. Convergence Analysis

To study the convergence properties of multirate SGD in the non-convex setting we make the following (standard) assumptions:

Assumption 3.1. We assume the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ to be L -smooth, i.e., f is continuously differentiable and its gradient is Lipschitz continuous with Lipschitz constant $L > 0$:

$$\|\nabla f(\varphi) - \nabla f(\theta)\|_2 \leq L\|\varphi - \theta\|_2, \quad \forall \theta, \varphi \in \mathbb{R}^n. \quad (3)$$

Assumption 3.2. We assume that the second moment of the stochastic gradient is bounded above, i.e., there exists a constant M for any sample x_i such that

$$\|\nabla f_{x_i}(\theta)\|_2^2 \leq M, \quad \forall \theta \in \mathbb{R}^n. \quad (4)$$

Assumption 3.2 guarantees that the variance of the stochastic gradient is bounded. Under Assumption 3.1 and 3.2 we show in Appendix B that Theorem 3.3 holds for our layer-wise partitioned multirate SGD approach:

Theorem 3.3. *We assume that 3.1 and 3.2 hold. Then*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(\theta^t)\|_2^2] \leq \frac{2(f(\theta^0) - f(\theta^*))}{hT} + hLM\ell \left(\frac{1}{3}hLk^2 + 1 \right), \quad (5)$$

where T is the number of iterations, L and M are as defined in Assumptions 3.1 and 3.2, ℓ is the number of layers, k is the additional hyperparameter associated with our multirate method, and θ^* is the optimal solution to $f(\theta)$.

From Theorem 3.3 one sees that as $T \rightarrow \infty$, the $hLM\ell \left(\frac{1}{3}hLk^2 + 1 \right)$ term controls the upper bound. The expression in Theorem 3.3 is very similar to that obtained for vanilla SGD where the rightmost term is replaced by $hLM/2$ (see Appendix B). Therefore, by decreasing the stepsize h , SGD can get closer to the neighborhood of a critical point. For our algorithm the choice of k (the additional hyperparameter introduced by our multirate method) also plays a role, where smaller values of k will lower the upper bound, but also increase the computational cost (in particular for our transfer learning application described in Section 4).

4. A Multirate Approach to Transfer Learning

We now discuss the application of our multirate scheme in the context of transfer learning, proposing a specific layer-wise division of the model parameters into fast and slow components within Algorithm 1. We will see that this can significantly reduce the computational cost of fine-tuning.

Background. The use of pre-trained deep neural networks has become a popular choice of initialization (Devlin et al., 2018; Yosinski et al., 2014). These pre-trained networks are readily available through popular machine learning libraries such as PyTorch (Paszke et al., 2017), and are usually trained on large datasets, such as ImageNet for vision applications (Huh et al., 2016) or large text corpora for natural language processing (Howard & Ruder, 2018). Using a pre-trained network as initialization has been shown to significantly accelerate training and typically improves the generalization performance of the resulting model (Yosinski et al., 2014; He et al., 2019; Radford et al., 2018). The procedure is typically as follows: start with a pre-trained model, remove task-specific layers, and then re-train (part of) the network on the new target task. Later layers of neural networks tend to capture more task-specific knowledge, while early layers encode more general features, which can be shared across tasks (Yosinski et al., 2014; Hao et al., 2019; Neyshabur et al., 2020; Raghu et al., 2019). Hence to speed up training (and, in low target-data scenarios, to prevent overfitting), one sometimes does not re-train the full neural network, but only the later layers, in particular the final fully connected layer. This process is called fine-tuning (Howard & Ruder, 2018; Dai & Le, 2015). There exists a delicate balance between computational cost and generalization performance of fine-tuned deep neural network architectures. ‘‘Fine-tuning the whole network usually results in better performance’’ (Li et al., 2020), but also increases the computational cost.

Methodology. We propose to split a neural network into two parts, the fully connected layer parameters – the fast part – and the other parameters of the deep neural network – the slow part. The fast part is updated with a stepsize h/k , while the other part (the slow part) is only updated

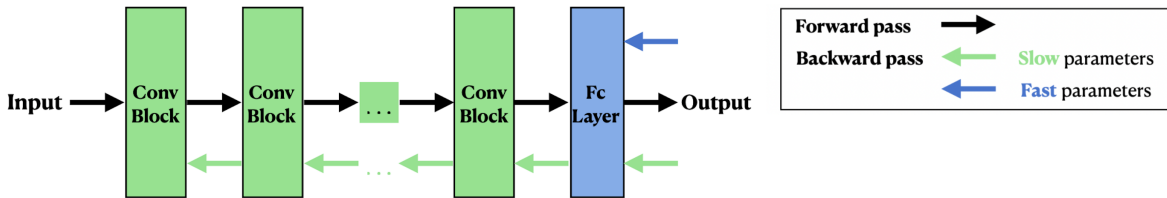


Figure 3. We indicate in blue the fast parameters and in green the slow parameters of a convolutional architecture, which consists of several convolutional blocks (conv block) and fully connected (fc) layer(s). When setting the fast parameters to be the final fc layer(s) (and optionally the conv block directly preceding it), the gradient computation for the backpropagation algorithm is very fast.

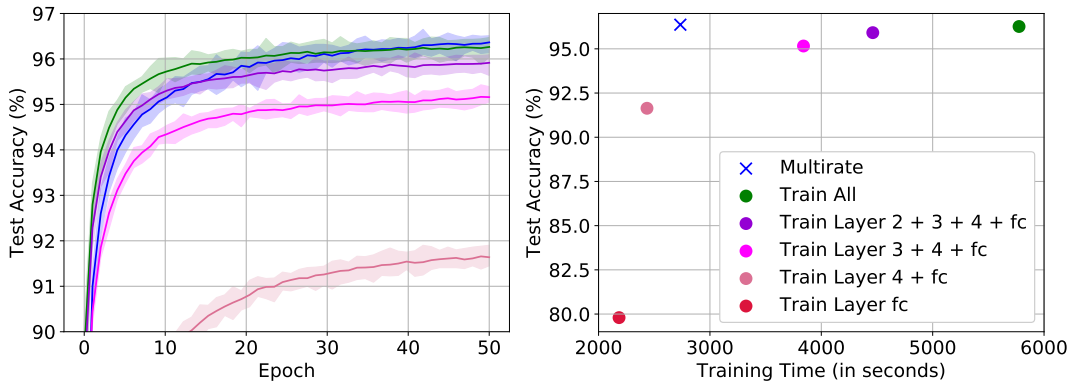


Figure 4. A pre-trained ResNet-34 being trained on CIFAR-10 data using different fine-tuning approaches and our multirate approach (blue). Results are averaged over 20 runs and all approaches are trained using SGD with momentum as base algorithm. We set $h/k = 0.001$, $k = 5$, and $\mu = 0.9$ in Algorithm 1. The highest test accuracy is reached using our multirate approach (blue), which can be used to train the net in almost half the time. Typical fine-tuning approaches only train the bottom layers of the network, e.g. just the fully connected (fc) layer (red) or layer 4 + fc, which results in a comparable speed-up, but much lower test accuracy.

every k steps with a stepsize h . The slow part is very large compared to the fast part. For example, for a ResNet-34 architecture (He et al., 2016), the fully connected layer parameters (the fast part) only constitute 0.024% of the total parameters. Because of the way the backpropagation algorithm works, for our fast parameter updates we do not need to compute gradients for the full network, because the fast part is the very last layer of the neural network. This is illustrated schematically in Figure 3. Assuming that computing the gradients constitutes the largest cost of neural network training, we obtain significant speed-up by only needing to compute the full network gradients every k steps. We show that by choosing an appropriate k we can maintain a good generalization performance for nearly half the computational cost.

4.1. Numerical Results

We study the computational speed-up and generalization performance of Algorithm 1 compared to standard fine-tuning approaches. We consider a ResNet-34 architecture (He et al., 2016), which has been pre-trained on ImageNet (Paszke et al., 2017), to classify CIFAR-10 data (Krizhevsky & Hinton, 2009). The standard procedure is to first replace the final fully connected layer of the architecture, to be

able to match the number of classes of the target dataset, and then to retrain either the full architecture on the target set or only some of the bottom layers (with layer we refer to convolutional blocks in this setting). In contrast our multirate approach only updates the final fully connected (fc) layer every step and updates the rest of the parameters every 5 steps (we have set $k = 5$ in Algorithm 1). We use as base algorithm SGD with momentum and performed a hyperparameter search to select the optimal learning rate for full network fine-tuning. We use pre-trained ResNet architectures from PyTorch (Paszke et al., 2017).

