Dimensionless Controls of Plasticity Under Alternating Tasks: From Evolutionary Biology to Continual Learning

Owen Skriloff OSKRILOFF@UCHICAGO.EDU
University of Chicago

Abstract

Plasticity under changing environments is central to both evolutionary biology and continual learning. Motivated by recent work on genotype–phenotype maps, we study a minimal deep-learning analogue where a network is trained alternately on two Boolean label sets, and ask which biological controls of plasticity survive the translation to gradient descent. Reinterpreting four proposed biological factors as quantities of training dynamics, we find the system reduces to two dimensionless controls: the *task disagreement* r, the fraction of disagreeing labels, and the *reach* ηT, the product of learning rate and switching period. We derive two bounds on plasticity: r alone fixes an extremal geometric floor on the utopia distance, while r and ηT jointly bound forgetting. Across 9,720 trajectories, an ANOVA confirms that r, η, and T dominate, while the effect of neutral-set size (emphasized in the biological setting) is negligible. The optimal reach itself follows an approximate inverse power law $\eta T^* \propto r^{-1.18}$, yielding a heuristic that sets the optimal reach ηT^* from the task disagreement alone. The analogy that survives is therefore dynamical rather than geometric, and our setting enables a view of plasticity through the lens of other driven systems in physics and engineering.

Keywords: Neural Network Plasticity, Continual Learning, Alternating Tasks, Dimensionless Analysis, Catastrophic Forgetting, Evolutionary Plasticity

1. Introduction

A longstanding question in the study of learning is how far universality extends across systems and substrates (Simon et al., 2026; van Rossem and Saxe, 2024; Huh et al., 2024; Zhang et al., 2024; Zhong, 2023). Recent work argues that the empirical successes of deep learning will give rise to a scientific *mechanics of learning* (Simon et al., 2026): just as physical mechanics is organized around canonical solvable models (the harmonic oscillator, the hydrogen atom, the Ising model) and a few dimensionless quantities (the Reynolds number), the mechanics of learning aims for a compact catalog of laws governing training. This work contributes one entry to that catalog, a toy model for plasticity in deep neural networks.

We study *plasticity*, a phenomenon investigated in parallel in continual learning and evolutionary biology, with mounting evidence suggesting shared structure across the two fields. We build on García-Galindo and Ahnert (2025), who show that biological plasticity emerges when environments alternate and identify four controlling factors (neutral set size (NSS), Hamming distance, switching period, and mutation rate). We translate their setting to a deep network trained alternately on two label sets and ask which of these controls survive the passage to gradient descent.

In continuous time, the system is a switched first-order gradient flow (Latz, 2021), structurally identical to the periodically forced overdamped systems studied in statistical

© 2026 O. Skriloff.

physics (Joubaud et al., 2015; Jung, 1993), electrical engineering (Oppenheim et al., 1997), and chemical engineering (Levenspiel, 1999). We exploit this analogy to identify the dimensionless quantities advocated by Simon et al. (2026). By reading the learning rate as an inverse resistance and the switching period as a relaxation time, their product, a dimensionless reach in parameter space, represents a Damköhler number (Levenspiel, 1999) that sets how far the network relaxes within one phase. Our contributions are as follows:

- **Two dimensionless controls.** We reduce the four biological factors of García-Galindo and Ahnert (2025) to two dimensionless controls of alternating training (task disagreement r and reach ηT), and show the remaining factors are subdominant. The optimal reach further obeys an approximate power law $\eta T^* \propto r^{-1.18} \approx 1/r$, so the optimal reach ηT^* can be set from r alone.
- **Bounds on plasticity.** We prove that r alone fixes a tight geometric floor on the utopia distance, $d_U \geq \sqrt{2}(1-2^{-r})$ (Theorem 2), while r and ηT jointly bound forgetting, $\Delta a_1 \lesssim e^{r\,\eta T\,M} - 1$ (Theorem 1).
- **Empirical validation and a divergence from biology.** Across 9,720 trajectories, an ANOVA shows that the *static* landscape quantity of neutral set size, critical in biology, has negligible effect, while the *dynamic* quantities of task disagreement, learning rate, and switching period dominate.
- **Toward a mechanics of learning.** Together, these results (disentangled hyperparameters, dimensionless controls, and scaling laws) provide a new solvable setting for plasticity that mirrors driven systems in physics and engineering, supporting the emerging mechanics of learning.

2. Related Work

Neural networks arose from a combination of biomimicry, through artificial neuron models, and statistical physics, through spin-glass models (Hopfield, 1982), spawning rich literatures linking machine learning to physics (Bahri et al., 2020) and neuroscience (Hassabis et al., 2017). Evolutionary biology and learning have likewise been connected through genetic algorithms (Katoch et al., 2021), shared simplicity bias (Johnston et al., 2022), and readings of the central dogma as a generative model (Mitchell and Cheney, 2025).

A strong connection has also formed between plasticity in evolutionary adaptation and in continual learning. The loss of plasticity, or a decline in a network's ability to learn new tasks, is well documented (Dohare et al., 2024; Lyle et al., 2023). The field has largely targeted catastrophic forgetting under task change (Kirkpatrick et al., 2017; Yu et al., 2020; Dohare et al., 2024), with recent work tying forgetting to the geometry of loss landscapes (Mirzadeh et al., 2020) and cyclical environmental change (Abbas et al., 2023).

These fields are further linked by geometric similarities between fitness landscapes (with genes as parameters) and loss landscapes of neural networks (with weights and biases as parameters). The equivalence of local minima (Choromanska et al., 2014; Laurent and von Brecht, 2017), connectedness of low-loss regions (Garipov et al., 2018; Yunis et al., 2022; Lubana et al., 2023; Altintas et al., 2023), and link between flatness and generalization (Hochreiter and Schmidhuber, 1997; Haddouche et al., 2025; Pouplin et al., 2023) all

recall the large, connected *neutral sets* of genes that map to the same optimal phenotype in genotype–phenotype (GP) maps (Ahnert, 2017). García-Galindo and Ahnert (2025) show that plasticity emerges when environments alternate, and identify neutral set size as one of four controlling factors, along with Hamming distance, switching period, and mutation rate. We follow the precedent of taking inspiration from biology and test whether plasticity emerges in an analogous continual learning setting, studying how controls survive in an alternative substrate.

3. Notation and Key Quantities

Plasticity. Plasticity has two related but distinct meanings. In continual learning it denotes a network's sustained ability to learn, whose decline is catastrophic forgetting. In the evolutionary setting of García-Galindo and Ahnert (2025) it denotes a single genotype performing well across multiple environments at once, or adapting quickly to changing conditions.

Accuracy proxies. We train a network alternately on two label sets y_1 and y_2 over N inputs, switching every T steps. Writing $\ell_j(t)$ for the loss on task j at step t, we form accuracy proxies $a_j(t) = e^{-\ell_j(t)} \in (0, 1]$, so $a_j = 1$ is perfect performance on task j. These proxies are bounded, monotone transformations of the loss, confine the trajectory to the unit square where the utopia geometry is defined, and echo the Boltzmann-probability reading of performance in García-Galindo and Ahnert (2025). The main conclusions are not specific to this choice: other monotone maps to the unit square (a reciprocal transform, direct 0/1 accuracy, and a Brier-style score) yield the same qualitative trends and rank-correlate strongly with a_j (Appendix C). The pair $(a_1(t), a_2(t))$ traces a trajectory in the unit square, and joint mastery corresponds to the *utopia* point $(1, 1)$ (Figure 1).

