Constructing artificial life and materials scientists with accelerated AI using Deep AndersoNN

Saleem Abdulfattah Ahmed Al Dajani and David E. Keyes

Applied Physics (AP) Program, Extreme Computing Research Center (ECRC), and KAUST Solar Center (KSC), Physical Science and Engineering (PSE) and Computer, Electrical, Mathematical Sciences and Engineering (CEMSE) Divisions, King Abdullah University of Science and Technology (KAUST), Thuwal, Makkah Province, Kingdom of Saudi Arabia (KSA) 23955-6900

Motivation

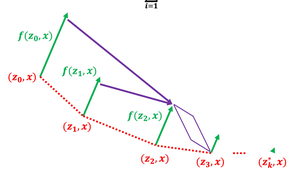
- **Leveraging ML and DFT for drug discovery** — Foundation models, large language models (LLMs), and multi-agent natural language societies of mind (NLSOMs) require significant computing resources to achieve practical accuracies with up to trillions of parameters using explicit neural networks. As the number of layers in a neural network approaches infinity, (Kolter, Duvenaud, Johnson; 2020) these models can be approximated with single-layer implicit models, known as deep equilibrium (DEQ) models. Solving for the parameters of a single implicit layer that takes both the input, x, and the output, y, as inputs are reduced to a fixed point iteration problem that is proven to converge to a deep equilibrium state with stable behavior under optimal hyperparameters where the fixed point converges. Vector-to-vector iterative solvers such as Anderson extrapolation (Anderson, 1965; 2019) can be employed to accelerate convergence (Al Dajani, Keyes; 2024), achieving similar performance similar to explicit networks while scaling neural networks and reducing the necessary compute resources.
- **Constructing artificial life and material scientists (ALMS)** — Drug discovery is at the intersection of life and materials science. Density functional theory (DFT) is at the forefront of materials modeling, yet suffers from computational limitations with high-atom biological systems needed in the life sciences. Responding to the COVID-19 pandemic required high-throughput, rapid methods to screen and discover drugs that could be pipelined into laboratory and, ultimately, clinical trials to treat different and evolving variants of the virus. Machine learning and scaling DFT enables high-throughput classification of candidate drugs based on their material properties, such as pore size for DNA/RNA capture and dipole moment for polarity.

State-of-the-art : Extrapolation

$$\mathbf{z_{k+1}} = \mathbf{f(z_k, x)}\,, \mathbf{z_k} \in \mathbb{R}^n\,, \mathbf{k = 0, 1, 2, \ldots}$$

$$\mathbf{r_k(x_k) = z_k - z(x_k), w_i^{(k)} = \frac{-r_{1k}^T(r_{ik} - r_{ik-1})}{\left\|r_{ik} - r_{ik-1}\right\|^2}}$$

$$\mathbf{z_{k+1} = f(z_k, x) + \sum_{i=1}^{min(k,m)} w_i^{(k)}(f(z_k, x) - f(z_{k-1}, x))}$$



$$z_3 = f(z_2) + w_1^{(2)}(f(z_2, x) - f(z_1, x)) + w_2^{(2)}(f(z_2, x) - f(z_0, x))$$

Mathematical formulation and vector representation. Adapted from Y. He & H. De Sterck (Copper Mountain, 2022. ICERM, 2023.)

Design of Fixed Point Acceleration

- Fixed point iteration $z^\star = f(z^\star, x)$
- Forward iteration $z^{k+1} = f(z^k, x)$
- Anderson acceleration $z^{k+1} = \sum_{i=1}^{m} \alpha_i f(z^{k-i+1}, x)$

$$\text{minimize}_\alpha \quad \|G\alpha\|_2^2, \quad \text{subject to} \quad 1^T\alpha = 1 \tag{1}$$

$$G = \left[\; f(z^k, x) - z^k, \quad \cdots \quad, (z^{k-m+1}, x) - z^{k-m+1} \;\right] \tag{2}$$

where ν is a Lagrange multiplier for the constraint $1^T\alpha = 1$ in the objective function: $L(\alpha, \nu) = \|G\alpha\|_2^2 - \nu(1^T\alpha - 1)$.

