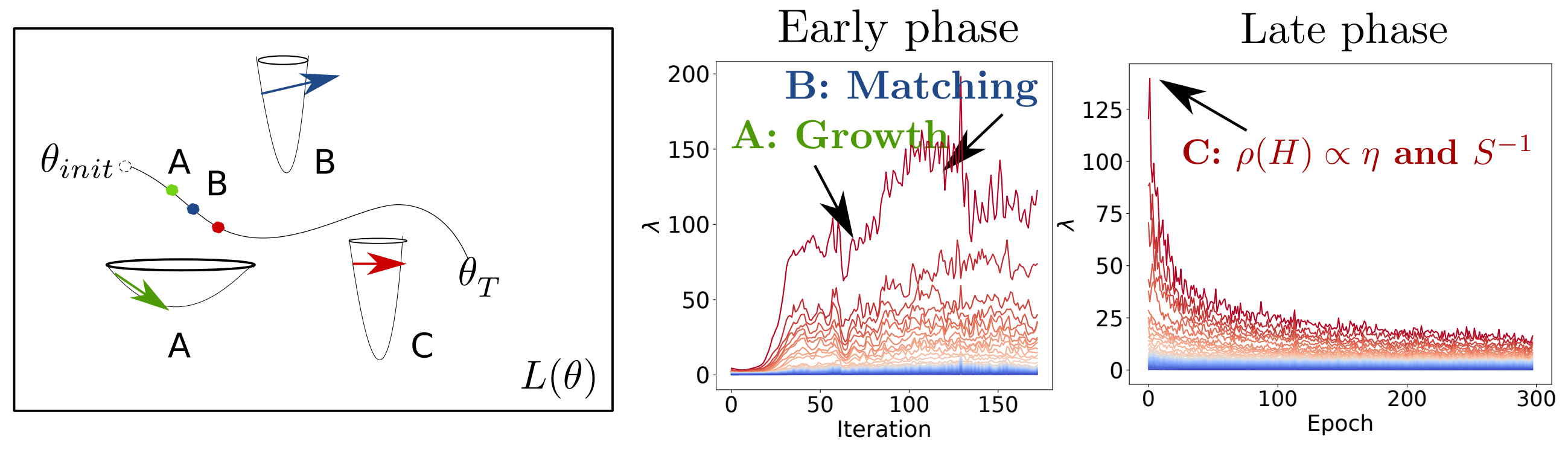
DNN's Sharpest Directions Along the SGD Trajectory

Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, Amos Storkey

Summary

Recent work has identified that a high learning rate or a small batch size in SGD training of DNNs encourages finding flatter minima. Flatness has been shown to correlate with good generalization performance.



Left: Schematic illustration of the evolution of the loss surface along one of the top eigenvectors during training. Curvature along this direction initially grows, and then stabilizes or decays. **Right two:** Evolution of the top 30 (decreasing, red to blue) eigenvalues of the Hessian for a simple CNN model during training (with $\eta = 0.005$).

Our key findings:

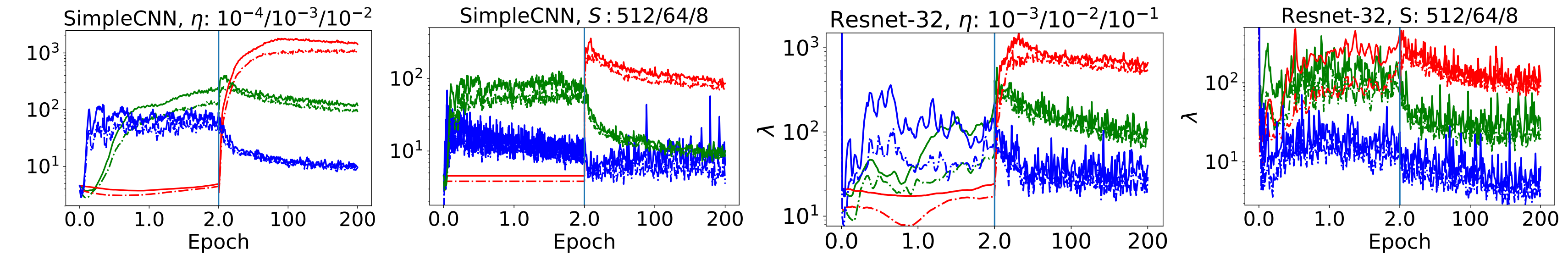
- At the **beginning of training**, a high learning rate or small batch size influences SGD to visit flatter loss regions.
- The evolution of the largest eigenvalues always follow a similar pattern, with a fast increase in the first epochs and a steady decrease thereafter, where the peak value is determined by the learning rate and batch size.
- By altering the learning rate in the sharpest direction, SGD can be steered towards regions which are an order of magnitude sharper with similar generalization.