We compare our multirate approach (blue) to different fine-tuning approaches in Figure 4. Our multirate approach can be used to train the network in almost half the time, without reducing the test accuracy of the resulting net. We show in Figure 5 and Figure A10 in Appendix A that the same observations hold when training using linear learning rate decay or weight decay, respectively. In Figure 6 we repeat the experiment for a ResNet-50 architecture (pre-trained on ImageNet), which is fine-tuned on CIFAR-100 data, and observe the same behaviour.

We also test our multirate approach on natural language data and consider a pre-trained DistilBERT (obtained from Hug-

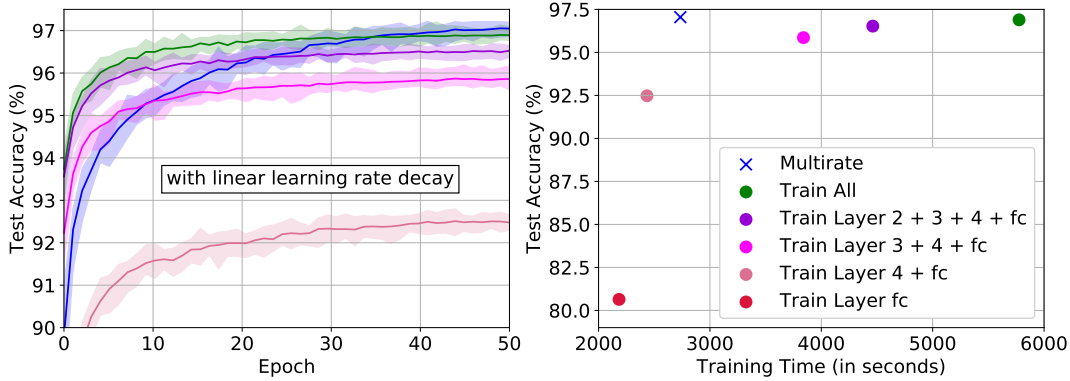


Figure 5. Fine-tuning a ResNet-34 architecture on CIFAR-10 data (same setting as in Figure 4), but using linear learning rate decay with initial learning rate set to $5e-3$. We again observe that the multirate approach (blue) can be used to train the network in about half the time, while maintaining (or even slightly improving) the test accuracy obtained when fine-tuning the full network (green).

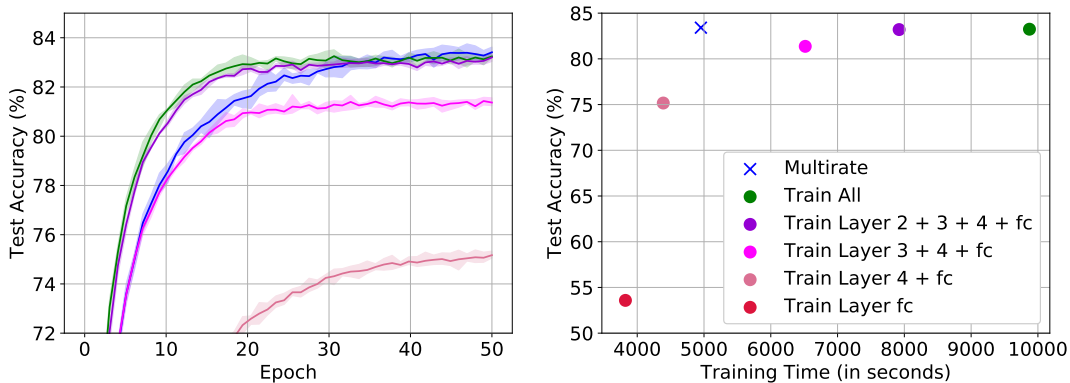


Figure 6. Same set-up as in Figure 4, but here we consider a pre-trained ResNet-50 architecture for CIFAR-100 data. The highest test accuracy is reached using our multirate approach (blue), which can be used to train the net in almost half the time.

gingFace, transformers library). We fine-tune DistilBERT on SST-2 data (Socher et al., 2013) and show the computational speed-up and maintained generalization performance obtained using our multirate approach in Figure 7. Just as for standard fine-tuning approaches, there exists a trade-off between generalization performance and training time. We find that also including the attention block directly preceding the final fully connected layer into the slow parameters further enhances the generalization performance, without significantly increasing the training time. Results for more GLUE benchmark tasks are provided in Appendix D.1.

4.2. Complexity Analysis

The number of floating point operations (FLOPs) for a forward pass through a neural network forms the lower bound of the execution time (Justus et al., 2018). The number of FLOPs required will depend on the architecture and amount of data, whereas the timing of the FLOPs depends on the hardware used (Qi et al., 2017). In our case, the number of

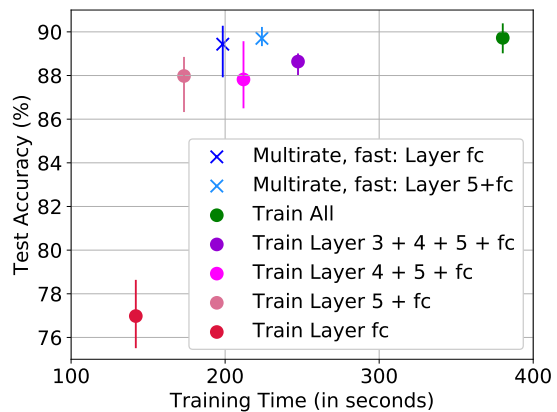


Figure 7. A pre-trained DistilBERT being trained on SST-2 data using different fine-tuning approaches, including our multirate approach (blue). Including the final attention block (together with the fc layer) in the slow parameters gives enhanced generalization performance for limited additional cost (light blue) and lowers the variance across multiple runs. We set $h/k = 1e-4$, $k = 5$, $\mu = 0.9$ in Algorithm 1, $batchsize = 16$, and average results over 10 runs.

FLOPs for the forward pass is the same for our multirate algorithm and standard approaches. The speed-up we obtain on ResNet architectures and DistilBERT arises from only needing to compute the gradients for the full net every k steps, while computing the gradients for the final fc layer(s) (and optionally the final convolutional/attention block) of the network every backward pass. The backward pass for the multirate method is hence a subset of the backward pass through the full net. Our multirate approach does require storing previous gradients of the slow parameters across iterations, which affects the amount of available memory.

Consider a neural network with L layers and our multirate scheme, where the fast parameters are set to be the final ℓ layers of the network with $\ell \ll L$. To get a relative idea of the speed-up obtained using the multirate approach, consider the ratio of the standard forward plus backward pass cost compared to the forward plus backward pass cost for our multirate approach over k steps:

$$\frac{\text{forward + backward pass} \quad \text{full net}}{\text{forward + backward pass} \quad \text{multirate}} \quad (6)$$

$$= \frac{kL + kL}{kL + L + (k - 1)\ell} = \frac{2kL}{(k + 1)L + (k - 1)\ell}, \quad \ell \ll L.$$

For comparison, when only fine-tuning the last ℓ layers of the network the cost is $k(L + \ell)$, but depending on the choice of ℓ this typically results in a lowered generalization performance. If the number of layers L is large, one can obtain a large speed-up using the multirate approach by only having to backpropagate through the full network every k steps, while maintaining a similar generalization performance as when fine-tuning the whole network. The exact speed-up obtained depends on the size of the layers and the hardware used. We performed our experiments in PyTorch on NVIDIA DGX-1 GPUs.

4.3. Ablation Studies

We consider a pre-trained DistilBERT being fine-tuned on SST-2 using our multirate approach (same set-up as in Figure 7) and perform ablation studies. In Table 1 we show that pushing the slow parameters along a linear path (as in Algorithm 1) improves the test accuracy compared to Algorithm 2. We also find that using the same stepsize $h_S = h_F$ for both the fast and slow parameters, but still only updating the slow parameters every k steps, does not lead to the same performance as using larger stepsize $h_S = k \cdot h_F$ for the slow parameters (Table 2). Finally, we provide a study on the role of k (Table A4 in Appendix E). Every epoch the slow parameters only see $1/k$ -th of the minibatches and are updated less frequently with a timestep k times larger than for the fast parameter update. We find optimal performance with $k = 5$, although training time can be decreased by choosing larger values of k . This trade-off needs to be taken into account when choosing k .