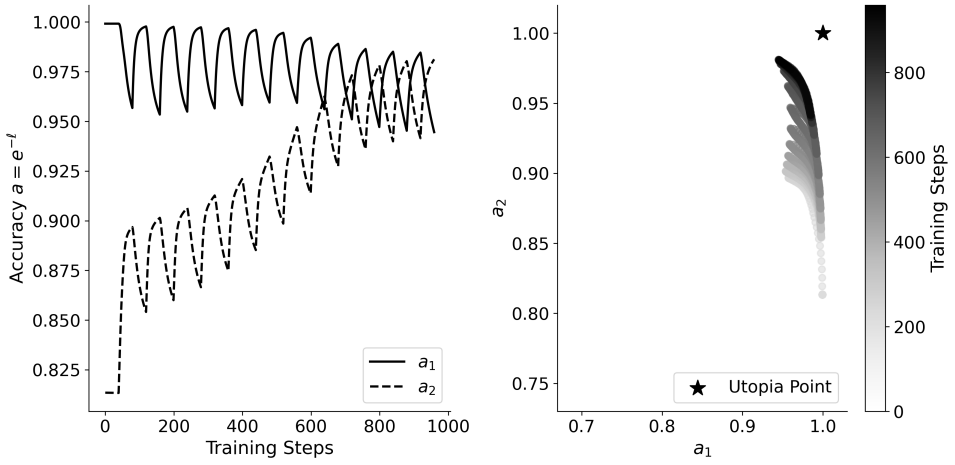

Figure 1: Example trajectory of alternating training on two functions. The left panel shows the accuracy proxies $a_1(t)$ and $a_2(t)$ over iterations. The right panel shows the corresponding phase plane trajectory and utopia point over iterations.

Utopia distance. The *instantaneous utopia distance* $\delta_U(t) = \|(a_1(t), a_2(t)) - (1,1)\|_2$ is the Euclidean distance from the current state to the utopia point. The *utopia distance* d_U captures how close the network comes to mastering both tasks at once: it is the time average of $\delta_U(t)$ at steady state, over the final two switching periods (ending at step T_*),

$$d_U = \frac{1}{2T} \sum_{t=T_*-2T}^{T_*-1} \delta_U(t),$$

so d_U is small only when both a_1 and a_2 stay near 1 throughout the cycle, penalizing a trajectory that excels at one task at the expense of the other. A lower d_U indicates higher plasticity.

Forgetting amplitude. The *forgetting amplitude* Δa_1 captures how much task-1 performance is lost when training switches to task 2: the peak-to-peak swing in a_1,

$$\Delta a_1 = \sup_{t,s \in [0,2T]} |a_1(t) - a_1(s)|,$$

analogous to catastrophic forgetting (Kirkpatrick et al., 2017; Yu et al., 2020) and to the robustness measure of García-Galindo and Ahnert (2025).

Dimensionless controls. Two dimensionless quantities govern these outcomes. The *task disagreement* $r = d_H/N \in [0,1]$ is the fraction of inputs on which y_1 and y_2 differ, with d_H their Hamming distance (see Metcalf and Casey, 2016). The *reach* ηT multiplies the learning rate η (step^{-1}) by the switching period T (steps), measuring how far the network travels in parameter space during a single phase.

Analysis of variance. To test which factors drive the empirical outcomes (Section 6), we use a Type II analysis of variance (ANOVA), reporting for each factor the effect size ω^2, the fraction of outcome variance it explains. A large ω^2 marks a dominant factor, letting us compare the influence of r, η, T, neutral set size, and complexity.

4. Experimental Setup

We approximate 7-bit Boolean functions ($N = 2^7 = 128$) with a neural network (architecture and training details in Appendix A, Table 2). For each function f, the neutral set size $\mathbb{P}(f) = \int_\Theta \mathbb{I}\{\ell_f(\theta) = 0\}\, p(\theta)\, d\theta$ is estimated by Monte Carlo sampling from a Gaussian prior $p(\theta)$ (Mingard et al., 2025; Zhang et al., 2021). Since simplicity bias couples Lempel–Ziv complexity $\tilde{K}(f)$ tightly to neutral set size (Dingle et al., 2018; Mingard et al., 2025), we decouple them by sampling function pairs within three complexity bins at varying probabilities. As all pairs have nonzero Hamming distance and cannot have overlapping neutral sets, a pair's total neutral set size is $\nu = \mathbb{P}(f_1) + \mathbb{P}(f_2)$.

We select 27 function pairs spanning low, medium, and high values of neutral set size, complexity, and Hamming distance d_H in a $3 \times 3 \times 3$ grid (Appendix A, Table 3). After a pretraining phase on the first function, each experiment alternates SGD between the two label sets every T steps until reaching steady state. Sweeping six learning rates and six switching periods (Table 2) with ten Monte Carlo trials per (η, T) pair gives 9,720 trajectories, for each of which we compute d_U and Δa_1.

5. Theory

All proofs are in Appendix B. We first upper bound the forgetting amplitude Δa_1, giving its scaling behavior:

Theorem 1 *Suppose the trajectory has reached a period-$2T$ steady-state orbit, so that $a_1(0) = a_1(2T)$; that $a_1(t)$ monotonically decreases during the y_2 training phase; and that the loss is L-smooth and the network f is C-smooth on the orbit. Let*

$$M = \sup_{\theta \in \text{Orbit}} \|\nabla \ell_1(\theta)\|_2 \max_{i \in \perp} \|J_i(\theta)\|_2 \leq LC < \infty,$$

where $\perp$ denotes the set of indices where the labels y_1 and y_2 differ, and $J_i(\theta)$ is the Jacobian of the network output with respect to parameters θ at index i. Here, M is the maximum amplitude of the forcing term in the linearized response of a_1 to training on y_2.

Then $\Delta a_1 \leq c_1$, where $c_1 = a_1(0)(e^{r\eta TM + O(\eta^2 T)} - 1)$. Further, for sufficiently small η, we have $c_1 \approx a_1(0)(e^{r\eta TM} - 1)$.

Two caveats temper the bound. First, the monotonic decrease of $a_1(t)$ during the y_2 phase is not guaranteed, though it holds in 97.5% of phases within a test grid of 56,748 steady-state y_2 phases (Appendix D). The assumption is thus a close but not universal description of the dynamics. Second, $M \leq LC$ is loose, so we use Theorem 1 for its scaling rather than as a quantitative bound, verified in Figure 2 (left) with the uniform proxy $M = 100$.

Next, we derive a tight lower bound on d_U depending only on r (Figure 2, right):

Theorem 2 *If $0 < r < 1$, then $d_U \geq \sqrt{2}(1 - 2^{-r})$.*

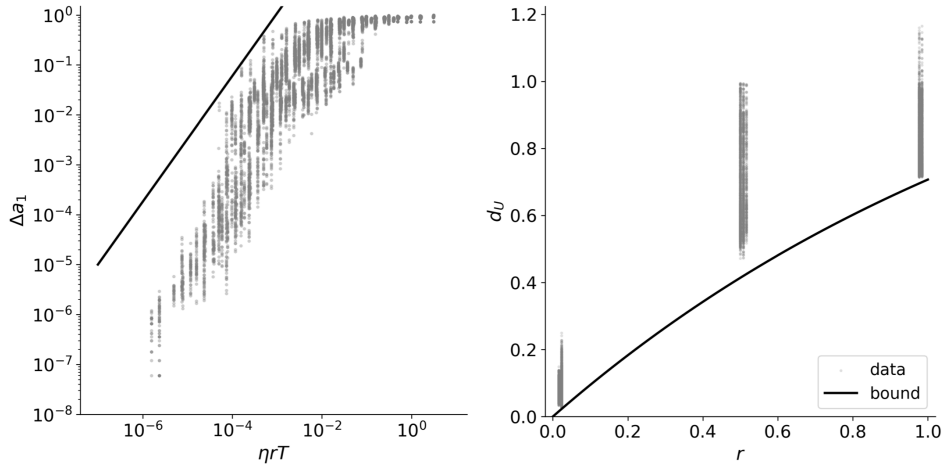

Figure 2: Empirical demonstration of the bounds of Theorems 1 (left) and 2 (right) across the 9,720 trajectories. For Theorem 1, we use the uniform estimate $M = 100$.

