Can Network Flatness Explain the Training Speed-Generalisation Connection?

Albert Q. Jiang*
Computer Laboratory, University of Cambridge

Lisa Schut **Clare Lyle** **Yarin Gal**
OATML Group, University of Oxford

Abstract

Recent work has shown that training speed, as estimated by the sum over training loss, is predictive of generalization performance. From a Bayesian perspective, this metric can be theoretically linked to marginal likelihood in linear models. However, it is unclear why the relationship holds for DNNs and what the underlying mechanisms are. We hypothesise that this relationship holds in DNNs because of network flatness, which causes both fast training speed and good generalization. We also investigated the hypothesis in varying settings and found that it might hold when the variance in the stochastic gradient estimation is moderate, with either logit averaging, or no data transformation at all. This paper specifies the conditions future works should impose when investigating the connecting mechanism.

1 Introduction

The problem of how neural networks generalise to unseen data is not considered fully answered. The neural network community is still in search of plausible explanations of generalisation, possibly with non-vacuous bounds, based on realistic assumptions. The large-scale investigation by Jiang et al. [2020] showed the potential of optimisation-based measures in predicting generalisation. Specifically, the optimisation-based measure training speed was studied by Lyle et al. [2020] and Ru et al. [2020]. Lyle et al. [2020] introduced the sum over training losses (SOTL) as a measure for training speed: $\text{SOTL} = \sum_{t=0}^{T-1} R(\theta_t)$, where $R(\theta_t)$ is the training loss of the model at the end of epoch t and T is the number of epochs. In this paper we empirically explore the training-speed explanation of generalisation.

They conjectured that the SOTL measures how well parameter updates transfer across mini-batches of training data, hence connecting training speed and generalisation. For linear models and infinitely wide networks, the SOTL estimates a lower bound of the marginal likelihood, which is a principal tool for Bayesian model selection. For deep neural networks, training speed's predictive power was illustrated by Ru et al. [2020], where it was used as a heuristic in Neural Architecture Search.

The flatness hypothesis We hypothesise that the connection between fast training speed and good generalisation can be explained by network flatness. Hochreiter and Schmidhuber [1997] suggested that flatter minima require less information to describe to a certain accuracy. The smaller minimum description length (MDL) of flatter minima can contribute to a better generalisation performance. The PAC-Bayes bound [McAllester, 1999] also connects the generalisation of neural networks to a notion of flatness, via the KL divergence between the weight prior and posterior. Later works [Dziugaite and Roy, 2017, Neyshabur et al., 2017] showed that one can derive a good measure of generalisation (PAC-Bayes sharpness, defined in Equation 7) using simple prior and posterior forms. If one takes a posterior of the isotropic Gaussian form, the PAC-Bayes sharpness is the inverse of the Gaussian variance. The variance can in turn be an indirect measure of posterior weight around the minimum found.

*Work done when the author was a Master student at OATML Group. Correspondence to qj213@cam.ac.uk.

Bayesian Deep Learning workshop, NeurIPS 2021.

2 Experiment

In the experiments we aim to answer two questions: (1) What are the conditions under which the connection between training speed and generalisation exists for deep neural networks? (2) What are the conditions under which the flatness hypothesis might hold? We trained a diverse set of neural networks with different training hyperparameters. To observe the effects of each hyperparameter, we set up 5 experimental groups beside 1 control group, varying one single property at a time.

2.1 Setup

We chose the CIFAR10 [Krizhevsky et al., 2009] visual classification task for our experiment. 270 networks were trained, split into 6 groups (1 for control and 5 for experiment). In each group. 5 learning rates $0.001, 0.0016, 0.003, 0.006, 0.01$, 3 network widths $8, 12, 16$, and 3 network depths $2, 3, 4$ are used. These settings represent the regime of hyperparameters often used in practice. Each group contains a total of $5 \times 3 \times 3 = 45$ training runs. We used the cross entropy as the risk function. A run is terminated when the training loss reaches 0.01 (the same stopping criterion was used by Jiang et al. [2020], Dziugaite et al. [2020]), or when it has been trained for 400 epochs.

In the control group we used an architecture free of batch normalization [Ioffe and Szegedy, 2015] with SGD as the optimiser and no data augmentation. We will explain the variations of the experimental groups in section 2.3.

For each run we monitor the final generalisation error and several complexity measures. We also monitor three measures related to training speed [Lyle et al., 2020, Ru et al., 2020]: Sum Over Training Loss (SOTL), Sum Over Training Loss of the last E epochs with E=50 (SOTL-E_50), and Training Speed Estimator-Exponential Moving Average (TSE-EMA). SOTL-E_50 and TSE-EMA are variants of the SOTL that focus on late-stage and early-stage training speed respectively. Two other measures are related to the network flatness are also monitored: PAC-Bayes Sharpness [McAllester, 1999, Dziugaite and Roy, 2017, Neyshabur et al., 2017], and Value sensitivity [Neu, 2021]. The former is known to provide a non-vacuous bound for generalisation, while the latter can be used to derive an information-theoretic generalisation bound for stochastic gradient descent. We define these quantities formally in Appendix A.

2.2 Evaluation

We used Kendall's ranking-correlation coefficient and Normalised Conditional Mutual Information (NCMI), similar to the setup of Jiang et al. [2020], as evaluation criteria for assessing the complexity measure's predictive power of the generalisation error. The Kendall coefficient between two entities a and b measures their correlation through ranking and is denoted by $\tau(a, b)$. The conditional mutual information between a and b measures how much information about variable a is revealed through b. We can then normalise the entity by the amount of information there is in variable a, usually the value of a complexity measure in our experiments. We denote the minimum NCMI of a complexity measure μ by $K(\mu)$, and the average NCMI by $\bar{K}(\mu)$. The larger the coefficient, the more correlated are the two entities. The higher the NCMI value, the more likely there is a causal connection between the complexity measure and the generalisation error. In preliminary experiments, we found TSE-EMA to be a robust predictor of generalisation so we calculate the correlation coefficients between it and other complexity measures too.

2.3 Results

We present the full results of the networks trained in Appendix C. In the main text we present a selection of results and the insights drawn from them. More analysis is in Appendix B.

Control Group In the control group we used the MLP architecture and initialisation [De and Smith, 2020] without batch normalization [Ioffe and Szegedy, 2015]. We chose this architecture because we wished to study the effect of batch normalization explicitly in one of our experimental groups. We present the important quantities in Table 1. We found that in the control group, the SOTL and the TSE-EMA are both very correlated with the generalisation error, according to the Kendall's ranking-correlation coefficient and the NCMI. Therefore training speed exhibits a strong connection with generalisation.

The flatness-based measures have relatively weaker correlations with the generalisation error, although still significant. They also have strong correlations with the training speed. Hence in this group, the flatness hypothesis might hold for the connection between training speed and generalisation.