Table 1. Effect of continuously pushing slow parameters along a linear path. Same setting as in Figure 7, where fast parameters θ_F are set to be the fully connected (fc) layer + optionally the final attention block (denoted as layer 5) of a DistilBERT. Results are presented over 10 runs. We compare Algorithm 1 (uses linear drift) to Algorithm 2. We find that pushing the slow parameters along a linear path during the fast parameter update improves the mean test accuracy.

θ_F are Layer	Linear path?	Test accuracy		
		Mean	Min	Max
fc	Yes	89.43%	87.92%	90.28%
	No	88.69%	87.53%	89.68%
5 + fc	Yes	89.70%	89.35%	90.23%
	No	89.54%	88.91%	90.44%

Table 2. Same learning rate for fast and slow parameters. Same setting as in Figure 7 for a DistilBERT. We study the effect of using the same learning rate for both the fast θ_F and slow θ_S parameters, but still only updating the slow parameters every k steps. We compare $h_S = h_F = 1e-4$ vs. using $h_S = k \cdot h_F = 5e-4$. Results are presented over 10 runs. We observe that using a larger learning rate for the slow parameters aids performance.

θ_F are Layer	Higher h for θ_S ?	Test accuracy		
		Mean	Min	Max
fc	Yes	89.43%	87.92%	90.28%
	No	88.78%	87.64%	89.90%
5 + fc	Yes	89.70%	89.35%	90.23%
	No	89.29%	88.08%	89.95%

5. Multirate Training From Scratch

Whereas the previous section focused on using a multirate approach to obtain computational speed-up in transfer learning settings, we will now discuss how multirate training can be used to enhance the generalization performance of neural networks trained from scratch. We already illustrated this for the patch-augmented CIFAR-10 data set in Figure 2, where a two-scale multirate approach can both memorize the patch and simultaneously obtain good performance on clean data, whereas fixed learning rate approaches fail to do both. In this section we will show the potential of using a multirate approach to regularize neural networks.

5.1. A Multirate Approach for Neural Network Regularization

Instead of using layer-wise partitioning, in this section we use randomly selected subsets of the neural network weight matrices (and optionally the biases) to form the slow parameters in Algorithm 1. Every k optimization steps a different subset of the network parameters is randomly selected. For the technique presented here we slightly modify Algorithm 1, by setting all the slow parameters θ_S to be zero during

the k fast weight updates. After the fast parameter update, the slow parameters resume their previous value. They are then updated together with the fast weights in a single step, but using a larger time-step $h \cdot k$ for the slow weights.

The base algorithm we use is again SGD. In Figure 8 we show that our multirate technique can be used to obtain enhanced performance on a single hidden layer perceptron applied to MNIST data. Our technique is inspired by Dropout (Srivastava et al., 2014), although there are some important differences: we do not modify the network architecture and keep the ‘slow’ weights deactivated for multiple steps after which we update them with a larger time-step. We show in Figure 8 the importance of the multirate aspect (blue), i.e. removing the multirate component from our approach results in worse performance (orange). Further, we compare our technique with dropout in Figure 9 for a small transformer trained on the Penn Treebank dataset (Marcus et al., 1993) and obtain enhanced validation loss. Ablation studies for k and uncoupled learning rates are provided in Table A5 and Table A6 in Appendix E.

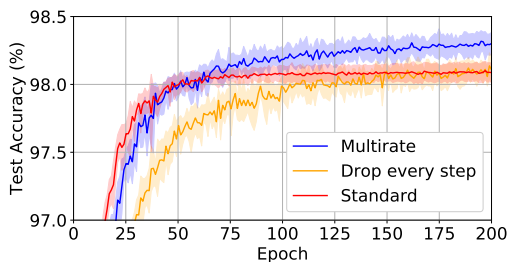


Figure 8. Single hidden layer perceptron trained on MNIST using SGD with $h = 0.1$. Our multirate technique with $k = 5$ (blue) deactivates weights in the input and hidden layer with a probability of 0.8 and 0.5, respectively, and obtains a higher test accuracy than standard SGD (red). We also test removing the multirate component from our approach, which results in an algorithm which sets a different part of the weights to zero every step (orange), and does not perform as well as the multirate technique (blue).

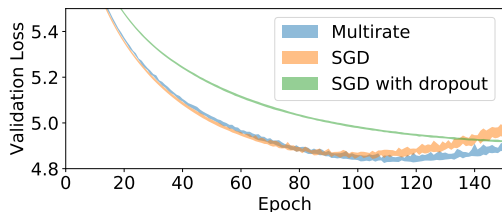


Figure 9. A transformer trained on Penn Treebank data (Marcus et al., 1993) using SGD with $h = 0.1$ and batchsize 128. Results are provided over 10 runs. The transformer has 2 encoder layers, where each encoder layer consists of self-attention with 2 heads and a feedforward network with 200 nodes followed by layer norms. We compare our multirate technique with $k = 5$ (blue) with vanilla SGD (orange) and SGD with appropriately tuned dropout (green) for the encoder layers. Our multirate approach obtains lower validation loss.

6. Related Work

Intuition for using fast and slow weights in a machine learning context can be found in neuroscience, as synapses operate at different time scales. One of the earliest mentions of fast and slow weights in the machine learning literature was by Hinton & Plaut (1987), who set each connection to have both a rapidly changing weight (which was supposed to act as a temporary memory) and a slowly changing weight which stores long-term knowledge. More recently, Ba et al. (2016) used fast weights as a temporary memory to improve recurrent neural networks.

Our multirate approach to transfer learning (Section 4) has similarities to multiple time-stepping techniques used in molecular dynamics, such as r-RESPA (Tuckerman et al., 1991; 1992), where the fast dynamics is typically cheap to compute in comparison with the slow dynamics. This is similar to our transfer learning application, where we set the final layer of the net to be the fast part to obtain computational speed-up. Although the use of more refined transfer learning schemes may lead to further test accuracy enhancement, the focus of our approach is to obtain significant computational speed-up, while maintaining the same test accuracy. Further, we introduce linear drift of the slow parameters during the fast parameter update. The use of linear drift draws inspiration from the reversible averaging approach to multiple time-stepping by Leimkuhler & Reich (2001), but forms a novel technique for multirate methods.

Further inspiration arises from the use of partitioned integrators for neural network training. It is well-known that different layers play different roles (Zhang et al., 2019) and that later layers capture more task-specific knowledge, while early layers capture more general features, which can be shared across tasks (Yosinski et al., 2014; Hao et al., 2019; Neyshabur et al., 2020; Raghu et al., 2019). It is hence natural to train the different layers of the neural network using layer-wise adaptive learning rates (You et al., 2017), layer-wise large-batch optimization techniques (You et al., 2020), using different optimizers (Leimkuhler et al., 2019), or by only being Bayesian for certain layers (Kristiadi et al., 2020; Murfet et al., 2020).

The multirate regularization technique (Section 5.1) has similarities to Dropout (Srivastava et al., 2014) and DropConnect (Wan et al., 2013). Dropout can enhance the robustness and generalization performance of neural networks, and is used widely, although its performance in combination with batch normalization (Ioffe & Szegedy, 2015) is an ongoing area of research (Luo et al., 2019; Li et al., 2018; Chen et al., 2019). In contrast to dropout and its variants we do not modify the network architecture but incorporate our technique inside the optimizer, which randomly selects a subset of the weights as the slow part and keeps these de-activated for multiple steps. These weights are then re-activated and

updated with a larger time-step, before de-activating a different subset. Instead of making strong claims, in this work we merely aim to illustrate the potential of using multirate techniques as a manner of regularization. We see an exploration of multirate variants of dropout as an exciting avenue for future work.

7. Discussion and Future Work

We outline possible directions for future work using multirate methods for neural network training along two axes: 1) different splitting choices of the neural network parameters into fast and slow parts and 2) using different optimizers or optimizer hyperparameter settings to train the different partitions. Our methods can be further generalized by combining them with well-known machine learning techniques, such as dropout or by exploring their behaviour under learning rate scheduling.

Splitting choices. For our multirate training approach we need to separate the neural network parameters into fast and slow parts. We illustrated the potential of this approach for different parameters splittings. In Section 5.1 we used random subgroups, where we randomly selected a different subset of the network parameters to be the slow parameters every k optimization steps. In Section 4 we used a layer-wise partitioning, where we set the final layer(s) to be the fast parameters and the remaining parameters to be the slow parameters, for transfer learning applications.

