A Minimal and Flexible Model for Investigating Critical Learning Periods

Sebastian Lee^{1,2}, Stefano Sarao Mannelli^{2}, and Andrew Saxe^{2};
sal14@ic.ac.uk, s.saraomannelli@ucl.ac.uk, a.saxe@ucl.ac.uk.
^{1}Imperial College London (UK), ^{2}University College London (UK).

Abstract

We present a mathematically solvable model of critical learning periods using the teacher-student setup from statistical physics. By separating training into 'disrupted' and 'true' learning regimes, we model several possible learning perturbations. Preliminary results in this model provide evidence of critical learning periods resembling those observed across neuroscience and deep learning, thereby laying the foundations for a theory of critical learning periods.

Keywords: critical learning periods; deep learning theory; statistical physics

Introduction

There is a large body of evidence across neuroscience that disruption to learning in early stages of development can have catastrophic long-term effects on cognitive abilities in animals Hubel & Wiesel (1970); Fine et al. (2003); Popescu & Polley (2010). These early phases have been termed *critical learning periods*, or simply *critical periods* (CPs) Hensch (2004, 2005); Cisneros-Franco et al. (2020); Maurer et al. (2007). Likewise, recent work has identified similar periods in training of artificial neural networks via stochastic gradient descent Achille et al. (2017); Kleinman et al. (2023). On a biological level, these CPs are broadly understood to be driven by strong E/I imbalance, which allows for greater sensitivity to sensory-evoked activity Fagiolini et al. (2004); Clopath et al. (2016); as this imbalance lessens during development, cortical representations stabilise Dorrn et al. (2010). On the other hand, despite some candidate theories, e.g. around information plasticity measures in deep learning, the computational principles underlying these CPs remain poorly understood. In this work we provide a different perspective on the problem through the language of learning dynamics in the tradition of statistical physics.

Deep neural networks might exhibit behaviour reminiscent of critical periods through a variety of mechanisms. One view of CPs suggests that early periods of heightened plasticity are—while sometimes activity dependent or tunable by the environment Fagiolini et al. (1994); de Villers-Sidani et al. (2008); Greifzu et al. (2014)—fundamentally genetically encoded, and once a genetic switch is flipped, the rules governing plasticity change permanently. This view has been likened to 'pre-training' protocols in deep learning systems, in which an initial unsupervised pre-training phase is followed by a supervised fine-tuning phase, such that the learning objective fundamentally changes at a predetermined point Saxe (2013); Zaadnoordijk et al. (2022). A second possibility revolves around the empirical observation of sleeper effects, in which disruption during early sensitive periods results in deficits which emerge only much later in life Maurer et al. (2007); Zeanah et al. (2011). These data are hard to reconcile with a change in learning objective, and instead might arise from the disruption placing network parameters into a different basin of attraction, ultimately yielding a different solution under the dynamics of a single learning algorithm. Furthermore, CPs have been observed across different sensory modalities Wiesel & Hubel (1963); Schreiner & Polley (2014), over a range of developmental timeframes de Villers-Sidani & Merzenich (2011), and studies involving manipulation of neuromodulatory populations (thought to play a role in regulating CPs) suggest the biological mechanisms governing CPs are also variegated Bear & Daniels (1983); Shepard et al. (2015). One aim of our model is to provide a framework in which to study the full gamut of phenomena that might fall under the umbrella definition of CPs.

Teacher-Student Framework for CP

We model critical learning periods using the teacher-student setup Zdeborová & Krzakala (2016), commonly used in analysing learning dynamics of deep neural networks in a range of settings Lee et al. (2022); Gerace et al. (2022); Patel et al. (2023); Margiotta et al. (2024). The teacher-student setup is a generative model used to construct a task—i.e. a training set with inputs $\mathbf{x}$ and labels y^{*}—that is precisely parameterisd and it is amenable of mathematical analysis. In this setup, randomly generated input x are given to a *teacher* that provides the correct label acting as an oracle, then the *student*'s goal is to match the label of the teacher by learning over several input-output examples.

