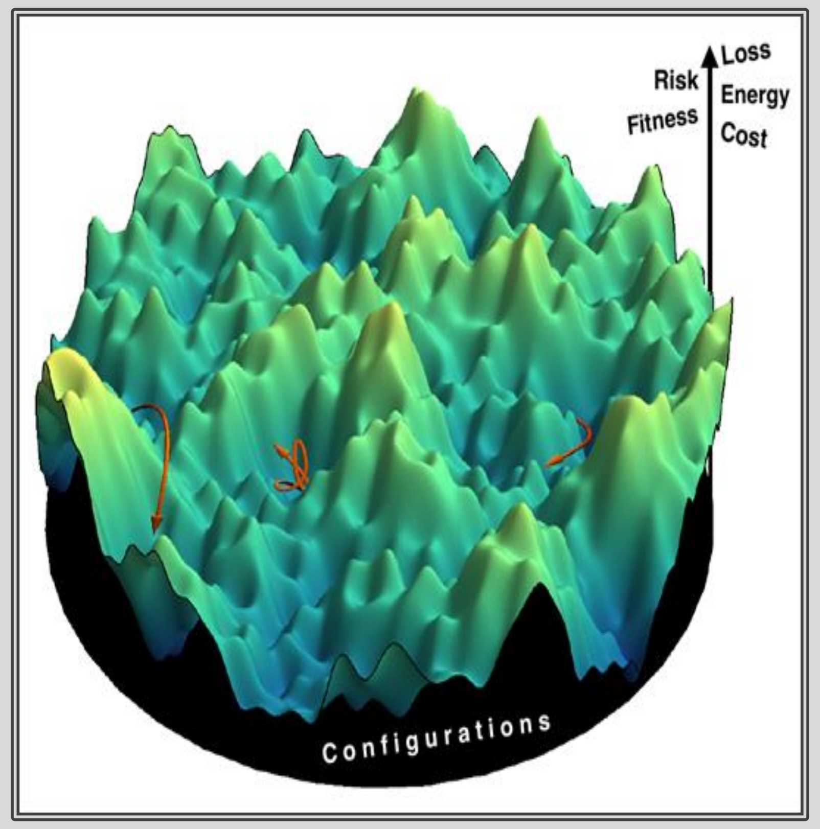
Swap Monte Carlo Methods in Deep Neural Network Training:

From Glassy Statistical Mechanics to Faster AI Learning

Juan Manuel Muñoz Hernández
King's College London

1. Original context: glassy energy landscape



- High-dimensional parameter spaces in disordered systems (spin-glass systems and DNNs alike) induce rough energy landscapes [1], which are nearly worst-case scenarios for optimisation.

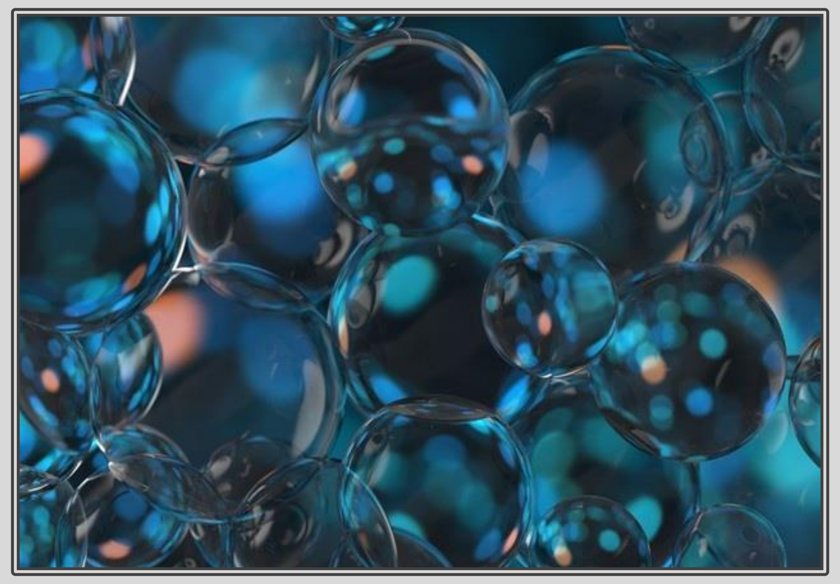
- Common consequent issues include slow convergence of stochastic gradient descent (SGD) methods and the presence of large sub-optimal basins.

- These traits are the cause of a class of properties known as 'glassy behaviour': permanent non-equilibrium, slow dynamics, aging, quenched disorder.

2. Swap Monte Carlo for glass simulation

- In simulated models of glassy materials, we can randomly swap bigger and smaller spheres during thermal relaxation, without affecting the final state [2].