An interesting direction for future work is to further explore layer-wise splitting when training networks from scratch, e.g., one could separate the early from later layers and train these with different time scales. It is important to note that the computational speed-up we obtained for the transfer learning setting by only computing the gradients for the final layer(s) at every step (Section 4), does not easily transfer to training from scratch, where the same approach significantly reduces generalization performance (earlier layers need to be updated more frequently to train well). Although for different choices of the layer-wise splitting the computational speed-up is lost, the use of layer-wise partitioned multirate algorithms may still enhance generalization performance compared to vanilla optimizers. You et al. (2017) found that layer-wise adaptive learning rates can aid training. Further, network layers were shown to have different sensitivities to re-initialization (Zhang et al., 2019) and to optimizer hyperparameter settings such as the learning rate (Vlaar & Frankle, 2022). This motivates training different layers with different initializations or learning rates.

Another splitting option is to set the biases of a multi-layer perceptron architecture to be the slow parameters, while keeping the weights on the fast time scale. In Appendix D.2 we show that using this approach we can obtain higher test

accuracies on spiral data and provide ablation studies. This illustrates the potential of other parameter splittings.

Hybrid optimization schemes. For our multirate approach we partition the network into multiple parts which we train on different time scales. A natural extension is to also use different optimizers or optimizer hyperparameters to train the different partitions, e.g., using SGD for the slow part, but SGD with momentum for the fast part(s), or using sampling techniques such as SGLD or discretized underdamped Langevin dynamics for certain parts. The latter was considered for layer-wise partitionings in Leimkuhler et al. (2019) and Murfet et al. (2020). In this work we used the same base algorithm for all partitions to keep the focus on the role of different time scales. However, we expect that further performance enhancement may be achieved by using hybrid optimization schemes.

8. Conclusion

This work illustrates the potential of multirate methods for various neural network training applications. In particular, we show that a multirate approach can be used to significantly reduce the computational cost for fine-tuning neural networks, without losing test accuracy or requiring extensive hyperparameter tuning. By introducing the use of multirate techniques to the machine learning community, showing their use in different training settings, and outlining various directions for future work, we hope to have built a strong foundation for further research in this area.

Acknowledgements

We thank the reviewers for many helpful comments! During the creation of this paper Tiffany Vlaar was supported by The Maxwell Institute Graduate School in Analysis and its Applications, a Centre for Doctoral Training funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016508/01), the Scottish Funding Council, Heriot-Watt University and the University of Edinburgh.

References

- Ba, J., Hinton, G., Mnih, V., Leibo, J. Z., and Ionescu, C. Using fast weights to attend to the recent past. *NeurIPS*, 2016.
- Bello, I., Fedus, W., Du, X., Cubuk, E. D., Srinivas, A., Lin, T.-Y., Shlens, J., and Zoph, B. Revisiting ResNets: Improved training and scaling strategies. *NeurIPS*, 2021.
- Chen, G., Chen, P., Shi, Y., Hsieh, C., Liao, B., and Zhang, S. Rethinking the usage of batch normalization and dropout in the training of deep neural networks. *arXiv:1905.05928*, 2019.

- Constantinescu, E. and Sandu, A. Extrapolated multirate methods for differential equations with multiple time scales. *Journal of Scientific Computing*, 56(1):28–44, 2013.
- Dai, A. and Le, Q. Semi-supervised sequence learning. *NIPS*, 2015.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Engstler, C. and Lubich, C. Multirate extrapolation methods for differential equations with different time scales. *Computing*, 58:173–185, 1997.
- Feldman, J. A. Dynamic connections in neural networks. *Biological Cybernetics*, 46(1):27–39, 1982.
- Gear, C. and Wells, D. Multirate linear multistep methods. *BIT*, 24:484–502, 1984.
- Gear, C. W. Multirate methods for ordinary differential equations. *Technical Report, University of Illinois*, 1974.
- Günther, M. and Rentrop, P. Multirate row methods and latency of electric circuits. *Appl. Numer. Math.*, 13:83–102, 1993.
- Hao, Y., Dong, L., Wei, F., and Xu, K. Visualizing and understanding the effectiveness of BERT. *EMNLP*, 2019.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- He, K., Girshick, R., and Dollár, P. Rethinking ImageNet pre-training. *ICCV*, 2019.
- Hinton, G. E. and Plaut, D. C. Using fast weights to deblur old memories. In *Proceedings of the ninth annual conference of the Cognitive Science Society*, pp. 177–186, 1987.
- Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. *ACL*, 2018.
- Huh, M., Agrawal, P., and Efros, A. A. What makes ImageNet good for transfer learning? *arXiv:1608.08614*, 2016.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML*, 2015.
- Justus, D., Brennan, J., Bonner, S., and McGough, A. Predicting the computational cost of deep learning models. *CoRR*, 2018.
- Kristiadi, A., Hein, M., and Hennig, P. Being Bayesian, even just a bit, fixes overconfidence in ReLU networks. *ICML*, 2020.
- Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.
- Leimkuhler, B. and Reich, S. A reversible averaging integrator for multiple time- scale dynamics. *Journal of Computational Physics*, 171(1):95–114, 2001.
- Leimkuhler, B., Matthews, C., and Vlaar, T. Partitioned integrators for thermodynamic parameterization of neural networks. *Foundations of Data Science*, 1(4):457–489, 2019.
- Li, H., Chaudhari, P., Yang, H., Lam, M., Ravichandran, A., Bhotika, R., and Soatto, S. Rethinking the hyperparameters for fine-tuning. *ICLR*, 2020.
- Li, X., Chen, S., Hu, X., and Yang, J. Understanding the disharmony between dropout and batch normalization by variance shift. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
- Li, Y., Wei, C., and Ma, T. Towards explaining the regularization effect of initial large learning rate in training neural networks. *NeurIPS*, 2019.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 2019.
- Luo, P., Wang, X., Shao, W., and Pen, Z. Towards understanding regularization in batch normalization. *ICLR*, 2019.
- Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.
- Mattingly, J., A.M.Stuart, and Higham, D. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185–232, 2002.
- Murfet, D., Wei, S., Gong, M., Li, H., Gell-Redman, J., and Quella, T. Deep learning is singular, and that’s good. *arXiv:2010.11560*, 2020.
- Neyshabur, B., Sedghi, H., and Zhang, C. What is being transferred in transfer learning? *NeurIPS*, 2020.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in PyTorch. 2017.

- Polyak, B. T. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.
- Qi, H., Sparks, E. R., and Talwalkar, A. PALEO: A performance model for deep neural networks. *ICLR*, 2017.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. *OpenAI*, 2018.
- Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. Transfusion: Understanding transfer learning for medical imaging. *NeurIPS*, 2019.
- Rice, J. R. Split runge-kutta methods for simultaneous equations. *Journal of Research of the National Institute of Standards and Technology*, 60, 1960.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. *Conference on Empirical Methods in Natural Language Processing*, 2013.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15:1929–1958, 2014.
- Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. *ICML*, 2013.
- Tuckerman, M. E., Berne, B. J., and Martyna, G. J. Molecular dynamics algorithm for multiple time scales: Systems with long range forces. *The Journal of Chemical Physics*, 94(10):6811–6815, 1991.
- Tuckerman, M. E., Berne, B. J., and Martyna, G. J. Reversible multiple time scale molecular dynamics. *The Journal of Chemical Physics*, 97(3):1990–2001, 1992.
- Vlaar, T. and Frankle, J. What can linear interpolation of neural network loss landscapes tell us? *ICML*, 2022.
- Wan, L., Zeiler, M., Zhang, S., Cun, Y. L., and Fergus, R. Regularization of neural networks using DropConnect. *ICML*, 2013.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *ICLR*, 2019.
- Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. *ICML*, pp. 681–688, 2011.
- Wenzel, F., Roth, K., Veeling, B. S., Swiatkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. How good is the Bayes posterior in deep neural networks really? *arXiv:2002.02405*, 2020.
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? *NIPS*, 2014.
- You, Y., Gitman, I., and Ginsburg, B. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.
- You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C. Large batch optimization for deep learning: training BERT in 76 minutes. *ICLR*, 2020.
- Zhang, C., Bengio, S., and Singer, Y. Are all layers created equal? *arXiv:1902.01996*, 2019.
- Zhang, J., Karimireddy, S., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? *NeurIPS*, 2020.

A. Variants of our Multirate Training Algorithms

Our multirate training scheme partitions the network parameters into multiple components (Algorithm 1 and 2). This setting lends itself naturally to training the different components (or copies) using different optimization strategies.