Table 1: The important analytical quantities of the complexity measures considered in the control group. Bold values indicate the largest value by magnitude. g is the generalisation error.

Control group	Average test accuracy: 71.1 %			
μ	$\mathcal{K}(\mu)$	$\bar{\mathcal{K}}(\mu)$	$\tau(\mu, g)$	$\tau(\mu, \text{TSE-EMA})$
SOTL	**0.130**	**0.325**	**0.640**	**0.949**
SOTL-E_50	0.015	0.050	-0.228	-0.125
TSE-EMA	0.123	0.310	0.610	-
PAC-Bayes sharpness	0.071	0.266	0.392	0.642
Value sensitivity	0.037	0.262	0.400	0.644

Experimental Groups 1 and 2: Data Augmentation We considered two forms of data augmentation. In experimental group 1 we applied to each image a standard transformation every epoch: a random cropping of 32×32 pixels, with a padding of 4 pixels, followed by a horizontal flip with probability 0.5; In experimental group 2, we used logit averaging [Nabarro et al., 2021]: we applied the same transformation 4 times to each image, resulting in 4 different images. Then the logits (unnormalised categorical probabilities) of the 4 images are averaged before loss is calculated.

The results from these two groups are presented in Table 2. We found that in both groups, the average test accuracies increased, compared to the control group. For all 45 different hyperparameter combinations, using data transformation improves both the network flatness and the generalisation performance. However, we found that the correlation between generalisation error and measures related to training speed are negligible or negative. This might be due to the distribution shift between original and transformed images: training loss is calculated on transformed images while generalisation error is calculated on original images. Using logit averaging fixes this problem: both the training loss and the generalisation error are calculated on transformed images. We found that in experimental group 2, training speed measures are strongly correlated with the generalisation error, and flatness measures are strongly correlated with TSE-EMA. Therefore, the connection between training speed and generalisation does not exist when data transformation is used. It exists, however, when logit averaging is used together with data transformation, and the flatness hypothesis might account for it then.

Table 2: The important analytical quantities of the complexity measures considered in the experimental groups 1 and 2. g is the generalisation error.

Experiemntal group 1	Average test accuracy: 85.0 %			
μ	$\mathcal{K}(\mu)$	$\bar{\mathcal{K}}(\mu)$	$\tau(\mu, g)$	$\tau(\mu, \text{TSE-EMA})$
SOTL	0.116	0.222	-0.097	0.820
TSE-EMA	0.032	0.140	0.046	-
Experiemntal group 2	Average test accuracy: 82.6 %			
μ	$\mathcal{K}(\mu)$	$\bar{\mathcal{K}}(\mu)$	$\tau(\mu, g)$	$\tau(\mu, \text{TSE-EMA})$
SOTL	0.035	0.154	0.414	0.907
TSE-EMA	0.029	0.150	0.428	-
PAC-Bayes sharpness	0.038	0.156	0.473	0.651
Value sensitivity	0.031	0.138	0.469	0.644

Experimental Group 3: Variance of the Stochastic Gradients In this group we used Stochastic Gradient Langevin Dynamics (SGLD) to increase the variance in the gradient estimator. The results are displayed in Table 3. The average test accuracy in this group is 78.4%, 7.3% higher than the control group. In this group we found that flatness measures fail to predict generalisation: they have negative correlation coefficients with the generalisation error. We found that for all 45 hyperparameter combinations, SGLD makes networks generalise better but the minima found sharper or flatter randomly.

Hence although we did find the connection between generalisation and training speed, network flatness cannot be the cause.

Table 3: The important analytical quantities of the complexity measures considered in the experimental group 3. g is the generalisation error.

Experiemntal group 3	Average test accuracy: 78.4 %			
μ	$\mathcal{K}(\mu)$	$\bar{\mathcal{K}}(\mu)$	$\tau\,(\mu, g)$	$\tau\,(\mu, \text{TSE-EMA})$
PAC-Bayes sharpness	0.015	0.048	-0.253	-0.022
Value sensitivity	0.041	0.060	-0.154	-0.085

3 Conclusion

In line with the work of Lyle et al. [2020], Ru et al. [2020], we showed that training speed exhibits a connection with the generalisation performance of deep neural networks. We extend the work by empirically showing that the connection exists when the neural architecture is SkipInit or ResNet, when the optimiser is SGD or SGLD, and when logit averaging is used for data augmentation or no data augmentation is used at all. Additionally, we showed that network flatness might explain this connection when SGD is used as the optimiser, when SkipInit or ResNet is used as the architecture, and when logit averaging or no data augmentation is used. Experimenting further under these conditions is a direction for future work in order to understand the true mechanism of the connection.

4 More Related Work

There have been numerous attempts at bounding the generalisation performances using classical methods from a learning theory perspective. Hardt et al. [2016] showed that when stochastic gradient descent is used as the optimisation method, parametric models have vanishing generalisation errors; Hochreiter and Schmidhuber [1997] and Keskar et al. [2017] demonstrated that the network flatness can explain generalisation from the perspective of minimum description length; McAllester [1999] drew from the probably approximately correct theory and the Bayesian theory of learning and derived a bound for generalisation; Jacot et al. [2018] used infinite-width network dynamics to explain generalisation. However, these methods suffer from serious flaws such as vacuous bounds, as pointed out by Dziugaite and Roy [2017] and Neyshabur et al. [2017]; they may exhibit paradoxical behaviours under equivalent network reparameterisation shown by Dinh et al. [2017]; or they may require unrealistic assumptions such as the network being infinitely wide; or that the connection might not exist when a different, but commonly used optimiser is used [Zhang et al., 2021]. These works inspire us to find the practical conditions under which the connection might exist.

Acknowledgement

We thank Mark van der Wilk and Andrew Jesson for helpful discussions.

References

Soham De and Samuel L. Smith. Batch normalization biases residual blocks towards the identity function in deep networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/e6b738eca0e6792ba8a9cbcba6c1881d-Abstract.html`.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR, 2017. URL `http://proceedings.mlr.press/v70/dinh17b.html`.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In Gal Elidan, Kristian Kersting, and Alexander T. Ihler, editors, *Proceedings of the Thirty-Third Conference on*

Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017. AUAI Press, 2017. URL http://auai.org/uai2017/proceedings/papers/173.pdf.

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/86d7c8a08b4aaa1bc7c599473f5dddda-Abstract.html.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1225–1234. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/hardt16.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Comput.*, 9(1):1–42, 1997. doi: 10.1162/neco.1997.9.1.1. URL https://doi.org/10.1162/neco.1997.9.1.1.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/ioffe15.html.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8580–8589, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Clare Lyle, Lisa Schut, Robin Ru, Yarin Gal, and Mark van der Wilk. A bayesian perspective on training speed and model selection. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/75a7c30fc0063c4952d7eb044a3c0897-Abstract.html.