6. Results

To test whether the theory's exclusive dependence on η, r, and T is an artifact of the proof techniques or a true phenomenon, we ran a Type II ANOVA (Table 1). It confirms both predicted dependencies: η and T have much larger effect sizes on Δa_1 than on d_U, while r dominates d_U far more than Δa_1. The bounds of Theorems 1 and 2 thus capture the dominant dependence of each outcome on the controls.

Table 1: Type II ANOVA effect sizes for the two log-transformed plasticity outcomes. Models use complexity bin, neutral-set-size class, Hamming-distance class, learning rate, and switching period as categorical predictors. Reported values are ω^2 effect sizes. Significance levels are denoted by *** for $p < 0.001$. Note that while $\tilde{K}$ and NSS yield statistically significant effects, their effect sizes are negligible compared to the other factors.

Factor	$\log \Delta a_1$	$\log d_U$
$\tilde{K}$ bin	1×10^{-4}	$3\times10^{-4***}$
NSS (ν) class	$3\times10^{-4***}$	$1\times10^{-4***}$
Hamming (r) class	0.4966^{***}	0.9422^{***}
η	0.2178^{***}	0.0019^{***}
T	0.1348^{***}	0.0030^{***}
Residual	0.1503	0.0525
R^2	0.8497	0.9475

Notably, the ANOVA rejects the hypothesis that neutral set size ν and complexity $\tilde{K}$ strongly influence plasticity as they do in biology: despite statistical significance from the large sample, their effect sizes are negligible compared to the other factors. This points to a structural difference between evolution on the fitness landscape of genotype-phenotype maps and learning on the loss landscape of neural networks.

Despite the dominant effect size of r, the spread of d_U above the theoretical floor within each r bin (Figure 2) shows that r alone does not fully determine d_U. Where a trajectory actually sits above the extremal floor given by Theorem 2 is governed by the reach ηT, as we show next. A heatmap of d_U over the (T, η) space (Figure 3, left) reveals a diagonal of optimal (lower) values along which ηT is approximately constant, suggesting that the product ηT governs optimal plasticity.

Plotting d_U against ηT across the r spectrum (Figure 3, center) confirms this. As r increases, the minimum attainable utopia distance d_U *rises*, tracking the Theorem 2 floor, while the optimal ηT (the value minimizing d_U) *falls*, and the optimum sharpens as the r-level curves grow more convex. Across 37 task pairs spanning 17 disagreement levels, this optimal reach follows an approximate power law (Figure 3, right), $\eta T^* \propto r^{-1.18}$ (pair-bootstrap 95% CI $[-1.34, -1.01]$, $R^2 = 0.86$). Together, theory and experiment show a clean division of labor: task disagreement alone sets the geometric floor of d_U, the reach governs where the trajectory sits above it, and their product $r\eta T$ governs the scaling of the forgetting amplitude Δa_1.

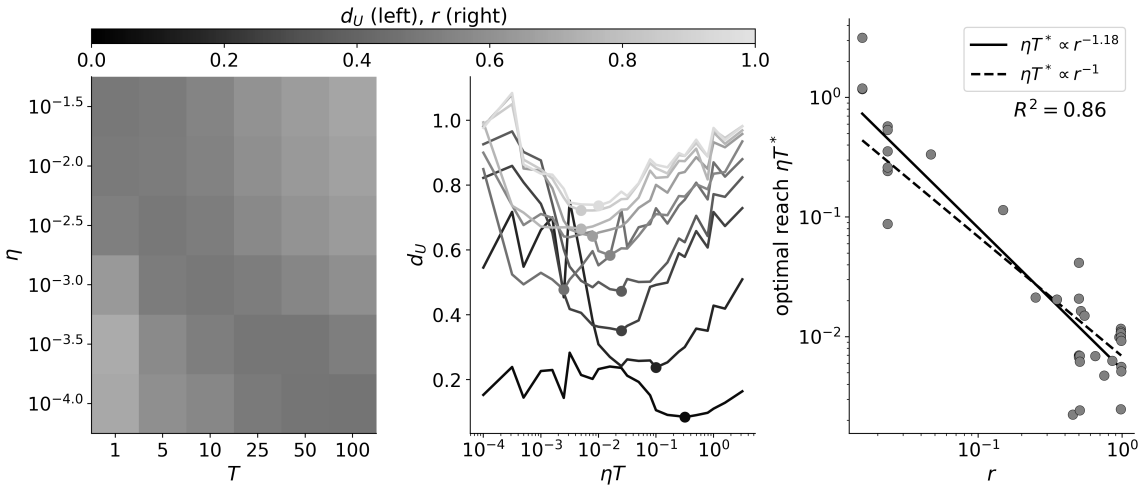

Figure 3: Heatmap of d_U values across (T, η) space with the most plastic trajectories (lowest d_U values) occurring along a diagonal (constant ηT) (left). Task disagreement r level curves of the utopia distance d_U versus nominal reach ηT (center) show that as r grows the optimum rises and shifts to smaller ηT while curves become more convex. The optimal reach ηT^* versus r over 37 task pairs (right) shows a power-law fit $\eta T^* \propto r^{-1.18}$ (95% CI $[-1.34, -1.01]$, $R^2 = 0.86$).

Beyond this primary grid, a set of smaller robustness studies (Appendix E) indicates the two controls are not artifacts of the specific setup. The reach-governed picture persists under minibatch SGD, momentum, and Adam, and under held-out continuous-input tasks where performance is measured on unseen test data (mean train–test gap in d_U below 0.01). Varying network width and depth over a 16× range in parameter count leaves d_U essentially unchanged, giving direct empirical support for network-size independence. These small studies are reported as consistency checks rather than conclusive evidence of generality.

7. Discussion

Plasticity across substrates. We translated the plasticity model of García-Galindo and Ahnert (2025) from evolutionary biology to deep learning, enabling a direct comparison across substrates. While Hamming distance, mutation (learning) rate, and switching period control plasticity in both settings, neutral set size plays a negligible role in the learning setting, suggesting a geometric divergence between the parameter spaces of genes and network weights despite the similarity of their fitness and loss landscapes.

The negligible role of neutral set size is the most surprising departure from biology. One plausible mechanism is that population dynamics deploy many trajectories that sample the neutral set volume, whereas a single SGD run occupies one point in it, governed by local forcing over global geometry. This predicts a larger role for neutral set size in ensemble methods that mimic population-based training as in Fort et al. (2020), or under large explicit noise injection. We leave the testing of this prediction to future work.

Reach as a physical control parameter. Prior work by Latz (2021) casts the continuous-time limit of SGD as phenotypic evolution under randomly fluctuating environments. We replace the random fluctuation with a deterministic square wave, a canonical forcing signal in physics and engineering. The linearized response to such forcing scales like $\tanh(x)$ (Chaffey and Forni, 2025), which Theorem 1 bounds via $\tanh(x) \leq e^x - 1$, suggesting room to tighten in the linear regime.