To discuss this more concretely, recall Langevin dynamics from Eq. (2) in the main paper:

$$d\theta_\alpha = p_\alpha dt,$$

$$dp_\alpha = \tilde{G}_\alpha(\theta)dt - \gamma_\alpha p_\alpha dt + \sqrt{2\gamma_\alpha\tau_\alpha} dW_\alpha, \text{ where } \alpha = F, S,$$

with neural network parameters $\theta = (\theta_F, \theta_S) \in \mathbb{R}^n$, momentum $p = (p_F, p_S) \in \mathbb{R}^n$, noisy (due to subsampling) gradient $\tilde{G}_\alpha(\theta)$ of the loss with respect to θ_α , Wiener process W , and hyperparameters $\gamma_\alpha, \tau_\alpha > 0$. Using discretized Langevin dynamics to train neural networks allows for incorporation of both momentum and additive noise, the size of which is controlled by the γ_α and τ_α hyperparameters, respectively. A straightforward variant is thus to use different values for γ_α and/or τ_α for the fast and slow components. The temperature hyperparameter τ_α controls the driving noise and thus the transition between a pure optimization and sampling approach. When small it can benefit neural network optimization (Leimkuhler et al., 2019; Wenzel et al., 2020). Using small values of τ for parts of the dynamics that require further exploration may thus benefit training. On the other hand, to train a component with stochastic gradient descent with momentum one can set $\tau = 0$. An example of a possible combination of Langevin dynamics with additive noise for the fast dynamics and without additive noise (corresponding to SGD) for the slow dynamics is then:

$$d\theta_F = p_F dt, dp_F = \tilde{G}_F(\theta)dt - \gamma_F p_F dt + \sqrt{2\gamma_F\tau} dW_F$$

$$d\theta_S = p_S dt, dp_S = \tilde{G}_S(\theta)dt - \gamma_S p_S dt.$$

Equivalently, one could change the value of the momentum hyperparameter γ_α for the different partitionings. Or use different optimizers for the different components, such as Adam and SGD. Finally, it would be interesting to study the effect of using different size initializations for the different components, which essentially starts off the different components on different scales. Of course, any of these suggestions require extra tuning of the algorithm, which is why we focused on SGD with the same momenta values and initializations for all components in the paper. We expect however that using hybrid optimization schemes may lead to even further performance enhancement, which we aim to explore in future work.

Weight decay. As a simple extension of Algorithm 1 we provide the case with weight decay in Algorithm A3, where we have used ω to denote the amount of weight decay. Our implementation of weight decay is the same as used in PyTorch for SGD (Paszke et al., 2017). Due to the use of linear drift all parameters are trained using the same learning rate (although slow parameter gradients are only updated every k steps), which is why no extra tuning of the weight decay is needed. We show in Figure A10 for a pre-trained ResNet-34 being fine-tuned on CIFAR-10 data (same set-up as in Figure 4) that using Algorithm A3 our multirate approach can train the network in almost half the time, without reducing the test accuracy.

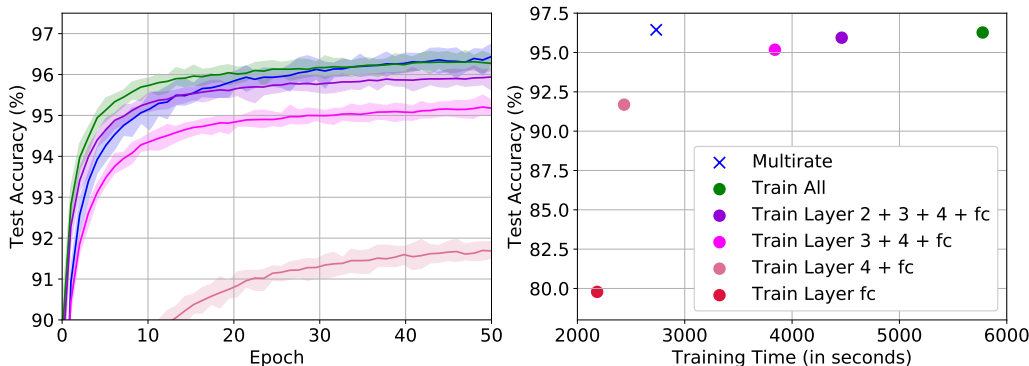


Figure A10. A pre-trained ResNet-34 being trained on CIFAR-10 data (same set-up as in Figure 4) using different fine-tuning approaches and our multirate approach Algorithm A3 (blue). Results are averaged over 20 runs and approaches are trained using SGD with momentum with weight decay. We set $h/k = 0.001, k = 5, \mu = 0.9$, and $\omega_S = \omega_F = 5e-4$ in Algorithm A3. Our multirate approach (blue) can be used to train the net in almost half the time, while maintaining generalization performance. Typical fine-tuning approaches only train the bottom layers of the network, e.g., layer 4 + fc, which results in a comparable speed-up, but much lower test accuracy.

Algorithm A3 Multirate SGD with linear drift and weight decay

```

 $p_S := \mu p_S + \nabla_{\theta_S} \mathcal{L}(\theta_S, \theta_F) + \omega_S \theta_S$ 
for  $i = 1, 2, \dots, k$  do
   $p_F := \mu p_F + \nabla_{\theta_F} \mathcal{L}(\theta_S, \theta_F) + \omega_F \theta_F$ 
   $\theta_F := \theta_F - \frac{h}{k} p_F$ 
   $\theta_S := \theta_S - \frac{h}{k} p_S$ 
end for

```

B. Convergence Analysis

Recall our main assumptions:

Assumption B.1. We assume function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ to be L -smooth, i.e., f is continuously differentiable and its gradient is Lipschitz continuous with Lipschitz constant $L > 0$:

$$\|\nabla f(\varphi) - \nabla f(\theta)\|_2 \leq L \|\varphi - \theta\|_2, \quad \forall \theta, \varphi \in \mathbb{R}^n. \quad (7)$$

Assumption B.2. We assume that the second moment of the stochastic gradient is bounded above, i.e., there exists a constant M for any sample x_i such that

$$\|\nabla f_{x_i}(\theta)\|_2^2 \leq M, \quad \forall \theta \in \mathbb{R}^n. \quad (8)$$

Assumption B.2 guarantees the variance of the stochastic gradient to be less than M , because

$$\begin{aligned} \text{Var}(\nabla f_{x_i}(\theta)) &= \mathbb{E} \|\nabla f_{x_i}(\theta) - \mathbb{E}[\nabla f_{x_i}(\theta)]\|_2^2 \\ &= \mathbb{E} \|\nabla f_{x_i}(\theta) - \nabla f(\theta)\|_2^2 \\ &= \mathbb{E} \|\nabla f_{x_i}(\theta)\|_2^2 - \|\nabla f(\theta)\|_2^2 \end{aligned} \quad (9)$$

where we used $\mathbb{E}[\nabla f_{x_i}(\theta)] = \nabla f(\theta)$ (unbiased gradient) for the second equality and $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

If we assume Assumption B.1 holds, we obtain the following Lemma, which we will need for the proof of the main theorem:

Lemma B.3. *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is L -smooth then $\forall \theta, \varphi \in \mathbb{R}^n$:*

$$|f(\varphi) - (f(\theta) + \nabla f(\theta)^T(\varphi - \theta))| \leq \frac{L}{2} \|\varphi - \theta\|_2^2. \quad (10)$$

Proof of Lemma B.3. From the fundamental theorem of calculus:

$$\int_0^1 \nabla f[\theta + t(\varphi - \theta)]^T(\varphi - \theta) dt = f(\varphi) - f(\theta)$$

So using Cauchy-Schwartz and the assumption that f is L -smooth we obtain:

$$\begin{aligned} |f(\varphi) - f(\theta) - \nabla f(\theta)^T(\varphi - \theta)| &= \left| \int_0^1 (\nabla f[\theta + t(\varphi - \theta)] - \nabla f(\theta))^T(\varphi - \theta) dt \right| \\ &\leq \int_0^1 \|\nabla f[\theta + t(\varphi - \theta)] - \nabla f(\theta)\|_2 \|\varphi - \theta\|_2 dt \\ &\leq L \|\varphi - \theta\|_2^2 \int_0^1 t dt = \frac{L}{2} \|\varphi - \theta\|_2^2. \end{aligned}$$

□

The multirate method update for base algorithm SGD is

$$\theta_\ell^{t+1} = \theta_\ell^t - h \nabla f_{\ell, x_i}(\theta^t), \quad (11)$$

where θ_ℓ^t are the parameters in layer ℓ at iteration t , h is the stepsize, and $\nabla f_{\ell, x_i}$ denotes the gradient of the loss of the i th training example for parameters in layer $l \in \{F, S\}$, where $\nabla f_{F, x_i}(\theta^t) = \nabla f_{F, x_i}(\theta^t)$ and with linear drift: for any $t \in [\tau, \tau + k - 1]$, where τ is divisible by k , $\nabla f_{S, x_i}(\theta^t) = \nabla f_{S, x_i}(\theta^\tau)$.

