Time Is All You Need: Temporal Translation and the Credit Assignment Problem

James Blight

Independent Researcher
james.blight64@gmail.com

Abstract

Learning algorithms assume meaningful input—state, context, relational structure. A continuous stream provides none of this. Before a system can learn from time, it must translate time into state. We argue that the central obstacle to embodied continual learning is not inadequate optimization but inadequate translation: credit assignment is ill-posed until temporal history has been rendered into a representation where responsibility is locally computable. We characterize the constraints on temporal translation—causal, bounded, continuous, locally interpretable—and show they admit essentially one solution class among linear, time-invariant, finite-state summaries: exponentially decaying measurements at geometrically spaced timescales. When spaced by the golden ratio to maximize incommensurability, this decomposition provides a minimal temporal language in which the past is present and credit assignment becomes tractable. We instantiate this in the Spectral Online Machine Architecture (SOMA), demonstrating continuous adaptation without catastrophic forgetting, under fixed resources, with no replay buffer and no sequence storage, reaching 1.87 bits per byte under streaming constraints with bounded memory. The architecture satisfies a requirement that theories of consciousness increasingly emphasize: temporal integration must be intrinsic to the system's state, not externalized as retrievable data. A system with translated time *is* its history; it does not merely *have* access to it. The failure of continual learning is not a failure of learning rules. It is a failure to give them a language for time.

The Translation Problem

Prediction as Substrate Arbitrage

The world we wish to predict—the three-dimensional world of molecules, bodies, and objects—operates in time. Mass produces inertia. Chemical reactions propagate at the speed of diffusion. Mechanical forces transmit at the speed of sound. Information in this world moves slowly, bound by the properties of the particular processes governing its dynamics. Any useful predictor must outrun this world. A prediction that arrives after the event it forecasts is not a prediction but a record. More precisely: useful prediction requires that, for the target dynamics, internal convergence time is less than external process time. The value of anticipation depends entirely on this inequality.

Biological intelligence exploits the fact that electrochemical signaling, while slow by electronic standards, is fast relative to mechanical and chemical processes governing an organism's environment. Artificial intelligence exploits a larger gap: electrons move faster than ions, silicon switches faster than synapses. The asymmetry between substrates is what makes prediction possible.

The Assumption Beneath Learning

This creates a translation problem. Every learning algorithm assumes its inputs have structure that learning can act on. Gradient descent assumes differentiable loss. Reinforcement learning assumes coherent states. Hebbian rules assume meaningful co-activation.

But there is a deeper assumption: that "the current input" is a well-formed object—that the system has access to context, to temporal relationships, to what just happened and what preceded it. This assumption is rarely examined because in standard settings (batched data, episodic tasks, stored sequences) it is satisfied by construction. For a system receiving a continuous stream—no resets, no episodes, no external memory—that assumption fails. The stream is undifferentiated. There is no intrinsic "current state." There is no "what has been happening" unless the system constructs it. Learning algorithms were not built to handle this. They were built for inputs that assume meaning is differentiable.

The Upstream Problem

We claim the difficulty of continual learning is not primarily about learning rules—it is about the representations they receive. The major approaches each fail upstream of learning: replay buffers store the raw stream, deferring translation indefinitely (Mnih et al. 2015); memory grows without bound. Backpropagation through time unrolls temporal structure spatially, translating post hoc rather than online (Werbos 1990). Attention mechanisms externalize time as queryable data, requiring storage proportional to context (Vaswani et al. 2017). Eligibility traces acknowledge something must be retained, but lack structure making responsibility precise (Sutton 1988). Regularization methods (EWC, etc.) protect old weights without translating new inputs—addressing forgetting by preventing learning (Kirkpatrick et al. 2017; French 1999). The learning rules may be sound; the representations they act upon are not.

The Problem, Precisely

A system embedded in time must translate the temporal stream into a representation that:

1. Is causal: Depends only on past and present
2. Is bounded: Does not grow with stream length
3. Is continuous: Updates incrementally without resets
4. Is locally interpretable: Supports learning using only current state

These constraints define embodied temporal existence. The question is what satisfies them.

What Can Be Retained from Infinity

The Constraint

An unbounded stream cannot be stored by a bounded system. The past must be compressed. The question is: what compression preserves what matters? The stream has no absolute reference—no zero, no origin, no canonical frame. The only quantities computable from within the stream are changes relative to itself.