David A. McAllester. Pac-bayesian model averaging. In Shai Ben-David and Philip M. Long, editors, *Proceedings of the Twelfth Annual Conference on Computational Learning Theory, COLT 1999, Santa Cruz, CA, USA, July 7-9, 1999*, pages 164–170. ACM, 1999. doi: 10.1145/307400.307435. URL https://doi.org/10.1145/307400.307435.

Seth Nabarro, Stoil Ganev, Adrià Garriga-Alonso, Vincent Fortuin, Mark van der Wilk, and Laurence Aitchison. Data augmentation in bayesian neural networks and the cold posterior effect. *CoRR*, abs/2106.05586, 2021. URL https://arxiv.org/abs/2106.05586.

Gergely Neu. Information-theoretic generalization bounds for stochastic gradient descent. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 3526–3545. PMLR, 2021. URL http://proceedings.mlr.press/v134/neu21a.html.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5947–5956, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/10ce03a1ed01077e3e289f3e53c72813-Abstract.html.

Binxin Ru, Clare Lyle, Lisa Schut, Mark van der Wilk, and Yarin Gal. Revisiting the train loss: an efficient performance estimator for neural architecture search. *CoRR*, abs/2006.04492, 2020. URL https://arxiv.org/abs/2006.04492.

Shuo Zhang, Isaac Reid, Guillermo Valle P'erez, and Ard A. Louis. Why flatness does and does not correlate with generalization for deep neural networks. 2021.

A Formal Definitions of Complexity Measures

Let T be the number of total epochs, θ be the network parameters, and R be the risk function. We can then denote the training loss at the end of epoch t by $R\left(\theta_t\right)$.

SOTL We can define the Sum Over Training Losses:

$$\text{SOTL} = \sum_{t=0}^{T-1} R\left(\theta_t\right). \tag{1}$$

TSE-EMA The Training Speed Estimator-Exponential Moving Average is a variation of the SOTL and it is a weighted average of the training losses. Training losses in earlier epochs are given more weight than those in later epochs:

$$\text{TSE-EMA} = \sum_{t=0}^{T-1} \eta^t R\left(\theta_t\right). \tag{2}$$

SOTL-E The Sum Over Training Losses of the last E epochs (SOTL-E) is another variation of the SOTL, which only considers the training losses of the last E epochs of training:

$$\text{SOTL-E} = \sum_{t=T-E}^{T-1} R\left(\theta_t\right). \tag{3}$$

PAC-Bayes sharpness To calculate the PAC-Bayes sharpness, we first define a 0-1 loss on a datapoint (x, y):

$$\hat{R}\left(f_\theta\left(x\right), y\right) = \mathbb{1}\left[\arg\max_i f_\theta(x)_i = y\right], \tag{4}$$

where f_θ is the function implemented by the network and $f_\theta\left(x\right)$ gives a vector of the probabilities for the categorical distribution.

We can then define the empirical 0-1 loss on the training set $\mathcal{D}_{train} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{|\mathcal{D}_{train}|}, y_{|\mathcal{D}_{train}|})\}$:

$$\hat{R}_{emp}\left(\theta, \mathcal{D}_{train}\right) = \frac{1}{\mid \mathcal{D}_{train} \mid} \sum_{i=1}^{|\mathcal{D}_{train}|} \hat{R}\left(f_\theta\left(x_i\right), y_i\right). \tag{5}$$

We can then have:

$$\hat{\sigma}^2 = \sup\left(\{\sigma^2 \mid \mathbb{E}_{\theta' \sim \mathcal{N}(\theta, \sigma^2 I)}\left[\hat{R}_{emp}(\theta', D_{train})\right] \leq 0.1\}\right), \tag{6}$$

and define the PAC-Bayes sharpness:

$$\text{PAC-Bayes sharpness} = \frac{1}{\hat{\sigma}^2}. \tag{7}$$

Value sensitivity The definition of the value sensitivity is closely related:

$$\text{Value sensitivity} = \mathbb{E}_{\theta' \sim \mathcal{N}(\theta, 0.01 I)}\left[\hat{R}_{emp}\left(\theta', \mathcal{D}_{train}\right)\right]. \tag{8}$$

B Additional Analysis of Experimental Results

Experimental Group 4: Convolutional Layers We used Multi-Layer Perceptrons (MLPs), an architecture without convolutional layers, in this experimental group. We found that this group has an average test accuracy of 47.4%. We found the generalisation performance of networks were very poor and most measures have relatively small correlation coefficients with the generalisation error.

Experimental Group 5: Batch Normalization In this group we used the Residual Network (ResNet) architecture [He et al., 2016], with batch normalization. They have an average test accuracy of 71.1%, the same as the control group. This validates our choice of the architecture in the control group that does indeed match the performance of ResNets, which are often used in practice.

Additionally, we found that measures related to training speed have better predictive power of the generalisation error than the control group in general. However, flatness-based measures have weaker correlations with the training-speed-based measures. This indicates that with batch normalization, it is less likely that network flatness causes the connection between training speed and generalisation.

