Langevin Algorithms for Very Deep Neural Networks with Application to Image Classification

Pierre BRAS

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université, Paris, France

Presented at the International Neural Network Society, Deep Learning Innovations and Applications INNS DLIA workshop,
part of the International Joint Conference on Neural Networks IJCNN 2023

Table of contents

Consider a training problem with parameter θ and data $\mathcal{D}$ and learning rate γ:

Gradient Descent versus Langevin Gradient Descent

$$\begin{aligned}
\text{(Stochastic) Gradient:} \quad & g_{n+1} = \nabla_\theta V(\theta_n; \mathcal{D}_{n+1}) \\
\text{(Stochastic) Gradient Descent:} \quad & \theta_{n+1} = \theta_n - \gamma_{n+1} g_{n+1}, \\
\text{Langevin (Stochastic) Gradient Descent:} \quad & \theta_{n+1} = \theta_n - \gamma_{n+1} g_{n+1} + \sigma \sqrt{\gamma_{n+1}} \mathcal{N}(0, I_d),
\end{aligned}$$

- Introduced in a Bayesian setting Welling and Teh (2011)
- The small white noise adds learning regularization
- Allows to escape from traps for the gradient descent: local minima, saddle points
- Adding noise is known to improve the learning in some cases Neelakantan et al. (2015); Anirudh Bhardwaj (2019); Gulcehre et al. (2016)

Preconditioned Langevin Gradient Descent Li et al. (2016)

For some preconditioner rule P_{n+1} depending on the previous updates of the gradient:

Preconditioned Gradient Descent: $\theta_{n+1} = \theta_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1}$,

Preconditioned Langevin: $\theta_{n+1} = \theta_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1} + \sigma \sqrt{\gamma_{n+1}} \mathcal{N}(0, \boldsymbol{P_{n+1}})$

- Per-dimension adaptive step size
- Typical examples: Adam, RMSprop, Adadelta...
- Li et al. (2016); Ma et al. (2015); Patterson and Teh (2013); Simsekli et al. (2016) compares the benefits of noisy and/or preconditioned optimizers

- Very deep neural networks are crucial, in particular in image classification He et al. (2016)
- However much more difficult to train: much more "non-linear", local traps, vanishing gradients
- Neelakantan et al. (2015): hints that noisy optimizers bring more improvements in this very deep setting

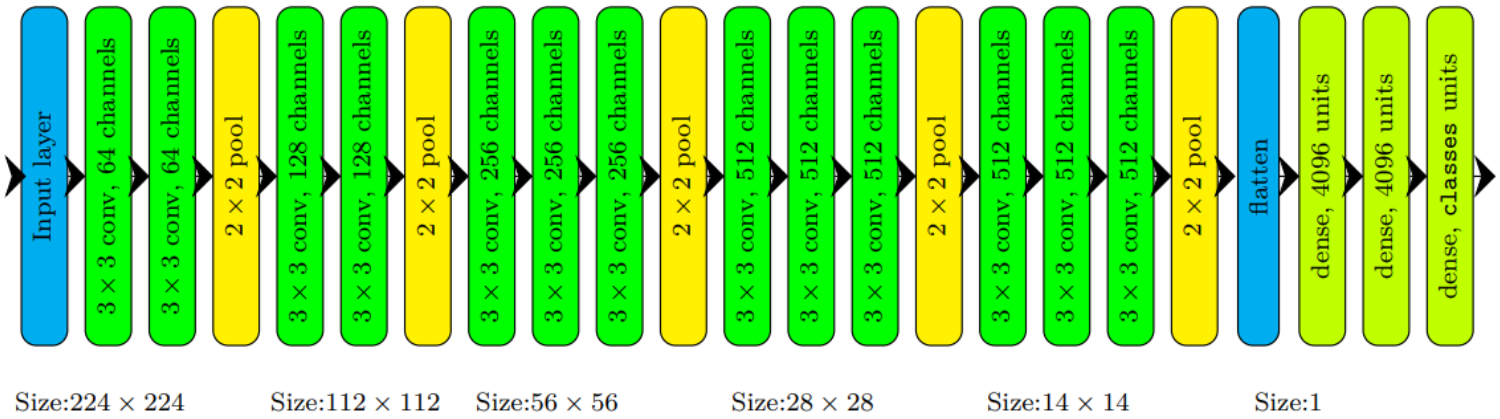

Figure: Architecture of the VGG-16 network for an input image of size 224 × 224.