Exponential Decay as Minimal Solution

Among linear, time-invariant (LTI), finite-state operators, the exponential moving average:

$$E(t) = (1 - \alpha) \cdot E(t - 1) + \alpha \cdot a(t) \tag{1}$$

is the irreducible basis of bounded causal linear memory. Any bounded, causal, stable LTI system decomposes spectrally into modes of this form; exponential traces are what remain when no further reduction is possible. Richer dynamics—including the nonlinear trace mechanisms present in biological systems—may satisfy the same constraints through different means, but they do not escape this decomposition at the linear limit. The LTI restriction is therefore a minimality claim: it identifies the simplest implementation.

Decomposition as Necessity

Structure at frequency ω requires a trace with timescale $\tau \sim 1/\omega$. A system that commits to one timescale is blind to others. An embodied system does not know in advance which frequencies matter. The robust solution is a bank of traces spanning the relevant range—a spectral decomposition projecting the stream onto a temporal frequency basis.

From Measurement to Language

Geometric Spacing and the Resonance Problem

Scale invariance argues for geometric timescale spacing: $\tau_k = \tau_0 \cdot r^k$. This ensures equal representation per octave—no privileged scale (Shankar and Howard 2012).

However, if r is rational (e.g., $r = 2$), timescales have harmonic relationships: structure at one scale creates artifacts at related scales. For faithful encoding, timescales must be incommensurate.

The golden ratio $\phi = (1 + \sqrt{5})/2$ has continued fraction $[1; 1, 1, \ldots]$. Its rational approximations converge more slowly than any other irrational—ϕ is "most irrational" in a precise sense (Appendix A). Timescales spaced by ϕ are therefore maximally incommensurate: no resonance, no harmonics, minimal crosstalk.

Given a fixed maximum observable window (slowest timescale $\tau_{\max}$ decaying over a prescribed number of steps), the ratio r directly controls spectral coverage. The bandpass features $B_k = E_{\cdot,k} - E_{\cdot,k+1}$ isolate frequency bands between successive cutoffs. If $r > \phi$, intermediate bands become under-sampled; spectral structure at those scales is irretrievably lost in the trace bank. Filling the window with golden-ratio spaced traces ($r = \phi$) provides a complete decomposition without such loss. Using a smaller ratio ($r < \phi$) adds redundant information, making the golden ratio the sparsest spacing that preserves all resolvable frequencies under the LTI constraints for any given resource budget.

This is not numerology. Given no prior on signal structure except continuity and boundedness, and a fixed trace budget for a given window, golden ratio spacing is the minimax solution to self-interference under LTI constraints—a principled uninformed default that simultaneously avoids spectral loss and harmonic artifacts. Optimal spacing is ultimately data-dependent, however ϕ is what the absence of prior knowledge recommends.

The Temporal Language

With K traces at decay rates $\alpha_k = 1/\phi^k$, the system possesses a vocabulary:

- Fast traces ($k \approx 0$): What is happening now, what just changed
- Medium traces: What has been happening recently
- Slow traces ($k \approx K$): What usually happens, the background

The trace vector $E(t)$ is a projection preserving temporal structure within the constraints of bounded existence. After translation, every input arrives contextualized. The system never sees raw data; it sees data embedded in temporal structure. Context is not retrieved. It is constitutive.

Translation as Temporal Differentiation

There is a deeper connection. Each exponential trace is a low-pass filter; differences between adjacent traces isolate frequency bands. Geometrically spaced traces thus perform a discrete wavelet-like decomposition, computing temporal derivatives at each resolvable scale. This reframes the relationship to gradient descent: we compute gradients in time rather than parameter space, and only at resolutions the signal admits. Temporal gradients, once computed, make credit assignment tractable.

Translation Enables Credit Assignment

The Reframe

Credit assignment—which parameters caused an outcome—has been treated as the hard problem (Minsky 1961). We claim it is downstream of the hard problem. Credit assignment is ill-posed on raw streams. There is no basis for

"which weights contributed" when the system has no representation of what happened. Credit assignment becomes straightforward once translation provides legible representations.

Contribution as Readable

A unit computing:

$$s = \sum_i \sum_k w_{i,k} \cdot E_i^{(k)} \tag{2}$$

has contribution from each weight:

$$c_{i,k} = w_{i,k} \cdot E_i^{(k)} \tag{3}$$

Weights encode "how much to listen"; traces encode "what there is to hear." The product is contribution. From the system's perspective, this is not approximation—it is the complete causal story available within the informational bounds on which it acts. The system has no access to counterfactuals; contribution *is* causation as far as the model can know.