- This effectively makes optimisation faster and less restricted to the shape of the landscape. Thermalisation times can be greatly shortened, depending on sphere size statistics (known as 'polydispersity') [3].



- The training stage of DNNs shares several characteristics with the dynamics of physical glassy systems [4]. If swap Monte Carlo can be successfully implemented in DNN training, then a new similarity is found and we can accelerate training; else, there is a significant difference to look into. However, as from the case of spheres it is not clear whether swapping should occur purely in parameter space or also permuting dynamic variables.

3. Swapping parameters

- Benchmarking DNN protocols typically involves image recognition systems. We have used a numerically normalised version of CIFAR-10.

- For parameter swapping, several levels are available: between different feature vectors, between colour channels of an image, or between pixels. Examples include "swap the values of all channels between two pixels of one image" and "swap two channels of all pixels of all images", thus greatly varying in swap intensity. This is equivalent to shuffling the system's disorder at a local, relatively low-dimensional scope, making new optimisation paths available and easing the process on average. Learning-wise, we are providing the system with slightly noisier training data, which is then reset so that simulated data corruption does not accumulate.

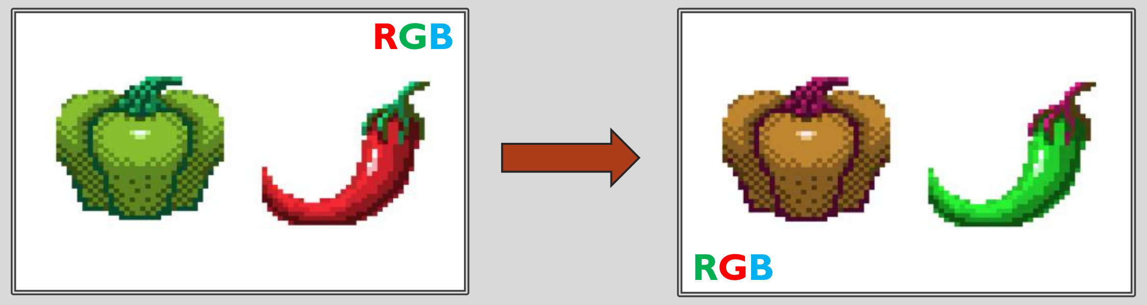


Figure A: visual representation of the 1SA protocol: one image is selected (1), from which two randomly selected channels are swapped (S), for all pixels (A) in the image.

4. Swapping weights and biases

- Swapping weights and biases effectively dashes dynamical trajectory along some dimensions, thus exploring nearby regions of the energy landscape that may not be accessible through standard SGD methods. In learning terms, we are affecting the 'memory' of the system, by re-arranging neuron information. The geometry of the loss function is not altered in this case.

- There are also sub-structures for swapping dynamic variables: in-row, in-column or fully mixed swapping, intra- or inter-layer permutation, mixing biases and weights.