C Full Results of Networks Trained

Exp. No.	Arch.	Optim.	Aug.	Depth	Width	LR	Test acc.(%)	Gen. error	SOTL	SOTL -E_50	TSE -EMA	PAC-Bayes sharpness	Value sens.
1	SkipInit	SGD	False	4	16	0.01	75.66	0.242	0.019	0.019	0.64	0.038	0.001
2	SkipInit	SGD	False	4	16	0.006	73.82	0.26	0.021	0.021	0.741	0.044	0.003
3	SkipInit	SGD	False	4	16	0.003	72.46	0.275	0.025	0.025	0.933	0.05	0.004
4	SkipInit	SGD	False	4	16	0.0016	71.61	0.284	0.03	0.028	1.152	0.059	0.006
5	SkipInit	SGD	False	4	16	0.001	70.35	0.296	0.035	0.024	1.335	0.064	0.007
6	SkipInit	SGD	False	4	12	0.01	75.02	0.248	0.021	0.021	0.718	0.036	0.001
7	SkipInit	SGD	False	4	12	0.006	73.95	0.259	0.023	0.023	0.837	0.041	0.002
8	SkipInit	SGD	False	4	12	0.003	73.55	0.264	0.027	0.025	1.046	0.048	0.003
9	SkipInit	SGD	False	4	12	0.0016	71.55	0.284	0.033	0.023	1.272	0.057	0.005
10	SkipInit	SGD	False	4	12	0.001	70.93	0.29	0.039	0.021	1.455	0.062	0.006
11	SkipInit	SGD	False	4	8	0.01	73.55	0.263	0.023	0.017	0.783	0.034	0.001
12	SkipInit	SGD	False	4	8	0.006	72.1	0.277	0.025	0.016	0.905	0.039	0.002
13	SkipInit	SGD	False	4	8	0.003	70.62	0.293	0.03	0.02	1.117	0.048	0.003
14	SkipInit	SGD	False	4	8	0.0016	70.03	0.299	0.036	0.02	1.345	0.055	0.005
15	SkipInit	SGD	False	4	8	0.001	69.59	0.304	0.043	0.018	1.521	0.061	0.006
16	SkipInit	SGD	False	3	16	0.01	74.11	0.257	0.019	0.019	0.669	0.037	0.001
17	SkipInit	SGD	False	3	16	0.006	73.04	0.269	0.021	0.021	0.785	0.042	0.002
18	SkipInit	SGD	False	3	16	0.003	72.18	0.278	0.026	0.026	0.995	0.05	0.004
19	SkipInit	SGD	False	3	16	0.0016	71.86	0.281	0.032	0.025	1.23	0.057	0.005
20	SkipInit	SGD	False	3	16	0.001	71.06	0.289	0.037	0.022	1.419	0.064	0.006
21	SkipInit	SGD	False	3	12	0.01	72.84	0.27	0.022	0.022	0.778	0.036	0.001
22	SkipInit	SGD	False	3	12	0.006	72.53	0.274	0.024	0.024	0.898	0.041	0.002
23	SkipInit	SGD	False	3	12	0.003	70.3	0.297	0.029	0.025	1.103	0.048	0.003
24	SkipInit	SGD	False	3	12	0.0016	69.49	0.305	0.035	0.022	1.319	0.057	0.005
25	SkipInit	SGD	False	3	12	0.001	68.85	0.311	0.042	0.019	1.481	0.062	0.006
26	SkipInit	SGD	False	3	8	0.01	70.5	0.293	0.024	0.015	0.828	0.035	0.001
27	SkipInit	SGD	False	3	8	0.006	69.17	0.306	0.026	0.016	0.958	0.039	0.002
28	SkipInit	SGD	False	3	8	0.003	68.84	0.31	0.031	0.02	1.164	0.048	0.003
29	SkipInit	SGD	False	3	8	0.0016	68.65	0.313	0.038	0.018	1.374	0.055	0.005
30	SkipInit	SGD	False	3	8	0.001	67.85	0.321	0.045	0.016	1.534	0.06	0.006
31	SkipInit	SGD	False	2	16	0.01	72.13	0.277	0.021	0.021	0.736	0.036	0.001
32	SkipInit	SGD	False	2	16	0.006	71.71	0.282	0.024	0.024	0.871	0.041	0.002
33	SkipInit	SGD	False	2	16	0.003	70.57	0.294	0.029	0.024	1.1	0.048	0.003
34	SkipInit	SGD	False	2	16	0.0016	69.76	0.302	0.037	0.018	1.329	0.055	0.004
35	SkipInit	SGD	False	2	16	0.001	69.72	0.303	0.044	0.016	1.493	0.06	0.006
36	SkipInit	SGD	False	2	12	0.01	70.83	0.289	0.022	0.022	0.807	0.036	0.001
37	SkipInit	SGD	False	2	12	0.006	70.12	0.297	0.025	0.023	0.936	0.041	0.002
38	SkipInit	SGD	False	2	12	0.003	69.94	0.3	0.03	0.022	1.161	0.048	0.003
39	SkipInit	SGD	False	2	12	0.0016	69.08	0.309	0.037	0.019	1.39	0.057	0.004
40	SkipInit	SGD	False	2	12	0.001	69.0	0.31	0.044	0.017	1.56	0.062	0.006
41	SkipInit	SGD	False	2	8	0.001	69.22	0.308	0.048	0.013	1.578	0.059	0.005
42	SkipInit	SGD	False	2	8	0.01	71.33	0.284	0.025	0.013	0.846	0.034	0.001
43	SkipInit	SGD	False	2	8	0.0016	69.98	0.3	0.04	0.015	1.405	0.052	0.004
44	SkipInit	SGD	False	2	8	0.006	71.02	0.289	0.027	0.016	0.97	0.038	0.002
45	SkipInit	SGD	False	2	8	0.003	70.74	0.291	0.033	0.016	1.186	0.046	0.003

Table 4: Results from the control group.