The local readability of contributions — $c_{i,k} = w_{i,k} \cdot E_i^{(k)}$ — is not an incidental property. Any system that maintains a bounded, causally distinct existence relative to its environment must solve the temporal translation problem as a constitutive requirement: it must map unbounded external history into a fixed-dimensional internal state that remains conditionally independent of unsummarized external variables given the current sensory and active states. In the framework of the Free Energy Principle, this requirement is formalized by the Markov blanket, which enforces that internal dynamics are causally mediated solely through the blanket states at each timestep.

Credit assignment, in this light, amounts to inferring the causal efficacy of internal parameters with respect to outcomes observable across the blanket. For such inference to be well-posed and local, the internal state must encode a sufficient statistic of past causes in a form that preserves predictive relevance while discarding extraneous temporal detail. The exponential trace bank with geometrically spaced timescales performs exactly this transformation: it projects the temporal stream onto a basis where relevant causes at multiple scales are intrinsically present in the current state, rendering responsibility computable from local quantities alone and ensuring causal closure within the system's intrinsic dynamics.

The Learning Rule Becomes Weakly Constrained

Once translation provides locally computable responsibility, the learning rule is weakly constrained. It must respect locality and boundedness—but within those constraints, any update that increases weights proportional to their responsibility for correct predictions, and decreases them for errors, will work. The translation does the heavy lifting; the learning rule exploits what translation provides.

This is our central claim: the learning problem is solved upstream. The architecture admits experimentation in the learning rule; the translation does not.

Biological Foundations

The trace bank is not analogy to neurobiology, but a derivation from it. The architecture emerges from asking: how do biological neurons perform temporal integration under the same constraints we have identified?

Neurotransmitters as Timescale-Specific Traces

Every excitatory neurotransmitter in the brain is a trace that decays at a characteristic rate. Glutamate operates on millisecond timescales—fast, immediate signaling. Acetylcholine persists longer, modulating attention over seconds. Dopamine signals over tens of seconds, encoding slower dynamics. Serotonin operates slower still, shaping mood and background state over minutes to hours.

These are not arbitrary biochemical facts. They constitute a multi-timescale decomposition of neural input—precisely the trace bank structure we have derived from first principles. The diversity of neurotransmitter kinetics is the translation layer (Buonomano and Maass 2009).

Receptors as Weights

Receptors are literal weights encoding how much to attend to each timescale. A neuron's receptor profile determines its sensitivity to fast versus slow dynamics. Receptor density is plastic—it changes with experience—while neurotransmitter kinetics are fixed. This is the same separation we implement: trace dynamics are fixed by design; only the mapping from traces to output is learned.

Inhibition as Boundary Detection

GABA, the primary inhibitory neurotransmitter, operates on fast timescales matching glutamate. Its role is not merely to decrease activity but to enable boundary detection—to allow the system to negate information, marking where one pattern ends and another begins. Typical learning rules permit signed weights, providing inhibition by construction. Biology, however, cannot signal using the absence of matter, and offers a minimal alternative solution.

Error as Metabolic

The error signal in biological learning is largely metabolic. Endocannabinoids are retrosignaled from postsynaptic to presynaptic neurons, damping activity proportional to downstream metabolic cost (Wilson and Nicoll 2002; Castillo et al. 2012). The presynaptic neuron is literally weakened by the excess firing it causes—activity is regulated by the metabolic cost of thought. This provides local error attribution without requiring backpropagated gradients.

Prediction error and metabolic cost are congruent signals expressed in different currencies: informational surprise in one domain manifests as thermodynamic expenditure in the other. Unresolved prediction error necessitates additional computation (belief updating, resource recruitment) that incurs real energetic cost. In this light, energy and information are measured in the same unit across time: minimizing prediction error is minimizing the metabolic demand of maintaining an adaptive, low-surprise exchange with the world.

The Spectral Online Machine Architecture

SOMA instantiates these principles. It is an existence proof: once temporal translation is solved, embodied continual learning works. The separation of memory and skill enables predictable learning over unbounded time.