More precisely, in the vanilla teacher-student, an input $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_N)$ is fed into a fixed teacher neural network—parameterised by an input-hidden weight matrix $\mathbf{W}^{*}$ and hidden-output weight vector $\mathbf{h}^{*}$—that generates the target label $y^{*} = \phi(\mathbf{x}; \mathbf{W}^{*}, \mathbf{h}^{*})$, where ϕ denotes the two-layer neural network function with sigmoidal non-linearity on the hidden layer. Likewise, the student is represented by and a trainable student neural network—parameterised by $(\mathbf{W},\ \mathbf{h})$—trained via stochastic gradient descent on the squared loss between the target y^{*} and the student's output $\phi(\mathbf{x}; \mathbf{W}, \mathbf{h})$. The quantity of interest for our study of these networks is the generalisation error defined by $\varepsilon_g = \langle (y^{*} - \phi(\mathbf{x}; \mathbf{W}, \mathbf{h}))^2 \rangle$ where the average is taken over the input distribution.

Several variations of early learning disruption can be modelled in this setting by modifying the training protocol for some time before proceeding with training as described above. In general we will refer to these two training phases as the *disrupted* and *true* regimes (denoted by $\sim$ and $*$ respectively);

we can on to investigate whether this disrupted phase yields the hallmarks of a CP. In particular:

Perturbed Teachers. Similarly to the approach of Lee et al. (2021) for continual learning modelling, consider two distinct teachers—in the two regimes—where the *disrupted teacher* is perturbation of the true one. The degree of disruption is controlled by the 'overlap' γ between the teachers such that $\tilde{\mathbf{W}} = \gamma\mathbf{W}^* + \sqrt{1-\gamma^2}\mathbf{Z}$, where Z_{ij} are iid Gaussian entries.

Frozen Units. A direct way to disrupt learning in the student is to freeze weights into some number of hidden units, i.e. gradients are not propagated back through the weights into some number of hidden units. Disruptions to the network that are not specifically in the stimulus or readout could be used to model equivalent deficiencies such as conductive hearing loss caused by ear infections Stephenson et al. (1995).

In this abstract we will focus on these specific kinds of perturbation to showcase the framework. However, our framework allows manipulations including different data (i.e. label and/or input) perturbations that closely represent experiments including de Villers-Sidani et al. (2008); Stephenson et al. (1995); Hubel & Wiesel (1970).

Noisy Inputs. We can also inject noise into the inputs observed by the students, i.e. the student output will be given by $\phi(\mathbf{x}+\sigma\mathbf{z};\mathbf{W},\mathbf{h})$, where $\mathbf{z}\in\mathbb{R}^N$ is a vector of samples from the standard normal, and σ controls the strength of disruption. Modifying the inputs in this way resembles studies with unstructured stimulus noise performed in rodent audition experiments de Villers-Sidani et al. (2008).

Frozen Units. Arguably the most direct way to disrupt learning in the student is to freeze weights into some number of hidden units, i.e. gradients are not propagated back through the weights into some number of hidden units. Disruptions to the network that are not specifically in the stimulus or readout could for instance be used to model equivalent deficiencies such as conductive hearing loss caused by ear infections Stephenson et al. (1995).

In the limit of large input dimension, it is possible to exactly characterise the dynamics of the generalisation error. For each of the modified settings described above we are able to extend that analysis and derive ordinary differential equations that exactly describe the dynamics of the student generalisation error during and after the disrupted period of learning. For space constraints we refer to Saad & Solla (1995) for details.

Results

Having outlined our general approach for analysing critical period using the teacher-student setup, we describe below early results we have obtained in this direction.

Teacher Perturbations Induce Saddles

If early on in learning a student network is trained on a slightly perturbed teacher, it severely hinders performance when switching to the true underlying task i.e. when the disruption ceases, as shown in Fig. 1. Although it is well known that for sigmoidal activations there is a tendency for student nodes to 'specialise' to the teacher nodes Goldt et al. (2019), the rate and degree to which this specialisation occurs is not consistent (see discussion below). If for instance the disrupted period of learning does not lead to full convergence on the perturbed teacher, the student may not have fully specialised. When learning proceeds on the true task, this lower entropy distribution of hidden layer activity leads to a significant slow down in learning, giving rise to a critical learning period.