Experiments setup

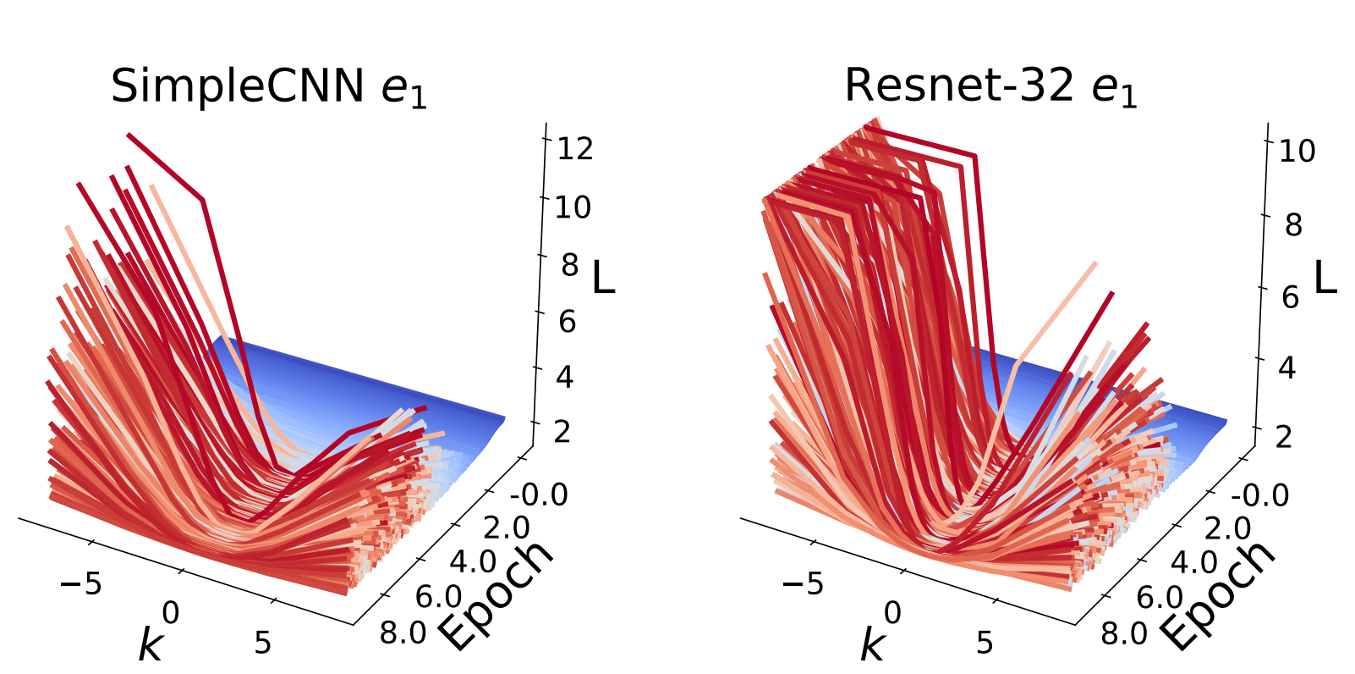
- We perform experiments mainly on Resnet-32 and a simple convolutional neural network (SimpleCNN) – a 4 layer CNN, and the CIFAR-10 dataset
- The Hessian of the training loss is approximated using the Lanczos algorithm on approx. 5% of the training set
- Vanilla SGD without momentum is used
- Additional results using momentum, different models, and datasets are provided in the paper