Objectives

- Side-by-side comparison of preconditioned Langevin versus their respective non-Langevin counterparts: Adam vs L-Adam, RMSprop vs L-RMSprop etc
- We progressively increase the depth of the network
- Based on this heuristic, we introduce the Layer Langevin algorithm: Add noise only to some layers of the network
- Test Langevin and Layer Langevin algorithms on deep image analysis architectures

We compare Preconditioned Langevin optimizers with their non-Langevin counterparts while increasing the depth of the network on:

- Fully connected (Dense) neural networks
- Convolutional layers followed by dense layers,

on the MNIST, CIFAR-10 and CIFAR-100 datasets.

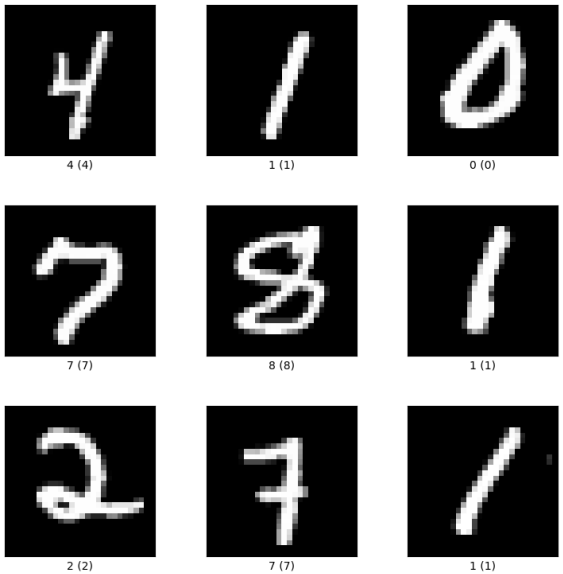

Figure: MNIST image dataset

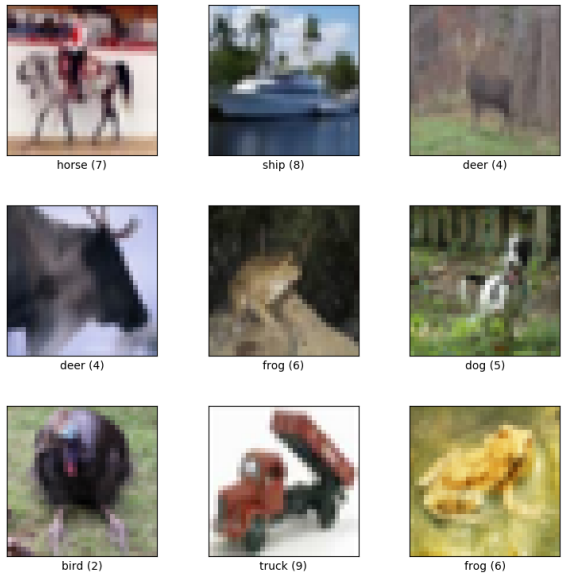

Figure: CIFAR-10 image dataset

Results for dense (fully connected) networks

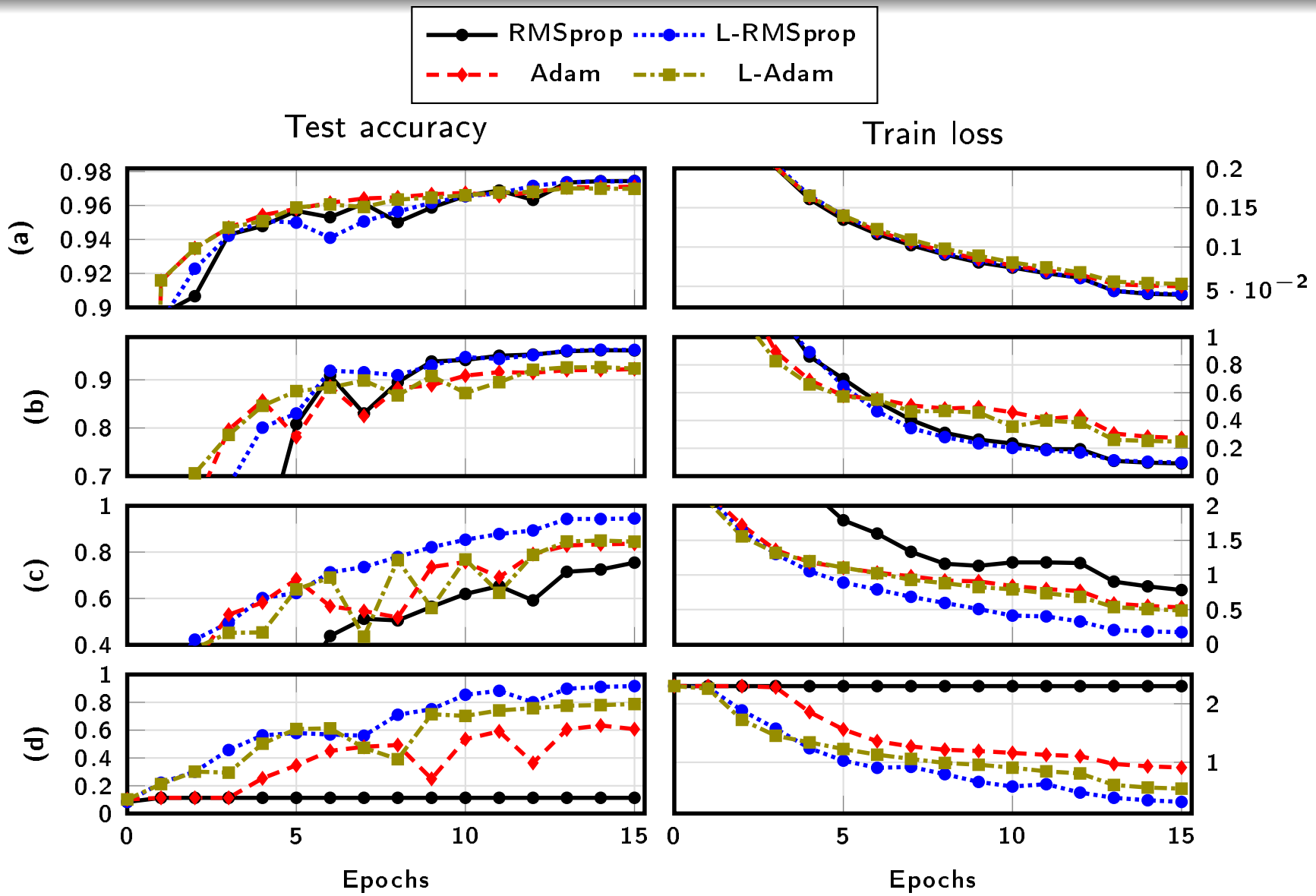

Figure: Training of neural networks of various depths on the MNIST dataset using Langevin algorithms compared with their non-langevin counterparts. (a): 3 hidden layers, (b): 20 hidden layers, (c): 30 hidden layers, (d): 40 hidden layers.