Exp. No.	Arch.	Optim.	Aug.	Depth	Width	LR	Test acc.(%)	Gen. error	SOTL	SOTL -E_50	TSE -EMA	PAC-Bayes sharpness	Value sens.
1	SkipInit	SGD	True	4	16	0.01	87.53	0.121	0.029	0.005	0.886	0.029	0.001
2	SkipInit	SGD	True	4	16	0.006	86.65	0.13	0.032	0.005	0.998	0.033	0.001
3	SkipInit	SGD	True	4	16	0.003	85.51	0.142	0.038	0.005	1.185	0.038	0.002
4	SkipInit	SGD	True	4	16	0.0016	84.91	0.147	0.047	0.005	1.377	0.042	0.003
5	SkipInit	SGD	True	4	16	0.001	83.68	0.157	0.055	0.006	1.524	0.048	0.004
6	SkipInit	SGD	True	4	12	0.01	88.06	0.116	0.032	0.005	0.931	0.028	0.001
7	SkipInit	SGD	True	4	12	0.006	87.71	0.12	0.036	0.005	1.056	0.032	0.001
8	SkipInit	SGD	True	4	12	0.003	86.93	0.128	0.043	0.005	1.252	0.036	0.002
9	SkipInit	SGD	True	4	12	0.0016	85.07	0.143	0.051	0.006	1.448	0.042	0.002
10	SkipInit	SGD	True	4	12	0.001	84.57	0.134	0.059	0.009	1.602	0.048	0.003
11	SkipInit	SGD	True	4	8	0.01	86.86	0.119	0.04	0.006	1.008	0.027	0.001
12	SkipInit	SGD	True	4	8	0.006	86.18	0.127	0.044	0.007	1.13	0.03	0.001
13	SkipInit	SGD	True	4	8	0.003	85.78	0.124	0.051	0.008	1.324	0.037	0.002
14	SkipInit	SGD	True	4	8	0.0016	83.65	0.124	0.06	0.011	1.52	0.041	0.003
15	SkipInit	SGD	True	4	8	0.001	83.02	0.1	0.068	0.015	1.669	0.046	0.003
16	SkipInit	SGD	True	3	16	0.01	87.04	0.126	0.03	0.005	0.896	0.029	0.001
17	SkipInit	SGD	True	3	16	0.006	86.6	0.131	0.033	0.005	1.017	0.032	0.001
18	SkipInit	SGD	True	3	16	0.003	85.94	0.137	0.04	0.005	1.21	0.037	0.002
19	SkipInit	SGD	True	3	16	0.0016	84.68	0.148	0.048	0.005	1.405	0.041	0.003
20	SkipInit	SGD	True	3	16	0.001	81.82	0.151	0.056	0.007	1.561	0.047	0.004
21	SkipInit	SGD	True	3	12	0.01	87.46	0.123	0.035	0.005	0.992	0.027	0.001
22	SkipInit	SGD	True	3	12	0.006	86.79	0.129	0.039	0.004	1.116	0.03	0.001
23	SkipInit	SGD	True	3	12	0.003	85.69	0.134	0.046	0.006	1.303	0.036	0.002
24	SkipInit	SGD	True	3	12	0.0016	83.79	0.142	0.056	0.008	1.486	0.042	0.003
25	SkipInit	SGD	True	3	12	0.001	83.04	0.134	0.064	0.011	1.622	0.048	0.004
26	SkipInit	SGD	True	3	8	0.01	85.12	0.133	0.043	0.008	1.065	0.027	0.001
27	SkipInit	SGD	True	3	8	0.006	83.01	0.124	0.047	0.008	1.18	0.032	0.001
28	SkipInit	SGD	True	3	8	0.003	83.72	0.132	0.054	0.01	1.355	0.038	0.002
29	SkipInit	SGD	True	3	8	0.0016	83.6	0.117	0.063	0.013	1.533	0.041	0.002
30	SkipInit	SGD	True	3	8	0.001	82.65	0.096	0.07	0.016	1.668	0.046	0.003
31	SkipInit	SGD	True	2	16	0.01	85.99	0.137	0.037	0.005	1.003	0.026	0.001
32	SkipInit	SGD	True	2	16	0.006	85.4	0.137	0.041	0.005	1.126	0.03	0.001
33	SkipInit	SGD	True	2	16	0.003	84.66	0.143	0.048	0.006	1.306	0.036	0.002
34	SkipInit	SGD	True	2	16	0.0016	83.99	0.145	0.057	0.009	1.487	0.042	0.003
35	SkipInit	SGD	True	2	16	0.001	83.06	0.131	0.066	0.012	1.623	0.048	0.004
36	SkipInit	SGD	True	2	12	0.01	86.13	0.122	0.039	0.006	1.02	0.027	0.001
37	SkipInit	SGD	True	2	12	0.006	86.33	0.126	0.042	0.006	1.146	0.03	0.001
38	SkipInit	SGD	True	2	12	0.003	85.04	0.133	0.05	0.008	1.344	0.036	0.002
39	SkipInit	SGD	True	2	12	0.0016	84.97	0.116	0.059	0.011	1.538	0.041	0.002
40	SkipInit	SGD	True	2	12	0.001	82.64	0.112	0.067	0.014	1.685	0.046	0.003
41	SkipInit	SGD	True	2	8	0.0016	83.65	0.08	0.066	0.016	1.533	0.041	0.002
42	SkipInit	SGD	True	2	8	0.003	83.76	0.1	0.058	0.013	1.34	0.036	0.002
43	SkipInit	SGD	True	2	8	0.006	84.1	0.108	0.051	0.011	1.159	0.03	0.001
44	SkipInit	SGD	True	2	8	0.01	85.57	0.111	0.048	0.011	1.045	0.027	0.001
45	SkipInit	SGD	True	2	8	0.001	82.34	0.066	0.072	0.018	1.685	0.044	0.003

Table 5: Results from experimental group 1 with straightforward data augmentation.

Exp. No.	Arch.	Optim.	Aug.	Depth	Width	LR	Test acc.(%)	Gen. error	SOTL	SOTL -E_50	TSE -EMA	PAC-Bayes sharpness	Value sens.
1	SkipInit	SGD	Logit avg.	4	16	0.01	84.23	0.089	0.022	0.022	0.814	0.034	0.001
2	SkipInit	SGD	Logit avg.	4	16	0.006	83.21	0.079	0.025	0.025	0.928	0.038	0.002
3	SkipInit	SGD	Logit avg.	4	16	0.003	83.44	0.077	0.03	0.026	1.11	0.041	0.002
4	SkipInit	SGD	Logit avg.	4	16	0.0016	82.5	0.091	0.037	0.022	1.305	0.048	0.003
5	SkipInit	SGD	Logit avg.	4	16	0.001	81.77	0.094	0.043	0.02	1.461	0.052	0.004
6	SkipInit	SGD	Logit avg.	4	12	0.01	84.36	0.068	0.023	0.023	0.862	0.034	0.001
7	SkipInit	SGD	Logit avg.	4	12	0.006	83.88	0.091	0.027	0.026	0.976	0.038	0.001
8	SkipInit	SGD	Logit avg.	4	12	0.003	83.3	0.083	0.032	0.022	1.168	0.041	0.002
9	SkipInit	SGD	Logit avg.	4	12	0.0016	82.68	0.096	0.04	0.02	1.37	0.046	0.003
10	SkipInit	SGD	Logit avg.	4	12	0.001	82.0	0.1	0.046	0.019	1.532	0.052	0.003
11	SkipInit	SGD	Logit avg.	4	8	0.01	83.22	0.083	0.026	0.026	0.916	0.032	0.001
12	SkipInit	SGD	Logit avg.	4	8	0.006	82.87	0.087	0.029	0.022	1.034	0.035	0.001
13	SkipInit	SGD	Logit avg.	4	8	0.003	82.93	0.094	0.036	0.019	1.235	0.039	0.002
14	SkipInit	SGD	Logit avg.	4	8	0.0016	81.77	0.099	0.045	0.019	1.442	0.046	0.003
15	SkipInit	SGD	Logit avg.	4	8	0.001	81.77	0.098	0.053	0.018	1.603	0.05	0.003
16	SkipInit	SGD	Logit avg.	3	16	0.01	84.19	0.081	0.022	0.022	0.833	0.034	0.001
17	SkipInit	SGD	Logit avg.	3	16	0.006	82.88	0.085	0.025	0.025	0.946	0.038	0.002
18	SkipInit	SGD	Logit avg.	3	16	0.003	82.93	0.091	0.031	0.024	1.137	0.041	0.002
19	SkipInit	SGD	Logit avg.	3	16	0.0016	81.77	0.103	0.038	0.02	1.339	0.047	0.003
20	SkipInit	SGD	Logit avg.	3	16	0.001	82.04	0.094	0.044	0.02	1.502	0.052	0.004
21	SkipInit	SGD	Logit avg.	3	12	0.01	83.79	0.075	0.025	0.025	0.915	0.033	0.001
22	SkipInit	SGD	Logit avg.	3	12	0.006	83.3	0.088	0.028	0.025	1.032	0.037	0.001
23	SkipInit	SGD	Logit avg.	3	12	0.003	83.27	0.086	0.035	0.022	1.223	0.041	0.002
24	SkipInit	SGD	Logit avg.	3	12	0.0016	81.75	0.097	0.043	0.019	1.411	0.044	0.003
25	SkipInit	SGD	Logit avg.	3	12	0.001	80.99	0.103	0.051	0.019	1.557	0.05	0.003
26	SkipInit	SGD	Logit avg.	3	8	0.01	82.12	0.093	0.028	0.023	0.97	0.031	0.001
27	SkipInit	SGD	Logit avg.	3	8	0.006	82.37	0.088	0.031	0.021	1.084	0.036	0.001
28	SkipInit	SGD	Logit avg.	3	8	0.003	82.83	0.09	0.038	0.019	1.268	0.039	0.002
29	SkipInit	SGD	Logit avg.	3	8	0.0016	81.54	0.099	0.047	0.018	1.457	0.044	0.003
30	SkipInit	SGD	Logit avg.	3	8	0.001	79.51	0.082	0.055	0.019	1.603	0.047	0.003
31	SkipInit	SGD	Logit avg.	2	16	0.01	82.41	0.09	0.026	0.026	0.921	0.033	0.001
32	SkipInit	SGD	Logit avg.	2	16	0.006	82.1	0.095	0.03	0.023	1.043	0.036	0.002
33	SkipInit	SGD	Logit avg.	2	16	0.003	82.07	0.093	0.036	0.022	1.233	0.041	0.002
34	SkipInit	SGD	Logit avg.	2	16	0.0016	82.03	0.093	0.044	0.02	1.421	0.044	0.003
35	SkipInit	SGD	Logit avg.	2	16	0.001	81.21	0.105	0.052	0.018	1.564	0.048	0.003
36	SkipInit	SGD	Logit avg.	2	12	0.01	83.68	0.077	0.026	0.026	0.928	0.032	0.001
37	SkipInit	SGD	Logit avg.	2	12	0.006	83.07	0.081	0.029	0.023	1.054	0.035	0.002
38	SkipInit	SGD	Logit avg.	2	12	0.0016	82.24	0.091	0.044	0.019	1.467	0.044	0.002
39	SkipInit	SGD	Logit avg.	2	12	0.001	81.99	0.093	0.052	0.018	1.625	0.048	0.003
40	SkipInit	SGD	Logit avg.	2	12	0.003	82.58	0.09	0.036	0.021	1.26	0.039	0.002
41	SkipInit	SGD	Logit avg.	2	8	0.0016	82.28	0.089	0.05	0.017	1.461	0.041	0.002
42	SkipInit	SGD	Logit avg.	2	8	0.01	84.07	0.073	0.028	0.021	0.951	0.03	0.001
43	SkipInit	SGD	Logit avg.	2	8	0.006	83.11	0.08	0.032	0.02	1.068	0.033	0.001
44	SkipInit	SGD	Logit avg.	2	8	0.003	83.61	0.083	0.04	0.018	1.259	0.038	0.002
45	SkipInit	SGD	Logit avg.	2	8	0.001	80.84	0.059	0.055	0.02	1.624	0.045	0.003