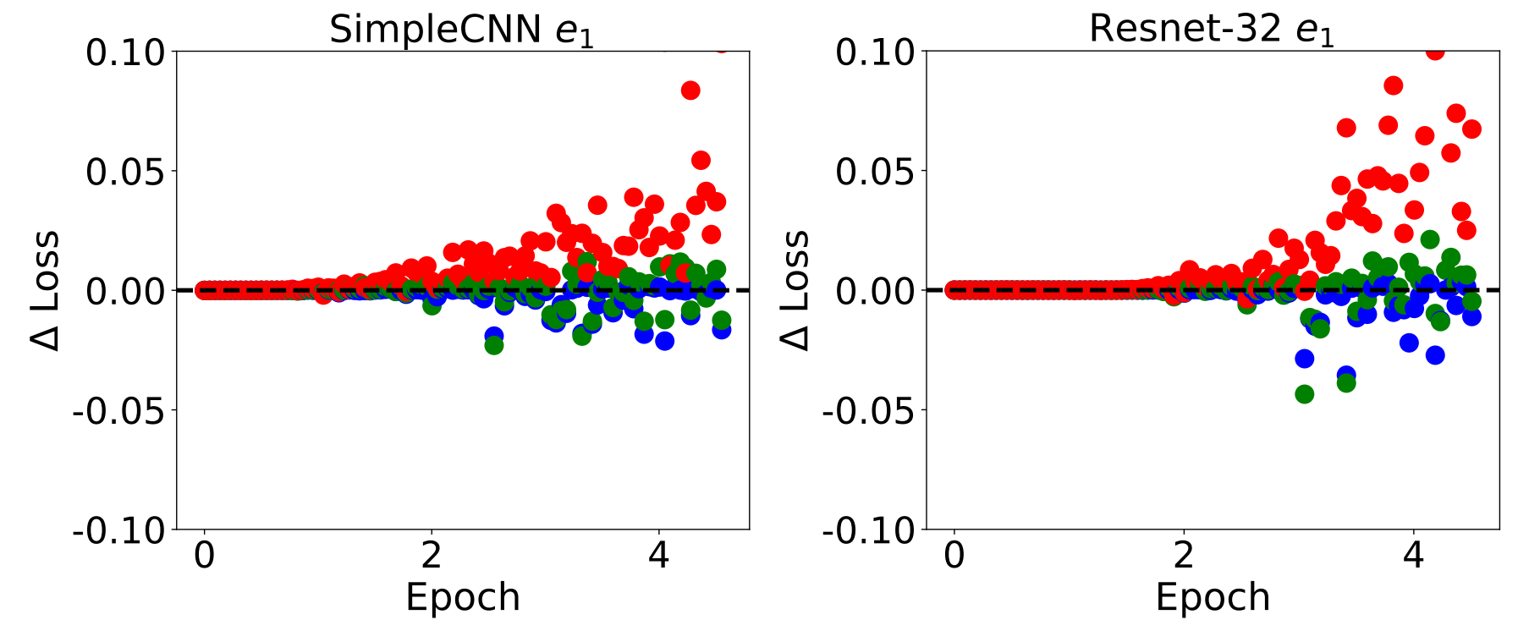
SGD is biased towards flat regions

We investigate how the curvature of the loss function along the path found by SGD evolves during training.