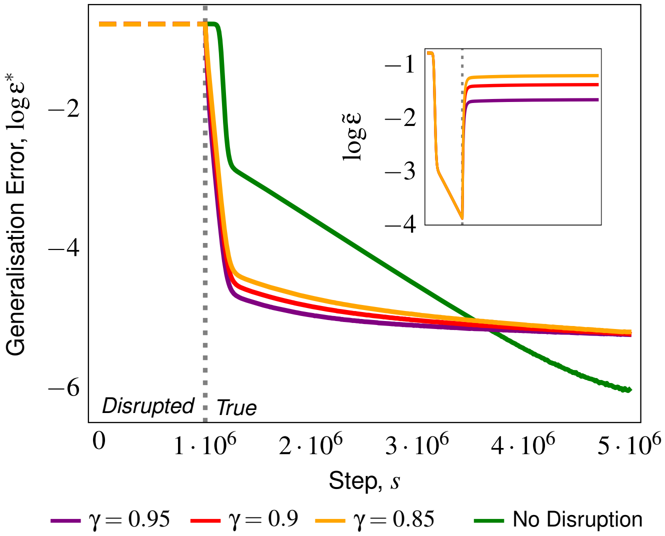

Figure 1: **Perturbed Teachers Reveal Critical Learning Period.** Students trained on perturbations ($\gamma=0.85, 0.9, 0.95$) of the *true* teacher for 1×10^6 steps before training for 4×10^6 steps on the true teacher all plateau significantly earlier than the student trained *ab initio* on the true teacher for the same time in the true regime (green line). Notice that the actual number of training steps for the 'no disrupted' line is shifted, while the other lines are learning a good approximation of the actual task from the first epoch. *(inset)* Generalisation error of the student on the perturbed teacher.

De-Noising Critical Periods

The second result we obtain is from the frozen node setup; here we focus on the importance of the length of disruption. Beyond some disruption time there is a collapse of trajectories in which units that are frozen in the impaired learning period fail to activate. The effective capacity of the network when learning the 'true' task is then the number of units that were unfrozen from the start. Note, this collapse is a function of the label noise: when the noise is increased beyond some non-zero threshold, the inactive nodes will also eventually activate when unfrozen. In the setting where a collapsed regime exists, we identify two types of critical learning periods:

1. A "de-noising" critical learning period: when the teacher is noisy, over-parameterisation can help to average out this noise and reach a better generalisation error that students that were disrupted and are operating with a lower capacity in the collapsed regime.

2. A “symmetry-breaking” critical learning period: the loss landscape is populated with saddle points, which largely govern the speed of learning. There is evidence to suggest that overparameterisation may help to: (a) reduce the number of saddle points in the landscape, (b) break the symmetries among hidden units that give rise to the saddle points.

Conclusions

Critical learning periods is a complex phenomenon that is observed in a variety of experiments, such complexity makes theoretical progresses hard. Here, we introduce a flexible and solvable framework able to reproduce features of critical learning periods in a minimal setting, paving the way to a deeper theoretical understanding of critical learning periods.

References

Achille, A., Rovere, M., & Soatto, S. (2017). Critical learning periods in deep neural networks. *arXiv preprint arXiv:1711.08856*.

Bear, M. F., & Daniels, J. (1983). The plastic response to monocular deprivation persists in kitten visual cortex after chronic depletion of norepinephrine. *Journal of Neuroscience*, *3*(2), 407–416.

Cisneros-Franco, J. M., Voss, P., Thomas, M. E., & de Villers-Sidani, E. (2020). Critical periods of brain development. In *Handbook of clinical neurology* (Vol. 173, pp. 75–88). Elsevier.

Clopath, C., Vogels, T. P., Froemke, R. C., & Sprekeler, H. (2016). Receptive field formation by interacting excitatory and inhibitory synaptic plasticity. *BioRxiv*, 066589.

de Villers-Sidani, E., & Merzenich, M. M. (2011). Lifelong plasticity in the rat auditory cortex: basic mechanisms and role of sensory experience. *Progress in brain research*, *191*, 119–131.

de Villers-Sidani, E., Simpson, K. L., Lu, Y., Lin, R. C., & Merzenich, M. M. (2008). Manipulating critical period closure across different sectors of the primary auditory cortex. *Nature neuroscience*, *11*(8), 957–965.

Dorrn, A. L., Yuan, K., Barker, A. J., Schreiner, C. E., & Froemke, R. C. (2010). Developmental sensory experience balances cortical excitation and inhibition. *Nature*, *465*(7300), 932–936.

Fagiolini, M., Fritschy, J.-M., Low, K., Mohler, H., Rudolph, U., & Hensch, T. K. (2004). Specific gabaa circuits for visual cortical plasticity. *Science*, *303*(5664), 1681–1683.

Fagiolini, M., Pizzorusso, T., Berardi, N., Domenici, L., & Maffei, L. (1994). Functional postnatal development of the rat primary visual cortex and the role of visual experience: dark rearing and monocular deprivation. *Vision research*, *34*(6), 709–720.