Table 6: Results from experimental group 2 with logit averaging.

Exp. No.	Arch.	Optim.	Aug.	Depth	Width	LR	Test acc.(%)	Gen. error	SOTL	SOTL -E_50	TSE -EMA	PAC-Bayes sharpness	Value sens.
1	SkipInit	SGLD	False	4	16	0.01	73.53	0.192	0.061	0.022	0.515	0.066	0.013
2	SkipInit	SGLD	False	4	16	0.006	83.14	0.165	0.017	0.01	0.358	0.064	0.009
3	SkipInit	SGLD	False	4	16	0.003	82.52	0.172	0.013	0.013	0.322	0.057	0.011
4	SkipInit	SGLD	False	4	16	0.0016	81.21	0.185	0.013	0.013	0.348	0.057	0.007
5	SkipInit	SGLD	False	4	16	0.001	80.52	0.192	0.014	0.014	0.389	0.034	0.001
6	SkipInit	SGLD	False	4	12	0.01	68.35	0.103	0.054	0.026	0.575	0.053	0.003
7	SkipInit	SGLD	False	4	12	0.006	81.73	0.18	0.019	0.01	0.416	0.067	0.006
8	SkipInit	SGLD	False	4	12	0.003	83.29	0.164	0.014	0.014	0.374	0.057	0.009
9	SkipInit	SGLD	False	4	12	0.0016	81.92	0.177	0.014	0.014	0.393	0.044	0.002
10	SkipInit	SGLD	False	4	12	0.001	80.97	0.188	0.015	0.015	0.437	0.029	0.001
11	SkipInit	SGLD	False	4	8	0.01	73.26	0.159	0.048	0.016	0.565	0.071	0.006
12	SkipInit	SGLD	False	4	8	0.006	81.28	0.184	0.022	0.01	0.419	0.061	0.008
13	SkipInit	SGLD	False	4	8	0.003	80.89	0.188	0.016	0.008	0.396	0.057	0.014
14	SkipInit	SGLD	False	4	8	0.0016	78.81	0.21	0.016	0.009	0.457	0.057	0.014
15	SkipInit	SGLD	False	4	8	0.001	78.4	0.213	0.018	0.011	0.526	0.041	0.015
16	SkipInit	SGLD	False	3	16	0.01	73.95	0.204	0.053	0.019	0.558	0.057	0.006
17	SkipInit	SGLD	False	3	16	0.006	82.78	0.169	0.017	0.01	0.381	0.062	0.012
18	SkipInit	SGLD	False	3	16	0.003	82.89	0.168	0.013	0.013	0.32	0.057	0.006
19	SkipInit	SGLD	False	3	16	0.0016	82.07	0.177	0.013	0.013	0.346	0.041	0.001
20	SkipInit	SGLD	False	3	16	0.001	80.52	0.192	0.014	0.014	0.39	0.036	0.0
21	SkipInit	SGLD	False	3	12	0.01	70.99	0.217	0.05	0.018	0.589	0.052	0.004
22	SkipInit	SGLD	False	3	12	0.006	80.77	0.189	0.021	0.01	0.452	0.061	0.01
23	SkipInit	SGLD	False	3	12	0.003	81.46	0.182	0.015	0.01	0.385	0.051	0.011
24	SkipInit	SGLD	False	3	12	0.0016	80.75	0.19	0.015	0.015	0.405	0.054	0.015
25	SkipInit	SGLD	False	3	12	0.001	79.52	0.202	0.016	0.016	0.462	0.03	0.002
26	SkipInit	SGLD	False	3	8	0.01	73.67	0.231	0.048	0.017	0.585	0.059	0.005
27	SkipInit	SGLD	False	3	8	0.006	79.24	0.205	0.021	0.01	0.455	0.063	0.008
28	SkipInit	SGLD	False	3	8	0.003	78.66	0.211	0.017	0.007	0.437	0.048	0.006
29	SkipInit	SGLD	False	3	8	0.0016	77.45	0.223	0.017	0.009	0.49	0.035	0.001
30	SkipInit	SGLD	False	3	8	0.001	76.42	0.234	0.018	0.012	0.569	0.027	0.0
31	SkipInit	SGLD	False	2	16	0.01	71.42	0.201	0.048	0.016	0.573	0.064	0.006
32	SkipInit	SGLD	False	2	16	0.006	81.4	0.183	0.016	0.011	0.403	0.064	0.011
33	SkipInit	SGLD	False	2	16	0.003	80.97	0.188	0.013	0.013	0.346	0.041	0.002
34	SkipInit	SGLD	False	2	16	0.0016	79.98	0.197	0.014	0.014	0.372	0.026	0.001
35	SkipInit	SGLD	False	2	16	0.001	79.48	0.202	0.015	0.015	0.432	0.022	0.0
36	SkipInit	SGLD	False	2	12	0.01	63.6	0.04	0.049	0.018	0.535	0.036	0.003
37	SkipInit	SGLD	False	2	12	0.006	80.03	0.197	0.019	0.01	0.415	0.057	0.007
38	SkipInit	SGLD	False	2	12	0.003	79.93	0.197	0.014	0.014	0.375	0.036	0.005
39	SkipInit	SGLD	False	2	12	0.0016	80.06	0.196	0.015	0.015	0.419	0.024	0.001
40	SkipInit	SGLD	False	2	12	0.001	78.9	0.209	0.016	0.016	0.498	0.024	0.0
41	SkipInit	SGLD	False	2	8	0.006	78.55	0.211	0.024	0.01	0.473	0.048	0.012
42	SkipInit	SGLD	False	2	8	0.003	78.14	0.216	0.017	0.008	0.465	0.044	0.005
43	SkipInit	SGLD	False	2	8	0.0016	76.7	0.23	0.018	0.009	0.533	0.028	0.001
44	SkipInit	SGLD	False	2	8	0.001	76.34	0.234	0.02	0.013	0.62	0.026	0.001
45	SkipInit	SGLD	False	2	8	0.01	73.21	0.231	0.053	0.017	0.591	0.064	0.007