Now we want to prove Theorem 3.3 in the main body of the paper:

Theorem B.4. *Assume that Assumptions B.1 and B.2 hold. Then*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(\theta^t)\|_2^2] \leq \frac{2(f(\theta^0) - f(\theta^*))}{hT} + hLM\ell \left(\frac{1}{3} hLk^2 + 1 \right), \quad (12)$$

where θ^* is the optimal solution to $f(\theta)$.

Proof of Theorem B.4. Because f is L -smooth, from Lemma B.3 it follows that:

$$\begin{aligned} f(\theta^{t+1}) &\leq f(\theta^t) + \nabla f(\theta^t)^{Tr}(\theta^{t+1} - \theta^t) + \frac{L}{2} \|\theta^{t+1} - \theta^t\|_2^2 \\ &\leq f(\theta^t) - h \nabla f(\theta^t)^{Tr} \left(\sum_{\ell} \nabla f_{\ell, x_i}(\theta^t) \right) + \frac{h^2 L}{2} \left\| \sum_{\ell} \nabla f_{\ell, x_i}(\theta^t) \right\|_2^2 \end{aligned} \quad (13)$$

Taking the expectation on both sides gives (because of unbiased gradient $\mathbb{E}[\nabla f_{x_i}(\theta)] = \nabla f(\theta)$ and Assumption B.2):

$$\mathbb{E}[f(\theta^{t+1}) - f(\theta^t)] \leq -h \nabla f(\theta^t)^{Tr} \left(\sum_{\ell} \nabla f_{\ell}(\theta^t) \right) + \frac{h^2 LM\ell}{2}$$

for number of layers ℓ . So in T iterations we have θ^T such that (using a telescoping sum):

$$\begin{aligned} f(\theta^*) - f(\theta^0) &\leq \mathbb{E}[f(\theta^T)] - f(\theta^0) \\ &\leq -h \underbrace{\sum_{t=0}^{T-1} \nabla f(\theta^t)^{Tr} \left(\sum_{\ell} \nabla f_{\ell}(\theta^t) \right)}_{\mathcal{A}} + \frac{h^2 LM\ell}{2} T. \end{aligned} \quad (14)$$

For term \mathcal{A} we get

$$\mathcal{A} = \sum_{t=0}^{T-1} a_t = \sum_{t=0}^{k-1} a_t + \sum_{t=k}^{2k-1} a_t + \cdots + \sum_{t=\tau}^{\tau+k-1} a_t + \cdots + \sum_{t=T-k}^{T-1} a_t, \quad (15)$$

where $\sum_{t=\tau}^{\tau+k-1} a_t$ is given by

$$\begin{aligned} \sum_{t=\tau}^{\tau+k-1} \nabla f(\theta^t)^{Tr} \left(\sum_{\ell} \nabla f_{\ell}(\theta^t) \right) &= \sum_{t=\tau}^{\tau+k-1} \nabla f(\theta^t)^{Tr} (\nabla f_F(\theta^t) + \nabla f_S(\theta^\tau)) \\ &= \sum_{t=\tau}^{\tau+k-1} \nabla f(\theta^t)^{Tr} (\nabla f_F(\theta^t) + \nabla f_S(\theta^\tau) - \nabla f_S(\theta^t) + \nabla f_S(\theta^t)) \\ &= \sum_{t=\tau}^{\tau+k-1} \|\nabla f(\theta^t)\|_2^2 + \underbrace{\sum_{t=\tau}^{\tau+k-1} \nabla f(\theta^t)^{Tr} (\nabla f_S(\theta^\tau) - \nabla f_S(\theta^t))}_{\mathcal{B}}. \end{aligned}$$

Because $xy \leq \frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|y\|_2^2$ (combination of Cauchy-Schwarz and Young's inequality) (gives 1st inequality) and Assumption B.1 (gives 2nd inequality) we get for term \mathcal{B} :

$$\begin{aligned} \mathcal{B} &\leq \frac{1}{2} \sum_{t=\tau}^{\tau+k-1} \|\nabla f(\theta^t)\|_2^2 + \frac{1}{2} \sum_{t=\tau}^{\tau+k-1} \|\nabla f_S(\theta^\tau) - \nabla f_S(\theta^t)\|_2^2 \\ &\leq \frac{1}{2} \sum_{t=\tau}^{\tau+k-1} \|\nabla f(\theta^t)\|_2^2 + \underbrace{\frac{L^2}{2} \sum_{t=\tau+1}^{\tau+k-1} \|\theta^\tau - \theta^t\|_2^2}_{\mathcal{C}}. \end{aligned}$$

We get for term \mathcal{C} from Eq. (11) (gives 2nd equality), $\|a_1 + \dots + a_m\|_2^2 \leq m(\|a_1\|_2^2 + \dots + \|a_m\|_2^2)$ (gives 1st inequality), Assumption B.2 (gives 2nd inequality), and $k > 1$ (final inequality):

$$\begin{aligned} \mathcal{C} &= \|\theta^\tau - \theta^{\tau+1}\|_2^2 + \|\theta^\tau - \theta^{\tau+2}\|_2^2 + \dots + \|\theta^\tau - \theta^{\tau+k-1}\|_2^2 \\ &= h^2 \left(\left\| \sum_{\ell} \nabla f_{\ell, x_i}(\theta^\tau) \right\|_2^2 + \left\| \sum_{\ell} \nabla f_{\ell, x_i}(\theta^\tau) + \sum_{\ell} \nabla f_{\ell, x_i}(\theta^{\tau+1}) \right\|_2^2 \right. \\ &\quad \left. + \dots + \left\| \sum_{\ell} \nabla f_{\ell, x_i}(\theta^\tau) + \dots + \sum_{\ell} \nabla f_{\ell, x_i}(\theta^{\tau+k-2}) \right\|_2^2 \right) \\ &\leq h^2 \left(\sum_{m=1}^{k-1} m \left\| \sum_{\ell} \nabla f_{\ell, x_i}(\theta^\tau) \right\|_2^2 + \sum_{m=2}^{k-1} m \left\| \sum_{\ell} \nabla f_{\ell, x_i}(\theta^{\tau+1}) \right\|_2^2 + \dots + (k-1) \left\| \sum_{\ell} \nabla f_{\ell, x_i}(\theta^{\tau+k-2}) \right\|_2^2 \right) \\ &\leq h^2 M \ell ((k-1)^2 + (k-2)^2 + \dots + 1) = h^2 M \ell \sum_{m=1}^{k-1} m^2 = h^2 M \ell (k/6 - k^2/2 + k^3/3) \leq h^2 M \ell k^3/3. \end{aligned}$$

So overall for term $-h\mathcal{A}$ we get

$$\begin{aligned} -h \sum_{t=0}^{T-1} \nabla f(\theta^t)^{Tr} \left(\sum_{\ell} \nabla f_{\ell}(\theta^t) \right) &\leq -h \sum_{t=0}^{T-1} \|\nabla f(\theta^t)\|_2^2 + h \left| \sum_{\tau} \mathcal{B} \right| \\ &\leq -\frac{h}{2} \sum_{t=0}^{T-1} \|\nabla f(\theta^t)\|_2^2 + \frac{1}{6} h^3 L^2 M \ell k^2 T. \end{aligned} \quad (16)$$

Substituting this into Eq. (14) and again taking the expectation gives:

$$\begin{aligned} f(\theta^*) - f(\theta^0) &\leq \mathbb{E}[f(\theta^T)] - f(\theta^0) \\ &\leq -\frac{h}{2} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(\theta^t)\|_2^2] + \frac{1}{6} h^3 L^2 M \ell k^2 T + \frac{h^2 L M \ell}{2} T \\ &= -\frac{h}{2} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(\theta^t)\|_2^2] + \frac{1}{2} h^2 L M \ell T \left(\frac{1}{3} h L k^2 + 1 \right). \end{aligned} \quad (17)$$

This gives Theorem B.4

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(\theta^t)\|_2^2] \leq \frac{2(f(\theta^0) - f(\theta^*))}{hT} + h L M \ell \left(\frac{1}{3} h L k^2 + 1 \right).$$

□

For comparison, the convergence analysis for vanilla SGD with fixed stepsize h update

$$\theta^{t+1} = \theta^t - h \nabla f_{x_i}(\theta^t), \quad (18)$$

where ∇f_{x_i} denotes the gradient of the loss of the i th training example, gives Theorem B.5.