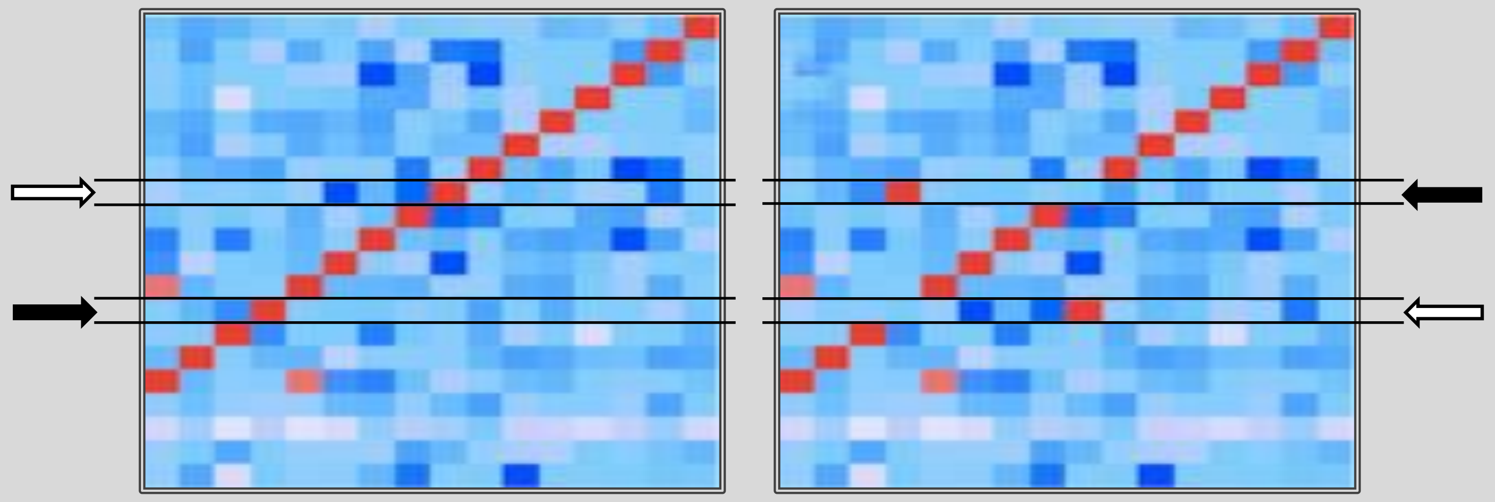
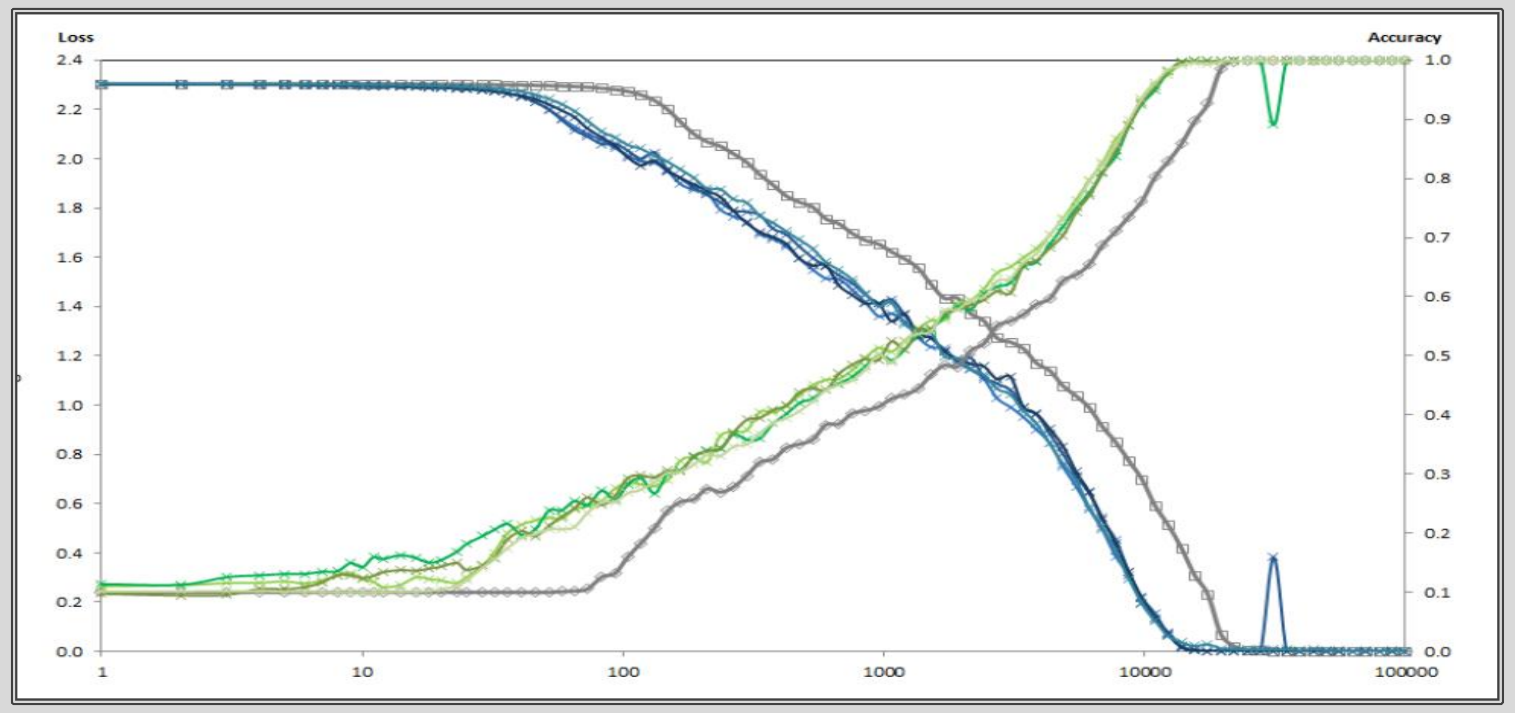
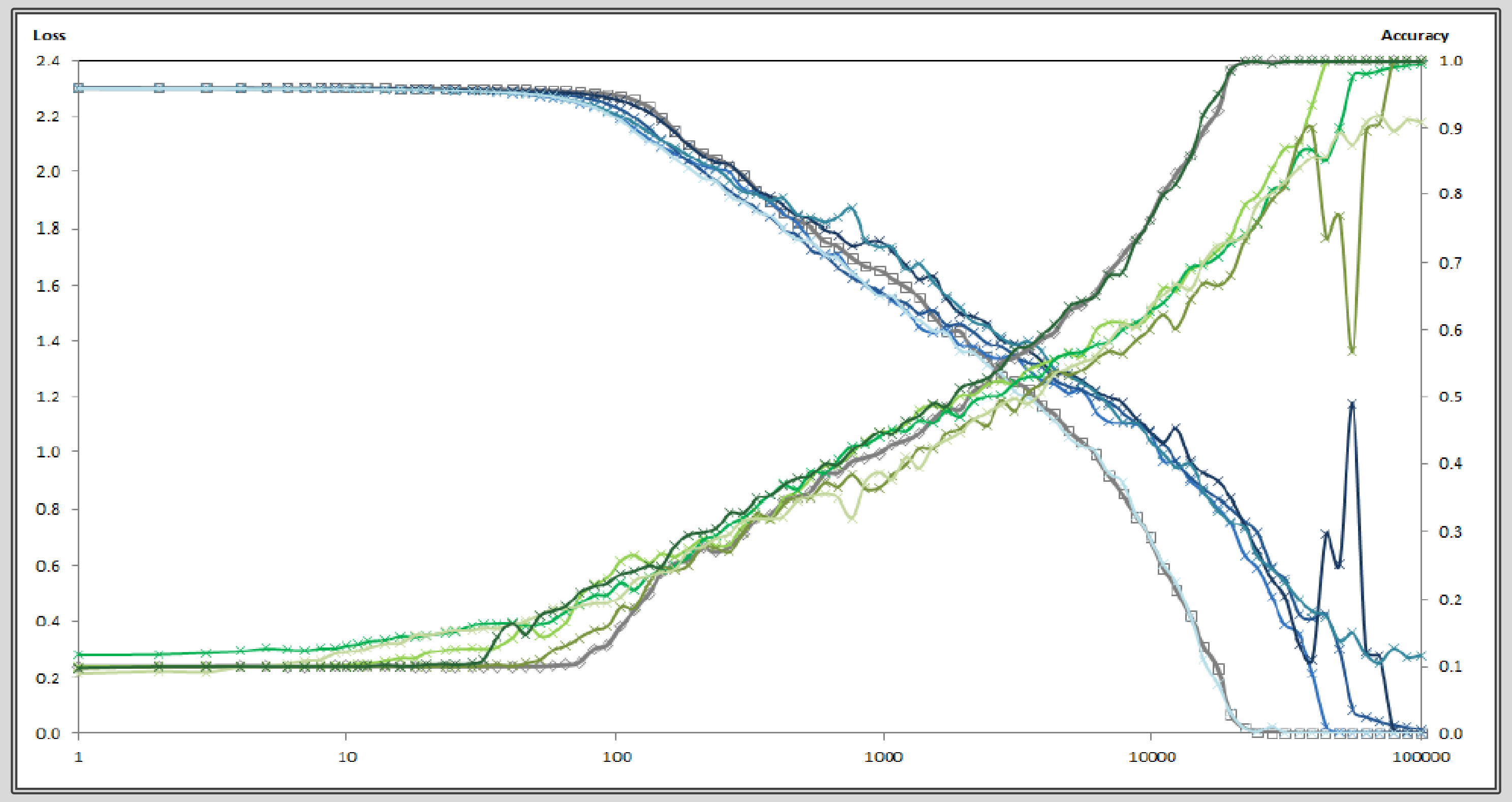