Trace Bank

A trace bank maintains K exponential traces per input symbol at geometrically spaced decay rates $\alpha_k = 1/r^k$ for $k \in \{0, \ldots, K-1\}$, where r is the geometric base. The base may be set to ϕ (the uninformed optimum) or computed from a desired temporal window: $r = W^{1/(K-1)}$ for maximum window W. On observing byte b at time t:

$$E_{b,k} \leftarrow (1 - \alpha_k) \cdot E_{b,k} + \alpha_k \tag{4}$$

$$E_{i \neq b,k} \leftarrow (1 - \alpha_k) \cdot E_{i,k} \tag{5}$$

Adjacent trace differences isolate frequency bands—bandpass features: $B_k = E_{\cdot,k} - E_{\cdot,k+1}$ for $k < K-1$, and $B_{K-1} = E_{\cdot,K-1}$ for the slowest band. The full feature vector $\text{vec}(B) \in \mathbb{R}^{256 \times K}$ is the intrinsic state representation: a bounded, causal, continuously updated sufficient statistic of recent history.

Architecture

The flattened feature vector feeds a two-layer network. A hidden layer $U \in \mathbb{R}^{H \times 256K}$ projects trace features to a hidden representation:

$$h = \text{ReLU}(U \cdot \text{vec}(B)) \tag{6}$$

The hidden representation is budget-normalized before the output projection:

$$\hat{h} = h \cdot \frac{\beta}{\|h\|_1 + \varepsilon} \tag{7}$$

where β is a fixed budget. This enforces bounded competition: total hidden activation is conserved, attending more to one feature requires attending less to another. An output matrix $W \in \mathbb{R}^{256 \times H}$ then produces logits:

$$p(y_t \mid \text{history}) = \text{softmax}(W \cdot \hat{h}) \tag{8}$$

A direct residual connection $W_d \in \mathbb{R}^{256 \times 256K}$ projects trace features to logits, enabling the model to exploit linear predictability and bypass hidden nonlinearity where advantageous. All weight matrices (U, W, W_d) are row-normalized to fixed ℓ_2 norms throughout training, and hidden activations are subject to budget normalization, ensuring bounded competition and preventing magnitude-driven feature dominance.

Experimentation across a range of learning rules has demonstrated the continual-learning property resides principally in the trace bank combined with weight bounding. Provided these, a linear readout over the trace features alone is sufficient to sustain continuous adaptation without interference. A hidden layer adds representational capacity and expressivity.

The reported configuration and results reflect the optimal instantiation under the imposed constraints achieved to date. Temporal translation lends itself to extensive experimentation in the learning rule.

Decimation

Each frequency band is sampled and updated at its Nyquist rate—fast bands every step, slow bands proportionally less often. Decimation factor $d_k = 2^{\lfloor \log_2(r^k/2) \rfloor}$ ensures each band receives gradient at the rate it can resolve, and no faster. Slow features encoding long-term statistics are not corrupted by high-frequency gradient noise. Gradient is accumulated within decimation groups and applied when each group's period elapses.

The decimation structure admits a natural training curriculum. At high downsample factors, only slow bands are active—the system processes the stream at coarse temporal resolution with proportionally higher throughput. At downsample 1, all bands are active and the system sees every transition.

What SOMA Is Not

Not BPTT. Backpropagation through time requires storage proportional to sequence length. SOMA has no temporal recurrence in learnable parameters. Traces evolve by a fixed rule; the network is feedforward on trace features.

Not an RNN. Recurrent networks transform hidden state through learned parameters at each step, creating temporal dependencies in the gradient graph. In SOMA, temporal state evolves independently of learned parameters.

Not a state space model. SSMs learn the dynamics governing state evolution. SOMA's trace dynamics are fixed by design—only the mapping from traces to predictions is learned.

Demonstrations

We report results under streaming constraints: no replay buffer, no stored sequences, bounded memory, no episode resets. We demonstrate that learning is possible, and catastrophic forgetting absent, under conditions that any temporally embedded system must satisfy.

Language Modeling

Typical language models report a single, final loss. SOMA does not have this constraint. Relevant comparisons remain thin given the sparsity of solutions to stable learning at scale under strict streaming constraints. We report results over a single pass of enwik9, and ongoing loss up to 2.5B bytes.

Robustness Across Distribution Shifts

The training run underlying Table 1 included multiple uncontrolled corpus switches. On each shift, loss spiked on the new distribution, then recovered; upon return to enwik9, performance quickly returned to prior levels, and learning continued. Old knowledge faded at trace-determined rates rather than via catastrophic overwriting—evidence that memory and skill remain architecturally separated. Interruptions and resampling produced no degradation in subsequent learning trajectory.