Identifying the reach ηT as a control parameter draws a further analogy. The optimal learning rate η is closely tied to the inverse local curvature (Lipschitz constant) of the loss landscape (Min and Lee, 2026; Cohen et al., 2021). Reading curvature as the system's "resistance" makes η an inverse resistance, speeding return to minima after switching. T then becomes the relaxation time, with their product ηT being the distance traveled in one phase, the reach. In chemical engineering this is exactly the first Damköhler number (Levenspiel, 1999), the ratio of residence to reaction time. The optimal reach spans roughly 0.005–0.02 at high r and is of order 0.1–1 at low r, where the optimum is only weakly identified because the reach curve flattens. Thus, $\eta T \lesssim 1$ throughout and $\eta T \ll 1$ wherever the optimum is sharp, identifying the system as reaction-limited: for optimal plasticity, the phase must be short enough that relaxation cannot complete to a local minimum.

The convexity of the reach curve follows from the same picture: at low r the system relaxes to a shared minimum and the curve is nearly flat, while at higher r too little reach underfits both tasks and too much reach overfits the current task and forgets the other, both inflating d_U. The optimal reach falls as disagreement grows since the system becomes more sensitive to overfitting due to the greater distance between the two tasks' minima.

Further, the optimal reach falls with r as an approximate power law $\eta T^* \propto r^{-1.18}$, close to the inverse law $\eta T^* \propto 1/r$ and suggesting a universal scaling for alternating tasks. This yields a design heuristic: given the many methods for estimating η from curvature (Min and Lee, 2026; Cohen et al., 2021), the optimal switching period follows from the task disagreement alone, without probing the loss landscape or architecture. Combined with the two theorems, it collapses both plasticity outcomes to functions of r: the utopia floor $d_U \geq \sqrt{2}(1 - 2^{-r})$ and the forgetting amplitude $\Delta a_1 \lesssim e^{r\eta T^* M} - 1 \approx e^{cM} - 1$ for some constant c.

Connections to continual learning. Our results also quantify continual-learning findings. Forgetting as curvature-weighted parameter displacement underlies the loss-landscape analysis of Mirzadeh et al. (2020), who show empirically that batch size, learning rate, and flatness affect forgetting. Theorem 1 gives a closed-form scaling in reach ηT and disagreement r. Elastic Weight Consolidation (EWC; Kirkpatrick et al., 2017) adds a remembering term penalizing changes to parameters important for the previous task, and Theorem 1 adapts directly to quantify the gain. Under a diagonal-Fisher approximation near an optimum, an EWC penalty of strength λ replaces the binary cross entropy (BCE) bound's exponent $\eta T(rM - E)$ with $\eta T(rM - E(1+\lambda))$, where E is the mean squared task-1 gradient norm over the phase, strictly lowering the forgetting bound (Corollary 4, Appendix B).

Continual Learning with Experience and Replay (CLEAR; Rolnick et al., 2018) interweaves data from both tasks, much like our task switching. Since the empirically optimal reach under d_U is nonzero, effective plasticity on disagreeing tasks requires a finite amount of learning within each alternating phase, giving quantitative support for CLEAR. Our results

further suggest that optimal hyperparameters for CLEAR, or for alternating environments used directly as a continual-learning method, could be estimated from theory alone using our hyperparameter heuristic when the tasks disagree.

Toward a mechanics of learning. Taken together, these results add direct evidence to four of the five strands of the mechanics-of-learning program (Simon et al., 2026): a solvable, idealized setting for plasticity, simple governing laws (Theorems 1 and 2), a reduction of hyperparameters to two dimensionless controls, and a link between evolutionary biology, chemistry, and learning hinting at substrate universality.

8. Conclusion and Future Work

We translated a model of biological plasticity to deep networks trained on alternating Boolean tasks and found that plasticity under task switching is governed by two dimensionless controls: the task disagreement r and the reach ηT. We proved that r alone fixes a tight geometric floor on the utopia distance, while r and ηT jointly bound forgetting. An ANOVA confirmed that these controls dominate the empirical behavior while the static landscape quantities central to the biological setting do not. The analogy that survives the translation is therefore dynamical rather than geometric, placing single-trajectory learning alongside driven systems in physics and engineering.

Since any computational data can in principle be written as a Boolean function, the setting is a candidate minimal model. Finer sweeps of the reach's interaction with r could sharpen the exponent, but we conjecture more strongly that setting the reach to the disagreement-matched value $\eta T \approx 1/r$ provides near-optimal conditions for continual learning on alternating tasks (phases short enough to avoid overfitting the current task, yet long enough to keep learning). Thus, hyperparameter constraints follow from r alone, echoing the transfer of maximal-update parameterization (μP) scaling heuristics (Yang et al., 2022).

The disagreement $r = d_H/N$ itself invites generalization. Because d_H is a metric on binary label vectors, the framework may be generalized using other dissimilarity measures. For BCE, information-theoretic divergences or distances are natural candidates. The discrete bound $a_1 a_2 \leq 4^{-r}$ (Lemma 3, Appendix B) suggests one route toward extending the framework to soft labels and continuous outputs. We present one alternative continuous metric in Appendix E where r is replaced by the angle between dividing hyperplanes of two linear classifiers.

The switching period T should be governed by the non-commutativity of the two gradient flows. The Lie bracket $[\nabla \ell_1, \nabla \ell_2]$ (Rukhovich et al., 2025), accumulated over one period, links the reach to Δa_1. Since this quantity vanishes when the flows commute, we recover the $r \to 0$ limit of Theorem 1, but further analysis could explicitly link the reach, forgetting amplitude, and the Lie bracket for general tasks. This would extend the framework to general loss functions and continuous outputs, and could provide a geometric interpretation of the reach as a curvature-weighted distance in parameter space.

The framework may also extend to generalization error. Since forgetting is a performance loss from distribution shift between tasks (Doan et al., 2021), Δa_1 and d_U may measure robustness to distribution shift and noisy labels, underscored by the out-of-distribution forgetting continual learners show under small intra-class shifts (Guo et al., 2023).

Acknowledgments

Thank you to Abiy Tassisa for helpful discussions and feedback on the manuscript.

References

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C. Machado. Loss of plasticity in continual deep reinforcement learning, 2023. URL https://arxiv.org/abs/2303.07507.

S. E. Ahnert. Structural properties of genotype–phenotype maps. *Journal of The Royal Society Interface*, 14(132):20170275, 2017. doi: 10.1098/rsif.2017.0275. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2017.0275.

Gul Sena Altintas, Gregor Bachmann, Lorenzo Noci, and Thomas Hofmann. Disentangling linear mode-connectivity, 2023. URL https://arxiv.org/abs/2312.09832.

Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S. Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11(Volume 11, 2020):501–528, 2020. ISSN 1947-5462. doi: https://doi.org/10.1146/annurev-conmatphys-031119-050745. URL https://www.annualreviews.org/content/journals/10.1146/annurev-conmatphys-031119-050745.

Thomas Chaffey and Fulvio Forni. Amplitude response and square wave describing functions. *European Journal of Control*, 86:101310, 2025. ISSN 0947-3580. doi: https://doi.org/10.1016/j.ejcon.2025.101310. URL https://www.sciencedirect.com/science/article/pii/S0947358025001396. Special Issue on the European Control Conference 2025.

