Embeddings and the Atomic Time Step

Hutter RNN Project

January 31, 2026

1 The Atomic Time Step

The Universal Model update:

$$t_{n+1} = p \times t_n$$

Each application of pattern p advances thought by one atomic timestep. This is the fundamental "tick" of the system.

1.1 Cost Per Tick

Each tick consumes entropy:

$$H_{\text{tick}} = -\log_2 p_{\text{avg}}$$

After n ticks:

$$H_{\text{total}} = n \cdot H_{\text{tick}}$$

The depth limit:

$$n_{\max} = \frac{24}{H_{\text{tick}}}$$

2 Embeddings

An embedding maps events to vectors:

$$\phi : E \to \mathbb{R}^d$$

The dimension d is the number of basis functions—like Fourier coefficients.

2.1 Thought as Distribution

Thought $t \in T$ is a distribution over events:

$$t : E \to [0, 255] \quad \text{(log-support values)}$$

In probability form:

$$p(e) = \frac{\exp(t(e))}{\sum_{e'} \exp(t(e'))}$$

2.2 Embedding the Thought

Project thought onto d basis functions:

$$h = \phi(t) \in \mathbb{R}^d$$

This is like taking d Fourier coefficients of the distribution t.

3 Time Evolution in Embedding Space

3.1 Abstract Form

In thought space:

$$t_{n+1} = p \times t_n$$

3.2 Embedded Form

In embedding space:

$$h_{n+1} = W \cdot h_n$$

The pattern p becomes a matrix $W \in \mathbb{R}^{d \times d}$.

3.3 The RNN Case

For an Elman RNN:

$$h_{n+1} = \tanh(W_{hh} \cdot h_n + W_{ih} \cdot x_n)$$

Here:

- W_{hh} is the embedded pattern (time evolution)
- W_{ih} incorporates new input x_n
- tanh is normalization (keeps values bounded)

Without input ($x_n = 0$) and for small values ($\tanh \approx \mathrm{id}$):

$$h_{n+1} \approx W_{hh} \cdot h_n$$

This is exactly $t_{n+1} = p \times t_n$ in the embedding.

4 Embedding Dimension as Bandwidth

4.1 Fourier Analogy

Fourier	**Embedding**
d frequency components	d hidden dimensions
Basis functions $e^{2\pi ikt}$	Basis vectors in $\mathbb{R}^d$
Bandwidth $= d \cdot f_0$	Representable complexity

4.2 What d Controls

Higher d:

- More frequencies / finer resolution
- More complex patterns representable
- More parameters (W is $d \times d$)
- More energy cost

Lower d:

- Fewer frequencies / coarser resolution
- Simpler patterns only
- Fewer parameters
- Lower energy cost

4.3 SVD and Dimension

From our SVD analysis:

- Component 0: 97.5% of variance (baseline frequency)
- Components 1–5: 2.5% (interpretable structure)
- Components 6–63: diminishing returns
- Components 64–255: noise

Truncating to rank-64 keeps the signal, discards the noise. The "bandwidth" of English bigrams is approximately 64 dimensions.

5 The Full Picture

Level	**Object**	**Time Step**
Events	$e \in E$	—
Thought	$t \in T = [0, 255]^{\lvert E \rvert}$	$t_{n+1} = p \times t_n$
Embedding	$h \in \mathbb{R}^d$	$h_{n+1} = W \cdot h_n$
RNN	$h \in \mathbb{R}^d$	$h_{n+1} = \tanh(W_{hh}h_n + W_{ih}x_n)$

The atomic time step $t_{n+1} = p \times t_n$ is the same operation at every level, just in different representations.

6 Back to Empirical

6.1 Pattern Injection Revisited

We injected bigram statistics into W_{ih} and W_{ho} via SVD:

$$P^T = U \cdot S \cdot V^T \implies W_{ho} = U\sqrt{S}, \quad W_{ih} = \sqrt{S}V^T$$

This embeds the pattern P (bigram log-probs) into the RNN weights.
Result: 1 bit/char head start (5.46 → 4.47 bpc).

6.2 Why W_{hh} Injection Fails

W_{hh} encodes the time evolution—how to carry information from h_n to h_{n+1}.

This requires patterns that span multiple ticks. But:

$$\text{bits per pattern} = k \cdot H_{\text{tick}}$$

For k-step patterns, we need $k \cdot H_{\text{tick}}$ bits of precision.

Trigrams ($k = 2$) at $H \approx 2$ bits/char need 4 bits per pattern. Seems fine.

But: the patterns must be *carried* through W_{hh} multiplication, which accumulates precision loss. After a few steps, the signal is washed out.

6.3 Prediction

Effective memory depth of RNN:

$$d_{\text{memory}} \approx \frac{24}{H_{\text{avg}}} \approx \frac{24}{2} = 12 \text{ steps}$$

For English text at $\sim$2 bits/char, the RNN should effectively "remember" about 12 characters back.

This is testable: probe how RNN predictions depend on context at various distances.

7 Summary

- The atomic time step $t_{n+1} = p \times t_n$ becomes $h_{n+1} = W \cdot h_n$ in embedding space
- Embedding dimension d = bandwidth = number of frequencies
- SVD shows English bigrams have effective dimension $\sim$64
- W_{hh} carries information through time, limited by precision
- Predicted memory depth: $\sim$12 characters for English