Figure B: visual representation of an all-column weight swap: the values of two randomly selected rows are swapped, for all columns in a given layer.

5. Conclusions

Feature swapping can accelerate training by as much of a factor of 2. Figure below: training loss (blue) and accuracy (green) over iteration time for several feature swapping protocols vs training without swapping (grey).





Weight swapping slows and impedes learning covariantly with swap intensity (figure above, analogous colour coding); however, swapped learning processes contain local information about the energy landscape.

References

- [1] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, Yann LeCun. The loss surfaces of multilayer networks. *Proceedings of Machine Learning Research*, 38: 192–204, 2015.
- [2] Tomás S. Grigera, Giorgio Parisi. Fast Monte Carlo algorithm for supercooled soft spheres. *Physical Review E, 63(4), 2001.*
- [3] Andrea Ninarello, Ludovic Berthier, Daniele Coslovich. Models and algorithms for the next generation of glass transition studies. *Physical Review X*, 7, 2017.
- [4] M. Baity-Jesi, M. Geiger, L. Sagun, S. Spigler, G. Ben Arous, C. Cammarota, Y. LeCun, M. Wyart and G. Biroli. Comparing dynamics: Deep neural networks versus glassy systems. *Proceedings of Machine Learning Research*, 80: 324–333, 2018.

Acknowledgements

This work constitutes the thesis of a Final Research Project for MSc Complex Systems Modelling at King's College London 2018 – 2019. Special thanks to prof. Chiara Cammarota for her counselling in regard to the scope of the work and the choice of the research topic.