Anna Choromanska, Mikael Henaff, Michaël Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surface of multilayer networks. *CoRR*, abs/1412.0233, 2014. URL http://arxiv.org/abs/1412.0233.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *CoRR*, abs/2103.00065, 2021. URL https://arxiv.org/abs/2103.00065.

Kamaludin Dingle, Chico Q. Camargo, and Ard A. Louis. Input–output maps are strongly biased towards simple outputs. *Nature Communications*, 9(1):761, 2018. doi: 10.1038/s41467-018-03101-6. URL https://doi.org/10.1038/s41467-018-03101-6.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 130:1072–1080, 2021. URL https://proceedings.mlr.press/v130/doan21a.html.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774, 2024. doi: 10.1038/s41586-024-07711-7. URL https://www.nature.com/articles/s41586-024-07711-7.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective, 2020. URL https://arxiv.org/abs/1912.02757.

Paula García-Galindo and Sebastian E. Ahnert. RNA plasticity emerges as an evolutionary response to fluctuating environments. *bioRxiv*, 2025. ISSN 2692-8205. doi: 10.1101/2024.10.02.614758. URL https://www.biorxiv.org/content/10.1101/2024.10.02.614758v2.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/be3087e74e9100d4bc4c6268cdbe8456-Paper.pdf.

Liangxuan Guo, Yang Chen, and Shan Yu. Out-of-distribution forgetting: Vulnerability of continual learning to intra-class distribution shift, 2023. URL https://arxiv.org/abs/2306.00427.

Maxime Haddouche, Paul Viallard, Umut Simsekli, and Benjamin Guedj. A PAC-Bayesian link between generalisation and flat minima. In Gautam Kamath and Po-Ling Loh, editors, *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pages 481–511. PMLR, 24–27 Feb 2025. URL https://proceedings.mlr.press/v272/haddouche25a.html.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2017.06.011. URL https://www.sciencedirect.com/science/article/pii/S0896627317305093.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554. URL https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL https://arxiv.org/abs/2405.07987.

Iain G. Johnston, Kamaludin Dingle, Sam F. Greenbury, Chico Q. Camargo, Jonathan P. K. Doye, Sebastian E. Ahnert, and Ard A. Louis. Symmetry and simplicity spontaneously emerge from the algorithmic nature of evolution. *Proceedings of the National Academy*

of Sciences, 119(11):e2113883119, 2022. doi: 10.1073/pnas.2113883119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2113883119.

Rémi Joubaud, Grigorios A. Pavliotis, and Gabriel Stoltz. Langevin dynamics with space–time periodic nonequilibrium forcing. *Journal of Statistical Physics*, 158(1):1–36, 2015.

Peter Jung. Periodically driven stochastic systems. *Physics Reports*, 234(4–5):175–295, 1993.

Sourabh Katoch, Sumit Singh Chauhan, and Vijay Kumar. A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, 80(5):8091–8126, 2021. doi: 10.1007/s11042-020-10139-6.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1611835114.

Jonas Latz. Analysis of stochastic gradient descent in continuous time. *Statistics and Computing*, 31(4):39, 2021.

Thomas Laurent and James von Brecht. Deep linear neural networks with arbitrary loss: All local minima are global. *CoRR*, abs/1712.01473, 2017. URL http://arxiv.org/abs/1712.01473.

Octave Levenspiel. *Chemical Reaction Engineering*. John Wiley & Sons, 3rd edition, 1999.

Ekdeep Singh Lubana, Eric J. Bigelow, Robert P. Dick, David Krueger, and Hidenori Tanaka. Mechanistic mode connectivity, 2023. URL https://arxiv.org/abs/2211.08422.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks, 2023. URL https://arxiv.org/abs/2303.01486.

Leigh Metcalf and William Casey. Chapter 2 - metrics, similarity, and sets. In *Cybersecurity and Applied Mathematics*, pages 3–22. Syngress, Boston, 2016. ISBN 978-0-12-804452-0. doi: https://doi.org/10.1016/B978-0-12-804452-0.00002-6. URL https://www.sciencedirect.com/science/article/pii/B9780128044520000026.

Chohong Min and Byungjoon Lee. A relaxation approach to layerwise determination of learning rates in deep neural networks. *Array*, 30:100807, 2026. ISSN 2590-0056. doi: https://doi.org/10.1016/j.array.2026.100807. URL https://www.sciencedirect.com/science/article/pii/S259000562600130X.

Chris Mingard, Henry Rees, Guillermo Valle-Pérez, and Ard A. Louis. Deep neural networks have an inbuilt Occam's razor. *Nature Communications*, 16(1), January 2025.

ISSN 2041-1723. doi: 10.1038/s41467-024-54813-x. URL http://dx.doi.org/10.1038/s41467-024-54813-x.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Dilan Görür, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. *CoRR*, abs/2010.04495, 2020. URL https://arxiv.org/abs/2010.04495.

Kevin J. Mitchell and Nick Cheney. The genomic code: the genome instantiates a generative model of the organism. *Trends in Genetics*, 41(6):462–479, 2025. doi: 10.1016/j.tig.2025.01.008. URL https://doi.org/10.1016/j.tig.2025.01.008.

Alan V. Oppenheim, Alan S. Willsky, and S. Hamid Nawab. *Signals and Systems*. Prentice Hall, 2nd edition, 1997.

Alison Pouplin, Hrittik Roy, Sidak Pal Singh, and Georgios Arvanitidis. On the curvature of the loss landscape, 2023. URL https://arxiv.org/abs/2307.04719.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. Experience replay for continual learning. *CoRR*, abs/1811.11682, 2018. URL http://arxiv.org/abs/1811.11682.

Alexey Rukhovich, Alexander Podolskiy, and Irina Piontkovskaya. Commute your domains: Trajectory optimality criterion for multi-domain learning, 2025. URL https://arxiv.org/abs/2501.15556.

Jamie Simon, Daniel Kunin, Alexander Atanasov, Enric Boix-Adserà, Blake Bordelon, Jeremy Cohen, Nikhil Ghosh, Florentin Guth, Arthur Jacot, Mason Kamb, Dhruva Karkada, Eric J. Michaud, Berkan Ottlik, and Joseph Turnbull. There will be a scientific theory of deep learning, 2026. URL https://arxiv.org/abs/2604.21691.

Loek van Rossem and Andrew M. Saxe. When representations align: Universality in representation learning dynamics, 2024. URL https://arxiv.org/abs/2402.09142.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer, 2022. URL https://arxiv.org/abs/2203.03466.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *CoRR*, abs/2001.06782, 2020. URL https://arxiv.org/abs/2001.06782.

David Yunis, Kumar Kshitij Patel, Pedro Henrique Pamplona Savarese, Gal Vardi, Jonathan Frankle, Matthew Walter, Karen Livescu, and Michael Maire. On convexity and linear mode connectivity in neural networks. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022. URL https://openreview.net/forum?id=TZQ3PKL3fPr.

Huijie Zhang, Jinfan Zhou, Yifu Lu, Minzhe Guo, Peng Wang, Liyue Shen, and Qing Qu. The emergence of reproducibility and consistency in diffusion models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 60558–60590. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhang24cn.html.

Shuofeng Zhang, Isaac Reid, Guillermo Valle Pérez, and Ard Louis. Why flatness does and does not correlate with generalization for deep neural networks, 2021. URL https://arxiv.org/abs/2103.06219.

Weishun Zhong. Non-equilibrium physics: from spin glasses to machine and neural learning, 2023. URL https://arxiv.org/abs/2308.01538.

Appendix A. Computational Experiment Details

Table 2: Architecture and training hyperparameters.