Theorem B.5. Assume that Assumptions B.1 and B.2 hold. Then:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(\theta^t)\|_2^2] \leq \frac{2(f(\theta^0) - f(\theta^*))}{hT} + \frac{hLM}{2} \quad (19)$$

where θ^* is the optimal solution to $f(\theta)$.

Proof of Theorem B.5. Because f is L -smooth, from Lemma 1 it follows that:

$$\begin{aligned} f(\theta^{t+1}) - f(\theta^t) &\leq \nabla f(\theta^t)^T r(\theta^{t+1} - \theta^t) + \frac{L}{2} \|\theta^{t+1} - \theta^t\|_2^2 \\ &\leq -h \nabla f(\theta^t)^T r \nabla f_{x_i}(\theta^t) + \frac{h^2 L}{2} \|\nabla f_{x_i}(\theta^t)\|_2^2 \end{aligned}$$

Taking the expectation on both sides gives (because of assumption 2 and unbiased gradient $\mathbb{E}[\nabla f_{x_i}(\theta)] = \nabla f(\theta)$):

$$\mathbb{E}[f(\theta^{t+1}) - f(\theta^t)] \leq -h \mathbb{E} [\|\nabla f(\theta^t)\|_2^2] + \frac{h^2 L}{2} M \quad (20)$$

So in T gradient steps we have θ^T such that:

$$f(\theta^*) - f(\theta^0) \leq \mathbb{E}[f(\theta^T)] - f(\theta^0) \leq -h \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(\theta^t)\|_2^2] + \frac{h^2 L M T}{2} \quad (21)$$

This gives:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(\theta^t)\|_2^2] \leq \frac{2(f(\theta^0) - f(\theta^*))}{hT} + \frac{hLM}{2} \quad (22)$$

□

C. Further Experimental Details

We run our experiments (unless indicated otherwise) with SGD with momentum set to 0.9. The learning rate varied per experiment and is detailed in the captions of the figures. For the transfer learning experiments (Section 4) it was set to $h = 0.001$ for the ResNet architectures and to $h = 1e-4$ for the DistilBERT and we did not use weight decay (except for Figure A10). The models were pre-trained on ImageNet, so we resized the CIFAR-10/CIFAR-100 images before training, e.g. as in Bello et al. (2021). In Algorithm 1 we set $k = 5$ and varied our partitionings of the network parameters into fast and slow parts. All our experiments were run in PyTorch using NVIDIA GPUs. We will discuss specific experiments that require further details below.

C.1. Patch-augmented CIFAR-10

The patch-augmented CIFAR-10 dataset that we used for Figure 2 was adapted from the paper by Li et al. (2019). To generate the dataset the 50000 CIFAR-10 training images are split into 10000 patch-free images and 40000 images which contain only a patch with probability 0.2 and contain a patch mixed with CIFAR-10 data with probability 0.8. The 7×7 pixel patch is located in the center of the images. Following Li et al. (2019) to generate the patch, sample $z \sim \mathcal{N}(0, 1.5625)$, a a random float in $[0, 1)$, and $\zeta_i \sim [-0.1, 0.1]$ for classes $i = 1, \dots, 10$. Then for patch-only images belonging to class i set everything to 0 and add $z \pm 1.75a\zeta_i$. To generate images containing both a patch and CIFAR-10 data add $z \pm \zeta_i$. For the multirate training approach we partitioned a composite network system into two parts, where each subnetwork was trained on a different timescale. The weights sampled from both parts were averaged and merged every k steps. The exact same learning rates were used as in the original paper by Li et al. (2019), so $h_F = 0.004$, $h_S = 0.1$, and thus $k = h_S/h_F = 25$.

D. Additional Experiments

D.1. GLUE Tasks

In Table A3 we provide results for fine-tuning a DistilBERT (same setting as in Section 4, Figure 7) on more tasks from the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019). We compare the performance and computational speed-up of multirate Algorithm 1 with full net fine-tuning. We see that a similar generalization performance is maintained using the multirate approach, while achieving computational speed-up. The focus of the experiment is solely on showing relative computational speed-up, not on beating state-of-the-art. For these experiments (and in the rest of this paper) we use as base algorithm SGD with momentum. However, adaptive optimizers such as Adam tend to be the method of choice in the natural language processing literature (Devlin et al., 2018; Liu et al., 2019; Zhang et al., 2020) and may lead to further performance enhancements. We see exploration of a multirate approach in this setting as an interesting direction for future work.

Table A3. Performance on dev sets (median over 5 runs) and averaged wallclock time per fine-tuning training run of a pre-trained DistilBERT on some GLUE tasks. For MNLI we report accuracy on matched and mismatched sets. Hyperparameter settings: batchsize = 16 (all, except 32 for MNLI), $k = 2$ and weight decay set to $5e-4$ (MNLI, QNLI), $k = 4$ (RTE, WNLI), $h/k = 3e-3$ (MNLI), $h/k = 5e-4$ (QNLI, RTE), $h/k = 1e-4$ (SST-2, WNLI) and the fast parameters being the linear head and attention block 5 (MNLI, SST-2, WNLI) + attention block 4 (QNLI, RTE).

		MNLI	QNLI	RTE	SST-2	WNLI
Accuracy (%)	Full net fine-tuning	75.3 / 76.7	86.6	57.8	89.7	54.2
	Multirate	75.4 / 76.7	86.4	57.8	89.7	54.9
Timing (sec.)	Full net fine-tuning	14617	14148	192	380	20
	Multirate	10797	10602	127	224	11

D.2. Slow Biases

We study the effect of putting all the biases of a neural network on the slow time scale, while keeping the weights on the fast time scale and only updating the slow parameters every k steps using Algorithm 1. Surprisingly, this gives big performance improvements on 4-turn spiral data (adapted from Leimkuhler et al. (2019)) as shown in Figure A11. Figure A12 confirms that this enhanced performance is caused by our multirate technique, i.e., freezing the biases for k steps and then boosting them with a larger time-step. Simply putting the biases on a different time scale or freezing the biases for k steps and using the same time-step does not lead to the same performance improvement. In Figure A13 we show that this effect is caused by the input layer biases, in particular. This seems to suggest a possible connection with data normalization (and the lack thereof for the spiral dataset). In Figure A14 we show that for a ResNet-34 on CIFAR-10 data one also obtains a small performance improvement by using slow biases for the fully connected layer, especially when no data augmentation is used.

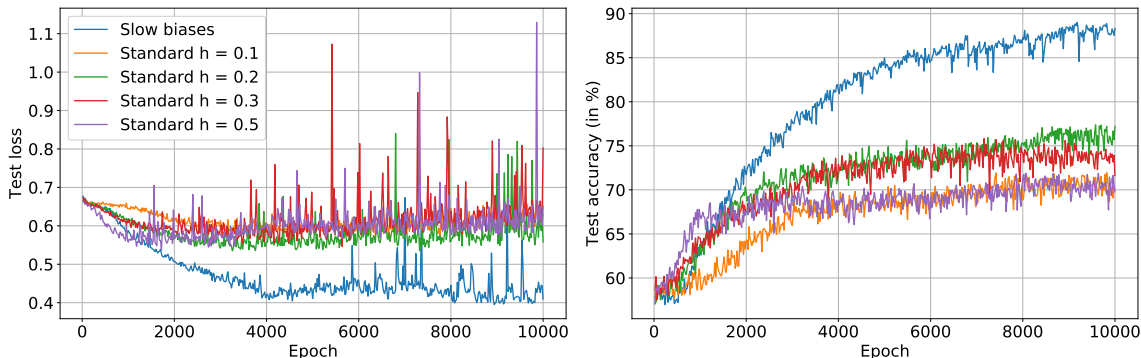


Figure A11. We use both standard SGD and our multirate approach as defined in Algorithm 1 to train a single hidden layer perceptron (SHLP) on a 4-turn spiral problem (adapted from (Leimkuhler et al., 2019)) with 5% subsampling. We set the biases of the neural network to be θ_S and the weights to be θ_F and set $k = 5$, $h = 1$ in Algorithm 1. Results are averaged over 5 runs.