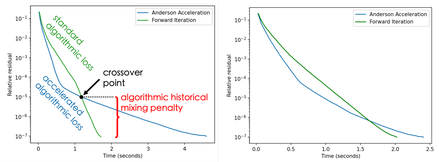
$$\begin{bmatrix} 0 & 1^T \\ 1 & H \end{bmatrix} \vec{y} = \begin{bmatrix} 0 & 1^T \\ 1 & G^TG + \lambda I \end{bmatrix} \begin{bmatrix} \nu \\ \alpha \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{3}$$

where λ is a regularization term, incorporating a mixing parameter $\beta > 0$ to balance between original and extrapolated iterates:

$$z^{k+1} = (1 - \beta)\sum_{i=1}^{m} \alpha_i z^{k-1+1} + \beta \sum_{i=1}^{m} \alpha_i f(z^{k-i+1}, x) \tag{4}$$

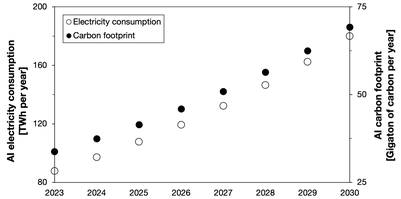
The function f, the forward pass through the implicit layer, ensures that despite the intermediate expansion of the channel depth to k_1 inner channels, the output tensor Z retains the same dimensions as the input tensor $X \in \mathbb{R}^{n \times d \times H \times W}$ by setting the number of outer channels k_2 to number of input and output channels, d

- Tuning trade-off between acceleration and accuracy with m



Evaluating relative residual, $\frac{\|f(z^k, x) - z^k\|_2}{\|f(z^k, x)+\lambda\|_2}$, for a random input x

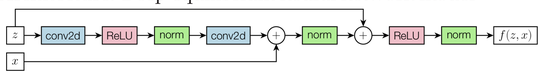
- Anderson extrapolation has fewer more expensive iterations



Anderson can save up to ∼30-88% of compute towards AI carbon footprint, projected to consume >2% of global electricity demand, >10% for data [3].

Learning Architecture & Hierarchy

- **Architecture:** Deep equilibrium neural network model



$$f(z, x) = \text{norm}(\text{ReLU}(z + \text{norm}(x + W_2 * (\text{norm}(\text{ReLU}(W_1 * z))))))$$

- **Forward pass:** Compute equilibrium point, logits, and loss

$$z^* = f(z^*, x), \quad \hat{y} = Az^* + d, \quad l(\hat{y}, y) = -\sum_i^n y_i \log \hat{y}_i$$

- **Backward pass:** Compute gradients by implicit function theorem

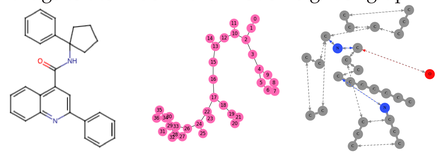
$$\frac{\partial l}{\partial \theta} = \frac{\partial l}{\partial z^*}(I - \frac{\partial f}{\partial z^*})^{-1}\frac{\partial f}{\partial \theta}$$

Learning Results

Algorithmic improvements to training and inference with augmentation vs. CIFAR10 benchmark without augmentation, accelerated over standard.