Setting	**Value**
Hidden layers	10
Layer width	40
Activation	Tanh
Initialization	i.i.d. $\mathcal{N}(0, 1/\text{width})$
Loss function	Binary cross-entropy (BCE)
Optimizer	SGD (full batch)
Learning rates η	$\{10^{-4}, 10^{-3.5}, 10^{-3}, 10^{-2.5}, 10^{-2}, 10^{-1.5}\}$
Switching periods T	$\{1, 5, 10, 25, 50, 100\}$
Max iterations	2500
Trials per (η, T)	10

Appendix B. Proofs for Theoretical Results

B.1. Proof of Theorem 1

Proof From $t = 0$ to $t = T$, training proceeds by full-batch gradient descent on the y_1 labels. Since ℓ_1 is L-smooth (an assumption of the theorem), the descent lemma guarantees that any step size $\eta < 2/L$ strictly decreases the loss at each iteration; hence $\ell_1(t)$ is monotonically decreasing and $a_1(t) = e^{-\ell_1(t)}$ is monotonically increasing on $[0, T]$. By the assumption of the theorem, from time $t = T$ to $t = 2T$ we know $a_1(t)$ is monotonically decreasing. Further, by periodicity we know that $a_1(0) = a_1(2T)$. Thus, we can conclude that

$$\Delta a_1 = \max_{s,t\in[0,2T]} |a_1(t) - a_1(s)| = a_1(T) - a_1(0)$$

Table 3: Attributes of the 27 label pairs (y_1, y_2) used in the computational experiments.

#	$\tilde{K}$	$\mathbb{P}(y_1)$	$\mathbb{P}(y_2)$	ν	d_H
1	56.0	1.80×10^{-5}	6.00×10^{-6}	2.40×10^{-5}	3
2	56.0	1.00×10^{-5}	1.20×10^{-5}	2.20×10^{-5}	64
3	56.0	1.80×10^{-5}	6.00×10^{-6}	2.40×10^{-5}	125
4	56.0	2.20×10^{-5}	4.60×10^{-5}	6.80×10^{-5}	3
5	56.0	3.00×10^{-5}	3.40×10^{-5}	6.40×10^{-5}	65
6	56.0	4.60×10^{-5}	2.20×10^{-5}	6.80×10^{-5}	125
7	56.0	9.80×10^{-5}	2.36×10^{-4}	3.34×10^{-4}	2
8	56.0	5.80×10^{-5}	3.94×10^{-4}	4.52×10^{-4}	65
9	56.0	9.80×10^{-5}	2.36×10^{-4}	3.34×10^{-4}	126
10	59.5	6.00×10^{-6}	1.20×10^{-5}	1.80×10^{-5}	3
11	59.5	1.20×10^{-5}	6.00×10^{-6}	1.80×10^{-5}	64
12	59.5	1.20×10^{-5}	6.00×10^{-6}	1.80×10^{-5}	126
13	59.5	3.00×10^{-5}	2.00×10^{-5}	5.00×10^{-5}	2
14	59.5	5.00×10^{-5}	6.00×10^{-6}	5.60×10^{-5}	64
15	59.5	3.00×10^{-5}	2.00×10^{-5}	5.00×10^{-5}	125
16	59.5	6.00×10^{-5}	3.80×10^{-4}	4.40×10^{-4}	2
17	59.5	4.56×10^{-4}	6.00×10^{-6}	4.62×10^{-4}	66
18	59.5	4.56×10^{-4}	5.00×10^{-5}	5.06×10^{-4}	126
19	66.5	6.00×10^{-6}	8.00×10^{-6}	1.40×10^{-5}	3
20	66.5	8.00×10^{-6}	6.00×10^{-6}	1.40×10^{-5}	64
21	66.5	6.00×10^{-6}	8.00×10^{-6}	1.40×10^{-5}	125
22	66.5	1.20×10^{-5}	8.00×10^{-6}	2.00×10^{-5}	3
23	66.5	1.20×10^{-5}	6.00×10^{-6}	1.80×10^{-5}	64
24	66.5	1.20×10^{-5}	8.00×10^{-6}	2.00×10^{-5}	125
25	66.5	1.20×10^{-5}	1.30×10^{-4}	1.42×10^{-4}	3
26	66.5	1.30×10^{-4}	6.00×10^{-6}	1.36×10^{-4}	65
27	66.5	1.30×10^{-4}	1.20×10^{-5}	1.42×10^{-4}	125

Since the trajectory evolves by gradient descent with respect to the y_1 labels during the $[0, T]$ phase, we define the quantity ς, the total change in loss over the phase, as

$$\varsigma := \eta \sum_{t=0}^{T-1} \|\nabla \ell_1(t)\|_2^2 + O\left(\eta^2 \sum_{t=0}^{T-1} \|\nabla \ell_1(t)\|_2^2\right)$$

Alternatively, we can derive that $\ell_1(T) = \ell_1(0) - \varsigma$ and $\varsigma = \ell_1(2T) - \ell_1(T)$ by periodicity. This allows us to write

$$\Delta a_1 = a_1(0)(e^{\varsigma} - 1). \tag{1}$$

We proceed by upper bounding ς. In this second phase where $a_1(t)$ is decreasing, we know that the increasing value of ℓ_1 is governed by the gradient disagreement between labels $\langle \nabla \ell_1(t), \nabla \ell_2(t) \rangle$. Using a Taylor expansion, we obtain

$$\varsigma = -\eta \sum_{t=T}^{2T-1} \langle \nabla \ell_1(t), \nabla \ell_2(t) \rangle + O\left(\eta^2 \sum_{t=T}^{2T-1} \|\nabla \ell_1(t)\|_2^2\right).$$

To bound $-\langle \nabla \ell_1(t), \nabla \ell_2(t) \rangle$, we first expand the gradient of the BCE loss giving

$$\nabla_\theta \ell(y, \theta) = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i) J_i,$$

where $p_i = \sigma(f(x_i))$ and $J_i = \nabla_\theta f(x_i)$. Since p is independent of y, subtracting the two gradients gives

$$\nabla \ell_2(t) = \nabla \ell_1(t) - \frac{1}{N} \sum_{i=1}^{N} (y_{2,i} - y_{1,i}) J_i.$$

Substituting into the inner product, we find that

$$\begin{aligned}
-\langle \nabla_\theta \ell_1(t), \nabla_\theta \ell_2(t) \rangle &= \left\langle \nabla \ell_1(t), \frac{1}{N} \sum_{i=1}^{N} (y_{2i} - y_{1i}) J_i \right\rangle - \|\nabla \ell_1(t)\|_2^2 && (2) \\
&\leq \left\langle \nabla \ell_1(t), \frac{1}{N} \sum_{i=1}^{N} (y_{2i} - y_{1i}) J_i \right\rangle && (3) \\
&\leq \frac{d_H}{N} \|\nabla \ell_1(t)\|_2 \max_{i \in \perp} \|J_i\|_2 && (4)
\end{aligned}$$

where $\perp$ denotes the set where the labels disagree, i.e. where $y_{1,i} - y_{2,i} \neq 0$. The last line is derived by restricting the sum to the d_H nonzero disagreement indices, applying Cauchy-Schwarz, and then a uniform Jacobian bound.

Plugging in and summing gives $\varsigma \leq r\eta T M + O(\eta^2 T)$. Applying this bound to Equation 1 proves the theorem. ∎

B.2. Proof of Theorem 2

To prove the lower bound, we begin with a lemma that provides a constraint for the optimization problem solved in the lower bound theorem that follows.