Results for convolutional layers

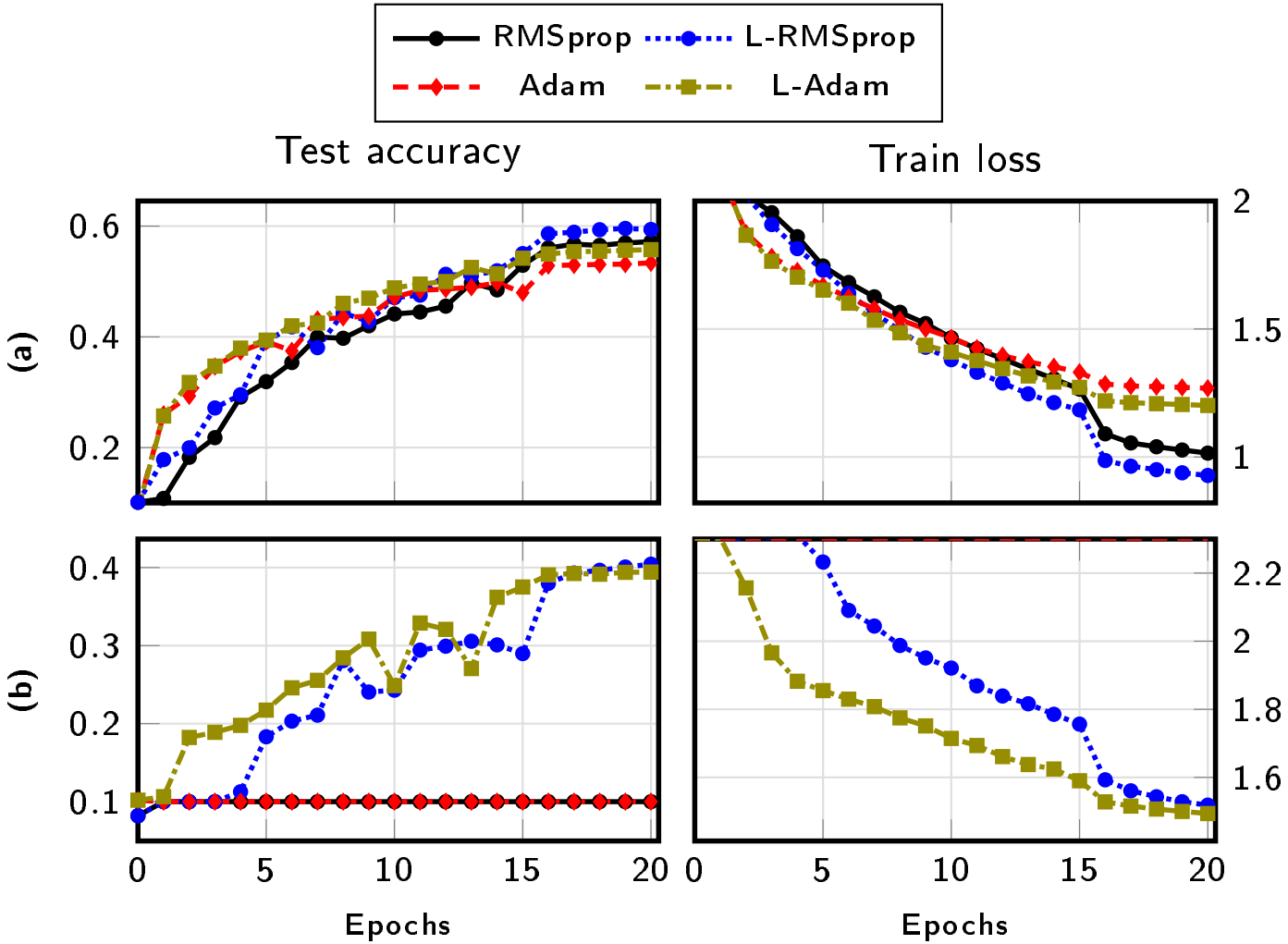

Figure: Training of convolutional neural networks on the CIFAR-10 dataset. (a): 10 hidden dense layers, (b): 30 hidden layers.