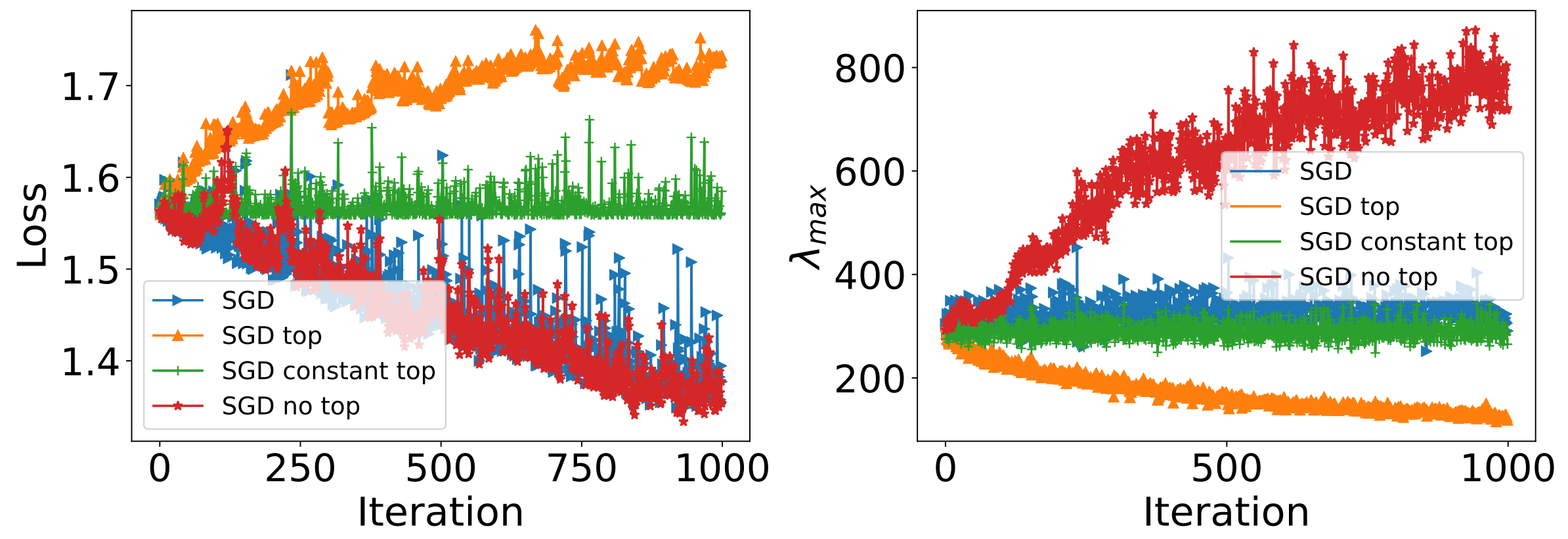
Evolution of the two largest eigenvalues of the Hessian for SimpleCNN on a log-scale for different learning rates (even columns) and batch-sizes (odd columns). **Larger learning rate or a smaller batch-size correlates with a smaller and earlier peak of the spectral norm and the subsequent largest eigenvalue.**





After a couple of epochs the loss in the subspace starts to have a bowl-like shape. Loss surface along the top eigenvector at the beginning of training . At iteration t we plot the loss $L(\vec{\theta}(t) + k\Delta\vec{\theta}_1(t))$, around the current parameters $\vec{\theta}(t)$, where $\Delta\vec{\theta}_1(t)$ is the expected norm of the SGD step along the top eigenvector. SimpleCNN (ResNet-32) is trained with $\eta = 0.005$, $S = 128$ (with $\eta = 0.025$, $S = 128$).

The SGD step-size in the direction of the sharpest direction does not minimize the loss, however, halving the step size can further minimize the loss in this direction. Average change in loss for $\alpha = 0.5, 1, 2$ corresponding to red, green, and blue, respectively. The red points further show that increasing the step-size by a factor of two consistently increases the loss, on average.