Lemma 3 *$a_1 a_2 \le c_2$ where $c_2 = 4^{-r}$.*

Proof We can expand the term-wise loss functions as

$$\ell_{1,i} + \ell_{2,i} = 2\ln(1 + e^z) - (y_{1,i} + y_{2,i})z,$$

where z is the neural network output. Then, we have two cases. If $y_{1,i} = y_{2,i}$, then we get $\ell_{1,i} + \ell_{2,i} \ge 0$. If not, then $y_{1,i} + y_{2,i} = 1$ and we have

$$2\ln(1 + e^z) - z = \ln(2 + 2\cosh z) \ge \ln 4.$$

Now, notice that

$$(y_{1,i} - y_{2,i})^2 = \begin{cases} 0, & \text{if} \quad y_{1,i} = y_{2,i} \\ 1, & \text{else} \end{cases} \tag{5}$$

$$= \mathbf{1}_{\{y_{1,i} \neq y_{2,i}\}}. \tag{6}$$

Next, we get the total loss by summing over the data points, which gives

$$\ell_1 + \ell_2 = \frac{1}{N}\sum_{i=1}^{N} (\ell_{1,i} + \ell_{2,i}) \tag{7}$$

$$\ge \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{y_{1,i} \neq y_{2,i}\}} \ln 4 \tag{8}$$

$$= \frac{\ln 4}{N} d_H. \tag{9}$$

Finally, expanding $a_1 a_2$ gives $a_1 a_2 = e^{-(\ell_1 + \ell_2)} \le e^{-\frac{\ln 4}{N} d_H} = 4^{-\frac{d_H}{N}}$. ■

Using the lemma, we can now prove the lower bound on the utopia distance.

Proof By Lemma 3, the constraint $a_1 a_2 \le c_2$ holds pointwise along the orbit. We therefore bound $\delta_U(t)$ pointwise by minimizing the squared instantaneous distance over admissible states $\mathbf{a} = (a_1, a_2)$:

$$\arg\min_{\mathbf{a}} \quad \|\mathbf{a} - \mathbf{1}\|_2^2 \quad \text{s.t.} \quad a_1 a_2 \le c_2. \tag{$\star$}$$

First, notice that $\|\mathbf{a} - \mathbf{1}\|_2^2$ decreases as a_1 or a_2 increase. Thus, the solution to ($\star$) will lie on the hyperbola $a_1 a_2 = c_2$. We proceed by using Lagrange multipliers:

$$\mathcal{L}(\mathbf{a}, \lambda) = \|\mathbf{a} - \mathbf{1}\|_2^2 - \lambda(a_1 a_2 - c_2) = (1 - a_1)^2 + (1 - a_2)^2 - \lambda(a_1 a_2 - c_2) \tag{10}$$

$$\nabla_{\mathbf{a}} \mathcal{L} = -2 \begin{bmatrix} 1 - a_1 \\ 1 - a_2 \end{bmatrix} - \lambda \begin{bmatrix} a_2 \\ a_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{11}$$

Solving, we find that $\lambda = \frac{-2}{a_2}(1-a_1) = \frac{-2}{a_1}(1-a_2)$. Cross multiplying and expanding gives

$$(a_1 - a_2)(1 - (a_1 + a_2)) = 0 \quad \Longrightarrow \quad a_1 = a_2 \quad \text{or} \quad a_1 + a_2 = 1.$$

The first case gives $a_1 = a_2 = \sqrt{c_2}$, resulting in $\|\mathbf{a} - \mathbf{1}\|_2 = \sqrt{2}(1 - \sqrt{c_2}) = \sqrt{2}(1 - 2^{-r})$. In the second case, Vieta's formulas give that a_1 and a_2 are the roots of $x^2 - x + c_2$. Hence, $a_1, a_2 = \frac{1 \pm \sqrt{1-4c_2}}{2}$. But we know that $a_1, a_2 \in \mathbb{R}$, so $c_2 \leq \frac{1}{4}$ which gives $4^{-r} \leq 4^{-1}$ which never happens since $0 < r < 1$. Thus, case 1 is the only possibility.

On the hyperbola $a_1 a_2 = c_2$, with $a_1, a_2 \in [0, 1]$, we parameterize $a_2 = c_2/a_1$ and then minimize $f(a_1) = (1 - a_1)^2 + (1 - c_2/a_1)^2$. By symmetry, the minimum is attained at $a_1 = a_2 = \sqrt{c_2}$ since $f''(\sqrt{c_2}) > 0$.

Thus, $\delta_U(t) \geq \sqrt{2}(1 - \sqrt{c_2}) = \sqrt{2}(1 - 2^{-r})$ pointwise, and averaging over the cycle gives $d_U \geq \sqrt{2}(1 - 2^{-r})$. ■

B.3. EWC Corollary

Corollary 4 *Suppose the Fisher information matrix is diagonal, the orbit remains sufficiently near some $\theta_1^* \in \Theta_1$ for ℓ_1 to admit a quadratic approximation, η is small enough that second-order update terms may be neglected, and all assumptions of Theorem 1 hold. Let $E = \frac{1}{T}\sum_{t=T}^{2T-1} \|\nabla \ell_1(t)\|_2^2$. If both losses use BCE and $rM \geq E(1+\lambda)$, then*

$$\Delta a_1 \leq a_1(0)\big(e^{\eta T(rM - E)} - 1\big);$$

if the loss for environment 2 is replaced with the EWC loss ℓ_2^{EWC}, then

$$\Delta a_1 \leq a_1(0)\big(e^{\eta T(rM - E(1+\lambda))} - 1\big).$$

Proof Recall that in proving Theorem 1, we needed to upper bound the inner product of the two loss gradients. Replacing ℓ_2 with ℓ_2^{EWC} gives

$$\begin{aligned}
-\langle \nabla \ell_1(\theta), \nabla \ell_2^{\mathrm{EWC}}(\theta) \rangle &= -\langle \nabla \ell_1(\theta), \nabla \ell_2(\theta) + \lambda F_1(\theta_1^*)(\theta - \theta_1^*) \rangle \\
&= -\langle \nabla \ell_1(\theta), \nabla \ell_2(\theta) \rangle - \lambda \langle \nabla \ell_1(\theta), F_1(\theta_1^*)(\theta - \theta_1^*) \rangle \\
&= \langle \nabla \ell_1(\theta), \frac{1}{N} \sum_{i=1}^{N} (y_{1,i} - y_{2,i}) J_i \rangle \\
&\quad - \|\nabla \ell_1(\theta)\|_2^2 - \lambda \langle \nabla \ell_1(\theta), F_1(\theta_1^*)(\theta - \theta_1^*) \rangle.
\end{aligned}$$

Because θ_1^* is a critical point of ℓ_1, we have $\nabla \ell_1(\theta_1^*) = 0$, so a first-order Taylor expansion of the *gradient field* about θ_1^* gives

$$\nabla \ell_1(\theta) = \nabla^2 \ell_1(\theta_1^*)(\theta - \theta_1^*) + O(\|\theta - \theta_1^*\|_2^2).$$

With BCE loss the Hessian at the optimum is approximately the (diagonal) Fisher information matrix, $\nabla^2 \ell_1(\theta_1^*) \approx F_1(\theta_1^*)$, so $F_1(\theta_1^*)(\theta - \theta_1^*) \approx \nabla \ell_1(\theta)$ as full vectors, aligned in direction and approximately equal in magnitude, rather than merely comparable in norm. Hence $\langle \nabla \ell_1(\theta), F_1(\theta_1^*)(\theta - \theta_1^*) \rangle \approx \|\nabla \ell_1(\theta)\|_2^2 \geq 0$, so the EWC term strictly lowers the bound, and

$$-\langle \nabla \ell_1(\theta), \nabla \ell_2^{\mathrm{EWC}}(\theta) \rangle \leq rM - (1+\lambda)\|\nabla \ell_1(\theta)\|_2^2.$$