Figure 1 shows SOMA adaptation to adversarial distribution shift. When trained on one distribution then exposed to another, performance on the new distribution improves while performance on the original degrades gracefully, at

Bytes seen	Loss (nats)	bpb	Acc.
1M	2.673	3.86	32.5%
10M	2.137	3.08	42.1%
100M	1.800	2.60	49.8%
250M	1.648	2.38	54.7%
500M	1.504	2.17	57.8%
750M	1.393	2.01	60.9%
∼2.5B	1.293	1.87	63.8%

Table 1: SOMA (K=512 traces, H=4096 hidden, geometric base 1.0447 for max window 5B bytes, 537.9M parameters, single RTX 5090) trained under streaming constraints over 2.5B total bytes, including two passes of enwik9. Each byte processed sequentially; no replay buffer, no stored sequences.

Method	bpb	Memory	Learning
Streaming bigram	3.89	Fixed	Stops
Streaming gzip (32KB)	2.92	Fixed	None
Order-5 Markov + backoff	2.95	Growing	Continues
SOMA (K=512, H=4096)	**2.60**	Fixed	Continues

Table 2: Bounded-memory comparison at 100M bytes on enwik8. Bigram and gzip use fixed memory but cannot continue learning. Order-5 Markov continues learning but requires unbounded growing memory. SOMA is the only method with both fixed memory and continuing improvement.

rates determined by trace timescales. Re-exposure to the original distribution yields faster relearning than initial acquisition.

Waveform Prediction

Figures 2 and 3 show waveform prediction after 500 and 10,000 steps respectively. The system maps plausible value regions at progressively finer resolution, first capturing slow structure then filling in detail. This demonstrates multi-scale learning: simultaneous adaptation at different temporal resolutions without interference.

Method	Time/step	Memory	Learning
Transformer	$O(nL^2)$	$O(L)$ grows	stops
RNN + BPTT	$O(nT)$	$O(T)$ grows	stops
Replay buffer	$O(n)$	$O(B)$ grows	continues
EWC	$O(n)$	$O(n)$ fixed	degrades
SOMA	$O(n)$	$O(n)$ fixed	continues

Table 3: Complexity comparison. n: parameters, L: context length, T: BPTT depth, B: buffer size. Only SOMA achieves $O(n)$ time per step, fixed memory, and continuous improvement without catastrophic forgetting.

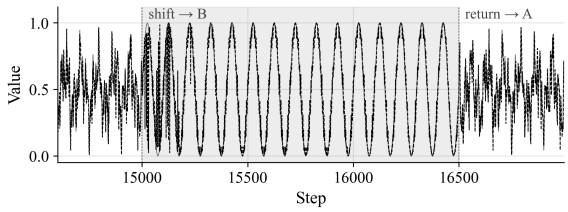

Figure 1: Adaptation to distribution shift. New statistics accumulate while old patterns fade at trace-determined rates. Re-exposure yields faster acquisition than initial learning.

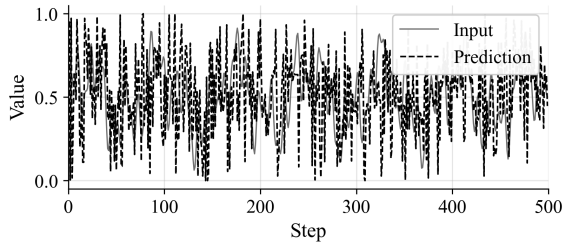

Figure 2: SOMA waveform prediction after 500 steps. Slow traces establish broad structure; fast traces have not yet accumulated sufficient statistics. Resolution emerges hierarchically.

Temporal Translation and the Free Energy Principle

The Free Energy Principle (FEP) states that self-organizing systems minimize variational free energy—an upper bound on surprise—by maintaining a generative model that is continuously updated through prediction error (Friston 2010, 2009):

$$\mathcal{F} = \mathbb{E}_{q(s)}[\log q(s) - \log p(o, s)] \geq -\log p(o). \quad (9)$$

At the heart of FEP is the Markov blanket: the partition of states into internal states μ, external states η, sensory states s, and active states a. The blanket $\mathcal{B} = \{s, a\}$ enforces conditional independence at every timestep:

$$p(\mu_t \mid s_t, a_t, \eta_t) = p(\mu_t \mid s_t, a_t). \quad (10)$$

Friston's claim is constitutive: a system exists as a bounded entity only if it actively maintains this boundary.