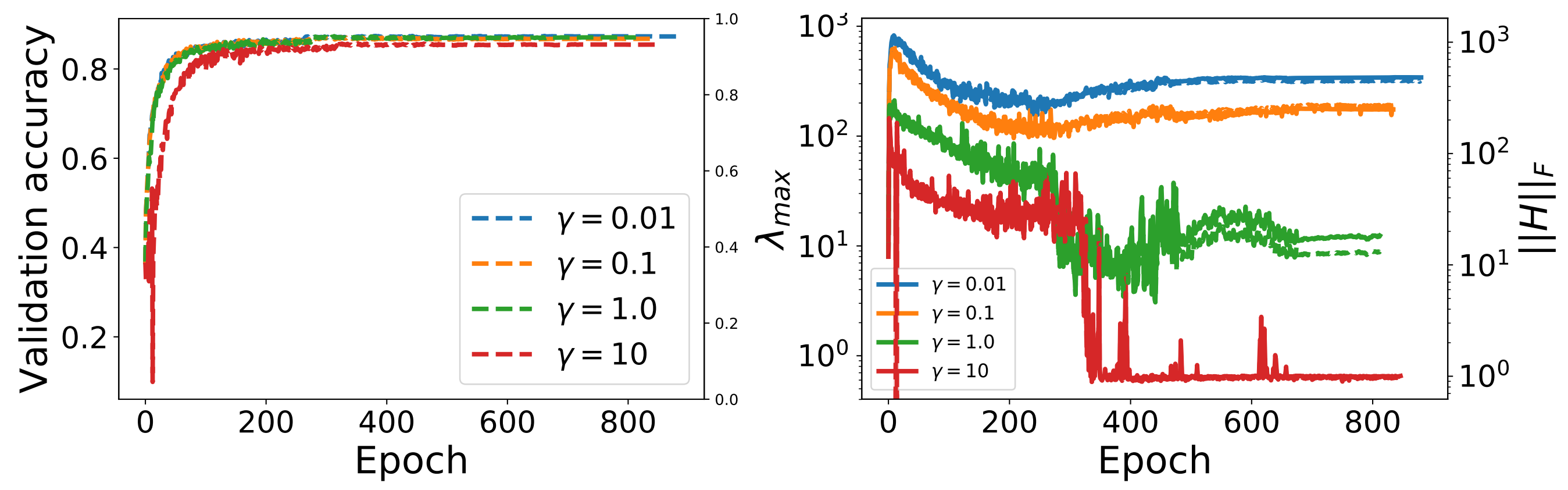


Evolution of the loss (left) and the magnitude of the largest eigenvalue of the Hessian for different SGD variants.

- **SGD variant (orange) following only the projection of the gradient on the top eigenvector finds a flatter region.**
- **SGD variant (red) subtracting this projection from the gradient, finds a sharper region while achieving a similar loss level as vanilla SGD (blue).**

Finding sharp regions that generalize well

- Our results suggest a straightforward way to steer SGD towards a sharp or wide minima
- In contrast to Dinh et al. the sharp regions that we investigate here are the endpoints of an optimization procedure, rather than a result of a reparametrization
- We investigate SGD variant, Nudged SGD, **using a different rescaled learning rate ($\eta' = \gamma\eta$) along the 5 sharpest directions.**



Results for SimpleCNN model. NSGD can steer towards an order of magnitude sharper or wider minima. **Left:** Validation accuracy. **Right**: Spectral norm (solid) and Frobenius norm (dashed), plotted on a log scale.

γ	$\lvert\lvert\mathbf{H}\rvert\rvert_F$	Test acc.	Val. acc. (50)	Loss	Dist.
0.01	672/1,559	87.4%	58.3%	0.06067	22.99
0.1	487/816	87.2%	57.0%	0.06128	23.41
1.0	119/121	86.6%	54.0%	0.06443	24.62
10	39/19	85.2%	43.8%	0.10194	29.39
SGD(0.1)	21/13	89.7%	82.6%	0.10787	34.90
SGD(0.0025)	890/1,190	84.8%	30.9%	0.10304	18.65
SGD(0.005)	377/511	85.6%	41.7%	0.08392	21.21

Results for Resnet-32 model. **We find sharper minima (second column) with similar generalization performance (third column)**.

Conclusions & Discussion

- SGD with a high learning rate or small batch size is biased towards a flat region early on in the training
- Loss surface along the sharpest direction assumes a bowl-like shape with curvature and *width* adapted to the learning rate and batch size
- "Escaping sharp minima" happens in the beginning of training
- SGD can be easily *nudged* towards much sharper regions, not affecting generalization, while slightly improving training speed

Acknowledgments: SJ was in part supported by Grant No. DI 2014/016644 from Ministry of Science and Higher Education, Poland and No. 2017/25/B/ST6/01271 from National Science Center, Poland. Work at MILA was funded by NSERC, CIFAR, and Canada Research Chairs. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732204. This work is supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 16.0159.