Table 7: Results from experimental group 3 with SGLD.

Exp. No.	Arch.	Optim.	Aug.	Depth	Width	LR	Test acc.(%)	Gen. error	SOTL	SOTL -E_50	TSE -EMA	PAC-Bayes sharpness	Value sens.
1	MLP	SGD	False	4	16	0.01	50.55	0.492	0.034	0.007	1.06	0.032	0.001
2	MLP	SGD	False	4	16	0.006	49.83	0.5	0.037	0.008	1.182	0.037	0.001
3	MLP	SGD	False	4	16	0.003	49.77	0.501	0.042	0.011	1.399	0.044	0.003
4	MLP	SGD	False	4	16	0.0016	48.77	0.511	0.05	0.01	1.636	0.052	0.004
5	MLP	SGD	False	4	16	0.001	47.39	0.525	0.059	0.009	1.821	0.059	0.006
6	MLP	SGD	False	4	12	0.01	49.24	0.505	0.04	0.008	1.105	0.03	0.001
7	MLP	SGD	False	4	12	0.006	48.42	0.514	0.042	0.009	1.206	0.036	0.001
8	MLP	SGD	False	4	12	0.003	48.52	0.514	0.047	0.011	1.387	0.044	0.003
9	MLP	SGD	False	4	12	0.0016	47.22	0.527	0.056	0.01	1.592	0.052	0.004
10	MLP	SGD	False	4	12	0.001	46.33	0.536	0.065	0.009	1.758	0.057	0.005
11	MLP	SGD	False	4	8	0.01	45.61	0.5	0.056	0.012	1.228	0.032	0.001
12	MLP	SGD	False	4	8	0.006	45.19	0.497	0.057	0.011	1.327	0.036	0.002
13	MLP	SGD	False	4	8	0.003	44.4	0.516	0.062	0.012	1.504	0.044	0.003
14	MLP	SGD	False	4	8	0.0016	44.62	0.497	0.07	0.015	1.709	0.055	0.004
15	MLP	SGD	False	4	8	0.001	44.34	0.474	0.076	0.017	1.87	0.06	0.005
16	MLP	SGD	False	3	16	0.01	50.3	0.495	0.033	0.009	0.996	0.032	0.001
17	MLP	SGD	False	3	16	0.006	51.29	0.486	0.034	0.011	1.09	0.038	0.001
18	MLP	SGD	False	3	16	0.003	49.67	0.502	0.04	0.012	1.27	0.046	0.003
19	MLP	SGD	False	3	16	0.0016	48.78	0.512	0.048	0.01	1.469	0.052	0.004
20	MLP	SGD	False	3	16	0.001	47.68	0.523	0.056	0.009	1.63	0.057	0.005
21	MLP	SGD	False	3	12	0.01	48.29	0.515	0.04	0.008	1.063	0.03	0.001
22	MLP	SGD	False	3	12	0.006	49.76	0.502	0.04	0.01	1.147	0.036	0.001
23	MLP	SGD	False	3	12	0.003	47.7	0.522	0.046	0.011	1.308	0.044	0.003
24	MLP	SGD	False	3	12	0.0016	46.79	0.532	0.054	0.009	1.485	0.052	0.004
25	MLP	SGD	False	3	12	0.001	45.94	0.54	0.062	0.008	1.633	0.057	0.005
26	MLP	SGD	False	3	8	0.01	45.68	0.484	0.056	0.013	1.184	0.032	0.001
27	MLP	SGD	False	3	8	0.006	45.35	0.503	0.057	0.011	1.263	0.036	0.002
28	MLP	SGD	False	3	8	0.003	44.94	0.542	0.061	0.012	1.404	0.044	0.003
29	MLP	SGD	False	3	8	0.0016	45.05	0.503	0.068	0.014	1.572	0.052	0.004
30	MLP	SGD	False	3	8	0.001	45.92	0.44	0.075	0.017	1.715	0.059	0.005
31	MLP	SGD	False	2	16	0.01	50.71	0.492	0.032	0.011	0.96	0.032	0.001
32	MLP	SGD	False	2	16	0.006	50.01	0.499	0.034	0.012	1.048	0.038	0.002
33	MLP	SGD	False	2	16	0.003	49.98	0.5	0.039	0.011	1.205	0.048	0.003
34	MLP	SGD	False	2	16	0.0016	48.8	0.512	0.047	0.009	1.377	0.053	0.004
35	MLP	SGD	False	2	16	0.001	47.76	0.522	0.055	0.008	1.516	0.059	0.005
36	MLP	SGD	False	2	12	0.01	49.4	0.505	0.038	0.012	1.036	0.032	0.001
37	MLP	SGD	False	2	12	0.006	49.12	0.508	0.039	0.012	1.111	0.038	0.002
38	MLP	SGD	False	2	12	0.003	47.34	0.526	0.045	0.01	1.251	0.046	0.003
39	MLP	SGD	False	2	12	0.0016	47.37	0.526	0.053	0.009	1.407	0.052	0.004
40	MLP	SGD	False	2	12	0.001	46.32	0.523	0.062	0.008	1.54	0.06	0.005
41	MLP	SGD	False	2	8	0.003	44.0	0.549	0.06	0.011	1.332	0.044	0.003
42	MLP	SGD	False	2	8	0.006	44.9	0.544	0.056	0.011	1.213	0.037	0.002
43	MLP	SGD	False	2	8	0.01	44.74	0.478	0.056	0.013	1.151	0.032	0.001
44	MLP	SGD	False	2	8	0.001	44.4	0.449	0.074	0.017	1.586	0.057	0.005
45	MLP	SGD	False	2	8	0.0016	44.79	0.499	0.067	0.014	1.47	0.053	0.004

Table 8: Results from experimental group 4 with MLPs.