For the blanket to hold continuously, internal dynamics must themselves be Markovian with respect to the current blanket:

$$\mu_t = f(\mu_{t-1}, s_t, a_t). \quad (11)$$

Any reconstruction of μ_t from external memory $\mathcal{M}$ (stored sequences, key-value caches, replay buffers) introduces a bypassing channel $\eta \to \mu_t$, violating the blanket. The system no longer maintains a boundary; it simulates one on demand. Catastrophic interference is the observable signature of this structural failure.

Under bounded physical capacity, the only solutions to the recursive filter $\mu_t = f(\mu_{t-1}, s_t)$ are (in the linear time-invariant case) banks of exponentially decaying traces. The

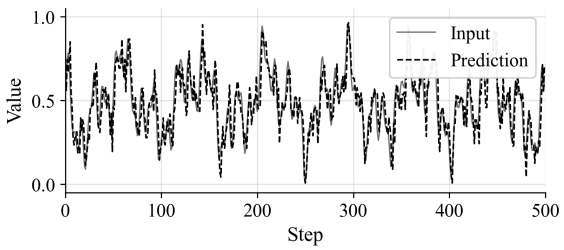

Figure 3: SOMA waveform prediction after 10,000 steps. Fine structure has filled in as medium and fast traces converge. Multiple resolutions predicted simultaneously, with no interference between timescales.

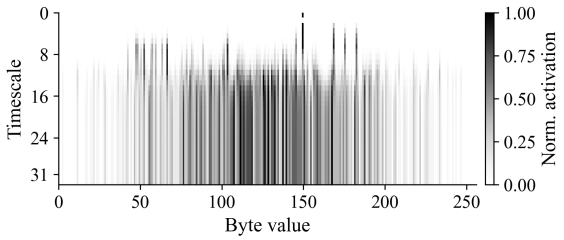

Figure 4: Trace bank visualization. Each row is a timescale; columns are input symbols. Fast traces (top) capture immediate structure; slow traces (bottom) encode long-run statistics. The multi-scale decomposition is the temporal language.

loss minimized by SOMA at each step,

$$\mathcal{L}_t = -\log p(y_t \mid E(t)), \tag{12}$$

is precisely the surprise term of variational free energy evaluated on an intrinsically maintained blanket state.

Temporal translation is not an architectural option but a necessary condition for any system that genuinely satisfies the FEP while embedded in time. All current large-scale sequence models fail this condition by externalizing history. SOMA satisfies it by construction, and provides a minimal existence proof that the FEP's conditional demands can be met in a practical architecture.

We do not claim SOMA is conscious—its scale is far too modest. We claim that coherent, intrinsic temporal translation is a binary prerequisite for any system that maintains a continuous Markov blanket over unbounded time. Whether it is sufficient for consciousness remains open, however the distinction constitutes a principled criterion for the search space of candidate systems.

Conclusion

Learning algorithms were given an impossible task: reason about time without a language for it. The solutions proposed—store more, compute backward, protect old weights—are workarounds for a missing translation layer. We have shown that temporal translation takes a specific form under embodied constraints: exponential traces at geometrically spaced timescales, providing a spectral decomposition where the past is present and credit assignment is locally computable. Once translation exists, the learning rule is weakly constrained. Standard gradient descent suffices.

SOMA demonstrates that this is sufficient for continuous adaptation with fixed resources, no replay, and no catastrophic forgetting. The biological parallels are not metaphorical: neurotransmitter kinetics, receptor plasticity, and retrograde metabolic signaling instantiate the same decomposition we derived from first principles. Translated time appears convergent among intelligent, embodied learners.

Whether this is sufficient for consciousness we cannot say. That it is necessary for continual, embodied learning, we have shown.

Time is all you need.

Optimality of Golden Ratio Spacing

The golden ratio $\phi = \frac{1+\sqrt{5}}{2}$ has continued fraction $[1; 1, 1, \ldots]$. Among irrationals, its rational approximations converge most slowly: $|\phi - p/q| > 1/(\sqrt{5} \cdot q^2)$, the tightest Hurwitz bound (Hurwitz 1891; Khinchin 1964). For timescale spacing, this makes ϕ^k maximally incommensurate with ϕ^j ($k \neq j$). Near-integer relationships are maximally suppressed under bounded rational approximation; resonance is minimized. Given uncertainty about signal structure, golden ratio spacing is the minimax solution under LTI constraints. Optimal spacing is data-dependent; ϕ is what the absence of prior knowledge recommends.