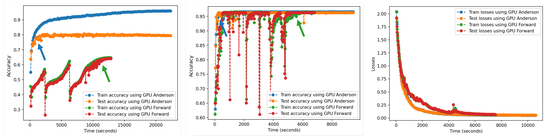
	Algorithm	**DEQ (ours)**	DEQ [Implicit, Bai et al. 2019]	ResNet-18 [**Explicit, He et al. 2016**]
Number of parameters	Standard CIFAR10	**64,842**	∼170,000	∼170,000
	Accelerated CIFAR10	**64,842**	-	-
	Standard ALMS	**64,842**	-	-
	Accelerated ALMS	**64,842**	-	-
Training accuracy	Standard CIFAR10	64.7%	-	-
	Accelerated CIFAR10	**87.6%**	-	-
	Standard ALMS	98.2%	-	-
	Accelerated ALMS	**98.4%**	-	-
Testing accuracy	Standard CIFAR10	64.2%	82.2%	81.6%
	Accelerated CIFAR10	**79.3%**	-	-
	Standard ALMS	97.9%	-	-
	Accelerated ALMS	**98.1%**	-	-
Training time [seconds]	Standard CIFAR10	1.2×10^4	-	-
	Accelerated CIFAR10	$\mathbf{1.4\times10^3}$	-	-
	Standard ALMS	2.3×10^3	-	-
	Accelerated ALMS	$\mathbf{5.3\times10^2}$	-	-
Inference time [seconds]	Standard	0.20	-	-
	Accelerated	**0.14**	-	-
Speedup relative to standard	Accelerated CIFAR10 ratio	**1.4-8.6**	-	-
	Accelerated ALMS ratio	**4.4-17.6**	-	-
	Compute saved	**77-94%**	-	-

- COVID drugs to atom-bond then node-neighbor graph images



Compound (left). Graphical representation (center). Node-neighbor (right).

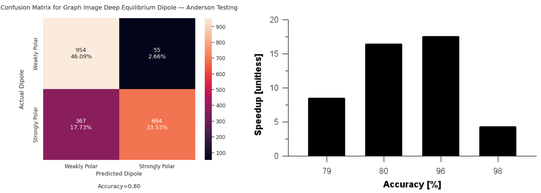
- Discovering drugs with artificial life and material scientists



Accelerating to fixed point deep equilibrium with accuracy near unity

Learning Analysis

- Anderson extrapolation has a higher cost per iteration, measured in function evaluations or epochs.



Deep equilibrium ∼1.4-17.6x faster to stable convergence with Anderson despite more expensive iterations due to instability of standard forward

- Anderson extrapolation exhibits less fluctuation in accuracy as seen in the test accuracy, whereas forward iteration shows more significant ups and downs in both training and testing accuracy, which suggest overfitting during training with forward iteration.
- Anderson acceleration reaches a higher accuracy plateau for both training and test datasets. This could indicate a better generalization capability when using Anderson acceleration.

Summary and Future Work

- This work shows that with accelerated deep equilibrium models, artificial life and materials scientists could be constructed for practical industry applications based on first principles theory.
- A speedup of up to an order of magnitude enables an order of magnitude larger models and/or 90% less computing resources for similar accuracies, paving the way for LLMs, NLSOMs, and foundation models for both training and running inferences.
- Future work: incorporate larger datasets to build multi-objective optimized models, LLMs, NLSOMs, and foundation models that make inferences for drug discovery and biocatalysis at scale, integrating life and materials science in a novel, unprecedented way.

Software Release and References

- Code and data found at https://tinyurl.com/Deep-AndersoNN

1 Anderson, D. G. (1965). Iterative procedures for nonlinear integral equations. *Journal of the ACM (JACM)*, 12(4), 547-560.
2 Anderson, D. G. (2019). Comments on "Anderson acceleration, mixing and extrapolation". *Numerical Algorithms*, 80, 135-234.
3 Al Dajani, Saleem A. A., and Keyes, David E., (2024). Accelerating AI Performance using Anderson Extrapolation on GPUs, (in review).
4 Kolter Z., Duvenaud D., and Johnson M. (2020) Deep Implicit Layers — Neural ODEs, Deep Equilibirum Models, and Beyond. *NeurIPS Tutorial*. Chapter 4. Available at: https://implicit-layers-tutorial.org/