Exp. No.	Arch.	Optim.	Aug.	Depth	Width	LR	Test acc.(%)	Gen. error	SOTL	SOTL -E_50	TSE -EMA	PAC-Bayes sharpness	Value sens.
1	ResNet	SGD	False	4	16	0.01	76.53	0.233	0.013	0.013	0.365	0.082	0.013
2	ResNet	SGD	False	4	16	0.006	74.71	0.252	0.015	0.015	0.447	0.085	0.016
3	ResNet	SGD	False	4	16	0.003	70.5	0.294	0.019	0.019	0.599	0.09	0.018
4	ResNet	SGD	False	4	16	0.0016	68.98	0.31	0.023	0.022	0.777	0.097	0.018
5	ResNet	SGD	False	4	16	0.001	65.84	0.341	0.028	0.018	0.939	0.1	0.018
6	ResNet	SGD	False	4	12	0.01	75.94	0.238	0.015	0.015	0.417	0.079	0.013
7	ResNet	SGD	False	4	12	0.006	73.62	0.262	0.017	0.017	0.505	0.085	0.015
8	ResNet	SGD	False	4	12	0.003	70.56	0.293	0.021	0.021	0.672	0.088	0.017
9	ResNet	SGD	False	4	12	0.0016	68.78	0.312	0.026	0.018	0.869	0.094	0.017
10	ResNet	SGD	False	4	12	0.001	66.43	0.335	0.031	0.015	1.033	0.095	0.018
11	ResNet	SGD	False	4	8	0.01	74.87	0.249	0.018	0.018	0.538	0.072	0.011
12	ResNet	SGD	False	4	8	0.006	72.33	0.275	0.02	0.019	0.641	0.082	0.013
13	ResNet	SGD	False	4	8	0.003	69.55	0.304	0.026	0.016	0.846	0.088	0.013
14	ResNet	SGD	False	4	8	0.0016	67.4	0.325	0.033	0.012	1.055	0.093	0.014
15	ResNet	SGD	False	4	8	0.001	65.91	0.341	0.04	0.011	1.21	0.095	0.015
16	ResNet	SGD	False	3	16	0.01	75.65	0.242	0.014	0.014	0.378	0.078	0.013
17	ResNet	SGD	False	3	16	0.006	74.74	0.251	0.016	0.016	0.456	0.082	0.014
18	ResNet	SGD	False	3	16	0.003	69.98	0.299	0.019	0.019	0.61	0.089	0.016
19	ResNet	SGD	False	3	16	0.0016	69.13	0.308	0.024	0.02	0.808	0.09	0.016
20	ResNet	SGD	False	3	16	0.001	67.46	0.325	0.029	0.017	0.969	0.095	0.016
21	ResNet	SGD	False	3	12	0.01	74.69	0.251	0.016	0.016	0.461	0.072	0.013
22	ResNet	SGD	False	3	12	0.006	74.06	0.258	0.018	0.018	0.555	0.079	0.014
23	ResNet	SGD	False	3	12	0.003	70.44	0.294	0.022	0.02	0.741	0.088	0.016
24	ResNet	SGD	False	3	12	0.0016	66.83	0.331	0.028	0.015	0.947	0.093	0.017
25	ResNet	SGD	False	3	12	0.001	65.19	0.348	0.034	0.014	1.105	0.094	0.018
26	ResNet	SGD	False	3	8	0.01	75.35	0.245	0.018	0.018	0.561	0.073	0.011
27	ResNet	SGD	False	3	8	0.006	72.54	0.273	0.021	0.017	0.68	0.08	0.013
28	ResNet	SGD	False	3	8	0.003	69.36	0.305	0.027	0.014	0.878	0.082	0.013
29	ResNet	SGD	False	3	8	0.0016	67.96	0.32	0.034	0.012	1.077	0.086	0.014
30	ResNet	SGD	False	3	8	0.001	66.16	0.338	0.042	0.011	1.224	0.092	0.015
31	ResNet	SGD	False	2	16	0.01	76.45	0.235	0.014	0.014	0.404	0.067	0.009
32	ResNet	SGD	False	2	16	0.006	74.29	0.256	0.016	0.016	0.494	0.073	0.01
33	ResNet	SGD	False	2	16	0.003	71.99	0.28	0.02	0.02	0.665	0.082	0.011
34	ResNet	SGD	False	2	16	0.0016	69.86	0.301	0.026	0.017	0.873	0.082	0.012
35	ResNet	SGD	False	2	16	0.001	68.44	0.315	0.032	0.014	1.037	0.088	0.012
36	ResNet	SGD	False	2	12	0.01	76.31	0.235	0.017	0.017	0.487	0.067	0.008
37	ResNet	SGD	False	2	12	0.006	74.65	0.253	0.019	0.019	0.596	0.07	0.008
38	ResNet	SGD	False	2	12	0.003	72.81	0.271	0.024	0.016	0.79	0.076	0.01
39	ResNet	SGD	False	2	12	0.0016	70.22	0.298	0.031	0.013	0.997	0.082	0.01
40	ResNet	SGD	False	2	12	0.001	69.27	0.307	0.038	0.011	1.15	0.083	0.011
41	ResNet	SGD	False	2	8	0.01	74.99	0.248	0.021	0.016	0.643	0.064	0.007
42	ResNet	SGD	False	2	8	0.006	73.18	0.267	0.024	0.014	0.763	0.073	0.008
43	ResNet	SGD	False	2	8	0.003	70.05	0.299	0.031	0.011	0.99	0.076	0.01
44	ResNet	SGD	False	2	8	0.0016	68.64	0.314	0.04	0.01	1.193	0.076	0.011
45	ResNet	SGD	False	2	8	0.001	67.76	0.322	0.049	0.009	1.332	0.082	0.012

Table 9: Results from experimental group 5 with ResNets.