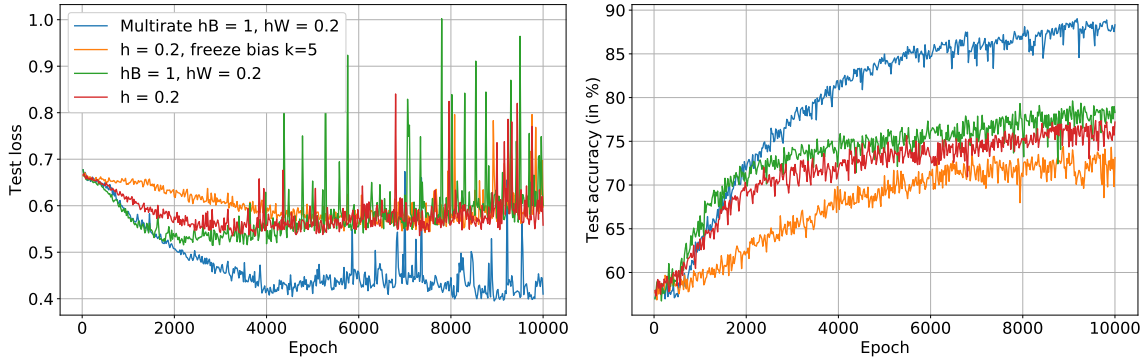


Figure A12. Same setting as in Figure A11, with a SHLP being trained on spiral data using SGD with $h = 0.2$ (red). We show that putting the biases on a different time scale (green) or freezing the biases for $k = 5$ steps and then updating them with the same stepsize as for the fast (weight) parameters (orange) both do not lead to the same performance improvement as our multirate technique (blue).

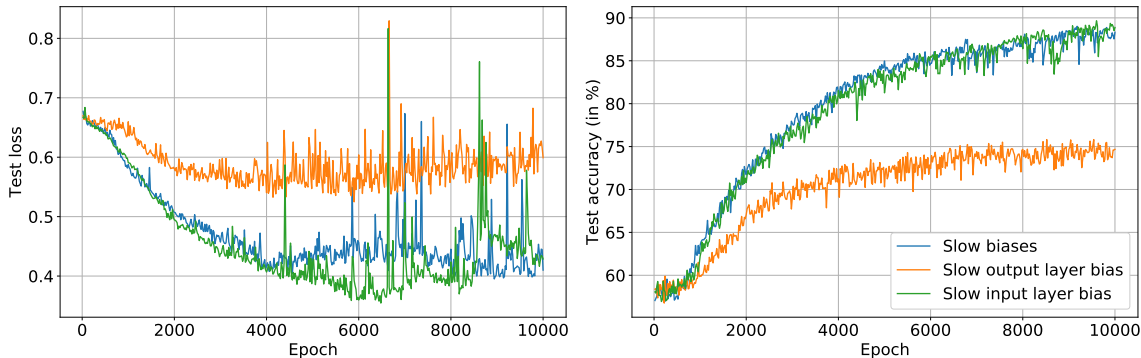


Figure A13. Same setting as in Figure A11, but here we study the effect of only putting the input biases (green) or only the output biases (orange) on the slow time scale. Clearly, using slow input biases (green) appears to be key to the enhanced generalization performance of the multirate approach (blue) in this setting.

E. Further Ablation Studies

The effect of k is studied in Table A4 for fine-tuning a pre-trained DistilBERT on SST-2 data using Algorithm 1. We find that although smaller values of k can improve the generalization performance, the training time gets increased. This trade-off needs to be taken into account when choosing k . Apart from this, recall that lower values of k also lower the stepsize used for the slow parameters, which can affect performance. Therefore, it may be beneficial to consider uncoupled learning rates, see discussion below and in Section 3.1. As a rule of thumb, we found that setting $k = 5$ often gives enough speed-up, without significantly affecting the accuracy.

We also provide ablation studies for the value of k for the multirate approach for neural network regularization used in Section 5.1 for training from scratch a small MLP on MNIST data (Table A5, left) and a transformer on the Penn Treebank dataset (Table A6, left). In this setting the aim is not computational speed-up, but enhanced generalization performance. We again find that setting $k = 5$ gives optimal performance. In Table A5 (right) and Table A6 (right) we also show that using an uncoupled h_S could lead to further performance enhancements, but this does introduce an additional hyperparameter to tune.

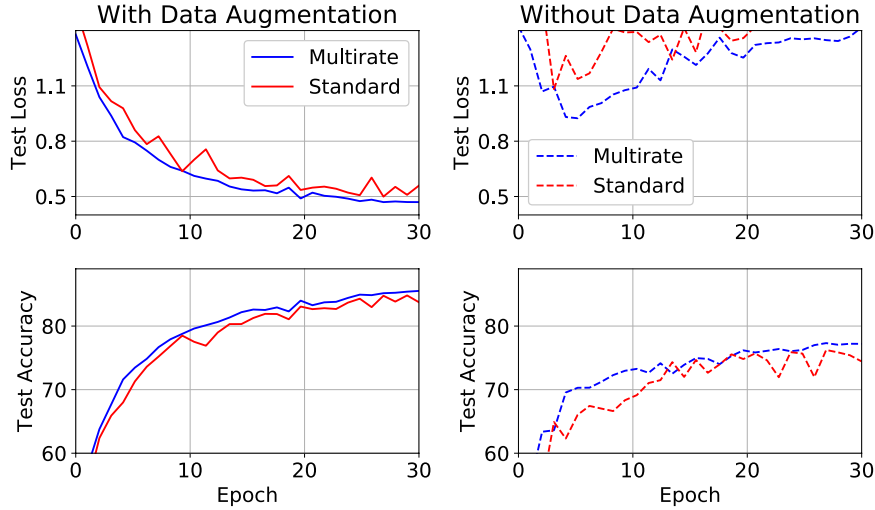


Figure A14. We use both standard SGD (with $h = 0.1$) and our multirate approach as defined in Algorithm 1 to train a ResNet-34 architecture on CIFAR-10 data. We set the biases of the fully connected layer of the neural network to be θ_S and the weights to be θ_F and set $k = 10$, $h = 1$ in Algorithm 1 and use $\text{batchsize} = 128$. Results are averaged over 10 runs.

Table A4. Effect of k . A pre-trained DistilBERT being trained on SST-2 data using our multirate approach for different values of k (same setting as in Figure 7), where the fast parameters are set to be the fully connected (fc) layer + optionally the final attention block (layer 5). We set $h/k = 1e-4$ and $\mu = 0.9$ in Algorithm 1 and use a batchsize of 16. Results are presented over 10 runs.

Fast params	k	Mean test acc	Min test acc	Max test acc	Avg Time (s)
Layer fc	$k = 3$	89.26%	88.69%	90.01%	245
	$k = 5$	89.43%	87.92%	90.28%	198
	$k = 10$	88.53%	84.79%	89.79%	180
Layer 5 + fc	$k = 3$	88.91%	87.42%	90.06%	264
	$k = 5$	89.70%	89.35%	90.23%	224
	$k = 10$	88.65%	87.97%	89.73%	207

Table A5. A single hidden layer perceptron trained on MNIST data using our multirate approach for neural network regularization (Section 5.1) with $h_F = 0.1$. Left: different values of k with coupled $h_S = kh_F$. Right: $k = 5$ and uncoupled h_S . Weights in the input and hidden layer are de-activated with a probability of 0.8 and 0.5, respectively (same setting as in Figure 8). Results are presented over 10 runs.

k	Mean test acc	Min test acc	Max test acc	h_S	Mean test acc	Min test acc	Max test acc
$k = 3$	76.16%	61.31%	89.41%	$h_S = 0.2$	98.26%	98.14%	98.37%
$k = 5$	98.30%	98.17%	98.44%	$h_S = 0.5$	98.30%	98.17%	98.44%
$k = 10$	98.22%	98.11%	98.29%	$h_S = 0.8$	98.31%	98.21%	98.44%
				$h_S = 1$	98.29%	98.19%	98.44%

Table A6. A transformer trained on Penn Treebank data (same setting as in Figure 9) using our multirate approach for neural network regularization (Section 5.1) with $h_F = 0.1$. Left: different values of k with coupled $h_S = kh_F$. Right: $k = 5$ and uncoupled h_S . Results are presented over 10 runs.

k	Minimum Validation Loss	h_S	Minimum Validation Loss
$k = 3$	4.870 ± 0.297	$h_S = 0.2$	4.824 ± 0.304
$k = 5$	4.825 ± 0.302	$h_S = 0.3$	4.823 ± 0.304
$k = 10$	4.838 ± 0.303	$h_S = 0.5$	4.825 ± 0.302
		$h_S = 0.8$	4.831 ± 0.298