$\Longrightarrow$ The deeper the network is, the greater are the gain provided by Langevin optimizers.

To deal with very deep networks, highway networks Srivastava et al. (2015) introduce parametrized residual connection:

$$y = T_{\theta_T}(x) \cdot D_{\theta_D}(x) + (1 - T_{\theta_T}(x)) \cdot x,$$

where T and D are dense or convolutional layers.

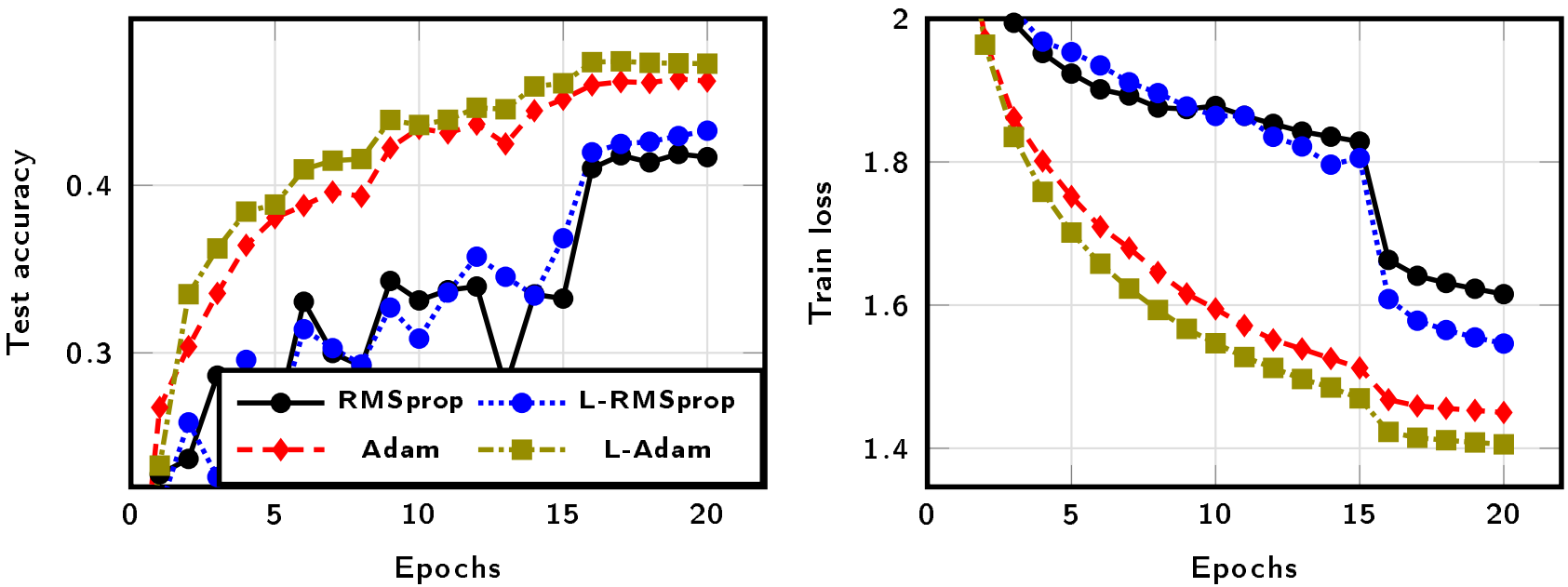

Figure: Training of a highway neural network with 80 highway hidden layers on the CIFAR-10 dataset.

$\Longrightarrow$ The previous conclusion is still true but only from a larger depth.

Idea: The deepest layers of the network bear the most non-linearities $\Longrightarrow$ are more subject to Langevin optimization

Layer Langevin Algorithm

$$\theta_{n+1}^{(i)} = \theta_n^{(i)} - \gamma_{n+1}[P_{n+1} \cdot g_{n+1}]^{(i)} + \mathbb{1}_{i \in \mathcal{J}} \sigma \sqrt{\gamma_{n+1}} \left[\mathcal{N}(0, P_{n+1})\right]^{(i)}, \tag{1}$$

where $\mathcal{J}$: subset of weight indices; P_n: preconditioner.