If we let $E = \frac{1}{T}\sum_{t=T}^{2T-1} \|\nabla \ell_1(t)\|_2^2$, following the same steps as in Theorem 1 gives

$$\Delta a_1 \le a_1(0)(e^{\eta T(rM - E(1+\lambda))} - 1).$$

The BCE result is recovered by setting $\lambda = 0$. The condition $rM \ge E(1+\lambda)$ ensures that the exponent is nonnegative, so the bound is meaningful. ■

Appendix C. Choice of Accuracy Proxy

The utopia geometry requires a monotone map from loss to the unit square. Besides $a = e^{-\ell}$ we compute three alternatives on every evaluation: a reciprocal transform $1/(1+\ell)$, direct 0/1 accuracy, and a Brier-style score $1 - \overline{(p-y)^2}$. Across 7,212 runs the four proxies rank-correlate strongly (Table 4), and the optimal-reach trend of Section 6 replicates for the exp and reciprocal proxies (power-law $R^2 \approx 0.97$–0.99), confirming that the conclusions are not artifacts of the specific $e^{-\ell}$ choice. The Theorem 2 floor is claimed only for $a = e^{-\ell}$.

Table 4: Spearman rank correlations between the four performance proxies, computed per run ($n = 7{,}212$; all $p < 10^{-3}$).

	$e^{-\ell}$	reciprocal	accuracy	Brier
$e^{-\ell}$	1	0.998	0.823	0.941
reciprocal	0.998	1	0.830	0.943
accuracy	0.823	0.830	1	0.939
Brier	0.941	0.943	0.939	1

Appendix D. Theorem 1 Monotonicity Audit

Theorem 1 assumes $a_1(t)$ decreases through the y_2 phase. We audited every steady-state y_2 phase in the Boolean reach grid (56,748 phases). Using the endpoint-inclusive convention (the value just before the first y_2 update, followed by every post-update value), strict pointwise decrease held in 97.5% of phases; the weaker endpoint condition—a_1 largest at the phase start and smallest at its end, which is what the bound actually needs—held in 98.8%; and the median normalized positive variation was 0. This variation—the total within-phase increase in a_1 divided by its total decrease, and zero for a strictly decreasing phase—is sharply concentrated at 0 with a thin right tail: the non-monotone phases show only small, transient upticks in a_1 rather than sustained increases, which is why the endpoint condition holds even where strict pointwise monotonicity fails. Violations are mild and heterogeneous in r. The assumption is thus a close but not universal description of the dynamics.

Appendix E. Robustness Experiments

E.1. Held-out generalization on continuous inputs

To test performance on unseen inputs rather than the training truth table, we replace Boolean pairs with linear-separator tasks on isotropic Gaussian inputs in $\mathbb{R}^{10}$. Two unit

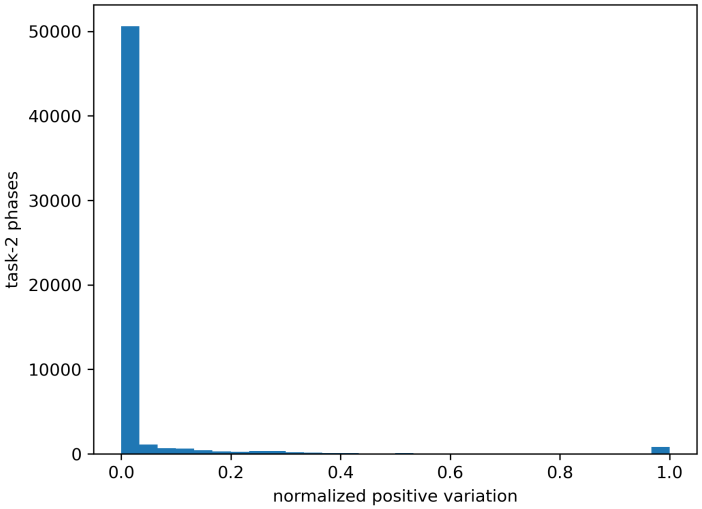

Figure 4: Monotonicity audit of Theorem 1. The y_2 phase is strictly decreasing in 97.5% of phases, and the weaker endpoint condition holds in 98.8%. The median normalized positive variation is 0, and the distribution is sharply concentrated at 0 with a heterogeneous right tail.

weight vectors w_1, w_2 are placed at angle $\theta = \pi r$, giving labels $y_j = \mathbb{I}[X \cdot w_j > 0]$ whose expected disagreement is exactly r, a continuous analogue of d_H/N. We draw independent train (4096), validation (2048), and test (4096) sets and evaluate on all splits. Over 454 runs the train–test gap in d_U is small (mean $|\Delta| = 0.007$, median 0.004, max 0.064), so the plasticity geometry is not a training-set memorization artifact. This grid is small and some small-model pretraining did not fully converge, so we treat it as a consistency check.

E.2. Optimizer and architecture

Table 5 sweeps the optimizer and the architecture while holding the reach sweep fixed. The reach-governed picture persists under minibatch SGD, momentum, and Adam. Varying width and depth over a 16× range in parameter count (3,641 to 59,041) leaves d_U essentially unchanged, giving direct empirical support for the network-size independence that our architecture-agnostic bounds assume but do not prove. Figure 5 shows the reach curves are preserved across architectures.

Table 5: Mean steady-state d_U (exp proxy) across optimizer and architecture variants, each with its own reach sweep over 9 task pairs and 5 seeds.

Study	**Variant**	**Params**	**Mean d_U**
Optimizer	full-batch SGD	15,121	0.565
	minibatch-16	15,121	0.539
	minibatch-32	15,121	0.548
	momentum 0.9	15,121	0.487
	Adam	15,121	0.456
Architecture	shallow (3×40)	3,641	0.626
	baseline (10×40)	15,121	0.568
	wide (10×80)	59,041	0.581

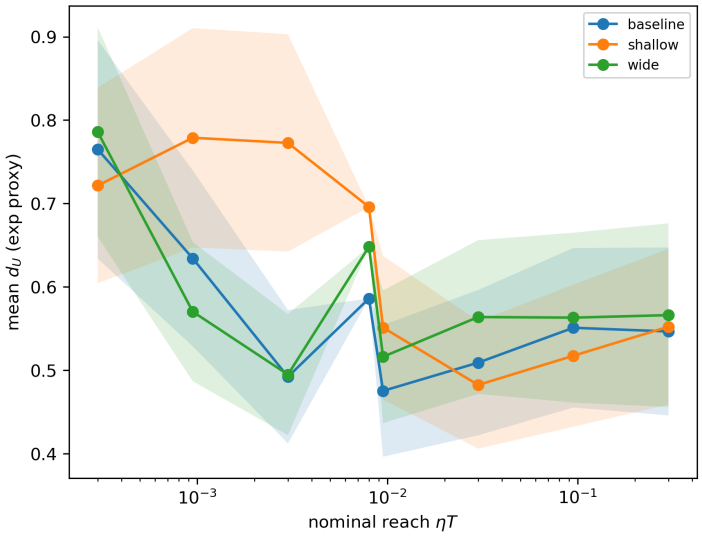

Figure 5: d_U versus nominal reach ηT for the three architectures; the optimum location and depth are preserved across a 16× parameter range.