Fine, I., Wade, A. R., Brewer, A. A., May, M. G., Goodman, D. F., Boynton, G. M., . . . MacLeod, D. I. (2003). Long-term deprivation affects visual perception and cortex. *Nature neuroscience*, *6*(9), 915–916.

Gerace, F., Saglietti, L., Mannelli, S. S., Saxe, A., & Zdeborová, L. (2022). Probing transfer learning with a model of synthetic correlated datasets. *Machine Learning: Science and Technology*, *3*(1), 015030.

Goldt, S., Advani, M., Saxe, A. M., Krzakala, F., & Zdeborová, L. (2019). Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, *32*.

Greifzu, F., Pielecka-Fortuna, J., Kalogeraki, E., Krempler, K., Favaro, P. D., Schlüter, O. M., & Löwel, S. (2014). Environmental enrichment extends ocular dominance plasticity into adulthood and protects from stroke-induced impairments of plasticity. *Proceedings of the National Academy of Sciences*, *111*(3), 1150–1155.

Hensch, T. K. (2004). Critical period regulation. *Annu. Rev. Neurosci.*, *27*, 549–579.

Hensch, T. K. (2005). Critical period plasticity in local cortical circuits. *Nature Reviews Neuroscience*, *6*(11), 877–888.

Hubel, D. H., & Wiesel, T. N. (1970). The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *The Journal of physiology*, *206*(2), 419–436.

Kleinman, M., Achille, A., & Soatto, S. (2023). Critical learning periods for multisensory integration in deep networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 24296–24305).

Lee, S., Goldt, S., & Saxe, A. (2021). Continual learning in the teacher-student setup: Impact of task similarity. In *International conference on machine learning* (pp. 6109–6119).

Lee, S., Mannelli, S. S., Clopath, C., Goldt, S., & Saxe, A. (2022). Maslow’s hammer for catastrophic forgetting: Node re-use vs node activation. *arXiv preprint arXiv:2205.09029*.

Margiotta, R. G., Goldt, S., & Sanguinetti, G. (2024). Attacks on online learners: A teacher-student analysis. *Advances in Neural Information Processing Systems*, *36*.

Maurer, D., Mondloch, C. J., & Lewis, T. L. (2007). Sleeper effects. *Developmental Science*, *10*(1), 40–47.

Patel, N., Lee, S., Mannelli, S. S., Goldt, S., & Saxe, A. (2023). The rl perceptron: Generalisation dynamics of policy learning in high dimensions. *arXiv preprint arXiv:2306.10404*.

Popescu, M. V., & Polley, D. B. (2010). Monaural deprivation disrupts development of binaural selectivity in auditory midbrain and cortex. *Neuron*, *65*(5), 718–731.

Saad, D., & Solla, S. A. (1995). On-line learning in soft committee machines. *Physical Review E*, *52*(4), 4225.

Saxe, A. M. (2013). *Precis of deep linear neural networks: A theory of learning in the brain and mind.*

Schreiner, C. E., & Polley, D. B. (2014). Auditory map plasticity: diversity in causes and consequences. *Current Opinion in Neurobiology*, *24*, 143–156.

Shepard, K. N., Liles, L. C., Weinshenker, D., & Liu, R. C. (2015). Norepinephrine is necessary for experience-dependent plasticity in the developing mouse auditory cortex. *Journal of Neuroscience*, *35*(6), 2432–2437.

Stephenson, H., Higson, J., & Haggard, M. (1995). Binaural hearing in adults with histories of otitis media in childhood. *Audiology*, *34*(3), 113–123.

Wiesel, T. N., & Hubel, D. H. (1963). Effects of visual deprivation on morphology and physiology of cells in the cat's lateral geniculate body. *Journal of neurophysiology*, *26*(6), 978–993.

Zaadnoordijk, L., Besold, T. R., & Cusack, R. (2022). Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, *4*(6), 510–520.

Zdeborová, L., & Krzakala, F. (2016). Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, *65*(5), 453–552.

Zeanah, C. H., Gunnar, M. R., McCall, R. B., Kreppner, J. M., & Fox, N. A. (2011). Vi. sensitive periods. *Monographs of the society for research in child development*, *76*(4), 147–162.

Acknowledgements

The authors thank Alessandro Achille, Sebastian Goldt and Michael Kleinman, for several fruitful discussions. This work was supported by a Sir Henry Dale Fellowship from the Wellcome Trust and Royal Society (216386/Z/19/Z) to A.S., and the Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z) and the Gatsby Charitable Foundation (GAT3755). A.S. is a CIFAR Azrieli Global Scholar in the Learning in Machines & Brains program.