Complete Algorithm Specification

TraceBank

For each of 256 input symbols, maintain K exponential traces at decay rates $\alpha_k = 1/r^k$ for $k \in \{0, \ldots, K-1\}$, where r is the geometric base. The base may be set to ϕ (the uninformed optimum) or computed from a desired temporal window: $r = W^{1/(K-1)}$ for maximum window W.

Trace update for observed byte b:

$$E_{b,k} \leftarrow (1 - \alpha_k) \cdot E_{b,k} + \alpha_k \qquad \text{(observed)} \tag{13}$$

$$E_{i,k} \leftarrow (1 - \alpha_k) \cdot E_{i,k} \qquad \text{(others)} \tag{14}$$

Bandpass features:

$$B_k = E_{\cdot,k} - E_{\cdot,k+1} \qquad \text{for } k < K-1 \tag{15}$$

$$B_{K-1} = E_{\cdot,K-1} \qquad \text{(lowest band)} \tag{16}$$

Decimation Schedule

Each band k is sampled and updated at its Nyquist rate: decimation factor $d_k = 2^{\lfloor \log_2(r^k/2) \rfloor}$, clipped to a minimum of 1. Bands are grouped by decimation factor; gradient is accumulated within groups and applied when the group's decimation period elapses.

Prediction

$$h = \text{ReLU}(U \cdot \text{vec}(B)), \quad U \in \mathbb{R}^{H \times 256K} \tag{17}$$

$$\hat{h} = h \cdot \frac{\beta}{\|h\|_1 + \varepsilon} \tag{18}$$

$$p(y_t \mid \text{history}) = \text{softmax}(W \cdot \hat{h}), \quad W \in \mathbb{R}^{256 \times H} \tag{19}$$

Optional direct residual: logits $= W\hat{h} + W_d \cdot \text{vec}(B)$, $W_d \in \mathbb{R}^{256 \times 256K}$.

Both U and W are row-normalized throughout training to fixed ℓ_2 norm. Weight decay ($\lambda = 10^{-4}$) is applied after each update. Updates are clipped to a maximum fractional change per step.

References

Buonomano, D. V.; and Maass, W. 2009. State-Dependent Computations: Spatiotemporal Processing in Cortical Networks. *Nature Reviews Neuroscience*, 10(2): 113–125.

Castillo, P. E.; Younts, T. J.; Chávez, A. E.; and Hashimotodani, Y. 2012. Endocannabinoid Signaling and Synaptic Function. *Neuron*, 76(1): 70–81.

French, R. M. 1999. Catastrophic Forgetting in Connectionist Networks. *Trends in Cognitive Sciences*, 3(4): 128–135.

Friston, K. 2009. The Free-Energy Principle: A Rough Guide to the Brain? *Trends in Cognitive Sciences*, 13(7): 293–301.

Friston, K. 2010. The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, 11(2): 127–138.

Hurwitz, A. 1891. Über die Angenäherte Darstellung der Irrationalzahlen durch Rationale Brüche. *Mathematische Annalen*, 39(2): 279–284.

Khinchin, A. Y. 1964. *Continued Fractions*. University of Chicago Press.

Kirkpatrick, J.; et al. 2017. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

Minsky, M. 1961. Steps Toward Artificial Intelligence. *Proceedings of the IRE*, 49(1): 8–30.

Mnih, V.; et al. 2015. Human-Level Control Through Deep Reinforcement Learning. *Nature*, 518(7540): 529–533.

Shankar, K. H.; and Howard, M. W. 2012. A Scale-Invariant Internal Representation of Time. *Neural Computation*, 24(1): 134–193.

Sutton, R. S. 1988. Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, 3(1): 9–44.

Vaswani, A.; et al. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Werbos, P. J. 1990. Backpropagation Through Time: What It Does and How to Do It. *Proceedings of the IEEE*, 78(10): 1550–1560.

Wilson, R. I.; and Nicoll, R. A. 2002. Endocannabinoid Signaling in the Brain. *Science*, 296(5568): 678–682.