We choose $\mathcal{J}$ to be the first k layers.

An example of Layer Langevin optimization

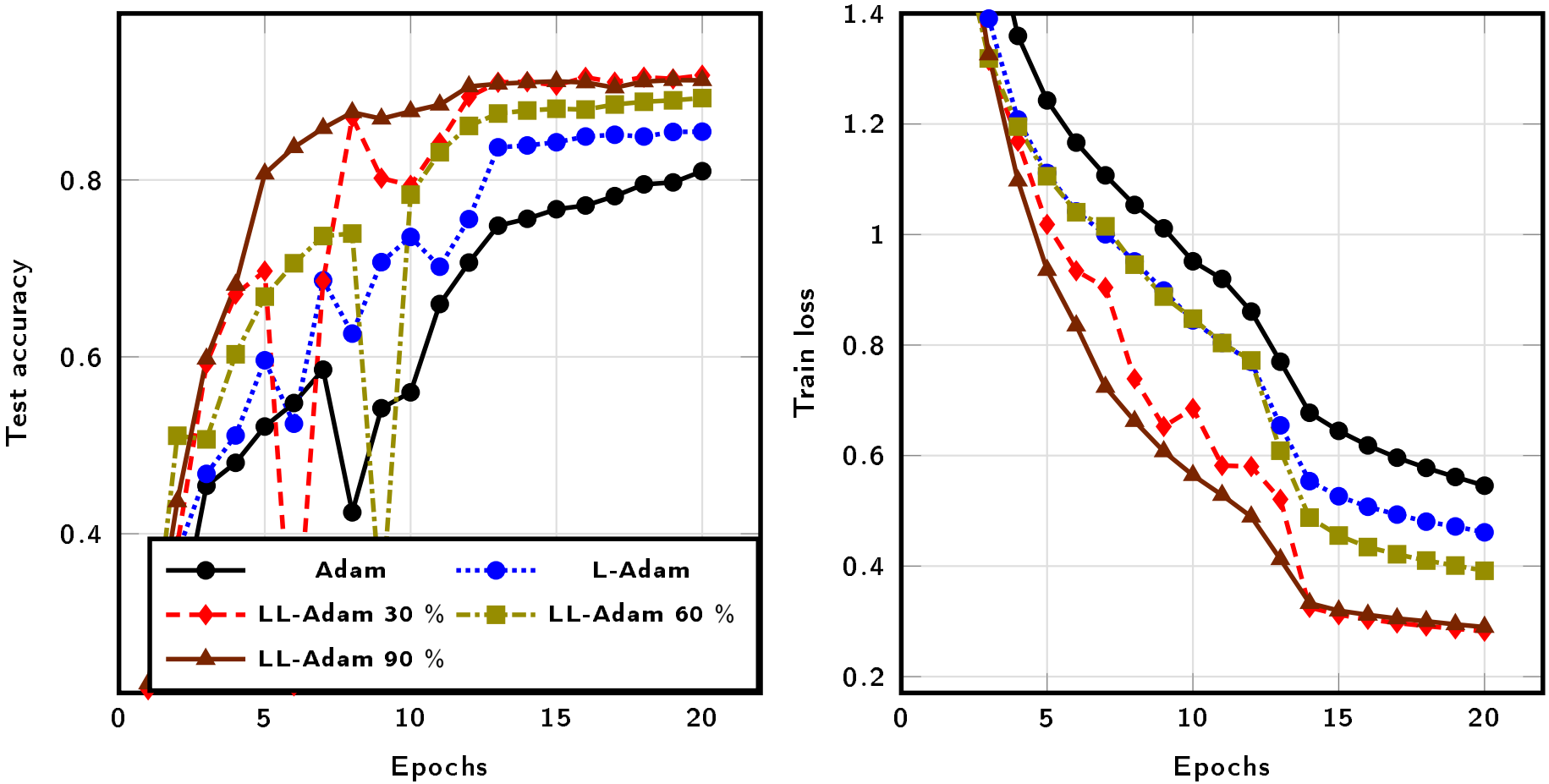

Figure: Layer Langevin method comparison on a dense neural network with 30 hidden layers on the MNIST dataset.

- Typical architecture in image recognition: Succession of convolutional layers with non-linearities (ReLU); the dimensions (width and height) of the image are progressively reduced while the number of channels is progressively augmented Simonyan and Zisserman (2015).
- Depth is crucial.
- Residual connections: each layer behaves in part like the identity layer to pass the information through the successive layers He et al. (2016); Huang et al. (2017).

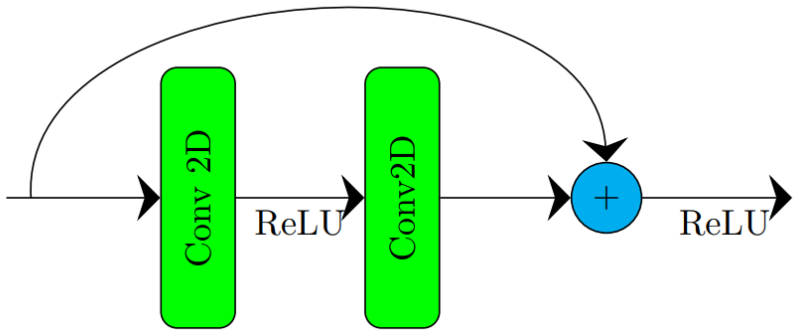

Figure: ResNet elementary block

Layer Langevin for training of ResNet-20

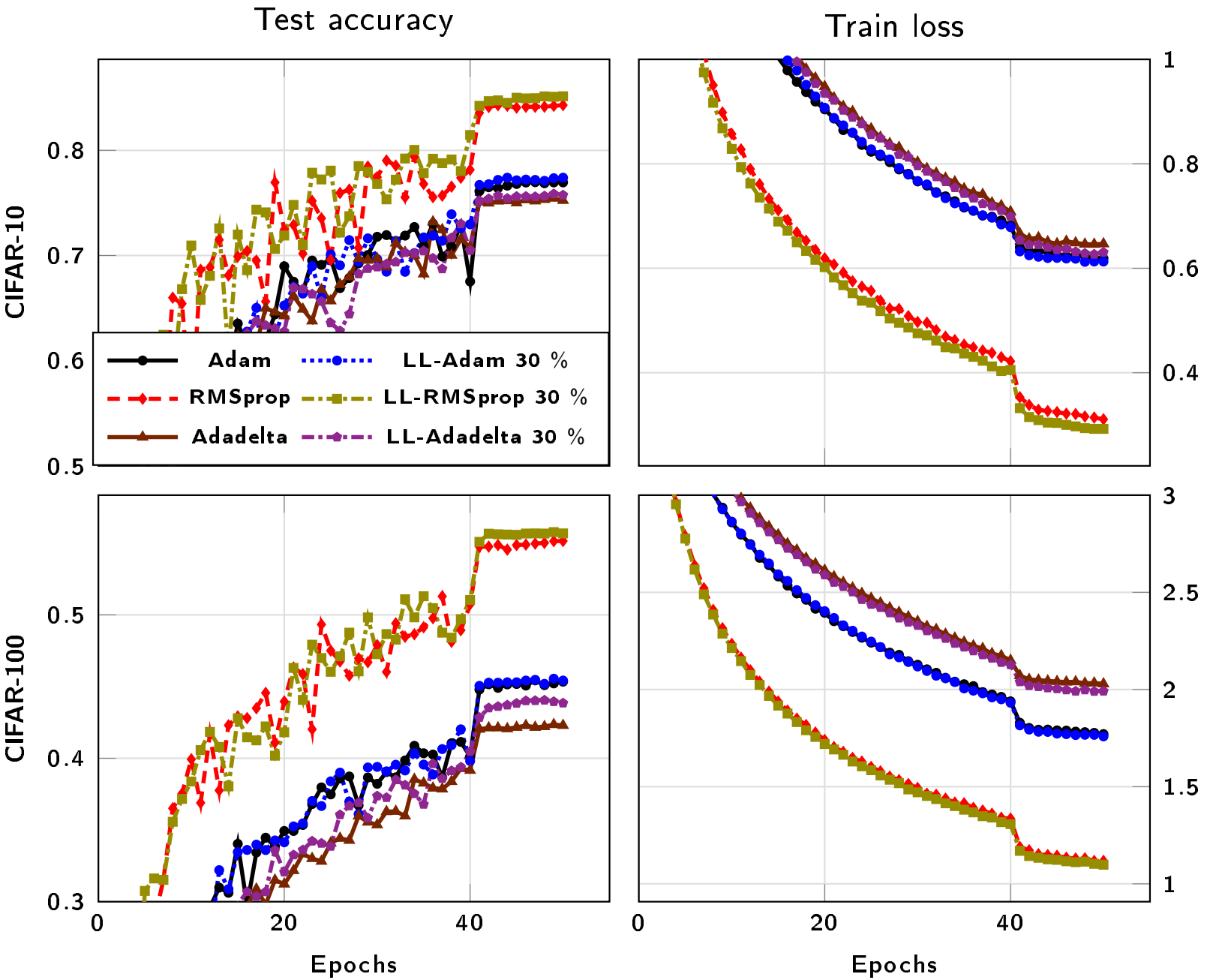

Table: Final test accuracy values obtained for ResNet

	Adam	LL-Adam	RMSprop	LL-RMSprop	Adadelta	LL-Adadelta
CIFAR-10	76.95 %	77.39 %	84.29 %	85.14 %	75.23 %	75.74 %
CIFAR-100	45.33 %	45.41 %	55.15 %	55.68 %	42.28 %	43.84 %

Table: Final test accuracy values on the CIFAR-10 dataset with DenseNet architecture.

	Adam	LL-Adam	RMSprop	LL-RMSprop	Adadelta	LL-Adadelta
CIFAR-10	87.81 %	88.16 %	57.59 %	57.56 %	71.64 %	72.72 %

Thank you for your attention !

Citations I

C. Anirudh Bhardwaj. Adaptively Preconditioned Stochastic Gradient Langevin Dynamics. *arXiv e-prints*, art. arXiv:1906.04324, June 2019.

C. Gulcehre, M. Moczulski, M. Denil, and Y. Bengio. Noisy activation functions. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 3059–3068. JMLR.org, 2016.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

C. Li, C. Chen, D. Carlson, and L. Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 1788–1794. AAAI Press, 2016.

Y. Ma, T. Chen, and E. B. Fox. A Complete Recipe for Stochastic Gradient MCMC. In *Neural Information Processing Systems*, 2015. URL http://terraswarm.org/pubs/646.html.

A. Neelakantan, L. Vilnis, Q. V. Le, I. Sutskever, L. Kaiser, K. Kurach, and J. Martens. Adding Gradient Noise Improves Learning for Very Deep Networks. *arXiv e-prints*, art. arXiv:1511.06807, Nov. 2015.

S. Patterson and Y. W. Teh. Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/309928d4b100a5d75adff48a9bfc1ddb-Paper.pdf.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

U. Simsekli, R. Badeau, A. T. Cemgil, and G. Richard. Stochastic Quasi-Newton Langevin Monte Carlo. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 642–651, 2016.

R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

M. Welling and Y. W. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688. Omnipress, 2011. ISBN 9781450306195.